Journal of Machine Learning Research 17 (2016) 1-52 Submitted 9/14; Revised 5/15; Published 12/16

Learning Latent Variable Models by Pairwise Cluster Comparison: Part I – Theory and Overview

Nuaman Asbeh ASBEH@POST.BGU.AC.IL
Department of Industrial Engineering and Management
Ben-Gurion University of the Negev
Beer Sheva, 84105, Israel

Boaz Lerner BOAZ@BGU.AC.IL
Department of Industrial Engineering and Management
Ben-Gurion University of the Negev
Beer Sheva, 84105, Israel

Editors: Isabelle Guyon and Alexander Statnikov

Abstract

Identification of latent variables that govern a problem and the relationships among them, given measurements in the observed world, are important for causal discovery. This identification can be accomplished by analyzing the constraints imposed by the latents in the measurements. We introduce the concept of *pairwise cluster comparison* (PCC) to identify causal relationships from clusters of data points and provide a two-stage algorithm called *learning PCC* (LPCC) that learns a latent variable model (LVM) using PCC. First, LPCC learns exogenous latents and latent colliders, as well as their observed descendants, by using pairwise comparisons between data clusters in the measurement space that may explain latent causes. Since in this first stage LPCC cannot distinguish endogenous latent non-colliders from their exogenous ancestors, a second stage is needed to extract the former, with their observed children, from the latter. If the true graph has no serial connections, LPCC returns the true graph, and if the true graph has a serial connection, LPCC returns a pattern of the true graph. LPCC's most important advantage is that it is not limited to linear or latent-tree models and makes only mild assumptions about the distribution. The paper is divided in two parts: Part I (this paper) provides the necessary preliminaries, theoretical foundation to PCC, and an overview of LPCC; Part II formally introduces the LPCC algorithm and experimentally evaluates its merit in different synthetic and real domains. The code for the LPCC algorithm and data sets used in the experiments reported in Part II are available *online*.

Keywords: causal discovery, clustering, learning latent variable model, multiple indicator model, pure measurement model

1. Introduction

Latent (unmeasured, hidden, unrecorded) variables, as opposed to observed (measured, manifest, recorded) variables, cannot usually be observed directly in a domain but only inferred from other observed variables or indicators (Spirtes, 2013). Concepts such as "quality of life," "economic stability," "gravitational fields," and "psychological stress"

play a key role in scientific theories and models, and yet such entities are latent (Klee, 1997).

Sometimes, latent variables correspond to aspects of physical reality that could, in principle, be measured but may not be for practical reasons, for example, "quarks". In this situation, the term hidden variables is commonly used, reflecting the fact that the variables are "really there", but hidden. On the other hand, latent variables may not be physically real but instead correspond to abstract concepts such as "psychological stress" or "mental states". The terms hypothetical variables or hypothetical constructs may be used in these situations.

Latent variable models (LVMs) represent latent variables and the causal relationships among them to explain observed variables that have been measured in the domain. These models are common and essential in diverse areas, such as in economics, social sciences, psychology, natural language processing, and machine learning. Thus, they have recently become the focus of an increasing number of studies. LVMs reduce dimensionality by aggregating (many) observed variables into a few latent variables, each of which represents a "concept" explaining some aspects of the domain that can be interpreted from the data. Latent variable modeling is a century-old enterprise in statistics. It originated with the work of Spearman (1904), who developed factor analytic models for continuous variables in the context of intelligence testing.

Learning an LVM exploits values of the measured variables as manifested in the data to make an inference about the causal relationships among the latent variables and to predict the value of these variables. Statistical methods for learning an LVM, such as factor analysis, are most commonly used to reveal the existence and influence of latent variables. Although these methods effectively reduce dimensionality and may fit the data reasonably well, the resulting models might not have any correspondence to real causal mechanisms (Silva et al., 2006). On the other hand, the focus of learning Bayesian networks (BNs) is on causal relations among observed variables, whereas the detection of latent variables and the interrelations among themselves and with the observed variables has received little attention. Learning an LVM using Inductive Causation* (IC*) (Pearl, 2000; Pearl and Verma, 1991) and Fast Causal Inference (FCI) (Spirtes et al., 2000) returns partial ancestral graphs, which indicate for each link whether it is a (potential) manifestation of a hidden common cause for the two linked variables. The structural EM algorithm (Friedman, 1998) learns a structure using a fixed set of previously given latents. By searching for "structural signatures" of latents, the FindHidden algorithm (Elidan et al., 2000) detects substructures that suggest the presence of latents in the form of dense subnetworks. Elidan and Friedman (2001) give a fast algorithm for determining the cardinality – the number of possible states – of latent variables introduced this way. However, Silva et al. (2006) suspected that FindHidden cannot always find a pure measurement sub-model,[1] which is a flaw in causal analysis. Also, the recovery of latent trees of binary and Gaussian variables has been suggested (Pearl, 2000). Hierarchical latent class (HLC) models, which are rooted trees where the leaf nodes are observed while all other nodes are latent, were proposed for the clustering of categorical data (Zhang, 2004). Two greedy algorithms are suggested (Harmeling

[1] A pure measurement model contains all graph variables and all and only edges directed from latent variables to observed variables, where each observed variable has only one latent parent and no observed parent.

and Williams, 2011) to expedite learning of both the structure of a binary HLC and the cardinalities of the latents. The BIN-G algorithm determines both the structure of the tree and the cardinality of the latent variables in a bottom-up fashion. The BIN-A algorithm first determines the tree structure using agglomerative hierarchical clustering and then determines the cardinality of the latent variables in the same manner as the BIN-G algorithm. Latent-tree models are also used to speed approximate inference in BN, trading the approximation accuracy with inferential complexity (Wang et al., 2008).

Models in which multiple latents may have multiple indicators (observed children), also known as multiple indicator models (MIMs) (Bartholomew et al., 2002; Spirtes, 2013), are a very important subclass of structural equation models (SEM), which are widely used, together with BNs, in applied and social sciences to analyze causal relations (Pearl, 2000; Shimizu et al., 2011). For these models, and others that are not tree-constrained, most of the mentioned algorithms may lead to unsatisfactory results. This is one of the most difficult problems in machine learning and statistics since, in general, a joint distribution can be generated by an infinite number of different LVMs. However, an algorithm that fills the gap between learning latent-tree models and learning MIMs is BuildPureClusters (BPC; Silva et al., 2006). BPC searches for the set (an equivalence class) of MIMs that best matches the set of vanishing tetrad differences (Scheines et al., 1995), but is limited to linear models (Spirtes, 2013).

In this study, we make another attempt in this direction and target the goal of Silva et al. (2006), but concentrate on the discrete case, rather than on the continuous case dealt with BPC. Towards this mission, we borrow ideas and principles of clustering and unsupervised learning. Interestingly, the same difficulty in learning MIMs is also faced in the domain of unsupervised learning that confronts similar questions such as: (1) How many clusters are there in the observed data? and (2) Which classes do the clusters really represent? Due to this similarity, our study suggests linking the two domains – learning a causal graphical model with latent variables and clustering analysis. We propose a concept and an algorithm that combine learning causal graphical models with clustering. According to the *pairwise cluster comparison* (PCC) concept, we compare pairwise clusters of data points representing instantiations of the observed variables to identify those pairs of clusters that exhibit major changes in the observed variables due to changes in their ancestor latent variables. Changes in a latent variable that are manifested in changes in the observed variables reveal this latent variable and its causal paths of influence in the domain. Using the *learning PCC* (LPCC) algorithm, we learn an LVM. We identify PCCs and use them to learn latent variables – exogenous and endogenous (the latter may be either colliders or non-colliders) – and their causal interrelationships as well as their children (latent variables and observed variables) and causal paths from latent variables to observed variables.

This paper is the first of two parts that introduce, describe, and evaluate LPCC. In this paper (Part I), we provide its foundations and theoretical infrastructure, from preliminaries to a broad overview of the PCC concept and LPCC algorithm. In the second paper (Part II), we formally introduce the two-stage LPCC algorithm, which implements the PCC concept, and evaluate LPCC, in comparison to state-of-the-art algorithms, using simulated and real-world data sets. The outline of the two papers is as follows:

Part I:

- **Section 2: Preliminaries to LVM learning** describes the assumptions of our approach and basic definitions of essential concepts of graphical models and SEM;
- **Section 3: Preliminaries to LPCC** formalizes our ideas and builds the theoretical basis for LPCC;
- **Section 4**: **Overview of LPCC** starts with an illustrative example and a broad description of the LPCC algorithm and then describes each step of LPCC in detail;
- **Section 5: Discussion and future research** summarizes and discusses the contribution of LPCC and suggests several new avenues of research;
- **Appendix A** provides proofs to all propositions, lemmas, and theorems for which the proof is either too detailed, lengthy, or impedes the flow of reading. All other proofs are given in the body of the paper;
- **Appendix B** sets a method to calculate a threshold in support of Section 4.4; and
- **Appendix C** supplies a detailed list of assumptions LPCC makes and the meaning of their violation.

Part II:

- **Section 2**: **The LPCC algorithm** introduces and formally describes a two-stage algorithm that implements the PCC concept;
- **Section 3: LPCC evaluation** evaluates LPCC, in comparison to state-of-the-art algorithms, using simulated and real-world data sets;
- **Section 4**: **Related works** compares LPCC to state-of-the-art LVM learning algorithms;
- **Section 5: Discussion** summarizes the theoretical advantages (from Part I) and the practical benefits (from this part) of using LPCC;
- **Appendix A** brings assumptions, definitions, propositions, and theorems from Part I that are essential to Part II;
- **Appendix B** supplies additional results for the experiments with the simulated data sets; and
- **Appendix C** provides PCC analysis for two example databases.

2. Preliminaries to LVM learning

The goal of our study is to reconstruct an LVM from i.i.d. data sampled from the observed variables in an unknown model. To accomplish this, we propose learning from pairwise cluster comparison using LPCC. First, we present the assumptions that LPCC makes and the constraints it applies on LVM and compare them to those required by other state-of-the-art methods.

Assumption 1 *The underlying model is a Bayesian network, BN=<**G**,**Θ**>, encoding a discrete joint probability distribution P for a set of random variables* **V**=**L**∪**O**, *where **G**=<**V**,**E**> is a directed acyclic graph (DAG) whose nodes* **V** *correspond to latents* **L** *and observed variables* **O**, *and **E** is the set of edges between nodes in **G**. **Θ** is the set of parameters, i.e., the conditional probabilities of variables in **V** given their parents.*

Assumption 2 *No observed variable in **O** is an ancestor of any latent variable in **L**.* This property is called the *measurement assumption* (Spirtes et al., 2000).

Before we present additional assumptions about the learned LVM, we need Definitions 1–4 (following Silva et al., 2006), which are specific to LVM:

Definition 1 *A model satisfying Assumptions 1 and 2 is a latent variable model.*

Definition 2 *Given an LVM **G** with a variable set **V**, the subgraph containing all variables in **V** and all and only those edges directed into variables in **O** is called the measurement model of **G**.*

Definition 3 *Given an LVM **G**, the subgraph containing all and only **G**'s latent nodes and their respective edges is called the structural model of **G**.*

When each model variable is a linear function of its parents in the graph plus an additive error term of positive finite variance, the latent variable model is linear; this is also known as SEM. Great interest has been shown in linear LVMs and their applications in social science, econometrics, and psychometrics (Bollen, 1989), as well as in their learning (Silva et al., 2006). The motivation to use linear models usually comes from social and related sciences. For example,[2] researchers give subjects a questionnaire with questions like: "On a scale of 1 to 5, how much do you agree with the statement: 'I feel sad every day'." The answer is measured by an observed variable, and the linearity of the influence of an unknown cause (say depression in this case) on the answer (value) to the question is assumed. By using other questions, which researchers assume also measure depression, they expect to discover a latent depression variable that is a parent of several observed variables (each measuring a question), together indicating depression. It is common to require several questions/observed variables for the identification of each latent variable and to consider the information revealed through only a single question as noise, which cannot guarantee the identification of the latent. Researchers expect that other questions will be clustered by another latent variable that measures another aspect in the domain, and thus

[2] P. Spirtes, private communication.

questions of one cluster will be independent of questions of another cluster conditioned on the latent variables. They also attribute unconditional independence that is detected between observed variables of different clusters to errors in the learning algorithm.

Adding the linearity assumption to Assumptions 1 and 2 allows for the transformation of Definition 1 into that of a linear LVM. Since assuming linearity means linearity is assumed in the measurement model, a key to learning a linear LVM is learning the measurement model and only then the structural model. In learning the measurement model of MIM, the linearity assumption entails constraints on the covariance matrix of the observed variables and thereby eliminates learning co-variants (dependences) between pairs of observed variables that "should" not be connected in the learned model (Silva et al., 2006; Spirtes, 2013). If, however, the linearity assumption does not hold, the algorithms suggested in Silva et al. (2006) may not find a model and would output a "can't tell" answer, which is, nevertheless, a better result than learning an incorrect model.

In this study, we dispense with the linearity assumption and apply the above concepts to learn not necessarily linear MIMs or latent-tree models. Our suggested algorithm, LPCC, is not limited by the linearity assumption and learns a model as long as it is MIM. In addition, we are interested in discrete LVMs.

Another important definition we need is:

Definition 4 *A pure measurement model is a measurement model in which each observed variable has only one latent parent and no observed parent.*

Assumption 3 *The measurement model of* $\boldsymbol{G}$ *is pure.*

As a principled way of testing conditional independence among latents, Silva et al. (2006) focus on MIMs, which are pure measurement models. Practically, these models have a smaller equivalence class of the latent structure than that of non-pure models and thus are easier to unambiguously learn. Consider, for example, that we are interested in learning the topic of a document (e.g., the first page in a newspaper) from anchor word (key phrase) distributions and that this document may cover several topics (e.g., politics, sports, and finance). Simplification of this topic modeling problem, following the representation of a topic using a latent variable in LVM, can be achieved by assuming and learning a pure measurement model representing a pure topic model for which each specific document covers only a single topic, which is reasonable in some cases, such as a sports or financial newspaper.

LPCC does not assume that the true measurement model is linear (which is a parametric assumption that, e.g., BPC makes), but rather assumes that the model is pure (a structural assumption). When the true causal model is pure, LPCC will identify it correctly (or find its pattern that represents the equivalence class of the true graph). When it is not pure, LPCC – similarly to BPC (Silva et al., 2006) – will learn a pure sub-model of the true model using two indicators for each latent (compared to three indicators per latent that are required by BPC). Part II of this paper presents several examples of real-world problems from different domains for which LPCC learns a pure (sometimes sub-) model, never less accurately than other methods.

That is, LPCC assumes that:

Assumption 4 *The true model **G** is MIM, in which each latent has at least two observed children and may have latent parents.*

Causal structure discovery – learning the number of latent variables in the model, their interconnections and connections to the observed variables, as well as the interconnections among the observed variables – is very difficult and thus requires making some assumptions about the problem. Particularly, MIMs, in which multiple observed variables are assumed to be affected by latent variables and perhaps by each other (Spirtes, 2013), are reasonable models but have attracted scant attention in the machine-learning community. As Silva et al. (2006) pointed out, factor analysis, principal component analysis, and regression analysis adapted to learning LVMs are well understood but have not been proven, under any general assumptions, to learn the true causal LVM, calling for better learning methods. By assuming that the true model manifests local influence of each latent variable on at least a small number of observed variables, Silva et al. (2006) showed that learning the complete Markov equivalence class of MIM is feasible. Similar to Silva et al. (2006), we assume that the true model is MIM; thus, this is where we place our focus on learning. Note also that based on Assumptions 3 and 4, the observed variables in ***G*** are d-separated, given the latents.

3. Preliminaries to LPCC

Figure 1 sketches a range of MIMs, which all exhibit pure measurement models, from basic to more complex models. Compared to G1, which is a basic MIM of two unconnected latents, G2 shows a structural model that is characterized by a latent collider. Note that such an LVM cannot be learned by latent-tree algorithms such as in Zhang (2004). G3 and G4 demonstrate serial and diverging structural models, respectively, that together with G2 cover the three basic structural models. G5 and G6 manifest more complex structural models comprising a latent collider and a combination of serial and diverging connections. As the structural model becomes more complicated, the learning task becomes more challenging; hence, G1–G6 present a spectrum of such challenges to an LVM learning algorithm. [3]

In Section 3.1, we build the infrastructure to pairwise cluster comparison that relies on understanding the influence of the exogenous latent variables on the observed variables in the LVM. This influence is divided into major and minor effects that are introduced and explained in Section 3.2. In Section 3.3, we link this structural influence to data clustering and introduce the pairwise cluster comparison concept for learning an LVM.

3.1 The influence of exogenous latents on observed variables is fundamental to learning an LVM

We distinguish between observed (**O**) and latent (**L**) variables and between exogenous (**EX**) and endogenous (**EN**) variables. **EX** have zero in-degree, are autonomous, and unaffected

[3] In Part II of the paper, we compare LPCC with BPC and exploratory factor analysis using these six LVMs. Since BPC requires three indicators per latent to identify a latent, we determined from the beginning three indicators per latent for all true models to recover. Nevertheless, in Part II, we evaluate the learning algorithms for increasing numbers of indicators.

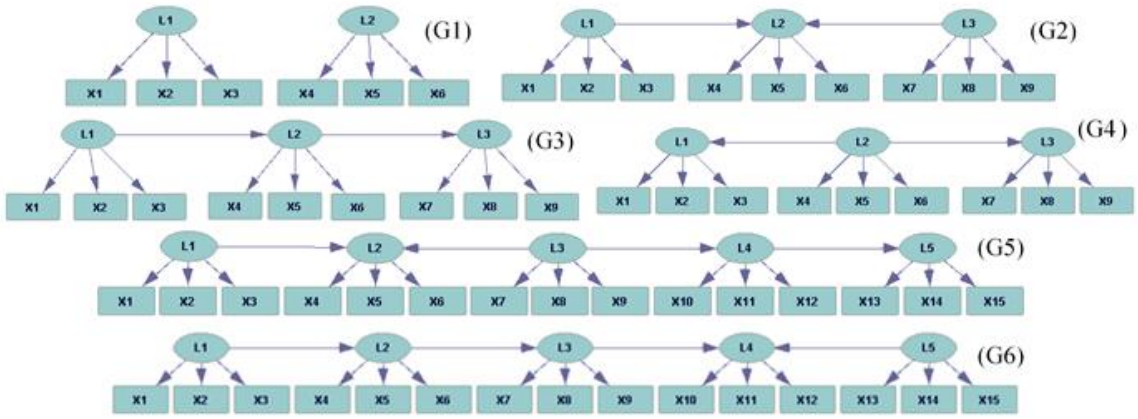

Figure 1: Example LVMs that are all MIMs. Each is based on a pure measurement model and a structural model of different complexity, posing a different challenge to a learning algorithm.

by the values of the other variables (e.g., L1 in all graphs but G4 in Figure 1), whereas **EN** are all non-exogenous variables in $\boldsymbol{G}$ (e.g., L2 in all graphs but G1 and G4, and X1 in all graphs in Figure 1). We identify three types of variables: (1) Exogenous latents, **EX**⊂(**L**∩**NC**) [all exogenous variables are latent non-colliders (**NC**)]; (2) Endogenous latents, **EL**⊂(**L**∩**EN**), which are divided into latent colliders **C**⊂**EL** (e.g., L2 in G2 and G5; note that all latent colliders are endogenous) and latent non-colliders (in serial and diverging connections) **S**⊂(**EL**∩**NC**) (e.g., L3 in G3, G4, and G6), thus **NC**= (**EX**∪**S**); and (3) Observed variables, **O**⊂**EN**, which are always endogenous and childless, that are divided into children of exogenous latents **OEX**⊂**O** (e.g., X1 and X9 in G2), children of latent colliders **OC**⊂**O** (e.g., X4, X5, and X6 in G2), and children of endogenous latent non-colliders **OS**⊂**O** (e.g., X4, X5, and X6 in G3). We denote value configurations of **EX**, **EN** (when we do not know whether the endogenous variables are latent or observed), **EL**, **C**, **NC** (when we do not know whether the non-collider variables are exogenous or endogenous), **S**, **O**, **OEX**, **OC**, and **OS** by **ex**, **en**, **el**, **c**, **nc**, **s**, **o**, **oex**, **oc**, and **os**, respectively.

