Robustly Disentangled Causal Mechanisms: Validating Deep Representations for Interventional Robustness

Raphael Suter [1] **Đorđe Miladinović** [1] **Bernhard Schölkopf** [2] **Stefan Bauer** [2]

Abstract

The ability to learn disentangled representations that split underlying sources of variation in high dimensional, unstructured data is important for data efficient and robust use of neural networks. While various approaches aiming towards this goal have been proposed in recent times, a commonly accepted definition and validation procedure is missing. We provide a causal perspective on representation learning which covers disentanglement and domain shift robustness as special cases. Our causal framework allows us to introduce a new metric for the quantitative evaluation of deep latent variable models. We show how this metric can be estimated from labeled observational data and further provide an efficient estimation algorithm that scales linearly in the dataset size.

1. Introduction

Learning deep representations in which different semantic aspects of data are structurally disentangled is of central importance for training robust machine learning models. Separating independent factors of variation could pave the way for successful transfer learning and domain adaptation (Bengio et al., 2013). Imagine the example of a robot learning multiple tasks by interacting with its environment. For data efficiency, the robot can learn a generic representation architecture that maps its high dimensional sensory data to a collection of general, compact features describing its surrounding. For each task, only a subset of features will be required. If the robot is instructed to grasp an object, it must know the shape and the position of the object, however, its color is irrelevant. On the other hand, when pointing to all red objects is demanded, only position and color are required.

Having a *disentangled* representation, where each feature captures only one factor of variation, allows the robot to build separate (simple) models for each task based on only a relevant and stable subselection of these generically learned features. We argue that robustness of the learned representation is a crucial property when this is attempted in practice. It has been proposed that features should be selected based on their robustness or invariance across tasks (e.g., Rojas-Carulla et al., 2018), we hence do not want them to be affected by changes in any other factor. In our example, the robot assigned with the grasping task should be able to build a model using features well describing shape and position of the object. For this model to be robust, however, these features must not be affected by changing color (or any other nuisance factor).

It is striking that despite the recent popularity of disentangled representation learning approaches, a commonly accepted definition and validation metric is missing (Higgins et al., 2018). We view disentanglement as a property of a causal process (Spirtes et al., 1993; Pearl, 2009) responsible for the data generation, as opposed to only a heuristic characteristic of the encoding. Concretely, we call a causal process disentangled when the parents of the generated observations do not affect each other (i.e., there is no total causal effect between them (Peters et al., 2017, Definition 6.12)). We call these parents *elementary ingredients*. In the example above, we view color and shape as elementary ingredients, as both can be changed without affecting each other. Still, there can be dependencies between them if for example our experimental setup is confounded by the capabilities of the 3D printers that are used to create the objects (e.g., certain shapes can only be printed in some colors).

Combining these *disentangled causal processes* with the encoding allows us to study interventional effects on feature representations and estimate them from observational data. This is of interest when benchmarking disentanglement approaches based on ground truth data (Locatello et al., 2018) or trying to evaluate robustness of a deep representations w.r.t. known nuisance factors (e.g., domain changes). In the example of robotics, knowledge about the generative factors

[1]Department of Computer Science, ETH Zurich, Switzerland [2]MPI for Intelligent Systems, Tübingen, Germany. Correspondence to: Raphael Suter <rasuter@icloud.com>, Stefan Bauer <bauers@inf.ethz.ch>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

(e.g., the color, shape, weight, etc. of an object to grasp) is often availabe and can be controlled in experiments.

We will start by first giving an overview of previous work in finding disentangled representations and how they have been validated in Section 2. In Section 3 we introduce our framework for the joint treatment of the disentangled causal process and its learned representation. We introduce our notion of interventional effects on encodings and the following *interventional robustness score* in Section 4 and show how this score can be estimated from *observational* data with an efficient $\mathcal{O}(N)$ algorithm in Section 5. Section 6 provides experimental evidence in a standard disentanglement benchmark dataset supporting the need of a robustness based disentanglement criterion.

OUR CONTRIBUTIONS:

- We introduce a unifying causal framework of disentangled generative processes and consequent feature encodings. This perspective allows us to introduce a novel validation metric, the *interventional robustness score*.
- We show how this metric can be estimated from observational data and provide an efficient algorithm that scales linearly in the dataset size.
- Our extensive experiments on a standard benchmark dataset show that our robustness based validation is able to discover vulnerabilities of deep representations that have been undetected by existing work.
- Motivated by this metric, we additionally present a new visualisation technique which provides an intuitive understanding of dependency structures and robustness of learned encodings.

NOTATION:

We denote the generative factors of high dimensional observations $\boldsymbol{X}$ as $\boldsymbol{G}$. The latent variables learned by a model, e.g., a variational auto-encoder (VAE) (Kingma & Welling, 2014), are denoted as $\boldsymbol{Z}$. We use the notation $E(\cdot)$ to describe the encoding which in case of VAEs corresponds to the posterior mean of $q_\phi(\boldsymbol{z}|\boldsymbol{x})$. Capital letters denote random variables, and lower case observations thereof. Subindices $\boldsymbol{Z}_J$ for a set J or Z_j for a single index j denote the selected components of a multidimensional variable. A backslash $\boldsymbol{Z}_{\setminus J}$ denotes all components except those in J.

2. Related Work

In the framework of variational auto-encoders (VAEs) (Kingma & Welling, 2014) the (high dimensional) observations $\boldsymbol{x}$ are modelled to be generated from some latent features $\boldsymbol{z}$ with chosen prior $p(\boldsymbol{z})$ according to the probabilistic model $p_\theta(\boldsymbol{x}|\boldsymbol{z})p(\boldsymbol{z})$. The generative model $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ as well as the proxy posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ can be estimated using neural networks by maximizing the variational lower bound (ELBO) of $\log p(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N)$:

$$\mathcal{L}_{VAE} = \sum_{i=1}^{N} \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})}[\log p_\theta(\boldsymbol{x}^{(i)}|\boldsymbol{z})] - D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}^{(i)})\|p(\boldsymbol{z})). \tag{1}$$

This objective function a priori does not encourage much structure on the latent space (except some similarity to the chosen prior $p(\boldsymbol{z})$ which is usually isotropic Gaussian). More precisely, for a given encoder E and decoder D any bijective transformation g of the latent space $\boldsymbol{z} = E(\boldsymbol{x})$ yields the same reconstruction $\hat{\boldsymbol{x}} = D(g(g^{-1}(E(\boldsymbol{x})))) = D(E(\boldsymbol{x}))$.

