Measurement Dependence Inducing Latent Causal Models

Alex Markham[1] **and Moritz Grosse-Wentrup**[1 2 3]
[1]Research Group Neuroinformatics, Faculty of Computer Science, University of Vienna
[2]Research Platform Data Science @ Uni Vienna
[3]Vienna Cognitive Science Hub

Abstract

We consider the task of causal structure learning over measurement dependence inducing latent (MeDIL) causal models. We show that this task can be framed in terms of the graph theoretic problem of finding edge clique covers, resulting in an algorithm for returning minimal MeDIL causal models (minMCMs). This algorithm is non-parametric, requiring no assumptions about linearity or Gaussianity. Furthermore, despite rather weak assumptions about the class of MeDIL causal models, we show that *minimality* in minMCMs implies some rather specific and interesting properties. By establishing MeDIL causal models as a semantics for edge clique covers, we also provide a starting point for future work further connecting causal structure learning to developments in graph theory and network science.

1 INTRODUCTION

Despite the many theoretical and practical difficulties, establishing and understanding causal relationships remains one of the fundamental goals of scientific research. Consequently, many different approaches have been developed, with applications spanning a diverse range of fields, e.g., from epidemiology to psychometrics to neuroimaging (Parascandola, 2001; Hoover, 2006; Seth et al., 2015). Some of the most well-known approaches include Granger causality (Granger, 1969) for time-series data, the Rubin causal model and potential outcomes framework (Holland, 1986) for randomized controlled trials, and functional causal models and the representation of their causal structure as directed acyclic graphs (Pearl, 2000; Spirtes et al., 2000). The last of these, the directed acyclic graph (DAG), provides the context for our approach to causal structure learning.

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

Roughly speaking, causal structure learning (CSL) typically focuses on identifying which variables are directly causally related and how these *direct causal relations form a structure* over which indirect causal relations exist. One way of characterizing CSL algorithms is according to which of the three following assumptions they rely on: (i) the *causal Markov assumption*, which says the random variables are (conditionally) independent (denoted by $\perp\!\!\!\perp$) if the corresponding vertices in the DAG are d-separated (denoted by $\perp$); (ii) the *causal faithfulness assumption*, which says the vertices in the DAG are d-separated if the corresponding random variables are (conditionally) independent; and (iii) the *causal sufficiency* of the set of variables, i.e. that there are no unobserved or latent common causes. The basic approach to CSL—namely the original constraint-based IC and PC algorithms (Verma and Pearl, 1990; Spirtes and Glymour, 1991)—rely on all three, while many of the algorithms developed in the 30 years since (as we will see in Section 1.1) relax these assumptions.

Considering applications of CSL to, for example, psychometrics and neuroimaging, the assumption of causal sufficiency seems implausible. For a data set consisting solely of answers to a depression diagnostic questionnaire or of voxel intensities in calcium imaging recordings (with random variables corresponding respectively to the questions or voxels), we think it is relatively uncontroversial to claim that not only are the random variables not causally sufficient, but indeed *every* dependence relation among them is induced by unobserved latent variables (respectively either cognitive processes related to, e.g., depression, or calcium signaling in cellular tissue, plus other confounders). In fields and applications such as these—where interventions are often difficult or unfeasible, and where the goal is to reason about underlying causes based on their measurable effects—a more tailored causal modeling framework may prove insightful. Thus, the main difference between the traditional approach outlined above and the one we present in this paper is that

we assume a strong causal *in*sufficiency of the random variables being modeled and therefore are able to represent a different (but not entirely disjoint) class of causal structures than is possible with DAGs in the traditional approach.

The rest of the paper is organized as follows: We begin by reviewing related work, emphasizing points of departure. In Section 2 we define measurement dependence inducing latent (MeDIL) causal models to be the class of latent measurement models in which measurement variables can only be effects (and not causes—contrary to the definition of measurement models explored by others), making no further assumptions about linearity or parametrizations of the distributions. We then introduce the notions of observational consistency and minimality, allowing us to, for a given (estimated) distribution of measurement variables, construct a minimal MeDIL causal model (minMCM). Then, in Section 3, by framing minMCMs as edge clique covers (ECCs) of the undirected dependency graph over measurement variables, we note how two notions of minimality emerge. Subsequently, despite our nonrestrictive assumptions and notion of minimality in minMCMs, we are able to prove (i) that a minMCM lower bounds the number of latent variables or the number of functional causal relations (depending on which notion of minimality is used), (ii) that the latent variables of the minMCM are all pairwise independent, and (iii) that (somewhat surprisingly) the minMCM can have more latent causes than measured variables. In Section 4 we describe an algorithm for learning minMCMs from only *unconditional* (in)dependencies. Finally, we demonstrate our approach with an application to a psychometric data set in Section 5, before concluding with a discussion of promising directions for future work.

1.1 RELATED WORK

Elaborating on the basic approach mentioned above, CSL without latents amounts to finding an *essential graph* (Andersson et al., 1997), a mixed graph with directed and undirected edges, which represents the Markov equivalence containing the true DAG. The essential graph is typically found by using either a score- or constraint-based approach. Score-based methods find an essential graph by directly optimizing a score of how well it fits the data samples (Chickering, 2002).Constraint-based methods take a set of conditional independence relations as input (which must be estimated or acquired somehow before applying the algorithm), and these relations constitute a set of constraints on the possible d-separations, which the output essential graph satisfies (Verma and Pearl, 1990; Spirtes and Glymour, 1991). Our approach in this paper is more closely related to constraint-based methods, especially their extensions to latent variable models.

Extensions of CSL to causal models including latent variables (i.e., relaxing the causal sufficiency assumption), such as the FCI algorithm and its variants (Spirtes et al., 1999), correspondingly extend the search space from essential graphs to partial ancestral graphs (PAGs), which have an additional three edge types (so five total), allowing them to represent the extended Markov equivalence class containing dependencies induced by latent variables.

