Journal of Machine Learning Research 27 (2026) 1-90 Submitted 5/24; Revised 2/26; Published 5/26

Nonparametric Partial Disentanglement via Mechanism Sparsity: Sparse Actions, Interventions and Sparse Temporal Dependencies

Sébastien Lachapelle S.LACHAPELLE@SAMSUNG.COM
Samsung AI Lab, Montreal; Mila & DIRO, Université de Montréal

Pau Rodríguez López
ServiceNow Research

Yash Sharma
Tübingen AI Center, University of Tübingen

Katie Everett
MIT CSAIL

Rémi Le Priol
Mila & DIRO, Université de Montréal

Alexandre Lacoste
ServiceNow Research

Simon Lacoste-Julien
Samsung AI Lab, Montreal; Mila & DIRO, Université de Montréal; Canada CIFAR AI Chair

Editor: Elias Bareinboim

Abstract

This work introduces a novel principle for disentanglement we call *mechanism sparsity regularization*, which applies when the latent factors of interest depend sparsely on observed auxiliary variables and/or past latent factors. We propose a representation learning method that induces disentanglement by *simultaneously* learning the latent factors and the sparse causal graphical model that explains them. We develop a nonparametric identifiability theory that formalizes this principle and shows that the latent factors can be recovered by regularizing the learned causal graph to be sparse, under some assumptions such as the absence of instantaneous causal effects between latent factors. More precisely, we show identifiability up to a novel equivalence relation we call *consistency*, which allows some latent factors to remain entangled (hence the term *partial* disentanglement). To describe the structure of this entanglement, we introduce the notions of *entanglement graphs* and *graph preserving functions*. We further provide a graphical criterion which guarantees *complete* disentanglement, that is identifiability up to permutations and element-wise transformations. We demonstrate the scope of the mechanism sparsity principle as well as the assumptions it relies on with several worked out examples. For instance, the framework shows how one can leverage multi-node interventions with unknown targets on the latent factors to disentangle them. We further draw connections between our nonparametric results and the now popular exponential family assumption. Lastly, we propose an estimation procedure based on variational autoencoders and a sparsity constraint and demonstrate it on various synthetic datasets. This work is meant to be a significantly extended version of Lachapelle et al. (2022).

Keywords: identifiable representation learning, causal representation learning, disentanglement, nonlinear independent component analysis, causal discovery

Contents

1 Introduction

It has been proposed that causal reasoning will be central in moving modern machine learning algorithms beyond their current shortcomings, such as their lack of *robustness*, *transferability*, and *interpretability* (Pearl, 2019; Schölkopf, 2019; Goyal and Bengio, 2021). To achieve this, the field of *causal representation learning* (CRL) (Schölkopf et al., 2021) aims to learn representations of high-dimensional observations, such as images, that are suitable to perform causal reasoning, such as predicting the effect of unseen interventions and answering counterfactual queries. A popular formalism to do so is to assume that the observations $\boldsymbol{x} \in \mathbb{R}^{d_z}$ are sampled from a generative model of the form $\boldsymbol{x} = \boldsymbol{f}(\boldsymbol{z})$ where $\boldsymbol{z} \in \mathbb{R}^{d_z}$ is a random vector of *unobserved* and *semantically meaningful* variables, also called latent factors, distributed according to an unknown *causal graphical model* (CGM) (Pearl, 2009; Peters et al., 2017) and transformed by a potentially highly nonlinear *decoder*, or *mixing function*, $\boldsymbol{f}$ (Kocaoglu et al., 2018; Volodin, 2021; Lachapelle et al., 2022; Lippe et al., 2023b; Brehmer et al., 2022; Ahuja et al., 2023; Buchholz et al., 2023; von Kügelgen et al., 2023; Zhang et al., 2023; Jiang and Aragam, 2023). The goal is then to recover the latent factors $\boldsymbol{z}_i$ up to permutation and rescaling, as well as the causal relationships that explain them. This is closely related to the problem of *disentanglement* (Bengio et al., 2013; Higgins et al., 2017; Locatello et al., 2020), which also aims to extract interpretable variables from high-dimensional observations, but without the emphasis on modeling their causal relations. Such problems are plagued by the difficult question of *identifiability*, which is of crucial importance in the classical settings of *causal discovery* (Pearl, 2009; Peters et al., 2017), where $\boldsymbol{f}$ is assumed to be the identity, and *independent component analysis* (ICA) (Hyvärinen et al., 2001, 2023), where the causal graph over latents is assumed empty. In the former, one can only identify the Markov equivalence class of the causal graph (assuming faithfulness), thus leaving some edge orientations ambiguous (Pearl, 2009), while in the latter, identifiability of the ground-truth latent factors is impossible when assuming a general nonlinear $\boldsymbol{f}$, (Hyvärinen and Pajunen, 1999). The general CRL problem inherits the difficulties from both of these settings, which makes identifiability especially challenging. Various strategies to improve identifiability have been contributed to the literature, such as assuming access to *interventional data* in which latent factors are targeted by interventions (Lachapelle et al., 2022; Lippe et al., 2022, 2023b; Ahuja et al., 2023), or access to an *auxiliary variable* $\boldsymbol{a}$ that makes factors $\boldsymbol{z}_i$ mutually independent when conditioned on (Hyvärinen et al., 2019; Khemakhem et al., 2020a,b). A valid auxiliary variable $\boldsymbol{a}$ must be observed and could correspond, for instance, to a time or environment index, an action in an interactive environment, or even a previous observation if the data have temporal structure. See Section 7 for a more extensive review of existing approaches to the identification of latent variables.

The present paper introduces[1] *mechanism sparsity regularization* as a new principle for the identification of latent variables. We show that if (i) an auxiliary variable $\boldsymbol{a}$ is observed and affects the latent variables *sparsely* and/or (ii) the latent variables $\boldsymbol{z}$ present *sparse* temporal dependencies, then the latent variables can be recovered by learning a graphical model for $\boldsymbol{z}$ and $\boldsymbol{a}$ and regularizing it to be sparse (Theorems 1, 2, 3 & 5). More specifically, we consider models of the form $\boldsymbol{x}^t = \boldsymbol{f}(\boldsymbol{z}^t) + \boldsymbol{n}^t$, where $\boldsymbol{n}^t$ is independent noise (Assumption 1) and the latent factors $\boldsymbol{z}_i^t$ are mutually independent given the past factors and the auxiliary variables, i.e. $p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = \prod_{i=1}^{d_z} p(\boldsymbol{z}_i^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ (Assumption 2). Crucially, we leverage the assumption that these mechanisms are sparse in the sense that $p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ factorizes according to a sparse

1. A shorter version of this work originally appeared in Lachapelle et al. (2022).

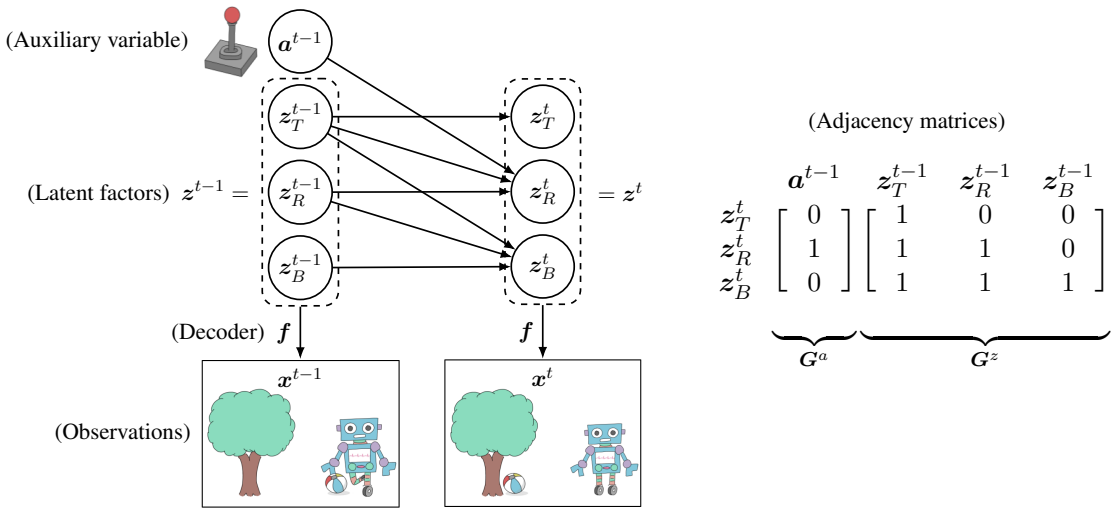


Figure 1: A minimal motivating example. The latent factors z_T^t, z_R^t and z_B^t represent the x-positions of the tree, the robot and the ball at time t, respectively. Only the image of the scene $\boldsymbol{x}^t$ and the action $\boldsymbol{a}^{t-1}$ are observed. See end of Section 2.1 for details.

causal graph $\boldsymbol{G}$ (Assumption 3). Interestingly, if $\boldsymbol{a}$ corresponds to an intervention index, our framework explains how interventions targeting unknown subsets of latent factors can identify them (Section 3.4.1). We emphasize that the settings where the data have no temporal dependencies or no auxiliary variable $\boldsymbol{a}$ are special cases of our framework. Our identifiability results are summarized in Table 1.

This work is meant to be an extended version of Lachapelle et al. (2022) in which we generalize along two main axes: First, we relax the *exponential family* assumption by providing a fully *nonparametric* treatment. Secondly, our results drop the graphical criterion of Lachapelle et al. (2022) and, thus, allow for *arbitrary* latent causal graphs. As a consequence of this relaxation, instead of guaranteeing identifiability up to permutation and element-wise transformation, we guarantee identifiability up to what we call $\boldsymbol{a}$-*consistency* or $\boldsymbol{z}$-*consistency* (Definitions 13 & 14), which might allow certain latent variables to remain entangled. Our results thus have the following flavor: Given a specific ground-truth causal graph $\boldsymbol{G}$ over $\boldsymbol{z}$ and $\boldsymbol{a}$, we describe precisely the structure of the entanglement between latent factors via what we call an *entanglement graph* (Definition 3) and *graph preserving functions* (Definition 12). See Figure 3 for examples. Interestingly, the stronger identifiability up to permutation and element-wise transformation arises as a simple consequence of our theory when the graphical criterion of Lachapelle et al. (2022) is assumed to hold. In addition to these two main axes of generalization, we provide extensive examples illustrating the scope of our framework, our assumptions, and the consequences of our results (see Table 2 for a list). When it comes to the learning algorithm, we replace the sparsity *penalty* by a sparsity *constraint*, which improves the learning dynamics and is more interpretable, which results in easier hyperparameter tuning.

The hypothesis that *high-level concepts are related by a sparse dependency graph* has been described and leveraged for out-of-distribution generalization by Bengio (2019) and Goyal et al. (2021b), which were early sources of inspiration for this work. To the best of our knowledge, the theory developed in the present work and in its preliminary version (Lachapelle et al., 2022) is the

	Parametric assumption	Continuous a	Discrete a (interventions)	Temporal dependencies	Sufficient influence	Identifiable up to	Examples
Thm. 1	None	Required	–	Optional	Ass. 6	Def. 13	3, 4, 5, 9, 10
Thm. 2	None	–	Required	Optional	Ass. 7	Def. 13	3, 4, 5, 11, 12, 13
Thm. 3	None	Optional	Optional	Required	Ass. 8	Def. 14	6, 7, 8, 14
Thm. 4	Exp. fam.	Optional	Optional	Optional	Ass. 10	Def. 17	15
Thm. 5	Exp. fam.	Optional	Optional	Required	Ass. 11	Def. 14, 17	16

Table 1: Summary of our identifiability results.

first to show formally that this inductive bias can sometimes be enough to recover ground-truth latent factors.

Figure 1 shows a minimal motivating example in which our approach could be used to extract the high-level variables (such as the x-position of the three objects) and learn their dynamics (how the objects move and affect each other) from a time series of images and agent actions, $(\boldsymbol{x}^t, \boldsymbol{a}^t)$. Theorems 1, 2, 3 & 5 show how the sparse dependencies between the objects and the action can be leveraged to identify the latent variables and the graph describing their dynamics. The learned CGM could be used subsequently to simulate interventions on semantic variables (Pearl, 2009; Peters et al., 2017), such as changing the torque of the robot or the weight of the ball. Moreover, disentanglement could be useful to interpret what caused the actions of an agent (Pearl, 2019). Following Lachapelle et al. (2022), empirical works demonstrated that disentangled representations with sparse mechanisms can adapt to unseen interventions faster in the context of single-cell biology (Lopez et al., 2023) and synthetic video data (Lei et al., 2023).

Summary of our contributions:

1. We introduce[1] a new principle for disentanglement based on *mechanism sparsity regularization* motivated by rigorous and novel *identifiability guarantees* (Theorems 1, 2, 3 & 5).
2. We extend Lachapelle et al. (2022) by providing a *nonparametric* treatment and allowing for *arbitrary latent graphs*, as long as it has no edges between latent factors from the same time step (no instantaneous causal links). Given a latent ground-truth graph, our theory predicts the structure of the entanglement between variables, which we formalize with *entanglement graphs* (Definition 3), *graph preserving maps* (Definition 12) and novel *equivalence relations* (Definitions 13 & 14).
3. We provide several examples to illustrate the generality of our results and to get a better understanding of their various assumptions and consequences (summarized in Table 2). For instance, we show how multi-node interventions with unknown-targets can yield disentanglement, both with and without temporal dependencies (Examples 12 & 13).
4. We introduce an evaluation metric denoted by R_{con} that quantifies how close two representations are to being $\boldsymbol{a}$-consistent or $\boldsymbol{z}$-consistent (Section 6).
5. We implement a learning approach based on variational autoencoders (VAEs) (Kingma and Welling, 2014) that learns the mixing function $\boldsymbol{f}$, the transition distribution $p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ and the causal graph $\boldsymbol{G}$. The latter is learned using binary masks and regularized for sparsity using a *constraint* as opposed to a penalty as in Lachapelle et al. (2022).
6. We perform experiments on synthetic datasets to validate the prediction of our theory.

Overview. Section 2 introduces the model (Section 2.1), entanglement maps and graphs (Section 2.2), the notion of identifiability (Section 2.3), equivalence up to diffeomorphism (Section 2.4)

Examples	Type of disentanglement	Auxiliary variable	Time dependencies
3	Complete	Yes (single target)	Optional
4	Partial	Yes (single target)	Optional
5	Complete	Yes (multi-target)	Optional
6	Complete	Optional	Yes (independent factors)
7	Complete	Optional	Yes (dependent factors)
8	Partial	Optional	Yes (dependent factors)
9	Partial	Yes (single-target continuous)	Yes
10	Complete	Yes (multi-target continuous)	No
11	Complete	Yes (single-target interventions)	No
12	Complete	Yes (multi-target interventions)	Yes
13	Complete	Yes (grouped multi-target interventions)	No
14	Complete	No	Yes (non-Markovian)
16	Complete	No	Yes (Markovian)

Table 2: List of examples illustrating the scope of our theory, its assumptions and its consequences.

and disentanglement formally (Section 2.5). Section 3 provides mathematical intuition for why mechanism sparsity yields disentanglement (Section 3.2); introduces the machinery of graph preserving maps (Section 3.3) which are key to establish identifiability up to $\boldsymbol{a}$-consistency (Section 3.4) and $\boldsymbol{z}$-consistency (Section 3.5), i.e. *partial* disentanglement. Section 3 also discusses the relationship with interventions (Section 3.4.1), provides a graphical criterion guaranteeing *complete* disentanglement (Section 3.7), and introduces and discusses extensively the *sufficient influence assumptions* on which these results are critically based (Sections 3.8 & 3.9). Section 4 draws connections between our *nonparametric* theory and the *exponential family* assumption sometimes used in the literature. Section 5 presents the VAE-based learning algorithm with sparsity constraint. Section 6 introduces our novel R_{con} metric. Section 7 reviews the literature on identifiability in representation learning. Section 8 presents the empirical results.

Notation. Scalars are denoted in lower-case and vectors in lower-case bold, e.g. $x \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^n$. Note that these will sometimes denote random variables depending on the context. We maintain an analogous notation for scalar-valued and vector-valued functions, e.g., f and $\boldsymbol{f}$. The ith coordinate of the vector $\boldsymbol{x}$ is denoted by $\boldsymbol{x}_i$. The set containing the first n integers excluding 0 is denoted by $[n]$. Given a subset of indices $S \subseteq [n]$, $\boldsymbol{x}_S$ denotes the subvector consisting of entries $\boldsymbol{x}_i$ for $i \in S$. Given a sequence of T random vectors $(\boldsymbol{x}^1, \dots, \boldsymbol{x}^T)$, the subsequence consisting of the first t elements is denoted by $\boldsymbol{x}^{\leq t} := (\boldsymbol{x}^1, ..., \boldsymbol{x}^t)$, and analogously for $\boldsymbol{x}^{<t}$. We will sometimes combine these notation to get $\boldsymbol{x}_S^{\leq t} := (\boldsymbol{x}_S^1, \dots, \boldsymbol{x}_S^t)$. Given a function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$, its Jacobian matrix evaluated at $\boldsymbol{x} \in \mathbb{R}^n$ is denoted by $D\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^{m \times n}$. See Table 5 in the Appendix for more.

2 Problem Setting, Entanglement Graphs & Disentanglement

In this section, we introduce the latent variable model under consideration (Section 2.1), entanglement graphs (Section 2.2), identifiability and observational equivalence (Section 2.3), equivalence up to diffeomorphism (Section 2.4) as well as permutation equivalence (Section 2.5).

2.1 An Identifiable Latent Causal Model

We now specify the setting under consideration. Assume that we observe the realization of a sequence of d_x-dimensional random vectors $\{\boldsymbol{x}^t\}_{t=1}^T$ and a sequence of d_a-dimensional auxiliary

vectors $\{\boldsymbol{a}^t\}_{t=0}^{T-1}$. The coordinates of $\boldsymbol{a}^t$ are discrete or continuous and can potentially represent, for example, an action taken by an agent or the index of an intervention or environment (see Section 3.4.1). Observations $\{\boldsymbol{x}^t\}$ are assumed to be explained by a sequence of continuous random vectors d_z of hidden dimensions $\{\boldsymbol{z}^t\}_{t=1}^{T}$ using a ground-truth decoder function $\boldsymbol{f}$.

Assumption 1 (Observation model) *For all $t \in [T]$, the observations $\boldsymbol{x}^t$ are given by*

$$\boldsymbol{x}^t = \boldsymbol{f}(\boldsymbol{z}^t) + \boldsymbol{n}^t \,,$$

where $\boldsymbol{n}^t \sim \mathcal{N}(0, \sigma^2 \boldsymbol{I})$ are mutually independent across time and independent of all $\boldsymbol{z}^t$ and $\boldsymbol{a}^t$ with $\sigma^2 \geq 0$. Moreover, $d_z \leq d_x$ and $\boldsymbol{f} : \mathbb{R}^{d_z} \to \mathbb{R}^{d_x}$ is a diffeomorphism onto its image[2]. Lastly, assume that $\boldsymbol{f}(\mathbb{R}^{d_z})$ is closed in $\mathbb{R}^{d_x}$.

Importantly, we suppose that each factor $\boldsymbol{z}_i^t$ contains interpretable information about the observation, e.g., for high-dimensional images, the coordinates $\boldsymbol{z}_i^t$ might be the position of an object, its color, or its orientation in space. This idea that there exists a *ground-truth decoder* $\boldsymbol{f}$ that captures the relationship between the so-called "natural factors of variations" and the observations $\boldsymbol{x}$ is of capital importance, since it is the very basis for a mathematical definition of disentanglement (Definition 7). Appendix D.1 discusses the implications of the diffeomorphism assumption (see also Mansouri et al. (2022)). We denote $\boldsymbol{z}^{\leq t} := [\boldsymbol{z}^1 \ \cdots \ \boldsymbol{z}^t] \in \mathbb{R}^{d_z \times t}$ and analogously for $\boldsymbol{z}^{<t}$ and other random vectors.

In a similar spirit to previous works on nonlinear ICA (Hyvärinen et al., 2019; Khemakhem et al., 2020a), we assume that the latent factors $\boldsymbol{z}_i^t$ are conditionally independent given the past.

Assumption 2 (Conditionally independent latent factors) *The latent factors $\boldsymbol{z}_i^t$ are conditionally mutually independent given $\boldsymbol{z}^{<t}$ and $\boldsymbol{a}^{<t}$:*

$$p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = \prod_{i=1}^{d_z} p(\boldsymbol{z}_i^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) \,, \tag{1}$$

where $p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ is a density function w.r.t. the Lebesgue measure on $\mathbb{R}^{d_z}$. We assume that the support of $p(\boldsymbol{z}_i^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ is $\mathbb{R}$ for all $\boldsymbol{z}^{<t}$ and $\boldsymbol{a}^{<t}$. The support of $p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ is thus given by $\mathbb{R}^{d_z}$. At $t = 1$, we set $p(\boldsymbol{z}^1 \mid \boldsymbol{z}^{<1}, \boldsymbol{a}^{<1}) = p(\boldsymbol{z}^1 \mid \boldsymbol{a}^0) = \prod_{i=1}^{d_z} p(\boldsymbol{z}_i^1 \mid \boldsymbol{a}^0)$.

We will refer to the l.h.s. of (1) as the *transition model* and to the factors $p(\boldsymbol{z}_i^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ as *mechanisms*. In a causal setting, this conditional independence assumption states that latent variables cannot exert influence on other latent variables instantaneously. We believe this to be a reasonable assumption when the time steps are sampled at a sufficiently high rate since, at a fundamental level, causal effects cannot travel faster than the speed of light. Of course in many practical scenarios, a low sampling rate can entail instantaneous dependencies which, if too strong, might

2. A *diffeomorphism* is a C^1 bijection with a C^1 inverse. Generally, given a map $\boldsymbol{h} : A \to \mathbb{R}^m$ where $A \subseteq \mathbb{R}^n$, saying $\boldsymbol{h}$ is C^k is typically only well defined if A is an open set of $\mathbb{R}^n$. Throughout, if $A \subseteq \mathbb{R}^n$ is arbitrary (not necessarily open), we say $\boldsymbol{h}$ is C^k if there exists a C^k map $\tilde{\boldsymbol{h}} : U \to \mathbb{R}^m$ defined on an open set U of $\mathbb{R}^n$ containing A such that $\boldsymbol{h} = \tilde{\boldsymbol{h}}$ on A. Note that it is then meaningful for $\boldsymbol{f}^{-1} : \boldsymbol{f}(\mathbb{R}^{d_z}) \to \mathbb{R}^{d_z}$ to be C^1 even when $\boldsymbol{f}(\mathbb{R}^{d_z})$ is not open in $\mathbb{R}^{d_x}$. Moreover, it can be shown that $\boldsymbol{f} : \mathbb{R}^{d_z} \to \mathbb{R}^{d_x}$ is a diffeomorphism onto its image if $\boldsymbol{f}$ is an homeomorphism onto its image, i.e. continuous in both directions, and has a full rank Jacobian everywhere on its domain (Munkres, 1991, Sec. 23 & Thm. 24.1).

result in model misspecification. In Section 7, we discuss and contrast with causal representation learning methods which allow for instantaneous influence.

Notice that we do not assume the dynamical system to be *Markovian*[3], i.e. the distribution over future states can depend on the whole history of latents and auxiliary variables $(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$. In addition, this model can represent *non-homogeneous* processes by taking the auxiliary variable $\boldsymbol{a}$ to be a time index (Hyvärinen et al., 2019).

We are going to describe the dependency structure of the latent and auxiliary variables over time using a *probabilistic directed graphical model* composed of two bipartite graphs, $\boldsymbol{G}^z \in \{0,1\}^{d_z \times d_z}$, which relates $\boldsymbol{z}^{<t}$ to $\boldsymbol{z}^t$, and $\boldsymbol{G}^a \in \{0,1\}^{d_z \times d_a}$, which relates $\boldsymbol{a}^{<t}$ to $\boldsymbol{z}^t$. A directed edge points from $\boldsymbol{z}_j^{<t}$ to $\boldsymbol{z}_i^t$ if and only if $\boldsymbol{G}^z_{i,j} = 1$. Analogously, a directed edge points from $\boldsymbol{a}_\ell^{<t}$ to $\boldsymbol{z}_i^t$, if and only if $\boldsymbol{G}^a_{i,\ell} = 1$. Figure 1 shows an example of such graphs together with its adjacency matrix $\boldsymbol{G} := [\boldsymbol{G}^z, \boldsymbol{G}^a]$. The following assumption specifies the relationship between these graphs and the transition model.

Assumption 3 (Transition model p follows $\boldsymbol{G}$) *For every mechanism $i \in [d_z]$,*

$$p(\boldsymbol{z}_i^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = p(\boldsymbol{z}_i^t \mid \boldsymbol{z}^{<t}_{\mathbf{Pa}_i^z}, \boldsymbol{a}^{<t}_{\mathbf{Pa}_i^a}) \,, \tag{2}$$

where $\mathbf{Pa}_i^z \subseteq [d_z]$ and $\mathbf{Pa}_i^a \subseteq [d_a]$ are the sets of parents of $\boldsymbol{z}_i^t$ in $\boldsymbol{G}^z$ and $\boldsymbol{G}^a$, respectively.

The graph $\boldsymbol{G}$ thus encodes a set of conditional independence statements about the latent and auxiliary variables.[4] We will say that *mechanisms are sparse* when the graphs $\boldsymbol{G}^a$ and $\boldsymbol{G}^z$ are sparse.

This model has three components that need to be learned: (i) the decoder function $\boldsymbol{f}$, (ii) the transition model over latent variables p, and (iii) the dependency graph $\boldsymbol{G}$. We gather all these components into $\boldsymbol{\theta} := (\boldsymbol{f}, p, \boldsymbol{G})$. Everything else in the model, i.e. d_z and σ^2, is assumed to be known. We assume that σ^2 is known here mainly for simplicity, since, when it is not, it can be identified as shown by Lachapelle et al. (2022, Appendix A.4.1), as long as $d_x > d_z$.

Notice how we have not specified any model for the auxiliary variable $\boldsymbol{a}^t$. We do not intend to do so in this work, as we are solely interested in modeling the *conditional* distribution of $\boldsymbol{x}^{\leq T}$ and $\boldsymbol{z}^{\leq T}$ given $\boldsymbol{a}^{<T}$. We denote by $\mathcal{A} \subseteq \mathbb{R}^{d_a}$ the set of possible values for the auxiliary variable $\boldsymbol{a}^t$. We thus have that, for all values of $\boldsymbol{a}^{<T} \in \mathcal{A}^T$, our model induces a conditional distribution

$$p(\boldsymbol{x}^{\leq T} \mid \boldsymbol{a}^{<T}) = \int \prod_{t=1}^{T} p(\boldsymbol{x}^t \mid \boldsymbol{z}^t) p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) d\boldsymbol{z}^{\leq T} \,,$$

where $p(\boldsymbol{x}^t \mid \boldsymbol{z}^t) = \mathcal{N}(\boldsymbol{x}^t; \boldsymbol{f}(\boldsymbol{z}^t), \sigma^2 \boldsymbol{I})$. We note that if $\sigma^2 = 0$, the conditional distribution of $\boldsymbol{x}^t$ given $\boldsymbol{z}^t$ is a Dirac centered at $\boldsymbol{f}(\boldsymbol{z}^t)$ and thus has no density w.r.t. the Lebesgue measure. Even if, in that case, the above integral does not make sense, the conditional distribution of $\boldsymbol{x}^{\leq T}$ given $\boldsymbol{a}^{<T}$ is still well-defined and all the results of this work still hold since none of the proofs requires $\sigma^2 > 0$.

3. We use the term "Markovian" in the stochastic processes sense, as in Markov chains, where the future is independent of the past given the present. We will use this meaning throughout. This is to be contrasted with the Markovian property in structural models which states that exogenous noises are mutually independent (Bareinboim et al., 2022).
4. This notion of graph is not standard since, typically, directed edges point from scalar variables to scalar variables, whereas here a directed edge points from $\boldsymbol{z}_j^{<t}$ (a vector) to $\boldsymbol{z}_i^t$ (a scalar). This can be understood as having all directed edges $\boldsymbol{z}_j^\tau \to \boldsymbol{z}_i^t$ for all $\tau < t$. This does not imply that all edges are necessarily "active". We made this choice to simplify the presentation and hypothesize that analogous results could be derived for standard graphs.

A motivating example. Figure 1 represents a minimal example in which our theory applies. The environment consists of three objects: a tree, a robot and a ball with x-positions $\boldsymbol{z}_T^t$, $\boldsymbol{z}_R^t$ and $\boldsymbol{z}_B^t$, respectively. Together, they form the vector $\boldsymbol{z}^t$ of high-level latent variables, i.e., $\boldsymbol{z}^t = (\boldsymbol{z}_T^t, \boldsymbol{z}_R^t, \boldsymbol{z}_B^t)$. A remote controls the direction in which the wheels of the robot turn. The vector $\boldsymbol{a}^t$ records these actions, which might be taken by a human or an artificial agent trained to accomplish some goal. The only observations are the actions $\boldsymbol{a}^t$ and the images $\boldsymbol{x}^t$ representing the scene which is given by $\boldsymbol{x}^t = \boldsymbol{f}(\boldsymbol{z}^t) + \boldsymbol{n}^t$. The dynamics of the environment is governed by the transition model p, which, e.g., could be given by a Gaussian model of the form $p(\boldsymbol{z}_i^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = \mathcal{N}(\boldsymbol{z}_i^t; \mu_i(\boldsymbol{z}^{t-1}, \boldsymbol{a}^{t-1}), \sigma_z^2)$. Plausible connectivity graphs $\boldsymbol{G}^z$ and $\boldsymbol{G}^a$ are shown in Figure 1 showing how the latent factors are related and how the controller affects them. For every object, its position at the time step t depends on its position at $t-1$. The position of the tree, $\boldsymbol{z}_T^t$, is not affected by anything, since neither the robot nor the ball can change its position. The robot, $\boldsymbol{z}_R^t$, changes its position based both on the action $\boldsymbol{a}^{t-1}$ and the position of the tree $\boldsymbol{z}_T^{t-1}$ (in case of collision). The position of the ball, $\boldsymbol{z}_B^t$, is affected by both the robot, which can move around it by running into it, and the tree, which can bounce. The key observations here are that (i) the different objects interact *sparsely* with one another and (ii) the action $\boldsymbol{a}^t$ affects very few objects (in this case, only one). The theorems of Section 3 show how one can leverage this sparsity for disentanglement.

2.2 Entanglement Maps & Entanglement Graphs

In this section, we define *entanglement maps*, which describe the functional relationship between the learned and ground-truth representations, and *entanglement graphs*, which describe their entanglement structure.

Definition 1 (Entanglement maps) *Let $\boldsymbol{f}$ and $\tilde{\boldsymbol{f}}$ be two diffeomorphisms from $\mathbb{R}^{d_z}$ to their images such that $\boldsymbol{f}(\mathbb{R}^{d_z}) = \tilde{\boldsymbol{f}}(\mathbb{R}^{d_z})$. The **entanglement map** of the pair $(\boldsymbol{f}, \tilde{\boldsymbol{f}})$ is given by*

$$\boldsymbol{v} := \boldsymbol{f}^{-1} \circ \tilde{\boldsymbol{f}}\,. \tag{3}$$

This map will be crucial throughout this work, especially in defining disentanglement. Intuitively, the entanglement map for a pair of decoders $(\boldsymbol{f}, \tilde{\boldsymbol{f}})$ translates the representation of one model to that of the other. In general, the entanglement maps of $(\boldsymbol{f}, \tilde{\boldsymbol{f}})$ and $(\tilde{\boldsymbol{f}}, \boldsymbol{f})$ are different.

We now define the *dependency graph* of some function $\boldsymbol{h}$ so that each edge indicates that some input i influences some output j:

Definition 2 (Functional dependency graph) *Let $\boldsymbol{h}$ be a function from $\mathbb{R}^n$ to $\mathbb{R}^m$. The **dependency graph** of $\boldsymbol{h}$ is a bipartite directed graph from $[n]$ to $[m]$ with adjacency matrix $\boldsymbol{H} \in \{0,1\}^{m \times n}$ such that*

$$\boldsymbol{H}_{i,j} = 0 \iff \text{There is a function } \bar{\boldsymbol{h}} \text{ such that, for all } \boldsymbol{a} \in \mathbb{R}^n,\ \boldsymbol{h}_i(\boldsymbol{a}) = \bar{\boldsymbol{h}}_i(\boldsymbol{a}_{-j})\,,$$

where $\boldsymbol{a}_{-j}$ is $\boldsymbol{a}$ with its jth coordinate removed.

Example 1 (Dependency graph of a linear map) *Let $\boldsymbol{h}(\boldsymbol{z}) := \boldsymbol{W}\boldsymbol{z}$ where $\boldsymbol{W} \in \mathbb{R}^{m \times n}$ and let $\boldsymbol{H}$ be the dependency graph of $\boldsymbol{h}$. Then $\boldsymbol{H}_{i,j} = 0 \iff \boldsymbol{W}_{i,j} = 0$.*

We will be particularly interested in the dependency graph of the entanglement map $\boldsymbol{v} := \boldsymbol{f}^{-1} \circ \hat{\boldsymbol{f}}$, denoted by $\boldsymbol{V}$.

Definition 3 (Entanglement graphs) *Let $\boldsymbol{f}$ and $\tilde{\boldsymbol{f}}$ be two diffeomorphisms from $\mathbb{R}^{d_z}$ to their images such that $\boldsymbol{f}(\mathbb{R}^{d_z}) = \tilde{\boldsymbol{f}}(\mathbb{R}^{d_z})$. The **entanglement graph** of the pair $(\boldsymbol{f}, \tilde{\boldsymbol{f}})$ is the dependency graph (Definition 2) of their entanglement map $\boldsymbol{v} := \boldsymbol{f}^{-1} \circ \tilde{\boldsymbol{f}}$, which we denote $\boldsymbol{V} \in \{0,1\}^{d_z \times d_z}$.*

We now relate the dependency graph of a function to the zeros of its Jacobian matrix. A proof can be found in Appendix A.2.

Proposition 1 (Linking dependency graph and Jacobian) *Let $\boldsymbol{h}$ be a C^1 function, i.e. continuously differentiable, from $\mathbb{R}^n$ to $\mathbb{R}^m$ and let $\boldsymbol{H}$ be its dependency graph (Definition 2). Then,*

$$\boldsymbol{H}_{i,j} = 0 \iff \text{For all } \boldsymbol{a} \in \mathbb{R}^n \,, D\boldsymbol{h}(\boldsymbol{a})_{i,j} = 0 \,. \tag{4}$$

The equivalence (4) can be seen as an equivalent definition of dependency graph for differentiable functions.

2.3 Identifiability and Observational Equivalence

To analyze formally whether a specific algorithm is expected to yield a disentangled representation, we will rely on the notion of *identifiability*. Before we define what we mean by identifiability, we need the notion of *observationally equivalent* models. Two models are observationally equivalent if both models represent the same distribution over observations. The following formalizes this definition.

Definition 4 (Observational equivalence) *We say that two models $\boldsymbol{\theta} := (\boldsymbol{f}, p, \boldsymbol{G})$ and $\tilde{\boldsymbol{\theta}} := (\tilde{\boldsymbol{f}}, \tilde{p}, \tilde{\boldsymbol{G}})$ satisfying Assumption 1 are **observationally equivalent**, denoted $\boldsymbol{\theta} \sim_{\text{obs}} \tilde{\boldsymbol{\theta}}$, if and only if, for all $\boldsymbol{a}^{<T} \in \mathcal{A}^T$ and all $\boldsymbol{x}^{\leq T} \in \mathbb{R}^{d_x \times T}$,*

$$p(\boldsymbol{x}^{\leq T} \mid \boldsymbol{a}^{<T}) = \tilde{p}(\boldsymbol{x}^{\leq T} \mid \boldsymbol{a}^{<T}) \,.$$

Formally, we say that a parameter $\boldsymbol{\theta}$ is **identifiable up to some equivalence relation** $\sim$, when

$$\boldsymbol{\theta} \sim_{\text{obs}} \tilde{\boldsymbol{\theta}} \implies \boldsymbol{\theta} \sim \tilde{\boldsymbol{\theta}} \,. \tag{5}$$

This work is mainly concerned with proving statements of the above form by making assumptions both on $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$. The stronger the assumptions on $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ are, the stronger the equivalence relation $\sim$ will be. The following sections present two equivalence relations over models, namely $\sim_{\text{diff}}$ and $\sim_{\text{perm}}$. We note that the equivalence relation $\sim_{\text{perm}}$ will help us formalize disentanglement.

Practically speaking, observational equivalence between the learned model $\hat{\boldsymbol{\theta}}$ and the ground-truth model $\boldsymbol{\theta}$ can be achieved via maximum likelihood estimation in the infinite data regime. Thus, identifiability results of the form of (5) guaranty that if the learned model is perfectly fitted to the data (assumed infinite), its parameter $\hat{\boldsymbol{\theta}}$ is $\sim$-equivalent to that of the ground-truth model, $\boldsymbol{\theta}$.

2.4 Equivalence up to Diffeomorphism

We start by defining *equivalence up to diffeomorphism*. This equivalence relation is important since we will show later on that it is actually the same as observational equivalence and will thus be our first step in all our identifiability results. In what follows, we overload the notation and write $\boldsymbol{v}(\boldsymbol{z}^{<t}) := [\boldsymbol{v}(\boldsymbol{z}^1), \ldots, \boldsymbol{v}(\boldsymbol{z}^{t-1})]$, and similarly for other functions.

Definition 5 (Equivalence up to diffeomorphism) *We say that two models $\boldsymbol{\theta} := (\boldsymbol{f}, p, \boldsymbol{G})$ and $\tilde{\boldsymbol{\theta}} := (\tilde{\boldsymbol{f}}, \tilde{p}, \tilde{\boldsymbol{G}})$ satisfying Assumption 1 are **equivalent up to diffeomorphism**, denoted $\boldsymbol{\theta} \sim_{\text{diff}} \tilde{\boldsymbol{\theta}}$, if and only if $\boldsymbol{f}(\mathbb{R}^{d_z}) = \tilde{\boldsymbol{f}}(\mathbb{R}^{d_z})$ and, for all $t \in [T]$, all $\boldsymbol{a}^{<t} \in \mathcal{A}^t$ and all $\boldsymbol{z}^{\leq t} \in \mathbb{R}^{d_z \times t}$,*

$$\tilde{p}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = p(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t}) |\det D\boldsymbol{v}(\boldsymbol{z}^t)|\,, \tag{6}$$

where $\boldsymbol{v} := \boldsymbol{f}^{-1} \circ \tilde{\boldsymbol{f}}$ (entanglement map) is a diffeomorphism and $D\boldsymbol{v}$ denotes its Jacobian matrix.

The fact that the relation $\sim_{\text{diff}}$ is indeed an *equivalence* comes from the fact that the set of diffeomorphisms from a set to itself forms a group under composition.

To better understand the above definition, let $\boldsymbol{z}^t := \boldsymbol{g}(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}; \epsilon^t)$ and $\tilde{\boldsymbol{z}}^t := \tilde{\boldsymbol{g}}(\tilde{\boldsymbol{z}}^{<t}, \boldsymbol{a}^{<t}; \tilde{\epsilon}^t)$ where ϵ^t and $\tilde{\epsilon}^t$ are noise variables and $\boldsymbol{g}$ and $\tilde{\boldsymbol{g}}$ are functions such that random variables $\boldsymbol{z}^t$ and $\tilde{\boldsymbol{z}}^t$ have conditional densities given by $p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ and $\tilde{p}(\tilde{\boldsymbol{z}}^t \mid \tilde{\boldsymbol{z}}^{<t}, \boldsymbol{a}^{<t})$, respectively. Using the change-of-variable formula for densities, one can rewrite (6) as

$$\tilde{\boldsymbol{g}}(\tilde{\boldsymbol{z}}^{<t}, \boldsymbol{a}^{<t}; \tilde{\epsilon}^t) \overset{d}{=} \boldsymbol{v}^{-1} \circ \boldsymbol{g}(\boldsymbol{v}(\tilde{\boldsymbol{z}}^{<t}), \boldsymbol{a}^{<t}; \epsilon^t)\,, \tag{7}$$

where "$\overset{d}{=}$" denotes equality in distribution. This equation has a nice interpretation: applying the latent transition model $\tilde{\boldsymbol{\theta}}$ to go from $(\tilde{\boldsymbol{z}}^{<t}, \boldsymbol{a}^{<t})$ to $\tilde{\boldsymbol{z}}^t$ is the same as first applying $\boldsymbol{v}$, then applying the latent transition model $\boldsymbol{\theta}$ and finally applying $\boldsymbol{v}^{-1}$. Equation (7) is reminiscent of Ahuja et al. (2022a), in which the mechanism $\tilde{\boldsymbol{g}}$ would be called an *imitator* of $\boldsymbol{g}$. Ahuja et al. (2022a) showed that $\sim_{\text{obs}}$ and $\sim_{\text{diff}}$ are actually one and the same. For completeness, we present an analogous argument here. We start by showing that $\boldsymbol{\theta} \sim_{\text{diff}} \tilde{\boldsymbol{\theta}}$ implies $\boldsymbol{\theta} \sim_{\text{obs}} \tilde{\boldsymbol{\theta}}$.

$$\begin{aligned}
p(\boldsymbol{x}^{\leq T} \mid \boldsymbol{a}^{<T}) &= \int \prod_{t=1}^{T} \left[p(\boldsymbol{x}^t \mid \boldsymbol{z}^t) p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})\right] d\boldsymbol{z}^{\leq T} \\
&= \int \prod_{t=1}^{T} \left[p(\boldsymbol{x}^t \mid \boldsymbol{v}(\boldsymbol{z}^t)) p(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t})\right] |\det D\boldsymbol{v}(\boldsymbol{z}^{\leq T})| d\boldsymbol{z}^{\leq T} \\
&= \int \prod_{t=1}^{T} \left[p(\boldsymbol{x}^t \mid \boldsymbol{v}(\boldsymbol{z}^t)) p(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t}) |\det D\boldsymbol{v}(\boldsymbol{z}^t)|\right] d\boldsymbol{z}^{\leq T} \\
&= \int \prod_{t=1}^{T} \left[\tilde{p}(\boldsymbol{x}^t \mid \boldsymbol{z}^t) \tilde{p}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})\right] d\boldsymbol{z}^{\leq T} = \tilde{p}(\boldsymbol{x}^{\leq T} \mid \boldsymbol{a}^{<T})\,,
\end{aligned}$$

where the second equality used the change-of-variable formula, the third equality used the fact that the Jacobian of $\boldsymbol{v}(\boldsymbol{z}^{\leq T})$ is block-diagonal (each block corresponds to a time step t) and the next to last equality used the definition of $\sim_{\text{diff}}$ and the fact that

$$p(\boldsymbol{x}^t \mid \boldsymbol{v}(\boldsymbol{z}^t)) = \mathcal{N}(\boldsymbol{x}^t; \boldsymbol{f}(\boldsymbol{f}^{-1} \circ \tilde{\boldsymbol{f}}(\boldsymbol{z}^t)), \sigma^2 \boldsymbol{I}) = \mathcal{N}(\boldsymbol{x}^t; \tilde{\boldsymbol{f}}(\boldsymbol{z}^t), \sigma^2 \boldsymbol{I}) = \tilde{p}(\boldsymbol{x}^t \mid \boldsymbol{z}^t)\,.$$

The following proposition establishes the converse, i.e. that $\boldsymbol{\theta} \sim_{\text{obs}} \tilde{\boldsymbol{\theta}}$ implies $\boldsymbol{\theta} \sim_{\text{diff}} \tilde{\boldsymbol{\theta}}$. Since its proof is more involved, we present it in the Appendix A.3. Note that this first identifiability result is relatively weak and should be seen as a first step towards stronger guarantees. A very similar result was shown by Ahuja et al. (2022a, Theorem 3.1) to highlight the fact that the representation $\boldsymbol{f}$ is identifiable up to the equivariances $\boldsymbol{v}$ of the transition model p.

Proposition 2 (Identifiability up to diffeomorphism) *Let $\boldsymbol{\theta} := (\boldsymbol{f}, p, \boldsymbol{G})$ and $\hat{\boldsymbol{\theta}} := (\hat{\boldsymbol{f}}, \hat{p}, \hat{\boldsymbol{G}})$ be two models satisfying Assumption 1. If $\boldsymbol{\theta} \sim_{\text{obs}} \hat{\boldsymbol{\theta}}$ (Def. 4), then $\boldsymbol{\theta} \sim_{\text{diff}} \hat{\boldsymbol{\theta}}$ (Def. 5).*

Intuitively, Proposition 2 shows that if two models agree on the distribution of the observations, then their "data manifold" $\boldsymbol{f}(\mathbb{R}^{d_z})$ and $\hat{\boldsymbol{f}}(\mathbb{R}^{d_z})$ are equal and their respective transition models are related via $\boldsymbol{v} := \boldsymbol{f}^{-1} \circ \hat{\boldsymbol{f}}$.

2.5 Disentanglement and Equivalence up to Permutation

A disentangled representation is often defined intuitively as a representation in which the coordinates are in one-to-one correspondence with *natural factors of variation* in the data. We are going to assume that these natural factors are captured by an unknown ground-truth decoder $\boldsymbol{f}$. Given a learned decoder $\hat{\boldsymbol{f}}$ such that $\boldsymbol{f}(\mathbb{R}^{d_z}) = \hat{\boldsymbol{f}}(\mathbb{R}^{d_z})$, the entanglement map $\boldsymbol{v} := \boldsymbol{f}^{-1} \circ \hat{\boldsymbol{f}}$ gives a correspondence between the learned representation $\hat{\boldsymbol{f}}$ and the natural factors of variations of $\boldsymbol{f}$. The following equivalence relation will help us define disentanglement.

Definition 6 (Equivalence up to permutation) *We say that two models $\boldsymbol{\theta} := (\boldsymbol{f}, p, \boldsymbol{G})$ and $\tilde{\boldsymbol{\theta}} := (\tilde{\boldsymbol{f}}, \tilde{p}, \tilde{\boldsymbol{G}})$ satisfying Assumptions 1, 2 & 3 are **equivalent up to permutation**, denoted $\boldsymbol{\theta} \sim_{\text{perm}} \tilde{\boldsymbol{\theta}}$, if and only if there exists a permutation matrix $\boldsymbol{P}$ such that*

1. *$\boldsymbol{\theta} \sim_{\text{diff}} \tilde{\boldsymbol{\theta}}$ (Def. 5) and $\tilde{\boldsymbol{G}}^a = \boldsymbol{P}\boldsymbol{G}^a$ and $\tilde{\boldsymbol{G}}^z = \boldsymbol{P}\boldsymbol{G}^z\boldsymbol{P}^\top$; and*
2. *The entanglement map $\boldsymbol{v} := \boldsymbol{f}^{-1} \circ \tilde{\boldsymbol{f}}$ can be written as $\boldsymbol{v} = \boldsymbol{d} \circ \boldsymbol{P}^\top$, where $\boldsymbol{d}$ is element-wise, i.e. $\boldsymbol{d}_i(\boldsymbol{z})$ depends only on $\boldsymbol{z}_i$, for all i. In other words, the entanglement graph is $\boldsymbol{V} = \boldsymbol{P}^\top$.*

The fact that the relation $\sim_{\text{perm}}$ is an equivalence relation is actually a special case of a more general result that we present later in Section 3.4.

This allows us to give a formal definition of (complete) disentanglement. Note that we use the term *complete* to contrast with *partial* disentanglement.

Definition 7 (Complete disentanglement) *Given a ground-truth model $\boldsymbol{\theta}$, we say that a learned model $\hat{\boldsymbol{\theta}}$ is **completely disentangled** when $\boldsymbol{\theta} \sim_{\text{perm}} \hat{\boldsymbol{\theta}}$.*

Intuitively, a learned representation is *completely disentangled* when there is a one-to-one correspondence between its coordinates and those of the ground-truth representation (see Figure 2).

We define *partial* disentanglement as something that lives strictly between equivalence up to diffeomorphism and equivalence up to permutation:

Definition 8 (Partial disentanglement) *Given a ground-truth model $\boldsymbol{\theta}$, we say that a learned model $\hat{\boldsymbol{\theta}}$ is **partially disentangled** when $\boldsymbol{\theta} \sim_{\text{diff}} \hat{\boldsymbol{\theta}}$ with an entanglement graph $\boldsymbol{V}$ (Definition 3) that is neither a permutation nor the complete graph.*

This definition of partial disentanglement ranges from models that are almost completely entangled, i.e. those with very dense entanglement graphs $\boldsymbol{V}$, to ones that are very close to being completely disentangled, i.e. those with a very sparse $\boldsymbol{V}$. The following section will make more precise how one can learn a completely or partially disentangled representation from data and exactly what form the entanglement graph is going to take.

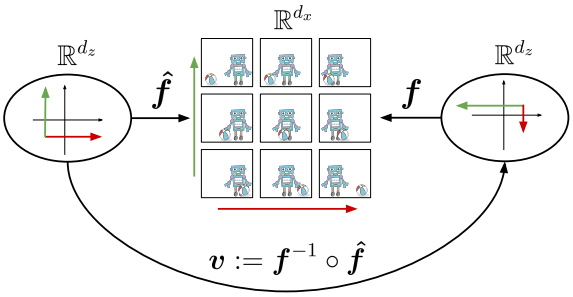


Figure 2: An illustration of disentanglement (Definition 7). The ground-truth decoder $\boldsymbol{f}$ captures the "natural factors of variations", which here are the x-positions of the robot and ball. The learned decoder $\hat{\boldsymbol{f}}$ is disentangled here because each of its latent coordinates corresponds exactly to one object in the scene, in a one-to-one fashion. Mathematically, this is captured by the special structure of the entanglement map $\boldsymbol{v} := \boldsymbol{f}^{-1} \circ \hat{\boldsymbol{f}}$, which is a permutation composed with an element-wise invertible transformation.

3 Nonparametric Partial Disentanglement via Mechanism Sparsity

In this section, we show that without further assumptions the model introduced so far is unidentifiable (Section 3.1), provide a first insight as to why mechanism sparsity can make the model identifiable and thus lead to disentanglement (Section 3.2), introduce the machinery of $\boldsymbol{G}$-preserving maps (Section 3.3) which leads up to theorems showing identifiability up to $\boldsymbol{a}$-consistency (Section 3.4) and $\boldsymbol{z}$-consistency (Section 3.5), which corresponds to partial disentanglement. We also relate these results to interventions (Section 3.4.1), show how to combine both regularization on $\hat{\boldsymbol{G}}^a$ and $\hat{\boldsymbol{G}}^z$ to obtain stronger guarantees (Section 3.6) and introduce a graphical criterion guaranteeing complete disentanglement (Section 3.7). Finally, we introduce the *sufficient influence assumptions* and prove the identifiability results (Section 3.8), and provide multiple examples to build intuition (Section 3.9).

3.1 Conditional Independence is Insufficient for Disentanglement and Graph Discovery

In the setting where the sequence $\{\boldsymbol{z}^t\}_{t=1}^T$ is fully observed, the absence of instantaneous causal effects (Assumption 2) combined with faithfulness (when the only conditional independencies in the distribution are those implied by the causal graph) makes the causal graph fully identifiable (Peters et al., 2017, Section 10.3.1). In our setting where only the sequence $\{\boldsymbol{f}(\boldsymbol{z}^t) + \boldsymbol{n}^t\}_{t=1}^T$ is observed, conditional independence alone is insufficient to identify the latent causal graph or the latent variables, as the following example demonstrates.