Various proposals for more structure imposing regularization have been made, either with some sort of supervision (e.g. Siddharth et al., 2017; Bouchacourt et al., 2017; Liu et al., 2017; Mathieu et al., 2016; Cheung et al., 2014) or completely unsupervised (e.g. Higgins et al., 2017; Kim & Mnih, 2018; Chen et al., 2018; Kumar et al., 2018; Esmaeili et al., 2018). Higgins et al. (2017) proposed the β-VAE penalizing the Kullback-Leibler divergence (KL) term in the VAE objective (1) more strongly, which encourages similarity to the factorized prior distribution. Others used techniques to encourage statistical independence between the different components in $\boldsymbol{Z}$, e.g., FactorVAE (Kim & Mnih, 2018) or β-TCVAE (Chen et al., 2018), similar to independent component analysis (e.g. Comon, 1994). With disentangling the *inferred prior* (DIP-VAE), Kumar et al. (2018) proposed encouraging factorization of $q_\phi(\boldsymbol{z}) = \int q_\phi(\boldsymbol{z}|\boldsymbol{x})p(\boldsymbol{x})\, d\boldsymbol{x}$.

A special form of structure in the latent space which has gained a lot of attention in recent time is referred to as *disentanglement* (Bengio et al., 2013). This term encompasses the understanding that each learned feature in $\boldsymbol{Z}$ should represent structurally different aspects of the observed phenomena (i.e., capture different sources of variation).

Various methods to validate a learned representation for disentanglement based on known ground truth generative factors $\boldsymbol{G}$ have been proposed (e.g. Eastwood & Williams, 2018; Ridgeway & Mozer, 2018; Chen et al., 2018; Kim & Mnih, 2018). While a universal definition of disentanglement is missing, the most widely accepted notion is that one feature Z_i should capture information of only one generative factor (Eastwood & Williams, 2018; Ridgeway & Mozer, 2018). This has for example been expressed as the mutual information of a single latent dimension Z_i with generative factors $G_1, \dots, G_K$ (Ridgeway & Mozer, 2018), where in the ideal case each Z_i has some mutual information with one generative factor G_k but none with all the others. Similarly, Eastwood & Williams (2018) trained predictors (e.g., Lasso or random forests) for a generative factor G_k based on the representation $\boldsymbol{Z}$. In a disentangled model, each dimension Z_i is only useful (i.e., has high feature importance) to predict one of those factors (see appendix D for details).

Validation without known generative factors is still an open research question and so far it is not possible to quantitatively validate disentanglement in an unsupervised way. The community has been using "latent traversals" (i.e., changing one latent dimension and subsequently re-generating the image) for visual inspection when supervision is not available (see e.g. Chen et al., 2018). This can be used to encounter physically meaningful interpretations of each dimension.

3. Causal Model

We will first consider assumptions for the causal process underlying the data generating mechanism. Following this, we discuss consequences for trying to match encodings $\boldsymbol{Z}$ with causal factors $\boldsymbol{G}$ in a deep latent variable model.

3.1. Disentangled Causal Model

As opposed to previous approaches that defined disentanglement heuristically as properties of the learned latent space, we take a step back and first introduce a notion of disentanglement on the level of the true causal mechanism (or data generation process). Subsequently, we can use this definition to better understand a learned probabilistic model for latent representations and evaluate its properties.

We assume to be given a set of observations from a (potentially high dimensional) random variable $\boldsymbol{X}$. In our model, the data generating process is described by K causes of variation (generative factors) $\boldsymbol{G} = [G_1, \dots, G_K]$ (i.e., $\boldsymbol{G} \to \boldsymbol{X}$) that do not cause each other. These factors $\boldsymbol{G}$ are generally assumed to be unobserved and are objects of interest when doing deep representation learning. In particular, knowledge about $\boldsymbol{G}$ could be used to build lower dimensional predictive models, not relying on the (unstructured) $\boldsymbol{X}$ itself. This could be classic prediction of a label Y, often in "confounded" direction (i.e., predicting effects from other effects) if $\boldsymbol{G} \to (\boldsymbol{X}, Y)$ or in anti-causal direction if $Y \to \boldsymbol{G} \to \boldsymbol{X}$. It is also relevant in a domain change setting when we know that the domain S has an impact on $\boldsymbol{X}$, i.e., $(S, \boldsymbol{G}) \to \boldsymbol{X}$.

Having these potential use cases in mind, we assume the generative factors themselves to be confounded by (multi-dimensional) $\boldsymbol{C}$, which can for example include a potential label Y or source S. Hence, the resulting causal model $\boldsymbol{C} \to \boldsymbol{G} \to \boldsymbol{X}$ allows for statistical dependencies between latent variables G_i and G_j, $i \neq j$, when they are both affected by a certain label, i.e., $G_i \leftarrow Y \to G_j$.

However, a crucial assumption of our model is that these latent factors should represent *elementary ingredients* to the causal mechanism generating $\boldsymbol{X}$ (to be defined below), which can be thought of as descriptive features of $\boldsymbol{X}$ that can be changed without affecting each other (i.e., there is no causal effect between them). A similar assumption on the underlying model is likewise a key requirement for the recent extension of identifiability results of non-linear ICA (Hyvarinen et al., 2018). We formulate this assumption of a disentangled causal model as follows (see also Figure 1):

Definition 1 (Disentangled Causal Process)**.** *Consider a causal model for $\boldsymbol{X}$ with generative factors $\boldsymbol{G}$, described by the mechanisms $p(\boldsymbol{x}|\boldsymbol{g})$, where $\boldsymbol{G}$ could generally be influenced by L confounders $\boldsymbol{C} = (C_1, \dots, C_L)$. This causal model for $\boldsymbol{X}$ is called disentangled if and only if it can be described by a structural causal model (SCM) (Pearl, 2009) of the form*

$$\begin{aligned} \boldsymbol{C} &\leftarrow \boldsymbol{N}_c \\ G_i &\leftarrow f_i(\boldsymbol{PA}_i^C, N_i), \quad \boldsymbol{PA}_i^C \subset \{C_1, \dots, C_L\},\ i = 1, \dots, K \\ \boldsymbol{X} &\leftarrow g(\boldsymbol{G}, N_x) \end{aligned}$$

with functions f_i, g and jointly independent noise variables $\boldsymbol{N}_c, N_1, \dots, N_K, N_x$. Note that $\forall i \neq j\ G_i \not\to G_j$.

