Proceedings of Machine Learning Research vol 140:1–50, 2022 1st Conference on Causal Learning and Reasoning

Disentanglement via Mechanism Sparsity Regularization: A New Principle for Nonlinear ICA

Author names withheld

Editor: Under Review for CLeaR 2022

Abstract

This work introduces a novel principle we call *disentanglement via mechanism sparsity regularization*, based on the idea that the dynamics of high-level concepts are often sparse. We propose a representation learning method that *induces* disentanglement by *simultaneously* learning the latent factors and the sparse causal graphical model that relates them. We develop a rigorous identifiability theory, building on recent nonlinear independent component analysis (ICA) results, that formalizes this principle and shows how the latent variables can be recovered up to permutation if one regularizes the latent mechanisms to be sparse and if some graph connectivity criterion is satisfied by the data generating process. As a special case of our framework, we show how one can leverage unknown-target interventions on the latent factors to disentangle them, thereby drawing further connections between ICA and causality. We also propose a VAE-based method in which the latent mechanisms are learned and regularized via binary masks, and validate our theory by showing it learns disentangled representations in simulations.

Keywords: Causal representation learning, disentanglement, nonlinear ICA, causal discovery

1. Introduction

It has been argued that causal reasoning will be central to move modern machine learning algorithms beyond their current shortcomings, such as their lack of *robustness*, *transferability* and *interpretability* (Pearl, 2019; Schölkopf, 2019). However, it is still unclear how to reconcile the causal graphical model (CGM) formalism (Pearl, 2009; Peters et al., 2017), which operates on semantically meaningful high-level variables, with deep neural networks, which excel on unstructured, low-level, high-dimensional observations, e.g. images. One way forward would be a two-step approach in which we first *disentangle* the high-level variables from low-level observations (Bengio et al., 2013; Locatello et al., 2019), then learn a CGM that relates them. Instead, this work proposes a method to do *both steps simultaneously*, and provides a rigorous theory that shows how doing so can *induce* disentanglement when the CGM is regularized to be sparse.

Our contribution is based on recent theoretical results in the nonlinear ICA literature (Hyvärinen et al., 2019; Khemakhem et al., 2020a,b) that assume the data is explained by *unobserved* and *meaningful* latent variables, or factors, Z, which are transformed by a *decoder*, or *mixing function*, $\mathbf{f}$, to produce the observation X. The problem of disentanglement can then be formulated as recovering, or reconstructing, the latent variables from the observation.

This problem is plagued by the difficult question of *identifiability*. Indeed, Hyvärinen and Pajunen (1999) showed that this task is impossible with general nonlinear mixing under the standard assumption of independent latent factors. Nevertheless, recent theoretical developments have shown identifiability of the latent factors is possible in the nonlinear setting, assuming the latent variables are *conditionally independent given an observed auxiliary variable* A (Hyvärinen et al., 2019; Khemakhem et al., 2020a,b). This auxiliary variable can be, for instance, a time or an environment

© 2022 .

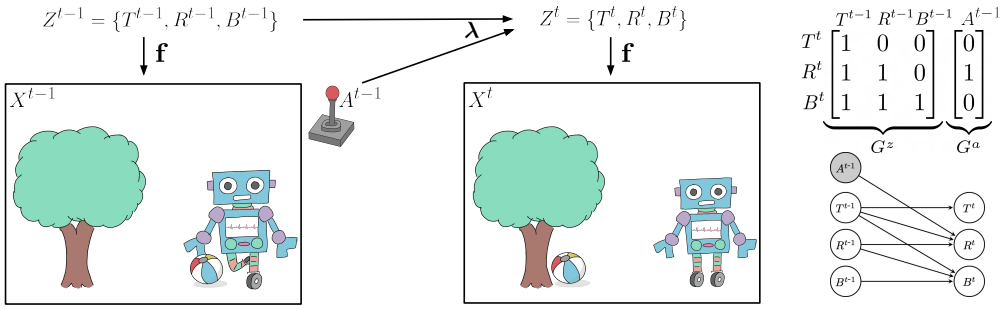

Figure 1: A minimal motivating example. The variables T^t, R^t and B^t represent the x-positions of the tree, the robot and the ball at time t, respectively. Only the image of the scene X^t and the action A^{t-1} are observed. See end of Sec. 2.1 for details.

index, an action in an interactive environment, or even a previous observation if the data has temporal structure, as long as its effect on the latent factors is "sufficiently strong".

The present paper introduces *mechanism sparsity regularization* as a new path to disentanglement. By building on the recent theoretical developments in ICA, we show that if the high-level variables have a *sparse* temporal structure and/or an action is observed and affects the high-level variables *sparsely*, then the latent variables can be recovered by regularizing the inferred graphical model to have sparse dependencies (Thm. 5). In estimating the latent variables, the presented methodology estimates the causal graph describing them and their relation to the action A (when available). Our theoretical result provides a precise graphical criterion to ensure regularized estimation enables identification. From the perspective of ICA, our approach extends the domain of known cases where latent variables can be recovered from observations. From the perspective of causal learning, we propose a principled solution to jointly extract the high-level variables and learn how they are causally related.

Fig. 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 one another) from a time series of images, X^t, and agent actions, A^t. Thm. 5 shows how the sparse dependencies between the objects can be leveraged to estimate the latent variables as well as the graph describing their dynamics. The learned CGM could subsequently be used 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. Interventions allow an agent to imagine situations it has never seen before, which would not be possible without a disentangled representation (Schölkopf, 2019). Moreover, disentanglement could be useful for interpretability by allowing for the extraction of a causal graph of the agent actions (Pearl, 2019).

Contributions:

1. A new principle to achieve disentanglement based on *mechanism sparsity regularization* motivated by a rigorous, novel identifiability theory (Thm. 5).
2. An estimation procedure which relies on variational autoencoders (VAEs) (Kingma and Welling, 2014) and learned causal mechanisms regularized for sparsity via binary masks.
3. An illustration of our theoretical predictions being satisfied in practice by our estimation procedure on synthetic datasets.