Since the underlying model is a BN, the joint probability over **V**, which is represented by the BN, is factored according to the local Markov assumption for $\boldsymbol{G}$. That is, any variable in **V** is independent of its non-descendants in $\boldsymbol{G}$ conditioned on its parents in $\boldsymbol{G}$:

$$P(\mathbf{V}) = \prod_{V_i \in \mathbf{V}} P(V_i|\mathbf{Pa}_i) \tag{1}$$

where $\mathbf{Pa}_i$ are the parents of V_i. It can be factorized under our assumptions as:

$$P(\mathbf{V}) = P(\mathbf{EX}, \mathbf{C}, \mathbf{S}, \mathbf{OEX}, \mathbf{OC}, \mathbf{OS}) = \prod_{EX_i \in \mathbf{EX}} P(EX_i) \prod_{C_j \in \mathbf{C}} P(C_j|\mathbf{Pa}_j) \prod_{S_t \in \mathbf{S}} P(S_t|Pa_t) \prod_{OEX_m \in \mathbf{OEX}} P(OEX_m|EX_m) \prod_{OC_k \in \mathbf{OC}} P(OC_k|C_k) \prod_{OS_v \in \mathbf{OS}} P(OS_v|S_v) \quad (2)$$

where $\mathbf{Pa}_j$ are the latent parents of the latent collider C_j, Pa_t is the latent parent of the latent non-collider S_t (in other words, $\mathbf{Pa}_j$, $Pa_t \subset \mathbf{NC}$), $C_k \in \mathbf{C}$ and $S_v \in \mathbf{S}$ are the latent collider and latent non-collider parents of observed variables OC_k and OS_v, respectively, and $EX_m \in \mathbf{EX}$ is the exogenous latent parent of observed variable OEX_m.

In this paper, we claim and demonstrate that the influence of exogenous (latent) variables on observed variables is fundamental to learning an LVM and introduce LPCC that identifies and exploits this influence to learn an MIM. In this section, we prove that changes in values of the observed variables are due to changes in values of the exogenous variables and thus the identification of the former indicates the existence of the latter. To do that, we analyze the propagation of influence along paths connecting both variables, remembering that the paths may contain latent colliders and latent non-colliders. First, however, we should analyze paths among the latents and only then paths ending in their sinks (i.e., the observed variables). To prove that all changes in the graph, and specifically those measured in the observed variables, are the result of changes in the exogenous latent variables, we will need to first provide some definitions (following Spirtes et al., 2000; Pearl, 1988, 2000) of paths and some assumptions about the possible paths between latents in the structural model.

Definition 5 *A path between two nodes V_1 and V_n in a graph $\boldsymbol{G}$ is a sequence of nodes $\{V_1, ..., V_n\}$, such that V_i and V_{i+1} are adjacent in $\boldsymbol{G}$, $1 \leq i < n$, i.e., $\{V_i, V_{i+1}\} \in \mathbf{E}$.*

Note that a unique set of edges is associated with each given path. Paths are assumed to be simple by definition; in other words, no node appears in a path more than once, and an empty path consists of a single node.

Definition 6 *A collider on a path $\{V_1, ..., V_n\}$ is a node $V_i, 1 < i < n$, such that V_{i-1} and V_{i+1} are parents of V_i.*

Definition 7 *A directed path T_{V_n} from V_1 to V_n in a graph $\boldsymbol{G}$ is a path between these two nodes, such that for every pair of consecutive nodes V_i and V_{i+1}, $1 \leq i < n$ on the path, there is an edge from V_i into V_{i+1} in $\mathbf{E}$. V_1 is the source, and V_n is the sink of the path. A directed path has no colliders.*

While BPC (Silva et al., 2006) needs to make a parametric assumption about the linearity of the model, LPCC makes assumptions about the model structure (Assumption 3 above and Assumption 5 below). This is also the approach of latent-tree algorithms (Zhang, 2004; Harmeling and Williams, 2011; Wang et al., 2008) that restrict the learned structure to a tree (note that LPCC is not limited to a tree because it allows latent variables

to be colliders). This shows a tradeoff between the structural and parametric assumptions that an algorithm for learning an LVM usually has to make; the fewer parametric assumptions the algorithm makes, the more structural assumptions it has to make and vice versa.

Assumption 5 *A latent collider does not have any latent descendants (and thus cannot be a parent of another latent collider).*

To distinguish between latent colliders and latent non-colliders, their observed children, and their connectivity patterns to their exogenous variables, we use Lemma 1. Latent colliders and their observed children are connected to several exogenous variables via several directed paths, whereas latent non-colliders and their observed children are connected only to a single exogenous variable via a single directed path. Use of these different connectivity patterns – from exogenous latents through endogenous latents (both colliders and non-colliders) to observed variables – simplifies (2) and the analysis of the influence of latents on observed variables.

Lemma 1

1. *Each latent non-collider NC_t has only one exogenous latent ancestor EX_{NC_t}, and there is only one directed path T_{NC_t} from EX_{NC_t}(source) to NC_t (sink).* (Note that we use the notation NC_t, rather than S_t, since the lemma applies to both exogenous and endogenous latent non-colliders.)

2. *Each latent collider C_j is connected to a set of exogenous latent ancestors $\mathbf{EX}_{C_j}$ via a set of directed paths $\mathbf{T}_{C_j}$ from $\mathbf{EX}_{C_j}$ (sources) to C_j (sink).*

Lemma 1 allows us to separate the influence of all exogenous variables to separate paths of influence, each from exogenous to observed variables. Proposition 1 quantifies the propagation of this influence along the paths through the joint probability distribution.

Proposition 1 *The joint probability over $\boldsymbol{V}$ due to value assignment $\boldsymbol{ex}$ to exogenous set $\boldsymbol{EX}$ is determined only by this assignment and the BN conditional probabilities.*

Proof The first product in (2) for assignment **ex** is of **ex**'s priors. In the other five products, the probabilities are of endogenous latents or observed variables conditioned on their parents, which, based on Lemma 1, are either on the directed paths from **EX** to the latents/observed variables or exogenous themselves. Either way, any assignment of endogenous latents or observed variables is a result of the assignment **ex** to **EX** that is mediated to the endogenous latents/observed variables by the BN probabilities:

$$P(\mathbf{V}|\mathbf{EX}=\mathbf{ex}) = P(\mathbf{EX},\mathbf{C},\mathbf{S},\mathbf{OEX},\mathbf{OC},\mathbf{OS}|\mathbf{EX}=\mathbf{ex}) =$$
$$\prod_{EX_i\in\mathbf{EX}} P(EX_i=ex_i) \prod_{C_j\in\mathbf{C}} P(C_j=c_j|\mathbf{Pa}_j=\mathbf{pa}_j^{\mathbf{ex}_{C_j}}) \prod_{S_t\in\mathbf{S}} P(S_t=nc_t|Pa_t=pa_t^{ex_{S_t}}) \tag{3}$$
$$\prod_{OEX_m\in\mathbf{OEX}} P(OEX_m=oex_m|EX_m=ex_m) \prod_{OC_k\in\mathbf{OC}} P(OC_k=oc_k|C_k=c_k^{\mathbf{ex}_{C_k}}) \prod_{OS_v\in\mathbf{OS}} P(OS_v=os_v|S_v=s_v^{ex_{S_v}})$$

where

- ex_i and ex_m are the values of EX_i and EX_m (the latter is the parent of the mth observed child of the exogenous latents), respectively, in the assignment **ex** to **EX**;
- $\mathbf{pa}_j^{\mathbf{ex}_{C_j}}$ is the configuration of C_j's parents due to configuration $\mathbf{ex}_{C_j}$ of C_j's exogenous ancestors in **ex**;
- $pa_t^{ex_{S_t}}$ is the value of S_t's parent due to the value ex_{S_t} of S_t's exogenous ancestor in **ex**;
- $c_k^{\mathbf{ex}_{C_k}}$ is the value of OC_k's collider parent due to the configuration $\mathbf{ex}_{C_k}$ of C_k's exogenous ancestors in **ex**; and
- $s_v^{ex_{S_v}}$ is the value of OS_v's non-collider parent due to the value ex_{S_v} of S_v's exogenous ancestor in **ex**.

■

Proposition 1 along with Lemma 1 are a key in our analysis because they show paths of hierarchical influence of latents on observed variables – from exogenous latents through endogenous latents (both colliders and non-colliders) to observed variables. Recognition and use of these paths of influence guides LPCC in learning LVMs.

To formalize our ideas, we introduce several concepts in Section 3.2. First, we define local influence on a single *EN* of its direct parents. Second, we use local influences and the BN Markov property to generalize the influence of **EX** on **EN**. Third, exploiting the connectivity between the exogenous ancestors and their endogenous descendants, as described by Lemma 1, we focus on the influence of a specific (partial) set of exogenous variables on the values of their endogenous descendants. Analysis of the influence of all configurations **ex**s on all **en**s and that of the configurations of specific exogenous ancestors in these **ex**s on their endogenous descendants enable learning the structure and parameters of the model and causal discovery. Finally, in Section 3.3, we show how these concepts can be exploited to learn an LVM from data clustering.

3.2 Major and minor effects and values

So far, we have analyzed the structural influences (path of hierarchies) of the latents on the observed variables. In this section, we complement this analysis with the parametric influences, which we divide into major and minor effects.

Definition 8 *A local effect on an endogenous variable EN is the influence of a configuration of EN's direct latent parents on any of EN's values.*

1. *A major local effect is the largest local effect on EN_i, and it is identified by the maximal conditional probability of a specific value en_i of EN_i given a configuration $\mathbf{pa}_i$ of EN_i's latent parents $\mathbf{Pa}_i$, which is $MAE_{EN_i}(\mathbf{pa}_i) = max_{en_i'} P(EN_i = en_i' | \mathbf{Pa}_i = \mathbf{pa}_i)$.*
2. *A minor local effect is any non-major local effect on EN_i, and it is identified by a conditional probability of any other value of EN_i given $\mathbf{pa}_i$ that is smaller than $MAE_{EN_i}(\mathbf{pa}_i)$. The minor local effect set, $MIES_{EN_i}(\mathbf{pa}_i)$, comprises all such probabilities.*

3. *A major local value is the en_i corresponding to $MAE_{EN_i}(\mathbf{pa}_i)$, i.e., the most probable value of EN_i due to $\mathbf{pa}_i$, $MAV_{EN_i}(\mathbf{pa}_i) = argmax_{en_i'} P(EN_i = en_i' | \mathbf{Pa}_i = \mathbf{pa}_i)$.*

4. *A minor local value is an en_i corresponding to a minor local effect, and $MIVS_{EN_i}(\mathbf{pa}_i)$ is the set of all minor values that correspond to $MIES_{EN_i}(\mathbf{pa}_i)$.*

   When EN_i is an observed variable or an endogenous latent non-collider, and thus has only a single parent Pa_i, the configuration $\mathbf{pa}_i$ is actually the value pa_i of Pa_i.

So far, we have listed our assumptions about the structure of the model. Following is a parametric assumption:

Assumption 6 *For every endogenous variable EN_i in $\boldsymbol{G}$ and every configuration $\mathbf{pa}_i'$ of EN_i's parents $\mathbf{Pa}_i$, there exists a certain value en_i' of EN_i, such that $P(EN_i = en_i' | \mathbf{Pa}_i = \mathbf{pa}_i') > P(EN_i = en_i'' | \mathbf{Pa}_i = \mathbf{pa}_i')$ for every other value en''_i of EN_i.* This assumption is related to the most probable explanation of a hypothesis given the data (Pearl, 1988).

Note that in the case that Assumption 6 is violated, in other words, if more than one value of EN_i gets the maximum probability value given a configuration of parents, LPCC still learns a model because the implementation will randomly choose a value that maximizes the probability as the most probable. However, the correctness of the algorithm is guaranteed only if all assumptions are valid; in other words, given the assumptions are valid, all causal claims made by the output graph are correct.

Proposition 2 *The major local value $MAV_{EN_i}\left(\mathbf{pa}_i'\right)$ of an endogenous variable EN_i given a certain configuration of its parents $\mathbf{pa}_i'$ is also certain.*

Proof Assumption 6 guarantees that given a certain configuration $\mathbf{pa}_i'$ of $\mathbf{Pa}_i$, there exists a certain value en_i' of EN_i, such that $P(EN_i = en_i' | \mathbf{Pa}_i = \mathbf{pa}_i') > P(EN_i = en_i'' | \mathbf{Pa}_i = \mathbf{pa}_i')$ for every other value en_i'' of EN_i. From the definition of a major local value, $MAV_{EN_i}\left(\mathbf{pa}_i'\right) = en_i'$. ■

We need one additional assumption about the model parameters that reflects parent-child influence in the causal model. Specifically, to identify parent-child relations, LPCC needs for each observed variable or endogenous latent non-collider to get different MAVs for different values of their latent parent. Similarly, LPCC needs a collider to get different values for each of its parents in at least two parent configurations in which this parent changes, whereas the other parents do not.

Assumption 7 *First, for every EN_i that is an observed variable or an endogenous latent non-collider and for every two values pa_i' and pa_i'' of Pa_i, $MAV_{EN_i}\left(pa_i'\right) \neq MAV_{EN_i}\left(pa_i'\right)$. Second, for every C_j that is a latent collider and for every $Pa_j \in \mathbf{Pa}_j$, there are at least two configurations $\mathbf{pa}_j'$ and $\mathbf{pa}_j''$ of $\boldsymbol{Pa}_j$ in which only the value of Pa_j is different and $MAV_{C_j}\left(\mathbf{pa}_j'\right) \neq MAV_{C_j}\left(\mathbf{pa}_j''\right)$.*

By aggregation over all local influences, we can now generalize these concepts through the BN parameters and Markov property from local influences on specific endogenous variables to influence on all endogenous variables in the graph.

Definition 9 *An effect on **EN** is the influence of a configuration **ex** of **EX** on **EN**. The effect is measured by a value configuration **en** of **EN** due to **ex**. A major effect (MAE) is the largest effect of **ex** on **EN** and a minor effect (MIE) is any non-MAE effect of **ex** on **EN**. Also, a major value configuration (MAV) is the configuration **en** of **EN** corresponding to MAE (i.e., the most probable **en** due to **ex**), and a minor value configuration is a configuration **en** corresponding to any MIE.*

[Note the difference between a major effect, MAE, and a major local effect, MAE_{EN_i}, and between a major value configuration, MAV, and a major local value, MAV_{EN_i} (and similarly for the "minors").]

Based on the proof of Proposition 1, we can quantify the effect of **ex** on **EN**. For example, a major effect of **ex** on **EN** can be factorized according to the product of major local effects on **EN** (weighted by the product of priors, $P(EX_i{=}ex_i)$):

$$MAE(\mathbf{ex}) = \prod_{EX_i \in \mathbf{EX}} P(EX_i = ex_i) \prod_{C_j \in \mathbf{C}} MAE_{C_j}(\mathbf{pa}_j^{\mathbf{ex}_{C_j}}) \prod_{S_t \in \mathbf{S}} MAE_{S_t}(pa_t^{ex_{S_t}})$$
$$\prod_{OEX_m \in \mathbf{OEX}} MAE_{OEX_m}(ex_m) \prod_{OC_k \in \mathbf{OC}} MAE_{OC_k}(c_k^{\mathbf{ex}_{C_k}}) \prod_{OS_v \in \mathbf{OS}} MAE_{OS_v}(s_v^{ex_{S_v}}) =$$
$$\prod_{EX_i \in \mathbf{EX}} P(EX_i = ex_i) \prod_{C_j \in \mathbf{C}} max_{c_j'} P(C_j = c_j' | \mathbf{Pa}_j = \mathbf{pa}_j^{\mathbf{ex}_{C_j}}) \prod_{S_t \in \mathbf{S}} max_{s_t'} P(S_t = s_t' | Pa_t = pa_t^{ex_{S_t}})$$
$$\prod_{OEX_m \in \mathbf{OEX}} max_{oex_m'} P(OEX_m = oex_m' | EX_m = ex_m) \prod_{OC_k \in \mathbf{OC}} max_{oc_k'} P(OC_k = oc_k' | C_k = c_k^{\mathbf{ex}_{C_k}})$$
$$\prod_{OS_v \in \mathbf{OS}} max_{os_v'} P(OS_v = os_v' | S_v = s_v^{ex_{S_v}}). \quad (4)$$

A configuration **en** of **EN** in which each variable in **EN** takes on the major local value is major or a MAV. Any effect in which at least one *EN* takes on a minor local effect is minor, and any configuration in which at least one *EN* takes on a minor local value is minor. We denote the set of all minor effects for **ex** with $MIES(\mathbf{ex})$ (with correspondence to $MIES_{EN_i}$) and the set of all minor configurations with $MIVS(\mathbf{ex})$ (with correspondence to $MIVS_{EN_i}$).

Motivated by Lemma 1 and Proposition 1, we are interested in representing the influence on a subset of the endogenous variables of the subset of the exogenous variables that impact these endogenous variables. This partial representation of MAE will enable LPCC to recover the relationships between exogenous ancestors and only the descendants that are affected by these exogenous variables. To achieve this, we first extend the concept of effect to the concept of partial effect of specific exogenous variables and then quantify it. Later, we shall formalize all of this in Lemma 2.

Definition 10 *A partial effect on a subset of endogenous variables* $\mathbf{EN}^{'} \subseteq \mathbf{EN}$ *is the influence of a configuration* $\mathbf{ex}^{'}$ *of* $\mathbf{EN}^{'}$*'s exogenous ancestors* $\mathbf{EX}^{'} \subseteq \mathbf{EX}$ *on* $\mathbf{EN}^{'}$*. We define a partial major effect* $MAE_{\mathbf{EN}^{'}}\left(\mathbf{ex}^{'}\right)$ *as the largest partial effect of* $\mathbf{ex}^{'}$ *on* $\mathbf{EN}^{'}$ *and a partial minor effect* $MIE_{\mathbf{EN}^{'}}\left(\mathbf{ex}^{'}\right)$ *as any non-*$MAE_{\mathbf{EN}^{'}}\left(\mathbf{ex}^{'}\right)$ *partial effect of* $\mathbf{ex}^{'}$ *on* $\mathbf{EN}^{'}$*. A partial major value configuration* $MAV_{\mathbf{EN}^{'}}\left(\mathbf{ex}^{'}\right)$ *is the* $\mathbf{en}^{'}$ *of* $\mathbf{EN}^{'}$ *corresponding to* $MAE_{\mathbf{EN}^{'}}\left(\mathbf{ex}^{'}\right)$*; in other words, the most probable* $\mathbf{en}^{'}$ *due to* $\mathbf{ex}^{'}$*, and a partial minor value configuration is an* $\mathbf{en}^{'}$ *corresponding to any* $MIE_{\mathbf{EN}^{'}}\left(\mathbf{ex}^{'}\right)$*.*

We are interested in representing the influence of exogenous variables on their observed descendants and all the variables in the directed paths connecting them. To do this, we separately analyze the (partial) effect of each exogenous variable on each observed variable for which the exogenous is its ancestor and all the latent variables along the path connecting these two. We distinguish between two cases (both are represented in Lemma 1): (1) Observed descendants in **OEX** and **OS** that are, respectively, children of exogenous latents and children of latent non-colliders that are linked to their exogenous ancestors, each via a single directed path; and (2) Observed descendants in **OC** that are children of latent colliders and linked to their exogenous ancestors via a set of directed paths through their latent collider parents. Thus, we are interested in:

1. The partial effect of a value of exogenous ancestor EX_{NC_v} to non-collider NC_v on any configuration of the set of variables $\{TS_{NC_v} \backslash EX_{NC_v}, ONC_v\}$, where ONC_v is an observed child of latent non-collider NC_v, and TS_{NC_v} is the set of variables in the directed path (recall Definition 7) T_{NC_v} from EX_{NC_v} to NC_v. The corresponding $MAE_{\{TS_{NC_v} \backslash EX_{NC_v}, ONC_v\}}\left(ex_{NC_v}\right)$ and $MAV_{\{TS_{NC_v} \backslash EX_{NC_v}, ONC_v\}}\left(ex_{NC_v}\right)$ are partial major effect and partial major value configuration, respectively. For example, we may be interested in the partial effect of a value of $EX_{NC_v} = EX_{\mathrm{L5}} = \mathrm{L3}$ in G5 (Figure 1) on $\{TS_{NC_v} \backslash EX_{NC_v}, ONC_v\}$= $\{TS_{\mathrm{L5}} \backslash \mathrm{L3}, \mathrm{X13}\}$=$\{\mathrm{L4}, \mathrm{L5}, \mathrm{X13}\}$. Note that we use here the notation NC_v since we are interested in both exogenous and endogenous latent non-colliders. When we are interested in the partial effect on an observed variable in **OEX**, its exogenous ancestor (which is also its direct parent) is also the latent non-collider, NC_v, and the effect is not measured on any other variable but this observed variable. This is *Case 1*, which is analyzed below;

2. The partial effect of a configuration of exogenous variables $\mathbf{EX}_{C_k}$ to collider C_k on any configuration of the set of variables $\{\mathbf{TS}_{C_k} \backslash \mathbf{EX}_{C_k}, OC_k\}$, where OC_k is an observed child of latent collider C_k, [4] and $\mathbf{TS}_{C_k}$ is the set of variables in the set of directed paths $\mathbf{T}_{C_k}$ from $\mathbf{EX}_{C_k}$ to C_k. The corresponding $MAE_{\{\mathbf{TS}_{C_k} \backslash \mathbf{EX}_{C_k}, OC_k\}}\left(\mathbf{ex}_{C_k}\right)$ and $MAV_{\{\mathbf{TS}_{C_k} \backslash \mathbf{EX}_{C_k}, OC_k\}}\left(\mathbf{ex}_{C_k}\right)$ are partial major effect and partial major value configuration, respectively. For example, we may be interested in the partial effect of a configuration of $\mathbf{EX}_{C_k} = \mathbf{EX}_{\mathrm{L4}} = \{\mathrm{L1}, \mathrm{L5}\}$ in G6 (Figure 1) on $\{\mathbf{TS}_{C_k} \backslash \mathbf{EX}_{C_k}, OC_k\}$= $\{\{\{\mathrm{L1}, \mathrm{L2}, \mathrm{L3}, \mathrm{L4}\} \backslash \{\mathrm{L1}\}, \{\mathrm{L5}\} \backslash \{\mathrm{L5}\}\}, \mathrm{X11}\}$=$\{\mathrm{L2}, \mathrm{L3}, \mathrm{L4}, \mathrm{X11}\}$. This is *Case 2*, which is analyzed below.

