# Neuro-Causal Factor Analysis

**Alex Markham** ALEX.MARKHAM@CAUSAL.DEV
*Department of Mathematical Sciences and Pioneer Centre for SMARTbiomed*
*University of Copenhagen, Denmark*

**Mingyu Liu**
*Stern School of Business, New York University, USA*

**Bryon Aragam**
*Booth School of Business, University of Chicago, USA*

**Liam Solus**
*Department of Mathematics, KTH Royal Institute of Technology, Sweden*

**Editors:** Gustau Camps-Valls, Manuele Leonelli and Gherardo Varando

## Abstract

Factor analysis (FA) is a statistical method for explaining how mutually dependent observed variables can be represented in terms of mutually independent latent factors, and it is widely used in the psychological, biological, and physical sciences. We revisit this classic method from the perspective of recent advances in causal structure learning and deep generative models, introducing a framework for *Neuro-Causal Factor Analysis (NCFA)*. Our approach is fully nonparametric: it learns a directed graph between latent and observed variables, and then fits a deep generative model constrained to respect the Markov factorization of the graph. Empirically, on synthetic and real data, NCFA attains better reconstruction error compared to standard FA and better latent distribution recovery compared to a standard variational autoencoder, all with the advantages of sparser architecture, lower model complexity, and causal interpretability.

**Keywords:** measurement model; representation learning; edge clique cover; one pure child.

## 1. Introduction

Since its development over a century ago, factor analysis (FA) [90] has been applied across scientific fields, including genomics, computational biology [77; 99], economics [26; 59], sociology [9] and many others. The goal of FA is to offer explanations of variability among dependent observables via (potentially) fewer latent variables that capture the joint dependence between the observables. For the sake of identifiability, it is common to assume linearity, although in practice it is well-known that many problems exhibit complex nonlinear latent structures. With the rise of nonparametric deep generative models that allow representing highly nonlinear relationships between dependent observables, one might hope to combine the best of both worlds.

Moreover, within applications such as those listed above, FA is considered useful because the learned factors (latents) may offer an interpretation of relevant observed correlations. Many applied FA studies provide an interpretation of the learned factors based on the observed variables whose joint correlation they encode. A natural tendency when trying to interpret these factors is to assume they reflect possible common causes linking observed variables. However, the models used in such studies are not necessarily built with causality

© 2026 A. Markham, M. Liu, B. Aragam & L. Solus.

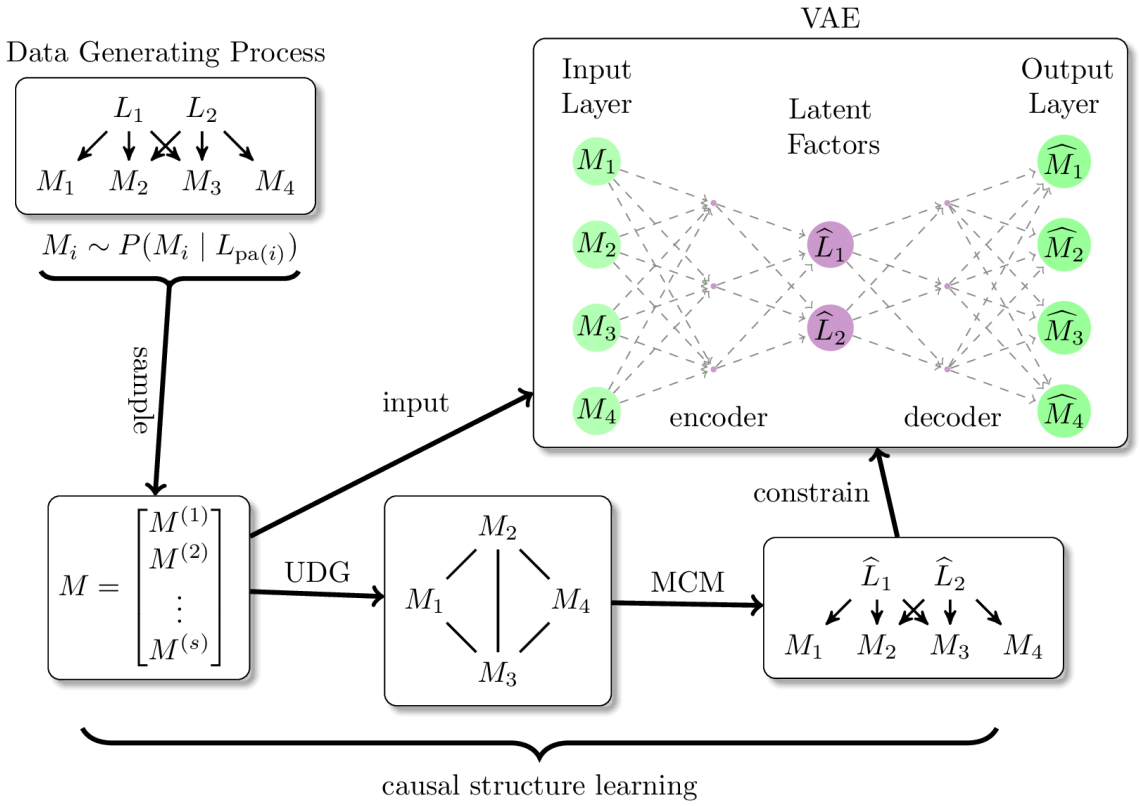



Figure 1: Learning a Neuro-Causal Factor Model.

in mind. Collectively, these considerations purport a need for a framework for nonlinear *causal* factor analysis that combines identifiability with flexibility through the use of modern advances in deep generative models and causality.

To this end, we propose *Neuro-Causal Factor Analysis (NCFA)*, augmenting classic FA on both fronts by leveraging advancements of the last few decades, including (i) causal discovery [91; 73] and (ii) deep generative models, such as variational autoencoders (VAEs) [49; 82]. To formalize this combination of ideas and apply it to settings where FA is typically invoked, we consider causal models that directly abide by *Reichenbach's common cause principle* [80, p. 157]: dependent variables in a system that do not share a direct causal relation should be explained by the existence of unobserved common causes which when conditioned upon render them independent. In particular, NCFA is applicable to problems where one can assume that the observed (or *measurement*) variables are rendered mutually independent when conditioning on a set of unobserved latent variables, which may be interpreted as *causally justifiable* factors from the FA perspective.

**Contributions** Our main contribution is the NCFA framework (Figure 1) for causally interpretable, identifiable FA models with the flexibility afforded by deep generative models. Unlike recent work (Section 2) that seeks to combine deep learning with causal models, NCFA is model-free and fully nonparametric, relying only on simple graphical assumptions. This is in contrast to approaches that rely on functional restrictions such as additive

noise, post-nonlinearity, or auxiliary information. Instead, we ensure identifiability through the graphical properties of the marginal independence structure induced by a latent factor model, which has not been previously leveraged for deep causal modeling. These assumptions arise in practical applications [e.g., 6; 8; 79] and are weaker than faithfulness. As a result, we obtain identifiability without imposing parametrization, functional relations, or conditions on data type (e.g., continuous, discrete, or mixed). Overall, NCFA provides a model-free perspective on latent causal models. Specifically:

1. NCFA (Algorithm 1) does not assume the underlying structure is known (i.e., it is learned entirely from data), allows flexible estimation of the latent space with deep generative models, and comes with fully nonparametric graph identifiability guarantees (Theorem 7; see Remark 9 for the distinction from full identifiability of the latent distribution). Unlike traditional causal discovery methods, NCFA relies only on *marginal* independence testing rather than general conditional independence testing, which is notoriously difficult [71; 12; 85]. As a result, the framework applies to any combination of discrete, continuous, or mixed variables as long as marginal dependence can be consistently tested. The advantages of this approach are illustrated with concrete examples (Appendix B) that cannot be captured by existing approaches.

2. We introduce the notion of *latent degrees of freedom* (Section 3), where additional representational capacity is achieved by giving each causal variable its own factorial prior. This separates neural network architectural choices from defining properties of the underlying causal graph.

3. We demonstrate (Section 4) on both synthetic and real data that NCFA injects generative models with interpretable structure and latent representations without significantly reducing reconstruction quality compared to unstructured generative models.

4. We provide an open source Python implementation as part of the `medil` package[1], along with a Snakemake workflow for reproducing all results[2].

## 2. Comparison to related work

We divide the vast amount of related work into three areas: factor analysis (Section 2.1), latent causal discovery (Section 2.2), and deep generative models (Section 2.3). Before describing each in more detail in its respective subsection, we first summarize their differing motivations and methods and provide a comparison to our proposed NCFA. A comprehensive comparison with recent related work on deep causal models can be found in Appendix C.

Factor analysis focuses on modeling measurement variables in terms of underlying factors (which can be interpreted as sources), focusing on model simplicity and interpretability, generally by assuming linear relations and jointly Gaussian random variables. Latent causal models focus on more detailed causal structure, not being limited to measurement variables and their latent sources, resulting in extremely interpretable models, but often at the expense of (arguably) strong, untestable assumptions like faithfulness. Deep generative models focus on learning as accurate a black box model as possible, optimizing a

1. https://medil.causal.dev
2. https://gitlab.com/alex-markham/medil/-/tree/v2.0.0/src/expt

highly overparameterized and nonlinear model to still achieve generalizability. Although the interpretation of deep generative models as nonlinear factor analysis is standard in the literature [e.g., 83; 70; 33], the additional dimensions of causality and identifiability are new to our approach. NCFA offers a unifying perspective on structured representation learning, incorporating the strengths of each of these approaches.

Like FA, we focus on modeling measurement variables in terms of their underlying sources; however, NCFA identifies these sources and their structural connections to the measurements through explicit latent causal structure learning, which is easier and requires weaker assumptions by focusing on sources cause measurements instead of more detailed intermediate causal structure. Furthermore, the source distributions and their corresponding functional relations to the measurements are estimated using a generative model whose architecture is constrained to respect the learned causal structure, gaining some of the expressiveness of deep generative models but regularized to maintain causal interpretability and generalizability. Hence, NCFA is motivated by the simplicity of FA, the causal interpretability of latent causal models, and the expressive power of deep generative models.

### 2.1. Factor analysis

We now give a brief introduction to FA, focusing on the key terms and mathematical ideas that we connect to latent causal discovery and deep generative models. For a more in-depth introduction, see Mulaik [69].