The paper is structured as follows. Sec. 2.1 introduces the model under consideration. Sec. 2.2 defines the notions of linear and permutation equivalence between representations. Sec. 2.3 provides conditions to identify the model up to linear equivalence. Sec. 2.4 shows how mechanism sparsity regularization can induce permutation-identifiability, i.e. disentanglement. Sec. 2.5 proposes a VAE-based approach to model estimation, as well as a practical way to induce sparsity based on binary masks. Sec. 3 situates this paper in the related literature. Sec. 4 illustrates the proposed identifiability theory and learning methods on synthetic data.

2. Disentanglement via Mechanism Sparsity Regularization

2.1. An identifiable latent causal model

We now specify the setting under consideration. Assume we observe the realization of a sequence of d_x-dimensional random vectors $\{X^t\}_{t=1}^T$ and a sequence of d_a-dimensional auxiliary vectors $\{A^t\}_{t=0}^{T-1}$. The coordinates of A^t are either discrete or continuous and can potentially represent, for example, an action taken by an agent, or the index of the environment the corresponding observation was taken from. Going forward, we will refer to A^t as the action vector. The observations $\{X^t\}$ are assumed to be explained by a sequence of hidden d_z-dimensional continuous random vectors $\{Z^t\}_{t=1}^T$ via the equation $X^t = \mathbf{f}(Z^t)+N^t$ where $N^t \sim \mathcal{N}(0, \sigma^2 I)$ are mutually independent across time and independent of all Z^t and A^t. Throughout, we assume $\mathbf{f} : \mathcal{Z} \to \mathcal{X}$ is a diffeomorphism[1] where $\mathcal{Z}$ is the support of Z^t for all t, and $\mathcal{X} := \mathbf{f}(\mathcal{Z})$, i.e. the image of $\mathcal{Z}$ under $\mathbf{f}$. We suppose that $d_z \leq d_x$ and that each coordinate Z_i^t contains semantic information about the observation. For example, if the observations $\{X^t\}$ represent high-dimensional images, the coordinates Z_i^t might be the position of an object, its color, or its orientation in space. We denote $Z^{\leq t} := [Z^1 \ \cdots \ Z^t] \in \mathbb{R}^{d_z \times t}$and analogously for $Z^{<t}$ and other random vectors.

Similar to previous work on nonlinear ICA (Hyvärinen et al., 2019; Khemakhem et al., 2020a), we assume the variables Z_i^t are mutually independent given $Z^{<t}$ and $A^{<t}$

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

Our theory holds for a rich family of conditional densities $p(z_i^t \mid z^{<t}, a^{<t})$ called the *exponential family* (Wainwright and Jordan, 2008), which has the following form:

$$p(z_i^t \mid z^{<t}, a^{<t}) = h_i(z_i^t) \exp\{\mathbf{T}_i(z_i^t)^\top \boldsymbol{\lambda}_i(G_i^z \odot z^{<t}, G_i^a \odot a^{<t}) - \psi_i(z^{<t}, a^{<t}))\} \,. \tag{2}$$

Well-known distributions which belong to this family include the Gaussian and beta distribution. In the Gaussian case, the *sufficient statistic* is $\mathbf{T}_i(z) := (z, z^2)$ and the *base measure* is $h_i(z) := \frac{1}{\sqrt{2\pi}}$. The function $\boldsymbol{\lambda}_i(G_i^z \odot z^{<t}, G_i^a \odot a^{<t})$ outputs the *natural parameter* vector 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 $(\lambda_1, \lambda_2) = (\frac{\mu}{\sigma^2}, -\frac{1}{\sigma^2})$. 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. It will be useful to keep this example in mind throughout the paper. The

1. A *diffeomorphism* is a differentiable bijection with a differentiable inverse.

remaining term $\psi_i(z^{<t}, a^{<t})$ acts as a normalization constant. The binary vectors $G_i^z \in \{0,1\}^{d_z}$ and $G_i^a \in \{0,1\}^{d_a}$ act as masks selecting the direct causal parents of z_i^t. The Hadamard product $\odot$ is applied element-wise and broadcasted along the time dimension.[2] We define

$$G^z := [G_1^z \ \cdots \ G_{d_z}^z]^\top \in \mathbb{R}^{d_z \times d_z} \qquad G^a := [G_1^a \ \cdots \ G_{d_a}^a]^\top \in \mathbb{R}^{d_z \times d_a} \tag{3}$$

as well as $G := [G^z \ G^a]$ which can be interpreted as the adjacency matrix of the causal graph. Indeed, (1) & (2) describe a CGM over the unobserved latent variables $Z^{\leq T}$ conditioned on the auxiliary variables $A^{<T}$. These masks will allow us to regularize the causal graph for sparsity.

We define $\boldsymbol{\lambda}(z^{<t}, a^{<t}) \in \mathbb{R}^{kd_z}$ to be the concatenation of all $\boldsymbol{\lambda}_i(G_i^z \odot z^{<t}, G_i^a \odot a^{<t})$ and similarly for $\mathbf{T}(z^t) \in \mathbb{R}^{kd_z}$. Note that $\boldsymbol{\lambda}(z^{<t}, a^{<t})$ depends on G, implicitly to simplify the notation.

The learnable parameters are $\theta := (\mathbf{f}, \boldsymbol{\lambda}, G)$, which induce a conditional probability distribution $\mathbb{P}_{X^{\leq T}|a;\theta}$ over $X^{\leq T}$, given $A^{<T} = a$. Let $\mathcal{A} \subset \mathbb{R}^{d_a}$ be the set of possible values A^t can take. We assume $p(a^{<T})$ has probability mass over all $\mathcal{A}^T$. This could arise, for instance, when A^t is sampled from a policy $\pi(a^t \mid z^t)$ distribution with probability mass everywhere in $\mathcal{A}$.