[4] Throughout the paper, we use a child index also for its parent, e.g., OC_k's parent is C_k, although generally, we use the index j for a collider, such as C_j.

Following, we provide detailed descriptions for these partial effects and partial values for observed children of latent non-colliders (*Case 1*) and observed children of latent colliders (*Case 2*) and formalize their properties in Propositions 3–7 to set the stage for Lemma 2.

Case 1: Observed children of latent non-colliders

If the latent non-collider NC_v is exogenous, $NC_v = EX_v$ and ONC_v=OEX_v, then, $\{TS_{NC_v} \backslash EX_{NC_v}, ONC_v\} = OEX_v$. Thus, the partial effect is simply the local effect, and the partial major effect is the major local effect $MAE_{OEX_v}(ex_v)$. If the latent non-collider NC_v is endogenous, then $NC_v = S_v$ and ONC_v=OS_v. Then, all variables in $\{TS_{S_v} \backslash EX_{S_v}, OS_v\}$ are d-separated by EX_{S_v} from $\mathbf{EX} \backslash EX_{S_v}$. For example, {L4, L5, X13} in G5 (Figure 1) are d-separated by L3 from L2 and its children. Thus, the effect of **ex** on the joint probability distribution (3) can be factored to the: a) joint probability over **EX=ex**; b) conditional probabilities of the influenced variables along a specific directed path that ends at OS_v on $EX_{S_v} = ex_{S_v}$ (note that the value ex_{S_t} for all $S_t \in TS_{S_v}$ is the same because $EX_{S_t} = EX_v$ is the same exogenous ancestor of all latent non-colliders on the path to S_v); and c) conditional probabilities of all the remaining variables in the graph on **EX=ex**:

$$P(\mathbf{V}|\mathbf{EX} = \mathbf{ex}) = P(\mathbf{EX} = \mathbf{ex}) P(\{TS_{S_v} \backslash EX_{S_v}, OS_v\} | EX_{S_v} = ex_{S_v}) \\ P(\mathbf{V} \backslash \{TS_{S_v} \backslash EX_{S_v}, OS_v\} | \mathbf{EX} = \mathbf{ex}) \tag{5}$$

in which the second factor corresponds to the partial effect of $EX_{S_v} = ex_{S_v}$ on $TS_{S_v} \backslash EX_{S_v}$ (the latent non-colliders on the path from EX_{S_v} to S_v) and S_v's observed child, OS_v, and the third factor corresponds to the influence of **EX=ex** on all the other (latent and observed) variables in the graph. We can write the second factor describing the partial effect of the value ex_{S_v} on the values of the variables $TS_{S_v} \backslash EX_{S_v}$ in the directed path from EX_{S_v} to OS_v (including) as:

$$P(\{TS_{S_v} \backslash EX_{S_v}, OS_v\} | EX_{S_v} = ex_{S_v}) = \\ \prod_{S_t \in \{TS_{S_v} \backslash EX_{S_v}\}} P(S_t = s_t | Pa_t = pa_t^{ex_{S_t}}) P(OS_v = os_v | S_v = s_v^{ex_{S_v}}) \tag{6}$$

The partial major effect in (4) for this directed path can be written as (note again that $ex_{S_t} = ex_{S_v}$):

$$MAE_{\{TS_{S_v} \backslash EX_{S_v}, OS_v\}}(ex_{S_v}) = MAE_{\{TS_{S_v} \backslash EX_{S_v}\}}(ex_{S_v}) \cdot MAE_{OS_v}(s_v^{ex_{S_v}}) = \\ \prod_{S_t \in \{TS_{S_v} \backslash EX_{S_v}\}} MAE_{S_t}(pa_t^{ex_{S_v}}) \cdot MAE_{OS_v}(s_v^{ex_{S_v}}) \tag{7}$$

Proposition 3 *The* $MAV_{\{TS_{NC_v} \backslash EX_{NC_v}, ONC_v\}}\left(ex_{NC_v}\right)$ *corresponding to* $MAE_{\{TS_{NC_v} \backslash EX_{NC_v}, ONC_v\}}\left(ex_{NC_v}\right)$ *is a certain value configuration for each certain value* ex_{NC_v}.

(Note that here we use the notation NC_v rather than S_v since the proposition applies to both exogenous and endogenous latent non-colliders.)

Proposition 4 *All corresponding values in* $MAV_{\{TS_{NC_v}\setminus EX_{NC_v},ONC_v\}}\left(ex'_{NC_v}\right)$ *and* $MAV_{\{TS_{NC_v}\setminus EX_{NC_v},ONC_v\}}\left(ex''_{NC_v}\right)$*, for two values* ex'_{NC_v} *and* ex''_{NC_v} *of* EX_{NC_v}*, are different.*

(Here also we use the notation NC_v, since the proposition applies to both exogenous and endogenous latent non-colliders.)

So far, we have analyzed the impact of an exogenous variable on a latent non-collider by "propagating" the exogenous (source) impact along the path to the latent non-collider (sink). Propositions 3 and 4, respectively, guarantee that a certain value of the exogenous variable is responsible for a certain value of the latent non-collider and different values of the exogenous are echoed through different values of the latent non-collider. Proposition 4 is based on the correspondence between changes in values of a latent non-collider and changes in values of its parent; a correspondence that is guaranteed by Assumption 7 (first part). Propositions 3 and 4, respectively, ensure the existence and uniqueness of the value a latent non-collider gets under the influence of an exogenous ancestor; one (Proposition 3) and only one (Proposition 4) value of the latent non-collider changes with a change in the value of the exogenous. We formalize this in the following Proposition 5.

Proposition 5 EX_{NC_v} *changes values (i.e., has two values* ex'_{NC_v} *and* ex''_{NC_v}*) if and only if* NC_v *changes values in the two corresponding major value configurations:* $MAV_{\{TS_{NC_v}\setminus EX_{NC_v},ONC_v\}}\left(ex'_{NC_v}\right)$ *and* $MAV_{\{TS_{NC_v}\setminus EX_{NC_v},ONC_v\}}\left(ex''_{NC_v}\right)$.

Case 2: Observed children of latent colliders

In the case of an observed variable OC_k that is a child of a latent collider C_k, all variables in $\{\mathbf{TS}_{C_k}\setminus\mathbf{EX}_{C_k}, OC_k\}$ are d-separated by $\mathbf{EX}_{C_k}$ from $\mathbf{EX}\setminus\mathbf{EX}_{C_k}$. Thus, the effect of **ex** on the joint probability distribution (3) can be factored (similarly to Case 1) to the: a) joint probability over **EX=ex**; b) conditional probabilities of the influenced variables along all directed paths that end at OC_k on $\mathbf{EX}_{C_k} = \mathbf{ex}_{C_k}$(note that all variables along each directed path T_{C_k} are influenced by the same ex_{C_k}); and c) conditional probabilities of all the remaining variables in the graph on **EX=ex**:

$$P(\mathbf{V}|\mathbf{EX}=\mathbf{ex}) = P(\mathbf{EX}=\mathbf{ex})P(\{\mathbf{TS}_{C_k}\setminus\mathbf{EX}_{C_k}, OC_k\}|\mathbf{EX}_{C_k}=\mathbf{ex}_{C_k})$$
$$P(\mathbf{V}\setminus\{\mathbf{TS}_{C_k}\setminus\mathbf{EX}_{C_k}, OC_k\}|\mathbf{EX}=\mathbf{ex}) \tag{8}$$

in which the second factor corresponds to the partial effect on $\{\mathbf{TS}_{C_k}\setminus\mathbf{EX}_{C_k}, OC_k\}$ of $\mathbf{EX}_{C_k}$, and the third factor corresponds to the partial effect on all variables other than $\{\mathbf{TS}_{C_k}\setminus\mathbf{EX}_{C_k}, OC_k\}$. We can decompose the second factor into a product of: a) a product over all directed paths into C_k of a product of partial effects over all variables (excluding C_k) in such a path; b) the partial effect on C_k; and c) the partial effect on its child OC_k:

$$P(\{\mathbf{TS}_{C_k}\setminus\mathbf{EX}_{C_k}, OC_k\}|\mathbf{EX}_{C_k}=\mathbf{ex}_{C_k}) =$$
$$\prod_{TS_{C_k}\in\mathbf{TS}_{C_k}}\prod_{S_t\in TS_{C_k}\setminus\{EX_{C_k},C_k\}} P(S_t=s_t|Pa_t=pa_t^{ex_{C_k}})P(C_k=c_k|\mathbf{Pa}_k=\mathbf{pa}_k^{\mathbf{ex}_{C_k}})P(OC_k=oc_k|C_k=c_k^{\mathbf{ex}_{C_k}}). \tag{9}$$

This factor can be rewritten as:

$$P(\{\mathbf{TS}_{C_k} \setminus \mathbf{EX}_{C_k}, OC_k\} | \mathbf{EX}_{C_k} = \mathbf{ex}_{C_k}) = \prod_{TS_{C_k} \in \mathbf{TS}_{C_k}} P(\{TS_{C_k} \setminus EX_{C_k}, C_k\} | EX_{C_k} = ex_{C_k}) P(C_k = c_k | \mathbf{Pa}_k = \mathbf{pa}_k^{\mathbf{ex}_{C_k}}) P(OC_k = oc_k | C_k = c_k^{\mathbf{ex}_{C_k}}). \tag{10}$$

It reflects the partial effects of a configuration $\mathbf{ex}_{C_k}$ on the values of the variables in $\{\mathbf{TS}_{C_k} \setminus \mathbf{EX}_{C_k}\}$ and the values C_k and OC_k get, and thus the partial major effect of the second factor can be represented as:

$$MAE_{\{\mathbf{TS}_{C_k} \setminus \mathbf{EX}_{C_k}, OC_k\}}(\mathbf{ex}_{C_k}) = \prod_{TS_{C_k} \in \mathbf{TS}_{C_k}} MAE_{\{TS_{C_k} \setminus EX_{C_k}, C_k\}}(ex_{C_k}) MAE_{C_k}(\mathbf{pa}_k^{\mathbf{ex}_{C_k}}) MAE_{OC_k}(c_k^{\mathbf{ex}_{C_k}}). \tag{11}$$

Proposition 6 *The $MAV_{\{\mathbf{TS}_{C_k} \setminus \mathbf{EX}_{C_k}, OC_k\}}\left(\mathbf{ex}_{C_k}\right)$ corresponding to $MAE_{\{\mathbf{TS}_{C_k} \setminus \mathbf{EX}_{C_k}, OC_k\}}\left(\mathbf{ex}_{C_k}\right)$ is a certain value configuration for each certain value configuration $\mathbf{ex}_{C_k}$.*

We wish to apply the same mechanism as in Case 1 to analyze the impact of more than a single exogenous ancestor on a latent collider, but here the impact is propagated toward the collider along more than a single path. To accomplish this, the following Proposition 7 analyzes the effect on a collider of each of its exogenous ancestors by considering the effect of such an exogenous on the corresponding collider's parent (using Proposition 5, similar to Case 1 for a latent non-collider) and then the effect of this parent on the collider itself (using the second part of Assumption 7).

Proposition 7 *For every exogenous ancestor $EX_{C_k} \in \mathbf{EX}_{C_k}$ of a latent collider C_k, there are at least two configurations $\mathbf{ex}'_{C_k}$ and $\mathbf{ex}''_{C_k}$ of $\mathbf{EX}_{C_k}$ in which only EX_{C_k} of all $\mathbf{EX}_{C_k}$ changes values when C_k changes values in the two corresponding major value configurations $MAV_{\{\mathbf{TS}_{C_k} \setminus \mathbf{EX}_{C_k}, OC_k\}}\left(\mathbf{ex}'_{C_k}\right)$ and $MAV_{\{\mathbf{TS}_{C_k} \setminus \mathbf{EX}_{C_k}, OC_k\}}\left(\mathbf{ex}''_{C_k}\right)$.*

Lemma 2

1. *A latent non-collider NC_v and its observed child ONC_v, both descendants of an exogenous variable EX_{NC_v}, change their values in any two major configurations if and only if EX_{NC_v} has changed its value in the corresponding two configurations of $\boldsymbol{EX}$.*
2. *A latent collider C_k and its observed child OC_k, both descendants of a set of exogenous variables $\mathbf{EX}_{C_k}$, change their values in any two major configurations only if at least one of the exogenous variables in $\mathbf{EX}_{C_k}$ has changed its value in the corresponding two configurations of $\boldsymbol{EX}$.*

3.3 PCC by clustering observational data

Practically, we use observational data that were generated from an unknown LVM and measured over the observed variables. Proposition 1 showed us that each configuration

of observed variables (which is part of a configuration of the endogenous variables) and their joint probability is a result of the assignment of a configuration **ex** to the exogenous variables **EX**. Therefore, we define:

Definition 11 *An observed value configuration, observed major value configuration, and observed minor value configuration due to* **ex** *are the parts in* **en**, *MAV, and a minor value configuration, respectively, that correspond to the observed variables.*

The following two propositions formalize the relationships between the observed major value configurations and the set of possible **ex**.

Proposition 8 *There is only a single observed major value configuration to each exogenous configuration* **ex** *of* **EX**.

Proof Based on Lemma 2, different observed major value configurations can be obtained if and only if there is more than a single exogenous configuration **ex** of **EX**. Thus, an exogenous configuration **ex** can only lead to a single observed major value configuration. ■

Proposition 9 *There are different observed major value configurations to different exogenous configurations* **ex***s.*

Proof Assume for the sake of contradiction that two different value configurations $\mathbf{ex}_1$ and $\mathbf{ex}_2$ led to the same observed major value configuration. Because the two configurations are different, there is at least one exogenous variable EX' that has different values in $\mathbf{ex}_1$ and $\mathbf{ex}_2$. Since based on Assumption 4, EX' has at least two observed children, then, based on Assumption 7, each of these children has different values in the two observed major value configurations due to the different value of EX' in $\mathbf{ex}_1$ and $\mathbf{ex}_2$. This is contrary to our assumption that there is only one observed major value configuration. ■

Due to the probabilistic nature of BN, each observed value configuration due to **ex** may be represented by several data points. Clustering these data points may produce several clusters for each **ex** and each cluster corresponds to another observed value configuration. Based on Propositions 8 and 9, one and only one of the clusters corresponds to each of the observed major value configurations, whereas the other clusters correspond to observed minor value configurations. We distinguish between these clusters using Definition 12.

Definition 12 *The single cluster that corresponds to the observed major value configuration, and thus also represents the major effect* MAE(**ex**) *due to configuration* **ex** *of* **EX**, *is the major cluster for* **ex**, *and all the clusters that correspond to the observed minor value configurations due to minor effects in* $MIES$(**ex**) *are minor clusters.*

To resolve between different types of minor effects/clusters, we make two definitions.

Definition 13 *A k-order minor effect is a minor effect in which exactly k endogenous variables in* **EN** *correspond to minor local effects. An* **en** *corresponding to a k-order minor effect is a k-order minor value configuration.*

Definition 14 *Minor clusters that correspond to k-order minor effects are k-order minor clusters.*

Based on Proposition 9 and Definition 12, the set of all major clusters (corresponding to all observed major value configurations) reflects the effect of all possible **ex**s, and thus the number of major clusters is expected to be equal to the number of **EX** configurations. Therefore, the identification of all major clusters is a key to the discovery of exogenous variables and their causal interrelations. For this purpose, we introduce the concept of *pairwise cluster comparison* (PCC). PCC measures the differences between two clusters; each represents the response of LVM to another **ex**.

Definition 15 *Pairwise cluster comparison is a procedure by which pairs of clusters are compared, for example through a comparison of their centroids. The result of PCC between a pair of cluster centroids of dimension* $|\mathbf{O}|$*, where* $\mathbf{O}$ *is the set of observed variables, can be represented by a binary vector of size* $|\mathbf{O}|$ *in which each element is 1 or 0 depending, respectively, on whether or not there is a difference between the corresponding elements in the compared centroids.*

When PCC is between clusters that represent observed major value configurations (i.e., PCC between major clusters), an element of 1 identifies an observed variable that has changed its value between the compared clusters due to a change in **ex**. Thus, the 1s in a major–major PCC provide evidence of causal relationships between **EX** and **O**. Practically, LPCC always identifies all observed variables that are represented by 1s *together* in *all* PCCs as the observed descendants of the same exogenous variable (Section 4.1). However, due to the probabilistic nature of BN and the existence of endogenous latents (mediating the connections from **EX** to **O**), some of the clusters are k-order minor clusters (in different orders), representing k-order minor configurations/effects. Minor clusters are more difficult to identify than major clusters because the latter reflect the major effects of **EX** on **EN** and, therefore, are considerably more populated by data points than the former. Nevertheless, minor clusters are important in causal discovery by LPCC even though a major–minor PCC cannot tell the effect of **EX** on **EN** because an observed variable in two compared (major and minor) clusters should not necessarily change its value as a result of a change in **ex**. Their importance is because a major cluster, which is a zero-order minor value configuration and thus has zero minor values, cannot indicate (when compared with another major cluster) the existence of minor values. On the contrary, PCC between major and minor clusters shows (through the number of 1s) the number of minor values represented in the minor cluster, and this is exploited by LPCC for identifying the endogenous latents and interrelations among them (Section 4.4). That is, PCC is the source to identify causal relationships in the unknown LVM; major–major PCCs are used for identifying the exogenous variables and their descendants, and major–minor PCCs are used for identifying the endogenous latents, their interrelations, and their observed children.

4. Overview of the LPCC concept[5]

Let us demonstrate the relations between clustering results and learning an LVM using LPCC through an example. G1 in Figure 1 shows a model having two exogenous variables,

[5]Preliminary versions of the PCC concept and LPCC algorithm are given in Asbeh and Lerner (2012).

