When are Post-hoc Conceptual Explanations Identifiable?

Tobias Leemann[1,2,†] **Michael Kirchhof**[1,†] **Yao Rong**[1,2] **Enkelejda Kasneci**[2] **Gjergji Kasneci**[2]

[1]University of Tübingen, Tübingen, Germany
[2]Technical University of Munich, Munich, Germany
[†]equal contribution

Abstract

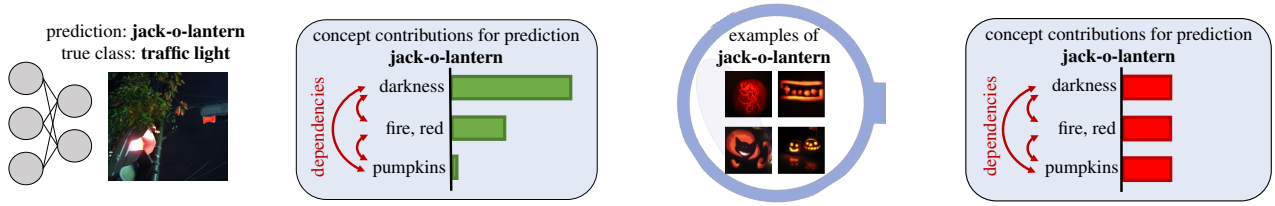
Interest in understanding and factorizing learned embedding spaces through conceptual explanations is steadily growing. When no human concept labels are available, concept discovery methods search trained embedding spaces for interpretable concepts like *object shape* or *color* that can provide post-hoc explanations for decisions. Unlike previous work, we argue that concept discovery should be *identifiable*, meaning that a number of known concepts can be provably recovered to guarantee reliability of the explanations. As a starting point, we explicitly make the connection between concept discovery and classical methods like Principal Component Analysis and Independent Component Analysis by showing that they can recover independent concepts under non-Gaussian distributions. For dependent concepts, we propose two novel approaches that exploit functional compositionality properties of image-generating processes. Our provably identifiable concept discovery methods substantially outperform competitors on a battery of experiments including hundreds of trained models and dependent concepts, where they exhibit up to 29 % better alignment with the ground truth. Our results highlight the strict conditions under which reliable concept discovery without human labels can be guaranteed and provide a formal foundation for the domain. Our code is available online.

1 INTRODUCTION

Modern computer vision systems represent images in embedding spaces. These are either constructed implicitly in higher-level layers of large models or explicitly through generative models such as Variational Autoencoders (Kingma and Welling, 2013) or recent Diffusion Models (Song and Ermon, 2019; Ho et al., 2020). To unveil why an image is considered similar to a certain class, interest in understanding these embeddings is increasing. Conceptual explanations (Crabbé and van der Schaar, 2022; Muttenthaler et al., 2022; Akula et al., 2020; Kazhdan et al., 2020; Yeh et al., 2019; Kim et al., 2018) are a popular explainable AI (XAI) technique for this purpose. They scrutinize a given encoder by decomposing its embedding space into interpretable concepts post-hoc, i.e., after training. Subsequently, these concepts form the basis of popular post-hoc explanations such as TCAV (Kim et al., 2018) or allow high-level interventions (Koh et al., 2020). Fig. 1 outlines a real-world example. A misclassification made by a pretrained model shipped with the `pytorch` library (Paszke et al., 2017) is to be explained. In the given example, the conceptual explanation allows identification of a spurious correlation that the model has picked up: Most jack-o-lanterns are found in combination with dark backgrounds, which causes it to mistake the traffic light at night for a jack-o-lantern.

Constructing such explanations is non-trivial. The key ingredient to all conceptual explanation techniques is a set of interpretable concepts, which is notoriously hard to specify (Leemann et al., 2022). It is frequently defined through human annotations (Crabbé and van der Schaar, 2022; Koh et al., 2020; Kim et al., 2018) on individual samples of the dataset that can be prohibitively expensive (Kazhdan et al., 2021). Furthermore, it is usually unknown which concepts will be leveraged by a machine learning model without a model at hand. Therefore, we consider fully unsupervised concept discovery (Ghorbani et al., 2019; Yeh et al., 2019), where the concepts are automatically discovered in the data. Concepts are frequently modeled as directions in a given embedding space (Ghorbani et al., 2019; Kim et al., 2018; Yeh et al., 2019), which have to be discovered without supervision. These embedding spaces can be highly distorted, making it hard to correctly separate the influences of individual concepts. However, this is essential to make the right inferences in practice (see Fig. 1d). This intuition is supported by prior work on generative models (Ross et al.,

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).



(a) **Misclassification:** A model makes an incorrect prediction. A user is interested in understanding why this incident happened.

(b) **Conceptual Explanation:** Concept contributions are computed that explain the prediction. In this example, the concept "darkness" is relevant for the outcome.

(c) **Inspection:** A closer inspection of samples from the predicted class reveals that most images in this class have a dark background; a spurious correlation picked up by the model.

(d) **Entangled Conceptual Explanation:** It is essential to correctly split up the contribution of individual concepts to allow for valid inferences.

Figure 1: Schematical use-case of conceptual explanations: A misclassification of an image classifier is explained. The example is based on a real explanation for a ResNet50 model. Details and the original explanation are provided in App. C.8.

2021), which has shown that user understanding is strongly linked to the representations' respective disentanglement.

While many methods have been empirically shown to work well, a rigorous theoretical analysis of the conditions under which concept discovery is possible is still lacking in previous works. We propose to consider concept discovery methods that are *identifiable*. This means when a known number of *ground truth components* generated the data, the concept discovery method provably yields concepts that correspond to the individual ground truth components and can correctly represent an input in the concept space. This is a crucial requirement: If a method is even incapable of recovering known components, there is no indication for its reliability in practice. In this work, we are the first to investigate identifiability results in the context of post-hoc concept discovery.

First, we find that identifiability results from Principal Component Analysis (PCA) and Independent Component Analysis (ICA) literature (Jolliffe, 2002; Comon, 1994; Hyvärinen et al., 2001) can be transferred to the conceptual explanation setup. We establish that they cover the case of independent ground truth components with non-Gaussian distributions. This is insufficient for two reasons: (1) In practice, concepts such as height and weight (Träuble et al., 2021) or wing and head colors of birds often follow complex dependency patterns. (2) Popular generative models (Kingma and Welling, 2013; Song and Ermon, 2019) frequently work with an embedding space with a Gaussian distribution.

As a second contribution, we seek to fill this void by providing an identifiable concept discovery approach that can handle dependent and Gaussian ground truth components. We can show that this is possible through taking the nature of the image-generating process into consideration. Specifically, we propose utilizing *visual compositionality properties*. These are based on the observation that tiny changes in the components frequently affect input images in orthogonal or even disjoint ways. These properties of image-generating processes also leave a "trace" in the encoders learned from a set of data samples. This insightful finding permits to construct two novel post-hoc concept discovery methods based on the *disjoint* or *independent mechanisms* criterion. We prove strong identifiability guarantees for recovering components, even if they are dependent. Our results highlight the strict and nuanced conditions under which identifiable concept discovery is possible.