In practice we assume that the dimensionality of the confounding L is significantly smaller than the number of factors K.

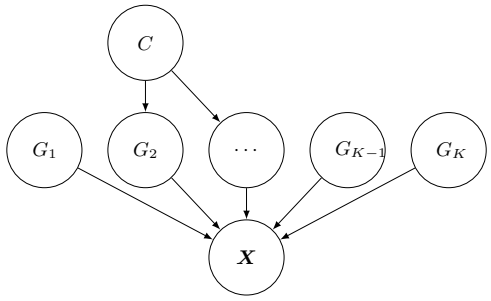

Figure 1. **Disentangled Causal Mechanism:** This graphical model encompasses our assumptions on a disentangled causal model. C stands for a confounder, $\boldsymbol{G} = (G_1, G_2, \dots, G_K)$ are the generative factors (or elementary ingredients) and $\boldsymbol{X}$ the observed quantity. In general, there can be multiple confounders affecting a range of elementary ingredients each.

This definition reflects our understanding of elementary ingredients $G_i, i = 1, \dots, K$, of the causal process. Each ingredient should work on its own and is changable without affecting others. This reflects the *independent mechanisms (IM)* assumption (Schölkopf et al., 2012). Independent mechanisms as components of causal models allow intervention on one mechanism without affecting the other modules and thus correspond to the notion of *independently controllable factors* in reinforcement learning (Thomas et al., 2017). Our setting is broader, describing any causal process and inheriting the generality of the notion of IM, pertaining to autonomy, invariance and modularity (Peters et al., 2017).

Based on this view of the data generation process, we can prove (see Appendix B) the following observations which will help us discuss notions of disentanglement and deep

latent variable models.

Proposition 1 (Properties of a Disentangled Causal Process)**.** *A disentangled causal process as introduced in Definition 1 fulfills the following properties:*

(a) $p(\boldsymbol{x}|\boldsymbol{g})$ *describes a causal mechanism invariant to changes in the distributions* $p(g_i)$.

(b) In general, the latent causes can be dependent

$$G_i \not\perp\!\!\!\perp G_j, i \neq j.$$

Only if we condition on the confounders in the data generation they are independent

$$G_i \perp\!\!\!\perp G_j | \boldsymbol{C} \quad \forall i \neq j.$$

(c) Knowing what observation of $\boldsymbol{X}$ *we obtained renders the different latent causes dependent, i.e.,*

$$G_i \not\perp\!\!\!\perp G_j | \boldsymbol{X}.$$

(d) The latent factors $\boldsymbol{G}$ *already contain all information about confounders* $\boldsymbol{C}$ *that is relevant for* $\boldsymbol{X}$*, i.e.,*

$$I(\boldsymbol{X}; \boldsymbol{G}) = I(\boldsymbol{X}; (\boldsymbol{G}, C)) \geq I(\boldsymbol{X}; C)$$

where I *denotes the mutual information.*

(e) There is no total causal effect from G_j *to* G_i *for* $j \neq i$*; i.e., intervening on* G_j *does not change* G_i*, i.e,*

$$\forall g_j^{\triangle} \quad p(g_i | do(G_j \leftarrow g_j^{\triangle})) = p(g_i) \quad \left(\neq p(g_i | g_j^{\triangle})\right)$$

(f) The remaining components of $\boldsymbol{G}$*, i.e.,* $\boldsymbol{G}_{\setminus j}$*, are a valid adjustment set (Pearl, 2009) to estimate interventional effects from* G_j *to* $\boldsymbol{X}$ *based on observational data, i.e.,*

$$p(\boldsymbol{x}|do(G_j \leftarrow g_j^{\triangle})) = \int p(\boldsymbol{x}|g_j^{\triangle}, \boldsymbol{g}_{\setminus j}) p(\boldsymbol{g}_{\setminus j})\, d\boldsymbol{g}_{\setminus j}.$$

(g) If there is no confounding, conditioning is sufficient to obtain the post interventional distribution of $\boldsymbol{X}$*:*

$$p(\boldsymbol{x}|do(G_j \leftarrow g_j^{\triangle})) = p(\boldsymbol{x}|g_j^{\triangle})$$

3.2. Disentangled Latent Variable Model

We can now understand generative models with latent variables (e.g., the decoder $p_\theta(\boldsymbol{x}|\boldsymbol{z})$ in VAEs) as models for the causal mechanism in (a) and the inferred latent space through $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ as proxy to the generative factors $\boldsymbol{G}$. Property (d) gives hope that under an adequate information bottleneck we can indeed recover information about causal parents and not the confounders. Ideally, we would hope for a one-to-one correspondance of Z_i to G_i for all $i = 1, \ldots, K$. In some situations it might be useful to learn multiple latent dimensions for one causal factor for a more natural description, e.g., describing an angle θ as $\cos(\theta)$ and $\sin(\theta)$

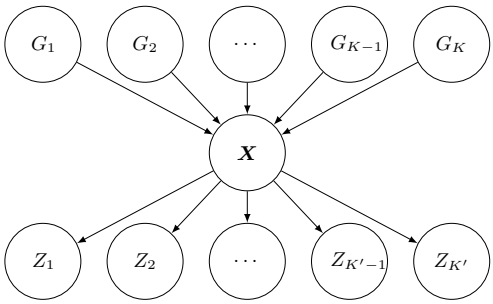

Figure 2. We assume that the data are generated by a process involving a set of unknown independent mechanisms G_i (which may themselves be confounded by other processes, see Figure 1). In the simplest case, disentangled representation learning aims to recover variables Z_i that capture the independent mechanisms G_i in the sense that they (i) represent the information contained in the G_i and (ii) respect the causal generative structure of $\boldsymbol{G} \to \boldsymbol{X}$ in an interventional sense: in particular, for any i, localized interventions on another cause G_j $(j \neq i)$ should not affect Z_i. In practice, there need not be a direct correspondence between G_i and Z_i variables (e.g., multiple latent variables may jointly represent one cause), hence our definitions deal with sets of factors rather than individual ones. Note that in the unsupervised setting, we do not know $\boldsymbol{G}$ nor the mapping from $\boldsymbol{G}$ to $\boldsymbol{X}$ (we do know, however, the "decoder" mapping from $\boldsymbol{Z}$ to $\boldsymbol{X}$, not shown in this picture). In experimental evaluations of disentanglement, however, such knowledge is usually assumed.