L1 and L2, each having three children X1, X2, X3 and X4, X5, X6, respectively. [6] For the example, let us assume that all variables are binary, [7] i.e., L1 and L2 have four possible **ex**s (L1L2= 00, 01, 10, 11). First, we generated a synthetic data set of 1,000 patterns from G1 over the six observed variables. We used a uniform distribution over L1 and L2 and set the probabilities of an observed child, X_i, $i = 1,\ldots,6$, given its latent parent, $L_k, k = 1,2$ (only if L_k is a direct parent of X_i, e.g., L1 and X1), to be $P(X_i = v \mid L_k = v) = 0.8, v = 0,1$. Second, using the self-organizing map (SOM) (Kohonen, 1997), we clustered the data set and found 16 clusters, of which four were major (see Section 4.3 for details on how to identify major clusters). This meets our expectation of four major clusters corresponding to the four possible **ex***s*. These clusters are presented in Table 1a by their centroids, which are the most prevalent patterns in the clusters, and in Table 1b by their PCCs. For example, *PCC1,2*, comparing clusters *C1* and *C2*, shows that when moving from *C1* to *C2*, only the values corresponding to variables X1, X2, and X3 have been changed (i.e., $\delta X_1 = \delta X_2 = \delta X_3 = 1$ in Table 1b). Lemma 2 guarantees that the three variables are descendants of the same *EX* that changed its value between two **ex***s* represented by *C1* and *C2*. *PCC1,4*, *PCC2,3*, and *PCC3,4* reinforce this conclusion. Indeed, we know from the true graph, G1, that this *EX* is latent L1. A similar conclusion can be deduced about X4, X5, and X6 as descendants of an exogenous latent, which we know, based on the true graph, is L2.

Centroid	**X1**	**X2**	**X3**	**X4**	**X5**	**X6**
C1	0	0	0	1	1	1
C2	1	1	1	1	1	1
C3	0	0	0	0	0	0
C4	1	1	1	0	0	0

(a)

PCC	δ**X1**	δ**X2**	δ**X3**	δ**X4**	δ**X5**	δ**X6**
PCC1,2	1	1	1	0	0	0
PCC1,3	0	0	0	1	1	1
PCC1,4	1	1	1	1	1	1
PCC2,3	1	1	1	1	1	1
PCC2,4	0	0	0	1	1	1
PCC3,4	1	1	1	0	0	0

(b)

Table 1: (a) Centroids of major clusters for G1 and (b) PCCs between these major clusters

LPCC is fed by data that is sampled from the observed variables in the unknown model. LPCC clusters the data using SOM (although any other clustering algorithm is good as well), and selects an initial set of major clusters (Section 4.3). Then, LPCC learns LVM in two stages. In the first stage, LPCC first identifies exogenous latent variables and latent colliders (without distinguishing them yet) and their corresponding observed descendants (Section 4.1) before distinguishing them (Section 4.2). LPCC iteratively improves the selection of the major clusters (Section 4.3), and the entire stage is repeated until convergence. In the second stage, LPCC identifies endogenous latent non-colliders with their children. Because this stage cannot distinguish from the outset between latent non-colliders and their latent ancestors, LPCC also needs to apply a mechanism to split these two types of latent variables from each other and to find the links between them after the split (Section 4.4). A flowchart of the LPCC algorithm is given in Figure 2.

[6] We remind that we determined three indicators per latent in all true models we demonstrate their learning (Figure 1) because BPC requires three indicators per latent to identify that latent; which makes the experimental evaluation we did in Part II of the paper fair.

[7] This is only for demonstration purposes. Part II of the paper shows evaluation results also for ternary latent variables and observed variables of different dimensions.

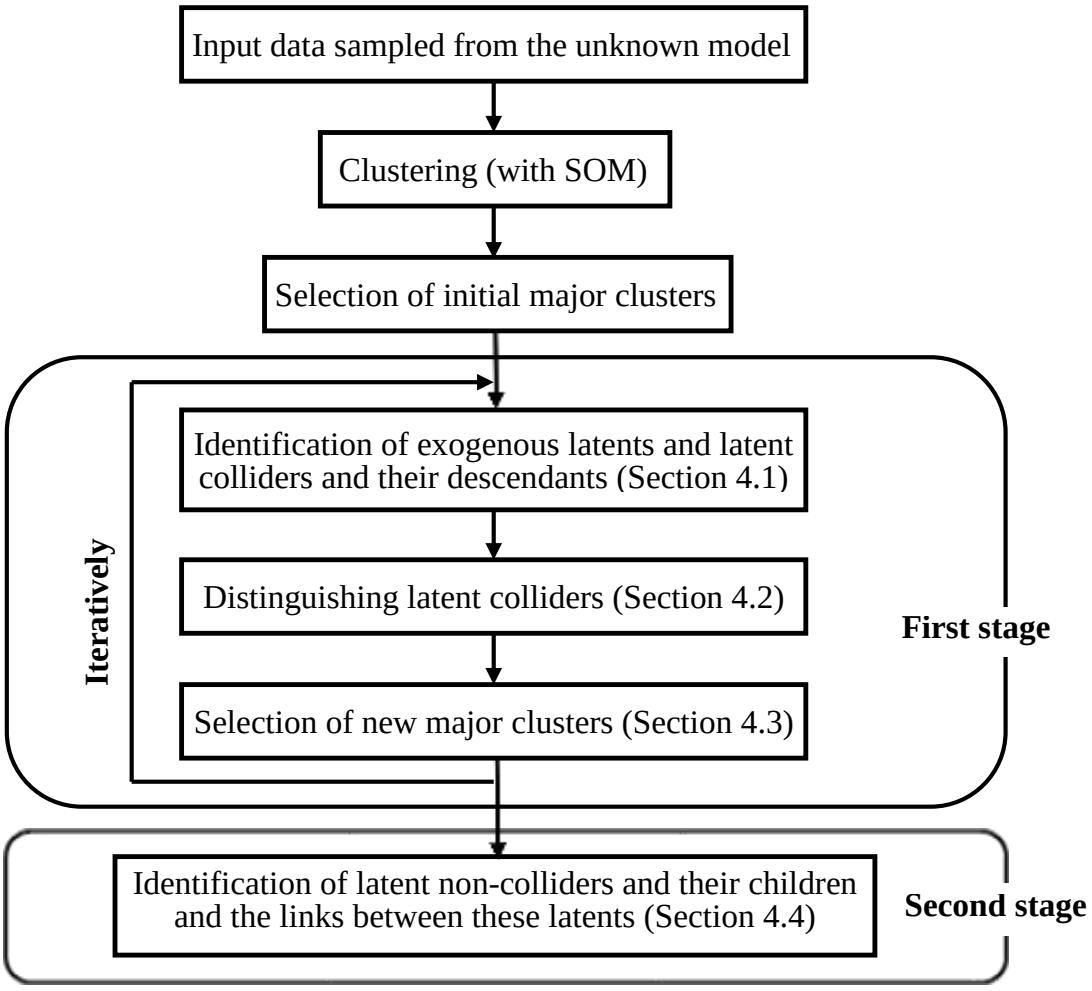

Figure 2: An overview of the LPCC algorithm.

4.1 Identification of exogenous latent variables and latent colliders and their descendants

Table 1b shows that *PCC1,2* (and *PCC3,4*) provides evidence that X1, X2, and X3 may be descendants of the same exogenous latent (L1, as we know) that has changed its value between the two **ex**s represented by *C1* and *C2* (and *C3* and *C4*). Relying only on one PCC may be inadequate when concluding that these variables are descendants of the same exogenous latent because there may be other exogenous latents that have changed their values too. Table 1b shows that *PCC2,3* (and *PCC1,4*) provides the same evidence about X1, X2, and X3. But, *PCC2,3* and *PCC1,4* also show that the values corresponding to X4, X5, and X6 have been changed together too, whereas these values did not change in *PCC1,2* and *PCC3,4*. Does this mean that X4, X5, and X6 are also descendants of the same latent ancestor of X1, X2, and X3? If we combine the two pieces of evidence provided by, e.g., *PCC1,2* and *PCC2,3*, we can answer this question with a "no". This is because X4, X5, and X6 changed their values only in *PCC2,3* but not in *PCC1,2*, and thus they cannot be descendants of L1. This insight strengthens the evidence that X1, X2, and X3 are the only descendants of L1. A similar analysis using *PCC1,3* and *PCC2,4* will identify that X4, X5, and X6 are descendants of another latent variable (L2, as we know). Therefore, we define:

Definition 16 *A maximal set of observed (**MSO**) variables is the set of variables that always changes its values together in each major–major PCC in which at least one of the variables changes value.*

That is, there is a particular interest in identifying the **MSO**s that always change their values together in each major–major PCC in which at least one of the variables changes value. For example, X1 (Table 1) changes its value in *PCC1,2*, *PCC1,4*, *PCC2,3*, and *PCC3,4* and always together with X2 and X3 (and vice versa). Thus {X1, X2, X3} (and similarly {X4, X5, X6}) is an **MSO**. Each **MSO** includes descendants of the same exogenous latent variable *L*, and after considering all PCCs, LPCC identifies an **MSO** for each exogenous latent variable.

Based on any identified **MSO**, LPCC introduces to the learned graph a new latent variable *L* together with all the observed variables that are included in this **MSO** as its children. At this stage, LPCC cannot yet distinguish between exogenous latents and latent colliders since the main goal at this stage is to identify latent variables. For now, LPCC focuses on the identification of the relations between the latents and the observed variables, but not on the identification of the interrelations between the latents. The latter task that is needed for distinguishing the latent colliders from the exogenous latents is performed in a further step (Section 4.2). Note, however, that the identification of endogenous latent non-colliders needs a different analysis that is based on major–minor PCCs and not on major–major PCCs, and thus it is described separately in Section 4.4.

The following Theorem 1 helps us formalize this identification step. For this theorem, we also need Definition 17 of equivalence relation/classes from set theory and Lemma 3, which is important by itself and for better understanding of LPCC, but also for proving Theorem 1.

Definition 17 *A given binary relation (i.e., between two elements) $\sim$ on a set $\boldsymbol{A}$ is said to be an equivalence relation if and only if it is reflexive ($a \sim a$), symmetric (if $a \sim b$ then $b \sim a$), and*

transitive (if $a \sim b$ and $b \sim c$, then $a \sim c$) for all a, b, and c in $\boldsymbol{A}$. The equivalence class of a under $\sim$, denoted $[a]$, is defined as: $[a] = \{b \in \boldsymbol{A} \mid b \sim a\}$ (Enderton, 1977).

Note that every two equivalence classes are either equal or disjoint. Therefore, the set of all equivalence classes of **A** forms a partition of **A**; every element of **A** belongs to one and only one equivalence class. It follows from the properties of an equivalence relation that: $a \sim b$ if and only if $[a] = [b]$. The following Lemma 3 is important since it shows that each **MSO** is an equivalence class, and thus **MSO**s corresponding to the learned latents are disjoint. At this stage, LPCC learns a set of at least two observed variables corresponding to a specific **MSO** for each latent where none of the observed variables is shared with other **MSO**s for other latents; in other words, a pure measurement model.

Lemma 3 *The relation "always changes together with" on the set **O** of all observed variables, such as "variable $O_i \in \mathbf{O}$ always changes together with variable $O_j \in \mathbf{O}$ in each PCC in which either O_i or O_j has changed" is an equivalence relation. Each equivalence class for this relation comprises an **MSO**.*

Proof All three conditions that are required for a binary relation to become equivalence are met:

1. O_i always changes with O_i (trivial).
2. If O_i always changes with O_j, then O_j always changes with O_i.
3. If O_i always changes with O_j, and O_j always changes with O_k, then O_i always changes with O_k.

Thus, the set of observed variables in a model can be represented by a set of equivalence classes for this relation, where each equivalence class includes all the variables that have the same equivalence relation, such as an **MSO**. ■

Theorem 1 *Variables of a particular **MSO** are children of a particular exogenous latent variable EX or its latent non-collider descendant or children of a particular latent collider C.*

Note that Theorem 1 guarantees that each of multiple latent variables (either an exogenous or any of its non-collider descendants or a collider) is identified by its own MSO, regardless of the latent cardinality.

4.2 Distinguishing latent collider variables

After identifying the exogenous latents and latent colliders together (Section 4.1), we need now to separate them. To demonstrate our concept for distinguishing latent colliders, we use graph G2 in Figure 1, which shows two exogenous latent variables, L1 and L3, that collide in one endogenous latent variable, L2. We assume that all latent variables are binary, [8] and each has three binary observed children X1, X2, and X3 (L1), X4, X5, and X6

[8] See footnote 7.

(L2), and X7, X8, and X9 (L3). Having two exogenous binary variables, we expect to find four major clusters in the data generated from G2. Each cluster will correspond to one of the four possible **ex**s (L1L3= 00, 01, 10, 11). In this case, as for G1 that was analyzed in the introduction to Section 4, we expect the values of X1, X2, and X3 to change together in all the PCCs following a change in the value of L1 , and the values of X7, X8, and X9 to change together in all the PCCs following a change in the value of L3. However, the values of X4, X5, and X6 will change together with those of X1, X2, and X3 in part of the PCCs and together with those of X7, X8, and X9 in the remaining PCCs, but always together in all of the PCCs. This will be evidence that X4, X5, and X6 are descendants of the same latent variable (L2, as we know), which is a collider of L1 and L3.

So far, LPCC learned latent variables but could not distinguish between exogenous latents and latent colliders (learning latent non-colliders will be described in Section 4.4). To learn that an already learned latent variable L is a collider for a set of other already learned (exogenous) latent ancestor variables $\mathbf{LA}\subset\mathbf{EX}$, LPCC requires that: (1) The values of the children of L will change with the values of descendants of different latent variables in **LA** in different parts of major–major PCCs; and (2) The values of the children of L will not change in any PCC unless the values of descendants of at least one of the variables in **LA** change. This insures that L does not change independently of latents in **LA** that are L's ancestors. We formalize this identification step in Theorem 2:

Theorem 2 *A latent variable L is a collider of a set of latent ancestors* $\mathbf{LA}\subset\mathbf{EX}$ *only if:*

1. *The values of the children of L change in different parts of some major–major PCCs each time with the values of descendants of another latent ancestor in* ***LA****; and*
2. *The values of the children of L do not change in any PCC unless the values of descendants of at least one of the variables in* ***LA*** *change too.*

4.3 Strategy for choosing major clusters

In this problem of unsupervised identification of latent variables given only observational data, LPCC has to deal with a lack of prior information regarding the distribution of each latent variable. Therefore, in its first iteration, LPCC assumes a uniform distribution over the latents and selects the major clusters based only on cluster size, which is the number of patterns clustered by the cluster. Clusters that are larger than the average cluster size are selected as majors. However, this initial selection may generate false negative errors (i.e., deciding a major cluster is minor). This may happen when a latent variable L has a skewed distribution over its values due to a low probability of L to take on any of its rare values. Then, the value configuration **ex** for which $L{=}v$, where v is a rare value, will be represented only by small clusters that could not be chosen as majors, although at least one of them should be major in representing v.

In addition, the initial selection may perform a false positive error (i.e., deciding a minor cluster is major), e.g., as a result of a very weak influence of L on any of its children (observed variables) X_i. In the discrete case, this weak influence can be represented as (almost) equal conditional probabilities of an observed variable to take on two different values $v_1{\neq}v_2$ given the same value v of its latent parent, $P(X_i{=}v_1 \mid L{=}v){\cong}P(X_i{=}v_2 \mid L{=}v)$. This may lead to splitting a data cluster that represents a configuration in which $L{=}v$ into

two clusters with almost the same size, and, when enough samples exist in both clusters, accepting both as major clusters instead of only one. For example, consider G1 in Figure 1, where all the variables are binary. Suppose that $P(X_2 = 0 \mid L1 = 0) = 0.6$ and $P(X_2 = 1 \mid L1 = 0) = 0.4$. This may split the cluster representing the configuration L1L2=00 into two clusters; in the first cluster $X_2 = 0$ and in the second cluster $X_2 = 1$. Due to the similar probabilities, both clusters may have approximately the same size, and if enough samples exist for L1L2=00 these two clusters may be larger than the average cluster size. Therefore, both may be accepted as major clusters in the initial selection. Recall that each **ex** should be represented by a single major cluster, which is the cluster that reflects the major effect of **ex** on the observed variables. In the example, only the cluster in which $X_2 = 0$ should be a major cluster, but due to the similar probabilities a false positive error could occur by also accepting the cluster in which $X_2 = 1$ as major.

To avoid these possible errors due to skewed data and circumstances that undermine identifiability, LPCC decides on major clusters iteratively. After learning a graph based on the initial selection of major clusters based on their sizes, it becomes possible to learn the cardinalities of the latent variables and consequently to find all possible **ex**s (Section 4.1). Then, for each **ex**, we can select the most probable cluster given the data and use it as an update to the major cluster that represents this **ex**. Using an EM-style procedure (Dempster et al., 1977), the set of major clusters can be updated iteratively and probabilistically and augment LPCC to learn more accurate graphs (see Section 2.1 in Part II for more details). This process can be repeated until convergence to a final graph (Figure 2). Since the final graph depends on the initial graph, the iterative approach cannot guarantee finding the optimal model, but only improving the initial graph.

4.4 Identification of latent non-collider variables

So far (Section 4.1), based on major–major PCCs, all the endogenous latent non-colliders that are descendants of an exogenous variable *EX* were temporarily combined with *EX*, and all the observed children of these latent non-colliders were temporarily combined with the direct children of *EX*. Thus, to identify latent non-colliders, LPCC needs to split them from their previously learned ancestor together with their observed children. We suggest that this identification stage be based on major–minor PCCs (recall that latent colliders were already identified separately, as described in Section 4.2).

To exemplify this need, let us observe G3 in Figure 1, which shows a serial connection of three latent variables L1, L2, and L3. Assume each of the latents is binary and has three binary observed children. L1 is the only *EX* with two possible ***ex***s (L1= 0, 1), and L2 and L3 are *NC*s; L2 is a child of L1 and a parent of L3. We synthetically generated a random data set of 1,000 patterns from G3 over the nine observed variables. We set the probabilities of: 1) L1 uniformly; 2) an observed child X_i, i= 1,…, 9, given its latent parent L_k, k= 1, 2, 3 (only if L_k is a direct parent of X_i, e.g., L1 and X1), as $P(X_i{=}v \mid L_k{=}v) = 0.8, v = 0, 1$; and 3) an endogenous latent L_j, j= 2, 3, given its latent parent L_k, k= 1, 2 (only if L_k is a direct parent of L_j, e.g., L1 and L2), as $P\left(L_j{=}v \;\middle|\; L_k{=}v\right) = 0.8, v = 0, 1$. Table 2 presents the seventeen largest clusters using their centroids and sizes, from which *C1* and *C2* were selected as major clusters (initially, *C1–C6* were selected, because they are larger than the average cluster size of 21, but then the iterative strategy described in Section 4.3 left only *C1* and *C2* as

major clusters). This meets our expectation of two major clusters corresponding to the two possible **ex**s of L1. However, because all the elements in *PCC1,2* are 1s (compare *C1* and *C2* in Table 2), the nine observed variables establish a single **MSO** and by Theorem 1 are considered descendants of the same exogenous variable. That is, the model G0 learned in the first phase of LPCC has only one exogenous latent variable (i.e., L1), and all of the nine observed descendants are learned as its direct children, which is contrary to G3. Since L2 and L3, which are latent non-colliders that are descendants of L1 in G3, were combined in G0 with L1, LPCC should split them from L1 along with their observed children in order to learn the true graph.

Thus, in the second phase, LPCC tests the assumption that G0 is true. If the assumption is rejected, LPCC infers that an exogenous latent *EX* has latent non-collider descendants, which were temporarily joined to *EX* in the first phase, and hence splits them from *EX*. To be able to reject the assumption about the correctness of G0, and thereby identify a possible split of an exogenous latent *EX*, we first define a first-order minor cluster (1-MC).