A motivating example. Fig. 1 represents a minimal example where our theory applies. The environment consists of three objects: a tree, a robot and a ball with x-positions T^t, R^t and B^t, respectively. Together, they form the vector Z^t of high-level latent variables, i.e. $Z^t = (T^t, R^t, B^t)$. A remote controls the direction in which the wheels of the robot turn. The vector 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 A^t and the images X^t representing the scene which is given by $X^t = \mathbf{f}(Z^t) + N^t$. The dynamics of the environment is governed by the transition function $\boldsymbol{\lambda}$. Assuming a Gaussian model with fixed variance for the latent factors $Z^{\leq T}$, $\boldsymbol{\lambda}$ would output the expected position of every object given their previous positions. Plausible connectivity graphs G^z and G^a are given in Fig. 1 showing how the latent factors are related, and how the controller affects them. For every object, its position at time step t depends on its position at $t-1$. The position of the tree, T^t, is not affected by anything, since neither the robot nor the ball can change its position. The robot, R^t, changes its position based on both the action, A^{t-1} and the position of the tree, T^{t-1} (in case of collision). The ball position, B^t, is affected by both the robot, which can kick it around by running into it, and the tree, on which it can bounce. The key observations here are that (i) the different objects interact *sparsely* with one another and (ii) the action A^t affects very few objects (in this case, only one). Thm. 5 will show how one can leverage this sparsity for disentanglement.

2.2. Identifiability and model equivalence

To formalize the problem of disentanglement, we will rely on the notion of *identifiability*, which is a property a model has when its parameters can be uniquely determined by the distribution that it represents. Formally, given some distribution $\mathbb{P}_\theta$ parameterized by θ, this means

$$\forall \theta, \tilde{\theta}, \ \mathbb{P}_\theta = \mathbb{P}_{\tilde{\theta}} \implies \theta = \tilde{\theta}\,. \tag{4}$$

For a model as flexible as the one described in the previous section, identifying the exact parameter θ is too strong a demand. Instead, we will be interested in identifying the parameter θ *up to an equivalence class*, which amounts to substituting some equivalence relation $\theta \sim \tilde{\theta}$ for $\theta = \tilde{\theta}$ in (4). We

2. This implies that if z_i^t is connected to z_j^{t-1}, it is also connected to $z_j^{<t-1}$. Our theory may be generalizable to G having different connectivity structure across time; but we keep this convention to simplify the notation.

now present two equivalence relations for the model presented in Sec. 2.1 adapted from Khemakhem et al. (2020a): linear and permutation equivalence. The latter will help us formalize disentanglement. In what follows, we overload the notation by defining $\mathbf{f}^{-1}(z^{<t}) := [\mathbf{f}^{-1}(z^1) \cdots \mathbf{f}^{-1}(z^{t-1})]$.

Definition 1 (Linear equivalence) *Let $\mathcal{X} := \mathbf{f}(\mathcal{Z})$ and $\tilde{\mathcal{X}} := \tilde{\mathbf{f}}(\mathcal{Z})$, i.e., the image of the support of Z^t under $\mathbf{f}$ and $\tilde{\mathbf{f}}$, respectively. We say θ is **linearly equivalent** to $\tilde{\theta}$ if and only if $\mathcal{X} = \tilde{\mathcal{X}}$ and there exists an invertible matrix $L \in \mathbb{R}^{kd_z \times kd_z}$ as well as vectors $b, c \in \mathbb{R}^{kd_z}$ such that*

1. $\mathbf{T}(\mathbf{f}^{-1}(x)) = L\mathbf{T}(\tilde{\mathbf{f}}^{-1}(x)) + b$, $\forall x \in \mathcal{X}$; *and*
2. $L^\top \boldsymbol{\lambda}(\mathbf{f}^{-1}(x^{<t}), a^{<t}) + c = \tilde{\boldsymbol{\lambda}}(\tilde{\mathbf{f}}^{-1}(x^{<t}), a^{<t})$, $\forall t \in \{1, ..., T\}, x^{<t} \in \mathcal{X}^{t-1}, a^{<t} \in \mathcal{A}^t$.

In this case, we write $\theta \sim_L \tilde{\theta}$.

Hence, two models are linearly equivalent if they entail the same data manifold $\mathcal{X}$ and their respective representations $\mathbf{f}^{-1}(x)$ and $\tilde{\mathbf{f}}^{-1}(x)$ transformed through the element-wise nonlinearity $\mathbf{T}$ are the same everywhere on $\mathcal{X}$ *up to an affine transformation*.

In the Gaussian case, with variance fixed to one, $\mathbf{T}(z) := z$ and $\boldsymbol{\lambda}$ outputs the usual mean parameter μ (here, $k = 1$). The first condition therefore means we can go from one representation to another via an affine transformation.

Suppose θ corresponds to the data generating process, while $\hat{\theta}$ is some learned model. Both being linearly equivalent is not enough to declare the learned representation disentangled, since the matrix L might still "mix up" the variables i.e. one component of $\mathbf{f}^{-1}$ corresponding to multiple components of $\hat{\mathbf{f}}^{-1}$. However, if L happens to have a (block-)permutation structure, we have a one-to-one correspondence between the ground truth latent factors of the data and the coordinates of the learned representation.

Definition 2 (Permutation equivalence) *We say θ is **permutation-equivalent** to $\tilde{\theta}$ if and only if $\theta \sim_P \tilde{\theta}$ (Def. 1) where P has a block-permutation structure respecting $\mathbf{T}$, i.e. there are d_z invertible $k \times k$ matrices $L_1, ..., L_{d_z}$ and a d_z-permutation π such that for all $y = [y_1 \dots y_{d_z}]^\top \in \mathbb{R}^{kd_z}$, $Py = [y_{\pi(1)} L_1^\top \dots y_{\pi(d_z)} L_{d_z}^\top]^\top$.*

In the Gaussian case with a fixed variance, permutation equivalence implies that each coordinate i of one representation is equal to the scaled and shifted coordinate $\pi(i)$ of the other, for some permutation π. Inspired by previous works on nonlinear ICA, we define *disentanglement* as follows.