**Definition 1** *A* factor model *is defined by* $\mathbf{M} = \mathbf{L}W + \boldsymbol{\epsilon}$. *A random (row) vector* $\mathbf{M} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ *consisting of* $n$ measurement variables *is represented as a linear transformation of a standard jointly normal random vector* $\mathbf{L} \sim \mathcal{N}(\mathbf{0}, I_K)$ *of* $K < n$ latent factors *via* factor loading weights $W \in \mathbb{R}^{K\times n}$ *plus a jointly normal random vector of* $n$ error terms $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, D)$, *where* $D \in \mathbb{R}_+^{n\times n}$ *is a diagonal matrix.*

Given a sample $M \in \mathbb{R}^{s,n} \sim \mathbf{M}$ and assuming that $\mathbf{L}$ and $\boldsymbol{\epsilon}$ are probabilistically independent, the factor model can be estimated [1] from the empirical covariance matrix $\widehat{\Sigma} = \frac{1}{s}M^\top M$ by finding $\widehat{W}$ and $\widehat{D}$ that minimize the squared Frobenius norm $\|\widehat{\Sigma} - \widehat{W}^\top\widehat{W} - \widehat{D}\|_F^2$. Such a solution is unique only up to orthogonal transformations of $\widehat{W}$, and so without further (e.g., in our case, causal) assumptions, finding a solution does not always warrant a meaningful interpretation of the resulting factor model. This unidentifiability poses a problem in *exploratory* FA, where there is no prior knowledge about $\widehat{\Sigma}, \widehat{W}$ or $\widehat{D}$, but less so in *confirmatory* FA, where experts incorporate domain knowledge to constrain and interpret solutions as well as test specific hypotheses.

Additionally, there are possibilities for either restricting or relaxing the FA model, including closely related methods like PCA [75; 38; 45], ICA [17; 41], and many others beyond our scope. Notably, compared to other related work, sparse FA [72; 96; 105], which penalizes $\widehat{W}$ according to the number of nonzero entries, produces solutions more closely related to those we find with NCFA. The two main differences between sparse FA and NCFA are that (i) rather than explicitly penalizing the solution to encourage sparsity, NCFA simply learns a causal structure that exhibits a structure typically sought in sparse FA, and (ii) like most FA methods, sparse FA still assumes linearity and Gaussianity, whereas NCFA can be highly nonlinear and nonparametric. More recently, Kim and Zhou [48] proposed a

method similar to the one proposed here, restricted to linear models. Their identifiability results rely on their Condition 4.2, and are a linear special case of our Theorem 8.

### 2.2. Latent causal models

Graphical causal modeling [91; 73] focuses on learning a directed acyclic graph (DAG) representation of the causal relations among variables. This typically requires a strengthening of the common cause principle (into what is sometimes called the causal Markov assumption), which additionally assumes causal sufficiency, i.e., that there are no latent variables, and hence that all probabilistic dependences among the observed variables are due to causal relations among them. Methods for learning latent causal models have classically focused on learning DAG-like structure (using mixed instead of only directed graphs) among the observed variables to the extent allowed by confounding latent variables, exemplified by algorithms such as FCI [91; 16] and IC [74], which relax the Causal Markov Assumption. We also mention early work on this problem by [64; 27; 25]. In contrast, research on causal measurement models [87] is more closely related to the goal of FA, in that it too focuses on factor-measurement relations. Recently, there has been a surge of interest in these models, with advances leveraging additive noise models [60; 107; 39; 104; 4], independent mechanisms [35], weak supervision [57; 10], and interventions [13; 14; 2; 92; 98].

### 2.3. Structured deep generative models

The past decade has seen a flurry of work on training large-scale deep latent variable models, fueled by advances in variational inference and deep learning [e.g. 52; 49; 82; 22; 32; 81; 88]. More recently, there has been a trend towards *structured* latent spaces, such as hierarchical, graphical, causal, and disentangled structures. Conceptually, NCFA provides a theoretically principled approach to automatically *learning* latent structure from data in a way that is causally meaningful. The related work here needs to be divided into two categories: *known* (e.g., from prior knowledge) vs. *learned* latent structure. These can be further divided into *non-causal* vs. *causal* approaches. Given that our main contribution is *learned causal structure*, we will focus the discussion on the latter: For causal structure, *identifiability* becomes crucial, as it is well-known that nonparametric latent variable models are unidentifiable in general [42; 58].

**Known structure** Early work looked at incorporating known structure into generative models, such as autoregressive, graphical, and hierarchical structure [31; 44; 89; 101; 103; 21; 68]. This was later translated into known *causal* structure [51].

**Learned structure** When the latent structure is unknown, several techniques have been developed to learn useful (not necessarily causal) structure from data [55; 37; 102; 50; 67]. More recently, based on growing interest in disentangled [5] and/or causal [84] representation learning, methods that learn causal structure have been developed [66; 106; 4; 86; 46]. Subramanian et al. [94] assumes a linear Gaussian additive noise model, whereas Moraffah et al. [66] uses GANs. Unlike NCFA, neither Moraffah et al. [66] nor Subramanian et al. [94] come with identifiability guarantees. The remaining papers (detailed in Appendix C) all rely on functional restrictions (e.g., additive noise models) or auxiliary information (e.g.,

as in iVAE) to prove identifiability. In contrast to this existing work, NCFA admits nonparametric identifiability guarantees without additional labels or specifying a particular parametric or functional form (see Section 3.1).

## 3. NCFA models

Consider a collection of jointly distributed measurement variables $(M_1, \dots, M_n)$ for which we assume that all dependences are explained by the existence of latent common causes of the measured variables, i.e., that no $M_i$ and $M_j$ share a direct causal relation. If we were able to observe these latent confounders and condition upon them, $M_1, \dots, M_n$ would become mutually independent. Since every $M_i$ has out-degree 0, any $M_k$ lying on a path between two other measurements must itself be a collider, and any two dependent measurement variables must share a latent parent; consequently the unconditional (in)dependencies among $M_1, \dots, M_n$ already fully determine their conditional (in)dependencies [61, Proposition 6]. Hence, the only causal structure encoded via conditional independence in the observed distribution is contained in their marginal independence structure, which can be represented as an undirected graph:

**Definition 2** *The* unconditional dependence graph *(UDG) for the jointly distributed random variables $(M_1, \dots, M_n)$ is the undirected graph $\mathcal{U}$ with node set $[n] = \{1, \dots, n\}$ and edge set $E = \{i \text{ — } j : M_i \not\perp\!\!\!\perp M_j\}$.*

To recover a causal interpretation of the relations that hold among the measurement variables, we extend a UDG graph to a *(minimum) MCM*[3] *graph*. Following the principle of Occam's Razor, we would like to explain the observed dependences in $(M_1, \dots, M_n)$ in the simplest possible way, i.e., using the fewest possible latents to serve as the common causes of the measurement variables that exhibit dependence. To do so, we identify a *minimum edge clique cover* (ECC) of the UDG $\mathcal{U}$, which is a collection $\mathcal{C} = \{C_1, \dots, C_K\}$ of *cliques* (i.e., complete subgraphs of $\mathcal{U}$) such that for every $i \text{ — } j \in E$ the pair $i, j$ is contained in at least one clique in $\mathcal{C}$ and there exists no set of cliques with this property that has cardinality smaller than $|\mathcal{C}|$:

**Definition 3** *Let $\mathcal{U}$ be an undirected graph with minimum edge clique cover $\mathcal{C} = \{C_1, \dots, C_K\}$. The* (minimum) MCM graph *$\mathcal{G}$ for $\mathcal{U}$ and $\mathcal{C}$ is the DAG with vertices $[n] \cup L$ where $L = \{l_1, \dots, l_K\}$ and edge set $E = \{l_i \to j : j \in C_i,\ i \in [K]\}$. We call $|L|$ the number of* causal degrees of freedom *of the model.*

Since we assumed all marginal dependencies in $(M_1, \dots, M_n)$ are explainable by latent common causes, then the observed distribution $(M_1, \dots, M_n)$ is realizable as the marginal distribution of $(M_1, \dots, M_n)$ in the joint distribution $(M_1, \dots, M_n, L_1, \dots, L_K)$ that is Markov to the DAG $\mathcal{G}$, where $L_i$ is the random variable represented by the node $l_i$ in $\mathcal{G}$. From a factor analysis perspective, the latents $L_1, \dots, L_K$ are the factors to be inferred.

The minimum MCM graph defines a putative causal graph that respects the independence structure of $(M_1, \dots, M_n)$, and our goal is to learn the associated latent representations from data using a deep generative model. Consider the DAG $\mathcal{G}$ depicted in Figure 2(*a*)

3. These were originally defined in the context of *measurement dependence inducing latent (MeDIL) causal models* [61]. A summary of this theory is given in Appendix A.

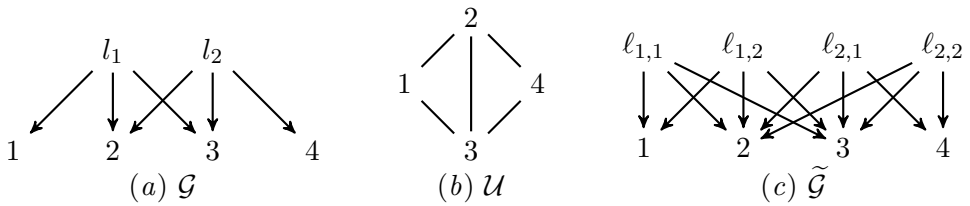



Figure 2: *(a)* shows an example minimum MCM graph with causal degrees of freedom $|L| = 2$, with corresponding UDG in *(b)*, and example NCFA-graph with latent degrees of freedom $\lambda = 4$ shown in *(c)*. Note that all three graphs encode the exact same set of (marginal) independencies among the measurement variables.

with two latents. A naïve approach would be to design a standard generative model, such as a variational autoencoder (VAE), such that the decoder respects the Markov properties implied by $\mathcal{G}$, however, it is unlikely that any generative model trained with a two-dimensional latent space will be able to represent the measurement variables accurately. The difficulty is that although the *true* causal structure involves only two latent variables, exactly fitting such a model is very difficult in practice. Thus, there is a tension between expressive capacity and respecting the causal structure.