Centroid	***X1***	***X2***	***X3***	***X4***	***X5***	***X6***	***X7***	***X8***	***X9***	**size**
C1	1	1	1	1	1	1	1	1	1	49
C2	0	0	0	0	0	0	0	0	0	47
C3	1	1	1	1	1	1	1	1	0	28
C4	0	0	0	0	0	0	0	1	0	24
C5	0	1	0	0	0	0	0	0	0	22
C6	1	1	1	1	1	1	0	0	0	22
C7	0	0	1	0	0	0	0	0	0	21
C8	0	0	0	1	1	1	1	1	1	19
C9	0	0	0	0	0	0	1	1	1	18
C10	1	1	1	0	0	0	0	0	0	16
C11	0	0	0	1	0	0	0	0	0	14
C12	0	0	0	0	0	0	1	0	0	14
C13	1	0	1	1	1	1	1	1	1	14
C14	1	1	1	0	1	1	1	1	1	14
C15	1	0	0	0	0	0	0	0	0	13
C16	1	1	1	1	1	1	0	1	1	12
C17	0	0	0	0	0	1	0	0	0	12

Table 2: The seventeen largest clusters for G3 represented by their centroids and sizes

A 1-MC is a cluster that corresponds to a 1-order minor value configuration (Definitions 13 and 14), which exists when exactly one endogenous variable in **EN** (either latent or observed) has a minor local value (Definition 13) as a response to a value $ex \in \mathbf{ex}$ that $EX \in \mathbf{EX}$ has obtained. By analyzing, for each exogenous *EX*, PCCs between 1-MCs and the major clusters that identified *EX*, LPCC reveals the existence of the latent non-colliders that were previously combined with *EX* (Section 4.1). Following that, LPCC splits these non-colliders from *EX*. We will show that if only one observed variable changes in such PCCs (e.g., X9 in *PCC1,3* in Table 3; *C1* is major and *C3* is 1-MC) as a response to *ex*, then the minor value in the 1-MC is of an observed descendant of *EX*. And, if two or more observed variables change in such PCCs (e.g., X7-X9 in *PCC1,6* in Table 4; *C1* is major and *C6* is 1-MC) as a response to *ex*, then the minor value in the 1-MC is due to a minor value of a

latent non-collider descendant of *EX*. Thus, PCCs between 1-MCs and major clusters that show a change in the values of two or more observed variables provide evidence of the existence of an *NC* that should be split from its exogenous ancestor. Following, we describe how LPCC finds the set of 1-MCs. Then, we elaborate why and how the analysis of the PCCs between 1-MCs and major clusters is used to identify and split latent non-colliders from their exogenous ancestor.

PCC	**δX1**	**δX2**	**δX3**	**δX4**	**δX5**	**δX6**	**δX7**	**δX8**	**δX9**
PCC1,3	0	0	0	0	0	0	0	0	1
PCC2,3	1	1	1	1	1	1	1	1	0

Table 3: PCCs for *C3* with *C1* and *C2* (Table 2) in learning G3

PCC	**δX1**	**δX2**	**δX3**	**δX4**	**δX5**	**δX6**	**δX7**	**δX8**	**δX9**
PCC1,6	0	0	0	0	0	0	1	1	1
PCC2,6	1	1	1	1	1	1	0	0	0
PCC1,8	1	1	1	0	0	0	0	0	0
PCC2,8	0	0	0	1	1	1	1	1	1
PCC1,9	1	1	1	1	1	1	0	0	0
PCC2,9	0	0	0	0	0	0	1	1	1
PCC1,10	0	0	0	1	1	1	1	1	1
PCC2,10	1	1	1	0	0	0	0	0	0

Table 4: All 2S-PCCs for G3

To find the set of 1-MCs, LPCC first calculates a threshold on the maximal size of 2-order minor clusters (2-MCs). This threshold represents the maximal size of a minor cluster that corresponds to a 2-order minor value configuration, i.e., a minor cluster that represents exactly two endogenous variables in **EN** that have minor values (Definition 13). This threshold is an approximation for the maximal probability of having minor values as a response to any *ex* in exactly two descendants of *EX*, where all other descendants of *EX* in **EN** have major values. This approximation is derived from the product of the maximal minor local effects (Definition B.1 in Appendix B) of two observed descendants of *EX* and the maximal major local effects (Definition B.1) of the other observed descendants in **EN** (Appendix B). Thus, the sizes of all 1-MCs lie between the maximal size of a 2-MC (i.e., the threshold) and the minimal size of a major cluster (note that a major cluster is also a zero-order minor cluster corresponding to a zero-order minor value configuration). For example, based on the analysis above, *C2* is the minimal major cluster in learning G3, and all the fifteen clusters (Table 2) that are smaller than *C2* and larger than the threshold (calculated as 11), i.e., *C3*-*C17*, are 1-MCs. Note that this procedure is separately applied to each $EX \in \mathbf{EX}$. That is, for each *EX*, there is a different set of 1-MCs, each representing a single minor value of a descendant of *EX* and used to identify this descendant, whereas the other descendants of *EX* have major values.

Recall that every 1-MC corresponds to a 1-order minor value configuration that is due to exactly a single minor value of either an observed variable *O* or a latent non-collider *NC*, where both *O* and *NC* are descendants of *EX* in **EN**. The main difference between

these two cases is that in the former, the minor value in *O* is reflected only in this value, whereas in the latter, the minor value in *NC* may affect the values of all descendant latents of *NC* together with those of all the direct children (observed variables) of *NC* and its descendant latents. A minor value in *O* is identified based on the probability of this value conditioned on a certain value of *O*'s direct parent that is smaller than the maximal probability achieved for another value of *O* (i.e., the major value) conditioned on the same value of *O*'s direct parent. This happens for each value of the direct parent and does not require a change in *EX* to happen. From definition, a minor value in *O* in a 1-order minor value configuration can only happen when all *EX*'s descendants, except *O*, obtain major values. Although the mechanism of obtaining a minor value in a latent descendant *NC* of *EX* is similar to that in *O*, the impact of such a minor value is not locally restricted to *NC*, as for *O*, but it simultaneously affects all the descendants (latent and observed) of *NC*, which again, from definition, obtain major values.

We are only interested in the second case of minor values of *NC*, because their identification helps split this *NC* from its ancestor *EX* to which it was initially combined (Section 4.1). Since the observed variables in both cases are among *EX*'s descendants, which were already used to identify *EX*, it is a challenge to distinguish between them. Following, we analyze 1-MCs to identify these two cases and concentrate on the second case.

Case 1: A minor value of an observed variable

When comparing, for a specific *EX*, two centroids – one of a major cluster and the other of a 1-MC that corresponds to an observed minor value configuration (Definition 11) in which an observed variable *O*, which is a descendant of *EX*, has a minor value – we can observe that when:

1. *EX* changes values between two **ex**s that correspond to the compared clusters, all observed descendants of *EX*, except *O*, change values together,

   and when

2. *EX* does not change values between two **ex**s that correspond to the compared clusters, the only observed descendant of *EX* that changes value is *O*.

Thus, a PCC – between the centroid of such 1-MC and a centroid of any of the major clusters – that shows the same value for all but one (i.e., *O*) of the observed descendants of *EX* (i.e., either 1 if *EX* changes values in the corresponding **ex**s or 0 if it does not) identifies a minor value in *O*. For example, in Table 3, *PCC1,3* and *PCC2,3* of *C3*, which is a 1-MC, with the two major clusters *C1* and *C2* (Table 2) show the set of observed variables X1–X8 that either do or do not change values together, whereas the single observed variable X9 acts contrariwise. This is evidence that *C3* is a 1-MC due to exactly a single minor value of an observed variable descendent (X9) of L1 in G3. Such an analysis helps LPCC ignore, on the one hand, observed descendants of L1 that cannot reflect minor values in L1's latent (non-collider) descendants, and focus, on the other hand, on the latent descendants that should be split from L1, as part of Case 2.

Case 2: A minor value of a latent non-collider

The minor value of a latent non-collider *NC*, which is a descendent of *EX*, can be reflected only via the values of its observed descendants in an observed minor value configuration that is represented by a certain 1-MC. By definition, all of these observed descendants

have major values in this 1-order minor configuration since only *NC* has a minor value in this configuration. The major value of each of these observed descendants is certain given the minor value of *NC* (Proposition 2) and different from the certain major value it would have if *NC* had a major value (Assumption 7) instead of its minor value.

When comparing for a specific *EX*, two centroids – one of a major cluster and the other of a 1-MC that corresponds to an observed minor value configuration in which a latent non-collider *NC*, which is a descendant of *EX*, has a minor value – we can observe that when:

1. *EX* changes values between two **ex**s that correspond to the compared clusters, all observed descendants of *EX*, but not observed descendants of *NC*, change values together,

   and when

2. *EX* does not change values between two **ex**s that correspond to the compared clusters, the only observed descendants of *EX* that change values are those of *NC*.

Thus, a PCC – between the centroid of such 1-MC and a centroid of any of the major clusters – that shows two sets of two or more observed variables, each set having a different value, identifies a minor value in *NC*. The first set in such a PCC comprises the descendants of *NC* (with a value of 0 if *EX* changes values in the corresponding **ex**s or 1 if it does not), and the second set comprises all other observed variables that are descendants of *EX*, but not *NC* (with a value of 1 if *EX* changes values in the corresponding **ex**s or 0 if it does not). For example, *PCC1,6* and *PCC2,6* (Table 4) of *C6*, which is a 1-MC, with the two major clusters *C1* and *C2* (Table 2), show two sets of observed variables for G3. The first set consists of X1–X6 and the second of X7–X9. This is evidence that *C6* is a 1-MC due to a minor value of a latent non-collider descendant of L1, and L1 should be split into two latents (each is responsible for one of the two sets). One latent (which we know is L3) is a parent of X7, X8, and X9, and the other latent is a parent of X1–X6 (which we will show is also split to L1 and L2, each with its three children).

Distinguishing between Case 1 and Case 2 gives us an instrument to identify latent non-colliders. We are interested in PCCs between 1-MCs and major clusters that show two sets of two or more elements corresponding to the observed variables. Variables in each set have the same value, which is different than that of the other set. Following, we infer that each set is of a different latent than the one that was expected to be sole. We denote such PCC by 2S-PCC (i.e., PCC of "two sets") and the corresponding 1-MC by 2S-MC (Definition 18). Thus, to identify a latent non-collider that was combined to an exogenous latent *EX*, we consider only the 2S-PCCs; these PCCs are the result of comparing all the 2S-MCs among the 1-MCs for *EX* with the major clusters that revealed *EX*. Table 4 represents all 2S-PCCs for G3.

Definition 18 *2S-PCC is PCC between 1-MC and a major cluster that shows two sets of two or more elements corresponding to the observed variables. Elements in each set have the same value, which is different than that of the other set. Accordingly, this 1-MC is defined as 2S-MC.*

The following Theorem 3 helps formalize this identification step, but to prove this theorem, we first need Lemma 4. Recall that the challenge here is to identify a latent non-collider *NC* that is a descendant of an exogenous latent *EX*, but was wrongly combined with this exogenous ancestor. To face this challenge, we need to find a circumstance in which *EX* and *NC* are involved that is different than that which led to the inability to distinguish between them. *NC* could not be distinguished from *EX* when we analyzed major value configurations. But, although a major value configuration is the most probable configuration (Definition 9), minor value configurations are possible too – according to the probability tables of the latents, each given its direct parent – albeit less likely. A minor value configuration in which only *NC* takes a minor value (i.e., a first-order minor value configuration) is exactly what we need.[9] This is because all *NC*'s latent ancestors, in the first-order minor value configuration, take the same major values they took in the major value configuration and thus influence their descendants the same. But, the minor value *NC* takes influences its (latent and observed) descendants differently than the major value *NC* took in the major value configuration. This influence is revealed in the different values the observed children of *NC* and its descendants take compared to the values they took when *NC* had a major value. Since the two value configurations are represented in two corresponding clusters – a major cluster and a 2S-MC for *NC* – the signature of *NC* can uniquely be detected by comparing the two clusters using 2S-PCC.[10]

Lemma 4 shows that it is possible to identify *NC* because: 1) Even when *EX* leads to major values in all *NC*'s ancestors (and in most cases also in *NC*), *NC* can still take a minor value; and 2) even when *EX* changes values, leading all *NC*'s ancestors to change values as well, *NC* can still keep the same (minor) value. Thereby, minor value configurations for *NC* demonstrate its autonomy, enabling its identification and its split from *EX*.

Lemma 4 *Let a latent non-collider NC be a descendant of an exogenous latent variable EX. 2S-PCC is a PCC between a "two-set" first-order minor cluster 2S-MC due to a minor value in NC and a major cluster that identified EX.* $\mathbf{ex}'$ *and* $\mathbf{ex}''$ *are two value configurations of* $\boldsymbol{EX}$ *that correspond to the compared clusters by 2S-PCC. When:*

1. *EX does not change values between* $\mathbf{ex}'$ *and* $\mathbf{ex}''$*, all the elements in 2S-PCC corresponding to the observed descendants of the latent ancestors of NC (including EX) show no change (i.e., are 0), whereas the elements corresponding to the observed descendants of NC show a change (i.e., are 1),*

   and when

2. *EX changes values between* $\mathbf{ex}'$ *and* $\mathbf{ex}''$*, all the elements in 2S-PCC corresponding to the observed descendants of the latent ancestors of NC (including EX) show a change (i.e., are 1), whereas the elements corresponding to the observed descendants of NC show no change (i.e., are 0).*

[9] All other first-order minor value configurations (due to other latent variables, which are also EX's descendants) or k-order minor value configurations (Definition 13) due to EX are irrelevant to the identification of NC, although the former – as will be shown in Theorem 3 – play a role in determining the direct observed children of NC among its observed descendants.

[10] Any 2S-PCC, which is detected for EX, will point to the NC that corresponds to the 2S-MC that is compared by this 2S-PCC.

Before moving to Theorem 3, let us illustrate the two cases discussed in Lemma 4 for G3. The "*EX* does not change values between **ex**′ and **ex**′′" case can be demonstrated, for example, when comparing *C1* and *C6* (Table 2). In response to *EX*(=L1)=1, *NC*'s (L3) parent (L2) takes a major value of 1 in both the value configurations of the latent variables in response to **ex**′ = **ex**′′.[11] Also, L3 takes a major value of 1 in the configuration that is represented by *C1*, which is one of the two major clusters. But, L3, in response to the same configuration of its latent ancestors (L1 and L2), takes a minor value of 0 in the value configuration that is represented by the 2S-MC *C6*. By comparing *C1* and *C6*, the corresponding 2S-PCC (i.e., *PCC1,6*; see Table 4) shows two sets of elements: the first of 0s that correspond to the observed variables X1–X6, which do not change values between the clusters, and the second of 1s that correspond to X7–X9, which do change values between the clusters. This is the evidence we are looking for that is needed to identify L3.

The "*EX* changes values between **ex**′ and **ex**′′" case can be demonstrated, for example, when comparing *C1* and *C9* (Table 2). In response to *EX*(=L1)=1 and *EX*(=L1)=0, *NC*'s (L3) parent (L2) takes a major value of 1 in response to L1=1 and a major value of 0 in response to L1=0. In the first instance, L3 takes a major value of 1 to create the major configuration that is represented by *C1*, and in the second instance, L3 takes a minor value of 1 in the value configuration that is represented by the 2S-MC *C9* (and although the first value is major and second is minor, they are both 1). By comparing *C1* and *C9*, the corresponding 2S-PCC shows two sets of elements, the first of 1s that correspond to the observed variables X1–X6, which changed values between the clusters, and the second of 0s that correspond to X7–X9, which did not change values between the clusters. This is additional support of the existence of L3. However, relying only on part of the 2S-PCCs may be inadequate to conclude on all possible splits. For example, *PCC1,8* and *PCC2,8* (Table 4) show that X1–X3 and X4–X9 are children of different latents, but do not suggest the split of X7–X9 as *PCC1,6* and *PCC2,6* do. Therefore, similarly to the **MSO** concept that was introduced for major–major PCCs to identify exogenous latents, it is necessary to introduce also for 2S-PCCs a *maximal set of observed variables* (**2S-MSO**) that always change their values together in all 2S-PCCs. We define:

Definition 19 *A **2S-MSO** is the maximal set of observed variables that always change their values together in all 2S-PCCs.*

For example, X1 in Table 4 changes its value in *PCC2,6*, *PCC1,8*, *PCC1,9*, and *PCC2,10* and always together with X2 and X3 (and the other way around). Thus, {X1, X2, X3} and similarly {X4, X5, X6} and {X7, X8, X9} are **2S-MSO**s. Each **2S-MSO** includes children of the same latent non-collider, which is a descendant of *EX*, or *EX* itself. After computing all 2S-PCCs for *EX*, LPCC detects **2S-MSO**s for all these latent variables and thereby identifies all possible splits for *EX*. Note that compared to **MSO** (Section 4.1), which is identified in major–major PCCs to reveal exogenous latents, **2S-MSO** is identified in PCCs between 2S-MCs and major clusters to reveal splits of latent non-colliders from the exogenous latent that was previously learned using these major clusters.

[11] Note that the values the three latents take in the two-value configurations can only be inferred from the values their children (X1–X3 for L1, X4–X6 for L2, and X7–X9 for L3) take.

Theorem 3 *Variables of a particular **2S-MSO** are children of an exogenous latent variable EX or any of its descendant latent non-colliders NC.*

After splitting the latent non-collider descendants from their exogenous latent ancestor *EX*, we need to identify the links between these latents. To identify these links, LPCC exploits the following Proposition 10 and Theorem 4. We will see that in the case of a serial connection, LPCC learns the undirected links among the latents, and in the case of a diverging connection, LPCC learns the directed links among the latents. That is, LPCC learns a pattern over the structural model of $\boldsymbol{G}$, which represents a Markov equivalence class of models among the latents. In the special case where $\boldsymbol{G}$ has no serial connection, LPCC learns the true graph.

Proposition 10 *In 2S-PCCs in which only the observed children of a single latent change, the latent is*

1. *EX or its leaf latent non-collider descendant, if the connection is serial; or*
2. *EX's leaf latent non-collider descendant, if the connection is diverging.*

Proof We already showed that at least a single 2S-PCC exists in the serial connection case in which only the observed children of *EX* change (Theorem 3). In addition, in the proof of Theorem 3 (Part II), we already showed that for any *NC* that is a latent non-collider descendent of *EX*, *NC*'s observed children change values in some 2S-PCCs with observed children of a latent non-collider descendant of *NC* and in the other 2S-PCCs with observed children of a latent non-collider ancestor of *NC*, but never alone. A special case in the proof of Theorem 3 is when *NC* is a leaf. Then, at least a single 2S-PCC exists in which only the children of *NC* change. ■

We will exemplify Proposition 10 using G3. Table 4 shows all the 2S-PCCs for G3 from which we can identify three **2S-MSO**s: {X1,X2,X3}, {X4,X5,X6}, and {X7,X8,X9}. If we consider only 2S-PCCs due to *C1* (the first major cluster), {X1,X2,X3} change alone in *PCC1,8*, and {X7,X8,X9} change alone in *PCC1,6*. By Proposition 10, these two **2S-MSO**s are observed children of an exogenous latent variable *EX* and its leaf latent non-collider descendant. From knowing G3, we know that these two latents are L1 and L3. Note that if more than a single leaf of *EX* exists (i.e., in the case of a diverging connection emerging from *EX*), then for each such leaf, there is a 2S-PCC in which only the observed children of this leaf change alone. This will help LPCC to identify a diverging connection and determine *EX* as the source in all paths leading to the leaves (sinks). As a result, LPCC could identify the correct direction of the links among the latents.

Proposition 10 guarantees that if the connection is serial, we find the source (*EX*) and sink of the path between them (but not who is who). To identify the directionality between any two latent non-collider variables on the path between the source and sink, we will need more. To motivate the need, suppose that when learning G3, we already identified L1 as *EX* and L3 as *EX*'s leaf descendant (Proposition 10), and now we have to split L2 from L1 using the two major clusters, *C1* and *C2* (Table 2), which revealed L1, and identify the directionality among these three latent variables. Lemma 4 (first part) guarantees that the

observed children of a latent non-collider *NC1*, which is a child of another non-collider *NC2* (both are descendants of *EX*), will change in all 2S-PCCs with the observed children of *NC2* except in a single additional 2S-PCC due to a minor value of *NC1*. That is, *NC1* is identified as a direct child of *NC2* if the observed children of *NC1* change in all 2S-PCCs (due to a specific major cluster and when *EX* does not change value), in which the children of *NC2* change plus an additional 2S-PCC in which they change without the children of *NC2*.[12] In our case, this means that the observed children of L3, which is a child of L2, will change values in all 2S-PCCs in which the observed children of L2 change values, and also in an additional 2S-PCC, which is due to a minor value in L3. Indeed, *PCC1*,*10* (Table 4), due to *C1*, shows that when *EX* does not change values and the observed children of L3, {X7,X8,X9}, change values, the observed children of L2, {X4,X5,X6}, also change values. In addition, *PCC1*,*6*, which is the result of comparing *C1* and 2-MC *C6* due to a minor value of L3, shows that {X7,X8,X9} change values without {X4,X5,X6} once. *PCC2*,*8* and *PCC2*,*9* demonstrate the same, when using major cluster *C2* instead of *C1* (and *C9* is the 2-MC that reveals the minor value of L3). This provides an indication that L3 is a child of L2.