In summary, our work advances current literature in multiple ways: **(1)** We present first identifiability results for post-hoc conceptual explanations. We find that results from ICA can be transferred under the assumption of independent ground truth components. **(2)** For the more intricate setting of dependent components, we propose the *disjoint mechanism analysis (DMA)* criterion and the less constrained *independent mechanism analysis (IMA)* criterion. We prove that they recover even dependent original components up to permutation and scale. **(3)** We construct DMA and IMA-based concept discovery algorithms for encoder embedding spaces with the same theoretical identifiability guarantees. **(4)** We test them (i) on embeddings of several autoencoder models learned from correlated data, (ii) with multiple and strong correlations, (iii) on discriminative encoders, and (iv) on the real-world CUB-200-2011 dataset (Wah et al., 2011). Our approaches maintain superior performance amidst increasingly severe challenges.

2 RELATED WORK

Works on the analysis and interpretation of embedding spaces touch a variety of subfields of machine learning.

Concept discovery for explainable AI. Conceptual explanations (Koh et al., 2020; Kim et al., 2018; Ghorbani et al., 2019; Yeh et al., 2019; Akula et al., 2020; Chen et al., 2020b) have gained popularity within the XAI community. They aim to explain a trained machine learning model post-hoc in terms of human-friendly, high-level concept directions

(Kim et al., 2018). These concepts are found via supervised (Koh et al., 2020; Kim and Mnih, 2018; Kazhdan et al., 2020) or unsupervised approaches (Yeh et al., 2019; Akula et al., 2020; Ren et al., 2022), such as clustering of embeddings (Ghorbani et al., 2019). However, their results are not always meaningful (Leemann et al., 2022; Yeh et al., 2019). Therefore, we suggest approaches with identifiability guarantees. We provide initial identifiability results and a novel approach, which can be used for unsupervised concept discovery under correlated components.

Independent Component Analysis (ICA). Independent Component Analysis (Comon, 1994; Hyvärinen and Pajunen, 1999; Hyvärinen et al., 2001) or blind source separation (BSS) consider a generative process $\boldsymbol{g}(\boldsymbol{z})$ as a mixture to undo and rely on traces that the distribution of the generating components $\boldsymbol{z}$ leaves in the mixture. In this work, we show that an identifiability result from ICA can be transferred to the conceptual explanation setup, but recovery is only possible under independent underlying components of which all but one are non-Gaussian. This result is not applicable to naturally correlated processes, which is why we design a novel method for this case.

Disentanglement Learning. Concurrently, literature on disentanglement learning is concerned with finding a data-generating mechanism $\boldsymbol{g}(\boldsymbol{z})$ and a latent representation $\boldsymbol{z}$ for a dataset, such that each of the original components (also known as factors of variation) is mapped to one (controllable) unit direction in $\boldsymbol{z}$ (Bengio et al., 2013). An alternative definition relies on group theory (Higgins et al., 2017) where certain group operations (symmetries) should be reflected in the learned representation (Painter et al., 2020; Yang et al., 2021). Most works in the domain enhance VAEs (Kingma and Welling, 2013) with additional loss terms (Higgins et al., 2017; Burgess et al., 2018; Kim and Mnih, 2018; Chen et al., 2018). Despite recent progress it is not always possible to construct disentangled embedding spaces from scratch: Locatello et al. (2019) have shown that the problem is inherently unidentifiable without additional assumptions. A more recent work by Träuble et al. (2021) shows that even if just two components of a dataset are correlated, current disentanglement learning methods fail. In this work, we focus on post-hoc explanations of embedding spaces of given models, which are usually entangled.

Identifiability results. Identifiability questions have been raised in domains such as Natural Language Processing (Carrington et al., 2019) or in disentanglement learning, which is most related to this work. It has been previously shown that unsupervised disentanglement, without further conditions, is impossible (Hyvärinen and Pajunen, 1999; Locatello et al., 2019; Moran et al., 2022). Hence, recent works aim to understand the conditions sufficient for identifiability. One strain of work relies on additional supervision, i.e., access to an additional observed variable (Hyvärinen et al., 2019; Khemakhem et al., 2020) or to tuples of observations that differ in only a limited number of components (Locatello et al., 2020). Gresele et al. (2021) and Zheng et al. (2022) proved identifiable disentanglement under independently distributed components and introduce a functional condition on the data generator. We also consider functional properties, but our setting is different as (1) we have access to a trained encoder only and (2) not even partial annotations or relations are available.

3 ANALYSIS

In this section, we formalize post-hoc concept discovery to provide an identifiability perspective. We find that Independent Component Analysis (ICA) and Principal Component Analysis (PCA) only guarantee identifiability when the ground-truth components are stochastically independent. We then study the intricate case of dependent components and propose using *disjoint* and *independent mechanisms analysis* (DMA / IMA) along with identifiability results. All proofs are provided in App. B.

3.1 PROBLEM FORMALIZATION

In post-hoc concept discovery, we are given a trained encoder $\boldsymbol{f} : \mathcal{X} \to \mathcal{E}$ with embeddings $\boldsymbol{e} = \boldsymbol{f}(\boldsymbol{x}) \in \mathcal{E} \subset \mathbb{R}^K$ of each image $\boldsymbol{x} \in \mathcal{X}$. We do not impose any restriction on how $\boldsymbol{f}$ was obtained; it can be the feature extractor part of a large classification model, or a feature representation learned through autoencoding, contrastive learning (Chen et al., 2020a), or related techniques. Interpretability literature seeks to understand the embedding space by factorizing it into concepts. Based on the observations that directions in the embedding space often correspond to meaningful features (Szegedy et al., 2013; Bau et al., 2017; Alain and Bengio, 2016; Bisazza and Tump, 2018), these concepts are frequently defined as direction vectors $\boldsymbol{m}_i$ (Kim et al., 2018; Ghorbani et al., 2019; Yeh et al., 2019). These are commonly referred to as concept activation vectors (CAVs). Hence, the combined output of a concept discovery algorithm is a matrix $\boldsymbol{M} = [\boldsymbol{m}_1, \dots, \boldsymbol{m}_K]^\top \in \mathbb{R}^{K \times K}$ where each row contains a concept direction.

We seek a theoretical guarantee on when these discovered concept directions align with ground truth components that generated the data. To this end, we formalize the data-generating process as shown in Fig. 2: There are K ground-truth components with scores $z_k, k = 1 \dots K$, summarized $\boldsymbol{z} \in \mathcal{Z} \subset \mathbb{R}^K$, that define an image. The term *components* always refers to the ground truth as opposed to the *concepts*, which denote the discovered directions. A data-generating process $\boldsymbol{g} : \mathcal{Z} \to \mathcal{X}$ generates images $\boldsymbol{x} = \boldsymbol{g}(\boldsymbol{z}) \in \mathcal{X} \subset \mathbb{R}^L$, $L \gg K$. A powerful algorithm should be able to recover the original components. That is, there should be a one-to-one mapping between entries of $\boldsymbol{M}\boldsymbol{e}$ and the entries in $\boldsymbol{z}$, up to the arbitrary scale and order