Definition 3 (Disentanglement) *Given a ground-truth model θ, we say a learned model $\hat{\theta}$ is **disentangled** when θ and $\hat{\theta}$ are permutation-equivalent.*

2.3. Conditions for linear identifiability

From now on, it will be useful to think of θ as the *ground-truth parameter* and $\hat{\theta}$ as a *learned parameter*. The following theorem provides conditions that ensure *linear identifiability*, which is defined as

$$\mathbb{P}_{X^{\leq T}|a;\theta} = \mathbb{P}_{X^{\leq T}|a;\hat{\theta}} \; \forall a \in \mathcal{A}^T \implies \theta \sim_L \hat{\theta}\,, \tag{5}$$

where L is some invertible matrix (not necessarily with a block-permutation structure). This theorem is strongly inspired by Thm. 1 from Khemakhem et al. (2020a), which we which we elaborate upon in Sec. 3, and is central to the stronger permutation-identifiability (disentanglement) theorems of the following section. A proof can be found in App. A.

Theorem 4 (Conditions for linear identifiability) *Suppose we have two models as described in Sec. 2.1 with parameters $\theta = (\mathbf{f}, \boldsymbol{\lambda}, G)$ and $\hat{\theta} = (\hat{\mathbf{f}}, \hat{\boldsymbol{\lambda}}, \hat{G})$ for a fixed T. Suppose the following assumptions hold:*

1. *For all $i \in \{1, ..., d_z\}$, the sufficient statistic $\mathbf{T}_i$ is minimal (see Definition 9).*
2. ***[Sufficient variability]** There exist $(z_{(p)}, a_{(p)})_{p=0}^{kd_z}$ in their respective supports such that the kd_z-dimensional vectors $(\boldsymbol{\lambda}(z_{(p)}, a_{(p)}) - \boldsymbol{\lambda}(z_{(0)}, a_{(0)}))_{p=1}^{kd_z}$ are linearly independent.*

Then, we have linearly identifiability: whenever $\mathbb{P}_{X^{\leq T}|a;\theta} = \mathbb{P}_{X^{\leq T}|a;\hat{\theta}}$ for all $a \in \mathcal{A}^T$, then $\theta \sim_L \hat{\theta}$.

The first assumption is a standard one saying that $\mathbf{T}_i$ is defined appropriately to ensure that the parameters of the exponential family are identifiable (see e.g. Wainwright and Jordan (2008, p. 40)). The second assumption is sometimes called the *assumption of variability* (Hyvärinen et al., 2019), and requires that the conditional distribution of Z^t depends "sufficiently strongly" on $Z^{<t}$ and/or $A^{<t}$. We stress the fact that this assumption concerns the ground-truth data generating model θ. Notice that the $z_{(p)}$ represent values of $Z^{<t}$ for potentially different values of t and can thus have different dimensions.

In the Gaussian case with variance fixed to one, the sufficient variability assumption requires that the values of the conditional mean $\mathbb{E}[Z^t|z^{<t}, a^{<t}]$ are not all contained in a *proper*[3] affine subspace of $\mathbb{R}^{d_z}$. This can be interpreted as having a *sufficiently nonlinear* transition model.

2.4. Permutation-identifiability via mechanism sparsity regularization

We are now ready to present the core contribution of this work, i.e. a novel permutation-identifiability result based on mechanism sparsity regularization (Thm. 5). The intuition for this result is that, under appropriate assumptions (that are satisfied in the motivating example of Fig. 1), models that have an entangled representation also have a denser adjacency matrix $\hat{G}$. Thus, by regularizing $\hat{G}$ to be sparse, we exclude entangled models, leaving us with only the disentangled ones. Thm. 5 gives precise conditions about the data-generating model θ under which fitting the model $\hat{\theta}$ and regularizing the graph $\hat{G} = [\hat{G}^z \ \hat{G}^a]$ to be sparse will be sufficient to obtain a disentangled model (Def. 3). Recall that $\hat{G}^z$ controls the connectivity between the latent variables from one time step to another and that $\hat{G}^a$ controls the connectivity between the action $A^{<t}$ and the latent variable Z^t. Sec. 3 will contrast these results with those introduced in the recent literature on nonlinear ICA.

Theorem 5 (Disentanglement via mechanism sparsity) *Suppose we have two models as described in Sec. 2.1 with parameters $\theta = (\mathbf{f}, \boldsymbol{\lambda}, G)$ and $\hat{\theta} = (\hat{\mathbf{f}}, \hat{\boldsymbol{\lambda}}, \hat{G})$ for a fixed T representing the same distribution, i.e. $\mathbb{P}_{X^{\leq T}|a;\theta} = \mathbb{P}_{X^{\leq T}|a;\hat{\theta}}$ for all $a \in \mathcal{A}^T$. Suppose the assumptions of Thm. 4 hold and:*

1. *The sufficient statistic $\mathbf{T}$ is d_z-dimensional ($k = 1$) and is a diffeomorphism from $\mathcal{Z}$ to $\mathbf{T}(\mathcal{Z})$.*
2. ***[Sufficient time-variability]** The Jacobian of the ground-truth transition function $\boldsymbol{\lambda}$ with respect to z varies "sufficiently", as formalized in Assumption 1 in the next section.*

Then, there exists a permutation matrix P such that $PG^zP^\top \subset \hat{G}^z$.[4] *Further assume that*

3. A subset $A \subset B$ is *proper* when $A \neq B$.
4. Given two binary matrices M^1 and M^2 with equal shapes, we say $M^1 \subset M^2$ when $M^1_{i,j} = 1 \implies M^2_{i,j} = 1$.

3. ***[Sufficient action-variability]*** *The ground-truth transition function $\boldsymbol{\lambda}$ is affected "sufficiently strongly" by each individual action a_ℓ, as formalized in Assumption 2 in the next section.*

Then $PG^a \subset \hat{G}^a$. Further assume that

4. ***[Sparsity]*** $||\hat{G}||_0 \leq ||G||_0$.

Then, $PG^z P^\top = \hat{G}^z$ and $PG^a = \hat{G}^a$. Further assume that

5. ***[Graphical criterion]*** *For all $p \in \{1, ..., d_z\}$, there exist sets $\mathcal{I}, \mathcal{J} \subset \{1, ..., d_z\}$ and $\mathcal{L} \subset \{1, ..., d_a\}$ such that*

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

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

Then θ and $\hat{\theta}$ are permutation-equivalent (Def. 2), i.e. the model $\hat{\theta}$ is disentangled.