But, Proposition 10 cannot guarantee distinguishing between *EX* and its leaf latent non-collider descendant (hereby a "leaf"); hence, what if we mistakenly identified them? In the G3 example, this means we identified L3 as *EX* and L1 as *EX*'s leaf. Lemma 4 demonstrates an interplay between *EX* and *NC* (and all of its descendants) as presented in 2S-PCCs due to a minor value in *NC*; when one of them changes, the other does not and vice versa. Because the leaf is one of *NC*'s descendants, Lemma 4 guarantees that the observed children of the leaf do not change if and only if *EX* changes value. That is, by the second part of Lemma 4, if *EX* changes, then the observed children of the leaf do not change. Thus, if we find 2S-PCCs that show that the observed children of the leaf do not change, then we have evidence that *EX* changed. This guarantees that the observed children of a latent non-collier *NC2* (or *EX* itself), which is a parent of another non-collider *NC1*, will change in all 2S-PCCs with the observed children of *NC1*, except in a single additional 2S-PCC due to a minor value of *NC2* (or if *NC2* is *EX*). In our case, this means that the observed children of L1, which is L2's parent, will change values in all 2S-PCCs in which the observed children of L2 change values, and also in an additional 2S-PCC. Indeed, *PCC1*,*9* (Table 4), due to *C1*, shows that when the leaf does not change value and the observed children of L1, {X1,X2,X3}, change values, the observed children of L2, {X4,X5,X6}, also change values. In addition, *PCC1*,*8* shows that {X1,X2,X3} change values without {X4,X5,X6} once. *PCC2*,*6* and *PCC2*,*10* demonstrate the same when using major cluster *C2* instead of *C1*. This provides an indication that L1 is a child of L2, which is the opposite direction between the two in G3. That is, the interplay between *EX* and its leaf lets LPCC identify the directionality between latent non-colliders on the path between

[12] Note that Lemma 4 makes a clear distinction between NC's ancestors (and their observed children) and NC's descendants (and their observed children), when NC gets a minor value. That is, all NC's ancestors follow EX (and change values or not with it) and all NC's descendants follow its change of value. This change of NC "breaks" the influence of EX on the latents on the path emerging from EX and "starts" NC's own influence on its latent descendants. And this is what is so important in finding the traces of minor values of endogenous latents through 2S-PCCs, that these traces identify the existence of the latents. Particularly, when EX does not change values and all its descendants get major values, the observed children of NC1 and NC2 will change together, and it is only a minor value that NC1 gets that can make a 2S-PCC in which NC1's observed children change without those of NC2, and thereby indicate that NC1 is NC2's child.

EX and the leaf, and in both directions. This means that LPCC can identify only the undirected links between the latents in the serial case.

In the diverging case, the children of *EX* never change alone, and every latent that its children change alone in some 2S-PCC is a leaf (Proposition 10). Therefore, by performing an analysis as for the serial case using 2S-PCCs in which the observed children of the leaf do not change for each leaf of the branches of the diverging connection, LPCC can identify the links among the latents in opposite directions on each branch. We formalize this by Theorem 4.

Theorem 4 *A latent non-collider NC1 is a direct child of another latent non-collider NC2 (both on the same path emerging in EX) only if:*

- *In all 2S-PCCs for which EX does not change, the observed children of NC1 always change with those of NC2 and also in a single 2S-PCC without the children of NC2; and*
- *In all 2S-PCCs for which a latent non-collider leaf descendant of EX does not change, the observed children of NC2 always change with those of NC1 and also in a single 2S-PCC without the children of NC1.*

LPCC uses Theorem 4 to identify the links between the split latents. In the serial connection, there are only two latents with observed children that change alone in some 2S-PCCs, that is, *EX* and its leaf latent non-collider descendant. However, LPCC cannot distinguish between them and thus finds all the links between these two latents as undirected. In the diverging connection, the observed children of *EX* never change alone (Proposition 10); thus, every latent with children that change alone in some 2S-PCCs can only be a leaf. Thereby, LPCC can identify the directed links among the latents repeatedly on each of the paths from *EX* to each of the leaves (Theorem 4). Still, LPCC needs to distinguish between the serial and diverging connections. In the case where the observed children of three or more latents change alone in some 2S-PCC, it is clear that it is a diverging connection. Then, LPCC treats these latents as leaves and returns directed paths from *EX* to each such leaf. However, in the case in which LPCC identifies that the observed children of exactly two latents change alone in some 2S-PCCs, it applies the analysis proposed in Theorem 4 to each of the latents. If it obtains the same path with opposite directions, then LPCC considers it as a serial connection and returns the undirected path; otherwise, it considers it as a diverging connection and returns the two directed paths from *EX*.

5. Discussion and Future Research

We introduced the PCC concept and LPCC algorithm for learning LVMs:

1. LPCC combines learning graphical models with data clustering by using the PCC concept to analyze clustering results of discrete variables for learning LVMs;
2. LPCC learns MIM, which is a large subclass of SEM. In MIM, multiple latent variables may have multiple indicators (observed children), and no observed variable may be an ancestor of any latent variable;

3. LPCC is not limited to latent-tree models, which are only a subclass of MIM, and does not make special assumptions, such as linearity, about the distribution;

4. LPCC assumes that the measurement model of the true graph is pure, but, if the true graph is not pure, LPCC learns a pure sub-model of the true model, if one exists. LPCC's only assumption about the structural model is that a latent collider does not have any latent descendants (a detailed list of assumptions LPCC makes is given in Appendix C);

5. LPCC is a two-stage algorithm. First, LPCC learns the exogenous latents and the latent colliders, as well as their observed descendants, by utilizing pairwise comparisons between data clusters in the measurement space that may explain latent causes. Second, LPCC learns the endogenous latent non-colliders and their children by splitting these latents from their previously learned latent ancestors;

6. LPCC learns an equivalence class of the structural model of the true graph; and

7. LPCC is formally expressed as an algorithm and evaluated using synthetic and real-world databases in Part II of the paper.

A number of open problems invite further research including:

1. Extending LPCC to identify observed variables that are effects of other observed variables;

2. Providing a formal analysis for the conditions of model identification and its sensitivity to parameterization. Learning by LPCC that an observed variable O is a descendent of a latent variable L depends on two factors. The first factor is the "graph distance", which means that the more edges that separate O from L, the less likely O would be grouped with other observed variables, descendants of L. The second factor is the conditional probabilities of an observed variable given its latent parent, which means that the stronger the probabilities are, the more likely the link will be identified by LPCC. Although the iterative strategy for choosing the major clusters (Section 4.3) improves the identification of observed children with weak associations with their latent parents, the final graph still depends on the initial graph. That is, the iterative approach alone cannot guarantee finding the optimal model. Future analysis should take into account both factors;

3. Analyzing LPCC complexity. Future research should dive into this topic and decompose LPCC complexity to those of clustering, identification of major–major PCCs, and identification of major–minor PCCs. Assume a set $\mathbf{V=(L\cup O)}$ with a variable maximal cardinality, $k = \max(|V_i|)$, a number of exogenous variables, $|\mathbf{EX}|$, and a number of major value configurations (major clusters), $|\mathbf{ex}| = k^{|\mathbf{EX}|}$. A preliminary analysis shows that LPCC complexity in identifying major–major PCCs is $O(|\mathbf{ex}|^2) = O(k^{2|\mathbf{EX}|})$. To compute the LPCC's complexity in identifying major–minor PCCs, we first have to identify 1-MC minor clusters (values), with complexity of $O((|V| - |\mathbf{EX}|)\ k^{|\mathbf{EX}|}(k-1))$ due to $(k-1)$ minor values for each of $k^{|\mathbf{EX}|}$ parent configurations of $(|V| - |\mathbf{EX}|)$ endogenous variables. Then, the complexity in identifying major–minor

PCCs is $O((|V| - |\mathbf{EX}|)k^{2|\mathbf{EX}|}(k-1))$, and the total complexity in computing PCCs is $O((|V| - |\mathbf{EX}|)k^{2|\mathbf{EX}|})$, which is exponential in $|\mathbf{EX}|$, but in most problems $|\mathbf{EX}| \ll |V|$. However, a more elaborated analysis that also includes the complexity of clustering is desired.

4. Exploring the impact of clustering – as is manifested by the clustering algorithm and its parameters – on the LPCC results. In Part II, we show a problem in which the data structure is hierarchical, and a clustering algorithm that is more sophisticated than SOM, which is suggested in Section 4, is needed to preprocess the data used to learn an LVM that is meaningful to the domain. Exploring the requirements on clustering and any guidelines about the best approach to take for clustering is a direction of further research; and

5. Suggesting ways to use the graphical model to cluster data points. Although we have established and exploited a link between cluster analysis and learning an LVM, in this work, we only studied learning (reconstructing) the graphical model by analyzing clusters of observational data. Another very interesting line of future research is in the opposite direction, extending previous studies such as that in Zhang (2004). Because MIM models learned by LPCC are richer than HLC models (which are only a subset of MIM), such a line of research may enable accurate clustering of observational data generated by a model also having collider nodes.

Acknowledgments

The authors thank Ricardo Silva from UCL for his helpful comments and suggestions given to improve an earlier version of this paper. The authors also thank the two anonymous reviewers for their comments and suggestions that helped strengthen the paper and the special issue editors: Isabelle Guyon and Alexander Statnikov. Nuaman Asbeh thanks the *Planning and Budgeting Committee* (PBC) of the *Israel Council for Higher Education* for its support by a scholarship for distinguished Ph.D. students.

Appendix A. Proofs of propositions, lemmas, and theorems

In this appendix, we give proofs of propositions, lemmas, and theorems for which the proof is too detailed, lengthy, or impedes the flow of reading. All other proofs are given in the body of the paper.

Lemma 1

1. *Each latent non-collider NC_t has only one exogenous latent ancestor EX_{NC_t}, and there is only one directed path T_{NC_t} from EX_{NC_t}(source) to NC_t (sink).*

   (Note that we use the notation NC_t, rather than S_t, since the Lemma applies to both exogenous and endogenous latent non-colliders.)

2. Each latent collider C_j is connected to a set of exogenous latent ancestors $\mathbf{EX}_{C_j}$ via a set of directed paths $\mathbf{T}_{C_j}$ from $\mathbf{EX}_{C_j}$ (sources) to C_j (sink).

Proof

1. If the latent non-collider is exogenous, $NC_t = EX_t$, then EX_{NC_t}=EX_t, and T_{NC_t} is the empty path consisting of EX_t. For example, EX_{L3}=L3 and T_{L3} = L3 in G2 and G5 in Figure 1. If, however, the latent non-collider is endogenous, $NC_t = S_t$, and we assume by contradiction that it has more than one exogenous latent ancestor and thus more than one directed path from each exogenous ancestor to S_t (and according to Assumption 5, none of the paths passes through a collider) that collide at S_t, then S_t is a collider. This is contrary to the assumption that NC_t is a non-collider. That is, EX_{S_t} is the only exogenous latent ancestor of S_t, and T_{S_t} is the only directed path from EX_{S_t} through S_t's parent Pa_t to S_t. For example, EX_{L5}=L3 and T_{L5} = {L3, L4, L5} in G5 (Figure 1).

   [Note that if S_t has no endogenous latent non-collider ancestors, then Pa_t=EX_{S_t} and T_{S_t} equals the ordered sequence $\{EX_{S_t}, S_t\}$, e.g., EX_{L4}=L3 and T_{L4} = {L3, L4} in G5 (Figure 1).]

2. Under Assumption 5, any parent Pa_j of latent collider C_j could be either a latent non-collider or an exogenous latent; in other words, $\mathbf{Pa}_j \subset (\mathbf{NC} \cup \mathbf{EX})$. If Pa_j is a latent non-collider, then it is on the directed path T_{C_j} from EX_{C_j} to C_j; and if Pa_j is an exogenous latent EX_{C_j}, then it is the source of a directed path T_{C_j}(or more than a single directed path) to C_j. $\mathbf{EX}_{C_j} = \cup EX_{C_j}$ is the set of exogenous ancestors of C_j, and $\mathbf{T}_{C_j} = \cup T_{C_j}$ is the set of directed paths from $\mathbf{EX}_{C_j}$ to C_j. For example, $\mathbf{EX}_{\mathrm{L4}}$={L1, L5} and $\mathbf{T}_{\mathrm{L4}}$ = {{L1, L2, L3, L4}, {L5, L4}} in G6 (Figure 1).

■

Proposition 3 *The $MAV_{\{TS_{NC_v} \setminus EX_{NC_v}, ONC_v\}}\left(ex_{NC_v}\right)$ corresponding to $MAE_{\{TS_{NC_v} \setminus EX_{NC_v}, ONC_v\}}\left(ex_{NC_v}\right)$ is a certain value configuration for each certain value ex_{NC_v}.*

(Note that here we use the notation NC_v rather than S_v since the proposition applies to both exogenous and endogenous latent non-colliders.)

Proof If the latent non-collider NC_v is exogenous, $NC_v = EX_v$ and ONC_v=OEX_v, then $\{TS_{NC_v} \backslash EX_{NC_v}, OEX_v\} = OEX_v$ and the partial major value is the local major value $MAV_{OEX_v}(ex_v)$, which by Proposition 2 is certain for a certain value ex_v.

If the latent non-collider NC_v is endogenous, $NC_v = S_v$ and ONC_v=OS_v, then we consider $\{TS_{S_v} \backslash EX_{S_v}, OS_v\}$, which is a set of ordered variables along the directed path T_{S_v} that ends in OS_v. The remainder of the proof is by induction:

Basis: Based on Proposition 2, $MAV_{S_1}(ex_{S_v})$, where S_1 is the first variable in $\{TS_{S_v} \backslash EX_{S_v}, OS_v\}$ and a direct child of EX_{S_v}, given a certain value ex_{S_v}, is also certain.

Step: If the major value of the ith variable, S_i, in the subset $\{TS_{S_v} \backslash EX_{S_v}, OS_v\}$, i.e., $MAV_{S_i}(pa_i^{ex_{S_v}})$, is certain for a certain value $pa_i^{ex_{S_v}}$, then the major value of the $(i+1)$th variable, S_{i+1}, in the subset (which is S_i's child), i.e., $MAV_{S_{i+1}}(pa_{i+1}^{ex_{S_v}})$, is by Proposition 2 certain too for a certain value $pa_{i+1}^{ex_{S_v}}$ (which is $MAV_{S_i}(pa_i^{ex_{S_v}})$). ■

Proposition 4 *All corresponding values in* $MAV_{\{TS_{NC_v} \backslash EX_{NC_v}, ONC_v\}}\left(ex'_{NC_v}\right)$ *and* $MAV_{\{TS_{NC_v} \backslash EX_{NC_v}, ONC_v\}}\left(ex''_{NC_v}\right)$*, for two values* ex'_{NC_v} *and* ex''_{NC_v} *of* EX_{NC_v}*, are different.*

(Here also we use the notation NC_v, since the proposition applies to both exogenous and endogenous latent non-colliders.)

Proof If the latent non-collider NC_v is exogenous, $NC_v = EX_v$ and ONC_v=OEX_v, then $\{TS_{NC_v} \backslash EX_{NC_v}, OEX_v\} = OEX_v$, and, by Assumption 7, the corresponding $MAV_{OEX_v}(ex'_v)$ and $MAV_{OEX_v}(ex''_v)$ are different for two values ex'_v and ex''_v.

If the latent non-collider NC_v is endogenous, $NC_v = S_v$ and ONC_v=OS_v, then we consider $\{TS_{S_v} \backslash EX_{S_v}, OS_v\}$, which is a set of ordered variables along the directed path T_{S_v} that ends in OS_v. The remainder of the proof is by induction:

Basis: The major local values $MAV_{S_1}(ex'_{S_v})$ and $MAV_{S_1}(ex''_{S_v})$ of the first variable, S_1, in $\{TS_{S_v} \backslash EX_{S_v}, OS_v\}$ (which is also a direct child of EX_{S_v}) and two values ex'_{S_v} and ex''_{S_v} of EX_{S_v} are different based on Assumption 7.

Step: If the major local values of the ith variable, S_i, in $\{TS_{S_v} \backslash EX_{S_v}, OS_v\}$ and two values ex'_v and ex''_v of EX_{S_v}, i.e., $MAV_{S_i}(ex'_{S_v})$ and $MAV_{S_i}(ex''_{S_v})$, are different, then the major local values of S_{i+1} (S_i's child), and the two values $MAV_{S_i}(ex'_{S_v})$ and $MAV_{S_i}(ex''_{S_v})$, i.e., $MAV_{S_{i+1}}\left(pa_{i+1}^{ex'_{S_v}}\right) = MAV_{S_{i+1}}\left(MAV_{S_i}(ex'_{S_v})\right)$ and $MAV_{S_{i+1}}(pa_{i+1}^{ex''_{S_v}}) = MAV_{S_{i+1}}\left(MAV_{S_i}(ex''_{S_v})\right)$ are different too based on Assumption 7. ■

Proposition 5 *EX_{NC_v} changes values (i.e., has two values ex'_{NC_v} and ex''_{NC_v}) if and only if NC_v changes values in the two corresponding major value configurations:* $MAV_{\{TS_{NC_v} \backslash EX_{NC_v}, ONC_v\}}\left(ex'_{NC_v}\right)$ *and* $MAV_{\{TS_{NC_v} \backslash EX_{NC_v}, ONC_v\}}\left(ex''_{NC_v}\right)$.

Proof ("if") Proposition 3 guarantees that NC_v has a certain value in $MAV_{\{TS_{NC_v} \backslash EX_{NC_v}, ONC_v\}}\left(ex_{NC_v}\right)$ for a certain value ex_{NC_v} of EX_{NC_v}. Thus, if NC_v has

different values in two $MAV_{\{TS_{NC_v} \setminus EX_{NC_v}, ONC_v\}}\left(ex_{NC_v}\right)$, then EX_{NC_v} should also have two corresponding values, say ex'_{NC_v} and ex''_{NC_v}.
("only if") Proposition 4 guarantees that NC_v will have different values in $MAV_{\{TS_{NC_v} \setminus EX_{NC_v}, ONC_v\}}\left(ex'_{NC_v}\right)$ and $MAV_{\{TS_{NC_v} \setminus EX_{NC_v}, ONC_v\}}\left(ex''_{NC_v}\right)$ for two values ex'_{NC_v} and ex''_{NC_v} of EX_{NC_v}. Thus, if NC_v has only a certain value in two $MAV_{\{TS_{NC_v} \setminus EX_{NC_v}, ONC_v\}}\left(ex_{NC_v}\right)$, then EX_{NC_v} should have also a certain value in the corresponding two ex_{NC_v}. ■

Proposition 6 *The* $MAV_{\{\mathbf{TS}_{C_k} \setminus \mathbf{EX}_{C_k}, OC_k\}}\left(\mathbf{ex}_{C_k}\right)$ *corresponding to* $MAE_{\{\mathbf{TS}_{C_k} \setminus \mathbf{EX}_{C_k}, OC_k\}}\left(\mathbf{ex}_{C_k}\right)$ *is a certain value configuration for each certain value configuration* $\mathbf{ex}_{C_k}$.

