Desiderata for Representation Learning from Identifiability, Disentanglement, and Group-Structuredness

Hamza Keurti [*123] **Patrik Reizinger** [*1456] **Bernhard Schölkopf** [1] **Wieland Brendel** [1]

Abstract

Machine learning subfields define *useful* representations differently: disentanglement strives for semantic meaning and symmetries, identifiability for recovering the ground-truth factors of the (unobservable) data-generating process, group-structured representations for symmetries We demonstrate that despite their merits, each approach has shortcomings. Surprisingly, joining forces helps overcome the limitations: we use insights from latent space statistics, geometry, and topology in our examples to elucidate how combining the desiderata of identifiability, disentanglement, and group structure yields more useful representations.

1. Introduction

Representation learning (Bengio et al., 2013) is in pursuit of a *useful* representation. However, usefulness depends on the (downstream) task and is generally ambiguous to define. latent variable models (Bishop, 2006; Murphy, 2012) rely on the latent manifold hypothesis (Bengio et al., 2013)—i.e., high-dimensional samples such as images belong to a low-dimensional manifold—to extract low-dimensional (latent) factors that sufficiently describe the data. Different machine learning approaches such as disentanglement (Cohen & Welling, 2014; Higgins et al., 2018), identifiability (Hyvärinen & Pajunen, 1999), and group-structured representations (Bronstein et al., 2021) impose distinct inductive biases on the learned representation. **Disentanglement** aims to uncover *semantically meaningful* latent factors. Intuitively, a disentangled representation should encode different object properties (e.g., color, size) into different latent factors. In practice, a representation is deemed disentangled if it excels w.r.t. a disentanglement metric, e.g., DCI (Eastwood & Williams, 2018) or MIG (Chen et al., 2018)—cf. (Locatello et al., 2019) for a comparison. However, there is no uniquely-accepted disentanglement definition. Other works try to learn a representation with a **group structure** (Cohen & Welling, 2014; Bronstein et al., 2021); this approach also inspired researchers to define disentanglement from a group-based perspective (Higgins et al., 2018). **Identifiability** constructs model classes that provably reconstruct the underlying latents (up to indeterminacies, such as permutations, or element-wise transformations). This is impossible without additional assumptions (Hyvärinen & Pajunen, 1999).

Our **contribution** illustrates the shortcomings of identifiability, disentanglement, and group-based representation and shows how combining desiderata from all three fields contributes to more useful representations.

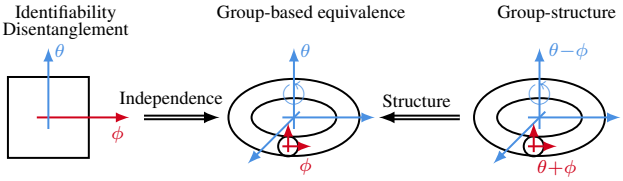

Figure 1. **Useful representations require desiderata from identifiability, disentanglement, and group-structured representations:** the latent space for two cyclic latents (color θ and orientation ϕ) is a torus with group structure $SO(2) \times SO(2)$. **Left:** independence is insufficient to capture the topology despite no information loss (i.e., encoding the cyclic latents in $[0; 2\pi)^2$, cf. Ex. 2.3); **right:** the group structure alone cannot ensure independent latents (a linear combination of θ, ϕ can also parameterize the torus, see Ex. 2.2); **middle:** a combination captures the topology and assigns latent factors encoding separate properties to distinct subspaces

2. Shortcomings of identifiability and disentanglement

Both identifiability and (group-based) disentanglement strive for a *useful* representation, though it is unclear which latent properties they can capture. Motivated by practical

[*]Equal contribution [1]Max Planck Institute for Intelligent Systems, Tübingen, Germany [2]Institute of Neuroinformatics, ETH Zürich, Switzerland [3]Max Planck ETH Center for Learning Systems [4]University of Tübingen, Tübingen, Germany [5]International Max Planck Research School for Intelligent Systems [6]European Laboratory for Learning and Intelligent Systems. Correspondence to: Hamza Keurti <hamza.keurti@tuebingen.mpg.de>, Patrik Reizinger <patrik.reizinger@tuebingen.mpg.de>.