(Ridgeway & Mozer, 2018). Hence, we will generally allow the encodings $\boldsymbol{Z}$ to be K' dimensional, where usually $K' \geq K$. The β-VAE (Higgins et al., 2017) encourages factorization of $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ through penalization of the KL to its prior $p(\boldsymbol{z})$. Due to property (c) other approaches were introduced making use of statistical independence (Kim & Mnih, 2018; Chen et al., 2018; Kumar et al., 2018). Esmaeili et al. (2018) allow dependence within groups of variables in a hierarchical model (i.e., with some form of confounding where property (b) becomes an issue) by specifically modelling groups of dependent latent encodings. In contrast to the above mentioned approaches, this requires prior knowledge on the generative structure. We will make use of property (f) to solve the task of using observational data to evaluate deep latent variable models for disentanglement and robustness. Figure 2 illustrates our causal perspective on representation learning which encompasses the data generating process ($\boldsymbol{G} \to \boldsymbol{X}$) as well as the subsequent encoding through $E(\cdot)$ ($\boldsymbol{X} \to \boldsymbol{Z}$). Based on this viewpoint, we define the interventional effect of a *group* of generative factors $\boldsymbol{G}_J$ on the implied latent space encodings $\boldsymbol{Z}_L$ with proxy posterior $q_\phi(\boldsymbol{z}|\boldsymbol{x})$ from a VAE, where $J \subset \{1, \ldots, K\}$ and $L \subset \{1, \ldots, K'\}$ as:

$$p(\boldsymbol{z}_I | \text{do}(\boldsymbol{G}_J \leftarrow \boldsymbol{g}_J^{\triangle})) := \int q_\phi(\boldsymbol{z}_I|\boldsymbol{x})\, p(\boldsymbol{x}|\text{do}(\boldsymbol{G}_J \leftarrow \boldsymbol{g}_J^{\triangle}))\, d\boldsymbol{x}$$

This definition is consistent with the above graphical model as it implies that $p(\boldsymbol{z}_I|\boldsymbol{x}, \text{do}(\boldsymbol{G}_J \leftarrow \boldsymbol{g}_J^{\triangle})) = q_\phi(\boldsymbol{z}_I|\boldsymbol{x})$.

4. Interventional Robustness

Building on the definition of interventional effects on deep feature representations in Eq. (3.2), we now derive a robustness measure of encodings with respect to changes in certain generative factors.

Let $L \subset \{1, \dots, K'\}$ and $I, J \subset \{1, \dots, K\}, I \cap J = \emptyset$ be groups of indices in the latent space and generative space. For generality, we will henceforth talk about robustness of *groups* of features $\boldsymbol{Z}_L$ with respect to interventions on *groups* of generative factors $\boldsymbol{G}_J$. We believe that having this general formulation of allowing disagreements between groups of latent dimensions and generative factors provides more flexibility, for example when multiple latent dimensions are used to describe one phenomenon (Esmaeili et al., 2018) or when some sort of supervision is available through groupings in the dataset according to generative factors (Bouchacourt et al., 2017). Below, we will also discuss special cases of how these sets can be chosen.

If we assume that the encoding $\boldsymbol{Z}_L$ captures information about the causal factors $\boldsymbol{G}_I$ and we would like to build a predictive model that only depends on those factors, we might be interested in knowing how robust our encoding is with respect to nuisance factors $\boldsymbol{G}_J$, where $I \cap J = \emptyset$. To quantify this robustness for specific realizations of $\boldsymbol{g}_I$ and $\boldsymbol{g}_J^{\triangle}$ we make the following definition:

Definition 2 (Post Interventional Disagreement)**.** *For any given set of feature indices $L \subset \{1, \dots, K'\}$, $\boldsymbol{g}_I$ and $\boldsymbol{g}_J^{\triangle}$, we call*

$$\text{PIDA}(L|\boldsymbol{g}_I, \boldsymbol{g}_J^{\triangle}) := d\left(\mathbb{E}[\boldsymbol{Z}_L|do(\boldsymbol{G}_I \leftarrow \boldsymbol{g}_I)], \mathbb{E}[\boldsymbol{Z}_L|do(\boldsymbol{G}_I \leftarrow \boldsymbol{g}_I, \boldsymbol{G}_J \leftarrow \boldsymbol{g}_J^{\triangle})]\right)$$

*the post interventional disagreement (*PIDA*) in $\boldsymbol{Z}_L$ due to $\boldsymbol{g}_J^{\triangle}$ given $\boldsymbol{g}_I$. Here, d is a suitable distance function (e.g., ℓ_2-norm).*

The above definition on its own is likewise a contribution to the defined but unused notion of extrinsic disentanglement in Besserve et al. (2018). PIDA now quantifies the shifts in our inferred features $\boldsymbol{Z}_L$ we experience when the generative factors $\boldsymbol{G}_J$ are externally changed to $\boldsymbol{g}_J^{\triangle}$ while the generative factors that we are actually interested in capturing with $\boldsymbol{Z}_L$ (i.e., $\boldsymbol{G}_I$) remain at the predefined setting of $\boldsymbol{g}_I$. Using expected values after *intervention* on the generative factors (i.e., Pearl's do-notation), as opposed to regular conditioning, allows for interpretation of the score also when factors are dependent due to confounding. The do-notation represents setting these generative values by external intervention. It thus isolates the *causal* effect that a generative factor has, which in general is not possible using standard conditioning (Pearl, 2009). This neglects the history that might have led to the observations in the collection phase of the *observational* dataset. For example, when a robot is trained with various objects of different colors, it might be the case that certain shapes occur more often in specific colors (e.g., due to 3D printer capabilities). When we would condition the feature encoding on on a specific color, the observed effects might as well be due to a change in object shape. The interventional distribution, on the other hand, measures by definition the change features experience due to externally setting the color while all other generative factors remain the same. If there was no confounding in the generative process, this definition is equivalent to regular conditioning (see Proposition 1 (g)).