Example 2 *Suppose* $p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) := \mathcal{N}(\boldsymbol{z}^t; \boldsymbol{W}\boldsymbol{z}^{t-1} + \boldsymbol{a}^{t-1}\boldsymbol{e}_1, \eta^2 I)$ *where* $\boldsymbol{a}^t \in \mathcal{A} = \mathbb{R}$, $\boldsymbol{W} \in \mathbb{R}^{d\times d}$ *is symmetric and* $\boldsymbol{e}_1 \in \mathbb{R}^d$ *is the first one-hot vector. Let* $\boldsymbol{G}^z$ *and* $\boldsymbol{G}^a$ *be the graph entailed by this model. Note that* $\boldsymbol{G}^z$ *can be very dense while* $\boldsymbol{G}^a$ *contains only a single edge. Choose* $\boldsymbol{f}$ *to be any diffeomorphism on its image. This defines a model* $\boldsymbol{\theta} = (\boldsymbol{f}, p, \boldsymbol{G})$ *that satisfies the conditional independence assumption. We now construct a second model* $\tilde{\boldsymbol{\theta}}$ *that is observationally equivalent to* $\boldsymbol{\theta}$*, but has a drastically different graph* $\tilde{\boldsymbol{G}}$ *and different latent factors. Since* $\boldsymbol{W}$ *is symmetric, it can be diagonalized by an orthogonal matrix, i.e.,* $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^\top$*, where* $\boldsymbol{U}$ *is orthogonal and* $\boldsymbol{D}$

is diagonal. Define $\tilde{\boldsymbol{f}} := \boldsymbol{f} \circ \boldsymbol{U}$*, so that* $\boldsymbol{f}(\mathbb{R}^d) = \tilde{\boldsymbol{f}}(\mathbb{R}^d)$ *and* $\boldsymbol{v} := \boldsymbol{f}^{-1} \circ \tilde{\boldsymbol{f}} = \boldsymbol{U}$*. Define*

$$\tilde{p}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) := p(\boldsymbol{U}\boldsymbol{z}^t \mid \boldsymbol{U}\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) \,,$$

which is a well-defined conditional density by the change-of-variable formula (using $|\det \boldsymbol{U}| = 1$*). We can see* $\tilde{p}$ *as the conditional density resulting from first applying* $\boldsymbol{U}$ *to* $\boldsymbol{z}^{t-1}$*, taking a step using* p *to get* $\boldsymbol{z}^t$ *and finally transforming* $\boldsymbol{z}^t$ *using* $\boldsymbol{U}^\top$*, analogously to (7). It is clear that* $\tilde{p}$ *satisfies the conditional independence (Assumption 2) since an orthogonal transformation of an isotropic Gaussian remains an isotropic Gaussian. Furthermore, the mean of* $\tilde{p}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ *is given by* $\boldsymbol{U}^\top \boldsymbol{W}\boldsymbol{U}\boldsymbol{z}^{t-1} + \boldsymbol{a}^{t-1}\boldsymbol{U}^\top \boldsymbol{e}_1 = \boldsymbol{D}\boldsymbol{z}^{t-1} + \boldsymbol{a}^{t-1}(\boldsymbol{U}^\top)_{:,1}$*. We find that* $\tilde{p}$ *is Markov w.r.t.* $\tilde{\boldsymbol{G}}^z = \boldsymbol{I}$*, since* $\boldsymbol{D}$ *is diagonal, and* $\tilde{\boldsymbol{G}}^a := \mathbb{1}$*, assuming* $\boldsymbol{U}$ *is dense. Note that* $\tilde{\boldsymbol{G}}^z$ *will be in general much sparser than* $\boldsymbol{G}^z$ *and* $\tilde{\boldsymbol{G}}^a$ *is potentially much denser than* $\boldsymbol{G}^a$ *(how extreme these differences are depends on the exact choice of* $\boldsymbol{W}$*). Moreover, in general, the representation is not disentangled, since* $\boldsymbol{v} = \boldsymbol{U}$ *can be chosen to be dense. Nevertheless, these two models are observationally equivalent since they are equivalent up to diffeomorphism by construction (Section 2.4).*

The preceding example illustrates some difficulties one might face when it comes to the identifiability of the model introduced so far. This serves as a motivation to explore additional inductive biases such as mechanism sparsity, which we do next.

3.2 A First Mathematical Insight for Disentanglement via Mechanism Sparsity

In this section, we derive a first insight pointing towards how mechanism sparsity regularization, i.e., regularizing $\hat{\boldsymbol{G}}$ to be sparse, can promote disentanglement.

Before going further, we briefly introduce an abuse of notation that will be handy throughout: we will sometimes use vectors and matrices as sets of indices corresponding to their supports.

Definition 9 (Vectors & matrices as index sets) *Let* $\boldsymbol{a} \in \mathbb{R}^n$ *and* $\boldsymbol{A} \in \mathbb{R}^{m \times n}$*. We will sometimes use* $\boldsymbol{a}$ *to denote the set of indices corresponding to the support of the vector* $\boldsymbol{a}$*, i.e.*

$$\boldsymbol{a} \sim \{i \in [n] \mid \boldsymbol{a}_i \neq 0\} \,.$$

This will allow us to write things like $i \in \boldsymbol{a}$ *or* $\boldsymbol{a} \subseteq \boldsymbol{b}$*, where* $\boldsymbol{b} \in \mathbb{R}^n$*. We will use an analogous convention for matrices, i.e.,*

$$\boldsymbol{A} \sim \{(i,j) \in [m] \times [n] \mid \boldsymbol{A}_{i,j} \neq 0\} \,,$$

This will allow us to write things like $(i,j) \in \boldsymbol{A}$ *and* $\boldsymbol{A} \subseteq \boldsymbol{B}$*, where* $\boldsymbol{B} \in \mathbb{R}^{m \times n}$*.*

Recall that we would like to show that $\boldsymbol{\theta} \sim_{\text{obs}} \hat{\boldsymbol{\theta}}$ implies $\boldsymbol{\theta} \sim_{\text{perm}} \hat{\boldsymbol{\theta}}$, i.e., disentanglement (or partial disentanglement). Our approach will be to start from (6), which is guaranteed by Proposition 2, and perform a series of algebraic manipulations to gain mathematical insight into how regularizing $\hat{\boldsymbol{G}}$ to be sparse (mechanism sparsity) can induce disentanglement. A key manipulation will be taking first and second order derivatives. For this to be possible, we require a certain level of smoothness for the transition models:

Assumption 4 (Smoothness of transition model) *When* $\boldsymbol{a}$ *is continuous, the transition densities* $p(\boldsymbol{z}_i^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ *are* C^2 *functions from* $\mathbb{R} \times \mathbb{R}^{d_z \times (t-1)} \times \mathcal{A}^t$ *to* $\mathbb{R}$ *and* $\mathcal{A} \subseteq \mathbb{R}^\ell$ *is regular closed*[5]*. When* $\boldsymbol{a}$ *is discrete (e.g. Section 3.4.1), for all* $\boldsymbol{a}^{<t}$*,* $p(\boldsymbol{z}_i^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ *are* C^2 *functions from* $\mathbb{R} \times \mathbb{R}^{d_z \times (t-1)}$ *to* $\mathbb{R}$*.*

5. A set $\mathcal{A} \subseteq \mathbb{R}^\ell$ is regular closed when it is equal to the closure of its interior, i.e. $\overline{\mathcal{A}^\circ} = \mathcal{A}$.

We start by taking the $\log$ on both sides of (6) and let $q := \log p$ and $\hat{q} := \log \hat{p}$:

$$\hat{q}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = q(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t}) + \log |\det D\boldsymbol{v}(\boldsymbol{z}^t)| \,.$$

We then take the derivative w.r.t. $\boldsymbol{z}^t$ on both sides:

$$D_z^t \hat{q}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = D_z^t q(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t}) D\boldsymbol{v}(\boldsymbol{z}^t) + \eta(\boldsymbol{z}^t) \in \mathbb{R}^{1 \times d_z} \,, \tag{8}$$

where $D_z^t q$ denotes the Jacobian of $q(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ w.r.t. $\boldsymbol{z}^t$ and analogously for $D_z^t \hat{q}$. The term $\eta(\boldsymbol{z}^t)$ is the derivative of $\log |\det D\boldsymbol{v}(\boldsymbol{z}^t)|$ w.r.t. $\boldsymbol{z}^t$.

We differentiate[6] yet once more w.r.t. $\boldsymbol{a}^\tau$ for some $\tau < t$ (assuming $\boldsymbol{a}^t$ is continuous for now) and obtain

$$H_{z,a}^{t,\tau} \hat{q}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = D\boldsymbol{v}(\boldsymbol{z}^t)^\top H_{z,a}^{t,\tau} q(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t}) \in \mathbb{R}^{d_z \times d_a} \,, \tag{9}$$

where $H_{z,a}^{t,\tau} q \in \mathbb{R}^{d_z \times d_a}$ is the Hessian matrix of second derivatives w.r.t. $\boldsymbol{z}^t$ and $\boldsymbol{a}^\tau$ and similarly for $H_{z,a}^{t,\tau} \hat{q}$.

We now look more closely at some specific entry (i, ℓ) of the Hessian $H_{z,a}^{t,\tau} q$. We first see that

$$\frac{\partial^2}{\partial \boldsymbol{a}_\ell^\tau \partial \boldsymbol{z}_i^t} q(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = \frac{\partial^2}{\partial \boldsymbol{a}_\ell^\tau \partial \boldsymbol{z}_i^t} \sum_{j=1}^{d_z} q(\boldsymbol{z}_j^t \mid \boldsymbol{z}_{\mathbf{Pa}_j^z}^{<t}, \boldsymbol{a}_{\mathbf{Pa}_j^a}^{<t}) \tag{10}$$

$$= \frac{\partial}{\partial \boldsymbol{a}_\ell^\tau} \sum_{j=1}^{d_z} \frac{\partial}{\partial \boldsymbol{z}_i^t} q(\boldsymbol{z}_j^t \mid \boldsymbol{z}_{\mathbf{Pa}_j^z}^{<t}, \boldsymbol{a}_{\mathbf{Pa}_j^a}^{<t}) \tag{11}$$

$$= \frac{\partial}{\partial \boldsymbol{a}_\ell^\tau} \frac{\partial}{\partial \boldsymbol{z}_i^t} q(\boldsymbol{z}_i^t \mid \boldsymbol{z}_{\mathbf{Pa}_i^z}^{<t}, \boldsymbol{a}_{\mathbf{Pa}_i^a}^{<t}) \,, \tag{12}$$

where the first equality holds by (1) & (2) and a basic property of logarithms. It is clear that (12) equals zero when $\ell \notin \mathbf{Pa}_i^a$. This is a crucial observation, since it implies that whenever $\boldsymbol{G}_{i,\ell}^a = 0$, we also have $(H_{z,a}^{t,\tau} q)_{i,\ell} = 0$. In other words, $H_{z,a}^{t,\tau} q \subseteq \boldsymbol{G}^a$. The same argument can also be applied to get $H_{z,a}^{t,\tau} \hat{q} \subseteq \hat{\boldsymbol{G}}^a$. Notice how this argument would fail without the conditional independence assumption, since $\boldsymbol{z}_i^t$ could appear in terms other than $q(\boldsymbol{z}_i^t \mid \boldsymbol{z}_{\mathbf{Pa}_i^z}^{<t}, \boldsymbol{a}_{\mathbf{Pa}_i^a}^{<t})$, as it could be the parent of some other $\boldsymbol{z}_j^t$, for $j \neq i$.

Intuitive argument. We can start to see why regularizing $\hat{\boldsymbol{G}}$ to be sparse might induce disentanglement. Intuitively, a sparse $\hat{\boldsymbol{G}}^a$ forces $D\boldsymbol{v}(\boldsymbol{z}^t)$ to be sparse, since otherwise the l.h.s. of (13) will not be sparse:

$$\underbrace{H_{z,a}^{t,\tau} \hat{q}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})}_{\subseteq \hat{\boldsymbol{G}}^a} = \underbrace{D\boldsymbol{v}(\boldsymbol{z}^t)^\top}_{\text{forced to be sparse}} \underbrace{H_{z,a}^{t,\tau} q(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t})}_{\subseteq \boldsymbol{G}^a} \,, \tag{13}$$

And, of course, the sparser $D\boldsymbol{v}(\boldsymbol{z}^t)$ is, the more disentangled $\hat{\boldsymbol{f}}$ is, since $D\boldsymbol{v}_{i,j} = 0$ everywhere implies $\boldsymbol{V}_{i,j} = 0$ under weak assumptions (Proposition 1). The above argument is not rigorous and is provided only to build intuition. It will be formalized later.

6. This derivative is well defined on $\mathcal{A}$ (in the sense that it does not depend on its C^k extension) since $\mathcal{A}$ is regular close. We prove this general fact in Lemma 4 in the appendix.

Sparse temporal dependencies. In what precedes, we made use of the sparsity of the graph $\hat{\boldsymbol{G}}^a$ to argue that $D\boldsymbol{v}$ must also be sparse. We now show a similar intuition based on the sparsity of $\hat{\boldsymbol{G}}^z$. Starting from (8), instead of differentiating w.r.t. $\boldsymbol{a}^\tau$, we will differentiate w.r.t. $\boldsymbol{z}^\tau$, for some $\tau < t$, which yields:

$$H^{t,\tau}_{z,z}\hat{q}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = D\boldsymbol{v}(\boldsymbol{z}^t)^\top H^{t,\tau}_{z,z} q(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t}) D\boldsymbol{v}(\boldsymbol{z}^\tau) \in \mathbb{R}^{d_z \times d_z}\,, \tag{14}$$

where $H^{t,\tau}_{z,z}q$ is the Hessian matrix of second derivatives of q w.r.t. $\boldsymbol{z}^t$ and $\boldsymbol{z}^\tau$, and analogously for $H^{t,\tau}_{z,z}\hat{q}$. Using an argument perfectly analogous to Equations (10) to (12), we can show that whenever $\boldsymbol{G}^z_{i,j} = 0$, we also have $(H^{t,\tau}_{z,z}q)_{i,j} = 0$, and similarly for $\hat{\boldsymbol{G}}^z$ and $H^{t,\tau}_{z,z}\hat{q}$. In other words, $H^{t,\tau}_{z,z}q \subseteq \boldsymbol{G}^z$ and $H^{t,\tau}_{z,z}\hat{q} \subseteq \hat{\boldsymbol{G}}^z$. Therefore, analogously to (13), regularizing $\hat{\boldsymbol{G}}^z$ to be sparse intuitively should force $D\boldsymbol{v}$ to be sparse as well, i.e., bringing us closer to disentanglement:

$$\underbrace{H^{t,\tau}_{z,z}\hat{q}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})}_{\subseteq \hat{\boldsymbol{G}}^z} = \underbrace{D\boldsymbol{v}(\boldsymbol{z}^t)^\top}_{\text{forced to be sparse}} \underbrace{H^{t,\tau}_{z,z} q(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t})}_{\subseteq \boldsymbol{G}^z} \underbrace{D\boldsymbol{v}(\boldsymbol{z}^\tau)}_{\text{forced to be sparse}}\,. \tag{15}$$

The crux of our technical contribution in this work is to make the above arguments formal and precisely characterize what the sparsity structure of $D\boldsymbol{v}(\boldsymbol{z})$ (hence, $\boldsymbol{V}$) will be based on the ground-truth graph $\boldsymbol{G}$ (Theorems 1, 2 & 3). We also provide conditions on $\boldsymbol{G}$ to guarantee complete disentanglement (Proposition 7).

3.3 Graph Preserving Maps

Theorems 1, 2, 3 & 5 will show how regularizing $\hat{\boldsymbol{G}}$ to be sparse can force the dependency graph of the entanglement map $\boldsymbol{v}$ to be sparse as well. These results characterize the functional dependency structure of the entanglement map $\boldsymbol{v}$ as a function of the ground-truth graph $\boldsymbol{G}$. This link will be made precise thanks to the notion of graph preserving maps, which we define next. Before going further, we need to set up the following notation.

Definition 10 (Aligned subspaces of $\mathbb{R}^m$ and $\mathbb{R}^{m\times n}$) *Given a binary vector $\boldsymbol{b} \in \{0,1\}^m$, let*

$$\mathbb{R}^m_{\boldsymbol{b}} := \{\boldsymbol{x} \in \mathbb{R}^m \mid \boldsymbol{b}_i = 0 \implies \boldsymbol{x}_i = 0\}$$

Given a binary matrix $\boldsymbol{B} \in \{0,1\}^{m\times n}$, let

$$\mathbb{R}^{m\times n}_{\boldsymbol{B}} := \{\boldsymbol{M} \in \mathbb{R}^{m\times n} \mid \boldsymbol{B}_{i,j} = 0 \implies \boldsymbol{M}_{i,j} = 0\}\,.$$

Note that $\mathbb{R}^m_{\boldsymbol{b}}$ and $\mathbb{R}^{m\times n}_{\boldsymbol{B}}$ are vector spaces under addition. This means that given $\boldsymbol{a}^{(1)}, \dots, \boldsymbol{a}^{(k)} \in \mathbb{R}^m_{\boldsymbol{b}}$, we have that $\operatorname{span}\{\boldsymbol{a}^{(1)}, \dots, \boldsymbol{a}^{(k)}\} \subseteq \mathbb{R}^m_{\boldsymbol{b}}$, where span denotes the subspace of all linear combinations. Similarly, given $\boldsymbol{A}^{(1)}, \dots, \boldsymbol{A}^{(k)} \in \mathbb{R}^{m\times n}_{\boldsymbol{B}}$, we have that $\operatorname{span}\{\boldsymbol{A}^{(1)}, \dots, \boldsymbol{A}^{(k)}\} \subseteq \mathbb{R}^{m\times n}_{\boldsymbol{B}}$.

To start formally reasoning about what will happen if $\hat{\boldsymbol{G}}^a$ is regularized sparse, we temporarily assume that $\hat{\boldsymbol{G}}^a = \boldsymbol{G}^a$. With this assumption, we can interpret (13) as implying that $D\boldsymbol{v}(\boldsymbol{z}^t)^\top$ must *preserve* the "sparsity structure" of the matrix $H^{t,\tau}_{z,a}q$. This observation motivates the following definitions, which will be central to our contribution.

Definition 11 ($\boldsymbol{G}$-preserving matrix) *Given $\boldsymbol{G} \in \{0,1\}^{m\times n}$, a matrix $\boldsymbol{C} \in \mathbb{R}^{m\times m}$ is **$\boldsymbol{G}$-preserving** when*

$$\boldsymbol{C}^\top \mathbb{R}^{m\times n}_{\boldsymbol{G}} \subseteq \mathbb{R}^{m\times n}_{\boldsymbol{G}}\,.$$

Definition 12 ($\boldsymbol{G}$-preserving functions) *Given $\boldsymbol{G} \in \{0,1\}^{m \times n}$, a function $\boldsymbol{c} : \mathbb{R}^m \to \mathbb{R}^m$ is **$\boldsymbol{G}$-preserving** when its dependency graph $\boldsymbol{C}$ (Definition 2) is $\boldsymbol{G}$-preserving.*

Without surprise, a linear map $\boldsymbol{c}(\boldsymbol{z}) := \boldsymbol{C}\boldsymbol{z}$ where $\boldsymbol{C} \in \mathbb{R}^{m \times m}$ is $\boldsymbol{G}$-preserving (Definition 12) if and only if the matrix $\boldsymbol{C}$ is $\boldsymbol{G}$-preserving (Definition 11).

We now show that $\boldsymbol{G}$-preserving functions can be defined alternatively in terms of a simple condition on their dependency graph. This characterization of $\boldsymbol{G}$-preserving functions is key to understanding how (partial) disentanglement results from sparsity regularization.

Proposition 3 *A function $\boldsymbol{c}$ with dependency graph $\boldsymbol{C}$ (Definition 2) is $\boldsymbol{G}$-preserving if and only*

$$\boldsymbol{G}_{i,\cdot} \not\subseteq \boldsymbol{G}_{j,\cdot} \implies \boldsymbol{C}_{i,j} = 0, \text{ for all } i, j\,.$$

Proof We start by showing the "only if" statement. We suppose $\boldsymbol{G}_{i,\cdot} \not\subseteq \boldsymbol{G}_{j,\cdot}$ and must now show that $\boldsymbol{C}_{i,j} = 0$. We know that there exists k such that $\boldsymbol{G}_{i,k} = 1$ but $\boldsymbol{G}_{j,k} = 0$. Since $\boldsymbol{C}^\top \mathbb{R}_{\boldsymbol{G}}^{m \times n} \subseteq \mathbb{R}_{\boldsymbol{G}}^{m \times n}$ and $\boldsymbol{e}_i \boldsymbol{e}_k^\top \in \mathbb{R}_{\boldsymbol{G}}^{m \times n}$, we must have $\boldsymbol{C}^\top \boldsymbol{e}_i \boldsymbol{e}_k^\top \in \mathbb{R}_{\boldsymbol{G}}^{m \times n}$. Since $\boldsymbol{G}_{j,k} = 0$, we must have $0 = (\boldsymbol{C}^\top \boldsymbol{e}_i \boldsymbol{e}_k^\top)_{j,k} = \boldsymbol{C}_{i,j}$.

We now show the "if" statement. Let $\boldsymbol{A} \in \mathbb{R}_{\boldsymbol{G}}^{m \times n}$. Take some (i, j) such that $\boldsymbol{G}_{i,j} = 0$. We must now show that $(\boldsymbol{C}^\top \boldsymbol{A})_{i,j} = 0$. We have $(\boldsymbol{C}^\top \boldsymbol{A})_{i,j} = \sum_k \boldsymbol{C}_{k,i} \boldsymbol{A}_{k,j}$. We now check that each term in this sum must be zero. If $\boldsymbol{A}_{k,j} = 0$, of course the corresponding term is zero. If $\boldsymbol{A}_{k,j} \neq 0$, it implies that $\boldsymbol{G}_{k,j} = 1$ and thus $\boldsymbol{G}_{k,\cdot} \not\subseteq \boldsymbol{G}_{i,\cdot}$. By assumption, this implies that $\boldsymbol{C}_{k,i} = 0$ and thus $\boldsymbol{C}_{k,i} \boldsymbol{A}_{k,j} = 0$. Hence $(\boldsymbol{C}^\top \boldsymbol{A})_{i,j} = 0$ as desired. ■

We now characterize differentiable $\boldsymbol{G}$-preserving functions in terms of their Jacobian matrices.

Lemma 1 *A differentiable function $\boldsymbol{c} : \mathbb{R}^m \to \mathbb{R}^m$ is $\boldsymbol{G}$-preserving if and only if, for all $\boldsymbol{z} \in \mathbb{R}^m$, $D\boldsymbol{c}(\boldsymbol{z})$ is $\boldsymbol{G}$-preserving.*

Proof Assume $\boldsymbol{c}$ is $\boldsymbol{G}$-preserving with dependency graph $\boldsymbol{C}$. By Proposition 3, this is equivalent to having that, for all $i, j \in [n]$,

$$\boldsymbol{G}_{i,\cdot} \not\subseteq \boldsymbol{G}_{j,\cdot} \implies \boldsymbol{C}_{i,j} = 0\,.$$

But by Proposition 1, this statement is equivalent to

$$\boldsymbol{G}_{i,\cdot} \not\subseteq \boldsymbol{G}_{j,\cdot} \implies \forall \boldsymbol{z} \in \mathbb{R}^m,\ D\boldsymbol{c}(\boldsymbol{z})_{i,j} = 0\,,$$

which is equivalent to saying that $D\boldsymbol{c}(\boldsymbol{z})$ is $\boldsymbol{G}$-preserving for all $\boldsymbol{z} \in \mathbb{R}^m$ (again by Proposition 3). ■

We will now show that $\boldsymbol{G}$-preserving diffeomorphisms form a group under composition. To do so, we start by showing that invertible $\boldsymbol{G}$-preserving matrices form a group under matrix multiplication (Proposition 4) and extend the result to diffeomorphisms in Proposition 5.

Proposition 4 *Invertible $\boldsymbol{G}$-preserving matrices form a group under matrix multiplication.*

Proof We must show that the set of invertible $\boldsymbol{G}$-preserving matrices contains the identity, is closed under matrix multiplication, and is closed under inversion.

Clearly, $\boldsymbol{I}$ is $\boldsymbol{G}$-preserving since $\boldsymbol{I}^\top \mathbb{R}_{\boldsymbol{G}}^{m\times n} = \mathbb{R}_{\boldsymbol{G}}^{m\times n}$.

Let $\boldsymbol{C}_1$ and $\boldsymbol{C}_2$ be $\boldsymbol{G}$-preserving. Then, $\boldsymbol{C}_1\boldsymbol{C}_2$ is $\boldsymbol{G}$-preserving because

$$(\boldsymbol{C}_1\boldsymbol{C}_2)^\top \mathbb{R}_{\boldsymbol{G}}^{m\times n} = \boldsymbol{C}_2^\top \boldsymbol{C}_1^\top \mathbb{R}_{\boldsymbol{G}}^{m\times n} \subseteq \boldsymbol{C}_2^\top \mathbb{R}_{\boldsymbol{G}}^{m\times n} \subseteq \mathbb{R}_{\boldsymbol{G}}^{m\times n}\,.$$

Let $\boldsymbol{C}$ be $\boldsymbol{G}$-preserving and invertible. Since $\boldsymbol{C}^\top$ is invertible as a map from $\mathbb{R}^{m\times n}$ to $\mathbb{R}^{m\times n}$, the dimensionality of the subspace $\mathbb{R}_{\boldsymbol{G}}^{m\times n}$ must be equal to the dimensionality of $\boldsymbol{C}^\top \mathbb{R}_{\boldsymbol{G}}^{m\times n}$. This fact combined with $\boldsymbol{C}^\top \mathbb{R}_{\boldsymbol{G}}^{m\times n} \subseteq \mathbb{R}_{\boldsymbol{G}}^{m\times n}$ implies that $\boldsymbol{C}^\top \mathbb{R}_{\boldsymbol{G}}^{m\times n} = \mathbb{R}_{\boldsymbol{G}}^{m\times n}$. Hence $\mathbb{R}_{\boldsymbol{G}}^{m\times n} = (\boldsymbol{C}^{-1})^\top \mathbb{R}_{\boldsymbol{G}}^{m\times n}$, i.e. $\boldsymbol{C}^{-1}$ is $\boldsymbol{G}$-preserving. ■

We now extend the above results to diffeomorphisms using Proposition 1.

Proposition 5 *The set of $\boldsymbol{G}$-preserving diffeomorphims forms a group under composition.*

Proof We must show that the set of $\boldsymbol{G}$-preserving diffeomorphisms contains the identity, is closed under matrix multiplication, and is closed under inversion.

The first statement is trivial since the entanglement graph of the identity diffeomorphism is the identity graph $\boldsymbol{C} := \boldsymbol{I}$, and of course it is $\boldsymbol{G}$-preserving.

We now prove the second statement. Let $\boldsymbol{c}$ and $\boldsymbol{c}'$ be two diffeomorphisms with dependency graph $\boldsymbol{C}$ and $\boldsymbol{C}'$ respectively. By the chain rule, we have that

$$D(\boldsymbol{c}\circ\boldsymbol{c}')(\boldsymbol{z}) = D\boldsymbol{c}(\boldsymbol{c}'(\boldsymbol{z}))D\boldsymbol{c}'(\boldsymbol{z})\,.$$

By Lemma 1, we have that $D\boldsymbol{c}(\boldsymbol{c}'(\boldsymbol{z}))$ and $D\boldsymbol{c}'(\boldsymbol{z})$ and $\boldsymbol{G}$-preserving matrices and, by Proposition 4 their product must also be $\boldsymbol{G}$-preserving. Hence $D(\boldsymbol{c}\circ\boldsymbol{c}')(\boldsymbol{z})$ is $\boldsymbol{G}$-preserving for all $\boldsymbol{z}$ and thus, by Lemma 1, $\boldsymbol{c}\circ\boldsymbol{c}'$ is $\boldsymbol{G}$-preserving.

The proof of the third statement has a similar flavor. By the inverse function theorem, we have

$$D\boldsymbol{c}^{-1}(\boldsymbol{z}) = D\boldsymbol{c}(\boldsymbol{c}^{-1}(\boldsymbol{z}))^{-1}\,.$$

Moreover, by Lemma 1, $D\boldsymbol{c}(\boldsymbol{c}^{-1}(\boldsymbol{z}))$ is $\boldsymbol{G}$ preserve. Furthermore, its inverse is also $\boldsymbol{G}$-preserving by Proposition 4. Similarly to the previous step, because $\boldsymbol{c}^{-1}$ is C^1, we can use Lemma 1 to conclude that $\boldsymbol{c}^{-1}$ is also $\boldsymbol{G}$-preserving. ■

3.4 Nonparametric Identifiability via Auxiliary Variables with Sparse Influence

In this section, we introduce our first identifiability results based on the sparsity of the graph $\boldsymbol{G}^a$ which describes the structure of the dependencies between $\boldsymbol{a}^{<t}$ and $\boldsymbol{z}^t$. We will see that, under some assumptions, regularizing the learned graph $\hat{\boldsymbol{G}}^a$ to be sparse will allow identifiability up to the following equivalence class:

Definition 13 ($\boldsymbol{a}$-consistency equivalence) *We say two models $\boldsymbol{\theta} := (\boldsymbol{f}, p, \boldsymbol{G})$ and $\tilde{\boldsymbol{\theta}} := (\tilde{\boldsymbol{f}}, \tilde{p}, \tilde{\boldsymbol{G}})$ satisfying Assumptions 1, 2 & 3 are **$\boldsymbol{a}$-consistent**, denoted $\boldsymbol{\theta} \sim^{\boldsymbol{a}}_{\text{con}} \tilde{\boldsymbol{\theta}}$, if and only if there exists a permutation matrix $\boldsymbol{P}$ such that*

1. $\boldsymbol{\theta} \sim_{\text{diff}} \tilde{\boldsymbol{\theta}}$ *(Def. 5), and* $\tilde{\boldsymbol{G}}^a = \boldsymbol{P}\boldsymbol{G}^a$ *; and*

2. *the entanglement map* $\boldsymbol{v} := \boldsymbol{f}^{-1} \circ \tilde{\boldsymbol{f}}$ *can be written as* $\boldsymbol{v} = \boldsymbol{c} \circ \boldsymbol{P}^\top$ *where* $\boldsymbol{c}$ *is a* $\boldsymbol{G}^a$*-preserving diffeomorphism (Def. 12).*

The main difference between $\boldsymbol{a}$-consistency (above definition) and permutation equivalence (Definition 6), is that instead of having $\boldsymbol{v} = \boldsymbol{d} \circ \boldsymbol{P}^\top$ where $\boldsymbol{d}$ is element-wise, we have $\boldsymbol{v} = \boldsymbol{c} \circ \boldsymbol{P}^\top$ where $\boldsymbol{c}$ is $\boldsymbol{G}^a$-preserving, which allows some mixing between the latent factors. Importantly, a $\boldsymbol{G}^a$-preserving map typically has missing edges in its dependency graph, as Proposition 3 shows. This means that this equivalence relation imposes structure on the entanglement map $\boldsymbol{v}$. Depending on the structure of $\boldsymbol{G}^a$, this can mean either complete, partial, or no disentanglement whatsoever. Note that the equivalence $\sim_{\text{perm}}$ is stronger than $\sim^{\boldsymbol{a}}_{\text{con}}$, in the sense that $\boldsymbol{\theta} \sim_{\text{perm}} \hat{\boldsymbol{\theta}} \implies \boldsymbol{\theta} \sim^{\boldsymbol{a}}_{\text{con}} \hat{\boldsymbol{\theta}}$. This is because element-wise transformations $\boldsymbol{d}$ are always $\boldsymbol{G}$-preserving, for any $\boldsymbol{G}$.

We demonstrate in Appendix A.4 that the $\boldsymbol{a}$-consistency relation is indeed an equivalence relation, as claimed in the above definition. This follows from the fact that the set of $\boldsymbol{G}^a$-preserving diffeomorphisms forms a *group* under composition (Proposition 5).

The first result provides conditions under which regularizing the learned graph $\hat{\boldsymbol{G}}^a$ to be as sparse as the ground-truth graph $\boldsymbol{G}^a$ will induce the learned model to be $\boldsymbol{a}$-consistent with the ground-truth one.

Theorem 1 (Nonparametric disentanglement from continuous a with sparse influence) *Let the parameters* $\boldsymbol{\theta} := (\boldsymbol{f}, p, \boldsymbol{G})$ *and* $\hat{\boldsymbol{\theta}} := (\hat{\boldsymbol{f}}, \hat{p}, \hat{\boldsymbol{G}})$ *correspond to two models satisfying Assumptions 1, 2, 3, & 4. Further assume that*

1. ***[Observational equivalence]*** $\boldsymbol{\theta} \sim_{\text{obs}} \hat{\boldsymbol{\theta}}$ *(Def. 4);*

2. ***[Sufficient influence of a]*** *The Hessian matrix* $H^{t,\tau}_{z,a} \log p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ *varies "sufficiently", as formalized in Assumption 6;*

Then, there exists a permutation matrix $\boldsymbol{P}$ *such that* $\boldsymbol{P}\boldsymbol{G}^a \subseteq \hat{\boldsymbol{G}}^a$*. Further assume that*

3. ***[Sparsity regularization]*** $||\hat{\boldsymbol{G}}^a||_0 \leq ||\boldsymbol{G}^a||_0$*;*

Then, $\boldsymbol{\theta} \sim^{\boldsymbol{a}}_{\text{con}} \hat{\boldsymbol{\theta}}$ *(Def. 13).*

The second assumption as well as a proof of this result is delayed to Section 3.8 for pedagogical reasons. We now describe and provide intuition about each assumption one by one.

Observational equivalence. The first assumption simply requires that both models agree on the observational model. In practice, this is achieved by fitting the model to data.

Sufficient influence. The second assumption requires that the "effect" of $\boldsymbol{a}^{<t}$ on $\boldsymbol{z}^t$ is "sufficiently strong". The assumption will be formalized and discussed in more detail later in Sections 3.8 & 3.9, but we can already see that it concerns the Hessian matrix $H^{t,\tau}_{z,a} \log p$ that we saw earlier in Eq. (13) of Sect. 3.2.

Sparsity regularization. The first two assumptions imply that the learned graph $\hat{\boldsymbol{G}}^a$ is a supergraph of some permutation of the ground-truth graph $\boldsymbol{G}^a$. By adding the *sparsity regularization* assumption, we find that the learned graph $\hat{\boldsymbol{G}}^a$ *is exactly* a permutation of the ground-truth graph $\boldsymbol{G}^a$ and that, more precisely, the learned model is $\sim^{\boldsymbol{a}}_{\text{con}}$-equivalent to the ground-truth. This assumption is satisfied if $\hat{\boldsymbol{G}}^a$ is a minimal graph among all graphs that allow the model to exactly match the ground-truth generative distribution. In Section 5, we suggest achieving this in practice by adding a sparsity penalty in the training objective or by constraining the optimization problem.

$\boldsymbol{a}$-consistency. The final conclusion of the result states that the learned model is $\sim^{\boldsymbol{a}}_{\text{con}}$-equivalent to the ground-truth, which means the entanglement map $\boldsymbol{v} := \boldsymbol{f}^{-1} \circ \hat{\boldsymbol{f}}$ can be written as $\boldsymbol{v} = \boldsymbol{c} \circ \boldsymbol{P}^\top$ where $\boldsymbol{c}$ is $\boldsymbol{G}^a$-preserving. This is important since the $\boldsymbol{G}^a$-preserving condition imposes structure on the entanglement graph $\boldsymbol{V}$ (Definition 3), as implied by Proposition 3. In other words, the result predicts precisely which latent factors are expected to remain entangled.

Remark 1 (Inverse of $\boldsymbol{v}$) *We defined $\boldsymbol{v}$ to be the mapping from the learned to the ground-truth representation, but in some context, it might be more telling to look at $\boldsymbol{v}^{-1}$, which maps from the ground-truth to the learned representation. If $\boldsymbol{v} = \boldsymbol{c} \circ \boldsymbol{P}^\top$ where $\boldsymbol{c}$ is $\boldsymbol{G}^a$-preserving (as predicted by Theorem 1), we know that its inverse is given by $\boldsymbol{v}^{-1} = \boldsymbol{P} \circ \boldsymbol{c}^{-1}$ where $\boldsymbol{c}^{-1}$ is $\boldsymbol{G}^a$-preserving, by closure under inversion (Proposition 5).*

The following result is the same as above, but for *discrete* auxiliary variables $\boldsymbol{a}$. This case is very important to cover the case where $\boldsymbol{a}$ indexes sparse interventions targeting latent factors, which we discuss in more detail in Section 3.4.1. Note that the only difference from the above theorem is the "sufficient influence" assumption, which we will present formally in Section 3.8 together with a proof of the result.

Theorem 2 (Nonparametric disentanglement via discrete $\boldsymbol{a}$ with sparse influence) *Let the parameters $\boldsymbol{\theta} := (\boldsymbol{f}, p, \boldsymbol{G})$ and $\hat{\boldsymbol{\theta}} := (\hat{\boldsymbol{f}}, \hat{p}, \hat{\boldsymbol{G}})$ correspond to two models satisfying Assumptions 1, 2, 3 & 4. Further assume that*

1. ***[Observational equivalence]*** $\boldsymbol{\theta} \sim_{\text{obs}} \hat{\boldsymbol{\theta}}$ *(Def. 4);*
2. ***[Sufficient influence of $\boldsymbol{a}$]*** *The vector of derivatives $D^t_z \log p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ depends "sufficiently strongly" on each component $\boldsymbol{a}_\ell$, as formalized in Assumption 7;*

Then, there exists a permutation matrix $\boldsymbol{P}$ such that $\boldsymbol{P}\boldsymbol{G}^a \subseteq \hat{\boldsymbol{G}}^a$. Further assume that

3. ***[Sparsity regularization]*** $||\hat{\boldsymbol{G}}^a||_0 \leq ||\boldsymbol{G}^a||_0$*;*

Then, $\boldsymbol{\theta} \sim^{\boldsymbol{a}}_{\text{con}} \hat{\boldsymbol{\theta}}$ (Def. 13).

We now provide a few examples to illustrate how Theorems 1 & 2 can be applied. Here, we concentrate on the relationship between the graph $\boldsymbol{G}^a$ and the entanglement graph $\boldsymbol{V}$ (Definition 3). The question of whether or not the sufficient influence assumption is satisfied will be delayed to Section 3.9, where the examples will be made more concrete by specifying latent models more explicitly.

Example 3 ($\boldsymbol{G}^a = \boldsymbol{I}$ implies complete disentanglement) *Assume that $d_a = d_z$ and $\boldsymbol{G}^a = \boldsymbol{I}$, i.e., each latent variable is affected by only one auxiliary variable, and each auxiliary variable affects only one latent variable. The graph $\boldsymbol{G}^a$ is depicted in Figure 3a and $\boldsymbol{G}^z$ could be anything (see the remark below). Assuming that the ground-truth transition model satisfies the sufficient influence assumption of Theorem 1 or 2, we have $\boldsymbol{\theta} \sim_{\text{obs}} \hat{\boldsymbol{\theta}}$ & $||\hat{\boldsymbol{G}}^a||_0 \leq ||\boldsymbol{G}^a||_0 \implies \boldsymbol{\theta} \sim^{\boldsymbol{a}}_{\text{con}} \hat{\boldsymbol{\theta}}$. This means that there exists a permutation matrix $\boldsymbol{P}$ such that $\hat{\boldsymbol{G}}^a = \boldsymbol{P}\boldsymbol{G}^a$ and such that the entanglement map is given by $\boldsymbol{v} = \boldsymbol{c} \circ \boldsymbol{P}^\top$ where $\boldsymbol{c}$ is a $\boldsymbol{G}^a$-preserving diffeomorphism (Definition 11). But since $\boldsymbol{G}^a = \boldsymbol{I}$, Proposition 3 tells us that the dependency graph of $\boldsymbol{c}$ is simply $\boldsymbol{C} := \boldsymbol{I}$ and thus the entanglement graph is $\boldsymbol{V} = \boldsymbol{P}^\top$, i.e., complete disentanglement holds. In fact, one could add more columns to $\boldsymbol{G}^a$ (i.e. adding auxiliary variables) without changing the conclusion. Example 11 will provide a concrete example satisfying the sufficient influence assumption of Theorem 2.*

Remark 2 (Temporal dependencies are not necessary) *The above example did not mention anything about the temporal graph $\boldsymbol{G}^z$. That is because this graph could be anything, in fact, we could be in the special case where there are no temporal dependencies whatsoever, i.e., $T = 1$ and the latent model is simply $p(\boldsymbol{z} \mid \boldsymbol{a}) = \prod_{i=1}^{d_z} p(\boldsymbol{z}_i \mid \boldsymbol{a})$. In that case, Theorems 1 & 2 could still be applied to prove the identifiability of the representation, as long as their assumptions hold. This remark also applies to the next two examples.*

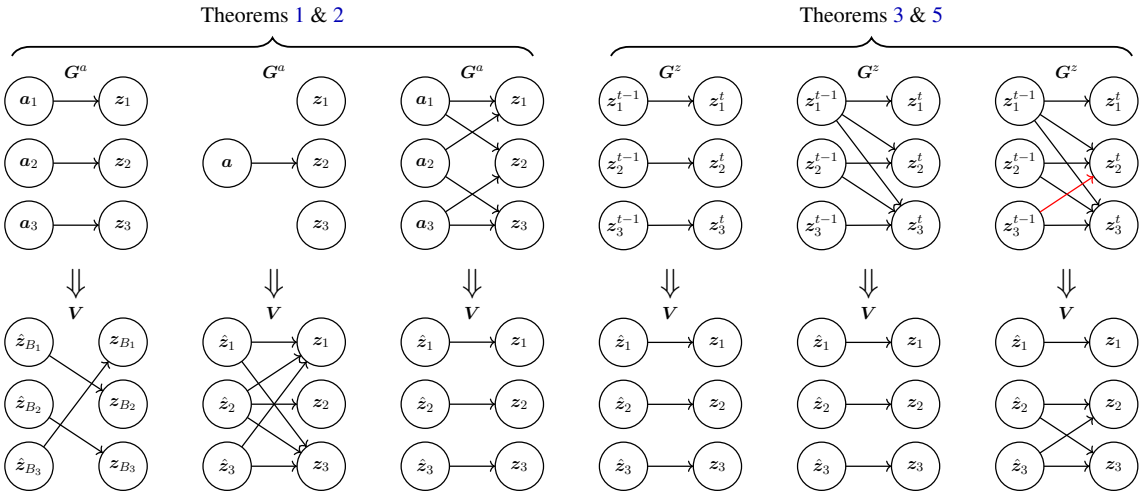
Example 4 (Action targeting a single latent variable identifies it) *Consider the situation depicted in Figure 1 where $\boldsymbol{z}_1$ is the tree position, $\boldsymbol{z}_2$ is the robot position and $\boldsymbol{z}_3$ is the ball position ($d_z = 3$). Assume $\boldsymbol{a} \in \mathbb{R}$ corresponds to the torque applied to the robot wheels ($d_a = 1$). We therefore have $\boldsymbol{G}^a = [0, 1, 0]^\top$, i.e., $\boldsymbol{a}$ affects only $\boldsymbol{z}_2$. For the sake of this example, $\boldsymbol{G}^z$ can be anything, i.e., it does not have to be lower triangular like in Figure 1 (see remark above).*

If the sufficient influence assumption of Theorem 1 or 2 is satisfied, we see that $\boldsymbol{\theta} \sim_{\text{obs}} \hat{\boldsymbol{\theta}}$ & $||\hat{\boldsymbol{G}}^a||_0 \leq ||\boldsymbol{G}^a||_0$ implies $\boldsymbol{v} = \boldsymbol{c} \circ \boldsymbol{P}^\top$ where $\boldsymbol{P}$ is a permutation and $\boldsymbol{c}$ is a $\boldsymbol{G}^a$-preserving diffeomorphism. Using Proposition 3, this means that the dependency graph of $\boldsymbol{c}$ is given by

$$\boldsymbol{C} = \begin{bmatrix} * & * & * \\ 0 & * & 0 \\ * & * & * \end{bmatrix}, \text{ since } \boldsymbol{G}^a_{2,\cdot} \not\subseteq \boldsymbol{G}^a_{1,\cdot} \text{ and } \boldsymbol{G}^a_{2,\cdot} \not\subseteq \boldsymbol{G}^a_{3,\cdot},$$

where "$*$" indicates a potentially nonzero value. This means that one of the components of the learned representation will be an invertible transformation of the ground-truth variable $\boldsymbol{z}_2$ (robot position), while the other components could be a mixture of $\boldsymbol{z}_1$, $\boldsymbol{z}_2$ and $\boldsymbol{z}_3$. Figure 3b shows both the graph $\boldsymbol{G}^a$ and the corresponding entanglement graph $\boldsymbol{V}$ assuming $\boldsymbol{P} = \boldsymbol{I}$. Example 9 will make this example more concrete by explicitly specifying a latent model that satisfies the sufficient influence assumption of Theorem 1.

Example 5 (Complete disentanglement from multi-target actions) *Assume $d_z = 3$ and $d_a = 3$ where $\boldsymbol{G}^a \in \mathbb{R}^{d_z \times d_a}$ is given by Figure 3c and the temporal graph $\boldsymbol{G}^z$ could be anything (see Remark 2 above). If the sufficient influence assumption of Theorem 1 or 2 is satisfied, then we have that $\boldsymbol{\theta} \sim_{\text{obs}} \hat{\boldsymbol{\theta}}$ & $||\hat{\boldsymbol{G}}^a||_0 \leq ||\boldsymbol{G}^a||_0$ implies $\boldsymbol{v} = \boldsymbol{c} \circ \boldsymbol{P}^\top$ where $\boldsymbol{P}$ is a permutation and $\boldsymbol{c}$ is a $\boldsymbol{G}^a$-preserving diffeomorphism. Proposition 3 implies that the dependency graph of $\boldsymbol{c}$ is simply $\boldsymbol{C} := \boldsymbol{I}$ because $\boldsymbol{G}^a_{i,\cdot} \not\subseteq \boldsymbol{G}^a_{j,\cdot}$ for all distinct i, j. This means that we have complete disentanglement (Definition 7). Examples 10, 12 and 13 will explore more concrete instantiations of this example by specifying concrete latent models satisfying the sufficient influence assumptions of Theorems 1 and 2.*



(a) Example 3 (b) Example 4 (c) Example 5 (d) Example 6 (e) Example 7 (f) Example 8

Figure 3: Graphs $\boldsymbol{G}^a$ and $\boldsymbol{G}^z$ from Examples 3, 4, 5, 6, 7 & 8 with their respective entanglement graphs $\boldsymbol{V}$ (Definition 3) guaranteed by Theorems 1, 2, 3 & 5 (assuming $\boldsymbol{P} = \boldsymbol{I}$ for simplicity). Recall, that $\boldsymbol{V}$ describes the dependency structure of $\boldsymbol{v} = \boldsymbol{f}^{-1} \circ \hat{\boldsymbol{f}}$, which maps $\hat{\boldsymbol{z}}$ to $\boldsymbol{z}$. By Remark 1, the functional dependency graph of $\boldsymbol{v}^{-1}$ is exactly the same except for $\boldsymbol{z}$ and $\hat{\boldsymbol{z}}$ being interchanged.

3.4.1 UNKNOWN-TARGET INTERVENTIONS ON THE LATENT FACTORS

An important special case of Theorem 2 is when $\boldsymbol{a}^{t-1}$ corresponds to a one-hot vector indexing an *intervention with unknown targets* on the latent variables $\boldsymbol{z}^t$. This specific kind of intervention has been explored previously in the context of causal discovery where the intervention occurs on *observed* variables instead of *latent* variables (Eaton and Murphy, 2007; Mooij et al., 2020; Squires et al., 2020; Jaber et al., 2020; Brouillard et al., 2020; Ke et al., 2019). Recently, multiple works in causal representation learning have considered interventions on latent variables (Lachapelle et al., 2022; Lippe et al., 2023b; Ahuja et al., 2023; Squires et al., 2023; Buchholz et al., 2023; von Kügelgen et al., 2023; Zhang et al., 2023; Jiang and Aragam, 2023) (see Section 7 for more). Here is how our framework can accommodate such interventions: Assume $\boldsymbol{a}^{t-1} \in \{\vec{0}, \boldsymbol{e}_1, ..., \boldsymbol{e}_{d_a}\}$, where each $\boldsymbol{e}_\ell$ is a one-hot vector. The action $\boldsymbol{a}^{t-1} = \vec{0}$ corresponds to the *observational setting*, i.e., when no intervention occurred, while $\boldsymbol{a}^{t-1} = \boldsymbol{e}_\ell$ corresponds to the ℓth intervention. In that context, the unknown graph $\boldsymbol{G}^a$ describes which latents are targeted by the intervention, i.e. $\ell \in \mathbf{Pa}_i^a$ if and only if $\boldsymbol{z}_i$ is targeted by the ℓth intervention. To see this, recall that, under Assumption 3, we have

$$p(\boldsymbol{z}_i^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = p(\boldsymbol{z}_i^t \mid \boldsymbol{z}^{<t}_{\mathbf{Pa}_i^z}, \boldsymbol{a}^{t-1}_{\mathbf{Pa}_i^a}) \,,$$

where we implicitly assumed that $p(\boldsymbol{z}_i^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ does not depend on $\boldsymbol{a}^{<t-1}$. In the observational setting, i.e., when $\boldsymbol{a}^{t-1} = \vec{0}$, the conditional becomes $p(\boldsymbol{z}_i^t \mid \boldsymbol{z}^{<t}, \vec{0})$. Now suppose that we are in the ℓth intervention, i.e., $\boldsymbol{a}^{t-1} = \boldsymbol{e}_\ell$. Then, if $\ell \notin \mathbf{Pa}_i^a$, we have that $\boldsymbol{a}^{t-1}_{\mathbf{Pa}_i^a} = \vec{0}$, which means that the conditional is also $p(\boldsymbol{z}_i^t \mid \boldsymbol{z}^{<t}, \vec{0})$, meaning that the variable $\boldsymbol{z}_i^t$ *is not* targeted by the ℓth intervention. When $\ell \in \mathbf{Pa}_i^a$, we have $\boldsymbol{a}^{t-1}_{\mathbf{Pa}_i^a} \neq \vec{0}$ and thus the conditional is allowed to change freely, i.e., $\boldsymbol{z}_i^t$ *is* targeted by the ℓth intervention.

Importantly, the assumption that $\boldsymbol{G}^a$ is sparse corresponds precisely to the *sparse mechanism shift* hypothesis from Schölkopf et al. (2021), i.e. that *only a few mechanisms change at a time.* Thm. 2 thus provides precise conditions for when sparse mechanism shifts induce disentanglement. Interestingly, our theory covers both hard and soft interventions, as long as the sufficient influence assumption is satisfied.