Proof $\{\mathbf{TS}_{C_k} \setminus \mathbf{EX}_{C_k}, OC_k\}$ comprises sets of variables $\{TS_{C_k} \setminus EX_{C_k}, OC_k\}$ along all directed paths through C_k that end at OC_k. We will divide each such set into three subsets $\{TS_{C_k} \setminus \{EX_{C_k}, C_k\}\}$, C_k, and OC_k and consider a value configuration for $\mathbf{ex}_{C_k}$ for each subset separately. First, since no latent collider can be a child of a latent collider (Assumption 5), a value configuration for the subset $\{TS_{C_k} \setminus \{EX_{C_k}, C_k\}\}$ is considered to be identical to a value configuration for $\{TS_{NC_v} \setminus EX_{NC_v}\}$, and thus according to Proposition 3, is a certain value configuration for a certain value ex_{C_k}. Because $MAV_{\{TS_{C_k} \setminus \{EX_{C_k}, C_k\}\}}\left(ex_{C_k}\right)$ is a certain value configuration for a certain ex_{C_k} for each directed path TS_{C_k} that is included in $\mathbf{TS}_{C_k}$, the product of these value configurations, which corresponds to the product of $MAE_{\{TS_{C_k} \setminus \{EX_{C_k}, C_k\}\}}\left(ex_{C_k}\right)$ in (11), is also certain. Second, since C_k's parents $\mathbf{Pa}_k \subset \bigcup_{TS_{C_k} \in \mathbf{TS}_{C_k}} \{TS_{C_k} \setminus \{EX_{C_k}, C_k\}\}$, $\mathbf{pa}_k{}^{\mathbf{ex}_{C_k}}$ are certain value configurations. Thus, based on Proposition 2, $MAV_{C_k}\left(\mathbf{pa}_k{}^{\mathbf{ex}_{C_k}}\right)$ is also a certain value and similarly $MAV_{OC_k}(C_k^{\mathbf{ex}_{C_k}})$ is certain, where $C_k^{\mathbf{ex}_{C_k}} = MAV_{C_k}\left(\mathbf{pa}_k{}^{\mathbf{ex}_{C_k}}\right)$. Therefore, all variables in $\{\mathbf{TS}_{C_k} \setminus \mathbf{EX}_{C_k}, OC_k\}$ are certain in the major configuration for a certain value configuration $\mathbf{ex}_{C_k}$. ■

Proposition 7 *For every exogenous ancestor* $EX_{C_k} \in \mathbf{EX}_{C_k}$ *of a latent collider* C_k*, there are at least two configurations* $\mathbf{ex}'_{C_k}$ *and* $\mathbf{ex}''_{C_k}$ *of* $\mathbf{EX}_{C_k}$ *in which only* EX_{C_k} *of all* $\mathbf{EX}_{C_k}$ *changes values when* C_k *changes values in the two corresponding major value configurations* $MAV_{\{\mathbf{TS}_{C_k} \setminus \mathbf{EX}_{C_k}, OC_k\}}\left(\mathbf{ex}'_{C_k}\right)$ *and* $MAV_{\{\mathbf{TS}_{C_k} \setminus \mathbf{EX}_{C_k}, OC_k\}}\left(\mathbf{ex}''_{C_k}\right)$.

Proof We divide the proof into two parts. In the first part, we prove that for each exogenous ancestor of a latent collider, there are at least two MAVs in which only the collider's parent on the path from the exogenous to the collider (of all collider's parents) changes values together with the exogenous. We are aided in this part of the proof by Proposition 5 after considering the collider's parent as a latent non-collider. In the second part, using Assumption 7, we show that each such collider's parent changes values together with the collider in the same two MAVs in which the parent changes values together with the exogenous. Thereby, we prove that for each exogenous ancestor of a latent collider, there are at least two MAVs in which the collider changes values only with this exogenous ancestor.

For the first part, Proposition 5 guarantees that any exogenous ancestor EX_{C_k} of a parent $Pa_k \in \mathbf{Pa}_k$ of collider C_k (and thus $EX_{Pa_k} = EX_{C_k}$ and Pa_k is also a latent non-collider) changes its value if and only if Pa_k changes its value in two value configurations $MAV_{\{TS_{Pa_k} \setminus EX_{Pa_k}, OPa_k\}}\left(ex_{Pa_k}\right)$. By the opposite of Proposition 5, any exogenous ancestor $EX^*_{C_k}$ of a parent $Pa^*_k \in \mathbf{Pa}_k \setminus Pa_k$ of C_k is certain if and only if Pa^*_k is certain in two value configurations $MAV_{\{TS_{Pa^*_k} \setminus EX^*_{Pa^*_k}, OPa^*_k\}}\left(ex^*_{Pa^*_k}\right)$.

For the second part, we know by Assumption 7 (second part) that for every C_k that is a latent collider and for every $Pa_k \in \mathbf{Pa}_k$, there are at least two configurations $\mathbf{pa}'_k$ and $\mathbf{pa}''_k$ of $\mathbf{Pa}_k$ in which only the value of Pa_k is different and $MAV_{C_k}\left(\mathbf{pa}'_k\right) \neq MAV_{C_k}\left(\mathbf{pa}''_k\right)$. That is, the collider (which is the only variable in MAV_{C_k}) changes values together with each of its parents in at least two parents' configurations.

Combining the two parts, we have proven that a collider changes values following a change in the value of each of its parents in at least two configurations of the parents, when the change of values of this parent is due to a change of values of its exogenous ancestor in two exogenous configurations. This means that the collider changes values with each of its exogenous ancestors in at least two exogenous configurations. That is, for two configurations $\mathbf{ex}'_{C_k}$ and $\mathbf{ex}''_{C_k}$ of $\mathbf{EX}_{C_k}$ in which only EX_{C_k} changes values, there are at least two configurations $\mathbf{pa}'_k$ and $\mathbf{pa}''_k$ of $\mathbf{Pa}_k$ in which $Pa_k \in \mathbf{Pa}_k$ changes values in $MAV_{\{\mathbf{TS}_{C_k} \setminus \mathbf{EX}_{C_k}, OC_k\}}\left(\mathbf{ex}'_{C_k}\right)$ and $MAV_{\{\mathbf{TS}_{C_k} \setminus \mathbf{EX}_{C_k}, OC_k\}}\left(\mathbf{ex}''_{C_k}\right)$ with EX_{C_k}. Since these values of Pa_k in $\mathbf{pa}'_k$ and $\mathbf{pa}''_k$ also change with values of C_k, C_k changes values with EX_{C_k} in $\mathbf{ex}'_{C_k}$ and $\mathbf{ex}''_{C_k}$. Therefore, there are at least two configurations $\mathbf{ex}'_{C_k}$ and $\mathbf{ex}''_{C_k}$ of $\mathbf{EX}_{C_k}$ in which only EX_{C_k} has changed values when C_k changes values in the two corresponding major value configurations $MAV_{\{\mathbf{TS}_{C_k} \setminus \mathbf{EX}_{C_k}, OC_k\}}\left(\mathbf{ex}'_{C_k}\right)$ and $MAV_{\{\mathbf{TS}_{C_k} \setminus \mathbf{EX}_{C_k}, OC_k\}}\left(\mathbf{ex}''_{C_k}\right)$. ■

Lemma 2

1. *A latent non-collider NC_v and its observed child ONC_v, both descendants of an exogenous variable EX_{NC_v}, change their values in any two major configurations if and only if EX_{NC_v} has changed its value in the corresponding two configurations of **EX**.*

2. *A latent collider C_k and its observed child OC_k, both descendants of a set of exogenous variables $\mathbf{EX}_{C_k}$, change their values in any two major configurations only if at least one of the exogenous variables in $\mathbf{EX}_{C_k}$ has changed its value in the corresponding two configurations of **EX**.*

Proof

1. First ("only if"), by Proposition 3, the major value configuration of a latent non-collider NC_v and its observed child ONC_v, both of which are descendants of an exogenous variable EX_{NC_v}, are certain for any certain ex_{NC_v}. That is, if NC_v and ONC_v changed their values in any two major configurations, it is only because EX_{NC_v} has changed its value in the corresponding two configurations of **EX**. Second ("if"), by Proposition 4, the major value configurations of NC_v and ONC_v are changed if EX_{NC_v} has changed its value between two configurations of **EX**.

2. By Proposition 6, the major value configuration of a collider C_k and its observed child OC_k, both of which are descendants of a set of exogenous variables $\mathbf{EX}_{C_k}$, are certain for a certain $\mathbf{ex}_{C_k}$. That is, if C_k and OC_k changed their values in any two major configurations, it is only because at least one of the variables in $\mathbf{EX}_{C_k}$ also changed its value in the corresponding two configurations of **EX**.

■

Theorem 1 *Variables of a particular **MSO** are children of a particular exogenous latent variable EX or its latent non-collider descendant or children of a particular latent collider C.*

Proof The proof is divided into two separate cases. In the first case, we show that the children of a particular exogenous latent variable or its non-collider descendant belong to the same **MSO**, and in the second case, we show that the children of a particular collider latent belong to the same **MSO**.

Case 1: MSO of observed children of an exogenous latent or its latent non-collider descendants

Let $\mathbf{ONC}_i$ [13] (in $\mathbf{OEX} \bigcup \mathbf{OS}$) be a set of observed variables that are children of an exogenous variable EX_i and any of its latent non-collider descendants (if they exist), and let $\mathbf{OC}_i$ be a set of observed variables that are children of latent colliders where each has EX_i as an exogenous ancestor with other exogenous variables. Note that $\mathbf{OC}_i$ may be empty, if EX_i does not have any collider descendants, but $\mathbf{ONC}_i$ is never empty because it includes at least $\mathbf{OEX}_i$ (Assumption 4). Because no observed child can be included in both $\mathbf{OC}_i$ and $\mathbf{ONC}_i$, these sets are disjoint. Their union, $\mathbf{OV}_i = \mathbf{ONC}_i \bigcup \mathbf{OC}_i$, includes all the observed variables that are affected by EX_i and thus should change their values when EX_i changes.

- First, by Lemma 2 (first part), any subset of variables in $\mathbf{ONC}_i$ (and thus also $\mathbf{ONC}_i$ itself, which is a maximal set) always changes together in all PCCs that correspond to a change in EX_i and never change together in any other PCC. These variables belong to the same **MSO** that represents EX_i.
- Second, let subset $\mathbf{OC}_{ij}$ of $\mathbf{OC}_i$ contain all variables that (1) have a shared exogenous ancestor EX_j (besides EX_i) and (2) change their values together in at least one PCC, which corresponds to a change only in the value of EX_j. By Lemma 2, the other variables in $\mathbf{OV}_i$ that are not descendants of EX_j do not change in that PCC. Thus, variables in $\mathbf{OC}_{ij}$ $\forall j$ do not belong to the same **MSO** for which variables in $\mathbf{ONC}_i$ belong.

Consequently, variables in $\mathbf{ONC}_i$ will change together only in all PCCs that correspond to a change in EX_i, and therefore, will establish a maximal set of the variables $\mathbf{MSO}_i = \mathbf{ONC}_i$ that corresponds to all and only observed variables that are children of exogenous variable EX_i and its latent non-collider descendants.

[13] So far, observed variables had their own indices and their parents/ancestors also had these indices. In Theorem 1, the index is associated with the exogenous variable (Case 1) and the collider latent (Case 2), since these are the central subjects of interest here.

Case 2: MSO of observed children of a latent collider
In this case, it is important to note that different colliders and their children are affected by different sets of exogenous variables. Thus, we assume:

Assumption 8 *Latent colliders do not share exactly the same sets of exogenous ancestors.*

(In case Assumption 8 is violated, for example, if several latent colliders share exactly the same set of exogenous ancestors, LPCC does not identify the latent colliders as separate and learns one collider as the parent of all children of the latent colliders. Nevertheless, we believe this assumption is very realistic.)

Let $\mathbf{OC}_i$ be the set of the observed variables that are children of latent collider C_i that is a descendant of a set of exogenous variables $\mathbf{EX}_{C_i}$. By Lemma 2, any variable in $\mathbf{OC}_i$ should not change in any PCC unless at least one of its exogenous ancestors changes. The sets of variables that should change together with variables in $\mathbf{OC}_i$ if any of the exogenous variables in $\mathbf{EX}_{C_i}$ change is represented by:

$$\mathbf{OV} = \bigcup_{EX_t \in \mathbf{EX}_{C_i}} \mathbf{OV}_t = \bigcup_{EX_t \in \mathbf{EX}_{C_i}} \{\mathbf{ONC}_t \bigcup \mathbf{OC}_t\} = \bigcup_{EX_t \in \mathbf{EX}_{C_i}} \{\mathbf{ONC}_t \bigcup \{\mathbf{OC}_t \backslash \mathbf{OC}_i\}\} \bigcup \mathbf{OC}_i \quad (12)$$

where the union is over all exogenous ancestors EX_t of C_i. We separate the proof to include three sets of observed variables: 1) $\mathbf{OC}_i$, which are children of C_i; 2) $\mathbf{ONC}_t$, which are children of an exogenous variable EX_t and any of its latent non-collider descendants; and 3) $\{\mathbf{OC}_t \backslash \mathbf{OC}_i\}$, which are children of latent colliders, other than $\mathbf{C}_i$, that are descendants of EX_t.

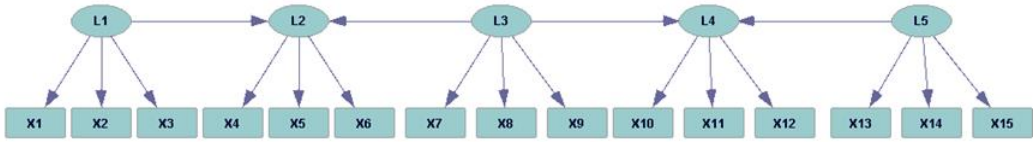

Figure 3: LVM with two latent colliders.

For example, for latent collider C_i = L2 in Figure 3, $\mathbf{EX}_{L_2}$ = {L1, L3}, $\mathbf{OC}_{\text{L2}}$={X4, X5, X6}, $\mathbf{ONC}_{\text{L1}}$={X1, X2, X3}, $\mathbf{ONC}_{\text{L3}}$={X7, X8, X9}, and $\{\mathbf{OC}_{\text{L4}} \backslash \mathbf{OC}_{\text{L2}}\}$ = {X10, X11, X12}.

Following, we analyze the three subsets of **OV**, specifically, $\mathbf{OC}_i$, $\mathbf{ONC}_t$, and $\{\mathbf{OC}_t \backslash \mathbf{OC}_i\}$, and show that only variables in $\mathbf{OC}_i$ (or any subset of $\mathbf{OC}_i$) will always change together, whereas other variables in **OV** will not. We analyze the subsets $\mathbf{ONC}_t$ and $\{\mathbf{OC}_{\text{t}} \backslash \mathbf{OC}_i\}$ for each exogenous $EX_t \in \mathbf{EX}_{C_i}$; thus, the analysis is also correct for their union (12).

- $\mathbf{OC}_i$: By Lemma 2 (second part), any subset of variables in $\mathbf{OC}_i$ always changes together in all PCCs that correspond to a change in at least one exogenous variable in $\mathbf{EX}_{C_i}$. In addition, none of the variables in $\mathbf{OC}_i$ has an exogenous ancestor that is not in $\mathbf{EX}_{C_i}$; therefore, no variable in $\mathbf{OC}_i$ ever changes in any PCC that corresponds to an exogenous variable that is not in $\mathbf{EX}_{C_i}$. These variables belong to the same **MSO** that represents C_i.

- $\mathbf{ONC}_t$: We previously showed in Case 1 that each $\mathbf{ONC}_t$ forms an **MSO** that corresponds to a single EX_t, and this is the only exogenous ancestor for $\mathbf{ONC}_t$. By Lemma 3, an **MSO** is an equivalence class; therefore, no other variable in a subset of **OV** (including $\mathbf{OC}_i$) can be added to $\mathbf{ONC}_t$, and it will remain an **MSO**. Similarly, no subset of variables in $\mathbf{ONC}_t$ can be added to any subset of **OV** to obtain an **MSO**.

- $\{\mathbf{OC}_t \backslash \mathbf{OC}_i\}$: Any subset of $\mathbf{OC}_t \backslash \mathbf{OC}_i$ does not change together with any subset of $\mathbf{OC}_i$ because (Assumption 8) for each variable OC_j in $\mathbf{OC}_t \backslash \mathbf{OC}_i$, there is an exogenous ancestor EX_j that is not an ancestor of variables in $\mathbf{OC}_i$. Thus, by Proposition 7, OC_j changes its value in a PCC that corresponds to a change only in the value of EX_j, whereas the variables in $\mathbf{OC}_i$, which are not descendants of EX_j, do not change in that PCC.

Consequently, all and only variables in $\mathbf{OC}_i$ (maximal subset of $\mathbf{OC}_i$) compose $\mathbf{MSO}_i$ that changes together in all PCCs that correspond to a change in $\mathbf{EX}_{C_i}$. ■

Theorem 2 *A latent variable L is a collider of a set of latent ancestors* **LA⊂EX** *only if:*

1. *The values of the children of L change in different parts of some major–major PCCs each time with the values of descendants of another latent ancestor in* ***LA****; and*

2. *The values of the children of L do not change in any PCC unless the values of descendants of at least one of the variables in* ***LA*** *change too.*

Proof Recall that by this point (Section 4.1), latent variables that have already been learned are either exogenous or colliders. Thus, first, we show that a latent variable L that satisfies (2) has to be a collider for a set of latent ancestors **LA⊂EX** by assuming by contradiction that L is not a collider but an exogenous variable. If L is an exogenous variable, then there exists at least a single major–major PCC_L that corresponds to two **ex**s in which only L changes its value. Thus, in PCC_L, only the values of descendants of L change, whereas descendants of other variables in any sub-set **LA⊂EX** do not change. This is in contrast to (2).

Second, we show that if L satisfies (1), then **LA** is the set of L's exogenous ancestors that collide in L. Let $\mathbf{ONC}_i$ (in $\mathbf{OEX} \bigcup \mathbf{OS}$) be the set of observed variables that are children of $LA_i \in \mathbf{LA}$ or children of its latent non-collider descendants. Let $\mathbf{OC}_i$ be the set of children of latent colliders where each has LA_i as its ancestor with other exogenous variables in **LA** or not. $\mathbf{OV}_i = \mathbf{ONC}_i \bigcup \mathbf{OC}_i$ includes all the observed variables that are affected by LA_i and thus may change their values when LA_i changes values. In addition, let $\mathbf{OC}_L$ be the set of children of L. We need to show that if L satisfies (1), then $\mathbf{OC}_L \subset \mathbf{OC}_i$ for each $LA_i \in \mathbf{LA}$. Since LA_i is an ancestor of L, (1) ensures that there exists a PCC in which only the values of descendants of LA_i including $\mathbf{OC}_L$ change, whereas the values of descendants of other variables in $\mathbf{LA} \backslash LA_i$ do not change. Thus, $\mathbf{OC}_L \subset \mathbf{OV}_i$. However, none of the children in $\mathbf{OC}_L$ belongs to $\mathbf{ONC}_i$; otherwise, it would have already been identified (Theorem 1) as a descendant of LA_i. Thus, $\mathbf{OC}_L \subset \mathbf{OC}_i$. ■

Lemma 4 *Let a latent non-collider NC be a descendant of an exogenous latent variable EX. 2S-PCC is PCC between a "two-set" first-order minor cluster 2S-MC due to a minor value in NC and a major cluster that identified EX.* **ex**′ *and* **ex**″ *are two value configurations of* ***EX*** *that correspond to the compared clusters by 2S-PCC. When:*

1. *EX does not change values between* **ex**′ *and* **ex**″*, all the elements in 2S-PCC corresponding to the observed descendants of the latent ancestors of NC (including EX) show no change (i.e., are 0), whereas the elements corresponding to the observed descendants of NC show a change (i.e., are 1),*

   and when

2. *EX changes values between* **ex**′ *and* **ex**″*, all the elements in 2S-PCC corresponding to the observed descendants of the latent ancestors of NC (including EX) show a change (i.e., are 1), whereas the elements corresponding to the observed descendants of NC show no change (i.e., are 0).*

Proof 2S-MC represents a 1-order minor configuration of **EN** in which only *NC* has a minor value, and all the other variables in **EN** have major values. Thus, when

- *EX* does not change values between **ex**′ and **ex**″ (i.e., $ex' = ex''$), then

  1. the major value configuration of the latent ancestors of *NC* is the same for both **ex**s (Proposition 3), and for each such latent, each of its observed children has the same major local value (Proposition 2) for both **ex**s. Thus, all the observed children of the latent ancestors of *NC* do not change values in both clusters, and all the corresponding elements in 2S-PCC are 0; and
  2. *NC* may take either a major or minor value in response to **ex**′(= **ex**″), depending on the probabilities of *NC* to take any of its values conditioned on the values *NC*'s direct parent takes. The result of the first case is a major cluster (*NC* and both its ancestors and descendants have major values) and that of the second case is 1-MC. Since all *NC*'s ancestors and descendants have major values, whereas *NC* has a minor value, this 1-MC is 2S-MC by definition. Using these two clusters, LPCC creates 2S-PCC. Since *NC* d-separates its descendants (both latents and observed) from its ancestors, the values of *NC*'s descendants are determined only by *NC* in a way similar to that which we used to prove Proposition 3. Since we are concerned with the case in which *NC* takes different values for **ex**′ and **ex**″, its descendants too have different values in the two corresponding configurations, and following Assumption 7, all of their observed children have different values in the corresponding observed configurations and clusters. Therefore, these children change their values between the clusters, as represented by 1s in the 2S-PCC.