For robustness reasons, we are interested in the worst case effect any change in nuisance parameters $\boldsymbol{g}_J^{\triangle}$ might have. We call this the maximal post interventional disagreement (MPIDA): $\text{MPIDA}(L|\boldsymbol{g}_I, J) := \sup_{\boldsymbol{g}_J^{\triangle}} \text{PIDA}(L|\boldsymbol{g}_I, \boldsymbol{g}_J^{\triangle})$. This metric is still computed for a specific realization of $\boldsymbol{G}_I$. Hence, we weight this score according to occurance probabilities of $\boldsymbol{g}_I$, which leads us to the expected MPIDA: $\text{EMPIDA}(L|I, J) := \mathbb{E}_{\boldsymbol{g}_I}[\text{MPIDA}(L|\boldsymbol{g}_I, J)]$. EMPIDA is now a (unnormalized) measure in $[0, \infty)$ quantifying the worst-case shifts in the inferred $\boldsymbol{Z}_L$ we have to expect due to changes in $\boldsymbol{G}_J$ even though our generative factors of interest $\boldsymbol{G}_I$ remain the same. This is for example of interest when the robot in our introductory example learns a generic feature representation $\boldsymbol{Z}$ of his environment from which he wants to make a subselection of features $\boldsymbol{Z}_L$ in order to perform a grasping task. For this model to work well, the generative factor of the object $I = \{\texttt{shape}, \texttt{weight}\}$ are important, however, factor $J = \{\texttt{color}\}$ is not. Now, the robot can evaluate how robust its features $\boldsymbol{Z}_L$ perform at the task requiring I but not J.

We propose to normalize this quantity with $\text{EMPIDA}(L|\emptyset, \{1, \dots, K\})$, which represents the expected maximal deviation from the mean encoding of $\boldsymbol{Z}_L$ without fixed generative factors as it is often useful to have a normalized score for comparisons. Hence, we define:

Definition 3 (Interventional Robustness Score)**.**

$$\text{IRS}(L|I, J) := 1 - \frac{\text{EMPIDA}(L|I, J)}{\text{EMPIDA}(L|\emptyset, \{1, \dots, K\})} \tag{2}$$

This score yields 1.0 for perfect robustness (i.e., no harm is done by changes in $\boldsymbol{G}_J$) and 0.0 for no robustness. Note that IRS has a similar interpretion to a R^2 value in regression. Instead of measuring the captured variance, it looks at worst case deviations of inferred values.

Special Case: Disentanglement One important special case includes the setting where $L = \{l\}$, $I = \{i\}$ and $J = \{1, \dots, i-1, i+1, \dots, K\}$. This corresponds to the degree to which Z_l is robustly isolated from any extraneous causes (assuming Z_l captures G_i), which can be interpreted as the

concept of *disentanglement* in the framework of Eastwood & Williams (2018). We define

$$D_l := \max_{i \in \{1,\dots,K\}} \mathrm{IRS}(\{l\}|\{i\}, \{1,\dots,K\}\backslash\{i\}) \quad (3)$$

as disentanglement score of Z_l. The maximizing $i^\star$ is interpreted as the generative factor that Z_l captures predominantly. Intuitively, we have robust disentanglement when a feature Z_l reliably captures information about the generative factor $G_{i^\star}$, where reliable means that the inferred value is always the same when $g_{i^\star}$ stays the same, regardless of what the other generative factors $\boldsymbol{G}_{\backslash i^\star}$ are doing.

In our evaluations of disentanglement, we also plot the full dependency matrix $\hat{\boldsymbol{R}}$ with $\hat{R}_{li} = \mathrm{IRS}(\{l\}|\{i\}, \{1,\dots,K\}\backslash\{i\})$ (see for example Figure 6 on page 16) next to providing the values D_l and their weighted average.

Special Case: Domain Shift Robustness If we understand one (or multiple) generative factor(s) $\boldsymbol{G}_S$ as indicating source domains which we would like to generalize over, we can use PIDA to evaluate robustness of a selected feature set $\boldsymbol{Z}_L$ against such domain shifts. In particular,

$$\mathrm{IRS}(L|\{1,\dots,K\}\backslash\{S\}, \{S\})$$

quantifies how robust $\boldsymbol{Z}_L$ is when changes in $\boldsymbol{G}_S$ occur. If we are building a model predicting a label Y based on some (to be selected) feature set L, we can use this score to make a trade-off between robustness and predictive power. For example, we could use the best performing set of features among all those that satisfy a given robustness threshold.