Remark 3 (Examples revisited) *Examples 3, 4 and 5 can be revisited while keeping in mind the "unknown-target intervention interpretation" in which $\boldsymbol{G}^a$ describes which latent variable is targeted by each intervention. For example, Example 3 tells us that if each latent variable is targeted by a single-node intervention, then complete disentanglement is guaranteed. Examples 11, 12 and 13 provide mathematically concrete latent models where $\boldsymbol{a}$ is interpreted to be an intervention.*

Remark 4 (Causal representation learning without temporal dependencies (static)) *The special case where $T = 1$, i.e., no temporal dependencies, is of special interest. In that case, the latent variable model is simply $p(\boldsymbol{z} \mid \boldsymbol{a}) = \prod_{i=1}^{d_z} p(\boldsymbol{z}_i \mid \boldsymbol{a})$. In other words, the causal graph that relates the latent factors $\boldsymbol{z}_i$ is empty. In contrast, recent work on learning causal representation showed how to obtain disentanglement in general latent causal graphical models without temporal dependencies, but are limited to single-node interventions (Ahuja et al., 2023; Squires et al., 2023; Buchholz et al., 2023; von Kügelgen et al., 2023; Zhang et al., 2023; Jiang and Aragam, 2023). These works can be roughly thought of as fitting our interventional setting described in this section where (i) $T = 1$; (ii) conditional independence (Assumption 1) is dropped (allowing instantaneous causal effects); and (iii) only single-node interventions are allowed, corresponding to a very sparse $\boldsymbol{G}^a$ with columns containing at most one non-zero element, like in Example 3. Although our framework with $T = 1$ assumes that the causal graph between latent variables is empty, it allows for multi-node interventions which are sometimes sufficient to disentangle (Example 13). See Section 3.9.2 for more on this.*

3.5 Nonparametric Identifiability via Sparse Temporal Dependencies

This section is analogous to the previous one, but instead of leveraging the sparsity of $\boldsymbol{G}^a$ to show identifiability, it leverages the sparsity of $\boldsymbol{G}^z$, which describes the structure of the dependencies between the latents from one time step to another. We will see that, under some assumptions, regularizing the learned graph $\hat{\boldsymbol{G}}^z$ to be sparse will allow identifiability up to the following equivalence class:

Definition 14 ($\boldsymbol{z}$-consistency equivalence) *We say two models $\boldsymbol{\theta} := (\boldsymbol{f}, p, \boldsymbol{G})$ and $\tilde{\boldsymbol{\theta}} := (\tilde{\boldsymbol{f}}, \tilde{p}, \tilde{\boldsymbol{G}})$ satisfying Assumptions 1, 2 & 3 are **$\boldsymbol{z}$-consistent**, denoted $\boldsymbol{\theta} \sim^{\boldsymbol{z}}_{\text{con}} \tilde{\boldsymbol{\theta}}$, if and only if there exists a permutation matrix $\boldsymbol{P}$ such that*

1. *$\boldsymbol{\theta} \sim_{\text{diff}} \tilde{\boldsymbol{\theta}}$ (Def. 5) and $\tilde{\boldsymbol{G}}^z = \boldsymbol{P}\boldsymbol{G}^z\boldsymbol{P}^\top$; and*
2. *the entanglement map $\boldsymbol{v} := \boldsymbol{f}^{-1} \circ \tilde{\boldsymbol{f}}$ can be written as $\boldsymbol{v} = \boldsymbol{c} \circ \boldsymbol{P}^\top$ where $\boldsymbol{c}$ is a $\boldsymbol{G}^z$-preserving and $(\boldsymbol{G}^z)^\top$-preserving diffeomorphism (Definition 12).*

This relation can be shown to be an *equivalence* relation, as was the case for $\sim^{\boldsymbol{a}}_{\text{con}}$. This is shown in Appendix A.4. Analogously to $\sim^{\boldsymbol{a}}_{\text{con}}$, the equivalence relation $\sim^{\boldsymbol{z}}_{\text{con}}$ relates the structure of the entanglement map $\boldsymbol{v}$ to the graph $\boldsymbol{G}^z$ via the notion of $\boldsymbol{G}$-preserving maps. It is also true that $\boldsymbol{\theta} \sim_{\text{perm}} \hat{\boldsymbol{\theta}} \implies \boldsymbol{\theta} \sim^{\boldsymbol{z}}_{\text{con}} \hat{\boldsymbol{\theta}}$.

The following result is analogous to Theorems 1 and 2 where, instead of regularizing $\hat{\boldsymbol{G}}^a$ to be sparse, we regularize $\hat{\boldsymbol{G}}^z$. The next theorem shows how this type of sparsity regularization can induce the learned model to be $\boldsymbol{z}$-consistent with the ground-truth one.

Theorem 3 (Nonparametric disentanglement via sparse temporal dependencies) *Let the parameters $\boldsymbol{\theta} := (\boldsymbol{f}, p, \boldsymbol{G})$ and $\hat{\boldsymbol{\theta}} := (\hat{\boldsymbol{f}}, \hat{p}, \hat{\boldsymbol{G}})$ correspond to two models satisfying Assumptions 1, 2, 3 & 4. Further assume that*

1. ***[Observational equivalence]*** *$\boldsymbol{\theta} \sim_{\text{obs}} \hat{\boldsymbol{\theta}}$ (Def. 4);*
2. ***[Sufficient influence of z]*** *The Hessian matrix $H^{t,\tau}_{z,z} \log p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ varies "sufficiently", as formalized in Assumption 8;*

Then, there exists a permutation matrix $\boldsymbol{P}$ such that $\boldsymbol{P}\boldsymbol{G}^z\boldsymbol{P}^\top \subseteq \hat{\boldsymbol{G}}^z$. Further assume that

3. ***[Sparsity regularization]*** *$||\hat{\boldsymbol{G}}^z||_0 \leq ||\boldsymbol{G}^z||_0$;*

Then, $\boldsymbol{\theta} \sim^{\boldsymbol{z}}_{\text{con}} \hat{\boldsymbol{\theta}}$ (Def. 14).

The structure of the above theorem is very similar to Theorem 1 & 2. For example, we still have a "sufficient influence" condition, but this time it concerns the Hessian matrix $H^{t,\tau}_{z,z} \log p$ which we saw in Section 3.2, Equation (15). The conclusion is that both models will be $\boldsymbol{z}$-consistent, which means that we recover the graph $\boldsymbol{G}^z$ up to permutation and have that the entanglement map $\boldsymbol{v}$ has a dependency graph given by $\boldsymbol{V} = \boldsymbol{C}\boldsymbol{P}^\top$ where $\boldsymbol{C}$ is $\boldsymbol{G}^z$- and $(\boldsymbol{G}^z)^\top$-preserving. Section 3.8 introduces the sufficient influence assumption formally as well as a proof of Theorem 3.

We now build intuition through some minimal examples that show how one can apply the above theorem to draw links between the graph $\boldsymbol{G}^z$ and the resulting entanglement graph $\boldsymbol{V}$ (Definition 3). For now we simply assume that the assumption of sufficient influence (Assumption 8) is satisfied and wait until Section 3.9.3 to present more concrete transition models that satisfy it.

Example 6 (Disentanglement via independent factors with temporal dependencies) *Consider the situation depicted in Figure 3d where the graph $\boldsymbol{G}^z = \boldsymbol{I}$, i.e., the latents $\boldsymbol{z}^t_i$ are dependent in time but independent across dimensions. For this example, actions are unnecessary. Assuming the sufficient influence assumption of Theorem 3 is satisfied, we have $\boldsymbol{\theta} \sim_{\text{obs}} \hat{\boldsymbol{\theta}}$ & $||\hat{\boldsymbol{G}}^z||_0 \leq ||\boldsymbol{G}^z||_0 \implies \boldsymbol{\theta} \sim^{\boldsymbol{z}}_{\text{con}} \hat{\boldsymbol{\theta}}$, which means there exists a permutation $\boldsymbol{P}$ such that $\hat{\boldsymbol{G}}^z = \boldsymbol{P}\boldsymbol{G}^z\boldsymbol{P}^\top$ and such that the entanglement map is given by $\boldsymbol{v} = \boldsymbol{c} \circ \boldsymbol{P}^\top$ where $\boldsymbol{c}$ is $\boldsymbol{G}^z$- and $(\boldsymbol{G}^z)^\top$-preserving. Using Proposition 3, one can verify that the dependency graph of $\boldsymbol{c}$ is $\boldsymbol{C} = \boldsymbol{I}$ and thus $\boldsymbol{V} = \boldsymbol{P}^\top$, i.e. the learned representation is completely disentangled. Example 14 will provide a concrete transition model where the sufficient influence assumption of Theorem 3 holds for this simple graph $\boldsymbol{G}^z$.*

Example 7 (Disentanglement via sparsely dependent factors with temporal dependencies) *The previous examples assumed independent latents, i.e., $\boldsymbol{G}^z = \boldsymbol{I}$. Instead, we now consider a more interesting "lower triangular" graph $\boldsymbol{G}^z$, as depicted in Figures 3e (This is the same graph as in the tree-robot-ball example in Figure 1). Again using Proposition 3, one can verify that $\boldsymbol{C} = \boldsymbol{I}$ and thus $\boldsymbol{V} = \boldsymbol{P}^\top$, i.e. the learned representation is completely disentangled. Example 14 will provide a concrete transition model where the sufficient influence assumption of Theorem 3 holds.*

Example 8 (Partial disentanglement via temporal sparsity) *Assume the same situation as previously, but add an additional edge from $\boldsymbol{z}_B^{t-1}$ to $\boldsymbol{z}_R^t$ (see Figure 3f). This could occur, for example, if the robot tries to follow the ball and is thus influenced by it. Using Proposition 3, one can show that $\boldsymbol{c}$ being $\boldsymbol{G}^z$- and $(\boldsymbol{G}^z)^\top$-preserving means that its dependency graph is given by*

$$\boldsymbol{C} = \begin{bmatrix} * & 0 & 0 \\ 0 & * & * \\ 0 & * & * \end{bmatrix}.$$

This means the robot and the ball remain entangled in the learned representation.

3.6 Combining Sparsity Regularization on $\hat{\boldsymbol{G}}^a$ & $\hat{\boldsymbol{G}}^z$

A natural question at this point is whether Theorem 1 (or Theorem 2) can be combined with Theorem 3 to obtain stronger guarantees. The answer is yes. In this section, we explain how this can be done. We would like to show how the combination of the assumptions of Theorem 1 and Theorem 3 can yield identifiability up to the following stronger equivalence relation.

Definition 15 ($(\boldsymbol{a}, \boldsymbol{z})$-consistency equivalence) *We say two models $\boldsymbol{\theta} := (\boldsymbol{f}, p, \boldsymbol{G})$ and $\tilde{\boldsymbol{\theta}} := (\tilde{\boldsymbol{f}}, \tilde{p}, \tilde{\boldsymbol{G}})$ satisfying Assumptions 1, 2 & 3 are $(\boldsymbol{a}, \boldsymbol{z})$-**consistent**, denoted $\boldsymbol{\theta} \sim_{\text{con}}^{\boldsymbol{z},\boldsymbol{a}} \tilde{\boldsymbol{\theta}}$, if and only if there exists a permutation matrix $\boldsymbol{P}$ such that*

1. *$\boldsymbol{\theta} \sim_{\text{diff}} \tilde{\boldsymbol{\theta}}$ (Def. 5) and $\tilde{\boldsymbol{G}}^a = \boldsymbol{P}^\top \boldsymbol{G}^a$ and $\tilde{\boldsymbol{G}}^z = \boldsymbol{P}^\top \boldsymbol{G}^z \boldsymbol{P}$; and*
2. *the entanglement map $\boldsymbol{v} := \boldsymbol{f}^{-1} \circ \tilde{\boldsymbol{f}}$ can be written as $\boldsymbol{v} = \boldsymbol{c} \circ \boldsymbol{P}^\top$ where $\boldsymbol{c}$ is a $\boldsymbol{G}^a$-, $\boldsymbol{G}^z$- and $(\boldsymbol{G}^z)^\top$-preserving diffeomorphism (Def. 12).*

Of course, if the assumptions of both theorems hold, we must have $\boldsymbol{\theta} \sim_{\text{con}}^{\boldsymbol{a}} \hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta} \sim_{\text{con}}^{\boldsymbol{z}} \hat{\boldsymbol{\theta}}$. As one might guess, this implies $\boldsymbol{\theta} \sim_{\text{con}}^{\boldsymbol{a},\boldsymbol{z}} \hat{\boldsymbol{\theta}}$, as the following proposition shows. The reason why this result is not completely trivial is that the permutations $\boldsymbol{P}$ given by $\sim_{\text{con}}^{\boldsymbol{a}}$ and $\sim_{\text{con}}^{\boldsymbol{z}}$ might not be the same. Its proof can be found in Appendix A.4.1.

Proposition 6 *Let $\boldsymbol{\theta} := (\boldsymbol{f}, p, \boldsymbol{G})$ and $\tilde{\boldsymbol{\theta}} := (\tilde{\boldsymbol{f}}, \tilde{p}, \tilde{\boldsymbol{G}})$ be two models satisfying Assumptions 1, 2 & 3. We have $\boldsymbol{\theta} \sim_{\text{con}}^{\boldsymbol{z},\boldsymbol{a}} \tilde{\boldsymbol{\theta}}$ if and only if $\boldsymbol{\theta} \sim_{\text{con}}^{\boldsymbol{a}} \tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta} \sim_{\text{con}}^{\boldsymbol{z}} \tilde{\boldsymbol{\theta}}$.*

We can thus combine both Theorems 1 (or Theorem 2) with Theorem 3 to obtain stronger guarantees. Practically, this means that regularizing both $\hat{\boldsymbol{G}}^a$ and $\hat{\boldsymbol{G}}^z$ to be sparse will lead to a more disentangled representation, i.e. a sparser entanglement graph $\boldsymbol{V}$, than if regularization was applied only on $\hat{\boldsymbol{G}}^a$ or only on $\hat{\boldsymbol{G}}^z$.

3.7 Graphical Criterion for Complete Disentanglement

The previous sections introduced results guaranteeing identifiability up to $\sim_{\text{con}}^{\boldsymbol{a}}$, $\sim_{\text{con}}^{\boldsymbol{z}}$ and $\sim_{\text{con}}^{\boldsymbol{z},\boldsymbol{a}}$, which all correspond to potentially *partial* disentanglement. This section provides an additional assumption to guarantee identifiability up to $\sim_{\text{perm}}$, i.e., *complete* disentanglement.

One can easily see from the definitions that $\boldsymbol{\theta} \sim_{\text{perm}} \hat{\boldsymbol{\theta}}$ holds precisely when $\boldsymbol{\theta} \sim_{\text{con}}^{\boldsymbol{a},\boldsymbol{z}} \hat{\boldsymbol{\theta}}$ with $\boldsymbol{C} = \boldsymbol{I}$. This condition can be achieved by making an additional assumption on $\boldsymbol{G}$. This assumption is taken directly from Lachapelle et al. (2022).

Assumption 5 (Graphical criterion, Lachapelle et al. (2022)) *Let $\boldsymbol{G} = [\boldsymbol{G}^z \; \boldsymbol{G}^a]$ be a graph. For all $i \in \{1, ..., d_z\}$,*

$$\left(\bigcap_{j \in \mathbf{Ch}_i^z} \mathbf{Pa}_j^z\right) \cap \left(\bigcap_{j \in \mathbf{Pa}_i^z} \mathbf{Ch}_j^z\right) \cap \left(\bigcap_{\ell \in \mathbf{Pa}_i^a} \mathbf{Ch}_\ell^a\right) = \{i\}\,,$$

where $\mathbf{Pa}_i^z$ and $\mathbf{Ch}_i^z$ are the sets of parents and children of node $\boldsymbol{z}_i$ in $\boldsymbol{G}^z$, respectively, while $\mathbf{Ch}_\ell^a$ is the set of children of $\boldsymbol{a}_\ell$ in $\boldsymbol{G}^a$.

The following proposition shows that when $\boldsymbol{G}$ satisfies the above criterion, the set of models that are $\sim_{\text{con}}^{\boldsymbol{a},\boldsymbol{z}}$-equivalent to $\boldsymbol{\theta}$ is equal to the set of models that are $\sim_{\text{perm}}$-equivalent to $\boldsymbol{\theta}$, thus allowing complete disentanglement. See Appendix A.6 for a proof.

Proposition 7 (Complete disentanglement as a special case) *Let $\boldsymbol{\theta} := (\boldsymbol{f}, p, \boldsymbol{G})$ and $\hat{\boldsymbol{\theta}} := (\hat{\boldsymbol{f}}, \hat{p}, \hat{\boldsymbol{G}})$ be two models satisfying Assumptions 1, 2 & 3. If $\boldsymbol{\theta} \sim_{\text{con}}^{\boldsymbol{z},\boldsymbol{a}} \hat{\boldsymbol{\theta}}$ and $\boldsymbol{G}$ satisfies Assumption 5, then $\boldsymbol{\theta} \sim_{\text{perm}} \hat{\boldsymbol{\theta}}$.*

The above result shows that our general theory can guarantee complete disentanglement as a special case. This is one way in which our work generalizes the work of Lachapelle et al. (2022), in addition to relaxing the exponential family assumption. The following section explores how the exponential family assumption fits into our nonparametric theory and how it allows one to simplify the "sufficient influence assumptions". But before, we provide some example to illustrate when Assumption 5 holds.

For example, the graphical criterion of Assumption 5 is trivially satisfied when $\boldsymbol{G}^z$ is diagonal, since $\{i\} = \mathbf{Pa}_i^{\boldsymbol{z}}$ for all i (actions are not necessary here). This simple case amounts to having mutual independence between the sequences $\boldsymbol{z}_i^{\leq T}$, which is a standard assumption in the ICA literature (Tong et al., 1990; Hyvarinen and Morioka, 2017; Klindt et al., 2021). The illustrative example we introduced in Fig. 1 has a more interesting "non-diagonal" graph satisfying our criterion. Indeed, we have $\{T\} = \mathbf{Pa}_T^{\boldsymbol{z}}$, $\{R\} = \mathbf{Ch}_R^{\boldsymbol{z}} \cap \mathbf{Pa}_R^{\boldsymbol{z}}$, and $\{B\} = \mathbf{Ch}_B^{\boldsymbol{z}}$. This example is actually part of an interesting family of graphs that satisfy our criterion:

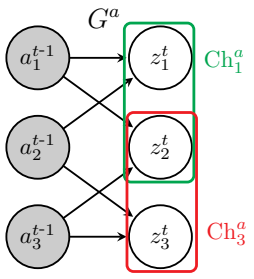


Figure 4: An example satisfying Assumption 5. Indeed, $\{\boldsymbol{z}_1\} = \mathbf{Ch}_1^a \cap \mathbf{Ch}_2^a$, $\{\boldsymbol{z}_2\} = \mathbf{Ch}_1^a \cap \mathbf{Ch}_3^a$ and $\{\boldsymbol{z}_3\} = \mathbf{Ch}_2^a \cap \mathbf{Ch}_3^a$.

Proposition 8 (Sufficient condition for the graphical criterion) *If $\boldsymbol{G}_{i,i}^{\boldsymbol{z}} = 1$ for all i (all nodes have a self-loop) and $\boldsymbol{G}^z$ has no 2-cycles, then $\boldsymbol{G}$ satisfies Assumption 5.*

Proof Self-loops guarantee $i \in \mathbf{Pa}_i^{\boldsymbol{z}} \cap \mathbf{Ch}_i^{\boldsymbol{z}}$ for all i. Suppose $j \in \mathbf{Pa}_i^{\boldsymbol{z}} \cap \mathbf{Ch}_i^{\boldsymbol{z}}$ for some $i \neq j$. This implies that i and j form a 2-cycle, which is a contradiction. Thus, $\{i\} = \mathbf{Pa}_i^{\boldsymbol{z}} \cap \mathbf{Ch}_i^{\boldsymbol{z}}$ for all i. ■

3.8 Proofs of Theorems 1, 2 & 3 and their Sufficient Influence Assumptions

In this section, we introduce the sufficient influence assumptions and use them to prove Theorems 1, 2 & 3. In the next section (Section 3.9), we provide multiple examples to gain intuition about the sufficient influence assumptions. Throughout, the following lemma will come in handy.

Lemma 2 (Invertible matrix contains a permutation) *Let $\boldsymbol{L} \in \mathbb{R}^{m \times m}$ be an invertible matrix. Then, there exists a permutation σ such that $\boldsymbol{L}_{i,\sigma(i)} \neq 0$ for all i, or, in other words, $\boldsymbol{P}^\top \subseteq \boldsymbol{L}$ where $\boldsymbol{P}$ is the permutation matrix associated with σ, i.e., $\boldsymbol{P}\boldsymbol{e}_i = \boldsymbol{e}_{\sigma(i)}$. Note that this implies that $\boldsymbol{PL}$ and $\boldsymbol{LP}$ have no zero on their diagonals.*

Proof Since the matrix $\boldsymbol{L}$ is invertible, its determinant is non-zero, i.e.

$$\det(\boldsymbol{L}) := \sum_{\sigma \in \mathfrak{S}_m} \operatorname{sign}(\sigma) \prod_{i=1}^{m} \boldsymbol{L}_{i,\sigma(i)} \neq 0\,,$$

where $\mathfrak{S}_m$ is the set of m-permutations. This equation implies that at least one term of the sum is non-zero, meaning there exists a permutation σ such that, for all i, $\boldsymbol{L}_{i,\sigma(i)} \neq 0$. ■

3.8.1 SUFFICIENT INFLUENCE ASSUMPTION OF THEOREM 1 AND ITS PROOF

We start by introducing the sufficient influence assumption of Theorem 1. Although it may seem intimidating at first, the reason why it is necessary will become clear when we prove the theorem.

Assumption 6 (Sufficient influence of $\boldsymbol{a}$ (nonparametric/continuous)) *For almost all $\boldsymbol{z} \in \mathbb{R}^{d_z}$ (i.e. except on a set with zero Lebesgue measure) and all $\ell \in [d_a]$, there exists*

$$\{(t_{(r)}, \tau_{(r)}, \boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)})\}_{r=1}^{|\mathbf{Ch}_\ell^{\boldsymbol{a}}|}\,,$$

such that $t_{(r)} \in [T]$, $\tau_{(r)} < t_{(r)}$, $\boldsymbol{z}_{(r)} \in \mathbb{R}^{d_z \times (t_{(r)}-1)}$, $\boldsymbol{a}_{(r)} \in \mathcal{A}^{t_{(r)}}$ and

$$\operatorname{span}\left\{ H_{z,a}^{t_{(r)},\tau_{(r)}} \log p(\boldsymbol{z} \mid \boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)})_{\cdot,\ell} \right\}_{r=1}^{|\mathbf{Ch}_\ell^{\boldsymbol{a}}|} = \mathbb{R}^{d_z}_{\mathbf{Ch}_\ell^{\boldsymbol{a}}}\,.$$

Proof [of Theorem 1] Recall equation (13), which we derived in Section 3.2:

$$\underbrace{H_{z,a}^{t,\tau} \hat{q}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})}_{\subseteq \hat{\boldsymbol{G}}^a} = D\boldsymbol{v}(\boldsymbol{z}^t)^\top \underbrace{H_{z,a}^{t,\tau} q(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t})}_{\subseteq \boldsymbol{G}^a}\,. \tag{16}$$

Notice that Assumption 6 holds only "almost everywhere", i.e. on a set $\mathbb{R}^{d_z} \setminus E_0$ where E_0 has zero Lebesgue measure. Fix an arbitrary $\boldsymbol{z} \in \mathbb{R}^{d_z} \setminus E_0$. For notational convenience, define

$$\Lambda(\boldsymbol{z}, \gamma) := H_{z,a}^{t,\tau} q(\boldsymbol{v}(\boldsymbol{z}) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t}) \qquad \hat{\Lambda}(\boldsymbol{z}, \gamma) := H_{z,a}^{t,\tau} \hat{q}(\boldsymbol{z} \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})\,,$$

where $\gamma := (t, \tau, \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$. This allows us to rewrite (16) with a much lighter notation:

$$\hat{\Lambda}(\boldsymbol{z}, \gamma) = D\boldsymbol{v}(\boldsymbol{z})^\top \Lambda(\boldsymbol{z}, \gamma)\,. \tag{17}$$

Now, notice that the sufficient influence assumption (Assumption 6) requires that, for all $\ell \in [d_a]$ there exists $\{\gamma_{(r)}\}_{r=1}^{|\mathbf{Ch}_\ell^a|}$ such that $\operatorname{span}\{\Lambda(\boldsymbol{z}, \gamma_{(r)})_{\cdot,\ell}\}_{r=1}^{|\mathbf{Ch}_\ell^a|} = \mathbb{R}^{d_z}_{\mathbf{Ch}_\ell^a}$. We can thus write

$$D\boldsymbol{v}(\boldsymbol{z})^\top \mathbb{R}^{d_z}_{\boldsymbol{G}^a_{\cdot,\ell}} = D\boldsymbol{v}(\boldsymbol{z})^\top \operatorname{span}\{\Lambda(\boldsymbol{z}, \gamma_{(r)})_{\cdot,\ell}\}_{r=1}^{|\mathbf{Ch}_\ell^a|} = \operatorname{span}\{\hat{\Lambda}(\boldsymbol{z}, \gamma_{(r)})_{\cdot,\ell}\}_{r=1}^{|\mathbf{Ch}_\ell^a|} \subseteq \mathbb{R}^{d_z}_{\hat{\boldsymbol{G}}^a_{\cdot,\ell}} \tag{18}$$

Since $D\boldsymbol{v}(\boldsymbol{z})$ is invertible, there exists a permutation $\boldsymbol{P}(\boldsymbol{z})$ such that $D\boldsymbol{v}(\boldsymbol{z})\boldsymbol{P}(\boldsymbol{z})$ has no zero on its diagonal (Lemma 2). Let $\boldsymbol{C}(\boldsymbol{z}) := D\boldsymbol{v}(\boldsymbol{z})\boldsymbol{P}(\boldsymbol{z})$. By left-multiplying (18) by $\boldsymbol{P}(\boldsymbol{z})^\top$, we get

$$\boldsymbol{C}(\boldsymbol{z})^\top \mathbb{R}^{d_z}_{\boldsymbol{G}^a_{\cdot,\ell}} \subseteq \mathbb{R}^{d_z}_{\boldsymbol{P}(\boldsymbol{z})^\top \hat{\boldsymbol{G}}^a_{\cdot,\ell}}. \tag{19}$$

We would like to show that $\boldsymbol{C}(\boldsymbol{z})$ is $\boldsymbol{G}^a$-preserving. Notice how the above equation is almost exactly the definition of $\boldsymbol{G}^a$-preserving. All that is left to prove is that $\boldsymbol{P}(\boldsymbol{z})^\top \hat{\boldsymbol{G}}^a = \boldsymbol{G}^a$.

We start by showing $\boldsymbol{P}(\boldsymbol{z})^\top \hat{\boldsymbol{G}}^a \supseteq \boldsymbol{G}^a$. Take $(i,\ell) \in \boldsymbol{G}^a$. Since $\boldsymbol{e}_i \in \mathbb{R}^{d_z}_{\boldsymbol{G}^a_{\cdot,\ell}}$, equation (19) implies

$$\boldsymbol{C}(\boldsymbol{z})^\top \boldsymbol{e}_i = \boldsymbol{C}(\boldsymbol{z})_{i,\cdot} \in \mathbb{R}^{d_z}_{\boldsymbol{P}(\boldsymbol{z})^\top \hat{\boldsymbol{G}}^a_{\cdot,\ell}}.$$

Since $\boldsymbol{C}(\boldsymbol{z})_{i,i} \neq 0$ (all elements on its diagonal are nonzero), we must have that $(i,\ell) \in \boldsymbol{P}(\boldsymbol{z})^\top \hat{\boldsymbol{G}}^a$.

Now, since $||\boldsymbol{P}(\boldsymbol{z})^\top \hat{\boldsymbol{G}}^a||_0 = ||\hat{\boldsymbol{G}}^a||_0 \leq ||\boldsymbol{G}^a||_0$, we have $\boldsymbol{P}(\boldsymbol{z})^\top \hat{\boldsymbol{G}}^a = \boldsymbol{G}^a$. This implies

$$\boldsymbol{C}(\boldsymbol{z})^\top \mathbb{R}^{d_z \times d_a}_{\boldsymbol{G}^a} \subseteq \mathbb{R}^{d_z \times d_a}_{\boldsymbol{G}^a},$$

i.e. $\boldsymbol{C}(\boldsymbol{z})$ is a $\boldsymbol{G}^a$-preserving matrix, as desired.

To recap, we now have that, for all $\boldsymbol{z} \in \mathbb{R}^{d_z} \setminus E_0$, there exists a permutation $\boldsymbol{P}(\boldsymbol{z})$ s.t. $D\boldsymbol{v}(\boldsymbol{z})\boldsymbol{P}(\boldsymbol{z})$ is $\boldsymbol{G}^a$-preserving. We are not done yet, since, a priori, the permutation $\boldsymbol{P}(\boldsymbol{z})$ can be different for different values of $\boldsymbol{z}$, and we do not know what happens on the measure-zero set E_0. What we need to show is that there exists a permutation $\boldsymbol{P}$ such that, for all $\boldsymbol{z}$, $D\boldsymbol{v}(\boldsymbol{z})\boldsymbol{P}$ is $\boldsymbol{G}^a$-preserving. Lemma 12 in Appendix A.5 shows precisely this, by leveraging the continuity of $D\boldsymbol{v}(\boldsymbol{z})$ ($\boldsymbol{v}$ is a diffeomorphism and thus C^1).

Notice that $D(\boldsymbol{v} \circ \boldsymbol{P})(\boldsymbol{z}) = D\boldsymbol{v}(\boldsymbol{P}\boldsymbol{z})\boldsymbol{P}$, which is $\boldsymbol{G}^a$-preserving everywhere. Using Lemma 1, we conclude that the function $\boldsymbol{c} := \boldsymbol{v} \circ \boldsymbol{P}$ is $\boldsymbol{G}^a$-preserving. This concludes the proof. ■

Remark 5 (Alternative view on sufficient influence assumptions) *Assumption 6, and all sufficient influence assumptions we present later on, can be thought of in terms of linear independence of functions. By definition, a family of functions $(f^{(i)} : X \to \mathbb{R})_{i=1}^n$ is linearly independent when $\sum_i \alpha_i f^{(i)}(x) = 0$ for all $x \in X$ implies $\alpha_i = 0$ for all i. It turns out that Assumption 6 is equivalent to requiring that, for all $\boldsymbol{z} \in \mathbb{R}^{d_z}$ and $\ell \in [d_a]$, the family of functions $(H^{t,\tau}_{z,a} \log p(\boldsymbol{z} \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})_{i,\ell})_{i \in \mathbf{Ch}^a_\ell}$ (seen as functions of $t, \tau, \boldsymbol{z}^{<t}$ and $\boldsymbol{a}^{<t}$) is linearly independent. To see this, note that, in general, $(f^{(i)} : X \to \mathbb{R})_{i=1}^n$ is linearly independent iff there exist $x_1, ..., x_n \in X$ s.t. the vectors $((f^{(1)}(x_i), ..., f^{(n)}(x_i)))_{i=1}^n$ are linearly independent (see Appendix A.1 for a proof).*

3.8.2 SUFFICIENT INFLUENCE ASSUMPTION OF THEOREM 2 AND ITS PROOF

One can see that, if $\boldsymbol{a}$ is discrete, Theorem 1 cannot be applied because its sufficient influence assumption (Assumption 6) refers to the cross derivative of $\log p$ w.r.t. $\boldsymbol{z}^t$ and $\boldsymbol{a}^\tau$, which, of course, is not well defined when $\boldsymbol{a}$ is discrete. The discrete case is important to discuss interventions with unknown-targets as we did in Section 3.4.1, which is why we have a specialized result (Theorem 2) which has an analogous sufficient influence assumption based on *partial differences*.

Definition 16 (Partial difference) *Let us define*

$$\Delta^{\tau,\epsilon}_{a,\ell} D^t_z \log p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) := D^t_z \log p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t} + \epsilon \boldsymbol{E}^{(\ell,\tau)}) - D^t_z \log p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) \,,$$

where $\epsilon \in \mathbb{R}$ and $\boldsymbol{E}^{(\ell,\tau)}$ is a matrix with a one at entry (ℓ, τ) and zeros everywhere else.

One can see that $\Delta^{\tau,\epsilon}_{a,\ell} D^t_z \log p$ is essentially the discrete analog of $(H^{t,\tau}_{z,a} \log p)_{\cdot,\ell}$. Apart from this difference, the sufficient influence assumption for discrete $\boldsymbol{a}$ is the same as for continuous $\boldsymbol{a}$.

Assumption 7 (Sufficient influence of $\boldsymbol{a}$ (nonparametric/discrete)) *For almost all $\boldsymbol{z} \in \mathbb{R}^{d_z}$ (i.e. except on a set with zero Lebesgue measure) and all $\ell \in [d_a]$, there exists*

$$\{(t_{(r)}, \tau_{(r)}, \boldsymbol{z}_{(r)}, \boldsymbol{a}^{<t}_{(r)}, \epsilon_{(r)})\}_{r=1}^{|\mathbf{Ch}^{\boldsymbol{a}}_\ell|} \,,$$

such that $t_{(r)} \in [T]$, $\tau_{(r)} < t_{(r)}$, $\boldsymbol{z}_{(r)} \in \mathbb{R}^{d_z \times (t_{(r)}-1)}$, $\boldsymbol{a}_{(r)} \in \mathcal{A}^{t_{(r)}}$, $\epsilon_{(r)} \in \mathbb{R}$, $(\boldsymbol{a}_{(r)})_{\cdot,\tau_{(r)}} + \epsilon_{(r)} \boldsymbol{e}_\ell \in \mathcal{A}$ and

$$\operatorname{span}\left\{ \Delta^{\tau_{(r)},\epsilon_{(r)}}_{a,\ell} D^{t_{(r)}}_z \log p(\boldsymbol{z} \mid \boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)}) \right\}_{r=1}^{|\mathbf{Ch}^{\boldsymbol{a}}_\ell|} = \mathbb{R}^{d_z}_{\mathbf{Ch}^{\boldsymbol{a}}_\ell} \,.$$

We can now provide a proof of Theorem 2. Note that it is almost identical to the proof of Theorem 1 except for the very first steps where we take a partial difference instead of a partial derivative.

Proof [of Theorem 2] We recall equation (8) derived in Section 3.2:

$$D^t_z \hat{q}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = D^t_z q(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t}) D\boldsymbol{v}(\boldsymbol{z}^t) + \eta(\boldsymbol{z}^t) \in \mathbb{R}^{1 \times d_z} \,. \tag{20}$$

Now, instead of differentiating w.r.t. $\boldsymbol{a}^\tau_\ell$ for some $\tau < t$ and $\ell \in [d_a]$, we are going to take a partial difference. That is, we evaluate the above equation on at $\boldsymbol{a}^{<t}$ and $\boldsymbol{a}^{<t} + \epsilon \boldsymbol{E}^{(\ell,\tau)}$ and $\epsilon \in \mathbb{R}$, where $\boldsymbol{E}^{(\ell,\tau)}$ is a "one-hot matrix", while keeping everything else constant, and take the difference. This yields:

$$\begin{aligned}
&[D^t_z \hat{q}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t} + \epsilon \boldsymbol{E}^{(\ell,\tau)}) - D^t_z \hat{q}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})]^\top \\
&\qquad = D\boldsymbol{v}(\boldsymbol{z}^t)^\top [D^t_z q(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t} + \epsilon \boldsymbol{E}^{(\ell,\tau)}) - D^t_z q(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t})]^\top \\
&\Delta^{\tau,\epsilon}_{a,\ell} D^t_z \hat{q}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})^\top = D\boldsymbol{v}(\boldsymbol{z}^t)^\top \Delta^{\tau,\epsilon}_{a,\ell} D^t_z q(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t})^\top \,,
\end{aligned} \tag{21}$$

where we used the notation for partial differences introduced in Definition 16. Notice that the difference on the left is $\subseteq \hat{\boldsymbol{G}}^a_{\cdot,\ell}$ and the difference on the right is $\subseteq \boldsymbol{G}^a_{\cdot,\ell}$. This equation is thus analogous to (16) from the continuous case. For that reason, we can employ a completely analogous strategy. Hence, we define

$$\hat{\Lambda}(\boldsymbol{z}^t, \gamma)_{\cdot,\ell} := \Delta^{\tau,\epsilon}_{a,\ell} D^t_z \hat{q}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})^\top \qquad \Lambda(\boldsymbol{z}^t, \gamma)_{\cdot,\ell} := \Delta^{\tau,\epsilon}_{a,\ell} D^t_z q(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t})^\top \,,$$

where $\gamma = (t, \tau, \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}, \vec{\epsilon})$. This notation allows us to rewrite (21) more compactly as

$$\underbrace{\hat{\Lambda}(\boldsymbol{z}^t, \gamma)}_{\subseteq \hat{\boldsymbol{G}}^a} = \boldsymbol{L}(\boldsymbol{z}^t)^\top \underbrace{\Lambda(\boldsymbol{z}^t, \gamma)}_{\subseteq \boldsymbol{G}^a} \,.$$

From here, the rest of the argument is exactly analogous to the proof of Theorem 1. ■

3.8.3 SUFFICIENT INFLUENCE ASSUMPTION OF THEOREM 3 AND ITS PROOF

We now introduce the sufficient influence assumption of Theorem 3, which showed how regularizing the temporal dependency graph $\hat{\boldsymbol{G}}^z$ to be sparse can result in disentanglement. Again, it is very similar to other sufficient influence assumptions we have seen so far.

Assumption 8 (Sufficient influence of $\boldsymbol{z}$ (nonparametric)) *For almost all $\boldsymbol{z} \in \mathbb{R}^{d_z}$ (i.e. except on a set with zero Lebesgue measure), there exists*

$$\{(t_{(r)}, \tau_{(r)}, \boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)})\}_{r=1}^{||\boldsymbol{G}^z||_0},$$

such that $t_{(r)} \in [T]$, $\tau_{(r)} < t_{(r)}$, $\boldsymbol{z}_{(r)} \in \mathbb{R}^{d_z \times (t_{(r)}-1)}$, $\boldsymbol{a}_{(r)} \in \mathcal{A}^{t_{(r)}}$, $\boldsymbol{z} = \boldsymbol{z}_{(r)}^{\tau_{(r)}}$ and

$$\operatorname{span}\left\{H_{z,z}^{t_{(r)},\tau_{(r)}} q(\boldsymbol{z} \mid \boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)})\right\}_{r=1}^{||\boldsymbol{G}^z||_0} = \mathbb{R}^{d_z}_{\boldsymbol{G}^z}.$$

Proof [of Theorem 3] We recall equation (15) derived in Section 3.2:

$$\underbrace{H_{z,z}^{t,\tau} \hat{q}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})}_{\subseteq \hat{\boldsymbol{G}}^z} = D\boldsymbol{v}(\boldsymbol{z}^t)^\top \underbrace{H_{z,z}^{t,\tau} q(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t})}_{\subseteq \boldsymbol{G}^z} D\boldsymbol{v}(\boldsymbol{z}^\tau).$$

This equation holds for all pairs of $\boldsymbol{z}^t$ and $\boldsymbol{z}^\tau$ in $\mathbb{R}^{d_z}$. We can thus evaluate it at a point such that $\boldsymbol{z}^t = \boldsymbol{z}^\tau$, which yields

$$H_{z,z}^{t,\tau} \hat{q}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = D\boldsymbol{v}(\boldsymbol{z}^t)^\top H_{z,z}^{t,\tau} q(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t}) D\boldsymbol{v}(\boldsymbol{z}^t). \tag{22}$$

Recall that Assumption 8 holds for all $\boldsymbol{z}^t \in \mathbb{R}^{d_z} \setminus E_0$ where E_0 has Lebesgue measure zero. Fix an arbitrary $\boldsymbol{z}^t \in \mathbb{R}^{d_z} \setminus E_0$ and set $\boldsymbol{z}^\tau = \boldsymbol{z}^t$. Let us define

$$\Lambda(\boldsymbol{z}^t, \gamma) := H_{z,z}^{t,\tau} q(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t}) \qquad \hat{\Lambda}(\boldsymbol{z}^t, \gamma) := H_{z,z}^{t,\tau} \hat{q}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}),$$

where $\gamma := (t, \tau, \boldsymbol{z}_{-\tau}^{<t}, \boldsymbol{a}^{<t})$ and $\boldsymbol{z}_{-\tau}^{<t}$ is $\boldsymbol{z}^{<t}$ but without $\boldsymbol{z}^\tau$. We can now rewrite (22) compactly as

$$\hat{\Lambda}(\boldsymbol{z}^t, \gamma) = D\boldsymbol{v}(\boldsymbol{z}^t)^\top \Lambda(\boldsymbol{z}^t, \gamma) D\boldsymbol{v}(\boldsymbol{z}^t).$$

Now, notice that the sufficient influence assumption (Assumption 8) requires that, there exists $\{\gamma_{(r)}\}_{r=1}^{||\boldsymbol{G}^z||_0}$ such that $\operatorname{span}\{\Lambda(\boldsymbol{z}^t, \gamma_{(r)})\}_{r=1}^{||\boldsymbol{G}^z||_0} = \mathbb{R}^{d_z \times d_z}_{\boldsymbol{G}^z}$. We can thus write

$$\begin{aligned} & D\boldsymbol{v}(\boldsymbol{z}^t)^\top \operatorname{span}\{\Lambda(\boldsymbol{z}^t, \gamma_{(r)})\}_{r=1}^{||\boldsymbol{G}^z||_0} D\boldsymbol{v}(\boldsymbol{z}^t) = \operatorname{span}\{\hat{\Lambda}(\boldsymbol{z}^t, \gamma_{(r)})\}_{r=1}^{||\boldsymbol{G}^z||_0} \subseteq \mathbb{R}^{d_z \times d_z}_{\hat{\boldsymbol{G}}^z} \\ \implies & D\boldsymbol{v}(\boldsymbol{z}^t)^\top \mathbb{R}^{d_z \times d_z}_{\boldsymbol{G}^z} D\boldsymbol{v}(\boldsymbol{z}^t) \subseteq \mathbb{R}^{d_z \times d_z}_{\hat{\boldsymbol{G}}^z} \end{aligned} \tag{23}$$

Since $D\boldsymbol{v}(\boldsymbol{z})$ is invertible, there exists a permutation $\boldsymbol{P}(\boldsymbol{z})$ such that $D\boldsymbol{v}(\boldsymbol{z})\boldsymbol{P}(\boldsymbol{z})$ has no zero on its diagonal (Lemma 2). Let $\boldsymbol{C}(\boldsymbol{z}) := D\boldsymbol{v}(\boldsymbol{z})\boldsymbol{P}(\boldsymbol{z})$. If we left and right-multiply (23) by $\boldsymbol{P}(\boldsymbol{z})^\top$ and $\boldsymbol{P}(\boldsymbol{z})$, respectively, we obtain

$$\boldsymbol{C}(\boldsymbol{z}^t)^\top \mathbb{R}^{d_z \times d_z}_{\boldsymbol{G}^z} \boldsymbol{C}(\boldsymbol{z}^t) \subseteq \mathbb{R}^{d_z \times d_z}_{\boldsymbol{P}(\boldsymbol{z})^\top \hat{\boldsymbol{G}}^z \boldsymbol{P}(\boldsymbol{z})}. \tag{24}$$

We now show that $\boldsymbol{G}^z \subseteq \boldsymbol{P}(\boldsymbol{z})^\top \hat{\boldsymbol{G}}^z \boldsymbol{P}(\boldsymbol{z})$. Take $(i,j) \in \boldsymbol{G}^z$. Since $\boldsymbol{e}_i \boldsymbol{e}_j^\top \in \mathbb{R}^{d_z \times d_z}_{\boldsymbol{G}^z}$, equation (24) implies

$$\boldsymbol{C}(\boldsymbol{z}^t)^\top \boldsymbol{e}_i \boldsymbol{e}_j^\top \boldsymbol{C}(\boldsymbol{z}^t) = (\boldsymbol{C}(\boldsymbol{z}^t)_{i,\cdot})^\top \boldsymbol{C}(\boldsymbol{z}^t)_{j,\cdot} \subseteq \mathbb{R}^{d_z \times d_z}_{\boldsymbol{P}(\boldsymbol{z})^\top \hat{\boldsymbol{G}}^z \boldsymbol{P}(\boldsymbol{z})} \tag{25}$$

Since $\boldsymbol{C}(\boldsymbol{z}^t)_{i,i} \boldsymbol{C}(\boldsymbol{z}^t)_{j,j} \neq 0$ (recall the diagonal of $\boldsymbol{C}(\boldsymbol{z}^t)$ has no zero), we must have $(i,j) \in \boldsymbol{P}(\boldsymbol{z})^\top \hat{\boldsymbol{G}}^z \boldsymbol{P}(\boldsymbol{z})$. This shows that $\boldsymbol{G}^z \subseteq \boldsymbol{P}(\boldsymbol{z})^\top \hat{\boldsymbol{G}}^z \boldsymbol{P}(\boldsymbol{z})$.

Since $||\boldsymbol{P}(\boldsymbol{z})^\top \hat{\boldsymbol{G}}^z \boldsymbol{P}(\boldsymbol{z})||_0 = ||\hat{\boldsymbol{G}}^z||_0 \leq ||\boldsymbol{G}^z||_0$, we must have $\boldsymbol{G}^z = \boldsymbol{P}(\boldsymbol{z})^\top \hat{\boldsymbol{G}}^z \boldsymbol{P}(\boldsymbol{z})$, which yields

$$\boldsymbol{C}(\boldsymbol{z}^t)^\top \mathbb{R}^{d_z \times d_z}_{\boldsymbol{G}^z} \boldsymbol{C}(\boldsymbol{z}^t) \subseteq \mathbb{R}^{d_z \times d_z}_{\boldsymbol{G}^z} . \tag{26}$$

We are now going to show that the above implies that $\boldsymbol{C}(\boldsymbol{z}^t)$ is both $\boldsymbol{G}^z$-preserving and $(\boldsymbol{G}^z)^\top$-preserving. Start by rewriting (25) as follows:

$$\text{for all } (i,j) \in \boldsymbol{G}^z,\ (\boldsymbol{C}(\boldsymbol{z}^t)_{i,\cdot})^\top \boldsymbol{C}(\boldsymbol{z}^t)_{j,\cdot} \subseteq \mathbb{R}^{d_z \times d_z}_{\boldsymbol{G}^z} . \tag{27}$$

We start by showing $\boldsymbol{G}^z$-preservation. To do so, we leverage the characterization of Proposition 3. We must show that $\boldsymbol{G}^z_{i,\cdot} \not\subseteq \boldsymbol{G}^z_{j,\cdot}$ implies $\boldsymbol{C}(\boldsymbol{z}^t)_{i,j} = 0$. Because $\boldsymbol{G}^z_{i,\cdot} \not\subseteq \boldsymbol{G}^z_{j,\cdot}$, there must exists k s.t. $\boldsymbol{G}^z_{i,k} = 1$ and $\boldsymbol{G}^z_{j,k} = 0$. We thus have, by (27), that $(\boldsymbol{C}(\boldsymbol{z}^t)_{i,\cdot})^\top \boldsymbol{C}(\boldsymbol{z}^t)_{k,\cdot} \subseteq \mathbb{R}^{d_z \times d_z}_{\boldsymbol{G}^z}$. Because $\boldsymbol{G}^z_{j,k} = 0$, we have $\boldsymbol{C}(\boldsymbol{z}^t)_{i,j} \boldsymbol{C}(\boldsymbol{z}^t)_{k,k} = 0$. But since $\boldsymbol{C}(\boldsymbol{z}^t)_{k,k} \neq 0$, we must have that $\boldsymbol{C}(\boldsymbol{z}^t)_{i,j} = 0$, as desired. To show $(\boldsymbol{G}^z)^\top$-preservation, one can use a completely analogous argument.

We showed that $\boldsymbol{C}(\boldsymbol{z}^t)$ is $\boldsymbol{G}^z$-preserving and $(\boldsymbol{G}^z)^\top$-preserving. It is easy to verify that this is equivalent to being $[\boldsymbol{G}^z\ (\boldsymbol{G}^z)^\top]$-preserving (where $[\cdot\ \cdot]$ stands for column concatenation). This remark will be useful below.

Similarly to the proof of Theorem 1, we must now show that there exists a single permutation that works for all $\boldsymbol{z}^t \in \mathbb{R}^{d_z}$. To achieve this, we use Lemma 12 with $\boldsymbol{G} := [\boldsymbol{G}^z\ (\boldsymbol{G}^z)^\top]$ and $\boldsymbol{L}(\boldsymbol{z}) := D\boldsymbol{v}(\boldsymbol{z})$. This allows us to say that there exists a permutation $\boldsymbol{P}$ such that $D\boldsymbol{v}(\boldsymbol{z})\boldsymbol{P}$ is $[\boldsymbol{G}^z\ (\boldsymbol{G}^z)^\top]$-preserving for all $\boldsymbol{z}$ (not "almost all").

Notice that $D(\boldsymbol{v} \circ \boldsymbol{P})(\boldsymbol{z}) = D\boldsymbol{v}(\boldsymbol{P}\boldsymbol{z})\boldsymbol{P}$, which is $[\boldsymbol{G}^z\ (\boldsymbol{G}^z)^\top]$-preserving everywhere. Using Lemma 1, we conclude that the function $\boldsymbol{c} := \boldsymbol{v} \circ \boldsymbol{P}$ is $[\boldsymbol{G}^z\ (\boldsymbol{G}^z)^\top]$-preserving. ■

3.9 Examples to Illustrate the Scope of the Theory

In this section, we provide several examples to gain better intuition as to when our results apply. Specifically, we will provide mathematically concrete examples of latent models $p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ illustrating the various sufficient influence assumptions we introduced. All these examples are summarized in Table 2.

Even though our results are nonparametric, we will concentrate on the special case of Gaussian models which are useful to get a good intuition of what the sufficient influence assumptions mean. The following simple lemma will be useful in the following examples. We present it without proof, as it can be derived from simple computations.

Lemma 3 *Let* $p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ *where* $\boldsymbol{\mu} \in \mathbb{R}^{d_z}$ *and* $\boldsymbol{\Sigma} := \text{diag}(\sigma_1^2, ..., \sigma_{d_z}^2)$. *Then*

$$D_z \log p(\boldsymbol{z}) = -\left[(\boldsymbol{z}_1 - \boldsymbol{\mu}_1)/\sigma_1^2, \ldots, (\boldsymbol{z}_{d_z} - \boldsymbol{\mu}_{d_z})/\sigma_{d_z}^2\right] \in \mathbb{R}^{1 \times d_z} .$$

3.9.1 CONTINUOUS AUXILIARY VARIABLES (THEOREM 1)

We start by illustrating Assumption 6 from Theorem 1. Example 9 assumes that we observe continuous actions that target each latent factor individually, while Example 10 gives a multi-target example.