The first assumption is satisfied for example by the Gaussian case with variance fixed to one since $\mathbf{T}(z) = z$ is a diffeomorphism. In contrast, it is not satisfied in the Gaussian case with fixed mean since $\mathbf{T}(z) = z \odot z$ is not invertible.

Sufficient variability. Thm. 5 has also two sufficient variability assumptions, one for G^z and one for G^a. Rigorous statements of these are delayed until Sec. 2.4.1. Intuitively, both assumptions requires that the ground-truth transition function $\boldsymbol{\lambda}$ is complex enough.

Sparsity. The first three assumptions imply that the learned graph $\hat{G}$ is a supergraph of some permutation of the ground-truth graph G. By adding the *sparsity* assumption, we have that the learned graph $\hat{G}$ *is exactly* a permutation of the ground-truth graph G. This assumption is satisfied if $\hat{G}$ is a minimal graph among all graphs that properly fit the data. In Sec. 2.5, we suggest achieving this by regularizing $\hat{G}$ to be sparse.

Graphical criterion. The very last assumption is a *graphical criterion* that guarantees disentanglement. This criterion is trivially satisfied when G^z is diagonal, since $\{i\} = \mathbf{Pa}_i^z$ for all i (actions are not necessary here). This simple case amounts to having mutual independence between the sequences $Z_i^{\leq T}$, 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 that $\{T\} = \mathbf{Pa}_T^z$, $\{R\} = \mathbf{Ch}_R^z \cap \mathbf{Pa}_R^z$ and $\{B\} = \mathbf{Ch}_B^z$. This example is actually part of an interesting family of graphs that satisfy our criterion:

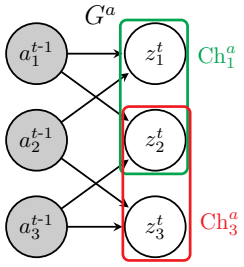

Figure 2: A non-trivial case where actions a are enough to satisfy the graphical criterion of Thm. 5. Indeed, we have $\{z_1\} = \mathbf{Ch}_1^a \cap \mathbf{Ch}_2^a$, $\{z_2\} = \mathbf{Ch}_1^a \cap \mathbf{Ch}_3^a$ and $\{z_3\} = \mathbf{Ch}_2^a \cap \mathbf{Ch}_3^a$.

Proposition 6 (Sufficient condition for the graphical criterion) *If $G_{i,i}^z = 1$ for all i (all nodes have a self-loop) and G^z has no 2-cycles, then G satisfies the graphical criterion of Thm. 5.*

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

Prop. 6 implies that actions are not necessary for permutation-identifiability. Conversely, it is also possible to satisfy the graphical criterion without temporal dependencies, e.g., see Fig. 2, which highlights the fact that the condition of Prop. 6 is not necessary.

In App. A.3.2 & A.3.3, we present two theorems which are versions of Thm. 5 specialized to time-sparsity and action-sparsity, respectively. Note that they are not special cases of Thm. 5.

2.4.1. Sufficient variability assumptions

We now present the two technical variability assumptions of Thm. 5. Intuitively, both assumptions require that the data generating model has a "sufficiently complex" transition function $\boldsymbol{\lambda}$.

Notation. Let $\mathbb{R}^{d_z \times d_z}_{G^z}$ be the set of matrices $M \in \mathbb{R}^{d_z \times d_z}$ such that $M_{i,j} = 0$ whenever $G^z_{i,j} = 0$. Similarly, let $\mathbb{R}^{d_z}_{\mathbf{Ch}^a_\ell}$ be the subspace of $\mathbb{R}^{d_z}$ where all coordinates outside $\mathbf{Ch}^a_\ell$ are zero.

Assumption 1 (Sufficient time-variability) *There exist* $\{(z_{(p)}, a_{(p)}, \tau_{(p)})\}_{p=1}^{||G^z||_0}$ *belonging to their respective support such that*

$$\operatorname{span}\left\{ D_z^{\tau_{(p)}} \boldsymbol{\lambda}(z_{(p)}, a_{(p)}) D_z \mathbf{T}(z_{(p)}^{\tau_{(p)}})^{-1} \right\}_{p=1}^{||G^z||_0} = \mathbb{R}^{d_z \times d_z}_{G^z} ,$$

where $D_z^{\tau_{(p)}} \boldsymbol{\lambda}$ *and* $D_z \mathbf{T}$ *are Jacobians with respect to* $z^{\tau_{(p)}}$ *and* z*, respectively.*

Notice the Jacobian $D_z^{\tau_{(p)}} \boldsymbol{\lambda}$ is always in $\mathbb{R}^{d_z \times d_z}_{G^z}$ because of how G^z masks the input of $\boldsymbol{\lambda}_i$ in (2). The *sufficient time-variability* assumption further requires that the Jacobian varies "enough" so that it cannot be contained in a *proper* subspace of $\mathbb{R}^{d_z \times d_z}_{G^z}$. The following *sufficient action-variability* assumption has an analogous interpretation.