Algorithm 1 EMPIDA Estimation

Input:
dataset $\mathcal{D} = \{(\boldsymbol{x}^{(i)}, \boldsymbol{g}^{(i)})\}_{i=1,\dots,N}$
trained encoder E
subsets of factors $L \subset \{1,\dots,K'\}$ and $I, J \subset \{1,\dots,K\}$
Preprocessing:
encode all samples to obtain $\{\boldsymbol{z}^{(i)} = E(\boldsymbol{x}^{(i)}) : i = 1,\dots,N\}$
estimate $p(\boldsymbol{g}^{(i)})$ and $p(\boldsymbol{g}^{(i)}_{\backslash(I\cup J)})$ $\forall i$ from relative frequencies in $\mathcal{D}$
Estimation:
find all realizations of $\boldsymbol{G}_I$ in $\mathcal{D}$: $\{\boldsymbol{g}_I^{(k)}, k = 1,\dots,N_I\}$
partition the dataset according to those realizations: $\mathcal{D}_I^{(k)} := \{(\boldsymbol{x}, \boldsymbol{g}) \in \mathcal{D} \text{ s.t. } \boldsymbol{g}_I = \boldsymbol{g}_I^{(k)}\}$
for $k = 1,\dots,N_I$ **do**
  estimate $\texttt{mean} \leftarrow \mathbb{E}[\boldsymbol{Z}_L|\mathrm{do}(\boldsymbol{G}_I \leftarrow \boldsymbol{g}_I^{(k)})]$ using Eq. (7) and samples $\mathcal{D}_I^{(k)}$
  partition $\mathcal{D}_I^{(k)}$ according to realizations of $\boldsymbol{G}_J$: $\mathcal{D}_{I,J}^{(k,l)} := \{(\boldsymbol{x}, \boldsymbol{g}) \in \mathcal{D}_I^{(k)} \text{ s.t. } \boldsymbol{g}_J = \boldsymbol{g}_J^{(l)}\}$
  initialize $\texttt{mpida}(k) \leftarrow 0.0$
  for $l = 1,\dots,N_{I,J}^{(k)}$ **do**
    $\texttt{mean}_{\texttt{int}} \leftarrow \mathbb{E}[\boldsymbol{Z}_L|\mathrm{do}(\boldsymbol{G}_I \leftarrow \boldsymbol{g}_I^{(k)}, \boldsymbol{G}_J \leftarrow \boldsymbol{g}_J^{(l)})]$ using Eq. (7) and samples $\mathcal{D}_{I,J}^{(k,l)}$ for estimation
    compute $\texttt{pida} \leftarrow d(\texttt{mean}, \texttt{mean}_{\texttt{int}})$
    update $\texttt{mpida}(k) \leftarrow \max(\texttt{mpida}(k), \texttt{pida})$
  end for
end for
Return $\texttt{empida} \leftarrow \sum_{k=1}^{N_I} \frac{|\mathcal{D}_I^{(k)}|}{|\mathcal{D}|} \texttt{mpida}(k)$

5. Estimation and Benchmarking Disentanglement

In the supplementary material A we provide the derivation of our estimation procedure for $\mathrm{EMPIDA}(L|I,J)$. Here we only present the specific algorithm how EMPIDA can be estimated from a generic observational dataset $\mathcal{D}$ in Algorithm 1. The main ingredient for this estimation to work is provided by our constrained causal model (i.e., a disentangled process) that implies that the backdoor criteria can be applied, which we showed in Proposition 1.

Even though the sampling procedure might look non-trivial at first sight, the algorithm 1 for estimating $\mathrm{EMPIDA}(L|I,J)$ has $\mathcal{O}(N)$ complexity as indicated by the following result:

Proposition 2 (Computational Complexity)**.** *The* EMPIDA *estimation algorithm described in Algorithm 1 scales* $\mathcal{O}(N)$ *in the dataset size* $N = |\mathcal{D}|$.

The proof of Proposition 2 can be found in Appendix C. Note that a dataset capturing all possible variations generally grows exponentially in the number of generative factors. While this is a general issue for all validation approaches and care needs to be taken when collecting such datasets in practice, we just remind that due to the generally large nature of N it is particularly important to have such an efficient validation procedure. In many benchmark datasets for disentanglement (e.g. dsprites) the observations are obtained noise-free and the dataset contains all possible combinations of generative factors exactly once. This makes the estimation of the disentanglement score even easier, as we have $|\mathcal{D}^{(k,l)}_{I=\{i\},J=\{1,\dots,K\}\backslash\{i\}}| = 1$. Furthermore, since no confounding is present, we can use conditioning to estimate the interventional effect, i.e., $p(\boldsymbol{x}|\mathrm{do}(G_i \leftarrow g_i)) = p(\boldsymbol{x}|g_i)$, as seen in Proposition 1 (g). The disentanglement score of Z_l, as discussed in Eq. (3) , follows (see A.1 for details) as:

$$D_l = \max_{i \in \{1,\dots,K\}} \left(1 - \frac{\mathrm{EMPIDA}_{li}}{\sup_{\tilde{\boldsymbol{x}} \in \mathcal{D}} d\left(\mathbb{E}[Z_l], E(\tilde{\boldsymbol{x}})\right)}\right).$$

6. Experiments

Our evaluations involve five different state of the art unsupervised disentanglement techniques (classic VAE, β-VAE, DIP-VAE, FactorVAE and β-TCVAE), each learning 10 features.

6.1. Methods Comparison

In Table 1 we provide a compact summary of our evaluation. Our objective is the analysis of various kinds of learned latent spaces and their characteristics, not primarily evaluating which methods work best under some metric. In particular, we used each method with the parameter settings that were indicated in the original publications (details are given in Appendix D) and did not tune them further in order to achieve a better robustness score, which is certainly feasible. Rather, we are interested in evaluating latent spaces as a whole, which encompasses both the method and its settings in combination. We can for example observe that β-TCVAE achieves a relatively low feature importance based measure by Eastwood & Williams (2018). This is due to the fact that Chen et al. (2018) did not consider shape to be a generative factor in their tuning (which also leads to a lower informativeness score in our evaluation that includes this factor), and also because their model ends up with only few active dimensions. The treatment of such inactive components can make a difference when averaging disentanglement scores of the single Z_i to an overall score. FI uses a simple average, MI weights the components with their overall feature importance and we weight them according to worst case deviation from mean (i.e., normalization of the IRS).

Table 1. **Metrics Overview:** IRS: (ours), FI: (Eastwood & Williams, 2018), MI: (Ridgeway & Mozer, 2018), INFO: informativeness score (Eastwood & Williams, 2018) (higher is better). The number in parentheses indicates the rank according to a particular metric. Experimental details are given in Section D.

Model	IRS	FI	MI	Info
VAE	0.33 (5)	0.23 (4)	0.90 (3)	0.82 (1)
Annealed β-VAE	0.57 (2)	0.35 (2)	0.86 (5)	0.79 (4)
DIP-VAE	0.43 (4)	0.39 (1)	0.89 (4)	0.82 (1)
FactorVAE	0.51 (3)	0.31 (3)	0.92 (1)	0.79 (4)
β-TCVAE	0.72 (1)	0.16 (5)	0.92 (1)	0.74 (5)

Believing that it is most insightful to look at scores for each dimension separately, which indicates the quality of a single feature, we included the full evaluations including plots of correspondance matrices (as in Figure 6) in Appendix E. For future extensions and applications our work is added to the `disentanglement_lib` of Locatello et al. (2018).