Example 9 (Sufficient influence for continuous single-target actions) *We make Example 4 more concrete by explicitly specifying a latent transition model. Recall the situation depicted in Figure 1 where $\boldsymbol{z}_1$ is the tree position, $\boldsymbol{z}_2$ is the robot position, and $\boldsymbol{z}_3$ is the ball position ($d_z = 3$). Assume $\boldsymbol{a} \in [-1, 1]$ corresponds to the amount of torque applied to the wheels of the robot. We thus have $\boldsymbol{G}^a = [0, 1, 0]^\top$, i.e., $\boldsymbol{a}$ affects only the robot position $\boldsymbol{z}_2$. For this example, $\boldsymbol{G}^z$ can be anything. Let $p(\boldsymbol{z}^t \mid \boldsymbol{z}^{t-1}, \boldsymbol{a}) = \mathcal{N}(\boldsymbol{z}^t; \boldsymbol{\mu}(\boldsymbol{z}^{t-1}, \boldsymbol{a}), \sigma^2 \boldsymbol{I})$ where*

$$\boldsymbol{\mu}(\boldsymbol{z}^{t-1}, \boldsymbol{a}) := \boldsymbol{z}^{t-1} + \boldsymbol{g}(\boldsymbol{z}^{t-1}) + \boldsymbol{a} \cdot \boldsymbol{G}^a \,.$$

where $\boldsymbol{g} : \mathbb{R}^{d_z} \to \mathbb{R}^{d_z}$ is some function that satisfies the dependency graph $\boldsymbol{G}^z$ (e.g. $\boldsymbol{g}(\boldsymbol{z}) := \boldsymbol{W}\boldsymbol{z}$ where $\boldsymbol{W} \in \mathbb{R}^{d_z \times d_z}_{\boldsymbol{G}^z}$). If no torque is applied ($\boldsymbol{a} = 0$), then the position of the robots is determined by the dynamics of the system. However, adding a positive or negative torque ($\boldsymbol{a} \neq 0$) nudges the robot to the right or to the left. Using Lemma 3, we can compute

$$H^t_{z,a} \log p(\boldsymbol{z}^t \mid \boldsymbol{z}^{t-1}, \boldsymbol{a}) = [0, 1/\sigma^2, 0]^\top \,,$$

which spans $\mathbb{R}^3_{\{2\}}$ and thus Assumption 6 holds.

Example 10 (Sufficient influence for continuous multi-target actions) *We make Example 5 more concrete by specifying an explicit latent model. Recall that $\boldsymbol{G}^a$ is given by Figure 3c with $d_z = d_a = 3$. Assume there are no temporal dependencies ($T = 1$), that $\boldsymbol{a} \in \mathbb{R}^3$ and that the latent model is given by $p(\boldsymbol{z} \mid \boldsymbol{a}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}(\boldsymbol{a}), \sigma^2 \boldsymbol{I})$ where*

$$\boldsymbol{\mu}(\boldsymbol{a}) := \begin{bmatrix} \boldsymbol{a}_1 \\ \boldsymbol{a}_1^2 \\ 0 \end{bmatrix} + \begin{bmatrix} \boldsymbol{a}_2 \\ 0 \\ \boldsymbol{a}_2^2 \end{bmatrix} + \begin{bmatrix} 0 \\ \boldsymbol{a}_3 \\ \boldsymbol{a}_3^2 \end{bmatrix} \,. \tag{28}$$

Using Lemma 3, we can compute

$$H_{z,a} q(\boldsymbol{z} \mid a) = \frac{1}{\sigma^2} \begin{bmatrix} 1 & 1 & 0 \\ 2\boldsymbol{a}_1 & 0 & 1 \\ 0 & 2\boldsymbol{a}_2 & 2\boldsymbol{a}_3 \end{bmatrix} \,.$$

Consider $\ell = 1$ so that $\mathbf{Ch}^a_1 = \{1, 2\}$. We can see that $H_{z,a} q(\boldsymbol{z} \mid a = 0)_{\cdot,1} = [1, 0, 0]^\top$ and $H_{z,a} q(\boldsymbol{z} \mid a = \boldsymbol{e}_1)_{\cdot,1} = [1, 2, 0]^\top$ span $\mathbb{R}^3_{\{1,2\}}$. Analogous conclusions can also be reached for $\ell = 2, 3$, which shows that Assumption 6 holds.

Now suppose that instead $\boldsymbol{\mu}(\boldsymbol{a})$ is a linear map, i.e., $\boldsymbol{\mu}(\boldsymbol{a}) := \boldsymbol{W}\boldsymbol{a}$ where $\boldsymbol{W} \in \mathbb{R}^{d_z \times d_a}_{\boldsymbol{G}^a}$. This would imply that $H_{z,a} q(\boldsymbol{z} \mid a) \propto \boldsymbol{W}$, which means it cannot satisfy the sufficient influence assumption (unless $||\boldsymbol{G}^a_{\cdot,\ell}||_0 \leq 1$ for all ℓ).

3.9.2 DISCRETE AUXILIARY VARIABLES OR INTERVENTIONS (THEOREM 2)

We now provide three concrete examples of latent models $p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ that satisfy Assumption 7, from Theorem 2. Here, we interpret the discrete auxiliary variable $\boldsymbol{a}$ as an *intervention index*, as discussed in Section 3.4.1, but note that other interpretations are possible (such as $\boldsymbol{a}$ as an action). Recall that our identifiability result does not require the knowledge of the targets of the interventions, these can be learned.

Example 11 shows how single target interventions can be used to obtain complete disentanglement without temporal dependencies, Example 12 shows how multi-target interventions can be leveraged for disentanglement if temporal dependencies are present, and Example 13 shows how *grouped* multi-target interventions allow disentanglement even when there are no time dependencies (Remark 6).

Example 11 (Single-target interventions for complete disentanglement without time) *We make Example 3 more concrete by specifying an explicit latent model. Assume $d_a = d_z$ and that $\boldsymbol{a} \in \mathcal{A} := \{\mathbf{0}, \boldsymbol{e}_1, \dots, \boldsymbol{e}_{d_a}\}$ is interpreted as an intervention index (see Section 3.4.1). Furthermore, Example 3 assumed $\boldsymbol{G}^a = \boldsymbol{I}$, i.e. each latent factor is targeted once by an intervention that targets only this factor (the example actually allowed to add arbitrary columns to $\boldsymbol{G}^a$, i.e. adding more interventions, without compromising complete disentanglement). Assume there are no temporal dependencies, i.e., $T = 1$, and that $p(\boldsymbol{z} \mid \boldsymbol{a}) := \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}(\boldsymbol{a}), \operatorname{diag}(\boldsymbol{\sigma}^2(\boldsymbol{a})))$ with*

$$\boldsymbol{\mu}(\boldsymbol{a}) := \boldsymbol{\mu} \odot \boldsymbol{a} \quad \text{and} \quad \boldsymbol{\sigma}^2(\boldsymbol{a}) := \mathbb{1} + \boldsymbol{\delta} \odot \boldsymbol{a}\,,$$

where $\odot$ denotes the Hadamard product (a.k.a. the element-wise product), $\boldsymbol{\mu} \in \mathbb{R}^{d_z}$ is the vector of means for each intervention and $\boldsymbol{\delta} \in \mathbb{R}^{d_z}$ is the vector of shifts in variance for all interventions. Thus, in the observational setting ($\boldsymbol{a} = \mathbf{0}$), we have $\mu(\boldsymbol{a}) = \mathbf{0}$ and $\boldsymbol{\sigma}(\boldsymbol{a}) = \mathbb{1}$ while in the ℓth intervention ($\boldsymbol{a} = \boldsymbol{e}_\ell$), the mean and variance of the targeted latent shift are the same, i.e. $\boldsymbol{\mu}(\boldsymbol{a}) = \boldsymbol{\mu}_\ell \boldsymbol{e}_\ell$ and $\boldsymbol{\sigma}^2(\boldsymbol{a}) = \mathbb{1} + \boldsymbol{\delta}\boldsymbol{e}_\ell$ (assuming that the shifted variance is > 0). Using Lemma 3, we can compute

$$\Delta^{\epsilon=1}_{a,\ell} D_z \log p(\boldsymbol{z} \mid \boldsymbol{a} = \mathbf{0}) := D_z \log p(\boldsymbol{z} \mid \boldsymbol{a} = \boldsymbol{e}_\ell) - D_z \log p(\boldsymbol{z} \mid \boldsymbol{a} = \mathbf{0}) = \frac{\boldsymbol{\mu}_\ell + \boldsymbol{\delta}_\ell \boldsymbol{z}_\ell}{1 + \boldsymbol{\delta}_\ell} \boldsymbol{e}_\ell\,,$$

which must span $\mathbb{R}^{d_z}_{\{\ell\}}$ unless $\boldsymbol{\mu}_\ell + \boldsymbol{\delta}_\ell \boldsymbol{z}_\ell = 0$. However, note that when, for all ℓ, $\boldsymbol{\mu}_\ell \neq 0$ or $\boldsymbol{\delta}_\ell \neq 0$ (i.e. all interventions truly have an effect), the set $\{\boldsymbol{z} \in \mathbb{R}^{d_z} \mid \boldsymbol{\mu}_\ell + \boldsymbol{\delta}_\ell \boldsymbol{z}_\ell = 0 \text{ for some } \ell\}$ has zero Lebesgue measure in $\mathbb{R}^{d_z}$, which is allowed by Assumption 7.

Remark 6 (Potential issues with multi-target interventions without time) *What if an intervention targets more than one latent at a time? Can it still satisfy the sufficient influence assumption? We will now see that, without time-dependencies ($T = 1$), it is impossible. Consider the simple situation where $d_z = 3$, $d_a = 1$, $\boldsymbol{a} \in \{0, 1\}$ and $\boldsymbol{G}^a = [1, 1, 0]^\top$, i.e., there is a single intervention targeting $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$. In that case, there is a single possible difference vector, which is*

$$\Delta^{\epsilon=1}_{a} D_z \log p(\boldsymbol{z} \mid \boldsymbol{a} = 0) = D_z \log p(\boldsymbol{z} \mid \boldsymbol{a} = 1) - D_z \log p(\boldsymbol{z} \mid \boldsymbol{a} = 0) \in \mathbb{R}^{d_z}_{\{1,2\}}\,.$$

Since this is the only difference vector, we can see that we cannot span the 2-dimensional space $\mathbb{R}^{d_z}_{\{1,2\}}$. Therefore, to leverage multi-target interventions in our framework, more "variability" is required. Example 12 below shows how temporal dependencies can provide this additional variability, while Example 13 shows how having "groups" of interventions known to have the same (unknown) targets can also provide the required variability.

Example 12 (Multi-target interventions for complete disentanglement with time) *We make Example 5 more concrete by specifying an explicit latent model that satisfies Assumption 7. Recall $d_z = 3$, $d_a = 3$ and $\boldsymbol{G}^a$ is depicted in Figure 3c. This time, we assume there are temporal dependencies, i.e. $T > 1$ and $\boldsymbol{G}^z$ is non-trivial. Suppose $\boldsymbol{a} \in \mathcal{A} := \{\mathbf{0}, \boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3\}$ where $\boldsymbol{e}_\ell$ is the ℓth one-hot and we interpret $\mathbf{0}$ to correspond to the observational setting and $\boldsymbol{e}_\ell$ to correspond to the ℓth intervention. Recall that in this interpretation $\boldsymbol{G}^a$ describes which latent variable is targeted by each intervention. Let $p(\boldsymbol{z}^t \mid \boldsymbol{z}^{t-1}, \boldsymbol{a}) = \mathcal{N}(\boldsymbol{z}^t; \boldsymbol{\mu}(\boldsymbol{z}^{t-1}, \boldsymbol{a}), \sigma^2 \boldsymbol{I})$ where*

$$\boldsymbol{\mu}(\boldsymbol{z}^{t-1}, \boldsymbol{a}) := \boldsymbol{z}^{t-1} + (\mathbb{1} - \boldsymbol{G}^a \boldsymbol{a}) \odot \boldsymbol{g}(\boldsymbol{z}^{t-1}) \,,$$

where $\boldsymbol{g} : \mathbb{R}^{d_z} \to \mathbb{R}^{d_z}$ is some function that respects the graph $\boldsymbol{G}^z$ (e.g. $\boldsymbol{g}(\boldsymbol{z}) = \boldsymbol{W}\boldsymbol{z}$ where $\boldsymbol{W} \in \mathbb{R}^{d_z \times d_z}_{\boldsymbol{G}^z}$). The observational dynamics is then $\boldsymbol{\mu}(\boldsymbol{z}^t, \boldsymbol{a} = \mathbf{0}) = \boldsymbol{z}^{t-1} + \boldsymbol{g}(\boldsymbol{z}^{t-1})$ and the interventional settings correspond to zeroing out the elements of $\boldsymbol{g}(\boldsymbol{z}^{t-1})$ targeted by the intervention. Using Lemma 3, we can compute

$$\begin{aligned} &\Delta^{\epsilon=1}_{\boldsymbol{a},\ell} D_z q(\boldsymbol{z}^t \mid \boldsymbol{z}^{t-1}, \boldsymbol{a} = \mathbf{0}) \\ &= D_z q(\boldsymbol{z}^t \mid \boldsymbol{z}^{t-1}, \boldsymbol{a} = \boldsymbol{e}_\ell) - D_z q(\boldsymbol{z}^t \mid \boldsymbol{z}^{t-1}, \boldsymbol{a} = \mathbf{0}) = -\frac{1}{\sigma^2} \boldsymbol{G}^a_{\cdot,\ell} \odot \boldsymbol{g}(\boldsymbol{z}^{t-1}) \,. \end{aligned}$$

One can see that, as soon as the image of $\boldsymbol{g}$ spans $\mathbb{R}^{d_z}$, Assumption 7 is satisfied since we can choose values $\boldsymbol{z}_{(1)}, \ldots, \boldsymbol{z}_{(d_z)} \in \mathbb{R}^{d_z}$ such that $\operatorname{span}\{\boldsymbol{g}(\boldsymbol{z}_{(1)}), \ldots, \boldsymbol{g}(\boldsymbol{z}_{(d_z)})\} = \mathbb{R}^{d_z}$, which implies $\operatorname{span}\{\boldsymbol{G}^a_{\cdot,\ell} \odot \boldsymbol{g}(\boldsymbol{z}_{(1)}), \ldots, \boldsymbol{G}^a_{\cdot,\ell} \odot \boldsymbol{g}(\boldsymbol{z}_{(d_z)})\} = \mathbb{R}^{d_z}_{\mathbf{Ch}^a_\ell}$. An example of a transition function $\boldsymbol{g}$ that satisfies this property is $\boldsymbol{g}(\boldsymbol{z}) := \boldsymbol{W}\boldsymbol{z}$ where $\boldsymbol{W} \in \mathbb{R}^{d_z \times d_z}_{\boldsymbol{G}^z}$ is invertible.

Note that even if the temporal dependencies are not sparse, they are still helpful for identifiability as they make it more likely to satisfy the sufficient influence assumption (Assumption 7).

Example 13 (Grouped multi-target interventions for disentanglement without time) *In this example, we assume that there are no temporal dependencies ($T = 1$) and that the learner has access to d_a groups of interventions where the interventions belonging to the ℓth group are known to target the same latent variables given by $\boldsymbol{G}^a_{\cdot,\ell}$ (these targets are unknown). Here is how our framework can accommodate this setting: Given that we have d_a groups of interventions where the ℓth group contains k_ℓ interventions, we set $\mathcal{A} := \{\mathbf{0}, 1\boldsymbol{e}_1, ..., k_1\boldsymbol{e}_1, 1\boldsymbol{e}_2, ..., k_2\boldsymbol{e}_2, ..., k_{d_a}\boldsymbol{e}_{d_a}\}$. In this setting, $\boldsymbol{a} = j\boldsymbol{e}_\ell$ corresponds to the jth intervention of the ℓth group. Moreover, the sufficient influence assumption requires that interventions within a group ℓ span $\mathbb{R}^{d_z}_{\mathbf{Ch}^a_\ell}$. More precisely, we need $\operatorname{span}\{\Delta^\epsilon_{a,\ell} D_z \log p(\boldsymbol{z} \mid \boldsymbol{a} = 0)\}_{\epsilon=1}^{k_\ell} = \mathbb{R}^{d_z}_{\mathbf{Ch}^a_\ell}$.*

3.9.3 TEMPORAL DEPENDENCIES (THEOREM 3)

Finally, we provide an example (Example 14) where temporal dependencies alone (no auxiliary variable $\boldsymbol{a}$) are enough to disentangle. We start with an important remark about the sufficient influence assumption of Theorem 3.

Remark 7 (Auxiliary variables or non-Markovianity are required) *An important observation is that, if the transition model does not have an auxiliary variable $\boldsymbol{a}$ and is Markovian, i.e., $p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = p(\boldsymbol{z}^t \mid \boldsymbol{z}^{t-1})$, then Assumption 8 cannot be satisfied (except in trivial circumstances). To see this, simply note that, in that case, $H^{t,t-1}_{z,z} q(\boldsymbol{z}^t \mid \boldsymbol{z}^{t-1})$ depends only on $\boldsymbol{z}^{t-1}$, which is*

forced to be equal to $\boldsymbol{z}^t$. This means that the span of the Hessian must be at most one-dimensional, which means that the assumption cannot hold as soon as $||\boldsymbol{G}^z||_0 > 1$. Therefore, when no auxiliary variable $\boldsymbol{a}$ is observed, Assumption 8 requires that the transition model be non-Markovian. In Example 14, we provide a concrete example of a transition model without auxiliary variable $\boldsymbol{a}$ that satisfies this assumption. We will also see in Section 4 that if the transition model $p(\boldsymbol{z}^t \mid \boldsymbol{z}^{t-1})$ is in the exponential family, this assumption can be relaxed so that non-Markovianity is no longer required.

Example 14 (Sparse temporal dependencies for disentanglement without auxiliary variables) *We continue with Examples 6 & 7 which were based on the graphs $\boldsymbol{G}^z$ depicted in Figures 3d & 3e, respectively. Assume that no action is observed, i.e., we can only take advantage of the sparsity of $\boldsymbol{G}^z$ to disentangle. Examples 6 & 7 already showed that these graph structures allow complete disentanglement, as long as the sufficient influence assumption for $\boldsymbol{z}$ (Assumption 8) is satisfied. We now provide concrete transition models $p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t})$ that satisfy this requirement. Similarly to previous examples, assume $p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}) = \mathcal{N}(\boldsymbol{z}^t \mid \boldsymbol{\mu}(\boldsymbol{z}^{t-1}, \boldsymbol{z}^{t-2}), \sigma^2 \boldsymbol{I})$ where*

$$\boldsymbol{\mu}(\boldsymbol{z}^{t-1}, \boldsymbol{z}^{t-2}) := \boldsymbol{z}^{t-1} + \boldsymbol{W}(\boldsymbol{z}^{t-2})\boldsymbol{z}^{t-1}\,,$$

where $\boldsymbol{W} : \mathbb{R}^{d_z} \to \mathbb{R}^{d_z \times d_z}_{\boldsymbol{G}^z}$ is some function of $\boldsymbol{z}^{t-2}$. Using Lemma 3, we can derive

$$H^{t,t-1}_{z,z} q(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}) = \frac{1}{\sigma^2}[\boldsymbol{I} + \boldsymbol{W}(\boldsymbol{z}^{t-2})]\,.$$

Thus, Assumption 8 holds when there exists $\{\boldsymbol{z}^{t-2}_{(r)}\}_{r=1}^{||\boldsymbol{G}^z||_0}$ such that

$$\operatorname{span}\{\boldsymbol{I} + \boldsymbol{W}(\boldsymbol{z}^{t-2}_{(r)})\}_{r=1}^{||\boldsymbol{G}^z||_0} = \mathbb{R}^{d_z \times d_z}_{\boldsymbol{G}^z}\,. \tag{29}$$

One can directly see that, if $\boldsymbol{W}(\boldsymbol{z}^{t-2})$ was actually constant in $\boldsymbol{z}^{t-2}$, the assumption could not hold (unless $||\boldsymbol{G}^z||_0 \leq 1$). This case would correspond to a simple linear model of the form $\boldsymbol{\mu}(\boldsymbol{z}^{t-1}) := \boldsymbol{z}^{t-1} + \boldsymbol{W}\boldsymbol{z}^{t-1}$. Our theory suggests that this transition function is "too simple" to allow disentanglement.

Nevertheless, we can find examples satisfying (29). For example, if $\boldsymbol{G}^z = \boldsymbol{I}$, we can take

$$\boldsymbol{W}(\boldsymbol{z}) = \begin{bmatrix} \boldsymbol{z}_1 & 0 & 0 \\ 0 & \boldsymbol{z}_2 & 0 \\ 0 & 0 & \boldsymbol{z}_3 \end{bmatrix}$$

and see that the family of functions $(1 + \boldsymbol{z}_1, 1 + \boldsymbol{z}_2, 1 + \boldsymbol{z}_3)$ is linearly independent (when seen as functions from $\mathbb{R}^3$ to $\mathbb{R}$). By Lemma 5 in the appendix, this is equivalent to the existence of $\boldsymbol{z}_{(1)}, \boldsymbol{z}_{(2)}, \boldsymbol{z}_{(3)} \in \mathbb{R}^{d_z}$ such that (29) holds (see also Remark 5). In other words, the sufficient influence assumption holds. In the case where $\boldsymbol{G}^z$ is lower triangular like in Figure 3e, one can take

$$\boldsymbol{W}(\boldsymbol{z}) = \begin{bmatrix} \boldsymbol{z}_1 & 0 & 0 \\ \boldsymbol{z}_2^2 & \boldsymbol{z}_2 & 0 \\ \boldsymbol{z}_3^3 & \boldsymbol{z}_3^2 & \boldsymbol{z}_3 \end{bmatrix}$$

and see that the family of functions $(1+\boldsymbol{z}_1, 1+\boldsymbol{z}_2, 1+\boldsymbol{z}_3, \boldsymbol{z}_2^2, \boldsymbol{z}_3^2, \boldsymbol{z}_3^3)$ is linearly independent, which similarly implies the existence of $\boldsymbol{z}_{(1)}, \ldots, \boldsymbol{z}_{(6)} \in \mathbb{R}^{d_z}$ such that (29) holds.

Example 16 will show how one can leverage the exponential family assumption to allow Markovianity even without auxiliary variables.

4 Partial Disentanglement via Mechanism Sparsity in Exponential Families

The goal of this section is to understand how restricting the transition model to be in the *exponential family* allows us to weaken the sufficient influence assumption of Theorem 3. Section 4.1 introduces the exponential family assumption. Section 4.2 follows Khemakhem et al. (2020a) and shows that this additional assumption guarantees that the entanglement map $\boldsymbol{v}$ is "quasi-linear", which means $\boldsymbol{v}(\boldsymbol{z}) := \boldsymbol{s}^{-1}(\boldsymbol{L}\boldsymbol{s}(\boldsymbol{z}) + \boldsymbol{b})$, where $\boldsymbol{L}$ is a matrix and $\boldsymbol{s}$ is an element-wise invertible function. Section 4.3 will introduce an identifiability result analogous to Theorem 3 for sparse $\hat{\boldsymbol{G}}^z$ that takes advantage of the quasi-linearity of $\boldsymbol{v}$ to weaken Assumption 8 (sufficient influence of $\boldsymbol{z}$). We also briefly discuss an additional result from Appendix B.4 that shows connections between the nonparametric sufficient influence assumptions of this work (Assumptions 7 & 8) and their counterparts in Lachapelle et al. (2022) (Assumptions 11 & 12).

4.1 Exponential Family Latent Transition Models

We will assume that the conditional densities $p(\boldsymbol{z}_i^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ are from an *exponential family* (Wainwright and Jordan, 2008):

Assumption 9 (Exponential family transition model) *For all $i \in [d_z]$, we have*

$$p(\boldsymbol{z}_i^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = h_i(\boldsymbol{z}_i^t) \exp\{\boldsymbol{s}_i(\boldsymbol{z}_i^t)^\top \boldsymbol{\lambda}_i(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) - \psi_i(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})\}\,. \tag{30}$$

Well-known distributions which belong to this family include the Gaussian and beta distributions. In the Gaussian case, the *sufficient statistic* is $\boldsymbol{s}_i(z) := (z, z^2)$ and the *base measure* is $h_i(z) := \frac{1}{\sqrt{2\pi}}$. The function $\boldsymbol{\lambda}_i(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ outputs the *vectors of natural parameters* for the conditional distribution and can be itself parametrized, for instance, by a multi-layer perceptron (MLP) or a recurrent neural network (RNN). We will refer to the functions $\boldsymbol{\lambda}_i$ as the *mechanisms* or the *transition functions*. In the Gaussian case, the natural parameter is two-dimensional and is related to the usual parameters μ and σ^2 via the equation $(\lambda_1, \lambda_2) = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})$. We will denote by k the dimensionality of the natural parameter and that of the sufficient statistic (which are equal). Thus, $k = 2$ in the Gaussian case. The remaining term $\psi_i(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ acts as a normalization constant.

We define $\boldsymbol{\lambda}(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) \in \mathbb{R}^{kd_z}$ as the concatenation of all $\boldsymbol{\lambda}_i(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ and similarly for $\boldsymbol{s}(\boldsymbol{z}^t) \in \mathbb{R}^{kd_z}$. Similarly to the nonparameteric case, the learnable parameters are $\boldsymbol{\theta} := (\boldsymbol{f}, \boldsymbol{\lambda}, \boldsymbol{G})$. Note that throughout, we assume that the sufficient statistic $\boldsymbol{s}$ is not learned and known in advance. With this notation, we can write the full transition model as

$$p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = h(\boldsymbol{z}^t) \exp\{\boldsymbol{s}(\boldsymbol{z}^t)^\top \boldsymbol{\lambda}(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) - \psi(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})\}\,,$$

where $h := \prod_{i=1}^{d_z} h_i$ and $\psi = \sum_{i=1}^{d_z} \psi_i$.

Remark 8 (Applying nonparametric identifiability results to exponential families) *One can apply the nonparametric results (Theorems 1, 2 & 3) to models satisfying the exponential family assumption. In fact, all examples of Section 3.9 were Gaussians and thus are in the exponential family.*

4.2 Conditions for Quasi-Linear Identifiability

In this section, we follow Khemakhem et al. (2020a) and show that the exponential family assumption combined with an additional sufficient variability assumption allows us to go from identifiability up to diffeomorphism (Definition 5) to identifiability up to quasi-linearity, which we define next:

Definition 17 (Quasi-linear equivalence) *We say that two models $\boldsymbol{\theta} := (\boldsymbol{f}, \boldsymbol{\lambda}, \boldsymbol{G})$ and $\tilde{\boldsymbol{\theta}} := (\tilde{\boldsymbol{f}}, \tilde{\boldsymbol{\lambda}}, \tilde{\boldsymbol{G}})$ satisfying Assumptions 1, 2 & 9 are **equivalent up to quasi-linearity**, denoted $\boldsymbol{\theta} \sim_{\text{lin}} \tilde{\boldsymbol{\theta}}$, if and only if $\boldsymbol{\theta} \sim_{\text{diff}} \tilde{\boldsymbol{\theta}}$ and there exist an invertible matrix $\boldsymbol{L} \in \mathbb{R}^{kd_z \times kd_z}$ and a vector $\boldsymbol{b} \in \mathbb{R}^{kd_z}$ such that the map $\boldsymbol{v} := \boldsymbol{f}^{-1} \circ \tilde{\boldsymbol{f}}$ satisfies*

$$\boldsymbol{s}(\boldsymbol{v}(\boldsymbol{z})) = \boldsymbol{L}\boldsymbol{s}(\boldsymbol{z}) + \boldsymbol{b}, \ \forall \boldsymbol{z} \in \mathbb{R}^{d_z} .$$

If the sufficient statistic $\boldsymbol{s}$ is invertible, one obtains

$$\boldsymbol{v}(\boldsymbol{z}) = \boldsymbol{s}^{-1}(\boldsymbol{L}\boldsymbol{s}(\boldsymbol{z}) + \boldsymbol{b}), \ \forall \boldsymbol{z} \in \mathbb{R}^{d_z} . \tag{31}$$

Equation (31) is particularly interesting, as it says that the mapping relating both representations is "almost" linear in the following sense: although the map is not necessarily linear because the sufficient statistic $\boldsymbol{s}$ might not be, the "mixing" between components is linear. Indeed, notice that the sufficient statistic $\boldsymbol{s}$ and its inverse operate "element-wise". The mixing between components is only due to the matrix $\boldsymbol{L}$. This specific form simplifies a few steps in the identifiability proof, which might explain the popularity of this assumption in the literature on nonlinear ICA (Hyvarinen and Morioka, 2016; Khemakhem et al., 2020a,b; Hälvä and Hyvärinen, 2020; Morioka et al., 2021; Yang et al., 2021; Lachapelle et al., 2022; Liu et al., 2023; Xi and Bloem-Reddy, 2023).

The following theorem provides conditions to guarantee identifiability up to quasi-linearity. This is an adaptation and minor extension of Theorem 1 from Khemakhem et al. (2020a). For completeness, we provide a proof in Appendix B.2.

Theorem 4 (Conditions for linear identifiability - Adapted from Khemakhem et al. (2020a)) *Let $\boldsymbol{\theta} := (\boldsymbol{f}, \boldsymbol{\lambda}, \boldsymbol{G})$ and $\hat{\boldsymbol{\theta}} := (\hat{\boldsymbol{f}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{G}})$ be two models satisfying Assumptions 1, 2 & 9. Further assume that*

1. ***[Observational equivalence]** $\boldsymbol{\theta} \sim_{\text{obs}} \hat{\boldsymbol{\theta}}$ (Definition 4);*
2. ***[Minimal sufficient statistics]** For all i, the sufficient statistic $\boldsymbol{s}_i$ is minimal (see below).*
3. ***[Sufficient variability]** The natural parameter $\boldsymbol{\lambda}$ varies "sufficiently" as formalized by Assumption 10 (see below).*

Then, $\boldsymbol{\theta} \sim_{\text{lin}} \hat{\boldsymbol{\theta}}$ (Def. 17).

The "minimal sufficient statistics" assumption is a standard one saying that $\boldsymbol{s}_i$ is defined appropriately to ensure that the parameters of the exponential family are identifiable (see e.g. Wainwright and Jordan (2008, p. 40)). See Definition 20 for a formal definition of minimality. The last assumption is sometimes called the *assumption of variability* (Hyvärinen et al., 2019), and requires that the conditional distribution of $\boldsymbol{z}^t$ depends "sufficiently strongly" on $\boldsymbol{z}^{<t}$ and/or $\boldsymbol{a}^{<t}$. We stress the fact that this assumption concerns the ground-truth data generating model $\boldsymbol{\theta}$.

Assumption 10 (Sufficient variability in exponential families) *There exist $(\boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)})_{p=0}^{kd_z}$ in their respective supports such that the kd_z-dimensional vectors $(\boldsymbol{\lambda}(\boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)}) - \boldsymbol{\lambda}(\boldsymbol{z}_{(0)}, \boldsymbol{a}_{(0)}))_{r=1}^{kd_z}$ are linearly independent.*

Notice that $\boldsymbol{z}_{(r)}$ represents values of $\boldsymbol{z}^{<t}$ for potentially different values of t and therefore can vary in shape.

The following example builds on Example 11 and shows that the sufficient variability of the above theorem might hold or not. The first case is interesting since it guarantees that $\boldsymbol{v}$ is linear while the second is interesting because it showcases a situation where the theory of Khemakhem et al. (2020a) and Lachapelle et al. (2022) do not apply (since they both rely on the above theorem), thus highlighting the importance of our nonparametric extension.

Example 15 (Satisfying or not the sufficient variability assumption of Theorem 4) *We recall Example 11 in which $d_a = d_z$, $\boldsymbol{a} \in \mathcal{A} := \{\mathbf{0}, \boldsymbol{e}_1, \dots, \boldsymbol{e}_{d_a}\}$ and $\boldsymbol{G}^a = \boldsymbol{I}$ without temporal dependencies: For all $i \in [d_z]$, $p(\boldsymbol{z}_i \mid \boldsymbol{a}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_i \boldsymbol{a}_i, 1 + \boldsymbol{\delta}_i \boldsymbol{a}_i)$ where $\boldsymbol{\mu}_i \in \mathbb{R}$ and $\boldsymbol{\delta}_i > -1$. We consider the cases where $\forall i,\ \boldsymbol{\delta}_i = 0$ (variances do not change) and $\forall i,\ \boldsymbol{\delta}_i \neq 0$ (variances change).*

If $\forall i,\ \boldsymbol{\delta}_i = 0$, we can represent $p(\boldsymbol{z}_i \mid \boldsymbol{a})$ in its exponential form with a one-dimensional sufficient statistic given by $\boldsymbol{s}_i(\boldsymbol{z}_i) = \boldsymbol{z}_i$ and a natural parameter given by $\boldsymbol{\lambda}_i(\boldsymbol{a}) = \boldsymbol{\mu}_i \boldsymbol{a}_i$. It can be easily seen that if $\forall i,\ \boldsymbol{\mu}_i \neq 0$ (i.e. the mean changes after the intervention), then the sufficient variability assumption of Theorem 4 holds since the vectors $\boldsymbol{\lambda}(\boldsymbol{e}_i) - \boldsymbol{\lambda}(\mathbf{0}) = \boldsymbol{\mu}_i \boldsymbol{e}_i$ span $\mathbb{R}^{d_z}$.

If $\forall i,\ \boldsymbol{\delta}_i \neq 0$, we can represent $p(\boldsymbol{z}_i \mid \boldsymbol{a})$ in its exponential form with a two-dimensional sufficient statistics given by $\boldsymbol{s}_i(\boldsymbol{z}_i) = (\boldsymbol{z}_i, \boldsymbol{z}_i^2)$ and natural parameter given by $\boldsymbol{\lambda}_i(\boldsymbol{a}) = \left(\frac{\boldsymbol{\mu}_i \boldsymbol{a}_i}{1+\boldsymbol{\delta}_i \boldsymbol{a}_i}, \frac{-1}{2(1+\boldsymbol{\delta}_i \boldsymbol{a}_i)}\right)$. Note that, because we only have d_z interventions, for any choice of $\boldsymbol{a}_{(0)} \in \mathcal{A}$, the vectors $\{\boldsymbol{\lambda}(\boldsymbol{a}) - \boldsymbol{\lambda}(\boldsymbol{a}_{(0)})\}_{\boldsymbol{a} \in \mathcal{A}}$ can span at most a d_z-dimensional subspace, which is insufficient variability according to Theorem 4 since it requires spanning $\mathbb{R}^{2d_z}$.

4.3 Partial Disentanglement via Sparse Time Dependencies in Exponential Families

We now provide a (partial) disentanglement guarantee that takes advantage of sparsity regularization of $\hat{\boldsymbol{G}}^z$ and is specialized for exponential families with a one-dimensional sufficient statistic ($k = 1$). We will see that this extra parametric assumption on the transition model allows us to weaken the sufficient influence assumption of Theorem 3 (Assumption 8). In particular, this is going to allow for Markovian transitions without auxiliary variables, which was not allowed by the nonparametric result (Remark 7).

The sufficient influence assumption for $\boldsymbol{z}$ specialized to exponential families with $k = 1$ is directly taken from Lachapelle et al. (2022):

Assumption 11 (Sufficient influence of $\boldsymbol{z}$ (Lachapelle et al., 2022)) *Assume $k = 1$ and $D\boldsymbol{s}(\boldsymbol{z})$ is invertible everywhere. There exist $\{(\boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)}, \tau_{(r)})\}_{r=1}^{||\boldsymbol{G}^z||_0}$ belonging to their respective support such that*

$$\operatorname{span}\left\{ D_z^{\tau_{(r)}} \boldsymbol{\lambda}(\boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)}) D\boldsymbol{s}(\boldsymbol{z}_{(r)}^{\tau_{(r)}})^{-1} \right\}_{r=1}^{||\boldsymbol{G}^z||_0} = \mathbb{R}^{d_z \times d_z}_{\boldsymbol{G}^z},$$

where $D_z^{\tau_{(r)}} \boldsymbol{\lambda}$ and $D\boldsymbol{s}$ are Jacobians with respect to $\boldsymbol{z}^{\tau_{(r)}}$ and $\boldsymbol{z}$, respectively.

In Appendix B.4, we show that the above assumption is implied by its nonparametric version (Assumption 8) when the transition model is in an exponential family with $k = 1$. However, Assumption 11 is *strictly weaker* than its nonparametric counterpart, Assumption 8. The reason is that in the former, $\boldsymbol{z}^{\tau_{(r)}}$ can vary for different p, whereas this is not allowed in the latter, since we require $\boldsymbol{z} = \boldsymbol{z}^{\tau_{(r)}}$ for all r.

The following theorem, extended from Lachapelle et al. (2022), shows that making stronger parametric assumptions on the transition model allows to weaken the sufficient influence assumption. Note that its structure is nearly identical to Theorem 3. Its proof can be found in Appendix B.3.

Theorem 5 (Disentanglement via sparse temporal dependencies in exponential families) *Let $\boldsymbol{\theta} := (\boldsymbol{f}, \boldsymbol{\lambda}, \boldsymbol{G})$ and $\hat{\boldsymbol{\theta}} := (\hat{\boldsymbol{f}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{G}})$ be two models satisfying Assumptions 1, 2, 3, 4, 9 as well as all assumptions of Theorem 4. Further suppose that*

1. *The sufficient statistic $\boldsymbol{s}$ is d_z-dimensional ($k = 1$) and is a diffeomorphism from $\mathbb{R}^{d_z}$ to $\boldsymbol{s}(\mathbb{R}^{d_z})$;*
2. ***[Sufficient influence of z]** The Jacobian of the ground-truth transition function $\boldsymbol{\lambda}$ with respect to $\boldsymbol{z}$ varies "sufficiently", as formalized in Assumption 11;*

Then, there exists a permutation matrix $\boldsymbol{P}$ such that $\boldsymbol{P}\boldsymbol{G}^z\boldsymbol{P}^\top \subseteq \hat{\boldsymbol{G}}^z$. Further assume that

3. ***[Sparsity regularization]** $||\hat{\boldsymbol{G}}^z||_0 \leq ||\boldsymbol{G}^z||_0$;*

Then, $\boldsymbol{\theta} \sim^{\boldsymbol{z}}_{\text{con}} \hat{\boldsymbol{\theta}}$ (Def. 14) & $\boldsymbol{\theta} \sim_{\text{lin}} \hat{\boldsymbol{\theta}}$ (Def. 17), which together implies that

$$\boldsymbol{v}(\boldsymbol{z}) = \boldsymbol{s}^{-1}(\boldsymbol{C}\boldsymbol{P}^\top \boldsymbol{s}(\boldsymbol{z}) + \boldsymbol{b}),$$

where $\boldsymbol{b} \in \mathbb{R}^{d_z}$ and $\boldsymbol{C} \in \mathbb{R}^{d_z \times d_z}$ is invertible, $\boldsymbol{G}^{\boldsymbol{z}}$- and $(\boldsymbol{G}^{\boldsymbol{z}})^\top$-preserving (Definition 11).

The reason we can simplify the sufficient influence assumption in the exponential family case has to do with the quasi-linear form of $\boldsymbol{v}$. Indeed, in that case, one can compute that the Jacobian of $\boldsymbol{v}$ takes a special form: $D\boldsymbol{v}(\boldsymbol{z}) = D\boldsymbol{s}(\boldsymbol{v}(\boldsymbol{z}))^{-1}\boldsymbol{L}D\boldsymbol{s}(\boldsymbol{z})$. Since $D\boldsymbol{s}$ is diagonal everywhere, one can see that the "non-diagonal part" of $D\boldsymbol{v}(\boldsymbol{z})$, i.e. $\boldsymbol{L}$, does not depend on $\boldsymbol{z}$, which simplifies the proof. See Appendix B.3 for more details.

In Appendix D.2, we discuss how Khemakhem et al. (2020a) & Yao et al. (2022b) obtain disentanglement guarantees and how their assumptions differ from ours.

Example 16 (Markovian sparse temporal dependencies without auxiliary variables) *Recall Remark 7 which pointed out that, without auxiliary variables, non-Markovianity was necessary to satisfy the nonparametric Assumption 8. We now illustrate that the analogous assumption specialized for exponential families with $k = 1$ (Assumption 11) is not as restrictive, i.e. it allows for Markovianity even when there are no auxiliary variables.*

We start from Example 7 which was based on the situation depicted in Figures 1 & 3e where the temporal graph $\boldsymbol{G}^z$ is lower triangular. Assume that no action is observed, i.e. we can only leverage the sparsity of $\boldsymbol{G}^z$ to disentangle. We now provide a concrete Markovian transition model $p(\boldsymbol{z}^t \mid \boldsymbol{z}^{t-1})$ that satisfies Assumption 11. Similarly to previous examples, assume $p(\boldsymbol{z}^t \mid \boldsymbol{z}^{t-1}) = \mathcal{N}(\boldsymbol{z}^t; \boldsymbol{\mu}(\boldsymbol{z}^{t-1}), \sigma^2\boldsymbol{I})$ where

$$\boldsymbol{\mu}(\boldsymbol{z}) := \boldsymbol{z} + \begin{bmatrix} \boldsymbol{z}_1^2/2 \\ \boldsymbol{z}_1^3/3 \\ \boldsymbol{z}_1^4/4 \end{bmatrix} + \begin{bmatrix} 0 \\ \boldsymbol{z}_2^2/2 \\ \boldsymbol{z}_2^3/3 \end{bmatrix} + \begin{bmatrix} 0 \\ 0 \\ \boldsymbol{z}_3^2/2 \end{bmatrix}.$$

Because the variance σ^2 is not influenced by $\boldsymbol{z}^{t-1}$, we can represent this transition model in an exponential family with $k = 1$ where the natural parameter is given by

$$\boldsymbol{\lambda}(\boldsymbol{z}^{t-1}) = \mu(\boldsymbol{z}^{t-1})/\sigma$$

and the sufficient statistic is given by $\boldsymbol{s}(\boldsymbol{z}) = \boldsymbol{z}/\sigma$. We can thus compute

$$D\boldsymbol{\lambda}(\boldsymbol{z})D\boldsymbol{s}(\boldsymbol{z})^{-1} = \boldsymbol{I} + \begin{bmatrix} \boldsymbol{z}_1 & 0 & 0 \\ \boldsymbol{z}_1^2 & \boldsymbol{z}_2 & 0 \\ \boldsymbol{z}_1^3 & \boldsymbol{z}_2^2 & \boldsymbol{z}_3 \end{bmatrix}$$

which spans the 6-dimensional space $\mathbb{R}^{3\times 3}_{\boldsymbol{G}^z}$, as shown in Example 14.

Connecting with the sufficient influence assumption of $\boldsymbol{a}$ in Lachapelle et al. (2022) The previous work of Lachapelle et al. (2022) could also leverage the sparse influence of $\boldsymbol{a}$ to disentangle and was based on exponential family and sufficient influence assumptions. In Appendix B.4, Proposition 12 shows that their sufficient influence assumption for $\boldsymbol{a}$ is actually equivalent to our nonparametric version (Assumption 7) in the exponential family case with $k = 1$. An important conclusion of this observation is that the identifiability result via sparse $\boldsymbol{G}^{\boldsymbol{a}}$ from Lachapelle et al. (2022), which was limited to the exponential family case with $k = 1$, can be derived from the more general nonparametric result of Theorem 2 we introduced earlier.

5 Model Estimation with a Sparsity Constraint

The identifiability results presented in this work are based on two crucial postulates: (i) the distribution over observations of both the learned and ground-truth models must match, i.e. $\hat{\boldsymbol{\theta}} \sim_{\text{obs}} \boldsymbol{\theta}$ (Definition 4), and (ii) the learned graphs $\hat{\boldsymbol{G}}^a$ and $\hat{\boldsymbol{G}}^z$ must be as sparse as their ground-truth counterparts, respectively $\boldsymbol{G}^a$ and $\boldsymbol{G}^z$. The theory suggests that in order to learn a (partially) disentangled representation, one should learn a model that satisfies these two requirements. In this section, we present one particular practical approach to achieve this. Appendix C.2 provides further details.

Data fitting. The first condition can be achieved by fitting a model to the data. Since the models discussed in this work present latent variable models, it is a natural idea to use a maximum likelihood approach based on the well-known framework of variational autoencoders (VAEs) (Kingma and Welling, 2014) in which the decoder neural network corresponds to the mixing function $\hat{\boldsymbol{f}}$. We consider an approximate posterior of the form

$$q(\boldsymbol{z}^{\leq T} \mid \boldsymbol{x}^{\leq T}, \boldsymbol{a}^{<T}) := \prod_{t=1}^{T} q(\boldsymbol{z}^t \mid \boldsymbol{x}^t)\,, \tag{32}$$

where $q(\boldsymbol{z}^t \mid \boldsymbol{x}^t)$ is a Gaussian distribution with mean and diagonal covariance outputted by a neural network $\texttt{encoder}(\boldsymbol{x}^t)$. In our experiments, the latent model $\hat{p}(\boldsymbol{z}_i^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ is a Gaussian distribution with mean $\hat{\boldsymbol{\mu}}_i(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ parameterized as a fully connected neural network that "looks" only at a fixed window of s lagged latent variables.[7] Furthermore, the variances are learned but does

7. The theory we developed would allow for a $\boldsymbol{\mu}$ function that depends on all previous time steps, not only the s previous ones. This could be achieved with a recurrent neural network or transformer, but we leave this to future work.

not depend on $(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ (see Appendix C.2 for details). This variational inference model induces the following evidence lower bound (ELBO) on $\log \hat{p}(\boldsymbol{x}^{\leq T}|\boldsymbol{a}^{<T})$:

$$\log \hat{p}(\boldsymbol{x}^{\leq T}|\boldsymbol{a}^{<T}) \geq \text{ELBO}(\hat{\boldsymbol{f}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{G}}, q; \boldsymbol{x}^{\leq T}, \boldsymbol{a}^{<T}) := \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q(\boldsymbol{z}^t|\boldsymbol{x}^t)} [\log \hat{p}(\boldsymbol{x}^t \mid \boldsymbol{z}^t)] - \mathop{\mathbb{E}}_{q(\boldsymbol{z}^{<t}|\boldsymbol{x}^{<t})} KL(q(\boldsymbol{z}^t \mid \boldsymbol{x}^t)||\hat{p}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}))\,. \quad (33)$$

We derive this fact in Appendix D.3. This lower bound can then be maximized using some variant of stochastic gradient ascent such as Adam (Kingma and Ba, 2015). We note that many works have proposed learning dynamical models with latent variables using VAEs (Girin et al., 2020), with various choice of architectures and approximate posteriors. Our specific choices were made out of a desire for simplicity, but the reader should be aware of other possibilities.

The learned distribution will exactly match the ground truth distribution if (i) the model has enough capacity to express the ground-truth generative process, (ii) the approximate posterior has enough capacity to express the ground-truth posterior $p(\boldsymbol{z}^t|\boldsymbol{x}^{\leq T}, \boldsymbol{a}^{<T})$, (iii) the dataset is sufficiently large, and (iv) the optimization finds the global optimum. If, in addition, the ground truth generative process satisfies the assumptions of Proposition 2, we can guarantee that the learned model $\hat{\boldsymbol{\theta}}$ will be equivalent to the ground truth model $\boldsymbol{\theta}$ up to diffeomorphism (Definition 5).

Learning $\hat{G}$ with sparsity constraints. To go from equivalence up to diffeomorphism to actual disentanglement (partial or not), Theorems 1, 2, 3 & 5 suggest that we should not only fit the data, but also choose the learned graph $\hat{\boldsymbol{G}}$ such that $||\hat{\boldsymbol{G}}^a||_0 \leq ||\boldsymbol{G}^a||_0$ and/or $||\hat{\boldsymbol{G}}^z||_0 \leq ||\boldsymbol{G}^z||_0$. In order to allow gradient-based optimization, our strategy consists in treating each edge $\hat{\boldsymbol{G}}_{i,j}$ as an independent Bernoulli random variable with probability of success $\sigma(\boldsymbol{\gamma}_{i,j})$, where σ is the sigmoid function and $\boldsymbol{\gamma}_{i,j}$ is a parameter learned using the Gumbel-Softmax trick (Jang et al., 2017; Maddison et al., 2017). Let $\text{ELBO}(\hat{\boldsymbol{f}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{G}}, q)$ be the ELBO objective of (33) averaged over the whole dataset. We tackle the following constrained optimization problem:

$$\max_{\hat{\boldsymbol{f}}, \hat{\boldsymbol{\mu}}, \boldsymbol{\gamma}, q} \mathbb{E}_{\hat{\boldsymbol{G}} \sim \sigma(\boldsymbol{\gamma})} \text{ELBO}(\hat{\boldsymbol{f}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{G}}, q) \;\text{ subject to }\; \mathbb{E}_{\hat{\boldsymbol{G}} \sim \sigma(\boldsymbol{\gamma})} ||\hat{\boldsymbol{G}}||_0 \leq \beta\,.$$

where β is a hyperparameter (which should ideally be set to $\beta^* := ||\boldsymbol{G}||_0$, i.e., the number of edges in the ground-truth graph) and $\hat{\boldsymbol{G}} \sim \sigma(\boldsymbol{\gamma})$ means that $\hat{\boldsymbol{G}}_{i,j}$ are independent and distributed according to $\sigma(\boldsymbol{\gamma}_{i,j})$. Because $\mathbb{E}_{\hat{\boldsymbol{G}} \sim \sigma(\boldsymbol{\gamma})} ||\hat{\boldsymbol{G}}||_0 = ||\sigma(\boldsymbol{\gamma})||_1$ where $\sigma(\boldsymbol{\gamma})$ is a matrix, the constraint becomes $||\sigma(\boldsymbol{\gamma})||_1 \leq \beta$. To solve this problem, we perform gradient descent-ascent on the Lagrangian function given by

$$\mathbb{E}_{\hat{\boldsymbol{G}} \sim \sigma(\boldsymbol{\gamma})} \text{ELBO}(\hat{\boldsymbol{f}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{G}}, q) - \alpha(||\sigma(\boldsymbol{\gamma})||_1 - \beta)$$

where the ascent step is performed w.r.t. $\hat{\boldsymbol{f}}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{G}}$ and q; and the descent step is performed w.r.t. Lagrangian multiplier α, which is forced to remain greater or equal to zero via a simple projection step. As suggested by Gallego-Posada et al. (2021), we perform *dual restarts* which simply means that, as soon as the constraint is satisfied, the Lagrangian multiplier is reset to 0. We used the `Cooper` library (Gallego-Posada and Ramirez, 2022), which implements many constrained optimization procedures in Python, including the one described above. Note that we use Adam (Kingma and Ba, 2015) for the ascent steps and standard gradient descent for the descent step on the Lagrangian multiplier α.

We also found empirically that the following schedule for β is helpful: We start training with $\beta = \max_{\hat{\boldsymbol{G}}} ||\hat{\boldsymbol{G}}||_0$ and linearly decrease its value until the desired number of edges is reached. This avoids getting a sparse graph too quickly while training, thus giving enough time to the model parameters to adjust. In each experiment, we trained for 300K iterations, and β takes 150K to go from its initial value to its desired value. We discuss how to select the hyperparameter β in Section 8.