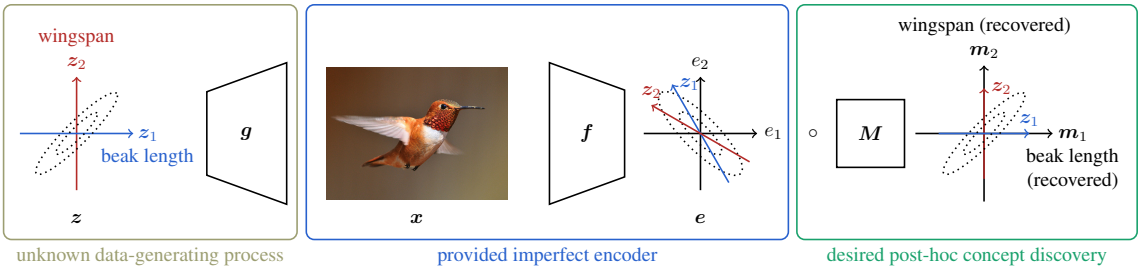


Figure 2: Overview over the concept discovery setup. We consider a process where data samples $\boldsymbol{x}$ are generated from possibly correlated ground truth components $\boldsymbol{z}$, e.g., a wingspan or beak length of a bird, by an unknown process $\boldsymbol{g}$ (left). The high-dimensional data is mapped to the to the embedding space of a given model $\boldsymbol{f}$ (center). A suitable post-hoc concept discovery yields concept vectors $\boldsymbol{m}_i$ that correspond to the original components (right).

of the entries. We say that a concept discovery algorithm *identifies* the true components if it is guaranteed to output directions $\boldsymbol{M}$ that satisfy $\boldsymbol{M}\boldsymbol{e} = \boldsymbol{M}\boldsymbol{f}(\boldsymbol{g}(\boldsymbol{z})) = \boldsymbol{P}\boldsymbol{S}\boldsymbol{z}$ $\forall \boldsymbol{z} \in \mathcal{Z}$, where $\boldsymbol{P} \in \mathbb{R}^{K\times K}$ is a permutation matrix that has one 1 per row and column and is 0 otherwise, and $\boldsymbol{S} \in \mathbb{R}^{K\times K}$ is an invertible diagonal scaling matrix.

To make the problem solvable in the first place, concept directions must exist in the embedding space of the given encoder, requiring $\boldsymbol{e} = \boldsymbol{D}\boldsymbol{z}$, where $\boldsymbol{D} \in \mathbb{R}^{K\times K}$ is of full rank. Depending on the scope of the conceptual explanation desired, it can be sufficient for the components to exist in a local region of the embedding space if the concept discovery algorithm is only applied around a region around a certain point of interest. This only changes the meaning of $\mathcal{E}, \mathcal{X}$, and $\mathcal{Z}$ but is formally equivalent.

3.2 IDENTIFIABILITY VIA INDEPENDENCE

Initially, we turn towards classical component analysis methods. We find that their identifiability results use non-correlation or even stronger stochastic independence assumptions of the ground truth components.

Principal Component Analysis (PCA) (Jolliffe, 2002) uses eigenvector decompositions to find orthogonal directions $\boldsymbol{M}$ that result in uncorrelated components $\boldsymbol{M}\boldsymbol{e}$. This means that PCA is only capable of identifying the original components if the ground truth components $\boldsymbol{z}$ were uncorrelated and exist as orthogonal directions in our embedding space. In our setup and notation, this leads to the following result:

Theorem 3.1 (PCA identifiability) *Let $z_k, k = 1, \dots, K$, be uncorrelated random variables with non-zero and unequal variances. Let $\boldsymbol{e} = \boldsymbol{D}\boldsymbol{z}$, where $\boldsymbol{D} \in \mathbb{R}^{K\times K}$ is an orthonormal matrix. If an orthonormal post-hoc transformation $\boldsymbol{M} \in \mathbb{R}^{K\times K}$ results in mutually uncorrelated components $(z'_1, \dots, z'_K) = \boldsymbol{z}' = \boldsymbol{M}\boldsymbol{e}$, then $\boldsymbol{M}\boldsymbol{e} = \boldsymbol{P}\boldsymbol{S}\boldsymbol{z}$, where $\boldsymbol{P} \in \mathbb{R}^{K\times K}$ is a permutation and $\boldsymbol{S} \in \mathbb{R}^{K\times K}$ is a diagonal matrix where $|s_{ii}| = 1$ for $i \in 1, \dots K$.*[1]

All proofs in this work are deferred to App. B. It is arguably a strong condition that the ground truth directions are encoded orthogonally in the embedding space. Independent Component Analysis (ICA) overcomes this limitation and allows for arbitrary directions. However, the classic result by Comon (1994) even demands stochastically independent components. Transferred to our setup and notation, the result can be stated as follows.

Theorem 3.2 (ICA identifiability) *Let $z_k, k = 1, \dots, K$, be independent random variables with non-zero variances where at most one component is Gaussian. Let $\boldsymbol{e} = \boldsymbol{D}\boldsymbol{z}$, where $\boldsymbol{D} \in \mathbb{R}^{K\times K}$ has full rank. If a post-hoc transformation $\boldsymbol{M} \in \mathbb{R}^{N\times N}$ results in mutually independent components $(z'_1, \dots, z'_K) = \boldsymbol{z}' = \boldsymbol{M}\boldsymbol{e}$, then $\boldsymbol{M}\boldsymbol{e} = \boldsymbol{P}\boldsymbol{S}\boldsymbol{z}$, where $\boldsymbol{P} \in \mathbb{R}^{K\times K}$ is a perm. and $\boldsymbol{S} \in \mathbb{R}^{K\times K}$ is a diag. matrix.*

This result shows that stochastic independence of the ground truth components leaves a strong trace in the embeddings that can be leveraged. Algorithms like `fastICA` (Hyvärinen and Oja, 1997) can find the concept directions $\boldsymbol{M}$ by searching for independence (Comon, 1994). We conclude that ICA is suited for post-hoc concept discovery under independent components.

In summary, we have transferred two results from the component analysis literature to the setup of post-hoc conceptual explanations. However, these results do not allow to recover components that are correlated or follow a Gaussian distribution. This limits their applicability in practice where concepts often appear pairwise (e.g., darkness and jack-o-lanterns, cf. Fig. 1). We will bridge this gap in the remainder of this paper by introducing two new identifiable discovery methods based on functional properties of the generation process that we term *disjoint* and *independent* mechanisms. A summary of identifiability results is provided in Tab. 1.

[1]To simplify notation, $\boldsymbol{P}$ and $\boldsymbol{S}$ mean *any* permutation and scale matrices. They do not have to be equal between the theorems.

Dependency	Marginal Dist.	Transform	Criterion
uncorr.	uneq. variances	orthogonal	non-correlation (PCA)
independent	non-Gaussian	invertible	independence (ICA)
arbitrary	arbitrary	invertible	disj. mechanisms (DMA)
arbitrary	arbitrary	invertible	indep. mechanisms (IMA)

Table 1: PCA and ICA provably identify concepts via their distributions. DMA and IMA utilize functional properties.

3.3 IDENTIFIABILITY VIA DISJOINT MECHANISMS

Instead of placing independence assumptions on $\boldsymbol{z}$, we propose a concept discovery algorithm that makes use of natural properties of the generative process $\boldsymbol{g}$. In particular, generative processes in vision are often compositional (Ommer and Buhmann, 2007): Different groups of pixels in an image, like a bird's wings, legs, and head, are each controlled by different components. Effects of tiny changes in components are visible in the Jacobian $\boldsymbol{J_g}$, where each row points to the pixels affected. Thus, a compositional process will follow the *disjoint mechanisms* principle.