We overcome this difficulty by replacing each causal latent with an overparameterized, factorial prior. The virtue of overparameterization is well-documented in the literature [78; 11]; in our setting this has the effect of increasing representational capacity without breaking the Markov structure encoded in $\mathcal{G}$. Formally, given a minimum MCM graph $\mathcal{G} = \langle [n] \cup L, E \rangle$, we replace each $l_i$ with a set of independent latent nodes $\mathcal{L}_i = \{\ell_{i,1}, \ldots, \ell_{i,k_i}\}$, for some $k_i \geq 1$, each with the same connectivity (i.e., children) as $l_i$. Thus, all told, we distribute $\lambda = \sum_{i \in [K]} k_i$ latents across the cliques, a hyperparameter called the *latent degrees of freedom*. It is easy to check that no matter how the $\lambda$ latent degrees of freedom are distributed, the resulting DAG has the same independence structure over the measurement variables as $\mathcal{G}$. This provides a rigorous device for increasing complexity without affecting the causal structure, and moreover, $\lambda$ is a flexible tuning hyperparameter that can be set arbitrarily large in practice, resulting in potentially overparameterized models. We call the resulting graph a *NCFA-graph* of $\mathcal{G}$ with $\lambda$ latent degrees of freedom:

**Definition 4** *Let $\mathcal{G}$ be a minimum MCM graph for the UDG $\mathcal{U} = \langle [n], E \rangle$ and the minimum edge clique cover $\mathcal{C} = \{C_1, \ldots, C_k\}$ of $\mathcal{U}$. A* NCFA graph *of $\mathcal{G}$ with $\lambda$ latent degrees of freedom is a graph $\widetilde{\mathcal{G}}$ with node set $[n] \cup \widetilde{\mathcal{L}}$ and edge set $\widetilde{E}$, where $\widetilde{\mathcal{L}} = \mathcal{L}_1 \cup \cdots \cup \mathcal{L}_k$ for $\mathcal{L}_i = \{\ell_{i,1}, \ldots, \ell_{i,k_i}\}$ with $k_i \geq 1 \;\; \forall i \in [k]$, and $\widetilde{E} = \{\ell_{i,m} \to j : j \in C_i, \; m \in k_i, \; i \in [k]\}$.*

Figure 2 shows an example of the three graph types considered so far in Definitions 2 to 4. Now we shift focus from graphs to models.

Each node $\ell_{i,m}$ represents a latent variable $Z_{i,m}$. Since the latent nodes in $\mathcal{L}_i$ all have the same connectivity as the single latent $l_i$, their joint distribution $f(L_i) = \prod_{m=1}^{k_i} f(Z_{i,m})$ represents the common cause of the measurement variables corresponding to the nodes in $C_i$, which was previously only represented by $l_i$ in $\mathcal{G}$. The factors to be inferred from a factor analysis perspective are now the random vectors $L_1, \ldots, L_K$ with $L_i = (Z_{i,1}, \ldots, Z_{i,k_i})$,

which still have the causal interpretation afforded by the minimum MCM graph. However, the multiple latents provide us flexibility to model the effects of the causal factors.

**Definition 5** *A* NCFA model *is a joint distribution* $(M_1, \dots, M_n)$ *for which there is an NCFA-graph* $\widetilde{\mathcal{G}} = \langle [n] \cup \widetilde{\mathcal{L}}, \widetilde{E} \rangle$ *and functions* $f_1, \dots, f_n$ *for which* $M_i := f_i(\mathrm{pa}_Z(i), \epsilon_i)$ *for all* $i \in [n]$*, where* $\mathrm{pa}_Z(i) := \{Z_{j,m} : \ell_{j,m} \in pa_{\widetilde{G}}(i)\}$*, and the measurement errors* $\epsilon_1, \dots, \epsilon_n$ *are jointly independent of each other and of* $Z$*.*

When modeling a distribution via an NCFA model, the functions $f_i$ are treated as unknowns to be inferred via a deep generative model. In our implementation, we used VAEs, however, other generative models can be used instead instead. The encoder maps the observations into the latent space as the joint posterior distribution $f(Z \mid M_1, \dots, M_n)$ where $Z$ is the random vector that collects the $Z_{j,m}$, and the decoder maps latents to observations according to

$$f(M_1, \dots, M_n \mid Z) = \prod_{i=1}^{n} f(M_i \mid \mathrm{pa}_Z(i)).$$

The joint distribution of the latent space is $f(Z) = \prod_{i=1}^{K} f(L_i)$; i.e., it is a product of the (joint) distributions we have specified to represent each of the latent common causes in the minimum MCM model $\mathcal{G}$ for $\mathcal{U}$. Following training of the generative model, the model may be used to generate predictions in the observation space via draws from the latent space. Since our representation of the latent space was constructed according to the minimum MCM graph $\mathcal{G}$, the resulting predictions can be viewed as causally informed; i.e., they are observations generated from the estimated distribution of the latent primary causes of the measurement variables.

### 3.1. Identifiability of minimum MCM graphs and ECC-model equivalence

While the UDG is identifiable, there may exist multiple minimum MCM graphs that yield the same UDG. This is because an undirected graph may have multiple, distinct minimum edge clique covers (see, for instance, Example 4 and surrounding discussion in Appendix B). In other words, similar to DAGs, minimum MCM graphs may be equivalent when provided with only observational data.

**Definition 6** *We say that two minimum MCM graphs* $\mathcal{G} = \langle [n] \cup L, E \rangle$ *and* $\mathcal{G}' = \langle [n] \cup L', E' \rangle$ *are* ECC-observationally equivalent *if* $i$ *and* $j$ *are* $d$*-separated given* $\emptyset$ *in* $\mathcal{G}$ *if and only if they are* $d$*-separated given* $\emptyset$ *in* $\mathcal{G}'$*.*

While there exist equivalence classes of minimum MCM graphs containing multiple elements, there also exist classes that are singletons; in other words, there exist undirected graphs (UDGs) with a unique minimum edge clique cover. For such UDGs, the minimum MCM graph is identifiable.

**Theorem 7** *Suppose that the data-generating distribution is Markov to a minimum MCM graph* $\mathcal{G}$*, and let* $\mathcal{U}$ *be the UDG induced by* $\mathcal{G}$*. Then the DAG* $\mathcal{G}$ *is identifiable from the data-generating distribution if:*

1. *$\mathcal{U}$ admits a unique minimum edge clique cover, and*

2. *$\mathcal{G}$ is* measurement-faithful *to the data-generating distribution [61, Def. 5].*

Proofs of this theorem and the following corollary are given in Appendix B. A concrete example of a class of practically relevant models that admit a unique minimum edge clique cover are those that satisfy the 1-pure-child assumption: For each latent $l_i$ in $\mathcal{G}$ there exists a measurement $i^*$ such that $\mathrm{pa}_{\mathcal{G}}(i^*) = \{l_i\}$. This assumption is widely used in machine learning [e.g., as in 23; 3; 7; 67], and also appears in biological applications of factor models [6; 8; 79]. This is also the main condition assumed by Kim and Zhou [48], who study linear factor models.

**Corollary 8** *If the data-generating distribution is Markov to a minimum MCM graph $\mathcal{G}$ satisfying the 1-pure-child assumption, and $\mathcal{G}$ is measurement-faithful to the data-generating distribution [61, Def. 5], then $\mathcal{G}$ is identifiable.*

Of course, this is just one example, and Theorem 7 shows that these are not the only models to which the identifiability result applies. An explicit construction of a model that admits a unique minimum edge clique cover but does not satisfy the 1-pure-child assumption is given in Example 2 (Appendix B). We emphasize once again that these results impose no functional restrictions, and in particular apply to general nonlinear models without additive noise or auxiliary information, including potentially mixed discrete variables.

**Remark 9** *Intuitively, a minimum MCM graph may be viewed as a summary of the confounding ancestors, and in particular the "causal sources", of a set of measurement variables. Imagine an arbitrary DAG, and consider its sink nodes to be the measurement variables—this is a (non-minimum) MCM graph. For each clique in a minimum edge clique cover of the corresponding UDG, the collection of common confounding ancestors that are relevant to that clique in the MCM graph get projected into a single latent factor in the minimum MCM graph. In this sense, the identified latent variables can be interpreted as a minimal projection of the underlying confounding structure onto the observational level.*

*At the distributional level, one could consider the pushforward measure induced by this projection. For example, in the linear Gaussian case, this would look like principal component analysis, and one could make use of linearity and Gaussianity to characterize identifiability of the minimal latent distribution; in the general NCFA setting, the corresponding latent representation may be nonlinear and nonparametric, but it still plays the same conceptual role as a compressed summary of the common causes.*

*Moreover, once the minimum MCM graph is identified, existing identifiability results [50; 67; 100] for deep generative models imply that, under suitable assumptions on the decoder and latent prior, the associated latent block distributions are identifiable up to the usual reparameterization ambiguities. Thus, graph identifiability provides the structural part of the argument, while the generative model identifiability theory supplies the distributional part.*

### 3.2. Learning

The data to UDG to NCFA-graph and model pipeline of the previous subsections can be formalized into Neuro-Causal Factor Analysis (NCFA) as Algorithm 1:

**Algorithm 1:** Neuro-Causal Factor Analysis (NCFA)

**Input:** sample $S$ of measurement variables $M$
**Parameter:** significance level $\alpha$, latent degrees of freedom $\lambda$
**Output:** neuro-causal factor model $\langle \widetilde{\mathcal{G}}, f_{[n]} \rangle$, with NCFA graph $\widetilde{\mathcal{G}}$ and loading functions $f_{[n]}$

1 Obtain undirected dependence graph $\mathcal{U}$ via pairwise marginal independence tests with threshold $\alpha$
2 Identify minimum edge clique cover $\mathcal{C}$ of $\mathcal{U}$ and construct corresponding minimum MCM graph $\mathcal{G}$
3 Assign remaining $\lambda - |\mathcal{C}|$ latents to the cliques in $\mathcal{C}$ to produce the NCFA-graph $\widetilde{\mathcal{G}}$
4 Estimate functions $f_{[n]}$ using a VAE constrained by $\widetilde{\mathcal{G}}$

To estimate the UDG, pairwise marginal independence tests are used. Starting with the complete graph, the edge $i - j$ is removed whenever $M_i$ and $M_j$ are deemed independent, i.e., according to a test with statistics such as distance covariance [95; 62] or xi correlation [15; 56]. Since these independence tests are estimated from finite data, they are subject to Type I and Type II errors, so the estimated UDG $\widehat{\mathcal{U}}$ may contain spurious edges or omit true ones; thus, although Theorem 7 establishes identifiability when the marginal independence relations are known, the graph recovered from a finite sample need not coincide with the true minimum MCM graph. A minimum edge clique cover is then identified for the estimated UDG $\widehat{\mathcal{U}}$. In general, this is an NP-hard problem, however there are both exact algorithms that work well for small graphs and heuristic algorithms that scale to large graphs [34; 18; 97].

Once a minimum edge clique cover is identified, the corresponding NCFA graph with $\lambda$ latent degrees of freedom is constructed. Here, we ensure that at every clique in the minimum edge clique cover of $\widehat{\mathcal{U}}$ is assigned at least one latent variable. The remaining $\lambda - k$ latents are then distributed uniformly over the cliques. Finally, a VAE for the functional relations specified by the NCFA-graph is trained. One could, in principle, alternatively use any deep generative model.

Since NCFA constructs its model via the MCM graph $\widehat{\mathcal{U}}$, the estimated factors (i.e., joint distributions) $f(L_i)$ in the factorization of the latent distribution represent the distributions for the primary causes of the measurement variables to which the latent nodes in $\mathcal{L}_i$ are connected. This yields a factor analysis model in which the latent factors can justifiably be causally interpreted. Furthermore, each causal latent $L_i = (Z_{i,1}, \ldots, Z_{i,k_i})$ is itself Gaussian, $L_i \sim \mathcal{N}(0, I_{k_i})$, for any choice of $k_i$. Increasing $k_i$ grows the dimension of this Gaussian input to the nonlinear decoder $f_i$, whose pushforward can represent arbitrarily non-Gaussian contributions to $M_i$. Hence, the estimated factors have a causal interpretation while additionally being as nonlinear as necessary.