6 Evaluation with R_{con} and SHD

In this section, we tackle the problem of quantitatively evaluating whether a learned representation $\hat{\boldsymbol{z}}$ is completely or partially disentangled with respect to the ground-truth representation $\boldsymbol{z}$, given a dataset of paired representations $\{(\boldsymbol{z}^i, \hat{\boldsymbol{z}}^i)\}_{i\in[N]}$. More precisely, we want to evaluate whether two models are $\boldsymbol{a}$-consistent or $\boldsymbol{z}$-consistent (Definitions 13 & 14). To achieve this, we must evaluate whether there exists a graph preserving map $\boldsymbol{c}$ (Definition 12) and a permutation matrix $\boldsymbol{P}$ such that for all $i \in [N]$, $\boldsymbol{z}^i = \boldsymbol{c}(\boldsymbol{P}^\top \hat{\boldsymbol{z}}^i)$. For evaluation purposes, we assume we observe the ground-truth latent representation for each observation, i.e., we have $\{(\boldsymbol{x}^i, \boldsymbol{z}^i)\}_{i\in N}$ sampled i.i.d. from the ground-truth data generating process. We will take $\hat{\boldsymbol{z}}^i := \texttt{encoder}(\boldsymbol{x}^i)$ where `encoder` is from the learned VAE model introduced in Section 5. For simplicity, we assume that $\boldsymbol{c}$ is affine.[8]

We start with evaluating complete disentanglement. A popular choice for this is the *mean correlation coefficient* (MCC), which is obtained by first computing the Pearson correlation matrix $\boldsymbol{K} \in \mathbb{R}^{d_z \times d_z}$ between the ground-truth representation and the learned representation ($\boldsymbol{K}_{i,j}$ is the correlation between $\boldsymbol{z}_i$ and $\hat{\boldsymbol{z}}_j$). Then $\text{MCC} := \max_{\boldsymbol{P}\in\text{permutations}} \frac{1}{d_z} \sum_{i=1}^{d_z} |(\boldsymbol{K}\boldsymbol{P})_{i,i}|$. We denote by $\hat{\boldsymbol{P}}$ the optimal permutation found by MCC.

To evaluate whether the learned representation is identified up to linear transformation (Definition 17), we perform linear regression to predict the ground-truth latent factors from the learned ones, and report the mean of the Pearson correlations between the predicted ground-truth latents and the actual ones. This metric is sometimes called the *coefficient of multiple correlation*, and happens to be the square root of the better known *coefficient of determination*, usually denoted by R^2. The advantage of using R instead of R^2 is that the former is comparable to MCC, and we always have $\text{MCC} \leq R$. Let us denote by $\hat{\boldsymbol{L}}$ the matrix of estimated coefficients, which should be thought of as an estimation of $\boldsymbol{L}$ in Definition 17 (assuming $\boldsymbol{s}(\boldsymbol{z}) = \boldsymbol{z}$, as is the case with Gaussian latents with fixed variance). Note that $\hat{\boldsymbol{L}}$ was fitted to the standardized $\boldsymbol{z}$ and $\hat{\boldsymbol{z}}$ (shifted and scaled to have mean 0 and variance 1). This yields coefficients $\hat{\boldsymbol{L}}_{i,j}$ that are directly comparable without changing the value of the R score. We visualize $\hat{\boldsymbol{L}}$ in Figures 6 & 8.

To evaluate whether the learned representation is $\boldsymbol{a}$-consistent or $\boldsymbol{z}$-consistent with the ground-truth (Definitions 13 & 14), as predicted by Theorems 1 & 3, we introduce a new metric, denoted by R_{con}. The idea behind R_{con} is to predict the ground-truth factors $\boldsymbol{z}$ from only the inferred factors $\hat{\boldsymbol{z}}$ that are allowed by the equivalence relations. For example, for $\boldsymbol{a}$-consistency (Definition 13), the relation between $\boldsymbol{z}$ and $\hat{\boldsymbol{z}}$ is given by $\boldsymbol{z} = \boldsymbol{c}(\boldsymbol{P}^\top \hat{\boldsymbol{z}})$ where $\boldsymbol{c}$ is a $\boldsymbol{G}^a$-preserving diffeomorphism. Since we assume for simplicity that $\boldsymbol{c}$ is affine, we have $\boldsymbol{z} = \boldsymbol{C}\boldsymbol{P}^\top \hat{\boldsymbol{z}} + \boldsymbol{b}$ where $\boldsymbol{C}$ is a $\boldsymbol{G}^a$-preserving matrix. The idea is then to estimate both $\boldsymbol{P}$ and $\boldsymbol{C}$ using samples $(\boldsymbol{z}, \hat{\boldsymbol{z}})$. The permutation $\boldsymbol{P}$ is estimated by $\hat{\boldsymbol{P}}$, which was found when computing MCC (Section 8). To estimate $\boldsymbol{C}$, we compute $\hat{\boldsymbol{z}}_{\text{perm}} := \hat{\boldsymbol{P}}^\top \hat{\boldsymbol{z}}$ and then compute the mask $\boldsymbol{M} \in \{0,1\}^{d_z \times d_z}$ specifying which entries of $\boldsymbol{C}$ are

8. This is not a simplification when the latent factors in the model and in the data-generating process are Gaussian with fixed variance and the assumptions of Theorem 4 hold. That is because the latent model is in the exponential family with sufficient statistic $\boldsymbol{s}(\boldsymbol{z}) = \boldsymbol{z}$ and, by Theorem 4, we must have $\boldsymbol{z} = \boldsymbol{s}^{-1}(\boldsymbol{L}\boldsymbol{s}(\hat{\boldsymbol{z}}) + \boldsymbol{b}) = \boldsymbol{L}\hat{\boldsymbol{z}} + \boldsymbol{b}$.

allowed to be nonzero, as required by the $\boldsymbol{G}^a$-preservation property (Proposition 3). Then, for every i, we predict the ground-truth $\boldsymbol{z}_i$ by performing linear regression only on the *allowed* factors, i.e., $\boldsymbol{M}_{i,\cdot} \odot \hat{\boldsymbol{z}}_{\text{perm}}$, and compute the associated coefficient of multiple correlations $R_{\text{con},i}$ and report the mean, i.e., $R_{\text{con}} := \frac{1}{d_z}\sum_{i=1}^{d_z} R_{\text{con},i}$. It is easy to see that we must have $R_{\text{con}} \leq R$, since R_{con} was computed with fewer features than R. Moreover, MCC $\leq R_{\text{con}}$, because MCC can be thought of as computing exactly the same thing as for R_{con}, but by predicting $\boldsymbol{z}_i$ only from $\hat{\boldsymbol{z}}_{\text{perm},i}$, i.e. with fewer features than R_{con}. This means that we always have $0 \leq \text{MCC} \leq R_{\text{con}} \leq R \leq 1$. This is a nice property allowing us to compare all three metrics together and reflects the hierarchy between equivalence relations. Note that R_{con} depends implicitly on the ground-truth graph, since the matrix $\boldsymbol{M}$ indicating which entries of $\boldsymbol{C}$ are forced to be zero by the equivalence relation depends on $\boldsymbol{G}$.

To compare the learned graph $\hat{\boldsymbol{G}}$ with the ground-truth $\boldsymbol{G}$, we report the (normalized) *structural Hamming distance* (SHD) between the ground-truth graph and the estimated graph *permuted by* $\hat{\boldsymbol{P}}$. More precisely, we report SHD $= (||\boldsymbol{G}^a - \hat{\boldsymbol{P}}^\top \hat{\boldsymbol{G}}^a||_0 + ||\boldsymbol{G}^z - \hat{\boldsymbol{P}}^\top \hat{\boldsymbol{G}}^z \hat{\boldsymbol{P}}||_0)/(d_a d_z + d_z^2)$, where $\hat{\boldsymbol{P}}$ is the permutation found by MCC and $(d_a d_z + d_z^2)$ is the maximal number of edges $\boldsymbol{G}$ can have.

7 Related Work

Causal discovery. Causal discovery is the problem of learning a causal graph from observational and possibly interventional data when all variables are observed, unlike in the present work where the representation $\boldsymbol{z}$ is latent. In this simpler setting, graph identifiability is still challenging since, without interventions, one can only identify the Markov equivalence class of the causal graph (thus leaving some edge orientation ambiguous), provided the distribution entailed by the causal graphical model is faithful to its causal graph, i.e. its conditional independences are precisely the ones induced by the causal graph. The faithfulness condition implies that we can infer precisely which d-separation statements hold in the causal graph by performing conditional independence tests on the observed distribution. This class of methods is known as *constrained-based* (Pearl, 2009; Peters et al., 2017). Another approach is to define a score function on the space of graphs, usually a sparsity-regularized maximum likelihood score, and search the space of graphs to maximize it. These are known as *score-based* methods (Chickering, 2003; Shimizu et al., 2006; Peters et al., 2014; Zheng et al., 2018). These approaches can be specialized for the interventional setting (Hauser and Bühlmann, 2012; Brouillard et al., 2020) and the temporal setting (Pamfil et al., 2020; Runge et al., 2019).

Linear and nonlinear ICA. The first results showing that latent variables can be identified up to permutation and rescaling at least date back to classical linear ICA which assumes a linear mixing function $\boldsymbol{f}$ and mutually independent and non-Gaussian latent variables (Jutten and Herault, 1991; Tong et al., 1993; Comon, 1994). Hyvärinen and Pajunen (1999) showed that when allowing $\boldsymbol{f}$ to be a general nonlinear transformation, a setting known as nonlinear ICA, mutual independence and non-Gaussianity alone are insufficient to identify the latent variables. This inspired multiple variations of nonlinear ICA that enabled identifiability by leveraging, e.g., nonstationarity (Hyvarinen and Morioka, 2016) and temporal dependencies (Hyvarinen and Morioka, 2017). Hyvärinen et al. (2019) generalized these works by introducing a data generating process in which latent variables are conditionally mutually independent given an observed *auxiliary variable* (corresponding to $\boldsymbol{a}$ in our work). These last three works rely on some form of *noise contrastive estimation* (NCE) (Gutmann and Hyvärinen, 2012), but similar identifiability results have also been shown for VAEs (Khe-

makhem et al., 2020a; Locatello et al., 2020; Klindt et al., 2021), normalizing flows (Sorrenson et al., 2020) and energy-based models (Khemakhem et al., 2020b). Kivva et al. (2022) showed that it is not necessary to observe the auxiliary variable to obtain disentanglement when the mixing function is piecewise affine and the latent factors are distributed according to a mixture of Gaussians with diagonal covariance.

Causal representation learning (static). Since the publication of the first iteration of this work in CLeaR 2022, the field now known as *causal representation learning* (CRL) (Schölkopf et al., 2021) gained significant traction. The prototypical problem of CRL is similar to nonlinear ICA in that the goal is to identify latent factors of variations but differs in that the latent variables are assumed to be related via a causal graphical model (CGM) and interventions on the latents are typically observed. Although some works assumed that the structure of the causal graph is known (Kocaoglu et al., 2018; Shen et al., 2022; Nair et al., 2019; Liang et al., 2023), significant progress has been achieved recently in the setting where the latent causal graph is unknown and must be inferred from single-node interventions targeting latent variables (Ahuja et al., 2023; Squires et al., 2023; Buchholz et al., 2023; von Kügelgen et al., 2023; Zhang et al., 2023; Jiang and Aragam, 2023; Varici et al., 2023b,a). See Remark 4 for a discussion contrasting these works with our interventional framework. In a similar spirit, Liu et al. (2023); Yang et al. (2021) leverage a form of nonstationarity that does not necessarily correspond to interventions, and Bengio et al. (2020) suggests using adaptation speed as a heuristic objective to disentangle latent factors in the bivariate case, although without identifiability guarantees. The above works do not support temporal dependencies, unlike the framework presented in this work. Although we do focus on temporal dependencies, the special case where $T = 1$ discussed in Remark 4 and fleshed out in Examples 10, 11 & 13 can be categorized as static CRL where the latent factors are independent (Assumption 2). This approach has been applied to single-cell data with gene perturbations (Lopez et al., 2023; Bereket and Karaletsos, 2023). Importantly, Example 13 illustrates how *multi-node* interventions on the latent factors can yield (partial) disentanglement in the independent factors regime. To the best of our knowledge, this constitutes the first identifiability guarantee from multi-node interventions with nonlinear mixing and should form an important step towards generalizing to arbitrary latent graphs. Note that Bing et al. (2023) recently proposed a disentanglement guarantee from multi-node interventions in the linear mixing setting.

CRL is closely related to methods that assume access to **paired observations** $(\boldsymbol{x}, \boldsymbol{x}')$ that are generated from a common decoder $\boldsymbol{f}$. These are in contrast with the works discussed above, which assume the samples from observational and interventional distributions are **unpaired**. In the *paired* data regime, Locatello et al. (2020) and Ahuja et al. (2022b) assume that only a small set of latent factors $S \subseteq [d_z]$ changes between $\boldsymbol{x}$ and $\boldsymbol{x}'$. Interestingly, Locatello et al. (2020) assume that for all i, $P(S \cap S' = \{i\}) > 0$ (for i.i.d S and S'), which resembles our graphical criterion for complete disentanglement (Definition 5). Karaletsos et al. (2016) proposed a related strategy based on triplets of observations and weak labels indicating which observation is closer to the reference in the (masked) latent space. Von Kügelgen et al. (2021) modeled the self-supervised setting with data augmentation using a similar idea and showed the block-identifiability of the latent variables shared among $\boldsymbol{x}$ and $\boldsymbol{x}'$. Brehmer et al. (2022) assumes that the latent variables are sampled from a structural causal model (SCM) (Peters et al., 2017) and that $\boldsymbol{x}'$ is *counterfactual* in the sense that it is generated using the same SCM and exogenous noise values as $\boldsymbol{x}$ except for some noises which are modified randomly. Similar approaches can also provide identifiability guarantees in the *multi-*

view setting where the decoders for the different views $\boldsymbol{x}$ and $\boldsymbol{x}'$ are allowed to be different (Gresele et al., 2020; Daunhawer et al., 2023). The paired observation setting bears some similarity with the temporal setting covered in this work since the pairs $(\boldsymbol{x}^t, \boldsymbol{x}^{t-1})$ are observed jointly. However, contrary to the above works, Theorems 3 & 5 allow all latent variables to change between $t-1$ and t, only the temporal dependencies between them are assumed to be sparse. Morioka and Hyvarinen (2023) can also be seen as paired CRL in which the latents of different views can interact causally in a restricted manner. Recently, Yao et al. (2023) generalized previous work by allowing more than two views.

Leveraging temporal dependencies or non-stationarity. Tong et al. (1990) proved the identifiability of linear ICA when latent factors $\boldsymbol{z}_i^t$ are correlated across time steps t but remain independent across components i, an idea that has been extended to nonlinear mixing (Hyvarinen and Morioka, 2017; Klindt et al., 2021; Schell and Oberhauser, 2023). Using our notation, these works assume a diagonal adjacency matrix $\boldsymbol{G}^z$ which contrasts with Theorems 3 & 5 which allow for general $\boldsymbol{G}^z$ (although some graphs might not yield complete disentanglement). Yao et al. (2022a, Theorem 1) also allows for general $\boldsymbol{G}^z$, but does not rely on the sparsity of $\boldsymbol{G}^z$ or sparse interventions on latent factors for identification. Instead, it relies on the conditional independence of $\boldsymbol{z}_i^t$ given $\boldsymbol{z}^{t-1}$ and on a "sufficient variability" condition involving the third cross-derivatives $\frac{\partial^3}{(\partial \boldsymbol{z}_i^t)^2 \partial \boldsymbol{z}_j^{t-1}} \log p(\boldsymbol{z}_i^t \mid \boldsymbol{z}^{t-1})$ which excludes simple Gaussian models with homoscedastic variance like the ones we considered in Examples 9, 10, 12 and in our experiments of Section 8. General non-stationarity of the latent distribution, i.e., that are not necessarily sparse like what is considered in this work, can also be used to identify the latent factors (Hyvarinen and Morioka, 2016; Hyvärinen et al., 2019; Khemakhem et al., 2020a; Hälvä and Hyvärinen, 2020; Morioka et al., 2021; Yao et al., 2022b,a), but these results require sufficient variability of higher-order derivatives/differences of the log-densities, which again typically exclude simple homoscedastic Gaussian models (see Appendix D.2 for more). Ahuja et al. (2022a) characterized the indeterminacies of the representation in latent dynamical models as the set of equivariances of the transition mechanism. In addition to temporal dependencies, one can also consider latent factors structured according to a spatial topology (Hälvä et al., 2021).

Dynamical causal representation learning: The previous iteration of this work (Lachapelle et al., 2022) concurrently with Lippe et al. (2022) introduced latent variables identifiability guarantees for dynamical latent models based on sparse interventions. Lippe et al. (2023b) later proposed a generalization in which instantaneous causal connections are allowed. The key differences from the present work are (i) Lippe et al. (2023b) considers interventions with *known targets*, while the present work (as well as Lachapelle et al. (2022)) considers interventions with *unknown targets*; (ii) Lachapelle et al. (2022, Theorem 5) and Theorems 3 & 5 do not necessarily need interventions to disentangle since they leverage sparsity of the temporal dependencies, contrary to Lippe et al. (2023b); (iii) Lippe et al. (2023b) allows instantaneous causal connections, unlike the present work; and (iv) Lippe et al. (2023b) demonstrates their approach on image data. We hypothesize that the *partially-perfect interventions* from Lippe et al. (2023b) in which instantaneous causal effects are disabled might be enough to get identifiability in our framework. We leave this for future work. The concurrent work of Volodin (2021) independently proposed an approach very similar to ours, which also learns a sparse latent causal graph over latent variables and actions using binary masks, but focuses on testing various algorithmic variants and empirically verifies that the approach works on interactive environments rather than formal identifiability guarantees. Lopez et al. (2023); Lei et al. (2023) found that such models adapt to sparse interventions more quickly than their entangled

counterparts. Keurti et al. (2023) discusses disentanglement in temporal regimes through the lens of group theory but does not provide identifiability guarantees. Recently, Lippe et al. (2023a) proposed a model similar to ours with disentanglement guarantees based on the constraint that the effect of the variable $\boldsymbol{a}^{t-1}$ (analogous to R in their work) on each $\boldsymbol{z}_i^t$ is mediated by a deterministic binary variable.

Constraining the decoder function $\boldsymbol{f}$. It is worth noting that one can also obtain disentanglement guarantees by constraining the decoder function $\boldsymbol{f}$ in some way (Taleb and Jutten, 1999; Gresele et al., 2021; Buchholz et al., 2022; Leemann et al., 2023; Lachapelle et al., 2023b; Horan et al., 2021). In particular, this can be achieved by imposing some form of sparsity on $\boldsymbol{f}$ (Moran et al., 2022; Zheng et al., 2022; Brady et al., 2023; Xi and Bloem-Reddy, 2023). In contrast, the present work assumes only that $\boldsymbol{f}$ is a general diffeomorphism onto its image. Note that Zheng et al. (2022) reused many proof strategies from the shorter version of this work (Lachapelle et al., 2022).

Disentanglement with explicit supervision. Some works leverage more explicit supervision to disentangle. For example, Ahuja et al. (2022c) assumes that the labels are given by a linear transformation of mutually independent and non-Gaussian latent factors. Instead of relying on independence, Lachapelle et al. (2023a); Fumero et al. (2023) leverage the sparsity of the linear map to disentangle.

Other relevant works on sparsity. The assumption that high-level variables are sparsely related to each other and/or to actions was discussed by Bengio (2019); Goyal and Bengio (2021); Ke et al. (2021). These ideas have also been leveraged by Goyal et al. (2021b,a); Madan et al. (2021) via attention mechanisms. Although these works are, in part, motivated by the same core assumption as ours, their focus is more on empirically verifying out-of-distribution generalization than it is on disentanglement (Definition 7) and formal identifiability results. The assumption that individual actions often affect only one factor of variation has been leveraged to disentangle Thomas et al. (2017). Loosely speaking, the theory we developed in the present work can be seen as a formal justification for such approaches.

8 Experiments

To illustrate our identifiability results and the benefit of mechanism sparsity regularization for disentanglement, we apply the sparsity regularized VAE method of Section 5 on various synthetic datasets. Section 8.1 focuses on graphs satisfying the criterion of Assumption 5 which, as we saw, guarantees complete disentanglement. We also verify experimentally that the sufficient influence assumptions are indeed important for disentanglement and explore latent model with both homoscedastic and heteroscedastic variance. Section 8.2 explores graphs that do not satisfy the criterion. Details about our implementation are provided in Appendix C.2 and the code used to run these experiments can be found here: `https://github.com/slachapelle/disentanglement_via_mechanism_sparsity`.

Synthetic datasets. The datasets we considered are separated into two groups: *Action* & *Time* datasets. The former group has only auxiliary variables, which we interpret as actions, without temporal dependence, so we fix $\hat{\boldsymbol{G}}^z = \boldsymbol{0}$. The latter group has only temporal dependence without actions, so we fix $\hat{\boldsymbol{G}}^a = \boldsymbol{0}$. In each dataset, the ground-truth mixing function $\boldsymbol{f}$ is a randomly initialized neural network. The dimensionalities of $\boldsymbol{z}$ and $\boldsymbol{x}$ are $d_z = 10$ and $d_x = 20$, respec-

tively. In the action datasets, the dimensionality of $\boldsymbol{a}$ is $d_a = 10$, unless otherwise specified. The ground-truth transition model $p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ is always a Gaussian with a mean outputted by some function $\boldsymbol{\mu}_{\boldsymbol{G}}(\boldsymbol{z}^{t-1}, a^{t-1})$ (the data is *Markovian*). For all datasets considered, the covariance matrix is given by $\sigma_z^2 \boldsymbol{I}$, i.e., the variance if *homoscedastic*, except for the datasets ActionNonDiag$_{k=2}$ and TimeNonDiag$_{k=2}$, which have *heteroscedastic* variance. Appendix C.1 provides more detailed descriptions of the datasets including the explicit form of $\boldsymbol{\mu}$ and $\boldsymbol{G}$ in each case. Note that the learned transition model $\hat{p}(\boldsymbol{z}^t \mid \boldsymbol{z}^{t-1}, \boldsymbol{a}^{t-1})$ is also a homoscedastic Gaussian where the mean function $\hat{\boldsymbol{\mu}}$ is an MLP.

Baselines. On the action datasets, we compare with TCVAE (Chen et al., 2018), iVAE (Khemakhem et al., 2020a). Only iVAE leverages the action. On the temporal datasets, we compare our approach with TCVAE, PCL (Hyvarinen and Morioka, 2017) and SlowVAE (Klindt et al., 2021). Only PCL and SlowVAE leverage the temporal dependencies. We also report the performance of a randomly initialized encoder (Random) and one trained via least-square regression directly on the ground-truth latent factors (Supervised). See Appendix C.3 for details on the baselines.

Unsupervised hyperparameter selection. In practice, the hyperparameters cannot be selected so as to optimize MCC, since this metric requires access to the ground-truth latent factors. Duan et al. (2020) introduced *unsupervised disentanglement ranking* (UDR) as a solution to unsupervised hyperparameter selection for disentanglement. Figures 5 & 7 show the performance of all approaches using UDR to select the hyperparameter (when it has one). For our approach, we show a range of sparsity bounds β and indicate the hyperparameter selected by UDR with a black star. Note that for our approach, we excluded from the UDR selection hyperparameters that yielded graphs with fewer edges than latent factors as a heuristic to prevent UDR from selecting overly sparse graphs. Figures 5 & 7 show that this *unsupervised* procedure selects a reasonable regularization level (as indicated by the black star), although not always the optimal one. See Appendix C.4 for more details.

8.1 Graphs Allowing Complete Disentanglement (Satisfying Assumption 5)

Satisfying sufficient influence assumptions. Figure 5 reports the MCC and R scores of all methods on four datasets that satisfy both the graphical criterion and the sufficient influence assumption: the datasets ActionDiag and TimeDiag have "diagonal" graphs, i.e., $\boldsymbol{G}^a = \boldsymbol{I}$ and $\boldsymbol{G}^a = \boldsymbol{I}$, while ActionNonDiag and TimeNonDiag present more involved graphs (depicted in Figure 6). **Observations:** We see that the sparsity constraint improves MCC and SHD on all datasets. Although most baselines obtain good R scores, which indicates that their representation encodes all the information about the factors of variations, they obtain poor MCC in comparison to our approach with a properly selected sparsity level, which indicates that they fail to disentangle. Moreover, the sparsity level selected by UDR (indicated by a black star) corresponds to the lowest SHD value for three out of four datasets and when it does not, it is still better than no sparsity at all. Figure 6 shows examples of estimated graphs. More details can be found in the caption.

Violating sufficient influence assumptions. The left column of Table 3 reports the performance of all methods on the ActionNonDiag$_{\text{NoSuffInf}}$ and TimeNonDiag$_{\text{NoSuffInf}}$ datasets, which are essentially the same as ActionNonDiag and TimeNonDiag but do not satisfy the sufficient influence assumptions (see Appendix C.1 for details). **Observations:** For the ActionNonDiag$_{\text{NoSuffInf}}$ dataset, we still see an improvement in MCC and SHD when regularizing for sparsity, but not as important

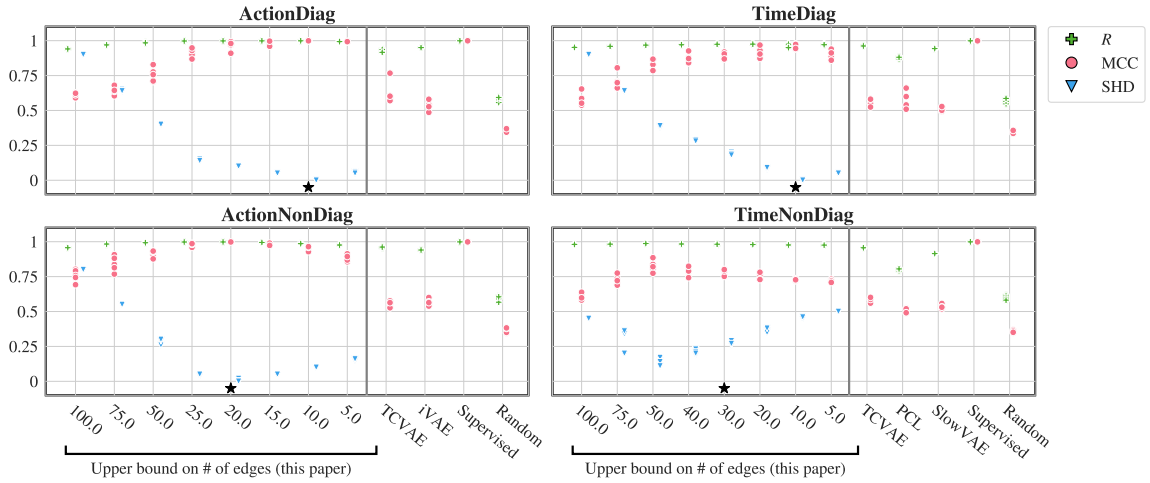


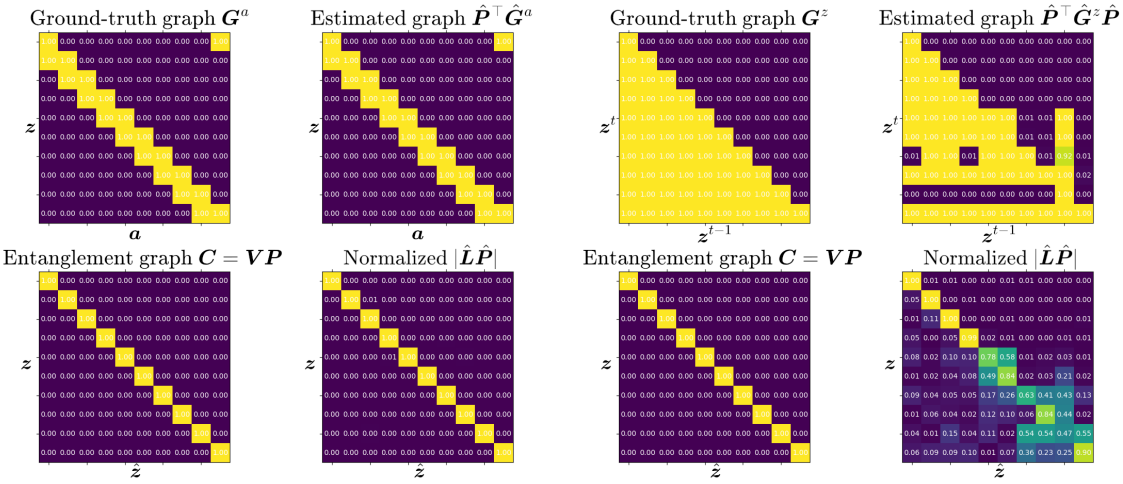
Figure 5: **Graphical criterion holds:** Datasets ActionDiag and TimeDiag have diagonal graphs while ActionNonDiag and TimeNonDiag have non-diagonal graphs. Sufficient influence is always satisfied. For our regularized VAE approach, we report performance for multiple sparsity levels β. In the left column, only $\hat{\boldsymbol{G}}^a$ is learned while in the right column, only $\hat{\boldsymbol{G}}^z$ is learned. For more details on the synthetic datasets, see Appendix C.1. The black star indicates which regularization parameter is selected by the filtered UDR procedure (see Appendix C.4). For R and MCC, higher is better. For SHD, lower is better. Performance is reported on 5 random seeds.

as for ActionNonDiag, which got MCC ≈ 1 and SHD ≈ 0. Nevertheless, our approach outperforms the baselines. For the $\text{TimeNonDiag}_{\text{NoSuffInf}}$ dataset, there is simply no improvement in MCC from sparsity regularization. In that case, SlowVAE (with the hyperparameter selected to maximize MCC) and PCL have a higher MCC. These observations confirm the importance of the sufficient influence assumptions.

Heteroscedastic variance ($k = 2$). The right column of Table 3 reports the performance of all methods on the $\text{ActionNonDiag}_{k=2}$ and $\text{TimeNonDiag}_{k=2}$ datasets, which are essentially the same as ActionNonDiag and TimeNonDiag but presents heteroscedastic variance, i.e., $\text{var}(\boldsymbol{z}^t \mid \boldsymbol{z}^{t-1}, \boldsymbol{a}^{t-1})$ is not a constant function of $(\boldsymbol{z}^{t-1}, \boldsymbol{a}^{t-1})$. This setting is interesting since it is not covered by the exponential family theory of Lachapelle et al. (2022) which assumed a one-dimensional sufficient statistic $\boldsymbol{s}$ ($k = 1$) whereas here we have $k = 2$. Both datasets fall under the umbrella of our nonparametric theory. However, $\text{TimeNonDiag}_{k=2}$ cannot satisfy the sufficient influence assumption because the data is Markovian and do not present an auxiliary variable (see Remark 7). **Observations:** Both datasets benefit from sparsity and outperform baselines. On $\text{ActionNonDiag}_{k=2}$ we obtain near perfect MCC and SHD while on $\text{TimeNonDiag}_{k=2}$ we obtain a performance similar to TimeNonDiag. We hypothesize that the performance bottleneck in both TimeNonDiag and $\text{TimeNonDiag}_{k=2}$ is the estimation of the graph, as in both cases SHD is always greater than $\approx 20\%$.

Datasets	$\text{ActionNonDiag}_{\text{NoSuffInf}}$			$\text{ActionNonDiag}_{k=2}$		
Metrics	SHD	MCC	R	SHD	MCC	R
iVAE	–	.61±.02	.97±.00	–	.59±.03	.94±.00
TCVAE (UDR)	–	.58±.03	.88±.01	–	.55±.02	.96±.00
TCVAE (MCC)	–	.61±.02	.96±.00	–	.55±.02	.96±.00
Ours (no sparsity)	.80±.00	.62±.02	.93±.00	.80±.00	.70±.03	.97±.00
Ours (sparsity)	.13±.03	.86±.04	1.0±.00	.03±.01	.98±.02	1.0±.00
Random	–	.37±.02	.63±.02	–	.37±.02	.60±.02
Supervised	–	1.0±.00	1.0±.00	–	1.0±.00	1.0±.00
Datasets	$\text{TimeNonDiag}_{\text{NoSuffInf}}$			$\text{TimeNonDiag}_{k=2}$		
Metrics	SHD	MCC	R	SHD	MCC	R
PCL	–	.66±.04	.96±.00	–	.58±.04	.83±.01
SlowVAE (UDR)	–	.59±.02	.98±.00	–	.57±.01	.93±.00
SlowVAE (MCC)	–	.71±.02	.98±.00	–	.58±.02	.95±.00
TCVAE (UDR)	–	.58±.03	.98±.00	–	.57±.01	.96±.00
TCVAE (MCC)	–	.58±.03	.98±.00	–	.57±.01	.96±.00
Ours (no sparsity)	.45±.00	.62±.04	.98±.00	.45±.00	.62±.01	.98±.00
Ours (sparsity)	.32±.05	.63±.03	.99±.00	.20±.07	.74±.04	.98±.00
Random	–	.40±.04	.67±.02	–	.36±.01	.59±.02
Supervised	–	1.0±.00	1.0±.00	–	1.0±.00	1.0±.00

Table 3: Datasets $\text{ActionNonDiag}_{\text{NoSuffInf}}$ and $\text{TimeNonDiag}_{\text{NoSuffInf}}$ do not satisfy their respective sufficient influence assumptions (Assumptions 6 & 11). Datasets $\text{ActionNonDiag}_{k=2}$ and $\text{TimeNonDiag}_{k=2}$ are such that $\text{var}(\boldsymbol{z}^t \mid \boldsymbol{z}^{t-1}, \boldsymbol{a}^{t-1})$ depends on $\boldsymbol{z}^{t-1}$ or $\boldsymbol{a}^{t-1}$ (which means the sufficient statistic has dimension $k = 2$, contrarily to all other datasets). For our method, we show performance both with and without the sparsity constraint. In the former case, the constraint is set to the number of edges in the ground-truth graph. For baselines that have hyperparameters, we report their performance with the hyperparameter configurations that maximize UDR and MCC.



(a) ActionNonDiag dataset, $\beta = 20$ (b) TimeNonDiag dataset, $\beta = 50$

Figure 6: For each dataset, we visualize the median SHD run among the five randomly initialized runs of Figure 5 with the sparsity level β that is the closest to the ground-truth sparsity level $||\boldsymbol{G}||_0$. For each dataset, we visualize (i) the ground-truth graph, (ii) the permuted estimated graph, (iii) the entanglement graph predicted by our theory, and (iv) the permuted matrix of regression coefficients in absolute value normalized by the maximum coefficient i.e. $|\hat{\boldsymbol{L}}\hat{\boldsymbol{P}}|/\max_{i,j}|\hat{\boldsymbol{L}}_{i,j}|$. In Figure 6a, the estimated graph is exactly the ground-truth and $|\hat{\boldsymbol{L}}\hat{\boldsymbol{P}}|$ is perfectly diagonal, indicating complete disentanglement. In Figure 6b, the learned graph is close but not equal to the ground-truth. We can see that the off-diagonal nonzero values in $|\hat{\boldsymbol{L}}\hat{\boldsymbol{P}}|$ align with the poorly estimated parts of the graph.

8.2 Graphs Allowing only Partial Disentanglement (Violating Assumption 5)

In this section, we explore datasets with graphs that do not satisfy the criterion of Assumption 5. This means that our theory can only guarantee a form of partial disentanglement. For this reason, we will report the R_{con} metric introduced in Section 6 which measures whether two representations are $\boldsymbol{a}$-consistent (Definition 13) or $\boldsymbol{z}$-consistent (Definition 14).

Satisfying sufficient influence assumptions. Figure 7 reports the MCC, R_{con} and R scores of all methods on four datasets that satisfy the sufficient influence assumption but not the graphical criterion: the datasets ActionBlockDiag and TimeBlockDiag have "block diagonal" graphs, while ActionBlockNonDiag and TimeBlockNonDiag have more intricate graphs (depicted in Figure 8). See Appendix C.1 for details on the datasets. **Observations:** In all four datasets, some sparsity level yields near perfect R_{con}, indicating the learned models are approximately $\boldsymbol{a}$-consistent or $\boldsymbol{z}$-consistent to the ground-truth. Moreover, SHD is correlated with R_{con}. Without surprise, MCC never comes close to one, since complete disentanglement is not guaranteed by our theory. Similarly to Figure 5, the baselines have decent R values but very low MCC and R_{con}, indicating that they cannot achieve partial disentanglement. Figure 8 shows examples of estimated graphs. When it comes to hyperparameter selection, UDR selects the hyperparameter with the lowest SHD on three out of four datasets, which indicates that UDR does reasonably well.

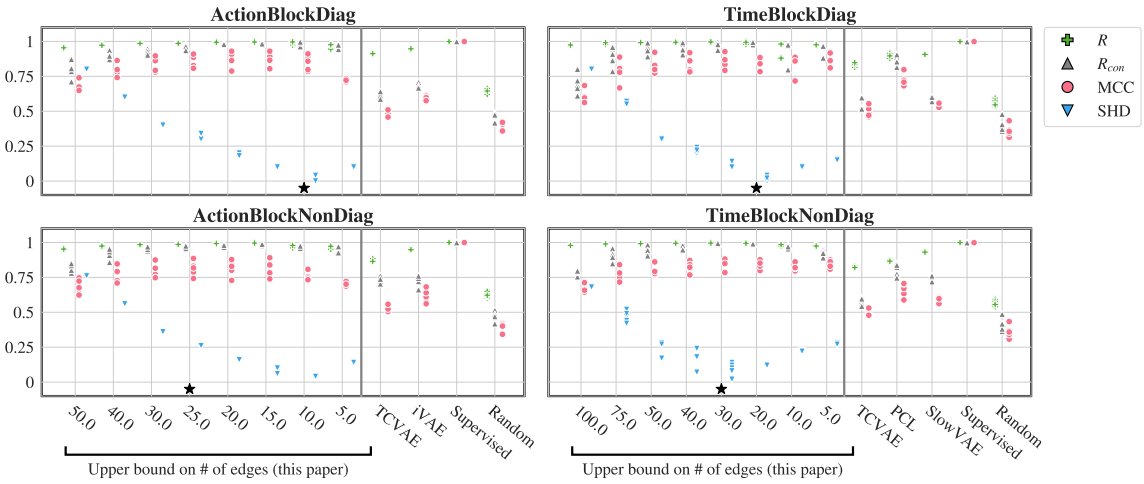
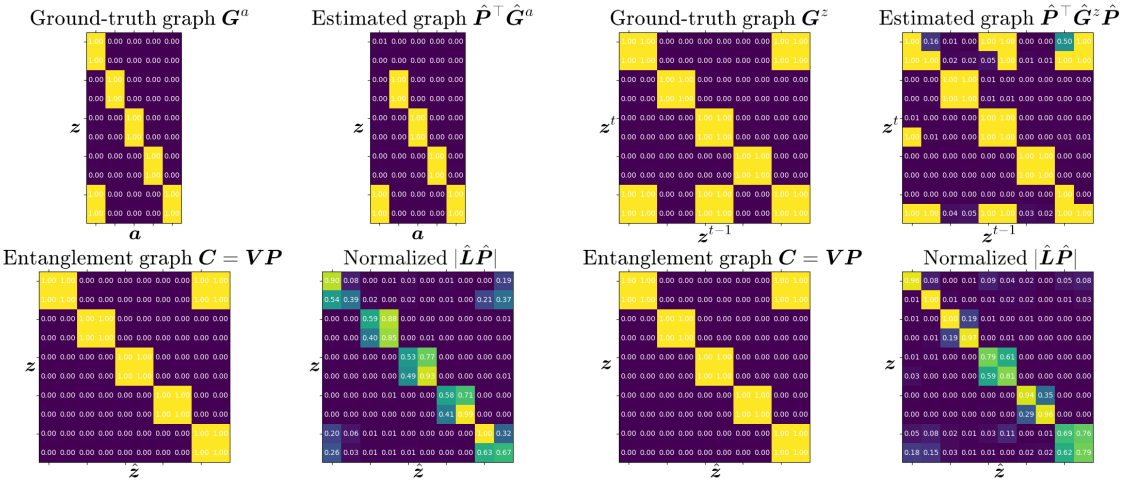


Figure 7: **Graphical criterion does not hold:** Datasets ActionBlockDiag and TimeBlockDiag have block-diagonal graphs while ActionBlockNonDiag and TimeBlockNonDiag have non-diagonal graphs. Sufficient influence is always satisfied. In the left column, only $\hat{\boldsymbol{G}}^a$ is learned and we vary β_a, and in the right column, only $\hat{\boldsymbol{G}}^z$ is learned and we vary β_z. For more details on the synthetic datasets, see Appendix C.1. The black star indicates which regularization parameter is selected by the filtered UDR procedure (see Appendix C.4). For R and MCC, higher is better. For SHD, lower is better. Performance is reported on 5 random seeds.

ActionRandomGraphs								
	Without sparsity			**With sparsity**				
p(edge)	MCC	R_{con}	R	MCC	R_{con}	R	SHD	$\mathbb{E}\|\|\boldsymbol{V}\|\|_0$
10%	.58±.13	.61±.11	.68±.13	.69±.14	.70±.12	.70±.12	.00±.00	45.0
20%	.67±.06	.69±.05	.83±.08	.85±.08	.86±.08	.86±.09	.01±.01	25.8
40%	.67±.03	.70±.03	.93±.04	.94±.05	.95±.05	.98±.04	.06±.05	15.8
60%	.69±.06	.73±.05	.96±.00	.88±.07	.91±.05	.99±.01	.14±.08	15.8
90%	.63±.04	.81±.08	.97±.00	.60±.01	.78±.07	.97±.00	.21±.07	45.0

TimeRandomGraphs								
	Without sparsity			**With sparsity**				
p(edge)	MCC	R_{con}	R	MCC	R_{con}	R	SHD	$\mathbb{E}\|\|\boldsymbol{V}\|\|_0$
10%	.66±.03	.66±.03	.98±.00	1.0±.00	1.0±.00	1.0±.00	.01±.02	10.2
20%	.63±.05	.63±.05	.98±.00	.99±.01	.99±.01	.99±.00	.08±.09	10.2
40%	.61±.02	.61±.02	.98±.00	.82±.16	.82±.16	.98±.01	.27±.13	10.2
60%	.58±.02	.58±.02	.98±.00	.71±.12	.71±.12	.98±.00	.33±.06	10.4
90%	.58±.03	.63±.08	.98±.00	.58±.02	.63±.08	.98±.00	.20±.07	26.1

Table 4: Experiments with randomly generated graphs. The probability of sampling an edge is p(edge). We report an estimation of $\mathbb{E}||\boldsymbol{V}||_0$ which is the average number of edges in the entanglement graph $\boldsymbol{V}$ entailed by the random ground-truth graph $\boldsymbol{G}$.



(a) ActionBlockNonDiag dataset, $\beta = 10$ (b) TimeBlockNonDiag dataset, $\beta = 30$

Figure 8: For each dataset, we visualize the median SHD run among the five randomly initialized runs of Figure 7 with the sparsity level β that is the closest to the ground-truth sparsity level $||\boldsymbol{G}||_0$. For each dataset, we visualize (i) the ground-truth graph, (ii) the permuted estimated graph, (iii) the entanglement graph predicted by our theory, and (iv) the permuted matrix of regression coefficients in absolute value normalized by the maximum coefficient i.e. $|\hat{\boldsymbol{L}}\hat{\boldsymbol{P}}|/\max_{i,j} |\hat{\boldsymbol{L}}_{i,j}|$. For both datasets, the learned graph is very close to the ground-truth. Furthermore, the match between the zero entries of $\hat{\boldsymbol{L}}\hat{\boldsymbol{P}}$ and those of the theoretical entanglement graph $\boldsymbol{C}$ is very good, although not perfect. Notice how certain blocks of latent factors remain entangled, as predicted by the theory.

Random graphs of varying sparsity levels. In Table 4, we consider the same $\boldsymbol{\mu}$ functions as in datasets ActionNonDiag and TimeNonDiag, but explore more diverse randomly generated ground-truth graphs with various degrees of sparsity. Edges are sampled i.i.d. with some probability $p(\text{edge})$. However, note that for the TimeRandomGraphs dataset, the self-loops are present with probability one. We report the performance of our approach both with and without sparsity regularization. When using sparsity, we set β equal to the ground-truth number of edges, $||\boldsymbol{G}||_0$. **Observations:** First, all datasets obtain an improvement in MCC and R_{con} from sparsity regularization, except for very dense graphs with $p(\text{edge}) = 90\%$, in which case regularization does nothing or slightly degrades performance. Secondly, we can see that the SHD tends to be higher for larger graphs, suggesting that these are harder to learn. Third, in the ActionRandomGraphs datasets, we can see a negative correlation between MCC and $\mathbb{E}||\boldsymbol{V}||_0$, which is expected since $||\boldsymbol{V}||_0$ close to 10 means complete disentanglement is possible (assuming the graph is learned properly). This pattern also appears to some extent in the TimeRandomGraphs datasets. Notice how, among the TimeRandomGraphs datasets, all datasets sparser than $p(\text{edge}) = 90\%$ always have $\mathbb{E}||\boldsymbol{V}||_0 \approx 10$, indicating complete disentanglement should be possible.[9] This is confirmed by very high MCC, at least for sparser graphs which are learned properly. Finally, we note that the R score is low for the very sparse action datasets. We suspect this is because very sparse graphs are less likely to satisfy the assumption of sufficient variability (Theorem 4), which guarantees quasi-linear equivalence (here it is actually *linear* equivalence, because of Gaussianity). Indeed, for very sparse graphs, some latent factors might end up without parents. This is not the case in the time datasets because of the self-loops which are always present.

9 Conclusion

This work proposed a novel principle for disentanglement based on *mechanism sparsity regularization*. The idea is based on the assumption that the mechanisms that govern the dynamics of high-level concepts are often sparse: actions usually affect only a few entities, and objects usually interact sparsely with each other. We provided novel nonparametric identifiability guarantees for this setting which give sufficient conditions for disentanglement, whether complete or partial. Given the dependency structure between latent factors and auxiliary variables, our theory predicts the entanglement graph that describes which of the estimated latent factors are expected to remain entangled. This constitutes a significant extension of the shorter version of this work (Lachapelle et al., 2022). We further provide various examples to illustrate the consequences of our guarantees as well as the assumptions they rely on. For instance, we show that multi-node interventions with unknown targets fall under the umbrella of our framework. Finally, we demonstrate the theory experimentally by training a sparsity-constrained variational autoencoder on synthetic data, which allows us to explore various settings. Our work establishes a solid theoretical foundation for further empirical investigations in more realistic scenarios, such as single-cell data with gene perturbations (Lopez et al., 2023) and video (Lei et al., 2023). Future works include relaxing assumptions such as conditional independence or considering more permissive settings such as "contextual sparsity", i.e., the assumption that objects only interact with each other in particular situations. We believe that the latter could be formalized and leveraged for disentanglement using the tools developed in this work.

9. We suspect this occurs because the self-loops which are present with probability one, unlike the Action dataset.

Acknowledgments and Disclosure of Funding

The authors would like to thank Aristide Baratin for important feedback on the manuscript. This research was partially supported by the Canada CIFAR AI Chair Program, by an IVADO excellence PhD scholarship, by a Google Focused Research award, the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039A, and by Mitacs through the Mitacs Accelerate program. The experiments were in part enabled by computational resources provided by Calcul Quebec and the Digital Research Alliance of Canada. The authors would like to thank Yoshua Bengio for inspiring mechanism sparsity regularization through various talks and discussions. The authors would also like to thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Yash Sharma. Simon Lacoste-Julien is a CIFAR Associate Fellow in the Learning in Machines & Brains program.

		Calligraphic & indexing conventions
$[n]$	:=	$\{1, 2, \ldots, n\}$
x		Scalar (random or not, depending on context)
$\boldsymbol{x}$		Vector (random or not, depending on context)
$\boldsymbol{X}$		Matrix
$\mathcal{X}$		Set/Support
f		Scalar-valued function
$\boldsymbol{f}$		Vector-valued function
Df, $D\boldsymbol{f}$		Jacobian of f and $\boldsymbol{f}$
$D^2 f$		Hessian of f
$B \subseteq [n]$		Subset of indices
$\boldsymbol{x}_B$		Vector formed with the ith coordinates of $\boldsymbol{x}$, for all $i \in B$
$\boldsymbol{X}_{B,B'}$		Matrix formed with the entries $(i, j) \in B \times B'$ of $\boldsymbol{X}$.
		Recurrent notation
$\boldsymbol{x}^t \in \mathbb{R}^{d_x}$		Observation at time t
$\boldsymbol{x}^{\leq t} \in \mathbb{R}^{d_x \times t}$		Matrix of observations at times $1, \ldots, t$
$\boldsymbol{z}^t \in \mathbb{R}^{d_z}$		Vector of latent factors of variations at time t
$\boldsymbol{z}^{\leq t} \in \mathbb{R}^{d_z \times t}$		Matrix of latent vectors at times $1, \ldots, t$
$\boldsymbol{a}^t \in \mathbb{R}^{d_a}$		Vector of auxiliary variables at time t
$\boldsymbol{a}^{<t} \in \mathbb{R}^{d_a \times t}$		Matrix of auxiliary vectors at times $0, 1 \ldots, t-1$
$\mathcal{A} \subseteq \mathbb{R}^{d_a}$		Support of $\boldsymbol{a}^t$
$\boldsymbol{f} : \mathbb{R}^{d_z} \to \mathbb{R}^{d_x}$		Ground-truth decoder function
$\hat{\boldsymbol{f}} : \mathbb{R}^{d_z} \to \mathbb{R}^{d_x}$		Learned decoder function
$p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$		Ground-truth latent transition model
$\hat{p}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$		Learned latent transition model
$\boldsymbol{G}^a \in \{0,1\}^{d_z \times d_a}$		Ground-truth adjacency matrix of graph connecting $\boldsymbol{a}^{<t}$ to $\boldsymbol{z}^t$
$\boldsymbol{G}^z \in \{0,1\}^{d_z \times d_z}$		Ground-truth adjacency matrix of graph connecting $\boldsymbol{z}^{<t}$ to $\boldsymbol{z}^t$
$\hat{\boldsymbol{G}}^a, \hat{\boldsymbol{G}}^z$		Learned adjacency matrices
$\mathbf{Pa}_i^a \subseteq [d_a]$		Parents of $\boldsymbol{z}_i^t$ in $\boldsymbol{G}^a$
$\mathbf{Ch}_\ell^a \subseteq [d_z]$		Children of $\boldsymbol{a}_\ell^t$ in $\boldsymbol{G}^z$
$\mathbf{Pa}_i^z \subseteq [d_z]$		Parents of $\boldsymbol{z}_i^t$ in $\boldsymbol{G}^z$
$\mathbf{Ch}_i^z \subseteq [d_z]$		Children of $\boldsymbol{z}_i^{t-1}$ in $\boldsymbol{G}^z$
$D_z^t \log p \in \mathbb{R}^{1 \times d_z}$		Jacobian vector of $\log p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ w.r.t. $\boldsymbol{z}^t$
$H_{z,a}^{t,\tau} \log p \in \mathbb{R}^{d_z \times d_a}$		Hessian matrix of $\log p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ w.r.t. $\boldsymbol{z}^t$ and $\boldsymbol{a}^\tau$
$H_{z,z}^{t,\tau} \log p \in \mathbb{R}^{d_z \times d_z}$		Hessian matrix of $\log p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ w.r.t. $\boldsymbol{z}^t$ and $\boldsymbol{z}^\tau$
$\sigma : [d_z] \to [d_z]$		A permutation
		Topology
$\overline{\mathcal{X}}$		Closure of the set $\mathcal{X} \subseteq \mathbb{R}^n$
$\mathcal{X}^\circ$		Interior of the set $\mathcal{X} \subseteq \mathbb{R}^n$

Table 5: Table of Notation.