Definition 3.1 (Disjoint mechanism analysis (DMA)) *$\boldsymbol{g}$ is said to generate $\boldsymbol{x}$ from its components $\boldsymbol{z}$ via disjoint mechanisms if the Jacobian $\boldsymbol{J_g}(\boldsymbol{z}) \in \mathbb{R}^{L\times K}$ exists and is a block matrix $\forall \boldsymbol{z} \in \mathcal{Z}$. That is, the columns of $\boldsymbol{J_g}(\boldsymbol{z})$ are non-zero at disjoint rows, i.e. $|\boldsymbol{J_g}(\boldsymbol{z})|^\top |\boldsymbol{J_g}(\boldsymbol{z})| = \boldsymbol{S}(\boldsymbol{z})$, where $\boldsymbol{S} \in \mathbb{R}^{K\times K}$ is a diagonal matrix that may be different for each $\boldsymbol{z}$ and $|\cdot|$ takes the element-wise absolute value.*

Note that this definition does not globally constrain the location of affected pixels. The components may still alter different but disjoint pixels for each image. In real concept discovery, we do not have access to the generative process $\boldsymbol{g}$ but can only access the encoder $\boldsymbol{f}$. However, an encoder corresponding to $\boldsymbol{g}$ will not be arbitrary and its Jacobian $\boldsymbol{J}_f \in \mathbb{R}^{K\times L}$ will have a distinct form in practice: First, to maintain the component information the composition $\boldsymbol{f} \circ \boldsymbol{g}$ will be of the form $\boldsymbol{f}(\boldsymbol{g}(\boldsymbol{z})) = \boldsymbol{D}\boldsymbol{z}$, with a yet unknown matrix $\boldsymbol{D} \in \mathbb{R}^{K\times K}$. Furthermore, we expect encoders to be rather lazy, meaning they only perform the changes to invert the data generation process but are almost invariant to input deviations not due to changes in the components. This is in line with the classic interpretability literature, where gradients of models were observed to noisily highlight the relevant input features (Baehrens et al., 2010; Simonyan et al., 2013) and form the basis of popular attribution methods such as Integrated Gradients (Sundararajan et al., 2017). Technically, the changes effected by the components form the linear $\text{span}(\boldsymbol{J_g}(\boldsymbol{z}))$, whereas entirely external changes are given in its orthogonal complement $\text{span}(\boldsymbol{J_g}(\boldsymbol{z}))^\perp$. Thus, for $\mathbf{v} \in \text{span}(\boldsymbol{J_g}(\boldsymbol{z}))^\perp \subset \mathbb{R}^L$ the encoder should not react to these change and the corresponding gradients of the encoder for these changes should be zero, i.e., $\boldsymbol{J_f}(\boldsymbol{g}(\boldsymbol{z}))\mathbf{v} = \mathbf{0} \Leftrightarrow \mathbf{v} \in \ker(\boldsymbol{J_f}(\boldsymbol{g}(\boldsymbol{z})))$.

Definition 3.2 (Faithful encoder) *$\boldsymbol{f}$ is a faithful encoder for the generative process $\boldsymbol{g}$ if the ground truth components remain recoverable, i.e., $\boldsymbol{f}(\boldsymbol{g}(\boldsymbol{z})) = \boldsymbol{D}\boldsymbol{z}$, for some $\boldsymbol{D} \in \mathbb{R}^{K\times K}$ with full rank. Furthermore, $\boldsymbol{f}$ is lazy and invariant to changes in $\boldsymbol{x}$ which cannot be explained by the ground truth components, requiring $\boldsymbol{J_f}(\boldsymbol{g}(\boldsymbol{z}))$ and $\boldsymbol{J_g}(\boldsymbol{z})$ to exist and $\text{span}(\boldsymbol{J_g}(\boldsymbol{z}))^\perp \subseteq \ker(\boldsymbol{J_f}(\boldsymbol{g}(\boldsymbol{z})))$, $\forall \boldsymbol{z} \in \mathcal{Z}$.*

Having defined what realistic encoders look like through the notion of faithful encoders, we find that there is distinct property which can be leveraged to discover the directions in $\boldsymbol{M}$ among faithful encoders: It is sufficient to find an encoder $\boldsymbol{M}\boldsymbol{f}$ whose Jacobian $\boldsymbol{M}\boldsymbol{J}_f$ will have disjoint rows. Intuitively, this means searching for components whose gradients affect disjoint image regions.

Theorem 3.3 (Identifiability under DMA) *Let $\boldsymbol{g}$ have disjoint mechanisms and $\boldsymbol{f}$ be a faithful encoder to $\boldsymbol{g}$. If a post-hoc transformation $\boldsymbol{M} \in \mathbb{R}^{K\times K}$ of full rank results in disjoint rows in the Jacobian $\boldsymbol{M}\boldsymbol{J_f}(\boldsymbol{g}(\boldsymbol{z}))$, i.e., $|\boldsymbol{M}\boldsymbol{J_f}(\boldsymbol{g}(\boldsymbol{z}))||\boldsymbol{M}\boldsymbol{J_f}(\boldsymbol{g}(\boldsymbol{z}))|^\top$ is invertible and diagonal for some $\boldsymbol{z} \in \mathcal{Z}$, then $\boldsymbol{M}\boldsymbol{e} = \boldsymbol{P}\boldsymbol{S}\boldsymbol{z}$ where $\boldsymbol{P} \in \mathbb{R}^{K\times K}$ is a permutation and $\boldsymbol{S} \in \mathbb{R}^{K\times K}$ is a scaling matrix.*

This theorem does not impose any restrictions on the distribution $\boldsymbol{z}$, making it applicable to realistic concept discovery scenarios through leveraging the nature of the generative process. The proof of this algorithm in App. B.5 also yields an analytical solution. We will use it to verify conditions in a controlled experiment in Sec. 4.1. We have thus identified the *DMA criterion* that is sufficient to discover the component directions when the rows of $\boldsymbol{M}\boldsymbol{J_f}$ point to disjoint image regions. We can formulate this as a loss function and optimize for $\boldsymbol{M}$ via off-the-shelf gradient descent:

$$\mathcal{L}(\boldsymbol{M}) = \mathbb{E}_{\boldsymbol{x}} \|\text{arn}\left[\boldsymbol{M}\boldsymbol{J}_f(\boldsymbol{x})\right] \text{arn}\left[\boldsymbol{M}\boldsymbol{J}_f(\boldsymbol{x})\right]^\top - \boldsymbol{I}\|_F^2. \tag{1}$$

The expectation is taken over a collection of real data samples $\boldsymbol{x} = \boldsymbol{g}(\boldsymbol{z})$. The arn-operator (absoute values, row normalization) takes the element-wise absolute value and subsequently normalizes the rows. This does not constrain the norms of the Jacobian's rows but only enforces disjointness.

3.4 CONCEPT DISCOVERY VIA INDEPENDENT MECHANISMS

We can perform an analogous derivation for a class of generating processes that is more general. Grounded by causal principles instead of compositionality, the independent mechanisms property has been argued to define a class of natural generators (Gresele et al., 2021).