In these terms, our latent CSL algorithm *is not* searching for a PAG. As we explain in sections 2 and 3, by making use of the strong causal insufficiency in this application space, we can directly represent the conditional independence constraints that form the input for our algorithm as an undirected dependence graph (UDG). This UDG is essentially a PAG with only bidirected edges. Or, put another way, it is a modified Markov random field (Kindermann and Snell, 1980) where the conditional independence relations are determined from the undirected edges by using strong causal insufficiency (see Proposition 6) instead of the Markov property, thereby allowing the UDG to represent latent induced dependence (which Markov random fields are usually incapable of representing).

With the conditional independence constraints input in the form of a UDG over measurement variables, our algorithm essentially adds the latent causes and directed edges necessary to construct the minimally causally sufficient DAG containing latent and measurement variables. Thus, instead of doing CSL in the presence of latent variables as is the case with FCI and similar algorithms, we *use CSL to reason about latent variables*.

Our approach is more related in this respect to other work on measurement models (Silva and Scheines, 2005; Silva et al., 2006; Kummerfeld et al., 2014; Kummerfeld and Ramsey, 2016). However, these other approaches utilize properties of the covariance matrix of the measurement variables, such as vanishing tetrad constraints, while we utilize graph theoretic properties of the UDG representation of conditional independencies. This results in connections between our approach and causal feature learning (Chalupka et al., 2016) and causal consistency and abstraction (Rubenstein et al., 2017; Beckers and Halpern, 2019), which will be discussed more with respect to future work in Section 6.2. Another closely related approach is factor analysis, especially when framed in terms of using the topology of a Bayesian network of observed variables to reason about hidden factors (Martin and VanLehn, 1994), with the main difference being our goal of a minimally causally sufficient DAG as opposed to a statistically convenient (but not necessarily as causally relevant) factor model.

Overall, our approach has several points of overlap in terms of motivations and formal methods in existing CSL, measurement model, and factor analysis approaches. However, we address the problem from a different perspective, utilizing the causal insufficiency property of our application space and graph theoretic edge clique cover methods to produce a novel algorithm.

2 MINIMAL MEDIL CAUSAL MODELS

We begin with a formal definition of *measurement dependence inducing latent (MeDIL) causal models*, before discussing the notion of observational consistency and its implications about minimality in such models.

We use functional causal models (FCMs) to describe causal relations in complex systems.

Definition 1 (Functional Causal Model)**.** A *functional causal model* is a triple $\mathcal{M} = \langle \mathbf{V}, \mathbf{F}, \boldsymbol{\epsilon} \rangle$, where

- $\mathbf{V}$ is the set of (endogenous) random variables,
- $\mathbf{F}$ is a set of functions defining each endogenous variable as a function of its direct causes (i.e., parents or $\mathrm{pa}()$) and its corresponding exogenous random variable, so that for each $V_i \in \mathbf{V}$, we have $V_i := f_i(\mathrm{pa}(V_i), \epsilon_i)$. Furthermore, $\mathbf{F}$ is constrained such that no V_i is a direct cause of itself or any of its causes, removing the possibility of causal cycles.
- $\boldsymbol{\epsilon}$ defines a joint probability distribution over the exogenous (or noise) variables, with a corresponding $\epsilon_i \in \boldsymbol{\epsilon}$ for each $V_i \in \mathbf{V}$, and with ϵ_i being independent with ϵ_j for each $\epsilon_i, \epsilon_j \in \boldsymbol{\epsilon}$

□

In particular, we are interested in latent CSL over measurement variables, so it is advantageous to move from the general FCM definition to a specifically structural/graphical definition that conceptually differentiates the set of endogenous variables into causally effective latent variables and their observed measurements, leading to the idea of MeDIL causal models:

Definition 2 (Measurement Dependence Inducing Latent Causal Model (MCM))**.** A graphical MCM is a DAG, given by the triple $\mathcal{G} = \langle \mathbf{L}, \mathbf{M}, \mathbf{E} \rangle$. $\mathbf{L}$ and $\mathbf{M}$ are disjoint sets of vertices, while $\mathbf{E}$ is a set of directed edges between these vertices, subject to the following constraints:

1. all vertices in $\mathbf{M}$ have in-degree of at least 1 and out-degree of 0
2. all vertices in $\mathbf{L}$ have out-degree of at least 1
3. $\mathbf{E}$ contains no cycles

□

There are no further constraints as to the variety of distributions and functional causal relations that MCMs can represent, i.e., they are non-parametric and their arrows can represent arbitrary functional relations between variables. The formal constraints 1. and 2. in Definition 2 are to ensure that MCMs are applicable to settings in which we can explicitly separate into disjoint sets the measured effect variables $\mathbf{M}$ whose probabilistic dependencies must therefore be mediated by latent causes $\mathbf{L}$.

However, the explicit separation of cause and effect and the corresponding latent structure in MCMs introduces its own difficulties for inference. Namely, many latent models are consistent with a given probability distribution over observed effects, making the task of inferring a single latent model ill-posed. In order to help explain this consistency of different latent models and illustrate our strategy for restricting the problem so that inference is well-posed, consider the following definition and example.