Appendix A. Identifiability Theory - Nonparametric Case

A.1 Useful Lemmas

Definition 18 (Regular closed set) *A set $A \subseteq \mathbb{R}^n$ is regular closed when it is equal to the closure of its interior, i.e. $\overline{A^\circ} = A$.*

Lemma 4 *Let $A \subseteq \mathbb{R}^n$ and $\boldsymbol{f} : A \to \mathbb{R}^m$ be a C^k function. Then, its k first derivatives is uniquely defined on $\overline{A^\circ}$ in the sense that they do not depend on the specific choice of C^k extension.*

Proof Let $\boldsymbol{g} : U \to \mathbb{R}^n$ and $\boldsymbol{h} : V \to \mathbb{R}^n$ be two C^k extensions of $\boldsymbol{f}$ to $U \subseteq \mathbb{R}^n$ and $V \subseteq \mathbb{R}^n$ both open in $\mathbb{R}^n$. By definition,

$$\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{h}(\boldsymbol{x}),\ \forall \boldsymbol{x} \in A\,.$$

The usual derivative is uniquely defined on the interior of the domain, so that

$$D\boldsymbol{g}(\boldsymbol{x}) = D\boldsymbol{f}(\boldsymbol{x}) = D\boldsymbol{h}(\boldsymbol{x}),\ \forall \boldsymbol{x} \in A^\circ\,.$$

Consider a point $\boldsymbol{x}_0 \in \overline{A^\circ}$. By definition of closure, there exists a sequence $\{\boldsymbol{x}_k\}_{k=1}^\infty \subseteq A^\circ$ s.t. $\lim_{k\to\infty} \boldsymbol{x}_k = \boldsymbol{x}_0$. We thus have that

$$\begin{aligned}\lim_{k\to\infty} D\boldsymbol{g}(\boldsymbol{x}_k) &= \lim_{k\to\infty} D\boldsymbol{h}(\boldsymbol{x}_k)\\ D\boldsymbol{g}(\boldsymbol{x}_0) &= D\boldsymbol{h}(\boldsymbol{x}_0)\,,\end{aligned}$$

where we used the fact that the derivatives of $\boldsymbol{g}$ and $\boldsymbol{h}$ are continuous to go to the second line. Thus, all the C^k extensions of $\boldsymbol{f}$ must have equal derivatives on $\overline{A^\circ}$. This means we can unambiguously define the derivative of $\boldsymbol{f}$ everywhere on $\overline{A^\circ}$ to be equal to the derivative of one of its C^k extensions.

Since $\boldsymbol{f}$ is C^k, its derivative $D\boldsymbol{f}$ is C^{k-1}, we can thus apply the same argument to get that the second derivative of $\boldsymbol{f}$ is uniquely defined on $\overline{\overline{A^\circ}^\circ}$. It can be shown that $\overline{\overline{A^\circ}^\circ} = \overline{A^\circ}$. One can thus apply the same argument recursively to show that the first k derivatives of $\boldsymbol{f}$ are uniquely defined on $\overline{A^\circ}$. ■

Lemma 5 *Let X be some set. A family of functions $(f_i : X \to \mathbb{R})_{i=1}^n$ is linearly independent if and only if there exists $x_1, ..., x_n \in X$ such that the family of vectors $((f_1(x_i), ..., f_n(x_i)))_{i=1}^n$ is linearly independent.*

Proof We start by proving the "if" part. Assume the functions are linearly dependent. Then there exists $\boldsymbol{\alpha} \neq 0$ such that, for all $x \in X$, $\sum_{i=1}^n \boldsymbol{\alpha}_i f_i(x) = 0$. Choose distinct $x_1, ..., x_n \in X$. We thus have that for all $j \in [n]$, $\sum_{i=1}^n \boldsymbol{\alpha}_i f_i(x_j) = 0$. This can be written in matrix form:

$$\begin{bmatrix} f_1(x_1) & \cdots & f_1(x_n) \\ \vdots & \ddots & \vdots \\ f_n(x_1) & \cdots & f_n(x_n) \end{bmatrix} \boldsymbol{\alpha} = \mathbf{0}\,,$$

which implies that the columns are linearly dependent.

We now show the "only if" part. Suppose that for all $\{x_1, \dots, x_n\} \subseteq X$, the family of vectors $((f_1(x_i), ..., f_n(x_i)))_{i=1}^n$ is linearly dependent. This means that the set $U = \text{span}\{(f_1(x), ..., f_n(x)) \mid x \in X\}$ is a proper linear subspace of $\mathbb{R}^n$. This means that there is a nonzero $u \in U^\perp$, the orthogonal complement of U. By definition, u is orthogonal to all elements in $\{(f_1(x), ..., f_n(x)) \mid x \in X\}$. In other words, for all $x \in X$, $\sum_{i=1}^n u_i f_i(x) = 0$. Hence the $\boldsymbol{f}_i$ are linearly dependent. ■

A.2 Proof of Proposition 1

Proposition 1 (Linking dependency graph and Jacobian) *Let $\boldsymbol{h}$ be a C^1 function, i.e. continuously differentiable, from $\mathbb{R}^n$ to $\mathbb{R}^m$ and let $\boldsymbol{H}$ be its dependency graph (Definition 2). Then,*

$$\boldsymbol{H}_{i,j} = 0 \iff \textit{For all } \boldsymbol{a} \in \mathbb{R}^n\,, D\boldsymbol{h}(\boldsymbol{a})_{i,j} = 0\,. \tag{4}$$

Proof The " $\implies$ " direction holds since we can simply differentiate $\boldsymbol{h}_i(\boldsymbol{a}) = \bar{\boldsymbol{h}}_i(\boldsymbol{a}_{-j})$ w.r.t. $\boldsymbol{a}_j$ to get zero.

We now show the " $\impliedby$ " direction. Suppose that for all $\boldsymbol{a} \in \mathbb{R}^n$, $D\boldsymbol{h}(\boldsymbol{a})_{i,j} = 0$. We must now show that $\boldsymbol{h}_i(\boldsymbol{a})$ is constant in $\boldsymbol{a}_j$ for all $\boldsymbol{a}_{-j}$. Choose any $\boldsymbol{a}^0, \boldsymbol{a}^1 \in \mathbb{R}^n$ such that $\boldsymbol{a}^0_{-j} = \boldsymbol{a}^1_{-j}$. Thanks to the fundamental theorem of calculus, we can write

$$\begin{aligned}
\boldsymbol{h}_i(\boldsymbol{a}^1) - \boldsymbol{h}_i(\boldsymbol{a}^0) &= \int_{[0,1]} \frac{d}{d\alpha} \boldsymbol{h}_i((1-\alpha)\boldsymbol{a}^0 + \alpha \boldsymbol{a}^1))d\alpha \\
&= \int_{[0,1]} \underbrace{D\boldsymbol{h}((1-\alpha)\boldsymbol{a}^0 + \alpha\boldsymbol{a}^1)_{i,\cdot}}_{\text{zero at } j} \cdot \underbrace{(\boldsymbol{a}^1 - \boldsymbol{a}^0)}_{\text{zero except at } j} d\alpha \\
&= 0\,.
\end{aligned}$$

Since $\boldsymbol{a}^0$ and $\boldsymbol{a}^1$ were arbitrary points such that $\boldsymbol{a}^0_{-j} = \boldsymbol{a}^1_{-j}$, this means the function $\boldsymbol{h}_i(\boldsymbol{a})$ is constant in $\boldsymbol{a}_j$ for all values of $\boldsymbol{a}_{-j}$. ■

A.3 Proof of Proposition 2

In this section, we prove Proposition 2. Before doing so, we first recall the definition of the support of a random variable (Definition 19) and prove a useful lemma (Lemma 6).

Definition 19 *(Support of a random variable) Let $\boldsymbol{x}$ be a random variable with values in $\mathbb{R}^n$ with distribution $\mathbb{P}_{\boldsymbol{x}}$. Let $\mathcal{O}_n$ be the standard topology of $\mathbb{R}^n$ (i.e. the set of open sets of $\mathbb{R}^n$). The support of $\boldsymbol{x}$ is defined as*

$$\text{supp}(\boldsymbol{x}) := \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{x} \in O \in \mathcal{O}_n \implies \mathbb{P}_{\boldsymbol{x}}(O) > 0\}\,.$$

Lemma 6 *Let $\boldsymbol{z}$ be a random variable with values in $\mathbb{R}^m$ with distribution $\mathbb{P}_{\boldsymbol{z}}$ and $\boldsymbol{y} := \boldsymbol{f}(\boldsymbol{z})$ where $\boldsymbol{f} : \text{supp}(\boldsymbol{z}) \to \mathbb{R}^n$ is a homeomorphism onto its image. Then*

$$\boldsymbol{f}(\text{supp}(\boldsymbol{z})) \subseteq \text{supp}(\boldsymbol{y}) \subseteq \overline{\boldsymbol{f}(\text{supp}(\boldsymbol{z}))}\,.$$

where the closure is taken w.r.t. the topology of $\mathbb{R}^n$.

Proof We first prove that $f(\text{supp}(\boldsymbol{z})) \subseteq \text{supp}(\boldsymbol{y})$. Let $\boldsymbol{y}^0 \in \boldsymbol{f}(\text{supp}(\boldsymbol{z}))$ and N be an open neighborhood of $\boldsymbol{y}^0$, i.e. $\boldsymbol{y}^0 \in N \in \mathcal{O}_n$. Note that there exists $\boldsymbol{z}^0 \in \text{supp}(\boldsymbol{z})$ such that $\boldsymbol{f}(\boldsymbol{z}^0) = \boldsymbol{y}^0$. Note that $\boldsymbol{z}^0 \in \boldsymbol{f}^{-1}(\{\boldsymbol{y}^0\}) \subseteq \boldsymbol{f}^{-1}(N)$ and that, by continuity of $\boldsymbol{f}$, $\boldsymbol{f}^{-1}(N)$ is an open neighborhood of $\boldsymbol{z}^0$. Since $\boldsymbol{z}^0 \in \text{supp}(\boldsymbol{z})$, we have

$$\begin{aligned} 0 &< \mathbb{P}_{\boldsymbol{z}}(f^{-1}(N)) \\ &= \mathbb{P}_{\boldsymbol{z}} \circ f^{-1}(N) \\ &= \mathbb{P}_{\boldsymbol{y}}(N)\,. \end{aligned}$$

Hence $\boldsymbol{y}^0 \in \text{supp}(\boldsymbol{y})$, which concludes the "$\subseteq$" part.

We now prove the other inclusion. Let $\boldsymbol{y}^0 \in \text{supp}(\boldsymbol{y})$ and suppose, by contradiction, that $\boldsymbol{y}^0 \notin \overline{\boldsymbol{f}(\text{supp}(\boldsymbol{z}))}$. Since $\overline{\boldsymbol{f}(\text{supp}(\boldsymbol{z}))}$ is closed in $\mathbb{R}^n$, there exists N s.t. $\boldsymbol{y}^0 \in N \in \mathcal{O}_n$ with $N \cap \boldsymbol{f}(\text{supp}(\boldsymbol{z})) = \emptyset$. Since $\boldsymbol{y}^0 \in \text{supp}(\boldsymbol{y})$,

$$\begin{aligned} 0 &< \mathbb{P}_{\boldsymbol{y}}(N) \\ &= \mathbb{P}_{\boldsymbol{z}}(\boldsymbol{f}^{-1}(N)) \\ &= \mathbb{P}_{\boldsymbol{z}}(\emptyset) = 0\,. \end{aligned}$$

The above contradiction implies that $\boldsymbol{y}^0 \in \overline{\boldsymbol{f}(\text{supp}(\boldsymbol{z}))}$. ■

Proposition 2 (Identifiability up to diffeomorphism) *Let $\boldsymbol{\theta} := (\boldsymbol{f}, p, \boldsymbol{G})$ and $\hat{\boldsymbol{\theta}} := (\hat{\boldsymbol{f}}, \hat{p}, \hat{\boldsymbol{G}})$ be two models satisfying Assumption 1. If $\boldsymbol{\theta} \sim_{\text{obs}} \hat{\boldsymbol{\theta}}$ (Def. 4), then $\boldsymbol{\theta} \sim_{\text{diff}} \hat{\boldsymbol{\theta}}$ (Def. 5).*

Proof

Equality of Denoised Distributions. Given an arbitrary $\boldsymbol{a}^{<T} \in \mathcal{A}^T$ and a parameter $\boldsymbol{\theta} = (\boldsymbol{f}, p, \boldsymbol{G})$, let $\mathbb{P}_{\boldsymbol{x}^{\leq T}|\boldsymbol{a}^{<T};\boldsymbol{\theta}}$ be the conditional probability distribution of $\boldsymbol{x}^{\leq T}$, let $\mathbb{P}_{\boldsymbol{z}^{\leq T}|\boldsymbol{a}^{<T};\boldsymbol{\theta}}$ be the conditional probability distribution of $\boldsymbol{z}^{\leq T}$ and let $\mathbb{P}_{\boldsymbol{n}^{\leq T}}$ be the probability distribution of $\boldsymbol{n}^{\leq T}$ (the Gaussian noises added on $\boldsymbol{f}(\boldsymbol{z}^{\leq T})$, defined in Sec. 2.1). Let $\boldsymbol{y}^t := \boldsymbol{f}(\boldsymbol{z}^t)$ and $\mathbb{P}_{\boldsymbol{y}^{\leq T}|\boldsymbol{a}^{<T};\boldsymbol{\theta}}$ be its conditional probability distribution. First, notice that

$$\mathbb{P}_{\boldsymbol{x}^{\leq T}|\boldsymbol{a}^{<T};\boldsymbol{\theta}} = \mathbb{P}_{\boldsymbol{y}^{\leq T}|\boldsymbol{a}^{<T};\boldsymbol{\theta}} * \mathbb{P}_{\boldsymbol{n}^{\leq T}}\,,$$

where $*$ is the convolution operator between two measures. We now show that if two models agree on the observations, i.e. $\mathbb{P}_{\boldsymbol{x}^{\leq T}|\boldsymbol{a}^{<T};\boldsymbol{\theta}} = \mathbb{P}_{\boldsymbol{x}^{\leq T}|\boldsymbol{a}^{<T};\hat{\boldsymbol{\theta}}}$, then $\mathbb{P}_{\boldsymbol{y}^{\leq T}|\boldsymbol{a}^{<T};\boldsymbol{\theta}} = \mathbb{P}_{\boldsymbol{y}^{\leq T}|\boldsymbol{a}^{<T};\hat{\boldsymbol{\theta}}}$. The following argument makes use of the Fourier transform $\mathcal{F}$ generalized to arbitrary probability measures. This tool is necessary to deal with measures which do not have a density w.r.t either the Lebesgue or the counting measure, as is the case of $\mathbb{P}_{\boldsymbol{y}^{\leq T}|\boldsymbol{a}^{<T};\boldsymbol{\theta}}$ (all its mass is concentrated on the set $\boldsymbol{f}(\mathbb{R}^{d_z})$). See Pollard (2001, Chapter 8) for an introduction and useful properties.

$$\mathbb{P}_{\boldsymbol{x}^{\leq T}|\boldsymbol{a}^{<T};\boldsymbol{\theta}} = \mathbb{P}_{\boldsymbol{x}^{\leq T}|\boldsymbol{a}^{<T};\hat{\boldsymbol{\theta}}} \tag{34}$$

$$\mathbb{P}_{\boldsymbol{y}^{\leq T}|\boldsymbol{a}^{<T};\boldsymbol{\theta}} * \mathbb{P}_{\boldsymbol{n}^{\leq T}} = \mathbb{P}_{\boldsymbol{y}^{\leq T}|\boldsymbol{a}^{<T};\hat{\boldsymbol{\theta}}} * \mathbb{P}_{\boldsymbol{n}^{\leq T}} \tag{35}$$

$$\mathcal{F}(\mathbb{P}_{\boldsymbol{y}^{\leq T}|\boldsymbol{a}^{<T};\boldsymbol{\theta}} * \mathbb{P}_{\boldsymbol{n}^{\leq T}}) = \mathcal{F}(\mathbb{P}_{\boldsymbol{y}^{\leq T}|\boldsymbol{a}^{<T};\hat{\boldsymbol{\theta}}} * \mathbb{P}_{\boldsymbol{n}^{\leq T}}) \tag{36}$$

$$\mathcal{F}(\mathbb{P}_{\boldsymbol{y}^{\leq T}|\boldsymbol{a}^{<T};\boldsymbol{\theta}})\mathcal{F}(\mathbb{P}_{\boldsymbol{n}^{\leq T}}) = \mathcal{F}(\mathbb{P}_{\boldsymbol{y}^{\leq T}|\boldsymbol{a}^{<T};\hat{\boldsymbol{\theta}}})\mathcal{F}(\mathbb{P}_{\boldsymbol{n}^{\leq T}}) \tag{37}$$

$$\mathcal{F}(\mathbb{P}_{\boldsymbol{y}^{\leq T}|\boldsymbol{a}^{<T};\boldsymbol{\theta}}) = \mathcal{F}(\mathbb{P}_{\boldsymbol{y}^{\leq T}|\boldsymbol{a}^{<T};\hat{\boldsymbol{\theta}}}) \tag{38}$$

$$\mathbb{P}_{\boldsymbol{y}^{\leq T}|\boldsymbol{a}^{<T};\boldsymbol{\theta}} = \mathbb{P}_{\boldsymbol{y}^{\leq T}|\boldsymbol{a}^{<T};\hat{\boldsymbol{\theta}}}\,, \tag{39}$$

where (36) & (39) use the fact that the Fourier transform is invertible, (37) is an application of the fact that the Fourier transform of a convolution is the product of their Fourier transforms and (38) holds because the Fourier transform of a Normal distribution is nonzero everywhere. Note that the latter argument holds because we assume σ^2, the variance of the Gaussian noise added to $\boldsymbol{y}^t$, is the same for both models. Notice that, since $\boldsymbol{f}(\mathbb{R}^{d_z})$ and $\hat{\boldsymbol{f}}(\mathbb{R}^{d_z})$ are closed in $\mathbb{R}^{d_x}$ and the support of $\boldsymbol{z}^{\leq T}$ is $\mathbb{R}^{d_z \times T}$, Lemma 6 implies that

$$\boldsymbol{f}(\mathbb{R}^{d_z \times T}) = \text{supp}(\mathbb{P}_{\boldsymbol{y}^{\leq T}|\boldsymbol{a}^{<T};\boldsymbol{\theta}}) \qquad \& \qquad \text{supp}(\mathbb{P}_{\boldsymbol{y}^{\leq T}|\boldsymbol{a}^{<T};\hat{\boldsymbol{\theta}}}) = \hat{\boldsymbol{f}}(\mathbb{R}^{d_z \times T})$$

where we overloaded the notation by defining $\boldsymbol{f}(\boldsymbol{z}^{\leq T}) := (\boldsymbol{f}(\boldsymbol{z}^1), ..., \boldsymbol{f}(\boldsymbol{z}^T))$ and analogously for $\hat{\boldsymbol{f}}(\boldsymbol{z}^{\leq T})$. Since both measure in (39) are equal, their supports must also be. This implies that $\boldsymbol{f}(\mathbb{R}^{d_z}) = \hat{\boldsymbol{f}}(\mathbb{R}^{d_z})$, which is part of the definition of equivalence up to diffeomorphism (Definition 5).

Equality of densities. Continuing with (39),

$$\begin{aligned}
\mathbb{P}_{\boldsymbol{y}^{\leq T}|\boldsymbol{a}^{<T};\boldsymbol{\theta}} &= \mathbb{P}_{\boldsymbol{y}^{\leq T}|\boldsymbol{a}^{<T};\hat{\boldsymbol{\theta}}} \\
\mathbb{P}_{\boldsymbol{z}^{\leq T}|\boldsymbol{a}^{<T};\boldsymbol{\theta}} \circ \boldsymbol{f}^{-1} &= \mathbb{P}_{\boldsymbol{z}^{\leq T}|\boldsymbol{a}^{<T};\hat{\boldsymbol{\theta}}} \circ \hat{\boldsymbol{f}}^{-1} \\
\mathbb{P}_{\boldsymbol{z}^{\leq T}|\boldsymbol{a}^{<T};\boldsymbol{\theta}} \circ \boldsymbol{f}^{-1} \circ \hat{\boldsymbol{f}} &= \mathbb{P}_{\boldsymbol{z}^{\leq T}|\boldsymbol{a}^{<T};\hat{\boldsymbol{\theta}}} \\
\mathbb{P}_{\boldsymbol{z}^{\leq T}|\boldsymbol{a}^{<T};\boldsymbol{\theta}} \circ \boldsymbol{v} &= \mathbb{P}_{\boldsymbol{z}^{\leq T}|\boldsymbol{a}^{<T};\hat{\boldsymbol{\theta}}}\,,
\end{aligned} \tag{40}$$

where $\boldsymbol{v} := \boldsymbol{f}^{-1} \circ \hat{\boldsymbol{f}}$ is a composition of diffeomorphisms and thus a diffeomorphism from $\mathbb{R}^{d_z}$ to itself. Note that this composition is well defined because $\boldsymbol{f}(\mathbb{R}^{d_z}) = \hat{\boldsymbol{f}}(\mathbb{R}^{d_z})$. We chose to work directly with measures (functions on sets), as opposed to manifold integrals in Khemakhem et al. (2020a), because it simplifies the derivation of (40) and avoids having to define densities w.r.t. measures concentrated on a manifold.

The density of $\mathbb{P}_{\boldsymbol{z}^{\leq T}|\boldsymbol{a}^{<T};\boldsymbol{\theta}} \circ \boldsymbol{v}$ w.r.t. to the Lebesgue measure is given by the change-of-variable rule for random vectors (which can be applied because $\boldsymbol{v}$ is a diffeomorphism) and is given by $\prod_{t=1}^{T} p(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t})|\det D\boldsymbol{v}(\boldsymbol{z}^t)|$, where p refers to the density model with parameter $\boldsymbol{\theta}$ and $D\boldsymbol{v}(\boldsymbol{z}^t)$ is the Jacobian matrix of $\boldsymbol{v}$. Since $\mathbb{P}_{\boldsymbol{z}^{\leq T}|\boldsymbol{a}^{<T};\boldsymbol{\theta}} \circ \boldsymbol{v} = \mathbb{P}_{\boldsymbol{z}^{\leq T}|\boldsymbol{a}^{<T};\hat{\boldsymbol{\theta}}}$, their respective densities w.r.t. Lebesgue must also agree:

$$\prod_{t=1}^{T} \hat{p}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = \prod_{t=1}^{T} p(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t})|\det D\boldsymbol{v}(\boldsymbol{z}^t)|\,,$$

where $\hat{p}$ refers to the conditional density of the model with parameter $\hat{\boldsymbol{\theta}}$.

For a given t_0, we have

$$\prod_{t=1}^{t_0} \hat{p}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = \prod_{t=1}^{t_0} p(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t})|\det D\boldsymbol{v}(\boldsymbol{z}^t)|\,, \tag{41}$$

by integrating first $\boldsymbol{z}^T$, then $\boldsymbol{z}^{t-1}$, then ..., up to $\boldsymbol{z}^{t_0+1}$. Note that we can integrate $\boldsymbol{z}^{t_0}$ and get

$$\prod_{t=1}^{t_0-1} \hat{p}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = \prod_{t=1}^{t_0-1} p(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t})|\det D\boldsymbol{v}(\boldsymbol{z}^t)|\,. \tag{42}$$

By dividing (41) by (42), we get

$$\hat{p}(\boldsymbol{z}^{t_0} | \boldsymbol{z}^{<t_0}, \boldsymbol{a}^{<t_0}) = p(\boldsymbol{v}(\boldsymbol{z}^{t_0}) \mid \boldsymbol{v}(\boldsymbol{z}^{<t_0}), \boldsymbol{a}^{<t_0}) |\det D\boldsymbol{v}(\boldsymbol{z}^{t_0})|\,, \tag{43}$$

which completes the proof. ■

A.4 The Consistency Relations (Definitions 13 & 14) are Equivalence Relations

In this section, we demonstrate that the relations $\sim^{\boldsymbol{a}}_{\text{con}}$ and $\sim^{\boldsymbol{z}}_{\text{con}}$ are equivalence relations by leveraging the fact that the set of $\boldsymbol{G}$-preserving diffeomorphisms form a group under composition (Proposition 5). We start by showing a fact that will be useful in the following.

Lemma 7 *Let $\boldsymbol{G} \in \{0,1\}^{m\times n}$.*

1. *A map $\boldsymbol{c} : \mathbb{R}^m \to \mathbb{R}^m$ is $\boldsymbol{G}$-preserving if and only if $\boldsymbol{c}$ is $\boldsymbol{GP}$-preserving, where $\boldsymbol{P}$ is an $n \times n$ permutation matrix.*
2. *A map $\boldsymbol{c} : \mathbb{R}^m \to \mathbb{R}^m$ is $\boldsymbol{G}$-preserving if and only if $\boldsymbol{P} \circ \boldsymbol{c} \circ \boldsymbol{P}^\top$ is $\boldsymbol{PG}$-preserving, where $\boldsymbol{P}$ is a $m \times m$ permutation matrix.*
3. *When m = n, a map $\boldsymbol{c} : \mathbb{R}^m \to \mathbb{R}^m$ is $\boldsymbol{G}$-preserving if and only if $\boldsymbol{P} \circ \boldsymbol{c} \circ \boldsymbol{P}^\top$ is $\boldsymbol{PGP}^\top$-preserving, where $\boldsymbol{P}$ is a $m \times m$ permutation matrix.*

Proof Let $\boldsymbol{C}$ be the dependency graph of $\boldsymbol{c}$.

1. $\boldsymbol{C}^\top \mathbb{R}^{m\times n}_{\boldsymbol{G}} \subseteq \mathbb{R}^{m\times n}_{\boldsymbol{G}} \iff \boldsymbol{C}^\top \mathbb{R}^{m\times n}_{\boldsymbol{G}} \boldsymbol{P} \subseteq \mathbb{R}^{m\times n}_{\boldsymbol{G}} \boldsymbol{P} \iff \boldsymbol{C}^\top \mathbb{R}^{m\times n}_{\boldsymbol{GP}} \subseteq \mathbb{R}^{m\times n}_{\boldsymbol{GP}}$
2. First, notice that the dependency graph of $\boldsymbol{P} \circ \boldsymbol{c} \circ \boldsymbol{P}^\top$ is $\boldsymbol{PCP}^\top$.

   $(\boldsymbol{PCP}^\top)^\top \mathbb{R}^{m\times n}_{\boldsymbol{PG}} \subseteq \mathbb{R}^{m\times n}_{\boldsymbol{PG}} \iff \boldsymbol{PC}^\top \boldsymbol{P}^\top \boldsymbol{P} \mathbb{R}^{m\times n}_{\boldsymbol{G}} \subseteq \boldsymbol{P} \mathbb{R}^{m\times n}_{\boldsymbol{G}} \iff \boldsymbol{C}^\top \mathbb{R}^{m\times n}_{\boldsymbol{G}} \subseteq \mathbb{R}^{m\times n}_{\boldsymbol{G}}$
3. This is a consequence of the first two statements.

■

We are now ready to show that the relation $\sim^{\boldsymbol{a}}_{\text{con}}$ (Definition 13) is an equivalence relation.

Proposition 9 *The consistency relation, $\sim^{\boldsymbol{a}}_{\text{con}}$ (Def. 13), is an equivalence relation.*

Proof

Reflexivity. It is easy to see that $\boldsymbol{\theta} \sim^{\boldsymbol{a}}_{\text{con}} \boldsymbol{\theta}$, simply by setting $\boldsymbol{v}(\boldsymbol{z}) := \boldsymbol{z}$ with $\boldsymbol{P} := \boldsymbol{I}$.

Symmetry. Assume $\boldsymbol{\theta} \sim^{\boldsymbol{a}}_{\text{con}} \tilde{\boldsymbol{\theta}}$. Hence, we have $\boldsymbol{PG}^a = \tilde{\boldsymbol{G}}^a$ as well as

$$\boldsymbol{f}(\mathbb{R}^{d_z}) = \tilde{\boldsymbol{f}}(\mathbb{R}^{d_z}), \text{and} \tag{44}$$

$$\tilde{p}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = p(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t}) |\det D\boldsymbol{v}(\boldsymbol{z}^t)|\,, \tag{45}$$

where $\boldsymbol{v} := \boldsymbol{f}^{-1} \circ \tilde{\boldsymbol{f}}$ can be written as $\boldsymbol{v} := \boldsymbol{c} \circ \boldsymbol{P}^\top$, where $\boldsymbol{c}$ is a $\boldsymbol{G}^a$-preserving diffeomorphism and $\boldsymbol{P}$ is a permutation. We can massage (45) to get

$$p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = \tilde{p}(\boldsymbol{v}^{-1}(\boldsymbol{z}^t) \mid \boldsymbol{v}^{-1}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t}) |\det D\boldsymbol{v}^{-1}(\boldsymbol{z}^t)|\,.$$

Of course, we also have that $\tilde{\boldsymbol{P}}\tilde{\boldsymbol{G}}^a = \boldsymbol{G}^a$, where $\tilde{\boldsymbol{P}} := \boldsymbol{P}^\top$. Now the only thing left to prove is that $\boldsymbol{v}^{-1}$ can be written as $\tilde{\boldsymbol{c}} \circ \tilde{\boldsymbol{P}}^\top$ where $\tilde{\boldsymbol{c}}$ is $\tilde{\boldsymbol{G}}^a$-preserving. We know that

$$\boldsymbol{v}^{-1} = \boldsymbol{P} \circ \boldsymbol{c}^{-1} = \underbrace{\boldsymbol{P} \circ \boldsymbol{c}^{-1} \circ \boldsymbol{P}^\top}_{\tilde{\boldsymbol{c}} :=} \circ \boldsymbol{P} = \tilde{\boldsymbol{c}} \circ \tilde{\boldsymbol{P}}^\top .$$

Note that $\boldsymbol{c}^{-1}$ is $\boldsymbol{G}^a$-preserving and thus, by Lemma 7, $\tilde{\boldsymbol{c}}$ is $\boldsymbol{P}\boldsymbol{G}^a$-preserving, i.e. $\tilde{\boldsymbol{G}}^a$-preserving. Hence, $\sim^{\boldsymbol{a}}_{\text{con}}$ is symmetric.

Transitivity. Suppose $\boldsymbol{\theta} \sim^{\boldsymbol{a}}_{\text{con}} \tilde{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}} \sim^{\boldsymbol{a}}_{\text{con}} \hat{\boldsymbol{\theta}}$. This means

$$\boldsymbol{P}_1 \boldsymbol{G}^a = \tilde{\boldsymbol{G}}^a , \tag{46}$$

$$\boldsymbol{f}(\mathbb{R}^{d_z}) = \tilde{\boldsymbol{f}}(\mathbb{R}^{d_z}), \text{and} \tag{47}$$

$$\tilde{p}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = p(\boldsymbol{v}_1(\boldsymbol{z}^t) \mid \boldsymbol{v}_1(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t}) |\det D\boldsymbol{v}_1(\boldsymbol{z}^t)| , \tag{48}$$

where $\boldsymbol{v}_1 := \boldsymbol{c}_1 \circ \boldsymbol{P}_1^\top$ with $\boldsymbol{c}_1$ being $\boldsymbol{G}^a$-preserving; and

$$\boldsymbol{P}_2 \tilde{\boldsymbol{G}}^a = \hat{\boldsymbol{G}}^a , \tag{49}$$

$$\tilde{\boldsymbol{f}}(\mathbb{R}^{d_z}) = \hat{\boldsymbol{f}}(\mathbb{R}^{d_z}), \text{and} \tag{50}$$

$$\hat{p}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = \tilde{p}(\boldsymbol{v}_2(\boldsymbol{z}^t) \mid \boldsymbol{v}_2(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t}) |\det D\boldsymbol{v}_2(\boldsymbol{z}^t)| , \tag{51}$$

where $\boldsymbol{v}_2 := \boldsymbol{c}_2 \circ \boldsymbol{P}_2^\top$ with $\boldsymbol{c}_2$ being $\tilde{\boldsymbol{G}}^a$-preserving.

To show that $\boldsymbol{\theta} \sim^{\boldsymbol{a}}_{\text{con}} \hat{\boldsymbol{\theta}}$, we first combine (46) with (49) to get

$$\underbrace{\boldsymbol{P}_2 \boldsymbol{P}_1}_{\boldsymbol{P} :=} \boldsymbol{G}^a = \hat{\boldsymbol{G}}^a . \tag{52}$$

Of course we also have that $\boldsymbol{f}(\mathbb{R}^{d_z}) = \tilde{\boldsymbol{f}}(\mathbb{R}^{d_z}) = \hat{\boldsymbol{f}}(\mathbb{R}^{d_z})$. By massaging both (48) and (51), we get:

$$\hat{p}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = p(\boldsymbol{v}_1 \circ \boldsymbol{v}_2(\boldsymbol{z}^t) \mid \boldsymbol{v}_1 \circ \boldsymbol{v}_2(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t}) |\det D(\boldsymbol{v}_1 \circ \boldsymbol{v}_2)(\boldsymbol{z}^t)| .$$

Define $\boldsymbol{v} := \boldsymbol{v}_1 \circ \boldsymbol{v}_2$. We now want to show that $\boldsymbol{v}$ can be written as $\boldsymbol{v} = \boldsymbol{c} \circ \boldsymbol{P}^\top$ where $\boldsymbol{c}$ is $\boldsymbol{G}^a$-preserving. We have that

$$\begin{aligned}
\boldsymbol{v}_1 \circ \boldsymbol{v}_2 &= \boldsymbol{c}_1 \circ \boldsymbol{P}_1^\top \circ \boldsymbol{c}_2 \circ \boldsymbol{P}_2^\top \\
&= \boldsymbol{c}_1 \circ \underbrace{\boldsymbol{P}_1^\top \circ \boldsymbol{P}_2^\top}_{\boldsymbol{P}^\top =} \circ \underbrace{\boldsymbol{P}_2 \circ \boldsymbol{c}_2 \circ \boldsymbol{P}_2^\top}_{\hat{\boldsymbol{c}} :=} \\
&= \boldsymbol{c}_1 \circ \boldsymbol{P}^\top \circ \hat{\boldsymbol{c}}
\end{aligned}$$

where, by Lemma 7, $\hat{\boldsymbol{c}}$ is $\boldsymbol{P}_2\tilde{\boldsymbol{G}}^a$-preserving, i.e. $\hat{\boldsymbol{G}}^a$-preserving. We continue and get

$$\begin{aligned}
\boldsymbol{v}_1 \circ \boldsymbol{v}_2 &= \boldsymbol{c}_1 \circ \boldsymbol{P}^\top \circ \hat{\boldsymbol{c}} \\
&= \boldsymbol{c}_1 \circ \underbrace{\boldsymbol{P}^\top \circ \hat{\boldsymbol{c}} \circ \boldsymbol{P}}_{\boldsymbol{c}' :=} \circ \boldsymbol{P}^\top \\
&= \boldsymbol{c}_1 \circ \boldsymbol{c}' \circ \boldsymbol{P}^\top ,
\end{aligned}$$

where, by Lemma 7, $\boldsymbol{c}'$ is $\boldsymbol{P}^\top \hat{\boldsymbol{G}}^a$-preserving, i.e. $\boldsymbol{G}^a$-preserving (by (52)). Since both $\boldsymbol{c}_1$ and $\boldsymbol{c}'$ are $\boldsymbol{G}^a$-preserving, $\boldsymbol{c} := \boldsymbol{c}_1 \circ \boldsymbol{c}'$ is $\boldsymbol{G}^a$-preserving, which concludes the proof. ■

The same can be shown for $\sim^{\boldsymbol{z}}_{\text{con}}$ (Definition 14).

Proposition 10 *The consistency relation, $\sim^{\boldsymbol{z}}_{\text{con}}$ (Def. 14), is an equivalence relation.*

Proof The proof is exactly analogous to the proof that $\sim^{\boldsymbol{a}}_{\text{con}}$ is an equivalence relation. Essentially, every statement of the form "$\boldsymbol{P}\boldsymbol{G}^a = \tilde{\boldsymbol{G}}^a$" becomes "$\boldsymbol{P}\boldsymbol{G}^z\boldsymbol{P}^\top = \tilde{\boldsymbol{G}}^z$" and statements of the form "$\boldsymbol{c}$ is $\boldsymbol{G}^a$-preserving" becomes "$\boldsymbol{c}$ is $\boldsymbol{G}^z$-preserving and $(\boldsymbol{G}^z)^\top$-preserving". The full proof is left as an exercise to the reader. ■

A.4.1 COMBINING EQUIVALENCE RELATIONS

Proposition 6 *Let $\boldsymbol{\theta} := (\boldsymbol{f}, p, \boldsymbol{G})$ and $\tilde{\boldsymbol{\theta}} := (\tilde{\boldsymbol{f}}, \tilde{p}, \tilde{\boldsymbol{G}})$ be two models satisfying Assumptions 1, 2 & 3. We have $\boldsymbol{\theta} \sim^{\boldsymbol{z},\boldsymbol{a}}_{\text{con}} \tilde{\boldsymbol{\theta}}$ if and only if $\boldsymbol{\theta} \sim^{\boldsymbol{a}}_{\text{con}} \tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta} \sim^{\boldsymbol{z}}_{\text{con}} \tilde{\boldsymbol{\theta}}$.*

Proof The "only if" part of the statement is trivial. We now show the "if" part.

Let $\boldsymbol{v} := \boldsymbol{f}^{-1} \circ \tilde{\boldsymbol{f}}$. Since $\boldsymbol{\theta} \sim^{\boldsymbol{a}}_{\text{con}} \tilde{\boldsymbol{\theta}}$, we have that $\tilde{\boldsymbol{G}}^a = \boldsymbol{P}\boldsymbol{G}^a$ and $\boldsymbol{v} = \boldsymbol{c} \circ \boldsymbol{P}^\top$ where $\boldsymbol{P}$ a permutation matrix and $\boldsymbol{c}$ is $\boldsymbol{G}^a$-preserving. Since $\boldsymbol{\theta} \sim^{\boldsymbol{z}}_{\text{con}} \tilde{\boldsymbol{\theta}}$, we have that $\tilde{\boldsymbol{G}}^z = \bar{\boldsymbol{P}}\boldsymbol{G}^z\bar{\boldsymbol{P}}^\top$ and $\boldsymbol{v} = \bar{\boldsymbol{c}} \circ \bar{\boldsymbol{P}}^\top$ where $\bar{\boldsymbol{P}}$ is a permutation matrix and $\bar{\boldsymbol{c}}$ is $\boldsymbol{G}^z$-preserving and $(\boldsymbol{G}^z)^\top$-preserving. Let $\boldsymbol{C}$ and $\bar{\boldsymbol{C}}$ be the dependency graphs $\boldsymbol{c}$ and $\bar{\boldsymbol{c}}$, respectively.

Choose an arbitrary $\boldsymbol{z}$. Since $D\boldsymbol{c}(\boldsymbol{z})$ is invertible, Lemma 2 implies that there exists a permutation $\boldsymbol{P}_0$ such that $\boldsymbol{P}_0^\top \subseteq D\boldsymbol{v}(\boldsymbol{z})$, which in turn implies that $\boldsymbol{P}_0^\top \subseteq \bar{\boldsymbol{C}}$. Because $\boldsymbol{P}_0^\top \subseteq \bar{\boldsymbol{C}}$, we have that $\boldsymbol{P}_0^\top$ is $\boldsymbol{G}^z$-preserving and $(\boldsymbol{G}^z)^\top$-preserving (Proposition 3). By closure under composition and inversion, $\bar{\boldsymbol{c}} \circ \boldsymbol{P}_0$ is $\boldsymbol{G}^z$- and $(\boldsymbol{G}^z)^\top$-preserving.

Note that

$$\begin{aligned} \boldsymbol{c} \circ \boldsymbol{P}^\top &= \bar{\boldsymbol{c}} \circ \bar{\boldsymbol{P}}^\top \\ \implies \boldsymbol{C}\boldsymbol{P}^\top &= \bar{\boldsymbol{C}}\bar{\boldsymbol{P}}^\top \\ \boldsymbol{C}\boldsymbol{P}^\top\bar{\boldsymbol{P}} &= \bar{\boldsymbol{C}} \supseteq \boldsymbol{P}_0^\top \\ \implies \boldsymbol{C} &\supseteq \boldsymbol{P}_0^\top\bar{\boldsymbol{P}}^\top\boldsymbol{P}\,. \end{aligned}$$

This means the permutation $\boldsymbol{P}_0^\top\bar{\boldsymbol{P}}^\top\boldsymbol{P}$ must be $\boldsymbol{G}^a$-preserving since $\boldsymbol{C}$ is (Proposition 3).

This further implies that $\boldsymbol{C}\boldsymbol{P}^\top\bar{\boldsymbol{P}}\boldsymbol{P}_0 = \bar{\boldsymbol{C}}\boldsymbol{P}_0$ is $\boldsymbol{G}^a$-preserving by closure under multiplication (recall $\boldsymbol{C}$ is $\boldsymbol{G}^a$-preserving too). Hence $\bar{\boldsymbol{c}} \circ \boldsymbol{P}_0$ is $\boldsymbol{G}^a$-preserving.

We thus have that $\boldsymbol{v} = (\bar{\boldsymbol{c}} \circ \boldsymbol{P}_0)(\bar{\boldsymbol{P}}\boldsymbol{P}_0)^\top$ where $(\bar{\boldsymbol{c}} \circ \boldsymbol{P}_0)$ is $\boldsymbol{G}^a$-, $\boldsymbol{G}^z$- and $(\boldsymbol{G}^z)^\top$-preserving. The only thing left to show is that $(\bar{\boldsymbol{P}}\boldsymbol{P}_0)\boldsymbol{G}^a = \tilde{\boldsymbol{G}}^a$ and that $(\bar{\boldsymbol{P}}\boldsymbol{P}_0)\boldsymbol{G}^z(\bar{\boldsymbol{P}}\boldsymbol{P}_0)^\top = \tilde{\boldsymbol{G}}^z$. The former holds since

$$(\bar{\boldsymbol{P}}\boldsymbol{P}_0)\boldsymbol{G}^a = \boldsymbol{P}(\boldsymbol{P}^\top\bar{\boldsymbol{P}}\boldsymbol{P}_0)\boldsymbol{G}^a = \boldsymbol{P}\boldsymbol{G}^a = \tilde{\boldsymbol{G}}^a\,,$$

where the second equality leverages the fact that $\boldsymbol{P}^\top\bar{\boldsymbol{P}}\boldsymbol{P}_0$ is $\boldsymbol{G}^a$-preserving. Furthermore,

$$(\bar{\boldsymbol{P}}\boldsymbol{P}_0)\boldsymbol{G}^z(\bar{\boldsymbol{P}}\boldsymbol{P}_0)^\top = \bar{\boldsymbol{P}}\boldsymbol{G}^z(\bar{\boldsymbol{P}}\boldsymbol{P}_0)^\top = \bar{\boldsymbol{P}}\boldsymbol{G}^z\boldsymbol{P}_0^\top\bar{\boldsymbol{P}}^\top = \bar{\boldsymbol{P}}(\boldsymbol{P}_0(\boldsymbol{G}^z)^\top)^\top\bar{\boldsymbol{P}}^\top = \bar{\boldsymbol{P}}\boldsymbol{G}^z\bar{\boldsymbol{P}}^\top = \tilde{\boldsymbol{G}}^z\,,$$

where the first and fourth equalities leveraged the fact that $\boldsymbol{P}_0$ is $\boldsymbol{G}^z$- and $(\boldsymbol{G}^z)^\top$-preserving. ■

A.5 Technical Lemmas Leading to Theorems 1, 2 & 3

The goal of this section is to introduce and prove Lemma 12 which was crucial in proofs of Theorems 1, 2 & 3. To prove it, we need a few more results, which we present next.

The following two lemmas are standard, but we provide them with proofs for completeness.

Lemma 8 *Let $f : \mathbb{R}^n \to \mathbb{R}$ be continuous and $A \subseteq \mathbb{R}^n$. If, for all $\boldsymbol{x} \in A$, $\boldsymbol{f}(\boldsymbol{x}) = 0$, then the equality holds on $\overline{A}$.*

Proof We have that $A \subseteq \boldsymbol{f}^{-1}(\{0\})$. Since $\{0\}$ is closed, $\boldsymbol{f}^{-1}(\{0\})$ is also closed by continuity of f. This means $\overline{A} \subseteq \boldsymbol{f}^{-1}(\{0\})$ (since the closure of $\boldsymbol{A}$ is the smallest closed set containing $\boldsymbol{A}$). ■

Lemma 9 *Let μ be the Lebesgue measure on $\mathbb{R}^{d_z}$ and let $E_0 \subseteq \mathbb{R}^{d_z}$ be a zero measure set, i.e. $\mu(E_0) = 0$. Then, $\overline{\mathbb{R}^{d_z} \setminus E_0} = \mathbb{R}^{d_z}$.*

Proof Clearly, $\overline{\mathbb{R}^{d_z} \setminus E_0} \subseteq \mathbb{R}^{d_z}$.

We now show that $\mathbb{R}^{d_z} \subseteq \overline{\mathbb{R}^{d_z} \setminus E_0}$. Take $\boldsymbol{z}_0 \in \mathbb{R}^{d_z}$ and let U be an open set of $\mathbb{R}^{d_z}$ containing $\boldsymbol{z}_0$. Every nonempty open sets have nonzero Lebesgue measure, so

$$0 \neq \mu(U) = \mu(U \cap \mathbb{R}^{d_z}) = \mu(U \cap (\mathbb{R}^{d_z} \setminus E_0)) \implies U \cap (\mathbb{R}^{d_z} \setminus E_0) \neq \emptyset \,.$$

Since U was arbitrary, this means $\boldsymbol{z}_0 \in \overline{\mathbb{R}^{d_z} \setminus E_0}$. ■

This simple lemma will come in handy when proving Lemma 11.

Lemma 10 *If a permutation $\boldsymbol{P}$ is not $\boldsymbol{G}$-preserving and $\boldsymbol{C}$ is a $\boldsymbol{G}$-preserving matrix, we have that $\boldsymbol{C}\boldsymbol{P}^\top$ and $\boldsymbol{P}^\top\boldsymbol{C}$ have a zero on their diagonal.*

Proof Assume $\boldsymbol{P}$ is not $\boldsymbol{G}$-preserving, hence there exists i, j such that $\boldsymbol{G}_{i,\cdot} \not\subseteq \boldsymbol{G}_{j,\cdot}$, but $\boldsymbol{P}_{i,j} = 1$. Now note that

$$(\boldsymbol{C}\boldsymbol{P}^\top)_{i,i} = \boldsymbol{C}_{i,\cdot}(\boldsymbol{P}_{i\cdot})^\top = \boldsymbol{C}_{i,\cdot}\boldsymbol{e}_j = \boldsymbol{C}_{i,j} \,,$$

which is equal to zero because $\boldsymbol{C}$ is $\boldsymbol{G}$-preserving and $\boldsymbol{G}_{i,\cdot} \not\subseteq \boldsymbol{G}_{j,\cdot}$. Similarly,

$$(\boldsymbol{P}^\top\boldsymbol{C})_{j,j} = (\boldsymbol{P}_{\cdot,j})^\top\boldsymbol{C}_{\cdot,j} = \boldsymbol{e}_i^\top\boldsymbol{C}_{\cdot,j} = \boldsymbol{C}_{i,j} = 0 \,.$$

which concludes the proof. ■

The following lemma is the same as Lemma 12 which is used to prove Theorems 1, 2 & 3, except it does not take into account the "almost everywhere" subtlety. Lemma 12 will extend it to deal with this difficulty.

Lemma 11 *Let $\boldsymbol{G} \in \{0,1\}^{m\times n}$, let $\mathcal{Z}$ be a connected subset of some topological space and let $\boldsymbol{L} : \mathcal{Z} \to \mathbb{R}^{m\times m}$ be a continuous function such that $\boldsymbol{L}(\boldsymbol{z})$ is invertible for all $\boldsymbol{z} \in \mathcal{Z}$. Suppose that, for all $\boldsymbol{z} \in \mathcal{Z}$, there exists a permutation matrix $\boldsymbol{P}(\boldsymbol{z})$ such that $\boldsymbol{L}(\boldsymbol{z})\boldsymbol{P}(\boldsymbol{z})$ is $\boldsymbol{G}$-preserving. Then, there exists a permutation matrix $\boldsymbol{P}$ such that, for all $\boldsymbol{z} \in \mathcal{Z}$, $\boldsymbol{L}(\boldsymbol{z})\boldsymbol{P}$ is $\boldsymbol{G}$-preserving.*

Proof The goal of this lemma is to show that in the statement above, one can change the order of the "for all $\boldsymbol{z} \in \mathcal{Z}$" and "there exists a permutation". To do that, we show that if $\mathcal{Z}$ is connected and the map $\boldsymbol{L}(\cdot)$ is continuous, then one can find a single permutation that works for all $\boldsymbol{z} \in \mathcal{Z}$.

Let $\mathcal{G}$ be the set of $\boldsymbol{G}^a$-preserving matrices. Recall that, by Proposition 3, $\mathcal{G}$ corresponds to all matrices that have some set of entries equal to zero.

First, since $\mathcal{Z}$ is connected and $\boldsymbol{L}$ is continuous, its image, $\boldsymbol{L}(\mathcal{Z})$, must be connected (by (Munkres, 2000, Theorem 23.5)).