Definition 3.3 (Independent mechanism analysis (IMA)) *$\boldsymbol{g}$ is said to generate $\boldsymbol{x}$ from its components $\boldsymbol{z}$ via independent mechanisms if the Jacobian $\boldsymbol{J_g}(\boldsymbol{z})$ of $\boldsymbol{g}$ exists and its columns (one per component) are orthogonal $\forall \boldsymbol{z} \in \mathcal{Z}$, i.e., $\boldsymbol{J_g}^\top(\boldsymbol{z})\boldsymbol{J_g}(\boldsymbol{z}) = \boldsymbol{S}(\boldsymbol{z})$, where $\boldsymbol{S} \in \mathbb{R}^{K\times K}$ is a diagonal matrix that may differ for each $\boldsymbol{z}$ (Gresele et al., 2021).*

Gresele et al. (2021) and Zheng et al. (2022) used this characteristic to find disentangled data generators, but we can again transfer characteristics via faithful encoders: This time we find that searching for an $\boldsymbol{M}\boldsymbol{J}_f$ with *orthogonal* (instead of disjoint) rows permits post-hoc discovery of concepts. We refer to is property of $\boldsymbol{M}\boldsymbol{J}_f$ as the *IMA criterion*.

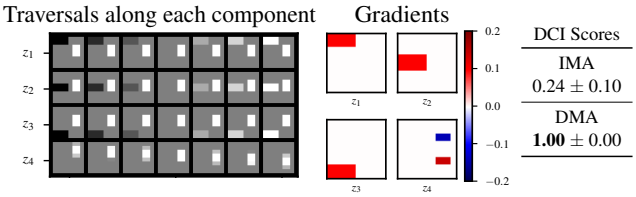
However, as the class of admissible processes has been increased, it is not strong enough to ensure identifiability in the most general case. This is prevented under an additional technical condition on the component magnitudes, which we refer to as *non-equal magnitude ratios* (NEMR). Intuitively, the magnitudes of the component gradients have to change non-uniformly between at least two points for the conditions to be sufficient. If there were two factors that always attribute to input pixels in the same way (imagine the sky being partitioned into two components termed "left sky" and "right sky"), they cannot be told apart anymore since there can be other mixtures which would result in orthogonality (they could equally be "lower sky" and "upper sky").

Theorem 3.4 (Identifiability under IMA) *Let $\boldsymbol{g}$ adhere to IMA. Let $\boldsymbol{f}$ be a faithful encoder to $\boldsymbol{g}$. Suppose we have obtained an $\boldsymbol{f}' = \boldsymbol{M}\boldsymbol{f}$ with a full-rank $\boldsymbol{M} \in \mathbb{R}^{K\times K}$ and orthogonal rows in its Jacobian $\boldsymbol{M}\boldsymbol{J_f}(\boldsymbol{g}(\boldsymbol{z})) := \boldsymbol{J}_{f'}(\boldsymbol{g}(\boldsymbol{z}))$, i.e, $\boldsymbol{J}_{f'}(\boldsymbol{g}(\boldsymbol{z}))\boldsymbol{J}_{f'}(\boldsymbol{g}(\boldsymbol{z}))^\top = \boldsymbol{\Sigma}(\boldsymbol{z})$ where $\boldsymbol{\Sigma}(\boldsymbol{z})$ is diagonal and full-rank at two points $\boldsymbol{z} \in \{\boldsymbol{z}_a, \boldsymbol{z}_b\}$. If additionally $\boldsymbol{\Sigma}(\boldsymbol{z}_a)\boldsymbol{\Sigma}(\boldsymbol{z}_b)^{-1}$ has unequal entries in its diagonal (NEMR condition), then $\boldsymbol{M}\boldsymbol{e} = \boldsymbol{P}\boldsymbol{S}\boldsymbol{z}$, where $\boldsymbol{P} \in \mathbb{R}^{K\times K}$ is a permutation and $\boldsymbol{S} \in \mathbb{R}^{K\times K}$ is a scaling matrix.*

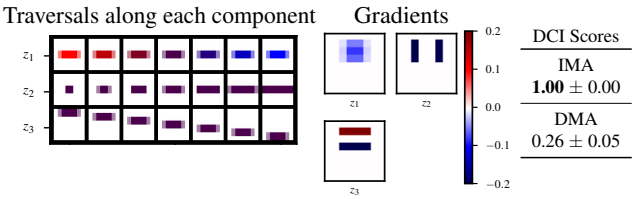
The constructive proof in App. B.6 can also be condensed into an analytical solution. Alternatively, one can again construct a suitable optimization objective for the IMA criterion, i.e., orthogonal Jacobians. This is achieved by removing the absolute value operation from the arn-operator in Eqn. (1), so that it solely performs a row-wise normalization. In summary, we have established the novel DMA and IMA criteria that allow concept discovery under dependent components.

4 EXPERIMENTS

In the following, we perform a battery of experiments of increasing complexity to compare the practical capabilities of approaches for identifiable concept discovery. We start by verifying the theoretical identifiability conditions (Sec. 4.1), then perform evaluation under increasing multi-component correlations for embedding spaces of generative and discriminative models (Sec. 4.2 to 4.4), and finally use a large-scale, discriminatively-trained ResNet50 encoder (Sec. 4.5).



(a) FourBars: DMA datasets can be solved by the DMA criterion.



(b) ColorBar: IMA datasets can be solved by the IMA criterion.

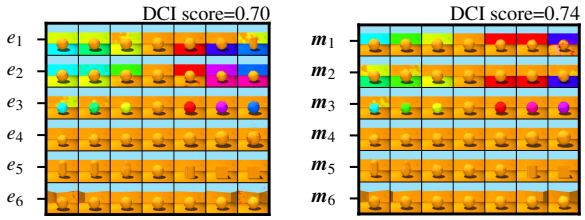
Figure 3: Experiments on two synthetic datasets: We confirm our analytical results and show that DMA (a) and IMA (b) cover visual concepts such as colors and translations.

We borrow the DCI metric (Eastwood and Williams, 2018) from disentanglement learning with scores in $[0, 1]$ to measure whether each discovered component predicts precisely one ground-truth component and vice versa. Following Locatello et al. (2020), we report additional metrics with similar results in App. D, along with results on additional datasets and ablations. For reproducibility, each experiment is repeated on five seeds and code is made available upon acceptance. In total, we train and analyze over 300 embedding spaces, requiring about 124 Nvidia RTX2080Ti GPU days. More implementation details are in App. C.

4.1 CONFIRMING IDENTIFIABILITY

We first confirm our identifiability guarantees with the analytical solutions. To this end, we implement two realistic synthetic datasets with differentiable generators. This allows computing the closed form of $\boldsymbol{J_g}$ and deliberately fulfilling or violating the DMA, IMA, and NEMR conditions.

FourBars consists of gray-scale images of four components: Three bars change their colors (black to white) and one bar moves vertically, showing that the image regions affected by each component may change in each image. The plot of $\boldsymbol{J_g}$ in Fig. 3a shows that each component maps to a disjoint image region. This fulfills DMA and thus also IMA. However, all factors have the same gradient magnitudes, making it impossible to find two points with NEMR. According to our theory, we expect DMA optimization to work and IMA to fail as NEMR is essential to make to proof of Thm. 3.4. The second dataset, ColorBar, contains a



(a) Autoencoder (DipVAE) (b) Autoencoder + DMA (ours)

Figure 4: DMA discovers directions $\boldsymbol{m}$ that control individual concepts (wall & floor color) of Shapes3D although they are confused in the original embedding space ($e_1, e_2, \ldots$).

single bar that undergoes realistic changes in color, width, and its vertical position, see Fig. 3b. It conforms to IMA and NEMR but not DMA. Our proofs indicate that IMA should work, and DMA should fail. Completing the problem formalization in Sec. 3.1, we compute analytical faithful encoders $\boldsymbol{f}$ for these datasets distorted by a random matrix $\boldsymbol{D}$. The solutions behave as expected: On `FourBars` only the DMA criterion delivers perfectly recovered components (DCI=1) whereas on `ColorBars` only IMA succeeds.