Definition 3 (Observational Consistency)**.** A MCM is *observationally consistent* with a probability distribution over measurement variables if it is capable of inducing the pairwise dependencies (which can estimated from samples) of that distribution. This can be seen as a weakening of the notion of observational equivalence corresponding to our extension from DAGs containing only observed variables to the notion of MCMs.[1]

Example 4 (Observational Consistency)**.** Suppose we have data consisting of peoples' answers to a questionnaire with four questions designed to measure depression and stress. We assume that the answer to one question cannot cause the answer to another and therefore that the observed answers as well as any observed association between answers are the result of latent causes, such as depression or stress. Define random variables $\mathbf{M} = \{M_1, M_2, M_3, M_4\}$ corresponding to answers to the four questions, and let them have only the following two pairwise independencies:

$$M_1 \perp\!\!\!\perp M_4 \quad \text{and} \quad M_2 \perp\!\!\!\perp M_4$$

The pairwise dependency structure between variables in $\mathbf{M}$ is shown in Figure 1(a), and three observationally consistent MCMs are shown in 1(b), 1(c), 1(d). As this example demonstrates, multiple latent models can give rise to the same set of observed dependencies.

□

[1] *observational* or *Markov equivalence* (Pearl, 2000, pp. 16–20) means two DAGs have the same skeletons and colliders, while observational consistency means that two MCMs have the same undirected dependency graphs over measurement variables (e.g., Figure 1)

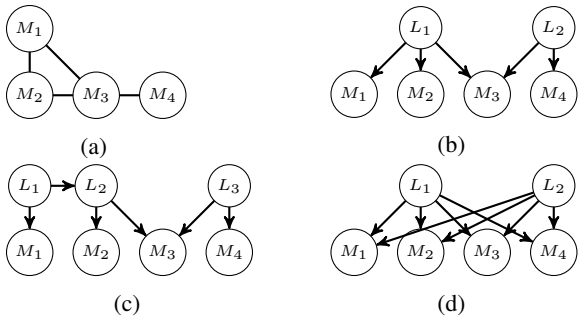

Figure 1: (a) undirected dependency graph over $\mathbf{M}$—notice two missing edges corresponding to independencies; (b) minimal MCM over $\mathbf{M}$; (c) non-minimal MCM observationally consistent with $\mathbf{M}$; (d) MCM corresponding to ICA or FA

We address this problem by employing Ockham's razor to pick a *minimal MCM (minMCM)* (e.g., Figure 1(b)).

Definition 5 (Minimal MeDIL causal model (minMCM))**.** A *minMCM* for a set of measurement variables $\mathbf{M}$ is any least expressive (i.e., minimal) MCM that is observationally consistent with $\mathbf{M}$. As Pearl and Verma (1995) note, a latent causal model's expressive power can be measured by the (in)dependencies it induces over the measured variables, with more dependencies corresponding to more expressive power. In our case, criteria can be given for minimality in modified terms of the causal faithfulness and causal Markov assumptions:

1. in addition to being observationally consistent with its set of measurements, a minMCM must graphically induce the measurements without violating faithfulness; the notion of faithfulness used here is concerned with conditional independencies only over measurements and not all variables in the MCM, so we call it *measurement-faithfulness*; note that Figure 1(b) is faithful to the conditional independencies in Example 4 while Figure 1(d) is not—the MCM in Figure 1(b) is minimal while that in 1(d) is not

2. considering arbitrary subsets of the latents, $\mathbf{Z} \subseteq \mathbf{L}$, there are as few d-separations of the form $M_i \not\perp\!\!\!\perp M_j \mid \mathbf{Z}$ as (faithfully) possible, i.e., such d-separations only exist in an minMCM if implied by the (in)dependencies and causal insufficiency of the distribution only over measurement variables; we call this *measurement-Markov* since it says the only d-separations in the minMCM are those implied by measurement-faithfulness[2]; note that Figure 1(c) does not satisfy this □

[2]just as is the case with the usual causal faithfulness and Markov conditions

Learning a minMCM for a data set only requires considering the *unconditional* independence relations among its variables, unlike the other methods mentioned in Section 1.1. This follows from Proposition 6.

Proposition 6. *In a MCM, the set of unconditional (in)dependencies over measurement variables fully determines the set of conditional (in)dependencies over measurement variables.*

Proof. The Causal Markov and Causal Faithfulness assumptions (CMA and CFA, respectively) imply that two variables are probabilistically independent if and only if they are d-separated (allowing us to use independence/d-separation and $\perp\!\!\!\perp$ / $\perp$ interchangeably). Recall from Definition 2 that all dependence relations (and therefore, by the CMA and CFA, d-connections) between measurement variables are mediated by latent variables. Hence, all measurement variables have out-degree 0, and so any measurement variable in a path between two other measurement variables must be a collider and any dependent measurement variables must share at least one latent parent. This means that the set of unconditional (in)dependencies over measurement variables fully determines the set of conditional (in)dependencies as follows: for all $M_i, M_j, M_k \in \mathbf{M}$,

- $M_i \not\perp\!\!\!\perp M_j \implies M_i \not\perp\!\!\!\perp M_j \mid M_k$
- $M_i \perp\!\!\!\perp M_j \implies$
$$\begin{cases} M_i \perp\!\!\!\perp M_j \mid M_k, & \text{if } M_i \perp\!\!\!\perp M_k \text{ or } M_j \perp\!\!\!\perp M_k \\ M_i \not\perp\!\!\!\perp M_j \mid M_k, & \text{otherwise} \end{cases}$$

□

As we will see in Section 4, even though estimating conditional independencies is not required for our method, doing so nevertheless can help determine whether any of the assumptions have been violated.

3 MINIMAL MEDIL CAUSAL MODELS AS EDGE CLIQUE COVERINGS

We can now present our main insight:

Proposition 7. *The problem of finding a minMCM for a set of measurement variables can be framed as the graph theoretical problem of finding a minimum edge clique covering (ECC)*[3] *(Erdős et al., 1966; Gramm et al., 2009; Ennis et al., 2012) over the corresponding undirected dependency graph of the measurement variables.*

[3]A minimum ECC over an undirected graph is a collection of cliques that exactly covers its edges, where an edge $E = (V_i, V_j)$ is covered by clique C iff $V_i, V_j \in C$.