Second, from the hypothesis of the lemma, we know that

$$\boldsymbol{L}(\mathcal{Z}) \subseteq \mathcal{L} := \left(\bigcup_{\pi \in \mathfrak{S}_m} \mathcal{G}\boldsymbol{P}_\pi \right) \setminus \{\text{singular matrices}\}\,,$$

where $\mathfrak{S}_m$ is the set of permutations and $\mathcal{G}\boldsymbol{P}_\pi = \{\boldsymbol{L}\boldsymbol{P}_\pi \mid \boldsymbol{L} \in \mathcal{G}\}$. We can rewrite the set $\mathcal{L}$ above as

$$\mathcal{L} = \left(\bigcup_{\pi \in \mathfrak{S}_m} \mathcal{G}\boldsymbol{P}_\pi \setminus \{\text{singular matrices}\} \right).$$

We now define an equivalence relation $\sim$ over permutations: $\pi \sim \pi'$ iff $\boldsymbol{P}_\pi \boldsymbol{P}_{\pi'}^\top$ is $\boldsymbol{G}$-preserving. One can verify that the relation $\sim$ is indeed an equivalence relation by using the fact that invertible $\boldsymbol{G}$-preserving matrices form a group (Proposition 4). We notice that

$$\pi \sim \pi' \implies \mathcal{G} = \mathcal{G}\boldsymbol{P}_\pi \boldsymbol{P}_{\pi'}^\top \implies \mathcal{G}\boldsymbol{P}_{\pi'} = \mathcal{G}\boldsymbol{P}_\pi\,, \tag{53}$$

where the first implication holds because $\boldsymbol{G}$-preserving matrices are closed under matrix multiplication (Proposition 4). Let $\mathfrak{S}_m/\sim$ be the set of equivalence classes induce by $\sim$ and let Π stand for one such equivalence class. Thanks to (53), we can define, for all $\Pi \in \mathfrak{S}_m/\sim$, the following set:

$$V_\Pi := \mathcal{G}\boldsymbol{P}_\pi \setminus \{\text{singular matrices}\},\ \text{for some } \pi \in \Pi\,,$$

where the specific choice of $\pi \in \Pi$ is arbitrary (any $\pi' \in \Pi$ would yield the same definition, by (53)). This construction allows us to write

$$\mathcal{L} = \bigcup_{\Pi \in \mathfrak{S}_m/\sim} V_\Pi\,.$$

We now show that $\{V_\Pi\}_{\Pi \in \mathfrak{S}_m/\sim}$ forms a partition of $\mathcal{L}$. Choose two distinct equivalence classes of permutations Π and Π' and let $\pi \in \Pi$ and $\pi' \in \Pi'$ be representatives. We will now prove that

$$\mathcal{G}\boldsymbol{P}_\pi \cap \mathcal{G}\boldsymbol{P}_{\pi'} \subseteq \{\text{singular matrices}\}\,. \tag{54}$$

To achieve this, we proceed by contradiction: Suppose there exists an invertible matrix $\boldsymbol{A} \in \mathcal{G}\boldsymbol{P}_\pi \cap \mathcal{G}\boldsymbol{P}_{\pi'}$. By Lemma 2, there exists a permutation $\boldsymbol{P}$ s.t. $\boldsymbol{A}\boldsymbol{P}^\top$ has no zero on its diagonal. Of course, the permutation $\boldsymbol{P}$ belongs to only one equivalence class and thus either $\boldsymbol{P} \notin \Pi$ or $\boldsymbol{P} \notin \Pi'$. Without loss of generality, assume the former. We thus have that $\boldsymbol{P} \not\sim \pi$ and thus $\boldsymbol{P}_\pi \boldsymbol{P}^\top$ is not $\boldsymbol{G}$-preserving. We can thus write

$$\boldsymbol{A} \in \mathcal{G}\boldsymbol{P}_\pi \cap \mathcal{G}\boldsymbol{P}_{\pi'} \implies \boldsymbol{A}\boldsymbol{P}^\top \in \mathcal{G}\boldsymbol{P}_\pi \boldsymbol{P}^\top \cap \mathcal{G}\boldsymbol{P}_{\pi'}\boldsymbol{P}^\top\,.$$

By Lemma 10, all matrices in $\mathcal{G}\boldsymbol{P}_\pi\boldsymbol{P}^\top$ have a zero on their diagonal. This is a contradiction with $\boldsymbol{A}\boldsymbol{P}^\top \in \mathcal{G}\boldsymbol{P}_\pi\boldsymbol{P}^\top$, since, as we said, $\boldsymbol{A}\boldsymbol{P}^\top$ has no zero on its diagonal. We thus conclude that no invertible matrix is in the intersection $\mathcal{G}\boldsymbol{P}_\pi \cap \mathcal{G}\boldsymbol{P}_{\pi'}$ and thus (54) holds.

We thus have that

$$V_\Pi \cap V_{\Pi'} = \emptyset\,,$$

which shows that $\{V_\Pi\}_{\Pi\in\mathfrak{S}_m/\sim}$ is indeed a partition of $\mathcal{L}$.

Each V_Π is closed in $\mathcal{L}$ (w.r.t. the subset topology inherited from $\mathbb{R}^{m\times m}$) since

$$V_\Pi = \mathcal{G}\boldsymbol{P}_\pi \setminus \{\text{singular matrices}\} = \mathcal{L} \cap \underbrace{\mathcal{G}\boldsymbol{P}_\pi}_{\text{closed in } \mathbb{R}^{m\times m}}\,.$$

Moreover, V_Π is open in $\mathcal{L}$, since

$$V_\Pi = \mathcal{L} \setminus \underbrace{\bigcup_{\Pi'\neq\Pi} V_{\Pi'}}_{\text{closed in } \mathcal{L}}\,.$$

Thus, for any $\Pi \in \mathfrak{S}(\mathcal{B})/\sim$, the sets V_Π and $\bigcup_{\Pi'\neq\Pi} V_{\Pi'}$ forms a *separation* (see (Munkres, 2000, Section 23)). Since $\boldsymbol{L}(\mathcal{Z})$ is a connected subset of $\mathcal{L}$, it must lie completely in V_Π or $\bigcup_{\Pi'\neq\Pi} V_{\Pi'}$, by (Munkres, 2000, Lemma 23.2). Since this is true for all Π, it must follow that there exists a Π^* such that $\boldsymbol{L}(\mathcal{Z}) \subseteq V_{\Pi^*}$. Choose any representative $\boldsymbol{P}_* \in \Pi^*$. We thus have that, for all $\boldsymbol{z} \in \mathcal{Z}$, $\boldsymbol{L}(\boldsymbol{z}) = \boldsymbol{C}(\boldsymbol{z})\boldsymbol{P}_*^\top$, where $\boldsymbol{C}(\boldsymbol{z})$ is $\boldsymbol{G}$-preserving, which completes the proof. ■

The goal of the next result is to relax the conditions of Lemma 11 so that $\boldsymbol{L}(\boldsymbol{z})\boldsymbol{P}(\boldsymbol{z})$ is $\boldsymbol{G}$-preserving for *almost all* $\boldsymbol{z} \in \mathbb{R}^{d_z}$.

Lemma 12 *Let $\boldsymbol{G} \in \{0,1\}^{m\times n}$ and let $\boldsymbol{L} : \mathbb{R}^{d_z} \to \mathbb{R}^{m\times m}$ be a continuous function such that $\boldsymbol{L}(\boldsymbol{z})$ is invertible for all $\boldsymbol{z} \in \mathbb{R}^{d_z}$. Suppose that, for almost all $\boldsymbol{z} \in \mathbb{R}^{d_z}$ (i.e. except on a set E_0 of Lebesgue measure zero), there exists a permutation matrix $\boldsymbol{P}(\boldsymbol{z})$ such that $\boldsymbol{L}(\boldsymbol{z})\boldsymbol{P}(\boldsymbol{z})$ is $\boldsymbol{G}$-preserving. Then, there exists a permutation matrix $\boldsymbol{P}$ such that, for all $\boldsymbol{z} \in \mathbb{R}^{d_z}$, $\boldsymbol{L}(\boldsymbol{z})\boldsymbol{P}$ is $\boldsymbol{G}$-preserving.*

Proof We know that for all $\boldsymbol{z} \in \mathbb{R}^{d_z} \setminus E_0$, where $\mu(E_0) = 0$ (Lebesgue measure zero), there exists a permutation matrix $\boldsymbol{P}(\boldsymbol{z})$ such that $\boldsymbol{L}(\boldsymbol{z})\boldsymbol{P}(\boldsymbol{z})$ is $\boldsymbol{G}$-preserving. For all permutations $\boldsymbol{P}$, define $\mathcal{Z}^{(\boldsymbol{P})} := \{\boldsymbol{z} \in \mathbb{R}^{d_z} \setminus E_0 \mid \boldsymbol{P}(\boldsymbol{z}) = \boldsymbol{P}\}$. The collection of all sets $\mathcal{Z}^{(\boldsymbol{P})}$ is finite (since there are finitely many permutations) and forms a partition of $\mathbb{R}^{d_z} \setminus E_0$. Of course, for all $\boldsymbol{P}$, $\boldsymbol{L}(\boldsymbol{z})\boldsymbol{P}$ is $\boldsymbol{G}$-preserving for all $\boldsymbol{z} \in \mathcal{Z}^{(\boldsymbol{P})}$. By Lemma 8, we can extend this statement to the closure, i.e. for all $\boldsymbol{z} \in \overline{\mathcal{Z}^{(\boldsymbol{P})}}$, $\boldsymbol{L}(\boldsymbol{z})\boldsymbol{P}$ is $\boldsymbol{G}$-preserving.

Furthermore, we have that $\bigcup_{\boldsymbol{P}} \overline{\mathcal{Z}^{(\boldsymbol{P})}} = \overline{\bigcup_{\boldsymbol{P}} \mathcal{Z}^{(\boldsymbol{P})}} = \overline{\mathbb{R}^{d_z} \setminus E_0} = \mathbb{R}^{d_z}$, where the first equality is a standard property of closure (which holds only for finite unions), and the last equality holds by Lemma 9. We thus have that, for all $\boldsymbol{z} \in \mathbb{R}^{d_z}$, there exists a permutation $\boldsymbol{P}(\boldsymbol{z})$ such that $\boldsymbol{L}(\boldsymbol{z})\boldsymbol{P}(\boldsymbol{z})$ is $\boldsymbol{G}$-preserving. Since $\mathbb{R}^{d_z}$ is connected we can apply Lemma 11 to get the desired conclusion. ■

A.6 Connecting to the Graphical Criterion of Lachapelle et al. (2022)

The goal of this section is to prove Proposition 7 which states if some graphical criterion holds (Assumption 5), then $\boldsymbol{\theta} \sim_{\text{con}}^{\boldsymbol{z},\boldsymbol{a}} \hat{\boldsymbol{\theta}}$ implies $\boldsymbol{\theta} \sim_{\text{perm}} \hat{\boldsymbol{\theta}}$, i.e. complete disentanglement. We recall Assumption 5.

Assumption 5 (Graphical criterion, Lachapelle et al. (2022)) *Let $\boldsymbol{G} = [\boldsymbol{G}^z \; \boldsymbol{G}^a]$ be a graph. For all $i \in \{1, ..., d_z\}$,*

$$\left(\bigcap_{j \in \mathbf{Ch}_i^z} \mathbf{Pa}_j^z\right) \cap \left(\bigcap_{j \in \mathbf{Pa}_i^z} \mathbf{Ch}_j^z\right) \cap \left(\bigcap_{\ell \in \mathbf{Pa}_i^a} \mathbf{Ch}_\ell^a\right) = \{i\}\,,$$

where $\mathbf{Pa}_i^z$ and $\mathbf{Ch}_i^z$ are the sets of parents and children of node $\boldsymbol{z}_i$ in $\boldsymbol{G}^z$, respectively, while $\mathbf{Ch}_\ell^a$ is the set of children of $\boldsymbol{a}_\ell$ in $\boldsymbol{G}^a$.

We note that the above assumption is slightly different from the original one from Lachapelle et al. (2022), since the intersections run over $\mathbf{Ch}_i^z$, $\mathbf{Pa}_i^z$ and $\mathbf{Pa}_i^{\boldsymbol{a}}$ instead of over some sets of indexes $\mathcal{I}, \mathcal{J} \subseteq \{1, ..., d_z\}$ and $\mathcal{L} \subseteq \{1, ..., d_a\}$. This slightly simplified criterion is equivalent to the original one, which we now demonstrate for the interested reader.

Proposition 11 *Let $\boldsymbol{G} = [\boldsymbol{G}^z \; \boldsymbol{G}^a] \in \{0,1\}^{d_z \times (d_z + d_a)}$. The criterion of Assumption 5 holds for $\boldsymbol{G}$ if and only if the following holds for $\boldsymbol{G}$: For all $i \in \{1, ..., d_z\}$, there exist sets $\mathcal{I}, \mathcal{J} \subseteq \{1, ..., d_z\}$ and $\mathcal{L} \subseteq \{1, ..., d_a\}$ such that*

$$\left(\bigcap_{j \in \mathcal{I}} \mathbf{Pa}_j^{\boldsymbol{z}}\right) \cap \left(\bigcap_{j \in \mathcal{J}} \mathbf{Ch}_j^{\boldsymbol{z}}\right) \cap \left(\bigcap_{\ell \in \mathcal{L}} \mathbf{Ch}_\ell^{\boldsymbol{a}}\right) = \{i\}\,,$$

Proof The direction " $\implies$ " is trivial, since we can simply choose $\mathcal{I} := \mathbf{Ch}_i^z$, $\mathcal{J} := \mathbf{Pa}_i^z$ and $\mathcal{L} := \mathbf{Pa}_i^a$.

To show the other direction, we notice that we must have $\mathcal{I} \subseteq \mathbf{Ch}_i^{\boldsymbol{z}}$, $\mathcal{J} \subseteq \mathbf{Pa}_i^z$ and $\mathcal{L} \subseteq \mathbf{Pa}_i^a$, otherwise one of the sets in the intersection would not contain i, contradicting the criterion. Thus, the criterion of Def. 5 intersects the same sets or more sets. Moreover, these potential additional sets must contain i because of the obvious facts that $j \in \mathbf{Ch}_i^z \iff i \in \mathbf{Pa}_j^{\boldsymbol{z}}$ and $\ell \in \mathbf{Pa}_i^a \iff i \in \mathbf{Ch}_\ell^{\boldsymbol{a}}$, thus they do not change the result of the intersection. ■

To prove Proposition 7, we will need the following lemma.

Lemma 13 *Let $\boldsymbol{G} \in \{0,1\}^{m \times n}$ and $\mathbf{c}$ be a diffeomorphism with a dependency graph given by $\boldsymbol{C} \in \{0,1\}^{m \times m}$ (Definition 2). The function $\mathbf{c}$ is $\boldsymbol{G}$-preserving (Definition 12) if and only if*

$$\forall i, \boldsymbol{C}_{i,\cdot} \subseteq \bigcap_{k \in \boldsymbol{G}_{i,\cdot}} \boldsymbol{G}_{\cdot,k}\,.$$

Proof We leverage Proposition 3.

$$\begin{aligned} \boldsymbol{G}_{i,\cdot} \not\subseteq \boldsymbol{G}_{j,\cdot} &\iff \exists k \text{ s.t. } \boldsymbol{G}_{i,k} = 1 \text{ and } \boldsymbol{G}_{j,k} = 0 \\ &\iff \exists k \in \boldsymbol{G}_{i,\cdot} \text{ s.t. } j \notin \boldsymbol{G}_{\cdot,k} \\ &\iff j \notin \bigcap_{k \in \boldsymbol{G}_{i,\cdot}} \boldsymbol{G}_{\cdot,k}\,. \end{aligned} \tag{55}$$

■

Proposition 7 (Complete disentanglement as a special case) *Let $\boldsymbol{\theta} := (\boldsymbol{f}, p, \boldsymbol{G})$ and $\hat{\boldsymbol{\theta}} := (\hat{\boldsymbol{f}}, \hat{p}, \hat{\boldsymbol{G}})$ be two models satisfying Assumptions 1, 2 & 3. If $\boldsymbol{\theta} \sim_{\text{con}}^{\boldsymbol{z},\boldsymbol{a}} \hat{\boldsymbol{\theta}}$ and $\boldsymbol{G}$ satisfies Assumption 5, then $\boldsymbol{\theta} \sim_{\text{perm}} \hat{\boldsymbol{\theta}}$.*

Proof By definition of $\sim_{\text{con}}^{\boldsymbol{z},\boldsymbol{a}}$, we know that the entanglement graph for $(\boldsymbol{f}, \hat{\boldsymbol{f}})$ is given by $\boldsymbol{V} = \boldsymbol{C}\boldsymbol{P}^\top$ where $\boldsymbol{P}$ is a permutation and $\boldsymbol{C}$ is a binary matrix that is $\boldsymbol{G}^a$-preserving, $\boldsymbol{G}^z$-preserving and $(\boldsymbol{G}^z)^\top$-preserving. Using Lemma 13, we have that, for all i,

$$\begin{aligned}
\boldsymbol{C}_{i,\cdot} &\subseteq \left(\bigcap_{j \in \boldsymbol{G}^{\boldsymbol{z}}_{i,\cdot}} \boldsymbol{G}^{\boldsymbol{z}}_{\cdot,j}\right) \cap \left(\bigcap_{j \in \boldsymbol{G}^{\boldsymbol{z}}_{\cdot,i}} \boldsymbol{G}^{\boldsymbol{z}}_{j,\cdot}\right) \cap \left(\bigcap_{j \in \boldsymbol{G}^{\boldsymbol{a}}_{i,\cdot}} \boldsymbol{G}^{\boldsymbol{a}}_{\cdot,j}\right) \\
&= \left(\bigcap_{j \in \mathbf{Pa}^z_i} \mathbf{Ch}^{\boldsymbol{z}}_j\right) \cap \left(\bigcap_{j \in \mathbf{Ch}^z_i} \mathbf{Pa}^{\boldsymbol{z}}_j\right) \cap \left(\bigcap_{\ell \in \mathbf{Pa}^{\boldsymbol{a}}_i} \mathbf{Ch}^{\boldsymbol{a}}_\ell\right) \\
&= \{i\}\,.
\end{aligned}$$

Thus, $\boldsymbol{C}$ is in fact the identity matrix, and hence $\boldsymbol{\theta} \sim_{\text{perm}} \hat{\boldsymbol{\theta}}$. ■

Appendix B. Identifiability Theory - Exponential Family Case

B.1 Technical Lemmas and Definitions

We recall the definition of a minimal sufficient statistic in an exponential family, which can be found in Wainwright and Jordan (2008, p. 40).

Definition 20 (Minimal sufficient statistic) *Given a parameterized distribution in the exponential family, as in (30), we say that its sufficient statistic $\boldsymbol{s}_i$ is minimal when there is no $v \neq 0$ such that $v^\top \boldsymbol{s}_i(z)$ is constant for all $z \in \mathcal{Z}$.*

The following Lemma gives a characterization of minimality, which will be useful in the proof of Theorem 4.

Lemma 14 (Characterization of minimal $\boldsymbol{s}$) *A sufficient statistic of an exponential family distribution $\boldsymbol{s} : \mathcal{Z} \to \mathbb{R}^k$ is minimal if and only if there exists $\boldsymbol{z}_{(0)}$, $\boldsymbol{z}_{(1)}$, ..., $\boldsymbol{z}_{(k)}$ belonging to the support $\mathcal{Z}$ such that the following k-dimensional vectors are linearly independent:*

$$\boldsymbol{s}(\boldsymbol{z}_{(1)}) - \boldsymbol{s}(\boldsymbol{z}_{(0)}), ..., \boldsymbol{s}(\boldsymbol{z}_{(k)}) - \boldsymbol{s}(\boldsymbol{z}_{(0)})\,. \tag{56}$$

Proof. We start by showing the "if" part of the statement. Suppose there exist $\boldsymbol{z}_{(0)}, ..., \boldsymbol{z}_{(k)}$ in $\mathcal{Z}$ such that the vectors of (56) are linearly independent. By contradiction, suppose that $\boldsymbol{s}$ is not minimal, i.e. there exist a nonzero vector v and a scalar b such that $v^\top \boldsymbol{s}(z) = b$ for all $z \in \mathcal{Z}$. Notice that

$b = v^\top \boldsymbol{s}(\boldsymbol{z}_{(0)})$. Hence, $v^\top(\boldsymbol{s}(\boldsymbol{z}_{(i)}) - \boldsymbol{s}(\boldsymbol{z}_{(0)})) = 0$ for all $i = 1, ..., k$. This can be rewritten in matrix form as

$$v^\top [\boldsymbol{s}(\boldsymbol{z}_{(1)}) - \boldsymbol{s}(\boldsymbol{z}_{(0)}) \; ... \; \boldsymbol{s}(\boldsymbol{z}_{(k)}) - \boldsymbol{s}(\boldsymbol{z}_{(0)})] = 0\,,$$

which implies that the matrix in the above equation is not invertible. This is a contradiction.

We now show the "only if " part of the statement. Suppose that there is no $\boldsymbol{z}_{(0)}$, ..., $\boldsymbol{z}_{(k)}$ such that the vectors of (56) are linearly independent. Choose an arbitrary $\boldsymbol{z}_{(0)} \in \mathcal{Z}$. We thus have that $U := \text{span}\{\boldsymbol{s}(z) - \boldsymbol{s}(\boldsymbol{z}_{(0)}) \mid z \in \mathcal{Z}\}$ is a proper subspace of $\mathbb{R}^k$. This means the orthogonal complement of U, $U^\perp$, has dimension 1 or greater. We can thus pick a nonzero vector $v \in U^\perp$ such that $v^\top(\boldsymbol{s}(z) - \boldsymbol{s}(\boldsymbol{z}_0)) = 0$ for all $z \in \mathcal{Z}$, which is to say that $v^\top \boldsymbol{s}(z)$ is constant for all $z \in \mathcal{Z}$, and thus, $\boldsymbol{s}$ is not minimal. ■

B.2 Proof of Linear Identifiability (Theorem 4)

Theorem 4 (Conditions for linear identifiability - Adapted from Khemakhem et al. (2020a)) *Let $\boldsymbol{\theta} := (\boldsymbol{f}, \boldsymbol{\lambda}, \boldsymbol{G})$ and $\hat{\boldsymbol{\theta}} := (\hat{\boldsymbol{f}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{G}})$ be two models satisfying Assumptions 1, 2 & 9. Further assume that*

1. ***[Observational equivalence]*** *$\boldsymbol{\theta} \sim_{\text{obs}} \hat{\boldsymbol{\theta}}$ (Definition 4);*
2. ***[Minimal sufficient statistics]*** *For all i, the sufficient statistic $\boldsymbol{s}_i$ is minimal (see below).*
3. ***[Sufficient variability]*** *The natural parameter $\boldsymbol{\lambda}$ varies "sufficiently" as formalized by Assumption 10 (see below).*

Then, $\boldsymbol{\theta} \sim_{\text{lin}} \hat{\boldsymbol{\theta}}$ (Def. 17).

Proof First, we apply Proposition 2 to get

$$\tilde{p}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = p(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t}) |\det D\boldsymbol{v}(\boldsymbol{z}^t)|\,. \tag{57}$$

Linear relationship between $\boldsymbol{s}(\boldsymbol{f}^{-1}(x))$ and $\boldsymbol{s}(\hat{\boldsymbol{f}}^{-1}(x))$. By taking the logarithm on each side of (57) and making explicit the exponential family form, we get

$$\begin{aligned}
&\sum_{i=1}^{d_z} \log h_i(\boldsymbol{z}_i^t) + \boldsymbol{s}_i(\boldsymbol{z}_i^t)^\top \boldsymbol{\lambda}_i(\boldsymbol{G}_i^z \odot \boldsymbol{z}^{<t}, \boldsymbol{G}_i^a \odot \boldsymbol{a}^{<t}) - \psi_i(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) \qquad (58)\\
=&\sum_{i=1}^{d_z} \log h_i(\boldsymbol{v}_i(\boldsymbol{z}^t)) + \boldsymbol{s}_i(\boldsymbol{v}_i(\boldsymbol{z}^t)))^\top \hat{\boldsymbol{\lambda}}_i(\hat{\boldsymbol{G}}_i^z \odot \boldsymbol{v}(\boldsymbol{z}^{<t}), \hat{\boldsymbol{G}}_i^a \odot \boldsymbol{a}^{<t}) - \hat{\psi}_i(\boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t})\\
&\qquad\qquad + \log|\det D\boldsymbol{v}(\boldsymbol{z}^t)|
\end{aligned}$$

Note that (58) holds for all $\boldsymbol{z}^{<t}$ and $\boldsymbol{a}^{<t}$. In particular, we evaluate it at the points given in the assumption of sufficient variability of Theorem 4. We evaluate the equation at $(\boldsymbol{z}^t, \boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)})$ and

$(\boldsymbol{z}^t, \boldsymbol{z}_{(0)}, \boldsymbol{a}_{(0)})$ and take the difference which yields[10]

$$\begin{aligned}
&\sum_{i=1}^{d_z} \boldsymbol{s}_i(\boldsymbol{z}_i^t)^\top [\boldsymbol{\lambda}_i(\boldsymbol{G}_i^z \odot \boldsymbol{z}_{(r)}, \boldsymbol{G}_i^a \odot \boldsymbol{a}_{(r)}) - \boldsymbol{\lambda}_i(\boldsymbol{G}_i^z \odot \boldsymbol{z}_{(0)}, \boldsymbol{G}_i^a \odot \boldsymbol{a}_{(0)})] - \psi_i(\boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)}) + \psi_i(\boldsymbol{z}_{(0)}, \boldsymbol{a}_{(0)}) \\
=&\sum_{i=1}^{d_z} \boldsymbol{s}_i(\boldsymbol{v}_i(\boldsymbol{z}^t))^\top [\hat{\boldsymbol{\lambda}}_i(\hat{\boldsymbol{G}}_i^z \odot \boldsymbol{v}(\boldsymbol{z}_{(r)}), \hat{\boldsymbol{G}}_i^a \odot \boldsymbol{a}_{(r)}) - \hat{\boldsymbol{\lambda}}_i(\hat{\boldsymbol{G}}_i^z \odot \boldsymbol{v}(\boldsymbol{z}_{(0)}), \hat{\boldsymbol{G}}_i^a \odot \boldsymbol{a}_{(0)})] \\
&\qquad - \hat{\psi}_i(\boldsymbol{v}(\boldsymbol{z}_{(r)}), \boldsymbol{a}_{(r)}) + \hat{\psi}_i(\boldsymbol{v}(\boldsymbol{z}_{(0)}), \boldsymbol{a}_{(0)})
\end{aligned} \tag{59}$$

We regroup all normalization constants ψ into a term $d(\boldsymbol{z}_{(r)}, \boldsymbol{z}_{(0)}, \boldsymbol{a}_{(r)}, \boldsymbol{a}_{(0)})$ and write

$$\begin{aligned}
&\boldsymbol{s}(\boldsymbol{z}^t)^\top [\boldsymbol{\lambda}(\boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)}) - \boldsymbol{\lambda}(\boldsymbol{z}_{(0)}, \boldsymbol{a}_{(0)})] \\
=&\boldsymbol{s}(\boldsymbol{v}(\boldsymbol{z}^t))^\top [\hat{\boldsymbol{\lambda}}(\boldsymbol{v}(\boldsymbol{z}_{(r)}), \boldsymbol{a}_{(r)}) - \hat{\boldsymbol{\lambda}}(\boldsymbol{v}(\boldsymbol{z}_{(0)}), \boldsymbol{a}_{(0)})] + d(\boldsymbol{z}_{(r)}, \boldsymbol{z}_{(0)}, \boldsymbol{a}_{(r)}, \boldsymbol{a}_{(0)})\,.
\end{aligned} \tag{60}$$

Define

$$\begin{aligned}
\boldsymbol{w}_{(r)} &:= \boldsymbol{\lambda}(\boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)}) - \boldsymbol{\lambda}(\boldsymbol{z}_{(0)}, \boldsymbol{a}_{(0)}) \\
\hat{\boldsymbol{w}}_{(r)} &:= \hat{\boldsymbol{\lambda}}(\boldsymbol{v}(\boldsymbol{z}_{(r)}), \boldsymbol{a}_{(r)}) - \hat{\boldsymbol{\lambda}}(\boldsymbol{v}(\boldsymbol{z}_{(0)}), \boldsymbol{a}_{(0)}) \\
d_{(r)} &:= d(\boldsymbol{z}_{(r)}, \boldsymbol{z}_{(0)}, \boldsymbol{a}_{(r)}, \boldsymbol{a}_{(0)})\,,
\end{aligned}$$

which yields

$$\boldsymbol{s}(\boldsymbol{z}^t)^\top \boldsymbol{w}_{(r)} = \boldsymbol{s}(\boldsymbol{v}(\boldsymbol{z}^t))^\top \hat{\boldsymbol{w}}_{(r)} + d_{(r)}\,. \tag{61}$$

We can regroup the $\boldsymbol{w}_{(r)}$ into a matrix and the $d_{(r)}$ into a vector:

$$\begin{aligned}
\boldsymbol{W} &:= [\boldsymbol{w}_{(1)} \dots \boldsymbol{w}_{(kd_z)}] \in \mathbb{R}^{kd_z \times kd_z} \\
\hat{\boldsymbol{W}} &:= [\hat{\boldsymbol{w}}_{(1)} \dots \hat{\boldsymbol{w}}_{(kd_z)}] \in \mathbb{R}^{kd_z \times kd_z} \\
\boldsymbol{d} &:= [d_{(1)} \dots d_{(kd_z)}] \in \mathbb{R}^{1 \times kd_z}\,.
\end{aligned}$$

Since (61) holds for all $1 \leq p \leq kd_z$, we can write

$$\boldsymbol{s}(\boldsymbol{z}^t)^\top \boldsymbol{W} = \boldsymbol{s}(\boldsymbol{v}(\boldsymbol{z}^t))^\top \hat{\boldsymbol{W}} + \boldsymbol{d}\,.$$

Note that $\boldsymbol{W}$ is invertible by the assumption of variability, hence

$$\boldsymbol{s}(\boldsymbol{z}^t)^\top = \boldsymbol{s}(\boldsymbol{v}(\boldsymbol{z}^t))^\top \hat{\boldsymbol{W}} \boldsymbol{W}^{-1} + \boldsymbol{d}\boldsymbol{W}^{-1}\,.$$

Let $\boldsymbol{b} := (\boldsymbol{d}\boldsymbol{W}^{-1})^\top$ and $\boldsymbol{L} := (\hat{\boldsymbol{W}}\boldsymbol{W}^{-1})^\top$. We can thus rewrite this as

$$\boldsymbol{s}(\boldsymbol{z}^t) = \boldsymbol{L}\boldsymbol{s}(\boldsymbol{v}(\boldsymbol{z}^t)) + \boldsymbol{b}\,. \tag{62}$$

10. Note that $\boldsymbol{z}_{(0)}$ and $\boldsymbol{z}_{(r)}$ can have different dimensionalities if they come from different time steps. It is not an issue to combine equations from different time steps, since (58) holds for all values of t, $\boldsymbol{z}^t$, $\boldsymbol{z}^{<t}$ and $\boldsymbol{a}^{<t}$.

Invertibility of $\boldsymbol{L}$. We now show that $\boldsymbol{L}$ is invertible. By Lemma 14, the fact that the $\boldsymbol{s}_i$ are minimal is equivalent to, for all $i \in \{1, ..., d_z\}$, having elements $\boldsymbol{z}_i^{(0)}$, ..., $\boldsymbol{z}_i^{(k)}$ in $\mathcal{Z}$ such that the family of vectors

$$\boldsymbol{s}_i(\boldsymbol{z}_i^{(1)}) - \boldsymbol{s}_i(\boldsymbol{z}_i^{(0)}),\ \dots\ , \boldsymbol{s}_i(\boldsymbol{z}_i^{(k)}) - \boldsymbol{s}_i(\boldsymbol{z}_i^{(0)})$$

is linearly independent. Define

$$\boldsymbol{z}^{(0)} := [\boldsymbol{z}_1^{(0)} \dots \boldsymbol{z}_{d_z}^{(0)}]^\top \in \mathbb{R}^{d_z}$$

For all $i \in \{1, ..., d_z\}$ and all $p \in \{1, ..., k\}$, define the vectors

$$\boldsymbol{z}^{(p,i)} := [\boldsymbol{z}_1^{(0)} \dots \boldsymbol{z}_{i-1}^{(0)}\ \boldsymbol{z}_i^{(p)}\ \boldsymbol{z}_{i+1}^{(0)} \dots \boldsymbol{z}_{d_z}^{(0)}]^\top \in \mathbb{R}^{d_z}\,.$$

For a specific $1 \leq p \leq k$ and $i \in \{1, ..., d_z\}$, we can take the following difference based on (62)

$$\boldsymbol{s}(\boldsymbol{z}^{(p,i)}) - \boldsymbol{s}(\boldsymbol{z}^{(0)}) = \boldsymbol{L}[\boldsymbol{s}(\boldsymbol{v}(\boldsymbol{z}^{(p,i)})) - \boldsymbol{s}(\boldsymbol{v}(\boldsymbol{z}^{(0)}))]\,, \tag{63}$$

where the left hand side is a vector filled with zeros except for the block corresponding to $\boldsymbol{s}_i(\boldsymbol{z}_i^{(p,i)}) - \boldsymbol{s}_i(\boldsymbol{z}_i^{(0)})$. Let us define

$$\begin{aligned}
\Delta\boldsymbol{s}^{(i)} &:= [\boldsymbol{s}(\boldsymbol{z}^{(1,i)}) - \boldsymbol{s}(\boldsymbol{z}^{(0)})\ \dots\ \boldsymbol{s}(\boldsymbol{z}^{(k,i)}) - \boldsymbol{s}(\boldsymbol{z}^{(0)})] \in \mathbb{R}^{kd_z \times k} \\
\Delta\hat{\boldsymbol{s}}^{(i)} &:= [\boldsymbol{s}(\boldsymbol{v}(\boldsymbol{z}^{(1,i)})) - \boldsymbol{s}(\boldsymbol{v}(\boldsymbol{z}^{(0)}))\ \dots\ \boldsymbol{s}(\boldsymbol{v}(\boldsymbol{z}^{(k,i)})) - \boldsymbol{s}(\boldsymbol{v}(\boldsymbol{z}^{(0)}))] \in \mathbb{R}^{kd_z \times k}\,.
\end{aligned}$$

Note that the columns of $\Delta\boldsymbol{s}^{(i)}$ are linearly independent and all rows are filled with zeros except for the block of rows $\{(i-1)k + 1, ..., ik\}$. We can thus rewrite (63) in matrix form

$$\Delta\boldsymbol{s}^{(i)} = \boldsymbol{L}\Delta\hat{\boldsymbol{s}}^{(i)}\,.$$

We can regroup these equations for every i by doing

$$[\Delta\boldsymbol{s}^{(1)}\ ...\ \Delta\boldsymbol{s}^{(d_z)}] = \boldsymbol{L}[\Delta\hat{\boldsymbol{s}}^{(1)}\ ...\ \Delta\hat{\boldsymbol{s}}^{(d_z)}]\,. \tag{64}$$

Notice that the newly formed matrix on the left hand side has size $kd_z \times kd_z$ and is block diagonal. Since every block is invertible, the left hand side of (64) is an invertible matrix, which in turn implies that $\boldsymbol{L}$ is invertible. This completes the proof. ■

B.3 Proof of Theorem 5

Lemma 15 *Let $\boldsymbol{\theta} := (\boldsymbol{f}, p, \boldsymbol{G})$ satisfy Assumptions 1, 2, 3, 4 & 9 and let $q := \log p$. Then*

$$\begin{aligned}
D_z^t q(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) &= \boldsymbol{\lambda}(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})^\top D\boldsymbol{s}(\boldsymbol{z}^t) + D(\log h)(\boldsymbol{z}^t) \\
H_{z,a}^{t,\tau} q(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) &= D\boldsymbol{s}(\boldsymbol{z}^t)^\top D_a^\tau \boldsymbol{\lambda}(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) \\
H_{z,z}^{t,\tau} q(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) &= D\boldsymbol{s}(\boldsymbol{z}^t)^\top D_z^\tau \boldsymbol{\lambda}(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})\,.
\end{aligned}$$

Proof We have

$$\log p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) := \log h(\boldsymbol{z}^t) + \boldsymbol{s}(\boldsymbol{z}^t)^\top \boldsymbol{\lambda}(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) - \psi(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$$
$$\log h(\boldsymbol{z}^t) + \boldsymbol{\lambda}(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})^\top \boldsymbol{s}(\boldsymbol{z}^t) - \psi(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) \,.$$

We can differentiate the above w.r.t. $\boldsymbol{z}^t$ to get

$$D_z^t q(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = \boldsymbol{\lambda}(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})^\top D\boldsymbol{s}(\boldsymbol{z}^t) + D(\log h)(\boldsymbol{z}^t)$$

Differentiating the above w.r.t. $\boldsymbol{z}^\tau$ or $\boldsymbol{a}^\tau$ yields the desired result. ■

Theorem 5 (Disentanglement via sparse temporal dependencies in exponential families) *Let $\boldsymbol{\theta} := (\boldsymbol{f}, \boldsymbol{\lambda}, \boldsymbol{G})$ and $\hat{\boldsymbol{\theta}} := (\hat{\boldsymbol{f}}, \hat{\boldsymbol{\lambda}}, \hat{\boldsymbol{G}})$ be two models satisfying Assumptions 1, 2, 3, 4, 9 as well as all assumptions of Theorem 4. Further suppose that*

1. *The sufficient statistic $\boldsymbol{s}$ is d_z-dimensional ($k = 1$) and is a diffeomorphism from $\mathbb{R}^{d_z}$ to $\boldsymbol{s}(\mathbb{R}^{d_z})$;*
2. ***[Sufficient influence of z]** The Jacobian of the ground-truth transition function $\boldsymbol{\lambda}$ with respect to $\boldsymbol{z}$ varies "sufficiently", as formalized in Assumption 11;*

Then, there exists a permutation matrix $\boldsymbol{P}$ such that $\boldsymbol{P}\boldsymbol{G}^z\boldsymbol{P}^\top \subseteq \hat{\boldsymbol{G}}^z$. Further assume that

3. ***[Sparsity regularization]** $||\hat{\boldsymbol{G}}^z||_0 \leq ||\boldsymbol{G}^z||_0$;*

Then, $\boldsymbol{\theta} \sim_{\text{con}}^{\boldsymbol{z}} \hat{\boldsymbol{\theta}}$ (Def. 14) & $\boldsymbol{\theta} \sim_{\text{lin}} \hat{\boldsymbol{\theta}}$ (Def. 17), which together implies that

$$\boldsymbol{v}(\boldsymbol{z}) = \boldsymbol{s}^{-1}(\boldsymbol{C}\boldsymbol{P}^\top \boldsymbol{s}(\boldsymbol{z}) + \boldsymbol{b}) \,,$$

where $\boldsymbol{b} \in \mathbb{R}^{d_z}$ and $\boldsymbol{C} \in \mathbb{R}^{d_z \times d_z}$ is invertible, $\boldsymbol{G}^{\boldsymbol{z}}$- and $(\boldsymbol{G}^{\boldsymbol{z}})^\top$-preserving (Definition 11).

Proof Recall the equation we derived in Section 3.2:

$$H_{z,z}^{t,\tau} \hat{q}(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = D\boldsymbol{v}(\boldsymbol{z}^t)^\top H_{z,z}^{t,\tau} q(\boldsymbol{v}(\boldsymbol{z}^t) \mid \boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t}) D\boldsymbol{v}(\boldsymbol{z}^\tau) \,.$$

Using Lemma 15, we get

$$D\boldsymbol{s}(\boldsymbol{z}^t)^\top D_z^\tau \hat{\boldsymbol{\lambda}}(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = D\boldsymbol{v}(\boldsymbol{z}^t)^\top D\boldsymbol{s}(\boldsymbol{v}(\boldsymbol{z}^t))^\top D_z^\tau \boldsymbol{\lambda}(\boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t}) D\boldsymbol{v}(\boldsymbol{z}^\tau) \,. \tag{65}$$

Note that Assumption 3 requires that $D_z^\tau \boldsymbol{\lambda}(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) \subseteq \boldsymbol{G}^z$ and $D_z^\tau \hat{\boldsymbol{\lambda}}(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) \subseteq \hat{\boldsymbol{G}}^z$. Theorem 4 implies that there exist an invertible matrix $\boldsymbol{L} \in \mathbb{R}^{d_z \times d_z}$ and a vector $\boldsymbol{b} \in \mathbb{R}^{d_z}$ such that

$$\boldsymbol{v}(\boldsymbol{z}) = \boldsymbol{s}^{-1}(\boldsymbol{L}\boldsymbol{s}(\boldsymbol{z}) + \boldsymbol{b}) \,.$$

Taking the derivative of the above w.r.t. $\boldsymbol{z}$, we obtain

$$D\boldsymbol{v}(\boldsymbol{z}) = D\boldsymbol{s}^{-1}(\boldsymbol{L}\boldsymbol{s}(\boldsymbol{z}) + \boldsymbol{b})\boldsymbol{L}D\boldsymbol{s}(\boldsymbol{z}) \tag{66}$$
$$= D\boldsymbol{s}^{-1}(\boldsymbol{s}(\boldsymbol{v}(\boldsymbol{z})))\boldsymbol{L}D\boldsymbol{s}(\boldsymbol{z}) \tag{67}$$
$$= D\boldsymbol{s}(\boldsymbol{v}(\boldsymbol{z}))^{-1}\boldsymbol{L}D\boldsymbol{s}(\boldsymbol{z}) \,, \tag{68}$$

where we used $\boldsymbol{s}(\boldsymbol{v}(\boldsymbol{z})) = \boldsymbol{L}\boldsymbol{s}(\boldsymbol{z}) + \boldsymbol{b}$ to go from the first to the second line and used the inverse function theorem to go from the second to the third line. Plugging (68) into (65) yields

$$\begin{aligned}
&D\boldsymbol{s}(\boldsymbol{z}^t)^\top D_z^\tau \hat{\boldsymbol{\lambda}}(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) \\
&= D\boldsymbol{s}(\boldsymbol{z}^t)^\top \boldsymbol{L}^\top D\boldsymbol{s}(\boldsymbol{v}(\boldsymbol{z}^t))^{-\top} D\boldsymbol{s}(\boldsymbol{v}(\boldsymbol{z}^t))^\top D_z^\tau \boldsymbol{\lambda}(\boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t}) D\boldsymbol{s}(\boldsymbol{v}(\boldsymbol{z}^\tau))^{-1} \boldsymbol{L} D\boldsymbol{s}(\boldsymbol{z}^\tau) \\
&= D\boldsymbol{s}(\boldsymbol{z}^t)^\top \boldsymbol{L}^\top D_z^\tau \boldsymbol{\lambda}(\boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t}) D\boldsymbol{s}(\boldsymbol{v}(\boldsymbol{z}^\tau))^{-1} \boldsymbol{L} D\boldsymbol{s}(\boldsymbol{z}^\tau)\,,
\end{aligned}$$

which implies

$$D\boldsymbol{s}(\boldsymbol{z}^t)^\top D_z^\tau \hat{\boldsymbol{\lambda}}(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) = D\boldsymbol{s}(\boldsymbol{z}^t)^\top \boldsymbol{L}^\top D_z^\tau \boldsymbol{\lambda}(\boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t}) D\boldsymbol{s}(\boldsymbol{v}(\boldsymbol{z}^\tau))^{-1} \boldsymbol{L} D\boldsymbol{s}(\boldsymbol{z}^\tau) \tag{69}$$

$$D_z^\tau \hat{\boldsymbol{\lambda}}(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) D\boldsymbol{s}(\boldsymbol{z}^\tau)^{-1} = \boldsymbol{L}^\top D_z^\tau \boldsymbol{\lambda}(\boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t}) D\boldsymbol{s}(\boldsymbol{v}(\boldsymbol{z}^\tau))^{-1} \boldsymbol{L}\,, \tag{70}$$

where we right- and left-multiplied by $D\boldsymbol{s}(\boldsymbol{z}^t)^{-\top}$ and $D\boldsymbol{s}(\boldsymbol{z}^\tau)$, respectively. Let us define

$$\Lambda(\gamma) := D_z^\tau \boldsymbol{\lambda}(\boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t}) D\boldsymbol{s}(\boldsymbol{v}(\boldsymbol{z}^\tau))^{-1} \quad \hat{\Lambda}(\gamma) := D_z^\tau \hat{\boldsymbol{\lambda}}(\boldsymbol{v}(\boldsymbol{z}^{<t}), \boldsymbol{a}^{<t}) D\boldsymbol{s}(\boldsymbol{v}(\boldsymbol{z}^\tau))^{-1}\,,$$

where $\gamma = (t, \tau, \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$. Note that because $D\boldsymbol{s}$ is diagonal, we have that $\Lambda(\gamma) \subseteq \boldsymbol{G}^z$ and $\hat{\Lambda}(\gamma) \subseteq \hat{\boldsymbol{G}}^z$. Using this notation, we can rewrite (70) as

$$\underbrace{\hat{\Lambda}(\gamma)}_{\subseteq \hat{\boldsymbol{G}}^z} = \boldsymbol{L}^\top \underbrace{\Lambda(\gamma)}_{\subseteq \boldsymbol{G}^z} \boldsymbol{L}\,.$$

Thanks to Assumption 11, we can apply the same argument as in Theorem 5 to show that $\boldsymbol{L} = \boldsymbol{C}\boldsymbol{P}^\top$ where $\boldsymbol{C}$ is a matrix that is both $\boldsymbol{G}^z$-preserving and $(\boldsymbol{G}^z)^\top$-preserving, as desired. ■

B.4 Connecting to Sufficient Influence Assumptions of Lachapelle et al. (2022)

In this section, we relate the nonparametric sufficient influence assumptions of this work, i.e. Assumptions 7 & 8, to the analogous assumptions of Lachapelle et al. (2022) for exponential families, i.e. Assumptions 11 & 12, the latter of which we recall below.

Assumption 12 (Sufficient influence of $\boldsymbol{a}$ (Lachapelle et al., 2022)) *Assume $k = 1$, i.e. the sufficient statistics $\boldsymbol{s}_i$ are one-dimensional. For all $\ell \in \{1, ..., d_a\}$, there exist $\{(\boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)}, \epsilon_{(r)}, \tau_{(r)})\}_{r=1}^{|\mathbf{Ch}_\ell^{\boldsymbol{a}}|}$ belonging to their respective support such that*

$$\operatorname{span}\left\{\boldsymbol{\lambda}(\boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)} + \epsilon_{(r)} \boldsymbol{E}^{(\ell,\tau)}) - \boldsymbol{\lambda}(\boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)})\right\}_{r=1}^{|\mathbf{Ch}_\ell^{\boldsymbol{a}}|} = \mathbb{R}^{d_z}_{\mathbf{Ch}_\ell^{\boldsymbol{a}}}\,,$$

where $\epsilon \in \mathbb{R}$ and $\boldsymbol{E}^{(\ell,\tau)} \in \mathbb{R}^{d_a \times t}$ is the one-hot matrix with the entry (ℓ, τ) set to one.

The following proposition shows that, when the exponential family holds with $k = 1$, we have that (i) for the "sufficient influence of $\boldsymbol{a}$" assumptions, the nonparametric and exponential family versions are actually equivalent, and (ii) for the "sufficient influence of $\boldsymbol{z}$" assumptions, the nonparametric version implies the exponential family version.

Proposition 12 (Sufficient influence assumptions: nonparametric v.s. exponential) *Let the parameter $\boldsymbol{\theta} := (\boldsymbol{f}, p, \boldsymbol{G})$ satisfy Assumptions 1, 2, 3 & 9. Further, assume that $k = 1$ and that $D\boldsymbol{s}(\boldsymbol{z}) \in \mathbb{R}^{d_z \times d_z}$ is invertible everywhere. Then,*

Sufficient influence of $\boldsymbol{a}$: Assumption 7 (nonparametric) $\iff$ Assumption 12 (exponential family)
Sufficient influence of $\boldsymbol{z}$: Assumption 8 (nonparametric) $\implies$ Assumption 11 (exponential family)

Proof We start by proving the first equivalence for the sufficient influence of $\boldsymbol{a}$ assumptions. By using Lemma 15 we see that

$$\begin{aligned}
&\operatorname{span}\left\{D_z^{t_{(r)}} \log p(\boldsymbol{z} \mid \boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)} + \epsilon_{(r)}\boldsymbol{E}^{(\ell,\tau_{(r)})}) - D_z^{t_{(r)}} \log p(\boldsymbol{z} \mid \boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)})\right\}_{r=1}^{|\mathbf{Ch}_\ell^{\boldsymbol{a}}|} \\
&= \operatorname{span}\left\{D\boldsymbol{s}(\boldsymbol{z})^\top \boldsymbol{\lambda}(\boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)} + \epsilon_{(r)}\boldsymbol{E}^{(\ell,\tau_{(r)})}) - D\boldsymbol{s}(\boldsymbol{z})^\top \boldsymbol{\lambda}(\boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)})\right\}_{r=1}^{|\mathbf{Ch}_\ell^{\boldsymbol{a}}|} \\
&= D\boldsymbol{s}(\boldsymbol{z})^\top \operatorname{span}\left\{\boldsymbol{\lambda}(\boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)} + \epsilon_{(r)}\boldsymbol{E}^{(\ell,\tau_{(r)})}) - \boldsymbol{\lambda}(\boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)})\right\}_{r=1}^{|\mathbf{Ch}_\ell^{\boldsymbol{a}}|}. \qquad (71)
\end{aligned}$$

We start by showing " $\impliedby$ ". Assumption 12 implies that (71) is equal to $D\boldsymbol{s}(\boldsymbol{z}^t)^\top \mathbb{R}^{d_z}_{\mathbf{Ch}_\ell^{\boldsymbol{a}}}$ which is equal to $\mathbb{R}^{d_z}_{\mathbf{Ch}_\ell^{\boldsymbol{a}}}$ since $D\boldsymbol{s}(\boldsymbol{z}^t)$ is invertible everywhere and is diagonal. To show " $\implies$ ", we can apply the same argument.

We now show that Assumption 8 implies Assumption 11. we again use Lemma 15 and see that

$$\begin{aligned}
\mathbb{R}^{d_z}_{\boldsymbol{G}^z} &= \operatorname{span}\left\{H_{z,z}^{t_{(r)},\tau_{(r)}} \log p(\boldsymbol{z} \mid \boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)})\right\}_{r=1}^{||\boldsymbol{G}^z||_0} \\
&= \operatorname{span}\left\{D\boldsymbol{s}(\boldsymbol{z})^\top D_z^{\tau_{(r)}} \boldsymbol{\lambda}(\boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)})\right\}_{r=1}^{||\boldsymbol{G}^z||_0} \\
&= D\boldsymbol{s}(\boldsymbol{z})^\top \operatorname{span}\left\{D_z^{\tau_{(r)}} \boldsymbol{\lambda}(\boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)}) D\boldsymbol{s}(\boldsymbol{z})^{-1}\right\}_{r=1}^{||\boldsymbol{G}^z||_0} D\boldsymbol{s}(\boldsymbol{z}).
\end{aligned}$$

Now recall that, in Assumption 8, we had that $\boldsymbol{z} = \boldsymbol{z}^{\tau_{(r)}}$ for all $r = 1, ..., ||\boldsymbol{G}^z||_0$, which allows us to write

$$\mathbb{R}^{d_z}_{\boldsymbol{G}^z} = D\boldsymbol{s}(\boldsymbol{z})^\top \operatorname{span}\left\{D_z^{\tau_{(r)}} \boldsymbol{\lambda}(\boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)}) D\boldsymbol{s}(\boldsymbol{z}^{\tau_{(r)}})^{-1}\right\}_{r=1}^{||\boldsymbol{G}^z||_0} D\boldsymbol{s}(\boldsymbol{z}),$$

which implies

$$\operatorname{span}\left\{D_z^{\tau_{(r)}} \boldsymbol{\lambda}(\boldsymbol{z}_{(r)}, \boldsymbol{a}_{(r)}) D\boldsymbol{s}(\boldsymbol{z}^{\tau_{(r)}})^{-1}\right\}_{r=1}^{||\boldsymbol{G}^z||_0} = D\boldsymbol{s}(\boldsymbol{z})^{-\top} \mathbb{R}^{d_z}_{\boldsymbol{G}^z} D\boldsymbol{s}(\boldsymbol{z})^{-1} = \mathbb{R}^{d_z}_{\boldsymbol{G}^z},$$

where the last equality holds because $D\boldsymbol{s}(\boldsymbol{z})$ is diagonal and invertible everywhere. ■

Appendix C. Experiments

C.1 Synthetic Datasets

We now provide a detailed description of the synthetic datasets used in the experiments of Section 8.