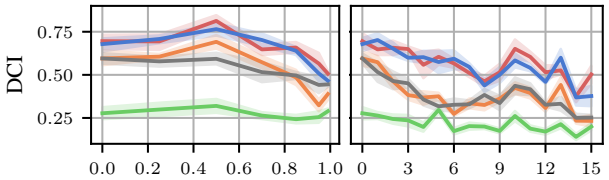
4.2 CORRELATED COMPONENTS

We now move to the common Shapes3D (Burgess and Kim, 2018) dataset. It shows geometric bodies that vary in their colors, shape, orientation, size, and background totaling six components. Compared to the previous section we train real encoders. We start our analysis where disentanglement learning is no longer possible: When components are correlated. Following Träuble et al. (2021), the dataset is resampled such that two components $z_i, z_j \in [0, 1]$ follow $z_i - z_j \sim \mathcal{N}(0, s^2)$. Lower s results in a stronger correlation where only few pairs of component values co-occur frequently. We choose a moderate correlation of $s = 0.4$ here and three pairs z_i, z_j that are nominal/nominal, nominal/ordinal, and ordinal/ordinal variables. We train four state-of-the-art disentanglement learning VAEs (BetaVAE (Higgins et al., 2017), FactorVAE (Kim and Mnih, 2018), BetaTCVAE (Chen et al., 2018), DipVAE (Kumar et al., 2018)) from a recent study (Locatello et al., 2019) and apply ICA, PCA, and our DMA and IMA discovery methods on their embedding spaces to post-hoc recover the original components. For DMA and IMA, we use the optimization-based algorithms (Eqn. 1) since they find approximate solutions through aggregation of many noisy sample gradients.

Tab. 2 shows the resulting DCI scores. In line with Träuble et al. (2021), we find that the disentanglement learning VAEs fail to recover the correlated components on their own due to their violated stochastic independence assumption (Fig. 4a). In eleven of the twelve model/correlation pairs, DMA or IMA identify better concepts than the VAE unit axes and the PCA/ICA components with improvements of

Correlated components	floor & background		orientation & background		orientation & size	
BetaVAE	0.497 ± 0.03		0.581 ± 0.04		0.491 ± 0.05	
+PCA	0.263 ± 0.03	-47%	0.310 ± 0.02	-47%	0.324 ± 0.04	-34%
+ICA	0.574 ± 0.04	+16%	0.540 ± 0.08	-7%	0.577 ± 0.04	+17%
+Ours (IMA)	0.617 ± 0.02	+24%	0.602 ± 0.05	+3%	0.579 ± 0.03	+18%
+Ours (DMA)	$\mathbf{0.641 \pm 0.03}$	**+29%**	$\mathbf{0.624 \pm 0.06}$	**+7%**	$\mathbf{0.627 \pm 0.03}$	**+28%**
FactorVAE	0.507 ± 0.11		0.502 ± 0.08		$\mathbf{0.712 \pm 0.01}$	
+PCA	0.358 ± 0.07	-29%	0.474 ± 0.05	-5%	0.556 ± 0.03	-22%
+ICA	0.294 ± 0.07	-42%	0.263 ± 0.05	-48%	0.340 ± 0.03	-52%
+Ours (IMA)	0.551 ± 0.04	+9%	0.498 ± 0.03	-1%	0.595 ± 0.05	-16%
+Ours (DMA)	$\mathbf{0.584 \pm 0.05}$	**+15%**	$\mathbf{0.510 \pm 0.05}$	**+2%**	0.556 ± 0.04	-22%
BetaTCVAE	0.619 ± 0.01		0.613 ± 0.04		0.659 ± 0.01	
+PCA	0.400 ± 0.03	-35%	0.421 ± 0.07	-31%	0.450 ± 0.07	-32%
+ICA	0.540 ± 0.02	-13%	0.497 ± 0.04	-19%	0.627 ± 0.02	-5%
+Ours (IMA)	0.623 ± 0.02	+1%	0.652 ± 0.03	+6%	0.638 ± 0.04	-3%
+Ours (DMA)	$\mathbf{0.666 \pm 0.01}$	**+8%**	$\mathbf{0.664 \pm 0.02}$	**+8%**	$\mathbf{0.748 \pm 0.03}$	**+14%**
DipVAE	0.631 ± 0.02		0.652 ± 0.02		0.548 ± 0.04	
+PCA	0.158 ± 0.01	-75%	0.160 ± 0.02	-75%	0.170 ± 0.02	-69%
+ICA	0.630 ± 0.02	-0%	0.651 ± 0.02	-0%	0.542 ± 0.03	-1%
+Ours (IMA)	0.644 ± 0.02	+2%	0.624 ± 0.01	-4%	0.558 ± 0.05	+2%
+Ours (DMA)	$\mathbf{0.684 \pm 0.01}$	**+8%**	$\mathbf{0.679 \pm 0.01}$	**+4%**	$\mathbf{0.601 \pm 0.05}$	**+10%**

Table 2: DMA recovers the components best in 11 out of 12 cases across different models and correlated components of Shapes3D. Mean $\pm$ std. err. of DCI across all components.



(a) Correlation strength (b) Number pairwise corr.

Figure 5: DMA and IMA recover the components even under strong and multiple correlations between them. ICA and PCA fail to return better components than the unit axes.

up to 29 %. This experiment shows that their concept discovery works regardless of (1) the model type and (2) the type of components correlated. On average, DMA delivers better results than IMA (+0.047), despite the generative process of Shapes3D only being roughly IMA or DMA-compliant. We therefore hypothesize that the DMA criterion might be more robustly optimizable in practice. Fig. 4b visualizes the performance achieved via DMA when traversing the embedding space. It also shows that small DCI differences can mean a significant improvement. This is because (1) the metric is computed across all six components and the strong baselines already identify many concepts and (2) a perfect score of 1.0 is usually not possible due to non-linearly encoded components. We investigate other correlation strengths with similar findings in App. D.3.

4.3 GAUSSIANITY AND MULTIPLE CORRELATIONS

In this section, we increase the distributional challenges to analyze whether our approaches are as distribution-agnostic

Method	$s = 0.1$	$s = 0.15$	$s = 0.2$	$s = \infty$
unit dirs.	0.238 ± 0.01	0.244 ± 0.01	0.247 ± 0.01	0.286 ± 0.02
PCA	0.238 ± 0.01	0.376 ± 0.03	0.373 ± 0.03	0.343 ± 0.03
ICA	0.409 ± 0.02	0.309 ± 0.02	0.311 ± 0.01	$\mathbf{0.652 \pm 0.00}$
(Ours) IMA	0.295 ± 0.01	0.302 ± 0.01	0.333 ± 0.04	0.266 ± 0.12
(Ours) DMA	$\mathbf{0.435 \pm 0.01}$	$\mathbf{0.411 \pm 0.03}$	$\mathbf{0.392 \pm 0.02}$	0.369 ± 0.05

Table 3: Without correlations ($s = \infty$), ICA is able to recover the components of a classification model. Under correlations, DMA works best. Mean $\pm$ std. err. of DCI.

as intended. We sample the components of Shapes3D from a (rotationally symmetric) Gaussian. Additionally, we introduce correlations between multiple components to its covariance matrix. Details on how covariance matrices are constructed are given in App. C.3.