## 4. Experiments

We now evaluate NCFA on synthetic and real data, comparing against classical factor analysis (FA), a fully connected variational autoencoder (VAE), and a linear Gaussian minimum

| | MSE ↓ | MCC ↑ | dCor ↑ | SFD ↓ | Time (s) | # Params |
|---|---|---|---|---|---|---|
| **FA** | 13.673 ± 7.358 | 0.459 ± 0.051 | 0.445 ± 0.036 | 508 ± 2 | 0.655 ± 0.132 | 120 ± 0 |
| **VAE** | 0.251 ± 0.027 | 0.756 ± 0.055 | 0.768 ± 0.017 | 508 ± 2 | 52.644 ± 7.254 | 7350 ± 0 |
| **LG-minMCM** | 0.965 ± 0.076 | 0.916 ± 0.016 | 0.908 ± 0.016 | 7 ± 6 | 0.061 ± 0.019 | 35 ± 1 |
| **NCFA** | 0.928 ± 0.078 | 0.898 ± 0.023 | 0.897 ± 0.018 | 5 ± 5 | 44.357 ± 9.329 | 6300 ± 0 |

Table 1: Results on linear Gaussian simulated data.

MCM model (LG-minMCM). Thus, we compare both sparse causal structure versus fully connected models (NCFA, LG-minMCM versus VAE, FA) as well as linear Gaussian vs nonparametric (FA, LG-minMCM versus VAE, NCFA). For quantitative evaluation we report, for each method:

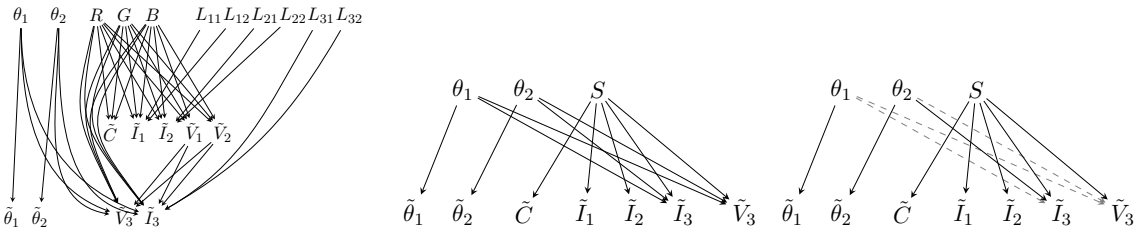
- **MSE**: mean squared reconstruction error on held-out data;
- **MCC**: mean correlation coefficient between matched pairs of learned and ground truth latents (matched to handle sign/permutation indeterminacy; see our implementation for details), a standard identifiability metric in the ICA and causal representation learning literature [40; 47; 28];
- **dCor**: distance correlation between joint learned and ground truth latent distributions—unlike MCC's linear matching, dCor also captures nonlinear reparameterizations, relevant given NCFA's nonlinear decoder;
- **SFD**: the shared factor distance between the learned and ground truth minimum MCM graphs (Appendix D), quantifying how well the causal structure is learned;
- **Time**: average wall-clock runtime; and
- **# Params**: the number of trainable parameters in the generative model.

Median values of all metrics are reported, plus or minus the median absolute deviation, computed over 101 random seeds. The first four metrics are computed on held out data, using a 70/30 train/test split.

LG-minMCM and NCFA both use the 1-pure-child assumption, ensuring fast recovery of the minimum edge clique cover. LG-minMCM estimates the BIC-optimal UDG, while NCFA uses xi correlation [15; 56] for independence tests.

All four methods compute held-out MSE identically (see our implementation for details); FA here is the saturated, fully connected linear Gaussian MCM ($K = n$), the dense counterpart to LG-minMCM's sparse structure, and its large MSE reflects this model's unidentifiability (Section 2.1).

For the synthetic experiments, we generate minimum MCM graphs over 10 measurement variables and 3 latents, with edge probability 0.3. On top of each such graph, we generate a dataset containing 10 000 observations from a linear Gaussian factor model. Table 1 shows the results. The assumptions for LG-minMCM are all satisfied, so it performs best on latent recovery (MCC and dCor), while being faster and using fewer parameters than the other methods. Nonetheless, the VAE is able to achieve better reconstruction (MSE), at the cost of worse latent recovery. This gap is not attributable to errors in graph estimation,




($a$) Ground-truth DAG. ($b$) Ground-truth minMCM. ($c$) Learned minMCM.

Figure 3: Panels show (a) the full ground truth DAG [29], (b) the true minMCM, and (c) an example estimated minMCM. False negatives in dashed gray; no false positives.

| | MSE ↓ | MCC ↑ | dCor ↑ | SFD ↓ | Time (s) | # Params |
|---|---|---|---|---|---|---|
| **FA** | 0.334 ± 0.048 | 0.346 ± 0.045 | 0.565 ± 0.065 | 169 ± 0 | 0.202 ± 0.048 | 63 ± 0 |
| **VAE** | 0.201 ± 0.004 | 0.503 ± 0.019 | 0.697 ± 0.002 | 169 ± 0 | 27.186 ± 6.171 | 6321 ± 0 |
| **LG-minMCM** | 0.566 ± 0.138 | 0.679 ± 0.014 | 0.570 ± 0.046 | 3 ± 2 | 0.051 ± 0.023 | 24 ± 1 |
| **NCFA** | 0.209 ± 0.008 | 0.885 ± 0.003 | 0.694 ± 0.003 | 10 ± 0 | 20.725 ± 4.986 | 5745 ± 0 |

Table 2: Results on real data from Causal Chamber's interventional light tunnel [29].

since LG-minMCM's assumptions are already satisfied here; rather, it reflects a capacity difference: the fully connected VAE has roughly 200× as many parameters as LG-minMCM (7350 vs. 35) and no sparsity constraint, letting it fit the reconstruction task more flexibly at the cost of its latents no longer corresponding to the true causal factors. NCFA attains comparable latent recovery to LG-minMCM and slightly better reconstruction, with run time and number of parameters comparable to the VAE. Thus, NCFA performs reasonably here, almost as well as the simple (correctly specified) parametric model and better than the VAE, which isn't designed for latent recovery.

For the real data [29], we treat sink nodes as measurement variables and the rest as latents, resulting in the ground truth minimum MCM graph in Figure 3($b$). The data is non-Gaussian, so the LG-minMCM is misspecified. Table 2 summarizes the results. Here, NCFA attains reconstruction performance comparable to VAE and the best latent recovery among all methods (in terms of MCC, though its dCor is nearly tied with VAE's, and LG-minMCM attains better structural recovery (SFD). The high MCC and low SFD are empirical corroboration of the identifiability results in Theorem 7 and Remark 9.

## 5. Conclusion

In summary, NCFA gives a graphical route to factor identifiability. Its main limitations are the interpretability of the unique minimum edge clique cover assumption and the extent to which the learned factors are genuinely causal. Future work includes exploring other graphical identifiability results for factor analysis [24; 93; 53; 36], replacing the constraint-based step with sparsity regularization for end-to-end training [67; 109], and using interventions to improve interpretability, identifiability, and out-of-distribution generalization [43; 63].

## Acknowledgments

The work by AM was partially supported by Novo Nordisk Foundation grant number NNF20OC0062897 and also by the Pioneer Centre for Statistical and computational Methods for Advanced Research to Transform Biomedicine (SMARTbiomed), DNRF grant number P4. Part of this research was performed while AM, BA, and LS were visiting the Institute for Mathematical and Statistical Innovation (IMSI), which is supported by the National Science Foundation (Grant No. DMS-1929348), and the authors thank the organizers of the IMSI long program *Algebraic Statistics and Our Changing World*. This research was supported in part through the computational resources and staff contributions provided for the Mercury high performance computing cluster at The University of Chicago Booth School of Business which is supported by the Office of the Dean.

## References

[1] Kohei Adachi. Factor analysis: Latent variable, matrix decomposition, and constrained uniqueness formulations. *Wiley Interdisciplinary Reviews: Computational Statistics*, 11(3):e1458, 2019. doi: 10.1002/wics.1458.

[2] Kartik Ahuja, Divyat Mahajan, Yixin Wang, and Yoshua Bengio. Interventional causal representation learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 372–407. PMLR, Jul 2023. URL https://proceedings.mlr.press/v202/ahuja23a.html.

[3] Sanjeev Arora, Rong Ge, and Ankur Moitra. Learning topic models—going beyond SVD. In *2012 IEEE 53rd annual symposium on foundations of computer science*, pages 1–10. IEEE, 2012. doi: 10.1109/focs.2012.49.

[4] Matthew Ashman, Chao Ma, Agrin Hilmkil, Joel Jennings, and Cheng Zhang. Causal reasoning in the presence of latent confounders via neural admg learning. In *The Eleventh International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=dcN0CaXQhT.

[5] Yoshua Bengio. Deep learning of representations: Looking forward. In *Statistical Language and Speech Processing: First International Conference, SLSP 2013, Tarragona, Spain, July 29-31, 2013. Proceedings 1*, pages 1–37. Springer, 2013. doi: 10.1007/978-3-642-39593-2_1.

[6] Xin Bing, Florentina Bunea, Martin Royer, and Jishnu Das. Latent model-based clustering for biological discovery. *Iscience*, 14:125–135, 2019. doi: 10.1016/j.isci.2019.03.018.

[7] Xin Bing, Florentina Bunea, Yang Ning, and Marten Wegkamp. Adaptive estimation in structured factor models with applications to overlapping clustering. *The Annals of Statistics*, 48(4), 2020. doi: 10.1214/19-aos1877.

[8] Xin Bing, Tyler Lovelace, Florentina Bunea, Marten Wegkamp, Sudhir Pai Kasturi, Harinder Singh, Panayiotis V Benos, and Jishnu Das. Essential regression: A generalizable framework for inferring causal latent factors from multi-omic datasets. *Patterns*, 3(5), 2022. doi: 10.1016/j.patter.2022.100473.

[9] Kenneth A Bollen. Instrumental variables in sociology and the social sciences. *Annual Review of Sociology*, 38:37–72, 2012. doi: 10.1146/annurev-soc-081309-150141.

[10] Johann Brehmer, Pim De Haan, Phillip Lippe, and Taco Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, volume 35, pages 38319–38331, 2022. doi: 10.52202/068431-2776.