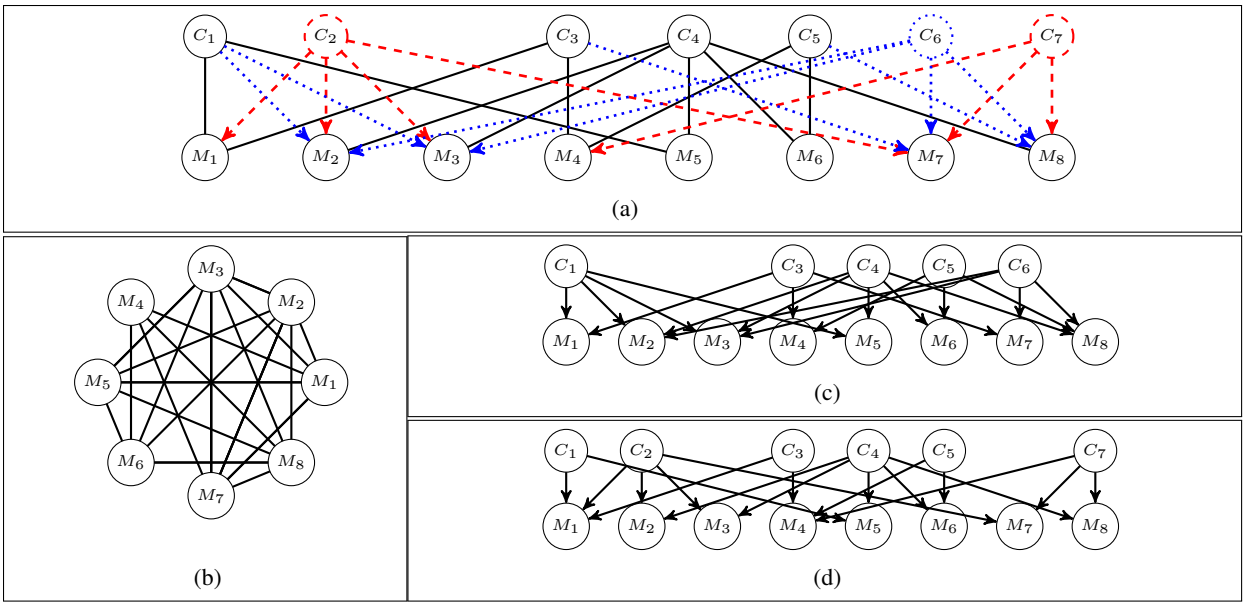

Figure 2: (a) MCM, where each C_i corresponds to a maximal clique in $\mathrm{D}(\mathbf{M})$—dashed red edges/vertices are redundant for vertex-minimality while blue dotted edges/vertices are redundant for edge-minimality; (b) $\mathrm{D}(\mathbf{M})$—undirected dependency graph of $\mathbf{M} = \{M_1, \ldots, M_8\}$; (c) vertex-minimal minMCM of $\mathrm{D}(\mathbf{M})$; (d) edge-minimal minMCM of $\mathrm{D}(\mathbf{M})$

Proof. For a given set of measurement variables $\mathbf{M}$, denote the *undirected dependency graph* as $\mathrm{D}(\mathbf{M})$, e.g., Figure 1(a), where an edge represents dependence and the lack of an edge represents independence. Proposition 6 tells us that $\mathrm{D}(\mathbf{M})$, though it only encodes unconditional (in)dependencies, contains all necessary information for characterizing observationally consistent MCMs. Consider the MCM $\mathcal{G} = \langle \mathbf{L}, \mathbf{M}, \mathbf{E} \rangle$ constructed from a set of cliques $\mathbf{C}$ comprising a minimum ECC over $\mathrm{D}(\mathbf{M})$ using the following procedure: (i) posit a latent $L_C \in \mathbf{L}$ iff $C \in \mathbf{C}$ and (ii) posit a directed edge $E \in \mathbf{E}$ from the latent L_C to the measurement variable M iff $M \in C$. In other words, $\mathbf{G}$ is a MCM with measurement variables $\mathbf{M}$, one latent for each clique in the minimum ECC over $\mathrm{D}(\mathbf{M})$, and an edge from each latent to exactly the measurement variables in the corresponding clique.

Note that $\mathbf{G}$ is not only observationally consistent with $\mathrm{D}(\mathbf{M})$ but also captures its independencies and is thus faithful, satisfying criterion 1. of Definition 5. Furthermore, the construction of $\mathbf{G}$ from a minimum ECC ensures that latents are only posited when necessitated by the dependencies between measurements, satisfying criterion 2. of Definition 5. Thus, $\mathbf{G}$ is an minMCM for $\mathrm{D}(\mathbf{M})$.

□

A minimum ECC can be minimal in two related but distinct ways: the original and more well-studied approach seeks the smallest number of cliques needed to cover all edges (this is equivalent to the *intersection number* (Erdős et al., 1966)), while another justifiable approach is to seek an ECC requiring the fewest assignments of vertices to cliques. The corresponding interpretation for minMCMs is vertex-minimal (fewer cliques imply fewer latents imply fewer total vertices) and edge-minimal (fewer assignments of measurement vertices to cliques implies fewer directed edges from latent to measurement vertices), resulting in Proposition 8. There are some undirected dependency graphs for which the vertex-minimal and edge-minimal minMCMs are identical, such as figures 1 and 3, but this identity does not hold generally (Ennis et al., 2012) (see Figure 2). In either approach to minimality, the resulting minMCM induces the same set of dependencies over measurement variables and thus has the same expressive power (w.r.t. the measurement variables). We thus see no straightforwardly principled way of picking one approach over the other, and so we present both in hopes that practitioners will use whichever one (or both) they judge most sensible/interesting for their particular application.