6.2. Robustness as Complementary Metric

As we could already see in Table 1, different metrics do not always agree with each other about which model disentangles best. This is consistent with the recent large scale evaluation provided by Locatello et al. (2018). In Figure 3 we further illustrate the dependency between MI score and our IRS on the finer granularity of considering the metrics of individual features (instead of the full latent space). There seems to be a clear positive correlation between the two evaluation metrics. However, there are features classified as well disentangled according to MI, but not robustly (according to IRS). These features are marked with the red rectangle in Figure 3. We explore one typical such example in more detail in Figures 4, 6 and 7 in the appendix, for the case of the DIP model.

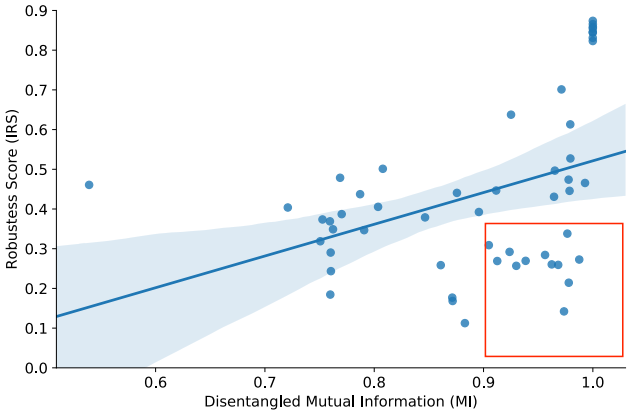

Figure 3. **Relationship Metrics:** Visualization of all learned features Z_i in our universe (5 models with 10 dimensions each) based on their MI disentanglement score on the x axis and interventional robustness (IRS) on the y axis. The red box indicates the features that obtained a high disentanglement score according to mutual information (i.e., they share high mutual information with only one generative factor), but still provide low robustness according to IRS. These are the cases where the robustness perspective delivers additional insight into disentanglement quality.

When there are rare events happening that still have a major impact on the features or when there is a cumulative effect from several generative factors (e.g., in Figure 4), pairwise information based methods (such as MI or FI) cannot capture this vulnerability of deeply learned features. IRS, on the other hand, looks specifically at these cases. For a well rounded view on disentanglement quality, we propose to use both types of measures in a manner that is complementary and use-case specific. Specificially when critical applications are designed on top of deep representations quantifying its robustness can be decisive.

6.3. Visualising Interventional Robustness

We further introduce a new visualization technique for latent space models based on ground truth factors which is

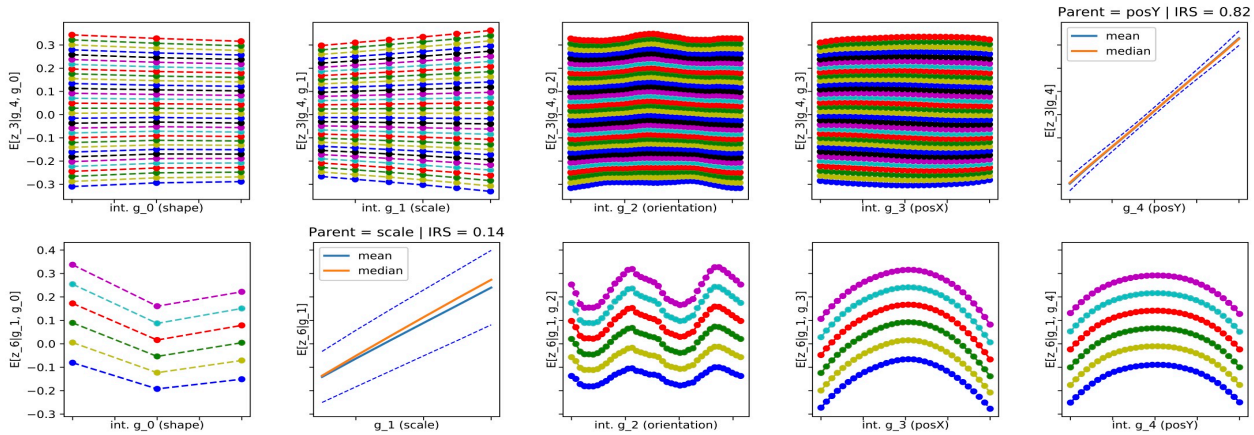

Figure 4. **Visualising Interventional Robustness:** Plots of $\mathbb{E}[Z_l|g_i*, \text{do}(G_j \leftarrow g_j^{\triangle})]$ as a function of $g_j^{\triangle}$ for different G_j per column as explained in Section 6.3. The upper row is an example of good, robust disentanglement (Z_3 from the DIP model discussed in Figure 6). The lower row illustrates Z_6 which is classified as well disentangled according to FI (top 18%) and MI (top 33%) but still has a low robustness score (bottom 4%). This stems from the fact that even though Z_6 is very informative about scale (almost a linear function in expectation), its value can still be changed remarkably by switching any of `posX`, `posY` or `orientation`. These additional dependencies are not discovered by mutual information (or feature importance) due to the higher noise in these relationships (see Figure 7) and because they are partly hidden in cumulated effects.