[11] Rares-Darius Buhai, Yoni Halpern, Yoon Kim, Andrej Risteski, and David Sontag. Empirical study of the benefits of overparameterization in learning latent variable models. In *International Conference on Machine Learning*, pages 1211–1219. PMLR, 2020. URL https://proceedings.mlr.press/v119/buhai20a.html.

[12] Clément L Canonne, Ilias Diakonikolas, Daniel M Kane, and Alistair Stewart. Testing conditional independence of discrete distributions. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 735–748, 2018. doi: 10.1109/ita.2018.8503255.

[13] Krzysztof Chalupka, Pietro Perona, and Frederick Eberhardt. Visual causal feature learning. In *Proceedings of the Thirty-First Conference on Uncertainty in Artificial Intelligence*, UAI'15, page 181–190, Arlington, Virginia, USA, 2015. AUAI Press. ISBN 9780996643108. URL https://www.auai.org/uai2015/proceedings/papers/109.pdf.

[14] Krzysztof Chalupka, Frederick Eberhardt, and Pietro Perona. Causal feature learning: An overview. *Behaviormetrika*, 44(1):137–164, 2017. doi: 10.1007/s41237-016-0008-2.

[15] Sourav Chatterjee. A new coefficient of correlation. *Journal of the American Statistical Association*, 116(536):2009–2022, 2021. doi: 10.1080/01621459.2020.1758115.

[16] Diego Colombo, Marloes H Maathuis, Markus Kalisch, and Thomas S Richardson. Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, pages 294–321, 2012. doi: 10.1214/11-aos940.

[17] Pierre Comon. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994. doi: 10.1016/0165-1684(94)90029-9.

[18] Alessio Conte, Roberto Grossi, and Andrea Marino. Large-scale clique cover of real-world networks. *Information and Computation*, 270:104464, Feb 2020. doi: 10.1016/j.ic.2019.104464.

[19] Danai Deligeorgaki, Alex Markham, Pratik Misra, and Liam Solus. Combinatorial and algebraic perspectives on the marginal independence structure of Bayesian networks. *Algebraic Statistics*, 14(2):233–286, 2023. doi: 10.2140/astat.2023.14.233.

[20] Michel Marie Deza and Elena Deza. *Encyclopedia of Distances*. Springer, 4th edition, 2016. doi: 10.1007/978-3-662-52844-0.

[21] Mucong Ding, Constantinos Daskalakis, and Soheil Feizi. GANs with conditional independence graphs: On subadditivity of probability divergences. In Arindam Banerjee and Kenji Fukumizu, editors, *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pages 3709–3717. PMLR, Apr 2021. URL https://proceedings.mlr.press/v130/ding21e.html.

[22] Laurent Dinh, David Krueger, and Yoshua Bengio. NICE: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014. URL https://arxiv.org/abs/1410.8516.

[23] David Donoho and Victoria Stodden. When does non-negative matrix factorization give a correct decomposition into parts? *Advances in Neural Information Processing Systems*, 16, 2003. URL https://papers.nips.cc/paper_files/paper/2003/hash/1843e35d41ccf6e63273495ba42df3c1-Abstract.html.

[24] Mathias Drton, Alexandros Grosdos, Irem Portakal, and Nils Sturma. Algebraic sparse factor analysis. *SIAM Journal on Applied Algebra and Geometry*, 9(2):279–309, 2025. doi: 10.1137/23m1626517.

[25] Gal Elidan, Noam Lotner, Nir Friedman, and Daphne Koller. Discovering hidden variables: A structure-based approach. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, pages 458–464, 2000. URL https://papers.nips.cc/paper_files/paper/2000/hash/95e6834d0a3d99e9ea8811855ae9229d-Abstract.html.

[26] Mario Forni and Lucrezia Reichlin. Let's get real: A factor analytical approach to disaggregated business cycle dynamics. *The Review of Economic Studies*, 65(3):453–473, 1998. doi: 10.1111/1467-937x.00053.

[27] Nir Friedman. Learning belief networks in the presence of missing values and hidden variables. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, page 125–133, San Francisco, CA, USA, 1997. Morgan Kaufmann Publishers Inc. ISBN 1558604863.

[28] Juan L. Gamella, Simon Bing, and Jakob Runge. Sanity checking causal representation learning on a simple real-world system. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 18143–18169. PMLR, 2025. URL https://proceedings.mlr.press/v267/gamella25a.html.

[29] Juan L. Gamella, Jonas Peters, and Peter Bühlmann. Causal chambers as a real-world physical testbed for AI methodology. *Nature Machine Intelligence*, 7(1):107–118, January 2025. ISSN 2522-5839. doi: 10.1038/s42256-024-00964-x.

[30] Luis David Garcia, Michael Stillman, and Bernd Sturmfels. Algebraic geometry of Bayesian networks. *Journal of Symbolic Computation*, 39(3-4):331–355, 2005. doi: 10.1016/j.jsc.2004.11.007.

[31] Mathieu Germain, Karol Gregor, Iain Murray, and Hugo Larochelle. MADE: Masked autoencoder for distribution estimation. In *International conference on machine learning*, pages 881–889. PMLR, 2015. URL https://proceedings.mlr.press/v37/germain15.html.

[32] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in Neural Information Processing Systems*, 27, 2014. URL https://papers.nips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html.

[33] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[34] Jens Gramm, Jiong Guo, Falk Hüffner, and Rolf Niedermeier. Data reduction and exact algorithms for clique cover. *ACM Journal of Experimental Algorithmics*, 13, Feb 2009. doi: 10.1145/1412228.1412236.

[35] Luigi Gresele, Julius Von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? *Advances in Neural Information Processing Systems*, 34, 2021. URL https://papers.nips.cc/paper_files/paper/2021/hash/edc27f139c3b4e4bb29d1cdbc45663f9-Abstract.html.

[36] Yuqi Gu. Blessing of dependence: Identifiability and geometry of discrete models with multiple binary latent variables. *Bernoulli*, 31(2):948–972, 2025. doi: 10.3150/24-bej1754.

[37] Jiawei He, Yu Gong, Joseph Marino, Greg Mori, and Andreas Lehrmann. Variational autoencoders with jointly optimized latent dependency structure. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SJgsCjCqt7.

[38] Harold Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, 24(6):417, 1933. doi: 10.1037/h0071325.

[39] Biwei Huang, Charles Low, Feng Xie, Clark Glymour, and Kun Zhang. Latent hierarchical causal structure discovery with rank constraints. In *Advances in Neural Information Processing Systems*, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/24d2dd6dc9b79116f8ebc852ddb9dc94-Abstract-Conference.html.

[40] Aapo Hyvarinen and Hiroshi Morioka. Nonlinear ICA of temporally dependent stationary sources. In *Artificial Intelligence and Statistics*, pages 460–469. PMLR, 2017. URL https://proceedings.mlr.press/v54/hyvarinen17a.html.

[41] Aapo Hyvärinen and Erkki Oja. Independent component analysis: Algorithms and applications. *Neural networks*, 13(4-5):411–430, 2000. doi: 10.1016/s0893-6080(00)00026-5.

[42] Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural networks*, 12(3):429–439, 1999. doi: 10.1016/s0893-6080(98)00140-3.

[43] Yibo Jiang and Bryon Aragam. Learning nonparametric latent causal graphs with unknown interventions. *Advances in Neural Information Processing Systems*, 36:60468–60513, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/bdeab378efe6eb289714e2a5abc6ed42-Abstract-Conference.html.

[44] Matthew J Johnson, David K Duvenaud, Alex Wiltschko, Ryan P Adams, and Sandeep R Datta. Composing graphical models with neural networks for structured representations and fast inference. *Advances in Neural Information Processing Systems*, 29, 2016. URL https://papers.nips.cc/paper_files/paper/2016/hash/7d6044e95a16761171b130dcb476a43e-Abstract.html.

[45] Ian T Jolliffe. *Principal component analysis for special types of data*. Springer, 2002. doi: 10.1007/0-387-22440-8_13.

[46] David Kaltenpoth and Jilles Vreeken. Nonlinear causal discovery with latent confounders. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 15639–15654. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/kaltenpoth23a.html.

[47] Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvarinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020. URL https://proceedings.mlr.press/v108/khemakhem20a.html.

[48] Dale Kim and Qing Zhou. Structure learning of latent factors via clique search on correlation thresholded graphs. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 16978–16996, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/kim23aa.html.

[49] Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2014. URL https://arxiv.org/abs/1312.6114.

[50] Bohdan Kivva, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Identifiability of deep generative models without auxiliary information. *Advances in Neural Information Processing Systems*, 35:15687–

15701, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/649f080d8891ab4d4b262cb9cd52e69a-Abstract-Conference.html.

[51] Murat Kocaoglu, Christopher Snyder, Alexandros G Dimakis, and Sriram Vishwanath. CausalGAN: Learning causal implicit generative models with adversarial training. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=BJE-4xW0W.

[52] Hugo Larochelle and Iain Murray. The neural autoregressive distribution estimator. In Geoffrey Gordon, David Dunson, and Miroslav Dudík, editors, *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 29–37. PMLR, Apr 2011. URL https://proceedings.mlr.press/v15/larochelle11a.html.

[53] Seunghyun Lee and Yuqi Gu. Identifiability of latent causal graphical models without pure children. *arXiv preprint arXiv:2505.18410*, 2025. URL https://arxiv.org/abs/2505.18410.

[54] Chunlin Li, Xiaotong Shen, and Wei Pan. Nonlinear causal discovery with confounders. *Journal of the American Statistical Association*, pages 1–10, 2023. doi: 10.1080/01621459.2023.2179490.

[55] Xiaopeng Li, Zhourong Chen, Leonard K. M. Poon, and Nevin L. Zhang. Learning latent superstructures in variational autoencoders for deep multidimensional clustering. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=SJgNwi09Km.

[56] Z Lin and F Han. On boosting the power of Chatterjee's rank correlation. *Biometrika*, 08 2022. doi: 10.1093/biomet/asac048. asac048.

[57] Yuhang Liu, Zhen Zhang, Dong Gong, Mingming Gong, Biwei Huang, Anton van den Hengel, Kun Zhang, and Javen Qinfeng Shi. Identifying weight-variant latent causal models. *Journal of Machine Learning Research*, 27(4):1–49, 2026. URL http://jmlr.org/papers/v27/23-1023.html.

[58] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning (ICML)*, pages 4114–4124. PMLR, 2019. URL https://proceedings.mlr.press/v97/locatello19a.html.

[59] Sydney C Ludvigson and Serena Ng. The empirical risk–return relation: A factor analysis approach. *Journal of financial economics*, 83(1):171–222, 2007. doi: 10.1016/j.jfineco.2005.12.002.

[60] Takashi Nicholas Maeda and Shohei Shimizu. Causal additive models with unobserved variables. In *Uncertainty in Artificial Intelligence*, pages 97–106. PMLR, 2021. URL https://proceedings.mlr.press/v161/maeda21a.html.