Regardless of which notion of minimality is used, minMCMs have some interesting properties. First, they lower bound (i) the number of causal concepts or (ii) the number of functional causal relations that are required to model measurements of a complex system at any level of granularity (Proposition 8). Second, minMCMs contain no

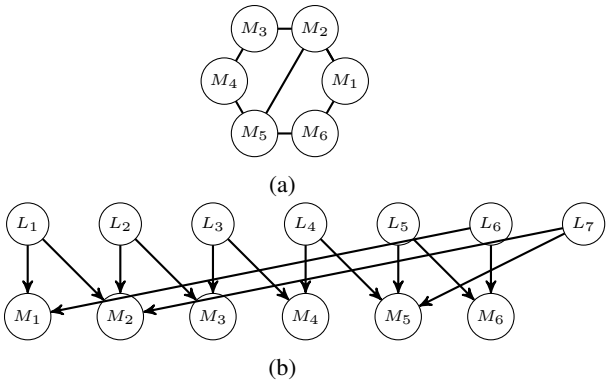

Figure 3: (a) example $\mathrm{D}(\mathbf{M})$ for which the minMCM (b) has 6 measurement variables and 7 latent variables

causal links between the latent variables (Proposition 9). Finally, in contrast to factor analysis, a minMCM may require more latent than measurement variables (Proposition 10).

Proposition 8. *For a given set of unconditional pairwise dependencies among measurement variables* $\mathbf{M}$*, a minMCM gives a lower bound on the number of latent variables or edges (depending on the measure of minimality is used) required in any (faithful and observationally consistent) MCM.*

Proof. This is a direct consequence of the construction of minMCMs from either the clique-minimum or assignment-minimum ECC of $\mathrm{D}(\mathbf{M})$, as described in Proposition 7. □

Proposition 9. *In a minMCM, each latent variable is d-separated from every other latent variable.*

Proof. Intuitively, this is a result of the definition of a minMCM being minimal in the sense of least expressive (and thus having as few latents or edges): if two latent variables are d-connected, then the dependencies among measurement variables that they induce could also instead be induced by a single latent variable (which also results in fewer edges). A minMCM has no redundant latent variables or edges and therefore no d-connected latent variables. For example, note in that the MCMs in figures 1(b) and 1(c) induce the same d-separations over the measurement variables, but that 1(b) with its d-separated latents has the fewer latents and fewer total edges. More formally, this follows directly from procedure for constructing an minMCM in Proposition 7 and Algorithm 1. □

Proposition 10. *There exist minMCMs containing more latent than measurement variables.*

Proof. This follows from the graph theoretical characterization of minMCMs: there are at least as many latent variables as the intersection number of $\mathrm{D}(\mathbf{M})$, which in a graph with n vertices is (non-trivially) upper bounded by $\frac{n^2}{4}$ (Erdős et al., 1966). A simple example can be found when $\mathrm{D}(\mathbf{M})$ is as in Figure 3, resulting in $n = 6$ nodes and an intersection number of $i = 7$. □

4 A minMCM-FINDING ALGORITHM AND ITS COMPLEXITY

The procedure in the proof of Proposition 7 for constructing a minMCM from an undirected dependency graph leads directly to Algorithm 1.

Algorithm 1: constructing a minimal MeDIL causal model (minMCM)

Input : undirected dependency graph, $\mathrm{D}(\mathbf{M})$, over the measurement variables $\mathbf{M}$

Output : vertex-minimal or assignment-minimal MCM $\mathcal{G}$ over $\mathbf{M}$

1 initialize edgeless graph with a vertex for each $M \in \mathbf{M}$;
2 find a clique-minimum or assignment-minimum edge clique cover of $\mathrm{D}(\mathbf{M})$, using the algorithm in Fig. 3 of (Gramm et al., 2009) or the algorithm `FIND-AM` of (Ennis et al., 2012), respectively;
3 **for** *each clique* C *in the cover* **do**
4 add vertex L with edges directed to each $M \in C$;
5 **end**

Notice that Line 2 in Algorithm 1 is to find a minimum ECC of $\mathrm{D}(\mathbf{M})$. Nearly all of the computational complexity of Algorithm 1 comes from this step, which is known to be an NP-hard problem, and so the choice of an efficient ECC-finding algorithm and implementation is especially important.

In case a clique-minimum ECC (and therefore vertex-minimum minMCM) is preferred, (Gramm et al., 2009) provides an exact algorithm. The exact algorithm finds an ECC in $\mathcal{O}(f(2^k) + n^4)$ time, where k is the number of cliques in the ECC and n is the number of vertices in the undirected graph, and is thus fixed-parameter tractable. Furthermore, (Cygan et al., 2016) gives a lower bound on the complexity of the clique-minimum ECC problem and argues that the algorithm is probably optimal. Gramm et al. (2009) also provide a free/libre implementation of their algorithm, though it has not been maintained for some time and does not easily run on most modern machines.

In case an assignment-minimum ECC (and therefore edge-minimum minMCM) is preferred, (Ennis et al., 2012) provides an exact algorithm. Though they do not offer an analyses of its complexity, it is essentially a backtracking algorithm based on (Bron and Kerbosch, 1973)'s maximal clique finding algorithm, which has time complexity of$\mathcal{O}(3^{n/3})$, and so this assignment-minimum ECC finding algorithm has an even larger complexity.

As far as we are aware, no other implementations of the clique-minimum or assignment-minimum ECC finding algorithms exist. To remedy this, we have implemented and released these and a few other related causal inference tools as a free/libre Python package at https://medil.causal.dev. Already Gramm et al. (2009) and Ennis et al. (2012) showed that their algorithms perform in a reasonable amount of time on moderately sized graphs, e.g., returning a solution containing 100 cliques in a matter of minutes. Unsurprisingly, given the hardware advancements of the past decade, our implementation performs even better, e.g. finding the 614 clique solution to the 61 node graph presented in the next section in only 39 seconds using an Intel Core i7-8700K CPU.