motivated by interventional robustness and illustrates how robust a learned feature is with respect to changes in nuisance factors. Figure 4 illustrates this approach on two features learned by the DIP model. Each row corresponds to a different feature Z_l. The upper row corresponds to a well disentangled and robust feature (Z_3) which gets classified as such by all three metrics. The lower (Z_6) also obtaines a high FI and MI score, however, IRS correctly discovers that this feature is not robust. This illustrates a case where having a robustness perspective on disentanglement is important. The columns correspond to different generative factors G_j (`shape`, `scale`, `orientation`, `posX`, `posY`) which potentially influence Z_l. For each latent variable Z_l we first find the generative factor G_{i^*} which is most related to it by choosing the maximizer of Eq. (3) (i.e., the factor that renders Z_l most invariant). In the column i^* we then plot the estimate of $\mathbb{E}[Z_l|g_{i^*}]$ together with its confidence bound in order to visualize the informativeness of Z_l about G_{i^*}. For example the upper row in plot 4 corresponds to Z_3 in the DIP model and mostly relates to `posY`. This is why we plot the dependence of Z_3 on `posY` in the fifth column. The remaining columns then illustrate how Z_3 changes when interventions on the other generative factor are made, even though `posY` is being kept at a fixed value. Each line with different color corresponds to a particular value `posY` can take on. More generally speaking, we plot in the jth column $\mathbb{E}[Z_l|g_i*, \text{do}(G_j \leftarrow g_j^{\triangle})]$ as a function of $g_j^{\triangle}$ for all possible realizations g_i* of G_i*. All values with constant g_i* are connected with a line. For a robustly disentangled feature, we would expect all of these colored lines to be horizontal (i.e., there is no more dependency on any G_j after accounting for G_i*). As such visualizations can provide a much more in depth understanding of learned representations than single numbers, we provide the full plots of various models in the appendix F.

7. Conclusion

We have proposed a framework for assessing disentanglement in deep representation learning which combines the generative process responsible for high dimensional observations with the subsequent feature encoding by a neural network. This perspective leads to a natural validation method, the *interventional robustness score*. We show how it can be estimated from observational data using an efficient algorithm that scales linearly in the dataset size. As special cases, this proposed measure captures robust disentanglement and domain shift stability. Extensive evaluations showed that the existing metrics do not capture the effects that rare events or cumulative influences from multiple generative factors can have on feature encodings, while our robustness based validation metric discovers such vulnerabilities.

We envision that the notion of interventional effects on encodings may give rise to the development of novel, robustly disentangled representation *learning* algorithms, for example in the interactive learning environment (Thomas et al., 2017) or when weak forms of supervision are available (Bouchacourt et al., 2017; Locatello et al., 2019). The exploration of those ideas, especially including confounding, is left for future research.

Acknowledgments

We thank Andreas Krause for helpful discussions and support. This research was partially supported by the Max Planck ETH Center for Learning Systems.

References

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8): 1798–1828, 2013.

Besserve, M., Sun, R., and Schölkopf, B. Counterfactuals uncover the modular structure of deep generative models. *arXiv preprint arXiv:1812.03253*, 2018.

Bouchacourt, D., Tomioka, R., and Nowozin, S. Multi-level variational autoencoder: Learning disentangled representations from grouped observations. *arXiv preprint arXiv:1705.08841*, 2017.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. Understanding disentangling in β-vae. *arXiv preprint arXiv:1804.03599*, 2018.

Chen, T. Q., Li, X., Grosse, R., and Duvenaud, D. Isolating sources of disentanglement in variational autoencoders. *arXiv preprint arXiv:1802.04942*, 2018.

Cheung, B., Livezey, J. A., Bansal, A. K., and Olshausen, B. A. Discovering hidden factors of variation in deep networks. *arXiv preprint arXiv:1412.6583*, 2014.

Comon, P. Independent component analysis, a new concept? *Signal processing*, 36(3):287–314, 1994.

Eastwood, C. and Williams, C. K. I. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

Esmaeili, B., Wu, H., Jain, S., Narayanaswamy, S., Paige, B., and Van de Meent, J.-W. Hierarchical disentangled representations. *arXiv preprint arXiv:1804.02086*, 2018.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. beta-vae: Learning basic visual concepts with a constrained variational framework. 2017.

Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., and Lerchner, A. Towards a definition of disentangled representations. *arXiv preprint arXiv:1812.02230*, 2018.

Hyvarinen, A., Sasaki, H., and Turner, R. E. Nonlinear ICA using auxiliary variables and generalized contrastive learning. *arXiv preprint arXiv:1805.08651*, 2018.

Kim, H. and Mnih, A. Disentangling by factorising. In *International Conference on Machine Learning*, pp. 2649–2658, 2018.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

Koller, D., Friedman, N., and Bach, F. *Probabilistic graphical models: principles and techniques*. MIT Press, 2009.

Kumar, A., Sattigeri, P., and Balakrishnan, A. Variational inference of disentangled latent concepts from unlabeled observations. In *International Conference on Learning Representations*, 2018.

Liu, Y.-C., Yeh, Y.-Y., Fu, T.-C., Chiu, W.-C., Wang, S.-D., and Wang, Y.-C. F. Detach and adapt: Learning cross-domain disentangled deep representation. *arXiv preprint arXiv:1705.01314*, 2017.

Locatello, F., Bauer, S., Lucic, M., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. *arXiv preprint arXiv:1811.12359*, 2018.

Locatello, F., Tschannen, M., Bauer, S., Rätsch, G., Schölkopf, B., and Bachem, O. Disentangling factors of variation using few labels. *arXiv preprint arXiv:1905.01258*, 2019.

Mathieu, M. F., Zhao, J. J., Zhao, J., Ramesh, A., Sprechmann, P., and LeCun, Y. Disentangling factors of variation in deep representation using adversarial training. In *Neural Information Processing Systems*, pp. 5040–5048, 2016.

Pearl, J. *Causality*. Cambridge University Press, 2009.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. MIT Press, 2017.

Ridgeway, K. and Mozer, M. C. Learning deep disentangled embeddings with the f-statistic loss. *arXiv preprint arXiv:1802.05312*, 2018.

Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 2018.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. On Causal and Anticausal Learning. In *Proceedings of the 29th ICML*, pp. 1255–1262, New York, NY, USA, 2012. Omnipress.

Siddharth, N., Paige, B., van de Meent, J.-W., Desmaison, A., Wood, F. D., Goodman, N. D., Kohli, P., and Torr, P. H. S. Learning disentangled representations with semi-supervised deep generative models. In *Neural Information Processing Systems*, 2017.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, prediction, and search*. Springer-Verlag. (2nd edition MIT Press 2000), 1993.

Thomas, V., Pondard, J., Bengio, E., Sarfati, M., Beaudoin, P., Meurs, M.-J., Pineau, J., Precup, D., and Bengio, Y. Independently controllable features. *arXiv preprint arXiv:1708.01289*, 2017.