Assumption 2 (Sufficient action-variability) *For all* $\ell \in \{1, ..., d_a\}$*, there exist* $\{(z_{(p)}, a_{(p)}, \epsilon_{(p)}, \tau_{(p)})\}_{p=1}^{|\mathbf{Ch}^a_\ell|}$ *belonging to their respective support such that*

$$\operatorname{span}\left\{ \Delta_\ell^{\tau_{(p)}} \boldsymbol{\lambda}(z_{(p)}, a_{(p)}, \epsilon_{(p)}) \right\}_{p=1}^{|\mathbf{Ch}^a_\ell|} = \mathbb{R}^{d_z}_{\mathbf{Ch}^a_\ell} ,$$

where $\Delta_\ell^\tau \boldsymbol{\lambda}(z^{<t}, a^{<t}, \epsilon)$ *is a partial difference defined by*

$$\Delta_\ell^\tau \boldsymbol{\lambda}(z^{<t}, a^{<t}, \epsilon) := \boldsymbol{\lambda}(z^{<t}, a^{<t} + \epsilon E_{\ell,\tau}) - \boldsymbol{\lambda}(z^{<t}, a^{<t}) , \tag{6}$$

where $\epsilon \in \mathbb{R}$ *and* $E_{\ell,\tau} \in \mathbb{R}^{d_a \times t}$ *is the one-hot matrix with the entry* (ℓ, τ) *set to one. Thus,* (6) *is the discrete analog of a partial derivative w.r.t.* a_ℓ^τ*.*

In App. A.5, we provide a plausible transition function $\boldsymbol{\lambda}$ based on the illustrative example of Fig. 1 and show that it has sufficient variability. Given the complex interactions which abound in the real world, we conjecture that "realistic" transition functions likely are "complex enough" to satisfy both assumptions.

2.4.2. ACTIONS AS INTERVENTIONS WITH UNKNOWN TARGETS

An interesting special case of Thm. 5 is when A^{t-1} corresponds to a one-hot vector indexing an *intervention with unknown targets* on the latent variables 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 like in our case (Eaton and Murphy, 2007; Mooij et al., 2020; Squires et al., 2020; Jaber et al., 2020). Specifically, assume $A^{t-1} \in \{\mathbf{0}, e_1, ..., e_{d_a}\}$, where each e_ℓ is a one-hot vector. The action $A^{t-1} = \mathbf{0}$ corresponds to the *observational setting*, i.e. when no intervention occurred, while $A^{t-1} = e_\ell$ corresponds the ℓth intervention. In that context, the unknown graph G^a describes which latents are targeted by the intervention, i.e. $G^a_{i,\ell} = 1$ if and only if z_i is targeted by the ℓth intervention. Here, the partial difference $\Delta^{t-1}_\ell \boldsymbol{\lambda}(z^{<t}, \mathbf{0}, 1)$ measures the difference of natural parameters between the observational setting and the ℓth intervention.

2.5. Regularized model estimation

In order to estimate from data the model presented in previous sections, we propose 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 $\mathbf{f}$. We consider an approximate posterior of the form

$$q(z^{\leq T} \mid x^{\leq T}, a^{<T}) := \prod_{t=1}^{T} q(z^t \mid x^t)\,, \tag{7}$$

where $q(z^t \mid x^t)$ is a Gaussian distribution with mean and diagonal covariance outputted by a neural network $\texttt{encoder}(x^t)$. In our experiments, the transition functions $\boldsymbol{\lambda}_i$ are parameterized by fully connected neural networks that look only at a fixed window of s lagged latent variables.[5] In all experiments, $\hat{p}(z_i^t \mid z^{<t}, a^{<t})$ is Gaussian with a learned variance that does not depend on $(z^{<t}, a^{<t})$ (see App. B.2 for details). This variational inference model induces the following evidence lower bound (ELBO) on $\log \hat{p}(x^{\leq T}|a^{<T})$:

$$\sum_{t=1}^{T} \mathop{\mathbb{E}}_{Z^t \sim q(\cdot|x^t)} [\log \hat{p}(x^t \mid Z^t)] - \mathop{\mathbb{E}}_{Z^{<t} \sim q(\cdot|x^{<t})} KL(q(Z^t \mid x^t)||\hat{p}(Z^t \mid Z^{<t}, a^{<t}))\,. \tag{8}$$

We derive this fact in App. A.6. 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 $p(z^t|x^{\leq T}, a^{<T})$, (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 Thm. 4, we can guarantee that the learned model $\hat{\theta}$ will be linearly equivalent to the ground truth model θ.

To go from linear identifiability (Def. 1) to permutation-identifiability (Def. 2), Thm. 5 suggests we should not only fit the data, but also choose the model such that $\hat{G}$ is sparse (or minimal). To achieve this in practice, we add regularization terms $-\alpha_z||\hat{G}^z||_0$ and $-\alpha_a||\hat{G}^a||_0$ to the ELBO objective, where $\alpha_z \geq 0$ and $\alpha_a \geq 0$ are hyperparameters. To make the objective amenable to gradient-based optimization, we treat $\hat{G}^z$ and $\hat{G}^a$ as independent Bernoulli random variables with

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

probabilities of success $\texttt{sigmoid}(\gamma^z_{i,j})$ and $\texttt{sigmoid}(\gamma^a_{i,j})$ and optimize the continuous parameters γ^z and γ^a using the Gumbel-Softmax gradient estimator (Jang et al., 2017; Maddison et al., 2017). This strategy has been used successfully in previous work to enable gradient-based causal discovery (Ng et al., 2019; Brouillard et al., 2020).

3. Related work

Recent theoretical results have shown that nonlinear ICA is possible when leveraging additional assumptions, e.g., nonstationarity (Hyvarinen and Morioka, 2016) and temporal dependencies (Hyvarinen and Morioka, 2017). Hyvärinen et al. (2019) generalized these works by introducing the notion of auxiliary variables (which correspond to A in our work). All of these methods rely on noise contrastive estimation (NCE) (Gutmann and Hyvärinen, 2012), which underlies the state-of-the-art in self-supervised representation learning (Oord et al., 2018; Chen et al., 2020), in which identifiability has been used as an analysis tool (Roeder et al., 2020; Zimmermann et al., 2021; von Kügelgen et al., 2021). Subsequent works have shown similar results using VAEs (Khemakhem 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).

Khemakhem et al. (2020a), which introduced iVAE, is likely the closest to the present work. Thm. 4 is quite similar to Thm. 1 from Khemakhem et al. (2020a), but iVAE's notion of linear equivalence is different in that it does not characterize the relationship between $\boldsymbol{\lambda}$ and $\hat{\boldsymbol{\lambda}}$, which is crucial for our proof of Thm. 5. The most significant distinction between the theory of (Khemakhem et al., 2020a) and ours is how *permutation-identifiability* is obtained: Thm. 2 & 3 from iVAE shows that if the assumptions of their Thm. 1 are satisfied and $\mathbf{T}_i$ has dimension $k > 1$ or is non-monotonic, then the model is not just linearly, but permutation-identifiable. In contrast, our theory covers the case where $k = 1$ and $\mathbf{T}_i$ is monotonic, like in the Gaussian case with fixed variance. Interestingly, Khemakhem et al. (2020a) mentioned this specific case as a counterexample to their theory in their Prop. 3. The extra power of our theory comes from the extra *structure* in the dependencies of the latent factors coupled with sparsity regularization.