Presented at the 2^{nd} Annual Workshop on Topology, Algebra, and Geometry in Machine Learning (TAG-ML) at the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. 2023.

considerations, we refer to a useful representation as one that encodes the latent factors necessary for the downstream task without information loss, assigns separate (independent) subspaces to factors encoding distinct properties such as position or color, and captures the latent space topology, e.g., to measure similarity between samples.[1] In the following, we demonstrate how and when identifiability and disentanglement fail to capture a useful representation, then propose a means to reconcile these shortcomings. We start with the classical example of Euler angles, showing that identifiability does not necessarily imply disentanglement:

Example 2.1 (Identifiability does not guarantee disentanglement)**.** *Euler angles describe 3D orientation by a non-commutative sequence of rotations around the $x, y, z-$axes. Rotating a 3D cuboid with unequal side lengths (e.g., a book) with $\pi/2$ around two axes in one (e.g., x, y) and the opposite (y, x) order yields different orientations (Higgins et al., 2018, Fig. 1B). The Euler angles cannot be disentangled from each other according to Defn. A.2, for no group affects only one Euler angle, but Euler angles can be disentangled from, e.g., position. Since identifiability is agnostic to group structure, Euler angles can be identified.*

Identifiability *can* imply group-based disentanglement. Assume that $\mathcal{Z} = [a; b]^2$ (e.g., 1D position and size and we have identifiability up to permutation and sign. By the identifiability guarantee, the inferred latents also factorize and have a corresponding group action (scalar addition); thus, the topology is also presevered. When (not group-based) disentanglement is measured by the DCI score (Eastwood & Williams, 2018), $D = C = 1$, and the inferred and true latents have the same dimensionality, then disentanglement does imply identifiability up to sign and permutation—which is a very strong identifiability class (Eastwood et al., 2022, Cor. 3.4). The group-based perspective does not imply identifiability though: capturing the latents' topology (which is not guaranteed by identifiability, cf. Ex. 2.3) does not ensure that semantic concepts are encoded in different latents.

Example 2.2 (Group-based disentanglement is insufficient to separate meaningful latents)**.** *Assume a torus latent space (i.e., $\mathcal{Z} = \mathcal{S}^1 \times \mathcal{S}^1 \subset \mathbb{R}^3$) with two cyclic latent factors and a group structure $G = SO(2) \times SO(2)$ such as orientation in the 2D plane and hue. Without considering the statistical perspective (i.e., the independence of latent factors), the correct group structure could be recovered without meaningful latents, due to the Abelian group structure. Consider the parametrization of the torus by $\theta, \phi \in [0; 2\pi)$ and $R, r > 0$, yielding the coordinates $x(\theta, \phi) = (R + r\cos\theta)\cos\phi$, $y(\theta, \phi) = (R + r\cos\theta)\sin\phi$, and $z(\theta, \phi) = \cos\theta$. Considering $\theta' = (\theta + \phi)$ and $\phi' = (\theta - \phi)$ such that θ', ϕ' are chosen to be in $[0; 2\pi)$, the topology still corresponds to a torus, despite the subgroups not corresponding to semantically meaningful latents, but a linear combination of orientation and hue.*

Ex. 2.2 also demonstrates that besides (group-based) disentanglement does not necessarily imply identifiability, it cannot ensure that the latent factors are independent.
Identifiability results ensure that inferred and ground-truth latent factors are related by a well-defined equivalence class. Despite capturing all information encoded in the latents, it might yield anomalous results when e.g. measuring sample similarity or producing latent interpolations.

Example 2.3 (Identifiability does not guarantee the correct topology)**.** *Assume $\mathcal{Z} = \mathbb{R}^2 \times [0; 2\pi)$ encoding x, y position and orientation θ, i.e., $\boldsymbol{z} = [x; y; \theta]$, and consider three points: $A_1(x, y, 0), A_2(x, y, 2\pi - \pi/12)$ and $A_3(x, y, \pi)$. Then there is no neighborhood of A_1 that contains A_2 without containing A_3, which violates the topology of the true latent space. This can also be appreciated by measuring the Euclidean distance (ℓ_2-distance) and the actual angular distance between the points. We find that the angular distances, $\sphericalangle(A; B) = (\theta_A - \theta_B \mod 2\pi)$, are ordered as $\sphericalangle(A_1; A_2) < \sphericalangle(A_1; A_3)$; and that $\mathcal{Z}$ captures the latent factors of position and orientation without loss of information. However, the ℓ_2-metric yields $\|A_1 - A_2\| > \|A_1 - A_3\|$, even for identifiable representations up to permutations and scalings.*