5 APPLICATION

In this section we demonstrate the necessary steps to get from a raw data set to a minMCM output from our algorithm. We then hint at how this output can be analyzed and suggest some conclusions that can be drawn from it. Note that our contribution in this paper is theoretical, and the point of the following application is to make some of our theoretical claims and the potential use cases more concrete.

5.1 THE DATA AND PREVIOUS ANALYSES

The *Stress, Religious Coping, and Depression* data set[4] was collected by Bongjae Lee from the University of Pittsburgh in 2003. There were 127 participants answering a total of 61 questions: 21 designed to measure stress, 20 for religious coping, and 20 for depression—see (Silva and Scheines, 2005) for the full questionnaire. This data has been analyzed by several other measurement model methods (Silva and Scheines, 2005; Silva et al., 2006; Kummerfeld et al., 2014; Kummerfeld and Ramsey, 2016), and their findings (which largely agree with each other) can be briefly summarized as follows: (i) in contrast to the design goal, most of the measurement variables are "impure" in that they are caused by multiple latent variables; (ii) they are able to find some subsets (ranging in number from three to nine) of "pure" measurement variables that passed their significance tests and some of which suggest a model similar to what Lee hypothesized containing three latent variables—the first of which causes only measurement variables of stress, the second only depression, and the third only coping; (iii) most of their models scoring the highest significance are more complex models than Lee's model (the most complex containing eight latents (Silva and Scheines, 2005)).

5.2 ANALYSIS USING minMCMS

Notice that the input to Algorithm 1 is an undirected dependency graph, while in practice one does not have direct knowledge of the (in)dependencies themselves but only samples of the measurement variables. It is therefore necessary to first estimate the independencies before applying this algorithm. Because the algorithm is agnostic to the test statistic, it is not constrained to linear methods such as Pearson correlation (for which "$X \perp\!\!\!\perp Y \implies \mathrm{corr}(X, Y) = 0$" but not the converse) but can leverage the power of nonlinear independence tests (Gretton et al., 2005; Székely et al., 2007). We used the distance correlation (Székely et al., 2007) as our test statistic (with the property "$X \perp\!\!\!\perp Y \iff \mathrm{dCorr}(X, Y) = 0$") and performed 1000 random permutations of the measurement variables to sample from the null-distribution (Dwass, 1957). The p-value for each pair was then calculated as the proportion of the permutation tests in which the absolute distance correlation of the pair of variables with permuted samples exceeded that of the original pair. Finally, independence between two variables was concluded if the distance correlation between them was less than 0.1 and the corresponding p-value was greater than 0.1.[5]

The binary-valued 61×61 matrix corresponding to the estimated independencies, with a 0 for independence and 1 for dependence thus forms the adjacency matrix for the UDG that is input for Algorithm 1. We decided to find a latent-minimal minMCM, and the result has 614 latent variables. It is thus too complex to be legibly displayed here, so we instead present figures 4 and 5 to facilitate analysis of the results.

Looking at the histogram in Figure 4(a), we find a median indegree (i.e., number of latent causes) of the measurement variables of 27, but with one in particular, M_{30}, having 425. The item in the questionnaire corresponding to M_{30} was the ninth in the set designed to measure depression, and it asked participants how frequently the event "I thought my life had been a failure" occurred in the preceding week. Semantically, it makes sense that this item would have many more latent causes than the

[4]We would like to thank David Danks and especially Joseph Ramsey at Carnegie Mellon University for providing us with a copy.

[5]As one would expect, using a nonlinear measure of dependence allows us to detect more dependencies: we found almost 31% of the over 1500 estimated nonlinear pairwise dependencies (i.e., edges in the UDG) to be undetectable using the linear Pearson correlation.

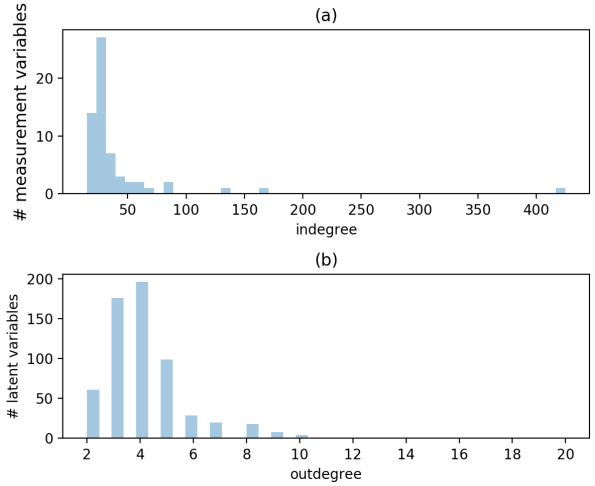

Figure 4: histograms showing (a) indegree of the measurement variables and (b) the outdegree of the latent variables

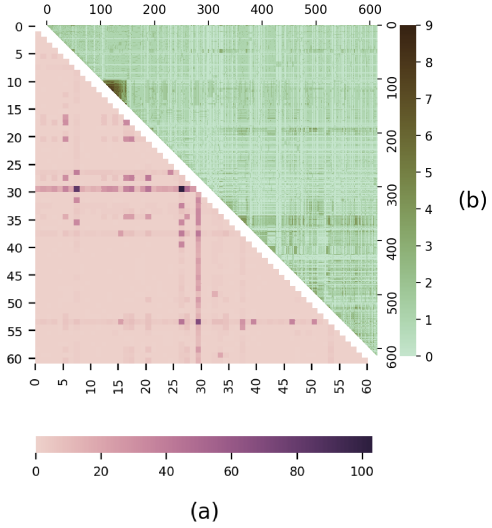

Figure 5: heatmaps showing (a) the number of latent variables each pair of the 61 measurement variables have in common and (b) the number of measurement variables each pair of the 614 latent variables have in common