Similar to our theory, PCL (Hyvarinen and Morioka, 2017) and SlowVAE (Klindt et al., 2021) leverage temporal dependence, but always assume mutual independence of the sequences $\{Z^t_i\}$. In our notation, this amounts to assuming the graph G^z is diagonal. Our theory allows for more flexibility by accounting for a variety of dependency structures, like a triangular graph G^z. However, we do not claim our theory is a strict generalization of these works, since, for instance, the latent Laplacian transition model assumed by SlowVAE is not in the exponential family.

Locatello et al. (2020) also leverages temporal dependence, but assumes each pair (Z^t, Z^{t+1}) shares a random subset S of its components. Our theory allows for every latent factor to change constantly and, thus, does not make this assumption. Interestingly, they assume that, for all i, $P(S \cap S' = \{i\}) > 0$ (for i.i.d S and S'), which resembles our graphical criteria.

While there is a significant amount of interest in learning CGMs between high-level latent variables extracted from low-level observations (Schölkopf, 2019; Schölkopf et al., 2021), there have been comparatively few solutions contributed to the literature. Of works which learn CGMs, a number assume the causal graph structure is known (Kocaoglu et al., 2018; Shen et al., 2021; Nair et al., 2019). Bengio et al. (2020) suggests using adaptation speed as a heuristic objective to disentangle causally related latent factors. Yang et al. (2021) learns the causal graph by incorporating a "causal model layer" into iVAE, but is limited to linear dependencies and does not apply to time-series.

4. Experiments

To illustrate Thm. 5 and the benefit of mechanism sparsity regularization for disentanglement, we apply the regularized VAE method of Section 2.5 on synthetic datasets that both satisfy (Fig. 3) and violate (Fig. 5, in App. B.3) the assumptions of Thm. 5. Details about the implementation of our approach are provided in App. B.2

Synthetic datasets. The datasets we considered are separated in two groups: *time-sparsity* and *action-sparsity* datasets. The former group has only temporal dependence without actions, we thus fix $\hat{G}^a = \mathbf{0}$, while the latter has only actions without temporal dependence, we thus fix $\hat{G}^z = \mathbf{0}$. In each dataset, the ground-truth mixing function $\mathbf{f}$ is a randomly initialized neural network. The dimensionality of Z and X are $d_z = 10$ and $d_x = 20$, respectively. In the action-sparsity datasets, the dimensionality of A is $d_a = 10$. The ground-truth transition model $p(z^t \mid z^{<t}, a^{<t})$ is always a Gaussian with covariance $\sigma_z^2 I$ and a mean outputted by some function $\mu_G(z^{t-1}, a^{t-1})$ (the data is *Markovian*). Hence, each dataset has a 1d sufficient statistic ($k = 1$) that is also monotonic and, thus, is not covered by the theory of Khemakhem et al. (2020a). App. B.1 provides a more detailed descriptions of the datasets including the explicit form of μ and G in each case. Note that the learned transition model $\hat{p}(z^t \mid z^{t-1}, a^{t-1})$ is also Gaussian where the mean is outputted by a MLP.

Performance metrics. To evaluate disentanglement, we will use $\texttt{encoder}(x)$ as a proxy for the learned $\hat{\mathbf{f}}^{-1}(x)$. To assess *linear identifiability*, we report linear regression to predict the ground-truth latent factors from the inferred ones, and report the *coefficient of determination* R^2. To assess *permutation-identifiability*, i.e. disentanglement, we report the *mean correlation coefficient* (MCC), which has been used in similar contexts, e.g. Khemakhem et al. (2020a). This metric is obtained by first computing the Pearson correlation matrix $C \in \mathbb{R}^{d_z \times d_z}$ between the learned representation and the ground truth latent variables. Then, MCC $= \max_{\pi \in \text{permutations}} \frac{1}{d_z} \sum_{i=1}^{d_z} |C_{i,\pi(i)}|$. For our method, which is the only one learning a graph, we also report the *structural hamming distance* (SHD) between the ground-truth graph and the learned graph permuted by π^*, the optimal permutation found when computing MCC. We normalize SHD by the maximal number of edges to ensure it is always between 0 and 1. The normalized SHD is thus the proportion of incorrectly estimated edges in the graph.

Baselines. On the temporal-sparsity datasets, we compare our approach with TCVAE (Chen et al., 2018), PCL (Hyvarinen and Morioka, 2017) and SlowVAE (Klindt et al., 2021). On the action-sparsity datasets, we compare with TCVAE and iVAE (Khemakhem et al., 2020a). 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 App. B.5 for details.

Unsupervised hyperparameter selection. The hyperparameters of the baselines were selected via *unsupervised disentanglement ranking* (UDR) (Duan et al., 2020). For our approach, Fig. 3 & 5 show performance for a range of regularization coefficients α_z and α_a. We suggest selecting it using UDR and excluding coefficients that yield graphs with less than $d_z = 10$ edges, as the graphical criterion cannot be achieved in that case. Fig. 3 & 5 show this *unsupervised* procedure selects a reasonable regularization coefficient (as indicated by the black star). See App. B.6 for details.