For all experiments, the dimensionality of $\boldsymbol{x}^t$ is $d_x = 20$ and the ground-truth $\boldsymbol{f}$ is a random neural network with three hidden layers of 20 units with Leaky-ReLU activations with a negative slope of 0.2. The weight matrices are sampled according to a 0-1 Gaussian distribution and, to ensure that $\boldsymbol{f}$ is injective, as assumed in all theorems of this paper, we orthogonalize its columns. Inspired by typical weight initialization in NN (Glorot and Bengio, 2010), we rescale the weight matrices by $\sqrt{\frac{2}{1+0.2^2}}\sqrt{\frac{2}{d_{in}+d_{out}}}$. The standard deviation of the Gaussian noise added to $\boldsymbol{f}(\boldsymbol{z}^t)$ is set to $\sigma = 10^{-2}$ throughout. Since the goal of the experiments is to validate our identifiability results, which assume infinite data, all datasets considered here are very large: 1 million examples.

We now present the different choices of ground-truth $p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ we explored in our experiments. In all cases considered (except for the experiment with $k = 2$ of Table 3), it is a Gaussian with covariance $0.0001I$ independent of $(\boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ and a mean given by some function $\mu(\boldsymbol{z}^{t-1}, \boldsymbol{a}^{t-1})$. Notice that we hence are in the case where $k = 1$ with monotonic sufficient statistics, which is not covered by the theory of Khemakhem et al. (2020a). Throughout, we set $d_z = 10$ and, unless explicitly specified otherwise, we set $d_a = 10$. In all *Time* datasets, sequences have length $T = 2$. In *Action* datasets, the value of T has no consequence since we assume there is no time dependence.

C.1.1 DATASETS SATISFYING THE GRAPHICAL CRITERION

The datasets of this section satisfy the graphical criterion of Section 3.7. This means our theory predicts complete disentanglement (Definition 7). Unless specified otherwise, all datasets satisfy their respective sufficient influence assumptions (Section 3.8). These can be checked using Remark 5 combined with standard facts about independence of the sine and cosine functions.

ActionDiag (Figure 5). In this dataset, $d_a = d_x$ and the connectivity matrix between $\boldsymbol{a}^{t-1}$ and $\boldsymbol{z}^t$ is diagonal, which trivially implies that the graphical criterion of Section 3.7 is satisfied. The mean function is given by

$$\mu(\boldsymbol{z}^{t-1}, \boldsymbol{a}^{t-1}) := \sin(\boldsymbol{a}^{t-1}) \,,$$

where $\sin$ is applied element-wise. Moreover, the components of the action vector $\boldsymbol{a}^{t-1}$ are sampled independently and uniformly between -2 and 2. The same sampling scheme is used for all following datasets. One can check that the sufficient influence assumption (Assumption 6) holds.

ActionNonDiag (Figure 5). We consider a case where the graphical criterion of Section 3.7 is satisfied non-trivially. Let

$$\boldsymbol{G}^a := \begin{pmatrix} 1 & & & & 1 \\ 1 & 1 & & & \\ & 1 & \ddots & & \\ & & \ddots & 1 & \\ & & & 1 & 1 \end{pmatrix} \tag{72}$$

be the adjacency matrix between $\boldsymbol{a}^{t-1}$ and $\boldsymbol{z}^t$. The ith row, denoted by $\boldsymbol{G}^a_i$, corresponds to the parents of $\boldsymbol{z}^t_i$ in $\boldsymbol{a}^{t-1}$. Note that it is analogous to the graph shown in Figure 4, which satisfies the

graphical criterion. The mean function is given by

$$\mu(\boldsymbol{z}^{t-1}, \boldsymbol{a}^{t-1}) := \begin{bmatrix} \boldsymbol{G}_1^{\boldsymbol{a}} \cdot \sin(\frac{3}{\pi}\boldsymbol{a}^{t-1}) \\ \boldsymbol{G}_2^{\boldsymbol{a}} \cdot \sin(\frac{4}{\pi}\boldsymbol{a}^{t-1} + 1) \\ \vdots \\ \boldsymbol{G}_{d_z}^{\boldsymbol{a}} \cdot \sin(\frac{d_z+2}{\pi}\boldsymbol{a}^{t-1} + d_z - 1) \end{bmatrix} . \tag{73}$$

One can check that the sufficient influence assumption (Assumption 6) holds, due to the independence of sines with different frequencies.

ActionNonDiag$_{\text{NoSuffInf}}$ (Table 3). This dataset has the same ground truth adjacency matrix as the above dataset (72), but a different transition function which does not satisfy the assumption of sufficient influence (Section 6). We sampled a matrix $\boldsymbol{W}$ with independent Normal 0-1 entries. The mean function is thus

$$\mu(\boldsymbol{z}^{t-1}, \boldsymbol{a}^{t-1}) := (\boldsymbol{G}^a \odot \boldsymbol{W})\boldsymbol{a}^{t-1} ,$$

where $\odot$ is the Hadamard product (a.k.a. element-wise product).

ActionNonDiag$_{k=2}$ (Table 3). This dataset has the "double diagonal" adjacency matrix of (72) and the same mean function of (73), but the variance of $\boldsymbol{z}^t$ (we assume diagonal covariance) depends on $\boldsymbol{a}^{t-1}$ via

$$\sigma^2(\boldsymbol{z}^{t-1}, \boldsymbol{a}^{t-1}) := \frac{1}{10 d_a} \begin{bmatrix} \exp\left(\boldsymbol{G}_1^{\boldsymbol{a}} \cdot \cos(\frac{3}{\pi}\boldsymbol{a}^{t-1})\right) \\ \exp(\boldsymbol{G}_2^{\boldsymbol{a}} \cdot \cos(\frac{4}{\pi}\boldsymbol{a}^{t-1} + 1)) \\ \vdots \\ \exp(\boldsymbol{G}_{d_z}^{\boldsymbol{a}} \cdot \cos(\frac{d_z+2}{\pi}\boldsymbol{a}^{t-1} + d_z - 1)) \end{bmatrix} . \tag{74}$$

TimeDiag (Figure 5). In this dataset, each $\boldsymbol{z}_i^t$ has only $\boldsymbol{z}_i^{t-1}$ as parent. This trivially satisfies the graphical criterion of Section 3.7. The mean function is given by

$$\mu(\boldsymbol{z}^{t-1}, \boldsymbol{a}^{t-1}) := \boldsymbol{z}^{t-1} + 0.5 \sin(\boldsymbol{z}^{t-1}) ,$$

where the $\sin$ function is applied element-wise. Notice that no auxiliary variables are required. One can check that the sufficient variability assumption (Assumption 10) and sufficient influence assumption (Assumption 11) of Theorem 5 (exponential family) holds.

TimeNonDiag (Figure 5). We consider a case where the graphical criterion of Section 3.7 is satisfied non-trivially. Let

$$\boldsymbol{G}^z := \begin{pmatrix} 1 & & & & \\ 1 & 1 & & & \\ \vdots & & \ddots & & \\ 1 & & & 1 & \\ 1 & 1 & \dots & 1 & 1 \end{pmatrix} \tag{75}$$

be the adjacency matrix between $\boldsymbol{z}^t$ and $\boldsymbol{z}^{t-1}$. The ith row of $\boldsymbol{G}^z$, denoted by $\boldsymbol{G}_i^z$, corresponds to the parents of $\boldsymbol{z}_i^t$. Notice that this connectivity matrix has no 2-cycles and all self-loops are present.

Thus, by Proposition 8, it satisfies the graphical criterion of Section 3.7. The mean function in this case is given by

$$\mu(\boldsymbol{z}^{t-1}, \boldsymbol{a}^{t-1}) := \boldsymbol{z}^{t-1} + 0.5 \begin{bmatrix} \boldsymbol{G}_1^{\boldsymbol{z}} \cdot \sin(\frac{3}{\pi}\boldsymbol{z}^{t-1}) \\ \boldsymbol{G}_2^{\boldsymbol{z}} \cdot \sin(\frac{4}{\pi}\boldsymbol{z}^{t-1} + 1) \\ \vdots \\ \boldsymbol{G}_{d_z}^{\boldsymbol{z}} \cdot \sin(\frac{d_z+2}{\pi}\boldsymbol{z}^{t-1} + d_z - 1) \end{bmatrix}, \tag{76}$$

which is analogous to (73). One can verify that this transition model satisfies the sufficient variability assumption (Assumption 10) and sufficient influence assumption (Assumption 10) of Theorem 5 (exponential family) holds.

TimeNonDiag$_{\text{NoSuffInf}}$ (Table 3). This dataset has the same ground truth adjacency matrix as in (75), but a different transition function that does not satisfy the assumption of sufficient influence. We sampled a transition matrix W with independent Normal 0-1 entries. The transition function is thus

$$\mu(\boldsymbol{z}^{t-1}, \boldsymbol{a}^{t-1}) := \boldsymbol{z}^{t-1} + 0.5(\boldsymbol{G}^z \odot W)\boldsymbol{z}^{t-1}\,.$$

TimeNonDiag$_{k=2}$ (Table 3). This dataset has the lower triangular adjacency matrix of (75) and the same mean function of (76), but the variance of $\boldsymbol{z}^t$ (we assume diagonal covariance) depends on $\boldsymbol{z}^{t-1}$ via

$$\sigma^2(\boldsymbol{z}^{t-1}, \boldsymbol{a}^{t-1}) := \frac{1}{10 d_z} \begin{bmatrix} \exp\left(\boldsymbol{G}_1^{\boldsymbol{z}} \cdot \cos(\frac{3}{\pi}\boldsymbol{z}^{t-1})\right) \\ \exp(\boldsymbol{G}_2^{\boldsymbol{z}} \cdot \cos(\frac{4}{\pi}\boldsymbol{z}^{t-1} + 1)) \\ \vdots \\ \exp(\boldsymbol{G}_{d_z}^{\boldsymbol{z}} \cdot \cos(\frac{d_z+2}{\pi}\boldsymbol{z}^{t-1} + d_z - 1)) \end{bmatrix}. \tag{77}$$

C.1.2 DATASETS THAT VIOLATE THE GRAPHICAL CRITERION

The transition mechanisms for the action and temporal datasets of Figure 7 and Table 4 are (73) and (76), respectively, except for the graphs which are different.

ActionBlockDiag and ActionBlockNonDiag (Figure 7). The left graph corresponds to ActionBlockDiag while the right one corresponds to ActionBlockNonDiag.

$$\boldsymbol{G}^a_{(1)} := \begin{bmatrix} 1 & & & & & \\ 1 & & & & & \\ & 1 & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & 1 & & & \\ & & & 1 & & \\ & & & 1 & & \\ & & & & 1 & \\ & & & & 1 & \end{bmatrix} \quad \boldsymbol{G}^a_{(2)} := \begin{bmatrix} 1 & & & & & \\ 1 & & & & & \\ & 1 & & & & \\ & 1 & & & & \\ & & 1 & & & \\ & & 1 & & & \\ & & & 1 & & \\ & & & 1 & & \\ 1 & & & & 1 & \\ 1 & & & & 1 & \end{bmatrix}$$

TimeBlockDiag and TimeBlockNonDiag (Figure 7). The left graph corresponds to TimeBlockDiag while the right one corresponds to TimeBlockNonDiag.

$$\boldsymbol{G}^z_{(1)} := \begin{bmatrix} 1 & 1 & & & & & & & & \\ 1 & 1 & & & & & & & & \\ & & 1 & 1 & & & & & & \\ & & 1 & 1 & & & & & & \\ & & & & 1 & 1 & & & & \\ & & & & 1 & 1 & & & & \\ & & & & & & 1 & 1 & & \\ & & & & & & 1 & 1 & & \\ & & & & & & & & 1 & 1 \\ & & & & & & & & 1 & 1 \end{bmatrix} \quad \boldsymbol{G}^z_{(2)} := \begin{bmatrix} 1 & 1 & & & & & & & 1 & 1 \\ 1 & 1 & & & & & & & 1 & 1 \\ & & 1 & 1 & & & & & & \\ & & 1 & 1 & & & & & & \\ & & & & 1 & 1 & & & & \\ & & & & 1 & 1 & & & & \\ & & & & & & 1 & 1 & & \\ & & & & & & 1 & 1 & & \\ 1 & 1 & & & 1 & 1 & & & 1 & 1 \\ 1 & 1 & & & 1 & 1 & & & 1 & 1 \end{bmatrix}$$

ActionRandomGraphs and TimeRandomGraphs (Table 4). The transition mechanisms are the same as in the ActionNonDiag and TimeNonDiag datasets, i.e. they are given by (73) & (76), respectively. However, the graphs are sampled randomly, with various levels of sparsity. For the ActionRandomGraphs dataset, we have $\boldsymbol{G}^a_{i,j} \sim \text{Ber}(p)$ and independent. For the TimeRandomGraphs datasets, it is the same except for the diagonal elements, which are forced to be active, i.e. $\boldsymbol{G}^z_{i,i} = 1$.

C.2 Implementation Details of our Constrained VAE Method

All details of our implementation match those of Lachapelle et al. (2022) (except for the constrained optimization introduced in Section 5).

Learned mechanisms. Every coordinate $\boldsymbol{z}_i$ of the latent vector has its own mechanism $\hat{p}(\boldsymbol{z}^t_i \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})$ that is Gaussian with mean outputted by $\hat{\boldsymbol{\mu}}_i(\boldsymbol{z}^{t-1}, \boldsymbol{a}^{t-1})$ (a multilayer perceptron with 5 layers of 512 units) and a learned variance which does not depend on the previous time steps. For learning, we use the typical parameterization of the Gaussian distribution with μ and σ^2 and not its exponential family parameterization. Throughout, the dimensionality of $\boldsymbol{z}^t$ in the learned model always matches the dimensionality of the ground-truth (same for the baselines). Learning the dimensionality of $\boldsymbol{z}^t$ is left for future work.

Prior of $\boldsymbol{z}^1$ in time-sparsity experiments. In *time-sparsity* experiments, the prior of the first latent $\hat{p}(\boldsymbol{z}^1)$ (when $t = 1$) is modeled separately as a Gaussian with learned mean and learned diagonal covariance. Note that this learned covariance at time $t = 1$ is different from the subsequent learned conditional covariance at time $t > 1$.

Encoder/Decoder. In all experiments, including baselines, both the encoder and the decoder is modeled by a neural network with 6 fully connected hidden layers of 512 units with LeakyReLU activation with negative slope 0.2. For all VAE-based methods, the encoder outputs the mean and a diagonal covariance. Moreover, $p(\boldsymbol{x}|\boldsymbol{z})$ has a *learned* isotropic covariance $\sigma^2 I$. Note that $\sigma^2 I$ corresponds to the covariance of the independent noise $\boldsymbol{n}^t$ in the equation $\boldsymbol{x}^t = \boldsymbol{f}(\boldsymbol{z}^t) + \boldsymbol{n}^t$.

C.3 Baselines

In synthetic experiments of Sec. 8, all methods used a minibatch size of 1024 and the same encoder and decoder architecture: A MLP with 6 layers of 512 units with LeakyReLU activations (negative slope of 0.2). We manually tuned the learning rate of each method to ensure proper convergence. For

VAE-based methods, i.e. TCVAE, SlowVAE and iVAE, we are always choosing $p(x|z)$ Gaussian with a covariance $\sigma^2 I$ and learn σ^2.

β-TCVAE. We used the implementation provided in the original paper by Chen et al. (2018), which is available at https://github.com/rtqichen/beta-tcvae. We used a learning rate of 1e-4.

iVAE. We used the implementation available at https://github.com/ilkhem/icebeem from Khemakhem et al. (2020a). In it, the mean of the prior $p(z|a)$ is fixed to zero, while its diagonal covariance is allowed to depend on a through an MLP. We change this to allow the mean to also depend on a through the neural network (with 5 layers and width 512). We also lower bounded its variance as well as the variance of $q(z \mid x, a)$ to improve the stability of learning. In the original implementation, the covariance of $p(x|z)$ was not learned. We found that learning it (analogously to what we do in our method) improved performance. We used a learning rate of 1e-4.

SlowVAE. We used the implementation provided in https://github.com/bethgelab/slow_disentanglement (Klindt et al., 2021). Like for other VAE-based methods, we modeled $p(x|z)$ as a Gaussian with covariance $\sigma^2 I$ and learned σ^2.

PCL. We used the implementation provided here: https://github.com/bethgelab/slow_disentanglement/tree/baselines. PCL (Hyvarinen and Morioka, 2017) stands for "permutation contrastive learning" and works as follows: Given sequential data $\{\boldsymbol{x}^t\}_{t=1}^T$, PCL trains a regression function $r((x', x))$ to discriminate between pairs of adjacent observations (positive pairs) and randomly matched pairs (negative pairs). The regression function has the form

$$r((x, x')) = \sum_{i=1}^{d_z} B_i(h_i(x), h_i(x')) \,,$$

where $h : \mathbb{R}^{d_x} \to \mathbb{R}^{d_z}$ is the encoder and $B_i : \mathbb{R}^2 \to \mathbb{R}$ are learned functions. In our implementation, the functions B_i are fully connected neural networks with 5 layers and 512 hidden units. We experimented with the less expressive function suggested in the original work, but found that the extra capacity improved performance across all datasets we considered.

C.4 Unsupervised Hyperparameter Selection

In practice, MCC cannot be measured since the ground-truth latent variables are not observed. Unlike in standard machine learning settings, hyperparameter selection for disentanglement cannot be performed simply by evaluating goodness of fit on a validation set and selecting the highest scoring model, since there is usually a trade-off between goodness of fit and disentanglement (Locatello et al., 2019, Sec. 5.4). To avoid this problem, Duan et al. (2020) introduced *an unsupervised disentanglement ranking* (UDR), which, for all combinations of hyperparameters, measures how consistent are different random initializations of the algorithm. The authors argue that hyperparameters yielding disentangled representation typically yields consistent representations. In our experiments, the consistency of a given hyperparameter combination is measured as follows: for every pair of models, we compute the MCC between their representations. Then, we report the median of all pairwise MCC. This gives a UDR score for every hyperparameter values considered. Figure 9 reports the ELBO (normalized between zero and one), the MCC and the UDR score for the experiments of Figure 5. We can visualize the trade-off between ELBO and MCC. That being said, MCC and UDR correlate nicely except for the TimeNonDiag dataset, in which this correlation breaks for stronger regularization. We noticed that these specific runs correspond to excessively

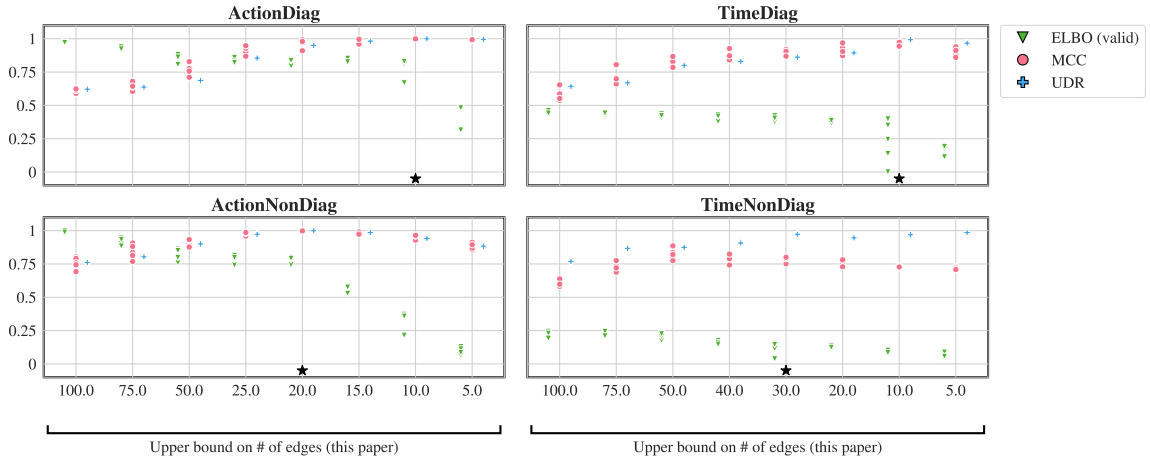


Figure 9: Investigating the link between goodness of fit (ELBO), disentanglement (MCC) and UDR. The ELBO is normalized so that it remains between 0 and 1.

sparse graph, with fewer than 10 edges (out of 100 possible edges). The black star indicates the hyperparameter selected by UDR when excluding coefficient values which yields graphs with less than 10 edges (on average).

Baselines. Two of the baselines considered had hyperparameters to tune, SlowVAE (Klindt et al., 2021) and TCVAE (Chen et al., 2018). For SlowVAE, we did a grid search on the following values, $\gamma \in \{1.0, 2.0, 4.0, 8.0, 16.0\}$ and $\alpha \in \{1, 3, 6, 10\}$. For TCVAE, we explored $\beta \in \{1, 2, 3, 4, 5\}$ but the optimal value in terms of disentanglement was almost always 1. Values of β larger than 5 led to instabilities during training. The hyperparameters were selected using UDR, as described in the paragraph above.

Appendix D. Miscellaneous

D.1 On the Invertibility of the Mixing Function

Throughout this work as well as many others (Hyvarinen and Morioka, 2016, 2017; Hyvärinen et al., 2019; Khemakhem et al., 2020a; Locatello et al., 2020; Klindt et al., 2021), it is assumed that the mixing function mapping the latent factors to the observation is a diffeomorphism onto its image. In this section, we briefly discuss the practical implications of this assumption.

Recall that a diffeomorphism is a differentiable bijective function with a differentiable inverse. We start by addressing the bijective part of the assumption. To understand it, we consider a plausible situation where the mapping $\boldsymbol{f}$ is not invertible. Consider the minimal example of Fig. 1 consisting of a tree, a robot and a ball. Assume that the ball can be hidden behind either the tree or the robot. Then, the mixing function $\boldsymbol{f}$ is not invertible because, given only the image, it is impossible to know whether the ball is behind the tree or the robot. Thus, this situation is not covered by our theory. Intuitively, one could infer, at least approximately, where the ball is hidden based on previous time frames. Allowing for this form of occlusion is left as future work. See also Mansouri et al. (2022) for further discussion about how one can relax this assumption.

We believe the differentiable part of this assumption is only a technicality that could probably be relaxed to being piecewise differentiable. Our experiments were performed with data generated

with a piecewise linear $\boldsymbol{f}$, which is not differentiable only on a set of (Lebesgue) measure zero, but this was not an issue in practice.

D.2 Contrasting with the Assumptions of Khemakhem et al. (2020a) & Yao et al. (2022b)

In this section, we discuss two identifiability results previously proposed in the literature that do not leverage sparsity (Khemakhem et al., 2020a; Yao et al., 2022a). We show that these results do not apply to the simple homoscedastic Gaussian latent models of the form $p(\boldsymbol{z}_i^t \mid \boldsymbol{z}^{t-1}) = \mathcal{N}(\boldsymbol{z}_i^t \mid \boldsymbol{\mu}_i(\boldsymbol{z}^{t-1}), \sigma_i^2)$, contrary to our theory, as we saw in Examples 9, 10 and 12. We will see that in the context of a Gaussian latent model, both results require the variance to vary sufficiently strongly. We believe that such a requirement is not well suited for nearly deterministic environments such as the one depicted in Figure 1.

Khemakhem et al. (2020a). The most significant distinction between the theory of Khemakhem et al. (2020a) (iVAE) and ours is how identifiability up to permutation is obtained: Theorems 2 & 3 from iVAE shows that if the assumptions of their Theorem 1 (which is essentially Theorem 4) are satisfied and $\boldsymbol{s}_i$ has dimension $k > 1$ **or** is non-monotonic, then the model is not just identifiable up to linear transformation but up to permutations (and rescalings). In contrast, our theory covers the case where $k = 1$ and $\boldsymbol{s}_i$ is monotonic, like in the homoscedastic Gaussian case. Interestingly, Khemakhem et al. (2020a) mentioned this specific case as a counterexample to their theory in their Proposition 3. The extra power of our theory comes from the extra *structure* in the dependencies of the latent factors coupled with sparsity regularization. We note that, assuming the latent factors are Gaussian, the variability assumption of Theorem 4 combined with $k > 1$ requires the variance to vary sufficiently, which is implausible in the nearly deterministic environment of Figure 1.

Yao et al. (2022a). This work (Theorem 1) requires that, for each value of $\boldsymbol{z}^t$, the $2d_z$ functions

$$\frac{\partial^2}{\partial \boldsymbol{z}_i^t \partial \boldsymbol{z}^{t-1}} \log p(\boldsymbol{z}_i^t \mid \boldsymbol{z}^{t-1}) \text{ and } \frac{\partial^3}{(\partial \boldsymbol{z}_i^t)^2 \partial \boldsymbol{z}^{t-1}} \log p(\boldsymbol{z}_i^t \mid \boldsymbol{z}^{t-1}) \text{ for } i = 1 \dots d_z \,,$$

seen as functions from $\mathbb{R}^{d_z}$ to $\mathbb{R}^{d_z}$ are linearly independent. Indeed, if $p(\boldsymbol{z}_i^t \mid \boldsymbol{z}^{t-1}) = \mathcal{N}(\boldsymbol{z}_i^t \mid \boldsymbol{\mu}_i(\boldsymbol{z}^{t-1}), \sigma_i^2)$, one can easily derive that

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{z}_i^t} \log p(\boldsymbol{z}_i^t \mid \boldsymbol{z}_i^{t-1}) &= -(\boldsymbol{z}_i - \boldsymbol{\mu}_i(\boldsymbol{z}^{t-1}))/\sigma_i^2 \\
\frac{\partial^2}{(\partial \boldsymbol{z}_i^t)^2} \log p(\boldsymbol{z}_i^t \mid \boldsymbol{z}_i^{t-1}) &= -1/\sigma_i^2 \\
\frac{\partial^2}{(\partial \boldsymbol{z}_i^t)^2 \partial \boldsymbol{z}^{t-1}} \log p(\boldsymbol{z}_i^t \mid \boldsymbol{z}_i^{t-1}) &= \mathbf{0} \,,
\end{aligned}$$

which shows that the assumption of Yao et al. (2022a, Theorem 1) does not hold for homoscedastic Gaussian latent models. We further notice that, had the variance σ_i^2 depend on $\boldsymbol{z}^{t-1}$, the identifiability result of Yao et al. (2022b) could have applied.

D.3 Derivation of the ELBO

In this section, we derive the evidence lower bound presented in Section 5.

$$
\begin{aligned}
&\log p(\boldsymbol{x}^{\leq T} \mid \boldsymbol{a}^{<T}) = \\
&\quad \mathbb{E}_{q(\boldsymbol{z}^{\leq T} \mid \boldsymbol{x}^{\leq T}, \boldsymbol{a}^{<T})} \left[\log \frac{q(\boldsymbol{z}^{\leq T} \mid \boldsymbol{x}^{\leq T}, \boldsymbol{a}^{<T})}{p(\boldsymbol{z}^{\leq T} \mid \boldsymbol{x}^{\leq T}, \boldsymbol{a}^{<T})} \right. \qquad (78) \\
&\qquad\qquad \left. + \log \frac{p(\boldsymbol{z}^{\leq T}, \boldsymbol{x}^{\leq T} \mid \boldsymbol{a}^{<T})}{q(\boldsymbol{z}^{\leq T} \mid \boldsymbol{x}^{\leq T}, \boldsymbol{a}^{<T})} \right] \\
&\quad \geq \mathbb{E}_{q(\boldsymbol{z}^{\leq T} \mid \boldsymbol{x}^{\leq T}, \boldsymbol{a}^{<T})} \left[\log \frac{p(\boldsymbol{z}^{\leq T}, \boldsymbol{x}^{\leq T} \mid \boldsymbol{a}^{<T})}{q(\boldsymbol{z}^{\leq T} \mid \boldsymbol{x}^{\leq T}, \boldsymbol{a}^{<T})} \right] \\
&\quad = \mathbb{E}_{q(\boldsymbol{z}^{\leq T} \mid \boldsymbol{x}^{\leq T}, \boldsymbol{a}^{<T})} \left[\log p(\boldsymbol{x}^{\leq T} \mid \boldsymbol{z}^{\leq T}, \boldsymbol{a}^{<T}) \right] \qquad (79) \\
&\qquad - KL(q(\boldsymbol{z}^{\leq T} \mid \boldsymbol{x}^{\leq T}, \boldsymbol{a}^{<T}) || p(\boldsymbol{z}^{\leq T} \mid \boldsymbol{a}^{<T})) \qquad (80)
\end{aligned}
$$

where the inequality holds because the term at (78) is a Kullback-Leibler divergence, which is greater or equal to 0. Notice that

$$
p(\boldsymbol{x}^{\leq T} \mid \boldsymbol{z}^{\leq T}, \boldsymbol{a}^{<T}) = p(\boldsymbol{x}^{\leq T} \mid \boldsymbol{z}^{\leq T}) = \prod_{t=1}^{T} p(\boldsymbol{x}^t \mid \boldsymbol{z}^t) \,. \qquad (81)
$$

Recall that we are considering a variational posterior of the following form:

$$
q(\boldsymbol{z}^{\leq T} \mid \boldsymbol{x}^{\leq T}, \boldsymbol{a}^{<T}) := \prod_{t=1}^{T} q(\boldsymbol{z}^t \mid \boldsymbol{x}^t) \,. \qquad (82)
$$

Equations (81) & (82) allow us to rewrite the term in (79) as

$$
\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\boldsymbol{z}^t \sim q(\cdot \mid \boldsymbol{x}^t)} \left[\log p(\boldsymbol{x}^t \mid \boldsymbol{z}^t) \right]
$$

Notice further that

$$
p(\boldsymbol{z}^{\leq T} \mid \boldsymbol{a}^{<T}) = \prod_{t=1}^{T} p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}) \,. \qquad (83)
$$

Using (82) & (83), the KL term (80) can be broken down as a sum of KL as:

$$
\sum_{t=1}^{T} \mathop{\mathbb{E}}_{\boldsymbol{z}^{<t} \sim q(\cdot \mid \boldsymbol{x}^{<t})} KL(q(\boldsymbol{z}^t \mid \boldsymbol{x}^t) || p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t}))
$$

Putting all together yields the desired ELBO:

$$
\begin{aligned}
\log p(\boldsymbol{x}^{\leq T} | \boldsymbol{a}^{<T}) \geq & \sum_{t=1}^{T} \mathop{\mathbb{E}}_{\boldsymbol{z}^t \sim q(\cdot \mid \boldsymbol{x}^t)} \left[\log p(\boldsymbol{x}^t \mid \boldsymbol{z}^t) \right] \qquad (84) \\
- & \mathop{\mathbb{E}}_{\boldsymbol{z}^{<t} \sim q(\cdot \mid \boldsymbol{x}^{<t})} KL(q(\boldsymbol{z}^t \mid \boldsymbol{x}^t) || p(\boldsymbol{z}^t \mid \boldsymbol{z}^{<t}, \boldsymbol{a}^{<t})) \,.
\end{aligned}
$$

Appendix E. Author Contributions

E.1 Contributions to the Extended Version

Sébastien Lachapelle developed the idea, the theory and proofs behind mechanism sparsity regularization for disentanglement, wrote the crux of the paper, developed the regularized VAE-based method, and performed most of the experiments. **Rémi Le Priol** provided valuable feedback on the clarity of the manuscript. **Simon Lacoste-Julien** helped with the overall presentation of the paper, clarified the conceptual framework and the motivation, and provided supervision.

E.2 Contributions to the CLeaR Version (Lachapelle et al., 2022)

Sébastien Lachapelle developed the idea, the theory and proofs behind mechanism sparsity regularization for disentanglement, wrote the first draft of the paper, and designed and implemented the regularized VAE-based method. **Pau Rodríguez López** ran all experiments appearing in the paper, produced associated figures, and ran experiments with image data that are still work in progress. **Yash Sharma** contributed to the research process, the experimental design in particular, implemented and ran experiments on image data that did not make it in the final version, and contributed to the writing and the literature review. **Katie Everett** implemented and ran experiments on image data that did not make it in the final version and contributed to the writing and figures. **Rémi Le Priol** reviewed the proofs of main theorems, simplified some arguments and the general presentation of the proof, and contributed to the writing and figures. **Alexandre Lacoste** produced image datasets that did not make it into the final version and provided supervision. **Simon Lacoste-Julien** helped with the overall presentation of the paper, clarified the conceptual framework and the motivation, and provided supervision.

References

K. Ahuja, J. Hartford, and Y. Bengio. Properties from mechanisms: an equivariance perspective on identifiable representation learning. In *International Conference on Learning Representations*, 2022a.

K. Ahuja, J. Hartford, and Y. Bengio. Weakly supervised representation learning with sparse perturbations. *arXiv preprint arXiv:2206.01101*, 2022b.

K. Ahuja, D. Mahajan, V. Syrgkanis, and I. Mitliagkas. Towards efficient representation identification in supervised learning. In *Conference on Causal Learning and Reasoning*, 2022c.

K. Ahuja, D. Mahajan, Y. Wang, and Y. Bengio. Interventional causal representation learning. In *International Conference on Machine Learning*, 2023.

Elias Bareinboim, Juan Correa, Duligur Ibeling, and Thomas Icard. On Pearl's hierarchy and the foundations of causal inference (1st edition). In *Probabilistic and Causal Inference: the Works of Judea Pearl*. ACM Books, 2022.

Y. Bengio. The consciousness prior. *arXiv preprint arXiv:1709.08568*, 2019.

Y. Bengio, A. Courville, and P. Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 2013.

Y. Bengio, T. Deleu, N. Rahaman, N. R. Ke, S. Lachapelle, O. Bilaniuk, A. Goyal, and C. Pal. A meta-transfer objective for learning to disentangle causal mechanisms. In *International Conference on Learning Representations*, 2020.

M. Bereket and T. Karaletsos. Modelling cellular perturbations with the sparse additive mechanism shift variational autoencoder. In *Advances in Neural Information Processing Systems*, 2023.

S. Bing, U. Ninad, J. Wahl, and J. Runge. Identifying linearly-mixed causal representations from multi-node interventions. *arXiv preprint arXiv:2311.02695*, 2023.

J. Brady, R. S. Zimmermann, Y. Sharma, B. Schölkopf, J. von Kügelgen, and W. Brendel. Provably learning object-centric representations. In *International Conference on Machine Learning*, 2023.

J. Brehmer, P. de Haan, P. Lippe, and T. Cohen. Weakly supervised causal representation learning. In *Advances in Neural Information Processing Systems*, 2022.

P. Brouillard, S. Lachapelle, A. Lacoste, S. Lacoste-Julien, and A. Drouin. Differentiable causal discovery from interventional data. In *Advances in Neural Information Processing Systems*, 2020.

S. Buchholz, M. Besserve, and B. Schölkopf. Function classes for identifiable nonlinear independent component analysis. In *Advances in Neural Information Processing Systems*, 2022.

S. Buchholz, G. Rajendran, E. Rosenfeld, B. Aragam, B. Schölkopf, and P. K. Ravikumar. Learning linear causal representations from interventions under general nonlinear mixing. In *Advances Neural Information Processing Systems*, 2023.

R. T. Q. Chen, X. Li, R. G., and D. Duvenaud. Isolating sources of disentanglement in vaes. In *Advances in Neural Information Processing Systems*, 2018.

D. M. Chickering. Optimal structure identification with greedy search. In *Journal of Machine Learning Research*, 2003.

P. Comon. Independent component analysis, a new concept? *Signal Processing*, 1994.

I. Daunhawer, A. Bizeul, E. Palumbo, A. Marx, and J. E Vogt. Identifiability results for multimodal contrastive learning. In *International Conference on Learning Representations*, 2023.

S. Duan, L. Matthey, A. Saraiva, N. Watters, C. Burgess, A. Lerchner, and I. Higgins. Unsupervised model selection for variational disentangled representation learning. In *International Conference on Learning Representations*, 2020.

D. Eaton and K. Murphy. Exact bayesian structure learning from uncertain interventions. In *International Conference on Artificial Intelligence and Statistics*, 2007.

M. Fumero, F. Wenzel, L. Zancato, A. Achille, E. Rodolà, S. Soatto, B. Schölkopf, and F. Locatello. Leveraging sparse and shared feature activations for disentangled representation learning. In *Advances in Neural Information Processing Systems*, 2023.

J. Gallego-Posada and J. Ramirez. Cooper: a toolkit for lagrangian-based constrained optimization. https://github.com/cooper-org/cooper, 2022.

J. Gallego-Posada, J. Ramirez De Los Rios, and A. Erraqabi. Flexible learning of sparse neural networks via constrained l_0 regularization. In *NeurIPS Workshop LatinX in AI*, 2021.

L. Girin, S. Leglaive, X. Bie, J. Diard, T. Hueber, and X. Alameda-Pineda. Dynamical variational autoencoders: A comprehensive review. *arXiv preprint arXiv:2008.12595*, 2020.

X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *International Conference on Artificial Intelligence and Statistics*, 2010.

A. Goyal and Y. Bengio. Inductive biases for deep learning of higher-level cognition. *arXiv preprint arXiv:2011.15091*, 2021.

A. Goyal, A. R. Didolkar, N. R. Ke, C. Blundell, P. Beaudoin, N. Heess, M. C. Mozer, and Y. Bengio. Neural production systems. In *Advances in Neural Information Processing Systems*, 2021a.

A. Goyal, A. Lamb, J Hoffmann, S. Sodhani, S. Levine, Y. Bengio, and B. Schölkopf. Recurrent independent mechanisms. In *International Conference on Learning Representations*, 2021b.

L. Gresele, P. K. Rubenstein, A. Mehrjou, F. Locatello, and B. Schölkopf. The incomplete rosetta stone problem: Identifiability results for multi-view nonlinear ica. In *Conference on Uncertainty in Artificial Intelligence Conference*, 2020.

L. Gresele, J. Von Kügelgen, V. Stimper, B. Schölkopf, and M. Besserve. Independent mechanism analysis, a new concept? In *Advances in Neural Information Processing Systems*, 2021.

M. U. Gutmann and A. Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research*, 2012.

H. Hälvä and A. Hyvärinen. Hidden markov nonlinear ica: Unsupervised learning from nonstationary time series. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2020.

H. Hälvä, S. Le Corff, L. Lehéricy, J. So, Y. Zhu, E. Gassiat, and A. Hyvarinen. Disentangling identifiable features from noisy data with structured nonlinear ICA. In *Advances in Neural Information Processing Systems*, 2021.

A. Hauser and P. Bühlmann. Characterization and greedy learning of interventional markov equivalence classes of directed acyclic graphs. *Journal of Machine Learning Research*, 2012.

I. Higgins, L. Matthey, A. Pal, C. P. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

D. Horan, E. Richardson, and Y. Weiss. When is unsupervised disentanglement possible? In *Advances in Neural Information Processing Systems*, 2021.

A. Hyvarinen and H. Morioka. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. In *Advances in Neural Information Processing Systems*, 2016.

A. Hyvarinen and H. Morioka. Nonlinear ICA of Temporally Dependent Stationary Sources. In *International Conference on Artificial Intelligence and Statistics*, 2017.

A. Hyvärinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 1999.

A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. Wiley, 2001.

A. Hyvärinen, H. Sasaki, and R. E. Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *International Conference on Artificial Intelligence and Statistics*, 2019.

A. Hyvärinen, I. Khemakhem, and H. Morioka. Nonlinear independent component analysis for principled disentanglement in unsupervised deep learning. *arXiv preprint arXiv:2303.16535*, 2023.

A. Jaber, M. Kocaoglu, K. Shanmugam, and E. Bareinboim. Causal discovery from soft interventions with unknown targets: Characterization and learning. In *Advances in Neural Information Processing Systems*, 2020.

E. Jang, S. Gu, and B. Poole. Categorical reparameterization with gumbel-softmax. *International Conference on Machine Learning*, 2017.

Y. Jiang and B. Aragam. Learning nonparametric latent causal graphs with unknown interventions. In *Advances in Neural Information Processing Systems*, 2023.

C. Jutten and J. Herault. Blind separation of sources, part 1: An adaptive algorithm based on neuromimetic architecture. *Signal Process.*, 1991.

T. Karaletsos, S. Belongie, and G. Rätsch. Bayesian representation learning with oracle constraints. In *International Conference on Learning Representations*, 2016.

N. R. Ke, O. Bilaniuk, A. Goyal, S. Bauer, H. Larochelle, C. Pal, and Y. Bengio. Learning neural causal models from unknown interventions. *arXiv preprint arXiv:1910.01075*, 2019.

N. R. Ke, A. R. Didolkar, S. Mittal, A. Goyal, G. Lajoie, S. Bauer, D. J. Rezende, M. C. Mozer, Y. Bengio, and C. Pal. Systematic evaluation of causal discovery in visual model based reinforcement learning. *arXiv preprint arXiv:2107.00848*, 2021.

H. Keurti, H.-R. Pan, M. Besserve, B. F. Grewe, and B. Schölkopf. Homomorphism autoencoder – learning group structured representations from observed transitions. In *International Conference on Machine Learning*, 2023.

I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, 2020a.

I. Khemakhem, R. Monti, D. Kingma, and A. Hyvarinen. Ice-beem: Identifiable conditional energy-based deep models based on nonlinear ica. In *Advances in Neural Information Processing Systems*, 2020b.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

B. Kivva, G. Rajendran, P. K. Ravikumar, and B. Aragam. Identifiability of deep generative models without auxiliary information. In *Advances in Neural Information Processing Systems*, 2022.

D. A. Klindt, L. Schott, Y Sharma, I Ustyuzhaninov, W. Brendel, M. Bethge, and D. M. Paiton. Towards nonlinear disentanglement in natural data with temporal sparse coding. In *International Conference on Learning Representations*, 2021.

M. Kocaoglu, C. Snyder, A. G. Dimakis, and S. Vishwanath. CausalGAN: Learning causal implicit generative models with adversarial training. In *International Conference on Learning Representations*, 2018.

S. Lachapelle, Rodriguez Lopez, P., Y. Sharma, K. E. Everett, R. Le Priol, A. Lacoste, and S. Lacoste-Julien. Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA. In *Conference on Causal Learning and Reasoning*, 2022.

S. Lachapelle, T. Deleu, D. Mahajan, I. Mitliagkas, Y. Bengio, S. Lacoste-Julien, and Q. Bertrand. Synergies between disentanglement and sparsity: Generalization and identifiability in multi-task learning. In *International Conference on Machine Learning*, 2023a.

S. Lachapelle, D. Mahajan, I. Mitliagkas, and S. Lacoste-Julien. Additive decoders for latent variables identification and cartesian-product extrapolation. In *Advances in Neural Information Processing Systems*, 2023b.

T. Leemann, M. Kirchhof, Y. Rong, E. Kasneci, and G. Kasneci. When are post-hoc conceptual explanations identifiable? *Conference on Uncertainty in Artificial Intelligence*, 2023.

A. Lei, B. Schölkopf, and I. Posner. Variational causal dynamics: Discovering modular world models from interventions. *Transactions on Machine Learning Research*, 2023.

W. Liang, A. Kekić, J. von Kügelgen, S. Buchholz, M. Besserve, L. Gresele, and B. Schölkopf. Causal component analysis. In *Advances in Neural Information Processing Systems*, 2023.

P. Lippe, S. Magliacane, S. Löwe, Y. M. Asano, T. Cohen, and E. Gavves. CITRIS: Causal identifiability from temporal intervened sequences. *arXiv preprint arXiv:2202.03169*, 2022.

P. Lippe, S. Magliacane, S. Löwe, Y. M Asano, T. Cohen, and E. Gavves. BISCUIT: Causal representation learning from binary interactions. In *Conference on Uncertainty in Artificial Intelligence*, 2023a.

P. Lippe, S. Magliacane, S. Löwe, Y. M Asano, T. Cohen, and E. Gavves. iCITRIS: Causal representation learning for instantaneous temporal effects. In *International Conference on Learning Representations*, 2023b.

Y. Liu, Z. Zhang, D. Gong, M. Gong, B. Huang, A. van den Hengel, K. Zhang, and J. Qinfeng Shi. Identifying weight-variant latent causal models. *arXiv preprint arXiv:2208.14153*, 2023.

F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, 2019.

F. Locatello, B. Poole, G. Raetsch, B. Schölkopf, O. Bachem, and M. Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, 2020.

R. Lopez, N. Tagasovska, S. Ra, K. Cho, J. K. Pritchard, and A. Regev. Learning causal representations of single cells via sparse mechanism shift modeling. *Conference on Causal Learning and Reasoning*, 2023.

K. Madan, N. R. Ke, A. Goyal, B. Schölkopf, and Y. Bengio. Fast and slow learning of recurrent independent mechanisms. In *International Conference on Learning Representations*, 2021.

C. J. Maddison, A. Mnih, and Y. W. Teh. The concrete distribution: A continuous relaxation of discrete random variables. *International Conference on Machine Learning*, 2017.

A. Mansouri, J. Hartford, K. Ahuja, and Y. Bengio. Object-centric causal representation learning. In *NeurIPS Workshop on Symmetry and Geometry in Neural Representations*, 2022.

J. M. Mooij, S. Magliacane, and T. Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 2020.

G. Elyse Moran, D. Sridhar, Y. Wang, and D. Blei. Identifiable deep generative models via sparse decoding. *Transactions on Machine Learning Research*, 2022.

H. Morioka and A. Hyvarinen. Connectivity-contrastive learning: Combining causal discovery and representation learning for multimodal data. In *International Conference on Artificial Intelligence and Statistics*, 2023.

H. Morioka, H. Hälvä, and A. Hyvärinen. Independent innovation analysis for nonlinear vector autoregressive process. In *International Conference on Artificial Intelligence and Statistics*, 2021.

J. R. Munkres. *Topology*. Prentice Hall, Inc., 2 edition, 2000.

J.R. Munkres. *Analysis On Manifolds*. Basic Books, 1991.

S. Nair, Y. Zhu, S. Savarese, and L. Fei-Fei. Causal induction from visual observations for goal directed tasks. *arXiv preprint arXiv:1910.01751*, 2019.

R. Pamfil, N. Sriwattanaworachai, S. Desai, P. Pilgerstorfer, P. Beaumont, K. Georgatzis, and B. Aragam. Dynotears: Structure learning from time-series data. *International Conference on Artificial Intelligence and Statistics*, 2020.

J. Pearl. *Causality*. Cambridge university press, 2009.

J. Pearl. The seven tools of causal inference, with reflections on machine learning. *Commun. ACM*, 2019.

J. Peters, J. M. Mooij, D. Janzing, and B. Schölkopf. Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, 2014.

J. Peters, D. Janzing, and B. Schölkopf. *Elements of Causal Inference - Foundations and Learning Algorithms*. MIT Press, 2017.

D. Pollard. *A User's Guide to Measure Theoretic Probability*. Cambridge University Press, 2001.

Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science Advances*, 2019.

A. Schell and H. Oberhauser. Nonlinear independent component analysis for discrete-time and continuous-time signals. *The Annals of Statistics*, 2023.

B. Schölkopf. Causality for machine learning. *arXiv preprint arXiv:1911.10500*, 2019.

B. Schölkopf, F. Locatello, S. Bauer, N. R. Ke, N. Kalchbrenner, A. Goyal, and Y. Bengio. Toward causal representation learning. *IEEE - Advances in Machine Learning and Deep Neural Networks*, 2021.

X. Shen, F. Liu, H. Dong, Q. Lian, Z. Chen, and T. Zhang. Weakly supervised disentangled generative causal representation learning. *Journal of Machine Learning Research*, 2022.

S. Shimizu, P.O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 2006.

P. Sorrenson, C. Rother, and U. Köthe. Disentanglement by nonlinear ica with general incompressible-flow networks (GIN). In *International Conference on Learning Representations*, 2020.

C. Squires, Y. Wang, and C. Uhler. Permutation-based causal structure learning with unknown intervention targets. *Conference on Uncertainty in Artificial Intelligence*, 2020.

C. Squires, A. Seigal, S. Bhate, and C. Uhler. Linear causal disentanglement via interventions. In *International Conference on Machine Learning*, 2023.

A. Taleb and C. Jutten. Source separation in post-nonlinear mixtures. *IEEE Transactions on Signal Processing*, 1999.

V. Thomas, J. Pondard, E. Bengio, M. Sarfati, P. Beaudoin, M.-J. Meurs, J. Pineau, D. Precup, and Y. Bengio. Independently controllable factors. *arXiv preprint arXiv:1708.01289*, 2017.

L. Tong, V.C. Soon, Y.F. Huang, and R. Liu. AMUSE: a new blind identification algorithm. In *IEEE International Symposium on Circuits and Systems*, 1990.

L. Tong, Y. Inouye, and R.-w. Liu. Waveform-preserving blind estimation of multiple independent sources. *IEEE Transactions on Signal Processing*, 1993.

B. Varici, E. Acarturk, K. Shanmugam, A. Kumar, and A. Tajer. Score-based causal representation learning from interventions: Nonparametric identifiability. In *Causal Representation Learning Workshop at NeurIPS 2023*, 2023a.

B. Varici, E. Acartürk, K. Shanmugam, and A. Tajer. General identifiability and achievability for causal representation learning. *arXiv preprint arXiv:2310.15450*, 2023b.

S. Volodin. CauseOccam : Learning interpretable abstract representations in reinforcement learning environments via model sparsity. Master's thesis, École Polytechnique Fédérale de Lausanne, 2021.

J. Von Kügelgen, Y. Sharma, L. Gresele, W. Brendel, B. Schölkopf, M. Besserve, and F. Locatello. Self-supervised learning with data augmentations provably isolates content from style. In *Advances in Neural Information Processing Systems*, 2021.

J. von Kügelgen, M. Besserve, W. Liang, L. Gresele, A. Kekić, E. Bareinboim, D. Blei, and B. Schölkopf. Nonparametric identifiability of causal representations from unknown interventions. In *Advances in Neural Information Processing Systems*, 2023.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 2008.

Q. Xi and B. Bloem-Reddy. Indeterminacy in generative models: Characterization and strong identifiability. In *International Conference on Artificial Intelligence and Statistics*, 2023.

M. Yang, F. Liu, Z. Chen, X. Shen, J. Hao, and J. Wang. CausalVAE: Disentangled representation learning via neural structural causal models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

D. Yao, D. Xu, S. Lachapelle, S. Magliacane, P. Taslakian, G. Martius, J. von Kügelgen, and F. Locatello. Multi-view causal representation learning with partial observability. *arXiv preprint arXiv:2311.04056*, 2023.

W. Yao, G. Chen, and K. Zhang. Temporally disentangled representation learning. In *Advances in Neural Information Processing Systems*, 2022a.

W. Yao, Y. Sun, A. Ho, C. Sun, and K. Zhang. Learning temporally causal latent processes from general temporal data. In *International Conference on Learning Representations*, 2022b.

J. Zhang, K. Greenewald, C. Squires, A. Srivastava, K. Shanmugam, and C. Uhler. Identifiability guarantees for causal disentanglement from soft interventions. In *Advances in Neural Information Processing Systems*, 2023.

X. Zheng, B. Aragam, P.K. Ravikumar, and E.P. Xing. DAGs with NO TEARS: Continuous optimization for structure learning. In *Advances in Neural Information Processing Systems*, 2018.

Y. Zheng, I. Ng, and K. Zhang. On the identifiability of nonlinear ICA: Sparsity and beyond. In *Advances in Neural Information Processing Systems*, 2022.