other items, because its scope is much larger, requiring reflection on the participants' entire life up to that point instead of just during the week in question, as is the case for other depression items, such as "I enjoyed life" (M_{37}, 24 latent causes) and "I felt sad" (M_{39}, 25 latent causes). Furthermore, looking at Figure 5(a), showing the number of latents each pair of measurement variables share, we see that M_{30} shares a relatively high amount of latent causes with the other measurement variables (median of 21), while for M_{37} and M_{39} the median of shared latent causes is one. Our analysis thus agrees with the previous analyses described in Section 5.1 insofar as we also find many "impure" measurement variables, but extends their insights by differentiating between measurement variables that are best considered a general or mixed measurement (M_{30}) and those that, even though they are also impure, span different subsets of the latent space (M_{37} and M_{39}).

Looking at the outdegree (i.e., the number of measurement variables a latent causes) in Figure 4(b) we find a median of four and a range from 2 to 20. The number of measurements shared by each pair of latent variables reveals further structure (Figure 5(b)). In particular, the incidence matrix representation of the latents corresponding to the block structure between approximately L_{105}–L_{145} reveals seven measurement variables that these latents mostly have in common, corresponding to four stress and three depression items. On the other side, 41% (roughly 74k) pairs of variables do not share any measurement variables. Such insights may be used to simplify models, e.g. by removing measurement variables that induce multiple latents, or to build subsets of "pure" measurement variables, in the sense that the resulting measurement subsets are caused by disjoint sets of latents[6].

Finally, we note that there is more structure to be explored in the minMCM and figures 4 and 5 , but that is beyond our present scope. Note that the type of structure analyzed here emerges only when considering an ECC (i.e. patterns in the UDG, which is an abstraction of the correlation matrix) and not from the correlation matrix itself—analogous to higher-moment statistics or higher-order logic.

Our findings are not inconsistent with previous analyses of this data set, as can be seen by their agreement with points (i) and (iii) in Section 5.1, and should rather be seen as complementary. More generally our algorithm and corresponding analyses do not subsume existing methods but rather provide a novel perspective that allows us to focus on otherwise unutilized structure in measurement data, which in addition to helping to model the data also aids in, e.g., assessing and revising questionnaires and instruments.

6 DISCUSSION

Having in the preceding sections presented our minMCM finding algorithm, its supporting theory, and a demonstration application, we now conclude with two main directions for future work: the first direction is primarily concerned with applications of Algorithm 1 in its current state or requiring only minor modifications, while the second is primarily concerned with significantly extending

[6]Note that this is a bit different from the notion of "pure" used in the other measurement literature

Algorithm 1 and with developing new methods based on insights gleaned during its development.

6.1 FUTURE APPLICATIONS AND MINOR MODIFICATIONS

Being constraint-based, the Algorithm 4 relies on estimated independencies. Thus, errors in the inference of minMCMs come not from Algorithm 1 itself but rather from the estimation of independencies that it (along with many other causal inference methods) requires as input. In this regard, a single incorrectly estimated independence can in the (unlikely) worst case[7]result in incorrectly doubling or halving the number of estimated latents or edges. In any case, as mentioned at the end of Section 2, further estimates of conditional independencies can help corroborate or refute the estimated unconditional independencies. More detailed examination is needed to make this more theoretically precise as well as to determine how much of a problem this is likely to pose for real data.

One final caveat for interpreting minMCMs is that, for complex graphs, there can be multiple minimum ECCs (for both types of minimality), each with the same minimum number of cliques or assignments. Thus, while using a minMCM to reason about the minimum number of edges or latents is always valid, stronger conclusions may require that the graph $\mathrm{D}(\mathbf{M})$ admits only one minMCM (which is simple enough to test) or that further assumptions or background knowledge are used to justify one minMCM over other observationally consistent ones. To this end, the (non-minimal) MCM corresponding to maximal cliques (e.g., Figure 2(a)) may be especially interesting, because it contains all observationally consistent MCMs (including the minMCMs in 2(c) and 2(d)).

Another promising aspect of our approach for future work is its extensibility, which results from establishing MeDIL causal models as a causal semantics for edge clique covers. Though we have so far focused on minimal ECCs, a MCM corresponding to *any* ECC for a given UDG is guaranteed to be measurement-faithful and causally sufficient (though not minimal or measurement-Markov) for the corresponding distribution of measurement variables. Using a different class of ECCs simply requires a different algorithm to be used in Line 2 of Algorithm 1. Just as we expressed simplicity of the causal model in terms of the number of latents (or edges) in the MCM and therefore the number of cliques (or assignments) in the ECC, *any* property of a causal model that can be expressed in terms of properties of an ECC can be used to repurpose an ECC-finding algorithm for the desired CSL task. For example, developments in network science (Conte et al., 2019) make it possible for ECC-based causal analysis of very large graphs, even containing up to millions of nodes.

[7]This is when the inclusion/exclusion of a single edge in an $n \geq 3$ vertex undirected dependency graph makes the difference between the graph having $2(n-2)$ maximal cliques that are all edges and $n-2$ maximal cliques that are all triangles. Fortunately, such precarious structures are easy to detect and can be removed by picking different sets of measurements.

6.2 EXTENSIONS AND FURTHER DEVELOPMENTS

Because Algorithm 1 returns a causally sufficient DAG, it should be possible to actually learn a corresponding fully specified functional causal model using, e.g., some version of nonlinear ICA or variational autoencoders (Khemakhem et al., 2019) that has been modified to take into account the conditional independence structure. This could potentially lead to the development of a causal, non-parametric generalization of factor analysis (Martin and VanLehn, 1994) which would still be interestingly different from similar existing work (Hoyer et al., 2008; Kummerfeld and Ramsey, 2016). Furthermore, since learning such a FCM would require the data set and not just its CI relations, it would be straight-forward to make a score-based adaptation of Algorithm 1 inspired by (Elidan et al., 2001), where cliques are picked according to maximizing a scoring criterion instead of (possibly misestimated) CI relations. This would help overcome the potential pitfall mentioned in Section 6.1.