[61] Alex Markham and Moritz Grosse-Wentrup. Measurement dependence inducing latent causal models. In *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI)*, volume 124 of *Proceedings of Machine Learning Research*, pages 590–599. PMLR, 03–06 Aug 2020. URL https://proceedings.mlr.press/v124/markham20a.html.

[62] Alex Markham, Richeek Das, and Moritz Grosse-Wentrup. A distance covariance-based kernel for nonlinear causal clustering in heterogeneous populations. In *Conference on Causal Learning and Reasoning (CLeaR)*, pages 542–558. PMLR, 2022. URL https://proceedings.mlr.press/v177/markham22a.html.

[63] Alex Markham, Isaac Hirsch, Jeri A. Chang, Liam Solus, and Bryon Aragam. Intervening to learn and compose causally disentangled representations. In Bijan Mazaheri and Niels Richard Hanson, editors, *Proceedings of the Fifth Conference on Causal Learning and Reasoning*, volume 323 of *Proceedings of Machine Learning Research*, pages 1481–1525. PMLR, 06–08 Apr 2026. URL https://proceedings.mlr.press/v323/markham26a.html.

[64] J Martin and Kurt VanLehn. Discrete factor analysis: Learning hidden variables in Bayesian networks. Technical report, Technical report, Department of Computer Science, University of Pittsburgh, 1995.

[65] Christopher Meek. Strong completeness and faithfulness in Bayesian networks. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 411–418, 1995.

[66] Raha Moraffah, Bahman Moraffah, Mansooreh Karami, Adrienne Raglin, and Huan Liu. Causal adversarial network for learning conditional and interventional distributions. *arXiv preprint arXiv:2008.11376*, 2020. URL https://arxiv.org/abs/2008.11376.

[67] Gemma Elyse Moran, Dhanya Sridhar, Yixin Wang, and David Blei. Identifiable deep generative models via sparse decoding. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=vd0onGWZbE.

[68] Jacobie Mouton and Rodney Stephen Kroon. Integrating Bayesian network structure into residual flows and variational autoencoders. *Transactions on Machine Learning Research*, 2023. URL https://openreview.net/forum?id=OsKXlWamTQ.

[69] Stanley A Mulaik. *Foundations of factor analysis*. CRC press, 2009. doi: 10.1201/b15851.

[70] Kevin P Murphy. *Probabilistic machine learning: An introduction*. MIT press, 2022.

[71] Matey Neykov, Sivaraman Balakrishnan, and Larry Wasserman. Minimax optimal conditional independence testing. *The Annals of Statistics*, 49(4):2151–2177, 2021. doi: 10.1214/20-aos2030.

[72] Lipeng Ning and Tryphon T Georgiou. Sparse factor analysis via likelihood and l1-regularization. In *2011 50th ieee conference on decision and control and european control conference*, pages 5188–5192. IEEE, 2011. doi: 10.1109/cdc.2011.6161415.

[73] Judea Pearl. *Causality*. Cambridge university press, 2009. doi: 10.1017/cbo9780511803161.

[74] Judea Pearl and Thomas S Verma. A theory of inferred causation. In *Studies in Logic and the Foundations of Mathematics*, volume 134, pages 789–811. Elsevier, 1995. doi: 10.1016/s0049-237x(06)80074-1.

[75] Karl Pearson. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin philosophical magazine and journal of science*, 2(11): 559–572, 1901. doi: 10.1080/14786440109462720.

[76] Jonas Peters, Joris M. Mooij, Dominik Janzing, and Bernhard Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 15(58):2009–2053, 2014. URL http://jmlr.org/papers/v15/peters14a.html.

[77] Iosifina Pournara and Lorenz Wernisch. Factor analysis for gene regulatory networks and transcription factor activity profiles. *BMC bioinformatics*, 8:1–20, 2007. doi: 10.1186/1471-2105-8-61.

[78] Adityanarayanan Radhakrishnan, Mikhail Belkin, and Caroline Uhler. Overparameterized neural networks implement associative memory. *Proceedings of the National Academy of Sciences*, 117(44):27162–27170, 2020. doi: 10.1073/pnas.2005013117.

[79] Javad Rahimikollu, Hanxi Xiao, Anna E. Rosengart, Aaron B. I. Rosen, Tracy Tabib, Paul Zdinak, Kun He, Xin Bing, Florentina Bunea, Marten Wegkamp, Amanda C. Poholek, Alok V Joglekar, Robert A Lafyatis, and Jishnu Das. SLIDE: Significant latent factor interaction discovery and exploration across biological domains. *Nature Methods*, 21(5):835–845, 2024. doi: 10.1038/s41592-024-02175-z.

[80] Hans Reichenbach. *The direction of time*. University of California Press, 1956.

[81] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *International conference on machine learning*, pages 1530–1538. PMLR, 2015. URL https://proceedings.mlr.press/v37/rezende15.html.

[82] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International conference on machine learning*, pages 1278–1286. PMLR, 2014. URL https://proceedings.mlr.press/v32/rezende14.html.

[83] Sam Roweis and Zoubin Ghahramani. A unifying review of linear Gaussian models. *Neural Computation*, 11(2):305–345, 1999. doi: 10.1162/089976699300016674.

[84] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/jproc.2021.3058954.

[85] Rajen D Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020. doi: 10.1214/19-aos1857.

[86] Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 23(241):1–55, 2022. URL http://jmlr.org/papers/v23/21-0080.html.

[87] Ricardo Silva, Richard Scheines, Clark Glymour, and Peter L Spirtes. Learning measurement models for unobserved variables. In *Proceedings of the 19th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 543–550, 2003.

[88] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. URL https://proceedings.mlr.press/v37/sohl-dickstein15.html.

[89] Casper Kaae Sønderby, Tapani Raiko, Lars Maaløe, Søren Kaae Sønderby, and Ole Winther. Ladder variational autoencoders. *Advances in Neural Information Processing Systems*, 29, 2016. URL https://papers.nips.cc/paper_files/paper/2016/hash/6ae07dcb33ec3b7c814df797cbda0f87-Abstract.html.

[90] C Spearman. "General intelligence," objectively determined and measured. *American Journal of Psychology*, 15:201, 1904. doi: 10.2307/1412107.

[91] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. *Causation, prediction, and search*. MIT press, 2000. doi: 10.7551/mitpress/1754.001.0001.

[92] Chandler Squires, Anna Seigal, Salil S Bhate, and Caroline Uhler. Linear causal disentanglement via interventions. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 32540–32560. PMLR, Jul 2023. URL https://proceedings.mlr.press/v202/squires23a.html.

[93] Nils Sturma, Miriam Kranzlmueller, Irem Portakal, and Mathias Drton. Matching criterion for identifiability in sparse factor analysis. *Psychometrika*, pages 1–39, 2025. doi: 10.1017/psy.2026.10079.

[94] Jithendaraa Subramanian, Yashas Annadani, Ivaxi Sheth, Nan Rosemary Ke, Tristan Deleu, Stefan Bauer, Derek Nowrouzezahrai, and Samira Ebrahimi Kahou. Learning latent structural causal models. *arXiv preprint arXiv:2210.13583*, 2022. URL https://arxiv.org/abs/2210.13583.

[95] Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, pages 2769–2794, 2007. doi: 10.1214/009053607000000505.

[96] Nickolay T Trendafilov, Sara Fontanella, and Kohei Adachi. Sparse exploratory factor analysis. *Psychometrika*, 82:778–794, 2017. doi: 10.1007/s11336-017-9575-8.

[97] Ahammed Ullah. Computing clique cover with structural parameterization. *arXiv preprint arXiv:2208.12438*, 2022. URL https://arxiv.org/abs/2208.12438.

[98] Burak Varici, Emre Acarturk, Karthikeyan Shanmugam, Abhishek Kumar, and Ali Tajer. Score-based causal representation learning with interventions. *arXiv preprint arXiv:2301.08230*, 2023. URL https://arxiv.org/abs/2301.08230.

[99] Britta Velten, Jana M Braunger, Ricard Argelaguet, Damien Arnol, Jakob Wirbel, Danila Bredikhin, Georg Zeller, and Oliver Stegle. Identifying temporal and spatial patterns of variation from multimodal data using mefisto. *Nature methods*, 19(2): 179–186, 2022. doi: 10.1038/s41592-021-01343-9.

[100] Yixin Wang, David Blei, and John P Cunningham. Posterior collapse and latent variable non-identifiability. *Advances in Neural Information Processing Systems*, 34, 2021. URL https://papers.nips.cc/paper_files/paper/2021/hash/2b6921f2c64dee16ba21ebf17f3c2c92-Abstract.html.

[101] Stefan Webb, Adam Golinski, Rob Zinkov, N. Siddharth, Tom Rainforth, Yee Whye Teh, and Frank Wood. Faithful inversion of generative models for effective amortized inference. *Advances in Neural Information Processing Systems*, 31, 2018. URL https://papers.nips.cc/paper_files/paper/2018/hash/894b77f805bd94d292574c38c5d628d5-Abstract.html.

[102] Antoine Wehenkel and Gilles Louppe. Graphical normalizing flows. In *International Conference on Artificial Intelligence and Statistics*, pages 37–45. PMLR, 2021. URL https://proceedings.mlr.press/v130/wehenkel21a.html.

[103] Christian Weilbach, Boyan Beronov, Frank Wood, and William Harvey. Structured conditional continuous normalizing flows for efficient amortized inference in graphical models. In *International Conference on Artificial Intelligence and Statistics*, pages 4441–4451. PMLR, 2020. URL https://proceedings.mlr.press/v108/weilbach20a.html.

[104] Feng Xie, Biwei Huang, Zhengming Chen, Yangbo He, Zhi Geng, and Kun Zhang. Identification of linear non-Gaussian latent hierarchical structure. In *International Conference on Machine Learning*, pages 24370–24387. PMLR, 2022. URL https://proceedings.mlr.press/v162/xie22a.html.

[105] Michio Yamamoto, Kei Hirose, and Haruhisa Nagata. Graphical tool of sparse factor analysis. *Behaviormetrika*, 44:229–250, 2017. doi: 10.1007/s41237-016-0007-3.

[106] Mengyue Yang, Furui Liu, Zhitang Chen, Xinwei Shen, Jianye Hao, and Jun Wang. CausalVAE: Disentangled representation learning via neural structural causal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9593–9602, 2021. doi: 10.1109/cvpr46437.2021.00947.

[107] Yuqin Yang, AmirEmad Ghassami, Mohamed Nafea, Negar Kiyavash, Kun Zhang, and Ilya Shpitser. Causal discovery in linear latent variable models subject to measurement error. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 874–886. Curran Associates, Inc., 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/05b63fa06784b71aab3939004e0f0a0d-Abstract-Conference.html.