First, we study a single pair of correlated components (floor and background color) with increasing correlation strength ρ. Fig. 5a shows that the BetaVAE handles low correlations well but starts deteriorating from a strength of $\rho > 0.5$, along with ICA. The DCI of our methods is an average constant of $+0.145$ above the BetaVAE's for $\rho \leq 0.85$. After this, it returns to the underlying BetaVAE's DCI, possibly because the two components collapsed in the BetaVAE's embedding space. For Fig. 5b, we gradually add more moderately correlated ($\rho \approx 0.7$) pairs to the Gaussian's covariance matrix until eventually all components are correlated. Again, our models show a constant benefit over the underlying BetaVAE's DCI curve. This experiment highlights that both DMA and IMA perform well with (1) strong and (2) multiple correlations and (3) Gaussian components.

4.4 DISCRIMINATIVE EMBEDDING SPACES

We highlight that our approach is also applicable to classification models that were trained in a purely discriminative manner, e.g., the feature space of a CNN model. To investigate this setting, we set up an 8-class classification problem on the Shapes3D dataset, where the combination of the four binarized components object color, wall color (blue/red vs. yellow/green), shape (cylinder vs. cube) and orientation (left vs. right) determines the class as visualized in App. C.4. To make the setting even more realistic, we artificially add labeling noise close to the decision boundary, correlations as in Sec. 4.2, and a small L2-regularizer on the embeddings to keep them in a reasonable range. We train a discriminative CNN with a K=6-dimensional embedding space.

The discriminative loss leads to a clustered distribution in the embedding space. ICA expectedly works very well in this highly non-Gaussian distribution, when no significant correlations are present which is in line with the result in Thm. 3.2. However, tables turn as we increasingly correlate the floor and background color: Starting at $s = 0.2$, DMA outperforms ICA and the other methods as can be seen in Tab. 3. While IMA leads to better concepts over the unit directions, it does not reach the level of DMA. We note that both ICA and our methods improve again for very strong correlations, where the setup approaches the case of three independent components (the other two components being treated as one) that is easier again. Overall, this demonstrates that our methods are applicable to purely discriminative embedding spaces and are more robust to high levels of correlations than ICA.

4.5 REAL-WORLD CONCEPT DISCOVERY

Last, we go beyond the traditional benchmarks and perform realistic concept discovery: We analyze the embedding space of a ResNet50 classifier (He et al., 2016) trained on the CUB-200-2011 (Wah et al., 2011) dataset consisting of high-resolution images of birds. This amplifies the challenges of the previous sections, i.e., a discriminative space, non-linear component dependencies of varying strengths across multiple components, and a large 512-dimensional embedding space. One restriction of this experiment is that CUB has no data-generating components to compare against, so we cannot report DCI scores. However, we qualitatively show that DMA can deliver interpretable concepts by matching them to annotated attributes of CUB.

We apply DMA and IMA to discover K=30 concepts of which the first two DMA concepts are shown exemplarily in Fig. 6. The images with the highest positive scores on the first component (on the right) consistently show white birds. The other end of the component comprises birds whose primary color is black. This gives a high Spearman rank correlation with the CUB attribute "primary color: white". The second concept is similarly interpretable. To quantify this across all K components, we provide an initial quantitative evaluation based on the Spearman rank correlation between components and attributes in App. D.7. It indicates that ICA and PCA have problems providing such components and the components identified by DMA usually correspond more closely to the attributes. The concepts provided by our method also compare favorably to those identified by ACE (Ghorbani et al., 2019) and ConceptSHAP (Yeh et al., 2019). While the construction of further quantitative evaluation schemes goes beyond the scope of this work, these promising results highlight that DMA also works for high-dimensional, real-world datasets.

5 DISCUSSION

We conclude by discussing the limitations of this work and related approaches and provide constructive guidance on which approach to choose in practice.

Limitations. In order to overcome distributional assumptions, our approach requires other forms of constraints. Most notably, we suppose that the generative processes comply

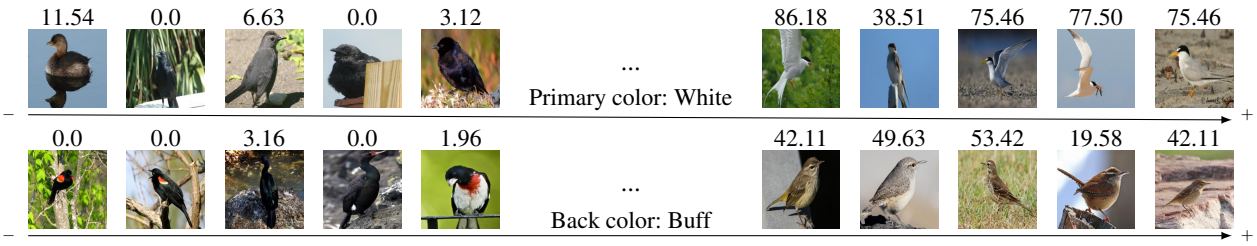


Figure 6: Components discovered by DMA on CUB correlate with interpretable ground truth attributes. Images are ordered by their concept scores $(\boldsymbol{M}\boldsymbol{e})_i$, and the numbers show their ground truth annotated attribute score.

with the functional properties of Disjoint or Independent Mechanisms. While they are intuitive and our empirical results suggest that they are a useful approximation of real-world images, we acknowledge that these requirements are not strictly fulfilled in most practical scenarios and the quality of the results depends on the extent to which these constraints are violated. We investigate the robustness of our methods to violations of the assumptions in App. D.5. Compared to the classical methods such as PCA or IMA, the gradient-based optimization requires additional resources. However, the runtime strongly depends on hyperparameters such as the number of optimization steps. We also show that improved results can still be obtained time budget comparable to that of PCA and IMA in App. D.5.

Choosing the right approach for concept discovery. Overall, our results show that unsupervised conceptual explanations with guarantees are only possible under specific sets of working assumptions. In this paragraph, we would like to briefly summarize them and give constructive suggestions on which approaches are best used when.

- PCA works with uncorrelated components that are orthogonally encoded. We believe that the assumption of an orthogonal encoding is rather unlikely in practice, even if non-correlation was possible.
- ICA works well under independently distributed components but fails under dependent components. We suggest using this method when there is evidence that the ground truth components are independent.
- DMA does not require independence, but instead requires a disjoint mechanisms process and a faithful encoder to this process. This assumption is particularly suitable for image-generating processes.
- IMA does not require independence as well, but requires a faithful encoder to an independent mechanisms process. The class of independent mechanism processes is larger and may also cover non-image processes (Gresele et al., 2021). However, it requires the additional NEMR condition. We further empirically observed that the objective derived from IMA is harder to optimize for with SGD optimizers.
- Other approaches like ConceptSHAP (Yeh et al., 2019) and ACE (Ghorbani et al., 2019) also come with certain restrictions: ACE requires a model that is scale and shift invariant, while ConceptSHAP is specifically designed for computer vision models with spatial feature maps such as ResNet (He et al., 2016). Further, these approaches come without formal guarantees.