Additionally, notice that formally, (though not semantically) *every* DAG is a MCM: any given DAG $\mathcal{G}$ can be partitioned into sink nodes $\mathbf{S}$ and non-sink nodes $\mathbf{N}$, in which case it is observationally consistent with respect to $\mathbf{S}$ to any other DAG $\mathcal{H}$ whose (sub)set of sink nodes $\mathbf{S}'$ has the same UDG as $\mathbf{S}$. This allows for some of the theory developed in sections 2 and 3 to be easily repurposed to characterizing subset-Markov equivalence classes for DAGs with different sets of variables, as long as they have some subset of sink nodes $\mathbf{S} = \mathbf{S}'$ in common. This may help connect causal coarsening (Chalupka et al., 2016) with causally consistent transformations between micro- and macro-models (Rubenstein et al., 2017) and causal abstraction (Beckers and Halpern, 2019).

References

Andersson, S. A., Madigan, D., and Perlman, Michael D, e. a. (1997). A characterization of markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541.

Beckers, S. and Halpern, J. Y. (2019). Abstracting causal models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 2678–2685.

Bron, C. and Kerbosch, J. (1973). Algorithm 457: finding all cliques of an undirected graph. *Communications of the ACM*, 16(9):575–577.

Chalupka, K., Eberhardt, F., and Perona, P. (2016). Causal feature learning: an overview. *Behaviormetrika*, 44(1):137–164.

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of machine learning research*, 3(Nov):507–554.

Conte, A., Grossi, R., and Marino, A. (2019). Large-scale clique cover of real-world networks. *Information and Computation*, 270:104464.

Cygan, M., Pilipczuk, M., and Pilipczuk, M. (2016). Known algorithms for edge clique cover are probably optimal. *SIAM Journal on Computing*, 45(1):67–83.

Dwass, M. (1957). Modified randomization tests for nonparametric hypotheses. *The Annals of Mathematical Statistics*, 28(1):181–187.

Elidan, G., Lotner, N., Friedman, N., and Koller, D. (2001). Discovering hidden variables: A structure-based approach. In *Advances in Neural Information Processing Systems*, pages 479–485.

Ennis, J. M., Fayle, C. M., and Ennis, D. M. (2012). Assignment-minimum clique coverings. *Journal of Experimental Algorithmics*, 17(1):1.1.

Erdős, P., Goodman, A. W., and Pósa, L. (1966). The representation of a graph by set intersections. *Canadian Journal of Mathematics*, 18:106–112.

Gramm, J., Guo, J., Hüffner, F., and Niedermeier, R. (2009). Data reduction and exact algorithms for clique cover. *Journal of Experimental Algorithmics*, 13:2.2.

Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424.

Gretton, A., Bousquet, O., Smola, A., and Schölkopf, B. (2005). Measuring statistical dependence with hilbert-schmidt norms. *Algorithmic Learning Theory*, page 63–77.

Holland, P. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Hoover, K. D. (2006). Causality in economics and econometrics. *SSRN Electronic Journal*.

Hoyer, P. O., Shimizu, S., Kerminen, A. J., and Palviainen, M. (2008). Estimation of causal effects using linear non-gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362–378.

Khemakhem, I., Kingma, D. P., and Hyvärinen, A. (2019). Variational autoencoders and nonlinear ica: A unifying framework. *arXiv preprint arXiv:1907.04809*.

Kindermann, R. and Snell, J. L. (1980). Ii. markov fields on graphs. *Contemporary Mathematics*, page 24–33.

Kummerfeld, E. and Ramsey, J. (2016). Causal clustering for 1-factor measurement models. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1655–1664. ACM.

Kummerfeld, E., Ramsey, J., Yang, R., Spirtes, P., and Scheines, R. (2014). Causal clustering for 2-factor measurement models. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 34–49. Springer.

Martin, J. and VanLehn, K. (1994). Discrete factor analysis: Learning hidden variables in bayesian networks (technical report no. lrdc-onr-94-1). *LRDC, University of Pittsburgh: Pittsburgh, Pennsylvania*.

Parascandola, M. (2001). Causation in epidemiology. *Journal of Epidemiology & Community Health*, 55(12):905–912.

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

Pearl, J. and Verma, T. (1995). A theory of inferred causation. In *Studies in Logic and the Foundations of Mathematics*, volume 134, pages 789–811. Elsevier.

Rubenstein, P., Weichwald, S., Bongers, S., Mooij, J., Janzing, D., Grosse-Wentrup, M., and Schölkopf, B. (2017). Causal consistency of structural equation models. In *Proceedings of the Thirty-Third Annual Conference on Uncertainty in Artificial Intelligence (UAI 2017)*, page ID11.

Seth, A. K., Barrett, A. B., and Barnett, L. (2015). Granger causality analysis in neuroscience and neuroimaging. *Journal of Neuroscience*, 35(8):3293–3297.

Silva, R. and Scheines, R. (2005). Generalized measurement models. Technical report, Carnegie Mellon University.

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

Spirtes, P. and Glymour, C. (1991). An algorithm for fast recovery of sparse causal graphs. *Social Science Computer Review*, 9(1):62–72.

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

Spirtes, P., Meek, C., and Richardson, T. (1999). An algorithm for causal inference in the presence of latent variables and selection bias. *Computation, causation, and discovery*, 21:1–252.

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. (2007). Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769–2794.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, pages 255–270. Elsevier Science Inc.