[108] Yue Yu, Jie Chen, Tian Gao, and Mo Yu. DAG-GNN: DAG structure learning with graph neural networks. In *International Conference on Machine Learning*, pages 7154–7163. PMLR, 2019. URL https://proceedings.mlr.press/v97/yu19a.html.

[109] Wenjin Zhang, Yixin Wang, and Yuqi Gu. Discrete causal representation learning. *arXiv preprint arXiv:2603.25017*, 2026. URL https://arxiv.org/abs/2603.25017.

## Appendix A. MeDIL Causal Models

The NCFA models used in this paper are a subfamily of the models known as measurement dependence inducing latent (MeDIL) causal models, originally introduced in [61]. For contextualization purposes, we give a brief description of MeDIL causal models here.

A *MCM graph* is a triple $\mathcal{G} = \langle M, L, E\rangle$ where $M = \{1, \dots, n\}$ and $L = \{l_1, \dots, l_K\}$ are disjoint sets of vertices corresponding, respectively, to observed variables and latent variables, and $E$ is the collection of directed edges between the nodes. In a MCM graph we require that the nodes in $M$ are all *sinks* (i.e., have out-degree 0) and that the edges $E$ are such that $\mathcal{G}$ is a *directed acyclic graph* (DAG). A distribution belongs to the *Measurement Dependence Inducing Latent (MeDIL) Causal Model* if it factorizes according to $\mathcal{G}$, i.e., if the probability density function (or probability mass function) satisfies

$$f(x_1, \dots, x_n, x_{l_1}, \dots, x_{l_K}) = \prod_{i=1}^{n} f(x_i | x_{\mathrm{pa}_{\mathcal{G}}(i)}) \prod_{j=1}^{K} f(x_{l_j} | x_{\mathrm{pa}_{\mathcal{G}}(l_i)}).$$

Since we do not observe the latent variables, but only the observed variables, $X_1, \dots, X_n$, the distribution of interest is the marginal obtained by integrating out the latents:

$$f(x_1, \dots, x_n) = \int_{\mathcal{X}_L} \prod_{i=1}^{n} f(x_i | x_{\mathrm{pa}_{\mathcal{G}}(i)}) \prod_{j=1}^{K} f(x_{l_j} | x_{\mathrm{pa}_{\mathcal{G}}(l_j)}) dx_L. \tag{1}$$

The MeDIL model $\mathcal{M}(\mathcal{G})$ consists of all observable distributions that arise as this marginal.

In general, MCM graphs may have a complex directed acyclic structure over the latents that induce correlations amongst the observables. However, since we only consider the observed variables, we may reduce to a minimum MCM representation of the distribution. In a *minimum MCM graph*, we further assume that the latent variables are *source nodes* (i.e., have in-degree 0) and out-degree at least 1. Hence, a *minimum MeDIL model*—i.e., a MeDIL model that factorizes as in equation (1)—explains the associations among the observed variables in the simplest possible way: via a collection of independent latents. These independent latents can be thought of as source nodes in the original MCM graph $G$, and the associated minimum MCM graph as the graph produced after we marginalize out all non-source latent variables. While learning the complex causal structure on the latents in a general MCM graph may be intractable, one can apply existing methods to estimate the simpler minimum MCM structure. Intuitively, one can think of learning the minimum MCM graph of a model as identification of the "primary causes" of the associations between the observed variables.

It follows from the factorization of a distribution according to a minimum MCM graph that the only conditional independence constraints amongst the observed variables are marginal independence constraints. These, in turn, encode the condition that two nodes do not have a shared (latent) parent. From the causal perspective, we see that two nodes in the system are marginally independent if and only if they share a latent common cause. Hence, minimum MeDIL models are the natural representation of Reichenbach's Common Cause Principle.

It follows that a natural representation of the model using only the observed variables is via an undirected graph $\mathcal{U} = \langle M, E\rangle$, where $i - j \notin E$ if and only if $i$ and $j$ are independent

in the joint distribution of the observed variables. We call this graph the *unconditional dependence graph* (UDG) of the model. A *minimum edge clique cover* of the undirected graph $U$ is a collection of cliques $\mathcal{C} = \{C_1, \ldots, C_K\}$ for which every pair $i, j$ satisfying $i - j \in E$ is contained in at least one clique in $\mathcal{C}$ and there exists no set of cliques with this property that has cardinality smaller than $|\mathcal{C}|$. Since a UDG depends only on the observed variables, it is possible to learn a UDG from the available data via pairwise marginal independence tests. A minimum MCM graph that captures the marginal independence structure encoded via the UDG is then the minimum MCM graph where we have one latent $l_i$ for each clique $C_i \in \mathcal{C}$ and $l_i$ has a directed arrow to each node $j \in C_i$ (and no other adjacencies).

To learn the assignment of latents producing a minimum MCM graph from a UDG $\mathcal{U}$, we must learn a minimum edge clique cover of $\mathcal{U}$. The problem of identifying a minimum edge clique cover of an undirected graph is NP-hard. There are exact algorithms [34; 97] showing it to be NP-complete and fixed-parameter tractable. However, there are also polynomial-time heuristic algorithms [18] that can efficiently handle very large graphs.

## Appendix B. Identifiability of minimum MCM graphs: Examples and Counterexamples

**Proof** [Proof of Theorem 7] Because the data-generating distribution is Markov to $\mathcal{G}$, the use of any consistent nonparametric test for independence (e.g., [15]) is guaranteed to identify all marginal independence statements in the limit. By measurement-faithfulness (Condition 2), these coincide exactly with the edges of $\mathcal{U}$, the UDG induced by $\mathcal{G}$. Using [61, Proposition 7] to associate $\mathcal{U}$ with a minimum edge clique cover recovering $\mathcal{G}$, the uniqueness of this minimum edge clique cover (Condition 1) guarantees the identifiability of $\mathcal{G}$. ■

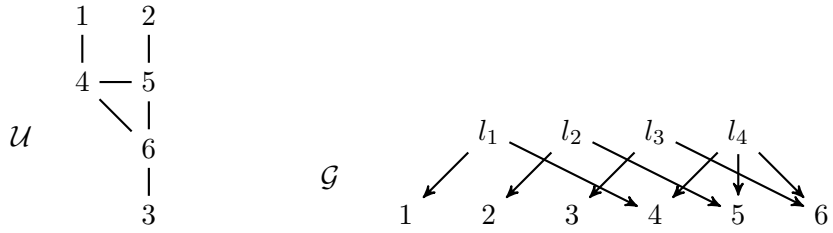
**Proof** [Proof of Corollary 8] We show the 1-pure-child assumption gives Condition 1 of Theorem 7; Condition 2 is assumed directly in the corollary's hypothesis. The association of a minimum MCM graph $\mathcal{G}$ with a minimum edge clique cover of $\mathcal{U}$ [61, Proposition 7] induces an association between a maximum independent set in $\mathcal{U}$ and a set of pure children in $\mathcal{G}$ with different parents. Observe that $\mathcal{G}$ satisfies the 1-pure-child assumption if and only if there exists a maximum independent set $\{i^*\}_{i=1}^K$ in $\mathcal{U}$ consisting of one pure child for each latent. In other words, $\mathcal{G}$ satisfies the 1-pure-child assumption if and only if the independence number (i.e., size of a maximum independent set, which is the number latents with at least one pure child) and intersection number (i.e., the number of cliques in a minimum edge clique cover, which is the number of latents) of its corresponding $\mathcal{U}$ are equal. It follows from Lemma 2.3.5 of Deligeorgaki et al. [19] that $\mathcal{U}$ has a unique minimum edge clique cover, giving Condition 1. Hence $\mathcal{G}$ is identifiable by Theorem 7. ■

The following examples illustrate the types of models that can be identified via Theorem 7.

**Example 1** *Graphs admitting a* pure measurement variable *in each clique of a minimum edge clique cover will have a unique minimum edge clique cover, making the resulting minimum MCM graph identifiable (see Theorem 7). From a graphical perspective, we call this condition the* 1-pure child constraint *since it insists that each latent node $l_i$ in the minimum*

*MCM graph has (at least) one child $i^*$ which has no other parents than $l_i$ (i.e., $l_i \to i^*$ is an edge of $\mathcal{G}$ and $i^*$ has no other adjacencies in $\mathcal{G}$).*

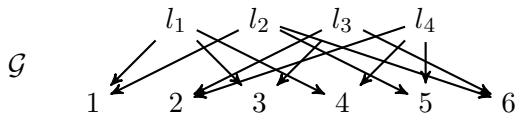
**Example 2** *This example shows that the 1-pure-child constraint in Example 1 is not necessary for a UDG to have a unique minimum edge clique cover, and thus Theorem 7 ensures identifiability of models violating this constraint. In particular, the following UDG $\mathcal{U}$ has the unique minimum edge clique cover $\mathcal{C} = \{\{1,4\},\{2,5\},\{3,6\},\{4,5,6\}\}$ and hence the corresponding minimum MCM graph $\mathcal{G}$ is identifiable:*



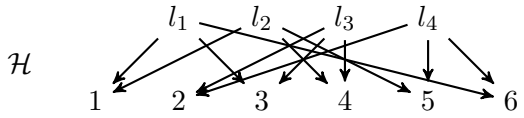

*Note here that the latent variables $l_1, l_2$ and $l_3$ each have a pure child ($1, 2$ and $3$, respectively). However, the latent $l_4$ does not. Hence $\mathcal{G}$ is identifiable but does not satisfy the 1-pure-child constraint.*

**Example 3** *Since previous work mainly focuses on continuous data, we point out that our assumptions are easily satisfied by models with discrete variables. For example, Assumption 2 in Theorem 7 is known to hold generically [65] for discrete (categorical) DAGs. Specifically, Meek [65] showed the existence of discrete distributions that are faithful to any given DAG for any choice of state spaces of the discrete variables. These same distributions necessarily fulfill Assumption 2 in Theorem 7. Moreover, any marginal independence relation $X_k \perp\!\!\!\perp X_\ell$ that is not encoded by the DAG is satisfied by the distributions represented as points in the open probability simplex that vanish on a collection of $2 \times 2$ minors of the matrix of parameters encoding the joint probabilities of $X_k$ and $X_\ell$ [30]. Using computational algebra software as described in [30], one can check for small examples of DAG models that the intersection of this set of points satisfying the polynomial constraints corresponding to $X_k \perp\!\!\!\perp X_\ell$ with the points vanishing on the polynomials corresponding to the CI relations encoded by the DAG has a lower dimension than the DAG model. This implies that almost all points in the DAG model, including points that are not faithful to the DAG, still satisfy Assumption 2 in Theorem 7. Hence, we may observe that Assumption 2 is strictly weaker than faithfulness. This, combined with the observations in Examples 1-2 provides a broad family of discrete models that satisfy our identifiability conditions, and that cannot be represented by existing work (see Appendix C).*