Effect of regularization. Fig. 3 shows that the right amount of mechanism sparsity regularization leads to improved disentanglement (as measured by MCC), which is in line with Thm. 5. Improvements in SHD indicate that regularization allows for estimation of the causal graph G (see App. B.4 for visualizations). When α_z and α_a are selected with the filtered UDR procedure, our approach outperforms the baselines by a significant margin, while without regularization, it performs

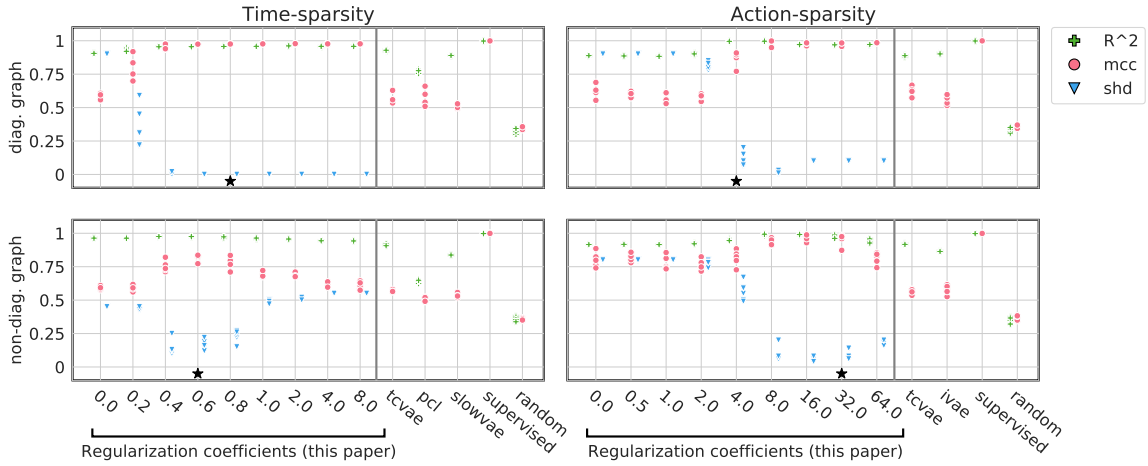

Figure 3: Top row datasets have a diagonal graph and bottom row datasets have a non-diagonal graph satisfying the graphical criterion of Thm. 5. Sufficient variability is always satisfied. In the left column, only $\hat{G}^z$ is learned and we vary α_z, and in the right column, only $\hat{G}^a$ is learned and we vary α_a. For more details on the synthetic datasets, see App. B.1. The black star indicates which regularization parameter is selected by the filtered UDR procedure (see App. B.6). For R^2 and MCC, higher is better. For SHD, lower is better. Performance is reported on 5 random seeds.

similarly. Most baselines obtain a high R^2 but a low MCC, indicating linear identifiability without disentanglement. These observations are consistent across all four datasets.

Violating assumptions. Fig. 5 in App. B.3 presents results on datasets that do not satisfy either sufficient variability or the graphical criterion. In the time-sparsity data that has insufficient variability, regularization does not help and performance is always similar to baselines. In all other datasets, regularization improves MCC, but by a smaller margin than when assumptions are met.

5. 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: objects usually interact sparsely with each other and actions usually affect only a few entities. Building on recent developments in nonlinear ICA, we constructed a rigorous theory which provides precise conditions for when regularizing the mechanisms to be sparse will result in disentanglement, e.g. the graphical criterion on the structure of the dependencies in the data generating process. A special case of our framework shows how one can leverage *unknown-target interventions* on the latent factors for disentanglement, thus bringing closer the language of causality and ICA. We proposed a regularized VAE-based approach and demonstrated that it can improve disentanglement in controlled synthetic settings, thereby preparing the stage for more realistic scenario, e.g. interactive environments. We believe this work opens up new possibilities at the intersection of causality and disentanglement that leverage *structural assumptions*. For instance, we posit that contextual sparsity, i.e. objects only interact with each other in particular situations, could be formalized using the tools developed in this work.

References

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.

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.

Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating sources of disentanglement in vaes. In *Advances in Neural Information Processing Systems*, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020.

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 *Artificial Intelligence and Statistics*, 2007.

X. Glorot and Y. Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, 2010.

Michael U. Gutmann and Aapo Hyvärinen. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *The Journal of Machine Learning Research*, 13:307–361, 2012.

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 *Proceedings of the 20th 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, H. Sasaki, and R. E. Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *AISTATS*. PMLR, 2019.

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. *Proceedings of the 34th International Conference on Machine Learning*, 2017.

I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen. Variational autoencoders and nonlinear ica: A unifying framework. In *Proceedings of the Twenty Third 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 M. Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations*, 2014.

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 *9th International Conference on Learning Representations*, 2021.

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

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 *Proceedings of the 36th 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 *Proceedings of the 37th International Conference on Machine Learning*, 2020.

C. J. Maddison, A. Mnih, and Y. W. Teh. The concrete distribution: A continuous relaxation of discrete random variables. *Proceedings of the 34th International Conference on Machine Learning*, 2017.

J. M. Mooij, S. Magliacane, and T. Claassen. Joint causal inference from multiple contexts. *Journal of Machine Learning Research*, 2020.

Suraj Nair, Yuke Zhu, Silvio Savarese, and Li Fei-Fei. Causal induction from visual observations for goal directed tasks, 2019.

I. Ng, Z. Fang, S. Zhu, Z. Chen, and J. Wang. Masked gradient-based causal structure learning. *arXiv preprint arXiv:1910.08527*, 2019.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

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, 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.

Geoffrey Roeder, Luke Metz, and Diedrik P. Kingma. On linear identifiability of learned representations. *arXiv preprint arXiv:2007.00810*, 2020.

B. Schölkopf. Causality for machine learning, 2019.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Towards causal representation learning, 2021.

Xinwei Shen, Furui Liu, Hanze Dong, Qing Lian, Zhitang Chen, and Tong Zhang. Disentangled generative causal representation learning, 2021.

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. *Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence*, 2020.

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.

Julius von Kügelgen, Yash Sharma, Luigi Gresele, Wieland Brendel, Bernhard Schölkopf, Michel Besserve, and Francesco Locatello. Self-supervised learning with data augmentations provably isolates content from style, 2021.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 2008.

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

Roland S. Zimmermann, Yash Sharma, Steffen Schneider, Matthias Bethge, and Wieland Brendel. Contrastive learning inverts the data generating process. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12979–12990. PMLR, 18–24 Jul 2021.