- *EX* changes values between **ex**′ and **ex**″, then

  1. by Proposition 4, all the latent ancestors of *NC* have different values for **ex**′ and **ex**″, and by Assumption 7, all the observed children of these latents have

different values for **ex**′ and **ex**′′. Thus, in any 2S-PCC between two clusters corresponding to **ex**′ and **ex**′′, all the elements that correspond to the observed children of the latent ancestors of *NC* (including *EX*) show a change (i.e., are 1); and

2. *NC* does not change values between **ex**′ and **ex**′′ because if it did, then by Proposition 4, all of its latent descendants have different values for **ex**′ and **ex**′′, and by Assumption 7, all of their observed children have different values in the two corresponding observed configurations. And following, in any 2S-PCC between two clusters corresponding to **ex**′ and **ex**′′, all the elements that correspond to *NC* and its descendants would show a change (i.e., are 1). But, since as we already showed that all the observed children of the ancestors of *NC* are equal to 1 in these 2S-PCCs, it is contrary to the definition of a 2S-PCC that needs two sets of two or more elements of different values. Thus, *NC* cannot change values between **ex**′ and **ex**′′. Following and by Proposition 3, all the latent descendants of *NC* have certain values for this certain value of *NC* in both configurations, and by Proposition 2, all the observed children of these latents have certain values in the corresponding observed configurations. Thus, all the elements in 2S-PCC that correspond to the observed children of *NC* and its descendants do not show a change (i.e., are 0).

Note that the proof implicitly assumes that *NC* is on a serial connection emerging from *EX*. In a diverging connection, all the latent variables that are on the paths other than the one that includes *NC* can be considered with *NC*'s ancestors because both the latents on the other paths and *NC*'s ancestors are d-separated (for these 2S-PCCs) by *NC* from its descendants. Thus, the analysis proposed above for a serial connection generalizes also to the diverging connection. ■

Theorem 3 *Variables of a particular **2S-MSO** are children of an exogenous latent variable EX or any of its descendant latent non-colliders NC.*

Proof *I. Variables of* **2S-MSO** *that are children of EX*

We need, first, to prove that the children of *EX* always change values together and second, that no other observed child of another latent can always change value with them. First, Lemma 4 guarantees that the observed children of *EX* always change values together since a value change of *EX* between two **exs** corresponds to the compared clusters in all 2S-PCCs of 2S-MCs with the major clusters for *EX*. The remainder of the proof is divided into two cases: 1) a serial connection and 2) a diverging connection. In case 1, there exists at least a single 2S-PCC in which only the observed children of *EX* change. This 2S-PCC is between a major cluster for *EX* and 2S-MC due to a minor value of the direct latent non-collider

child *NC*[14] of *EX* (e.g., L2 is the direct latent non-collider child of L1 in G3).[15] Thus, only the elements in 2S-PCC that correspond to the observed children of *EX* show a change and are equal to 1 (e.g., *PCC2,10* in Table 4), which guarantees that the observed children of *EX* establish a **2S-MSO**.

In case 2, the same analysis proposed in case 1 is repeated for each of the direct latent non-collider children of *EX* in each of the paths that emerges from *EX*. Let us use the same notation *NC* for each such direct child in each path in turn. In this case, not only do the observed children of *EX* change each time *EX* changes, but also the observed descendants of the other direct latent non-collider children of *EX* (in all paths except that which includes *NC*) change with *EX*. This shows that the observed children of *EX* change with the observed descendants of the direct latent non-collider children of *EX* (all but the descendants of *NC*), but never together with all of them (as at each time, another *NC* is excluded). This guarantees that the observed children of *EX* establish a **2S-MSO**.

II. Variables of **2S-MSO** *that are children of EX's descendant NC*

In a serial connection, we identify three possible situations in which either *NC*, its latent descendant, or its latent ancestor takes a minor value. In each of these situations, no other latent or observed variable can take a minor value because we focus the analysis on 2S-MC through the evaluation of 2S-PCC between this minor cluster and a major cluster for *EX*. For each of the three situations, *EX* may change its value or not, so we have to consider six cases:

1. 2S-MC is due to a minor value of any of *NC*'s latent non-collider descendants, *NC1*, and *EX* does not change value between two **ex**s that correspond to the compared clusters. Then, by Lemma 4 (first part), all of *NC1*'s observed descendants do change values, but all the observed children of *NC1*'s latent ancestors, including those of *NC*, do not change values.

2. 2S-MC is due to a minor value of any of *NC*'s latent non-collider descendants, *NC1*, and *EX* changes value between two **ex**s that correspond to the compared clusters. Then, by Lemma 4 (second part), all of *NC1*'s observed descendants do not change values, but all the observed children of *NC1*'s latent ancestors, including those of *NC*, change values.

3. 2S-MC is due to a minor value of *NC*, and *EX* does not change value between two **ex**s that correspond to the compared clusters. Then, by Lemma 4 (first part), all of *NC*'s

[14] A) We focus on the latent non-collider *NC* that is the direct child of *EX* since only a minor value that *NC* takes can d-separate *EX* and its observed children from NC's observed children and the observed children of the remaining latent non-colliders, and partition the elements in the corresponding 2S-PCC into two sets in which the first consists of the observed children of EX and the second consists of the observed children of all EX's latent descendants. B) In our circumstances, where at least a single latent non-collider has been combined with EX, the existence of such a latent variable is guaranteed. C) It is also guaranteed that the 1-MC due to the minor value of the direct latent child of EX is 2S-MC because it cannot be due to an observed variable (see Case 2 above).

[15] We assume that all possible 1-MCs, including the one corresponding to a minor value of the direct latent non-collider child NC of EX, are found. Practically, if we err in estimating the threshold on the maximal 2-MC (as described above and in Appendix B), we may miss this 1-MC, but this is an identification issue that does not affect the correctness of the theorem.

observed descendants do change values, but all the observed children of its ancestors do not.

4. 2S-MC is due to a minor value of *NC*, and *EX* changes value between two **ex**s that correspond to the compared clusters. Then, by Lemma 4 (second part), all of *NC*'s observed descendants do not change values, but all the observed children of its ancestors do.

5. 2S-MC is due to a minor value of *NC*'s latent non-collider ancestor, *NC1*, and *EX* does not change value between two **ex**s that correspond to the compared clusters. Then, by Lemma 4 (first part), all the observed children of *NC1* and of its latent descendants, including those of *NC*, change values.

6. 2S-MC is due to a minor value of *NC*'s latent non-collider ancestor, *NC1*, and *EX* changes value between two **ex**s that correspond to the compared clusters. Then, by Lemma 4 (second part), all the observed children of *NC1* and of its latent descendants, including those of *NC*, do not change values.

That is, in all six cases, *NC*'s observed children change values together; in some 2S-PCCs they change values with observed children of a latent non-collider ancestor of *NC* and in some other 2S-PCCs with observed children of a latent non-collider descendant of *NC*. Thus, not only will the set of all the observed children of *NC* always change values together, but also no observed child of any of *NC*'s latent non-collider ancestors or descendants can be part of this set. This means that the set of observed children of *NC* is a maximal set of variables that always change together, i.e., **2S-MSO**.

Note that if *NC* does not have a latent non-collider descendant or ancestor, then Cases 1 and 2 and Cases 5 and 6, respectively, do not exist. In the special case where *NC* is a leaf (i.e., does not have a latent descendant), Case 3 guarantees that there exists at least a single 2S-PCC in which only the observed children of *NC* change.

In a diverging connection, all the latent variables that are on paths other than the one that includes *NC* can be considered with *NC*'s ancestors because *NC* d-separates them all from its descendants. Thus, the same analysis proposed in the serial case also holds in the diverging case. ■

Theorem 4 *A latent non-collider NC1 is a direct child of another latent non-collider NC2 (both on the same path emerging in EX) only if:*

- *In all 2S-PCCs for which EX does not change, the observed children of NC1 always change with those of NC2 and also in a single 2S-PCC without the children of NC2; and*

- *In all 2S-PCCs for which a latent non-collider leaf descendant of EX does not change, the observed children of NC2 always change with those of NC1 and also in a single 2S-PCC without the children of NC1.*

Proof Let *NC1* and *NC2* be latent non-collider descendants of *EX* (both on the same path emerging from *EX*), and *NC1* be a direct child of *NC2*. A 2S-PCC may result from a 2S-MC

due to a minor value in: 1) a latent ancestor of *NC1* (including *NC2* itself), 2) *NC1*, or 3) a latent descendent of *NC1*. In the first type of such 2S-PCC (for which *EX* does not change), Lemma 4 (first part) guarantees that the children of *NC1* and *NC2* change together in (1) and do not change at all in (3), whereas in (2) only the observed children of *NC1* change. Thus, the children of *NC1* always change with the children of *NC2*, and in addition also in a single 2S-PCC in which the children of *NC2* do not change.

In the second type of such 2S-PCC for which the observed children of the leaf latent non-collider descendant of *EX* do not change, Lemma 4 (second part) guarantees that *EX* changes value, and the children of *NC1* and *NC2* do not change at all in (1) and change together in (3), whereas in (2) only the observed children of *NC2* change. Thus, the children of *NC2* always change with the children of *NC1*, and in addition also in a single 2S-PCC in which the children of *NC1* do not change. The same analysis is true for both a serial and diverging connection. ■

Appendix B. Setting a threshold for the maximal size of *2*-order minor clusters (Section 4.4)

In this appendix, we describe the calculation of a *2*-order minor cluster threshold (*2MCT*) on the maximal size of *2*-order minor clusters (2-MCs) that were introduced in Section 4.4.

This threshold represents the maximal size of a minor cluster that corresponds to a 2-order minor value configuration (Definition 13), i.e., a minor cluster that represents exactly two endogenous variables in **EN** that have minor values. This threshold is separately calculated to each $EX_i \in \mathbf{EX}$, when all endogenous variables in **EN**, except the two mentioned, have major values. This threshold is an approximation for the maximal probability of having minor values as a response to any **ex** in exactly two descendants of *EX*, where all other descendants of *EX* and the other exogenous variables in **EX** have major values. This approximation is derived from the product of the maximal minor local effects (Definition B.1) of two observed descendants of EX_i, the maximal major local effects (Definition B.1) of the other observed descendants of EX_i, the maximal major local effects of the descendants of the other exogenous variables in **EX**, and the maximal prior of all exogenous variables in **EX**. We define:

Definition B.1 A *maximal major local effect* on an observed child O_i of a latent parent Pa_i is the maximal major effect on O_i over all values pa_i' of Pa_i, such that $MaxMAE_i = max_{pa_i'} MAE_i\left(pa_i'\right)$. Similarly, a *maximal minor local effect* is the maximal minor effect over all values pa_i' of Pa_i, such that $MaxMIE_i = max_{pa_i'} MIE_i\left(pa_i'\right)$.

First, we find $MaxMAEV_i$ and $MaxMIEV_i$, which are the sorted vectors of $MaxMAE_t$ and $MaxMIE_t$ (Definition B.1) of all $O_t \in \mathbf{Ch}_i$ (observed descendants of EX_i), respectively. These vectors include the maximal major local effects and the maximal minor local effects on the observed descendants of EX_i sorted from the highest to the lowest. Note that EX_i replaces the actual direct parent of an observed variable for calculating the maximal major and minor effects since the direct parent has not been identified and split yet from EX_i at this stage.

Using these maximal major and minor effects and their sorted vectors for EX_i, we can calculate the approximation of the threshold. First, the maximal probability of exactly two minor values among the descendants of EX_i can be approximated by:

$$\prod_{t=1}^{2} MaxMIEV_i(t).$$

Second, the maximal probability of the other descendants of EX_i to have major values can be approximated by:

$$\prod_{t=1}^{|\boldsymbol{Ch}_i|-2} MaxMAEV_i(t).$$

Third, the maximal probability of the other descendants of the other exogenous variables to have major values can be approximated by:

$$\prod_{EX_j \in \mathbf{EX} \setminus EX_i} \prod_{o_t \in \mathbf{Ch}_j} MaxMAE_t.$$

Fourth, the maximal prior of all the exogenous variables is represented by:

$$\prod_{EX_k \in \boldsymbol{EX}} max_{ex'_k} P(EX_i{=}ex'_k).$$

Then, the threshold for the maximal size of *2*-order minor clusters (measured by the number of patterns in such a cluster) for EX_i can be approximated by the product of all the above approximations multiplied by the data size N:

$$2MCT_i = N \prod_{t=1}^{2} MaxMIEV_i(t) \prod_{t=1}^{|\boldsymbol{Ch}_i|-2} MaxMAEV_i(t)$$

$$\prod_{EX_j \in \mathbf{EX} \setminus EX_i} \prod_{o_t \in \mathbf{Ch}_j} MaxMAE_t \prod_{EX_k \in \boldsymbol{EX}} max_{ex'_k} P(EX_i{=}ex'_k).$$

Appendix C. Assumptions LPCC makes and the meaning of their violation

Assumption	Essential?	If violated
Assumption 1 The underlying model is a Bayesian network, BN=<$\boldsymbol{G}$,$\boldsymbol{\Theta}$>, encoding a discrete joint probability distribution P for a set of random variables **V**=**L**∪**O**, where $\boldsymbol{G}$=<**V**,**E**> is a directed acyclic graph (DAG) whose nodes **V** correspond to latents **L** and observed variables **O**, and **E** is the set of edges between nodes in $\boldsymbol{G}$. $\boldsymbol{\Theta}$ is the set of parameters, i.e., the conditional probabilities of variables in **V** given their parents. **Assumption 2** No observed variable in **O** is an ancestor of any latent variable in **L** (the *measurement assumption*; Spirtes et al., 2000).	Yes [made also in similar algorithms, e.g., that of Silva et al. (2006) for continuous joint probability distributions].	It neither was investigated theoretically nor studied experimentally what LPCC returns if the underlying model is not a BN (i.e., there are cycles in $\boldsymbol{G}$), no latent variables exist in the domain, or an observed variable is an ancestor of a latent variable.
Assumption 3 The measurement model of $\boldsymbol{G}$ is pure.	No (only needed for the correctness of the learned model).	When the true causal model is pure, LPCC will identify it correctly (or find its pattern). However, when it is not pure, LPCC – similarly to BPC (Silva et al., 2006) – will learn a pure sub-model of the true model using two indicators for each latent (compared to three indicators per latent that are required by BPC).
Assumption 4 The true model $\boldsymbol{G}$ is MIM, in which each latent has at least two observed children and may have latent parents.	Yes [made also by Silva et al. (2006), which requires three indicators per latent].	If a latent has only one observed child, LPCC will not identify this latent.
Assumption 5 A latent collider does not have any latent descendants (and thus cannot be a parent of another latent collider).	No (only needed for the correctness of the learned model).	If this assumption is violated, and a latent collider has latent descendants, but none of them is a collider, LPCC does not identify the latent descendants as separate and join them, along with their observed children, to the learned ancestor latent collider. The case in which this assumption is violated, and at least one of the latent descendants of the collider is a latent collider itself, needs further investigation.
Assumption 6 For every endogenous variable EN_i in $\boldsymbol{G}$ and every configuration $\mathbf{pa}_i'$ of EN_i's parents $\mathbf{Pa}_i$, there exists a certain value en_i' of EN_i, such that $P\left(EN_i = en_i' \mid \mathbf{Pa}_i = \mathbf{pa}_i'\right) > P\left(EN_i = en_i'' \mid \mathbf{Pa}_i = \mathbf{pa}_i'\right)$ for every other value en_i'' of EN_i. This assumption is related to the most probable explanation of a hypothesis given the data (Pearl, 1988).	No (only needed for the correctness of the learned model).	If more than one value of EN_i gets the maximal probability value given a configuration of parents, LPCC still learns a model because the implementation will randomly choose one of the values that maximize the probability as the most probable. However, the correctness of the algorithm will not be guaranteed.
Assumption 7 First, for every EN_i that is an observed variable or an endogenous latent non-collider and for every two values pa_i' and pa_i'' of Pa_i, $MAV_{EN_i}\left(pa'_i\right) \neq MAV_{EN_i}\left(pa''_i\right)$. Second, for every C_j that is a latent collider and for every $Pa_j \in \mathbf{Pa}_j$, there are at least two configurations $\mathbf{pa}_j'$ and $\mathbf{pa}_j''$ of $\boldsymbol{Pa}_j$ in which only the value of Pa_j is different and $MAV_{C_j}\left(\mathbf{pa}'_j\right) \neq MAV_{C_j}\left(\mathbf{pa}''_j\right)$.	Not essential but very reasonable.	Regarding the second part of the assumption first: if this assumption is violated, and a collider has the same major value for any value of one of its parents (while the values of the other parents are the same), then its correlation to this parent should be very weak, which challenges the existence of their connection in the domain, and of course, the ability of any learning algorithm to identify this connection. Although the first part of the assumption may be considered similarly (based on a parent-child correlation), it also invites further investigation.
Assumption 8 Latent colliders do not share exactly the same sets of exogenous ancestors.	Not essential but very reasonable.	If this assumption is violated, and several latent colliders share exactly the same set of exogenous ancestors, LPCC does not identify the latent colliders as separate and learns a single collider as the parent of all children of the latent colliders.

References

N. Asbeh and B. Lerner. Learning latent variable models by pairwise cluster comparison. In *Proceedings of the 4th Asian Conference on Machine Learning, JMLR Workshop & Conference Proceedings*, pages 25:33–48, 2012.

D. J. Bartholomew, F. Steele, I. Moustaki, and J. I. Galbraith. *The Analysis and Interpretation of Multivariate Data for Social Scientists (Texts in Statistical Science Series)*. Chapman & Hall/CRC Press, Boca Raton, Florida, USA, 2002.

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

A. Dempster, N. Laird, and D. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of Royal Statistical Society*, B 39:1–39, 1977.

G. Elidan and N. Friedman. Learning the dimensionality of hidden variables. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, pages 144–151, Seattle, Washington, 2001.

G. Elidan, N. Lotner, N. Friedman, and D. Koller. Discovering hidden variables: A structure-based approach. In *Advances in Neural Information Processing Systems*, pages 13:479–485, 2000.

H. B. Enderton. *Elements of Set Theory*. Academic Press, New York, New York, 1977.

N. Friedman. The Bayesian structural EM algorithm. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, pages 129–138, San Francisco, CA, 1998.

S. Harmeling and C. K. I. Williams. Greedy learning of binary latent trees. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33:1087–1097, 2011.

R. Klee. *Introduction to the Philosophy of Science: Cutting Nature at its Seams*. Oxford University Press, New York, New York, 1997.

T. Kohonen. *Self-Organizing Maps*. Springer-Verlag, New York, New York, 1997.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann Press, San Mateo, California, 1988.

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

J. Pearl and T. Verma. A theory of inferred causation. In *Proceedings of the 2nd International Conference on Principles of Knowledge Representation and Reasoning*, pages 441–452, Cambridge, MA, 1991.

R. Scheines, P. Spirtes, C. Glymour, C. Meek, and T. Richardson. The tetrad project: Constraint based aids to causal model specification. Technical report, Department of Philosophy, Carnegie-Mellon University, Pittsburgh, Pennsylvania, 1995.

S. Shimizu, T. Inazumi, Y. Sogawa, A. Hyvarinen, Y. Kawahara, T. Washiok, P. Hoyer, and K. Bollen. DirectedLiNGAM: A direct method for learning a linear non-Gaussian structural equation model. *Journal of Machine Learning Research*, 12:1225–1248, 2011.

R. Silva, R. Scheines, C. Glymour, and P. Spirtes. Learning the structure of linear latent variable models. *Journal of Machine Learning Research*, 7:191–246, 2006.

C. Spearman. General intelligence objectively determined and measured. *American Journal of Psychology*, 15:201–293, 1904.

P. Spirtes. Calculation of entailed rank constraints in partially non-linear and cyclic models. In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, pages 606–615, Bellevue, Washington, 2013.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction and Search*. MIT Press, New York, New York, 2nd edition, 2000.

Y. Wang, N. L. Zhang, and T. Chen. Latent-tree models and approximate inference in Bayesian networks. *Journal of Artificial Intelligence Research*, 32:879–900, 2008.

N. Zhang. Hierarchical latent class models for cluster analysis. *Journal of Machine Learning Research*, 5:697–723, 2004.