6 CONCLUSION

Summary. We proposed identifiability as a minimal requirement for concept discovery algorithms. Furthermore, we suggested the two functional paradigms of disjoint and independent mechanisms and proved that they can recover known components in visual embedding spaces. Extensive experiments confirmed that they offer substantial improvements on various generative and discriminative models and remain unaffected by distributional challenges.

Outlook. We believe our work to be a valuable step towards a rigorous formalization of concept discovery. However, the considered setup can be generalized in the future, for instance to components that are not linearly encoded. This would permit even stronger guarantees. While we have taken a technical perspective here, future work is required to investigate the effect of improved concepts on upstream explanations.

Acknowledgements

The authors thank Frederik Träuble, Luigi Gresele, and Julius von Kügelgen for insightful discussions during early development of this project. This work was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC number 2064/1 – Project number 390727645. We thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Michael Kirchhof.

References

Arjun Akula, Shuai Wang, and Song-Chun Zhu. Cocox: Generating conceptual and counterfactual explanations

via fault-lines. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 2594–2601, 2020.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

David Baehrens, Timon Schroeter, Stefan Harmeling, Motoaki Kawanabe, Katja Hansen, and Klaus-Robert Müller. How to explain individual classification decisions. *The Journal of Machine Learning Research*, 11:1803–1831, 2010.

David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6541–6549, 2017.

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

Arianna Bisazza and Clara Tump. The lazy encoder: A fine-grained analysis of the role of morphology in neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2871–2876. Association for Computational Linguistics, 2018.

Chris Burgess and Hyunjik Kim. 3d shapes dataset. https://github.com/deepmind/3dshapes-dataset/, 2018.

Christopher P Burgess, Irina Higgins, Arka Pal, Loic Matthey, Nick Watters, Guillaume Desjardins, and Alexander Lerchner. Understanding disentangling in β-VAE. *arXiv preprint arXiv:1804.03599*, 2018.

Rachel Carrington, Karthik Bharath, and Simon Preston. Invariance and identifiability issues for word embeddings. *Advances in Neural Information Processing Systems*, 32, 2019.

Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020a.

Zhi Chen, Yijie Bei, and Cynthia Rudin. Concept whitening for interpretable image recognition. *Nature Machine Intelligence*, 2(12):772–782, 2020b.

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

Jonathan Crabbé and Mihaela van der Schaar. Concept activation regions: A generalized framework for concept-based explanations. In *Advances in Neural Information Processing Systems*, 2022.

Cian Eastwood and Christopher KI Williams. A framework for the quantitative evaluation of disentangled representations. In *International Conference on Learning Representations*, 2018.

Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. Towards automatic concept-based explanations. In *Advances in Neural Information Processing Systems*, volume 32, pages 9277–9286, 2019.

Luigi Gresele, Julius von Kügelgen, Vincent Stimper, Bernhard Schölkopf, and Michel Besserve. Independent mechanism analysis, a new concept? In *Advances in Neural Information Processing Systems*, 2021.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. Beta-VAE: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations*, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Aapo Hyvärinen and Erkki Oja. A fast fixed-point algorithm for independent component analysis. *Neural computation*, 9(7):1483–1492, 1997.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent component Analysis*. John Wiley & Sons, Inc, 2001.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999. ISSN 0893-6080.

Aapo Hyvärinen, Hiroaki Sasaki, and Richard Turner. Nonlinear ica using auxiliary variables and generalized contrastive learning. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 859–868. PMLR, 2019.

Ian T Jolliffe. *Principal component analysis*. Springer, 2nd edition, 2002.

Dmitry Kazhdan, Botty Dimanov, Mateja Jamnik, Pietro Liò, and Adrian Weller. Now you see me (cme): concept-based model extraction. *AIMLAI workshop at the 29th ACM International Conference on Information and Knowledge Management (CIKM)*, 2020.

Dmitry Kazhdan, Botty Dimanov, Helena Andres Terre, Mateja Jamnik, Pietro Liò, and Adrian Weller. Is disentanglement all you need? comparing concept-based & disentanglement approaches. *RAI, WeaSul, and RobustML workshops at The Ninth International Conference on Learning Representations 2021*, 2021.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *International Conference on Artificial Intelligence and Statistics*, pages 2207–2217. PMLR, 2020.

Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning*, pages 2668–2677. PMLR, 2018.

Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International Conference on Machine Learning*, pages 2649–2658. PMLR, 2018.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International Conference on Machine Learning*, pages 5338–5348. PMLR, 2020.

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

Tobias Leemann, Yao Rong, Stefan Kraft, Enkelejda Kasneci, and Gjergji Kasneci. Coherence evaluation of visual concepts with objects and language. In *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality*, 2022.

Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Raetsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *International Conference on Machine Learning*, pages 4114–4124. PMLR, 2019.

Francesco Locatello, Ben Poole, Gunnar Rätsch, Bernhard Schölkopf, Olivier Bachem, and Michael Tschannen. Weakly-supervised disentanglement without compromises. In *International Conference on Machine Learning*, pages 6348–6359. PMLR, 2020.

Gemma Elyse Moran, Dhanya Sridhar, Yixin Wang, and David Blei. Identifiable deep generative models via sparse decoding. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=vd0onGWZbE.

Lukas Muttenthaler, Charles Yang Zheng, Patrick McClure, Robert A. Vandermeulen, Martin N Hebart, and Francisco Pereira. VICE: Variational interpretable concept embeddings. In *Advances in Neural Information Processing Systems*, 2022.

Bjorn Ommer and Joachim M Buhmann. Learning the compositional nature of visual objects. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

Matthew Painter, Adam Prugel-Bennett, and Jonathon Hare. Linear disentangled representations and unsupervised action estimation. In *Advances in Neural Information Processing Systems*, volume 33, pages 13297–13307. Curran Associates, Inc., 2020.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Xuanchi Ren, Tao Yang, Yuwang Wang, and Wenjun Zeng. Learning disentangled representation by exploiting pretrained generative models: A contrastive learning view. In *International Conference on Learning Representations*, 2022.

Andrew Ross, Nina Chen, Elisa Zhao Hang, Elena L Glassman, and Finale Doshi-Velez. Evaluating the interpretability of generative models by interactive reconstruction. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–15, 2021.

Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pages 3319–3328. PMLR, 2017.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Frederik Träuble, Elliot Creager, Niki Kilbertus, Francesco Locatello, Andrea Dittadi, Anirudh Goyal, Bernhard Schölkopf, and Stefan Bauer. On disentangled representations learned from correlated data. In *International Conference on Machine Learning*, pages 10401–10412. PMLR, 2021.

C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. The Caltech-UCSD Birds-200-2011 Dataset. Technical Report CNS-TR-2011-001, 2011.

Tao Yang, Xuanchi Ren, Yuwang Wang, Wenjun Zeng, and Nanning Zheng. Towards building a group-based unsupervised representation disentanglement framework. In *International Conference on Learning Representations*, 2021.

Chih-Kuan Yeh, Been Kim, Sercan O Arik, Chun-Liang Li, Tomas Pfister, and Pradeep Ravikumar. On completeness-aware concept-based explanations in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yujia Zheng, Ignavier Ng, and Kun Zhang. On the identifiability of nonlinear ICA with unconditional priors. In *ICLR2022 Workshop on the Elements of Reasoning: Objects, Structure and Causality*, 2022.