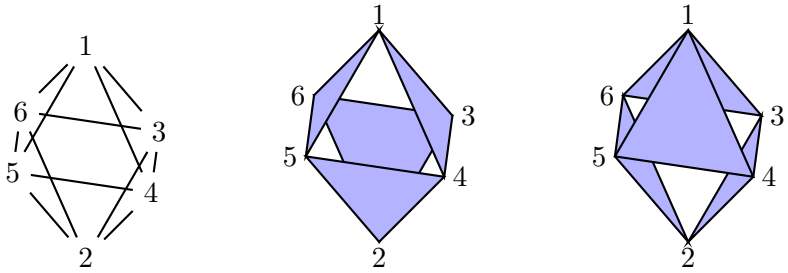
**Example 4** *An example of a minimum MCM graph that cannot be identified because it admits multiple edge clique covers is the following:*



*The UDG of $\mathcal{G}$ is the edge graph $\mathcal{U}$ of the octahedron which admits exactly two minimum edge clique covers: $\mathcal{C}_1 = \{\{1,3,4\},\{1,5,6\},\{2,3,6\},\{2,4,5\}\}$ and $\mathcal{C}_2 = \{\{1,3,6\},\{1,4,5\},\{2,3,4\},\{2,5,6\}\}$. $\mathcal{C}_1$ recovers the minimum MCM graph $\mathcal{G}$, whereas $\mathcal{C}_2$ produces the minimum MCM graph $\mathcal{H}$, depicted below:*




*This issue arises due to the spherical structure of the UDG $\mathcal{U}$, depicted on the left below. In the remaining two figures, the each shaded triangle represents the children of a latent node (i.e., a clique in a minimum edge clique cover) of a possible MCM graph with associated UDG $\mathcal{U}$. The middle figure represents the minimum MCM graph $\mathcal{G}$ and the rightmost figure represents the alternative $\mathcal{H}$.*




*The spherical structure of this graph creates symmetry allowing for multiple minimum edge clique covers. In particular, this gives an example to where we lack identifiability and Theorem 7 does not apply.*

Interestingly, Example 4 illustrates that how identifiability breaks down in practice, in the sense that if the true minimum MCM graph is $\mathcal{G}$, but the wrong edge clique cover is chosen, we could learn $\mathcal{H}$ instead. Future work, possibly incorporating interventional knowledge, to address such identifiability concerns would be of interest.

## Appendix C. Detailed comparison to related work

Below we describe in detail recent related work, and make comparisons between the assumptions made there and in our work. We focus on work with nonparametric identifiability guarantees, and make comparisons with our Theorem 7.

First, we highlight an important distinction between our work and previous work related to confounding. The problem of confounding, broadly construed, refers to the estimation of causal relationships *between observables* (i.e., measurement variables) when there are unobserved latent variables (i.e., unmeasured confounders). Crucially, identifying and estimating the latent variables themselves is not an issue. This is orthogonal to our setting, where causal relationships between observables are either meaningless or not the object of interest (e.g. pixels). By contrast, we are interested specifically in *latent factor recovery*, i.e., the identification and estimation of latent structure in causal problems. The latter is

the typical setting in factor analysis, or more broadly, causal representation learning. To summarize:

- *Confounding:* Learn edges $M_i \to M_j$, but not latents $L_i$
- *Factor analysis / representation learning:* Learn latents $L_i$, influences $L_i \to M_j$, but not edges $M_i \to M_j$

With this in mind, all existing results require various functional restrictions or parametric assumptions, and generally require auxiliary information, additivity, and/or no interaction effects. NCFA does not require these assumptions, and thus our framework applies more broadly, although it does not consider the most general forms of confounding, as in Maeda and Shimizu [60] and described above. The tradeoff is that NCFA guarantees model-free latent factor recovery, whereas papers focused on confounding impose model-specific functional assumptions to handle more general forms of confounding (i.e., to solve a different problem). We emphasize once again that NCFA does not even need assumptions on the type of data, as it can be continuous, discrete, or mixed.

- Maeda and Shimizu [60] (CAM-UV) study causal additive models with unobserved variables, with a focus on causal discovery under unmeasured confounding. In order to identify confounders, this paper assumes an additive noise model that is further additive over the parents, similar to CAM [76], along with faithfulness-type assumptions. As dscribed above, their focus on confounding is orthogonal to the present work: Maeda and Shimizu [60] study causal discovery *between* measurement variables where latent variables appear as nuisance confounders, whereas in latent factor models (i.e., here) the latent variables are the target of interest.
- Yang et al. [106] (CausalVAE) introduce CausalVAE for structured causal disentanglement. Built upon the framework of auxiliary variables introduced by Khemakhem et al. [47], this paper makes several assumptions that we do not require: 1) The latent SEM is linear, and 2) They observe additional signal in the form of auxiliary variables $u$ associated with the true causal concepts. The latter assumption is crucial for identification.
- Shen et al. [86] (DEAR) again use auxiliary variables (i.e., iVAE) as weak supervision for disentangled generative causal representation learning. In addition to extra auxiliary information in the form of annotated labels of the ground-truth factors, a super-structure of the latent DAG is required, as the authors "leave incorporating causal discovery from scratch to future work". Moreover, they make parametric assumptions including a linear latent SEM and a post-nonlinear model inspired by DAG-GNN [108].
- Ashman et al. [4] (Neural-ADMG / DDECI) learn latent causal models in the form of an acyclic directed mixed graph (ADMG). As with Yang et al. [106], they make several strong identifying assumptions that we do not require: 1) Additive noise models, 2) No interaction between parents and confounders, and 3) At least two pure children for each latent. In our case, only one pure child *suffices*, but is not necessary, for identification (see Theorem 7 and its special case Corollary 8).

- Kaltenpoth and Vreeken [46] (No-cadillac) perform nonlinear causal discovery with latent confounders, similar to Maeda and Shimizu [60]. As in Maeda and Shimizu [60], the focus is on confounding (i.e., relationships between measurement variables), which are restricted by a post-nonlinear model similar to DAG-GNN [108]. The latent variables are assumed to be i.i.d. Gaussian, and their results impose four additional nontrivial graphical assumptions. By contrast, NCFA requires two simple graphical constraints (expressed in Theorem 7), and imposes no parametric assumptions.
- Li et al. [54] (DeFuSE) also consider nonlinear causal discovery with confounders by assuming the latent structure is completely known (and in particular, $\dim(Z) = \dim(X)$) along with Gaussian noise and Gaussian confounders. They further assume an additive noise model and that each nonlinearity is of sublinear growth. None of these assumptions are required by our approach, and in particular we allow for any number of latents, independent of $\dim(X)$.

## Appendix D. Evaluation Metrics

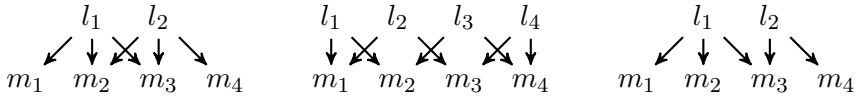
**Distance measures** Unlike classical VAE methods, NCFA has a causal discovery step in which it infers the UDG upon which the VAE in the model is based. The structural Hamming distance (SHD) is commonly used to quantify the difference between graph structures and is simply defined to be the number of edges that appear in one graph but not the other. SHD, however, is not applicable in our case as (i) it is defined for graphs having the same number of vertices, whereas for a fixed finite sample from a set of measurement variables, differently estimated minimum MCMs may have differing numbers of latent variables, and (ii) even if we compute, e.g., the SHD between the undirected graphs $\mathcal{U}, \mathcal{U}'$ (which *do* have the same number of vertices), this aligns poorly with intuitions about distance between their respective generating MCM graphs $\mathcal{G}, \mathcal{G}'$. To remedy this, we introduce the following distance:

**Definition 10** *The* shared factor distance *between biadjacency matrices* $B_1 \in \{0,1\}^{K_1,n}$ *and* $B_2 \in \{0,1\}^{K_2,n}$*, corresponding to minimum MCM graphs* $\mathcal{G}_1, \mathcal{G}_2$ *each with* $n = |M|$ *measurement variables and* $K_i$ *latents, respectively, for* $i = 1, 2$*, is defined as*

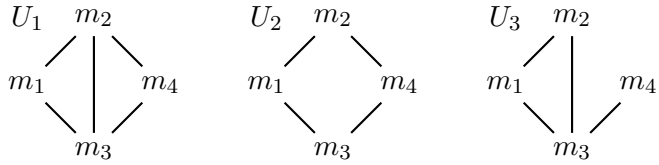
$$\mathrm{SFD}(B_1, B_2) \coloneqq \sum_{i \leq j} \left| (B_1^\top B_1)_{ij} - (B_2^\top B_2)_{ij} \right|.$$

The SFD is a proper distance metric [20], satisfying nonnegativity, symmetry, the triangle inequality, and (as detailed below) identity of indiscernibles. The intuition behind this definition is that we transform a given minimum MCM structure into a weighted undirected graph, allowing it to be easily compared to other minimum MCM structures over the same set of measurement variables. Via this transformation, the weights keep track of how many latent parents each pair of measurement variables have in common as the $(i, j)$-th entry in the matrix $B_1^\top B_1$ records the number of colliderless paths between measurement variables $i$ and $j$ in the graph. For our graphs, two nodes admit such a path between them if and only if they are connected by a latent, and this connection adds exactly 1 to the $(i, j)$-th matrix entry. Hence, these matrices will be equal if and only if the MeDIL models are identical (for a consistent ordering of measurement variables) [19, Section 2.2]. Figure D.1

provides an example comparing the SFD between minimum MCMs and the SHD between their corresponding (unweighted) undirected graphs.




$(a)$ $\mathcal{G}_1$ $(b)$ $\mathcal{G}_2$ $(c)$ $\mathcal{G}_3$




$(d)$ $(e)$ $(f)$

$$B_1 = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix} \quad (g) \qquad B_2 = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad (h) \qquad B_3 = \begin{bmatrix} 1 & 1 & 1 & 0 \\ 0 & 0 & 1 & 1 \end{bmatrix} \quad (i)$$

| | $\mathcal{G}_1$ and $\mathcal{G}_2$ | $\mathcal{G}_1$ and $\mathcal{G}_3$ | $\mathcal{G}_2$ and $\mathcal{G}_3$ |
|---|---|---|---|
| structural Hamming distance | 1 | 1 | 2 |
| shared factor distance | 10 | 5 | 7 |

$(j)$

Figure D.1: Pairwise comparison of structural Hamming distance and shared factor distance on three causal factor structures, $\mathcal{G}_1, \mathcal{G}_2$, and $\mathcal{G}_3$. For each structure $\mathcal{G}_i$, we also show its associated undirected graph $U_i$ and its biadjacency matrix $B_i$.