Ex. 2.3 can be resolved when $\mathcal{Z}$, all else being equal, is structured as $\mathbb{R}^2 \times [-1; 1]^2$ to capture the periodicity of θ via $[\cos\theta; \sin\theta]$. Since the unit circle is embedded $\mathbb{R}^2$, we can use the ℓ_2-metric to compare orientations.
An additional restriction of identifiability is that it treats the latent factors as homogeneous, i.e., it restricts the equivalence class for the latent components jointly. Our next example shows that it is not strictly necessary: when inferring position and orientation, we should "separate" one from the other, but e.g. any coordinate system should suffice for the position.

Example 2.4 (More strict identifiability classes can be unnecessary when considering the latent space structure)**.** *Consider a ground-truth latent space $\mathcal{Z}^* = \mathbb{R}^2 \times [0; 2\pi)$ encoding x, y position and orientation θ, i.e., $\boldsymbol{z}^* = [x; y; \theta]$. Equip the inferred spaces $\mathcal{Z}, \mathcal{Z}'$ with the ℓ_2-metric, assume identifiability up to a linear map i.e., $\mathcal{Z} = \mathbf{A}\mathcal{Z}^*, \mathcal{Z}' = \mathbf{A}'\mathcal{Z}^*$ such that:*

$$\mathbf{A} = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 0 & -1 \end{pmatrix}, \qquad \mathbf{A}' = \begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

yielding $\boldsymbol{z} = [x + \theta; y; x - \theta]$ and $\boldsymbol{z}' = [x + y; x - y; \theta]$. Euclidean distance does not express the similarity of meaningful properties for $\mathcal{Z}$, for $\mathcal{Z}$ is not structured by a group; however, there is no such problem for $\mathcal{Z}'$: though $\mathbf{A}'$ can be thought as a change of basis up to scaling, it does not combine latent factors with different semantics. $\mathcal{Z}'$ has a

[1] A factor such as color can be encoded in multiple dimensions, which is necessary to encode its cyclic property

group structure with linear group actions for 2D-position and orientation:

$$g^{xy} = \begin{pmatrix} \mathbf{I}_2 & \boldsymbol{\delta} \\ \mathbf{0}_{1\times 2} & 1 \end{pmatrix}; \qquad g^{\theta} = \begin{pmatrix} \cos\delta & \sin\delta \\ -\sin\delta & \cos\delta \end{pmatrix},$$

where $\boldsymbol{\delta} = [\delta_x; \delta_y] \in \mathbb{R}^2, \delta \in [0; 2\pi)$ *and* g^{xy}, g^{θ} *are applied to* x, y *parametrized as homogeneous coordinates* $[x; y; 1]$ *and* θ *encoded as* $[\cos\theta; \sin\theta]$. *By not capturing latent semantics (as group-based disentanglement does), identifiability might unnecessarily restrict the equivalence class. E.g., identifying the latent "blocks" of* $(x; y)$ *and* θ *(subspaces of* $\mathcal{Z}$*) can suffice without further constraints in the subspace. Requiring, e.g., linear latent interpolations might rule out elementwise nonlinearities, but this approach is still permits the change of basis within a subspace.*

3. Equivalence of Group-Structured Representations

Higgins et al. (2018) introduced a notion of disentanglement of group structured representations w.r.t. a given group decomposition. However, without knowing which group decomposition provides semantically meaningful direct subgroups, there are often infinitely many possible group decompositions and, therefore, infinitely many ways to disentangle the group-structured representations. In this section, we define an equivalence class over group-structured and over disentangled group-structured representations, then we highlight the challenge of finding the right disentanglement.

Criteria for the equivalence class: Assume that the representations $\boldsymbol{f}_\theta : \mathcal{X} \to \mathcal{Z}$ and $\boldsymbol{f}_{\theta'} : \mathcal{X} \to \mathcal{Z}'$ capture the structure of the generative factors (or a part thereof) through equivariance. We consider them as equivalent if (1) they are both invariant/equivariant to the same group elements; and (2) information orthogonal to the group action (e.g., the object's identity when acted on by some displacement group) is encoded in the latent spaces the same way. We consider the group homomorphisms ρ, ρ' induced by the group actions on $\mathcal{Z}, \mathcal{Z}'$, respectively. Condition (1) corresponds to $\mathrm{Ker}(\rho) = \mathrm{Ker}(\rho')$, whereas condition (2) can be translated into the quotient spaces of the representations being homeomorphic, i.e., $\boldsymbol{f}_\theta(\mathcal{X})/G \sim \boldsymbol{f}_{\theta'}(\mathcal{X})/G$. We can now define an equivalence class that satisfies these criteria:

Definition 3.1 (Equivalence of group-structured representations)**.** Two group-structured representations $\boldsymbol{f}_\theta, \boldsymbol{f}_{\theta'}$ are equivalent w.r.t. the group G if there is a mapping $\psi : \mathcal{Z} \to \mathcal{Z}'$ such that: $\psi_{|\boldsymbol{f}_\theta(\mathcal{X})}$ is injective and Fig. 2 commutes, i.e., $\boldsymbol{f}_{\theta'}(\boldsymbol{x}) = \psi(\boldsymbol{f}_\theta(\boldsymbol{x}))$.

From Defn. 3.1, we derive an equivalence class for disentangled representations.

Definition 3.2 (Equivalence of Disentangled Group-Structured Representations)**.** Let $\boldsymbol{f}_\theta$ and $\boldsymbol{f}_{\theta'}$ be disentangled w.r.t. the group decompositions $G = G_1 \otimes \cdots \otimes G_n$ and $G = G'_1 \otimes \cdots \otimes G'_{n'}$ (cf. Definition A.2). Let I (resp. I') be the subset of indices i for which the action of G_i (resp. G'_i) is not trivial on $\mathcal{Z}_i$ (resp. $\mathcal{Z}'_i$). Then the two disentangled representations are equivalent if and only if there exists an injective mapping $\varphi : I \mapsto I'$ such that $G_i \cong G'_{\varphi(i)}$ and the projected representations $\boldsymbol{f}_{\theta,i}$ and $\boldsymbol{f}_{\theta',\varphi(i)}$ are equivalent with regard to the action of the group G_i.

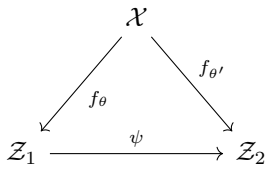

Figure 2. Commutativity diagram for Defn. 3.1

We can learn such decompositions unsupervisedly if we enforce block-diagonality of any group element in matrix form (Keurti et al., 2022). However, when the decomposition's subgroups are commutative or contain a non-trivial center, multiple decompositions may fit the same block-diagonal structure. Similar to (Higgins et al., 2018), assume that a 2D shape is transitively acted on by the group of cyclic translations $G = SO(2) \times SO(2)$, where the first subgroup corresponds to cyclic translations along the x axis and the second subgroup to color shifts on the hue wheel h. This is the semantically meaningful decomposition that the group-structured representation is expected to learn. Let $\boldsymbol{f}_\theta$ be such a disentangled representation with the associated induced group homomorphism $\rho = \rho_1 \oplus \rho_2$. However, G admits an identical decomposition $SO(2) \times SO(2)$ where the subgroups correspond to cyclic translations along the $x + h$ and $x - h$ axes. If $\boldsymbol{f}_{\theta'}$ is disentangled along this second decomposition with the associated induced group homomorphism $\rho' = \rho'_1 \oplus \rho'_2$ then the two representations are not equivalent disentangled group representations. Indeed, looking at the kernels we find that $\mathrm{Ker}(\rho'_1) = \{(x, h) \in G | x = -h\}$ which is different than both kernels for the subrepresentations of ρ: $\mathrm{Ker}(\rho_1) = \{(x, 0) | x \in G_1\}$ and $\mathrm{Ker}(\rho_2) = \{(0, h) | h \in G_2\}$. This ambiguity does not concern centerless subgroups, e.g., $G = SO(3) \times SO(2)$ acting on the $3D$ orientation (α, β, γ) and the color hue h of a $3D$ shape. If we find a disentangled group-structured representation with the associated induced group homomorphism $\rho = \rho_1 \oplus \rho_2$ such that both $\rho_{1/2}$ have a non-trivial kernel, then ρ_1 represents $SO(3)$ and ρ_2 represents $SO(2)$ or the other way. The proposition 3.3 summarizes these insights.

Proposition 3.3. *A group decomposition* $G = G_1 \times G_2$ *is identifiable up to a mixing of the centers of* G_1 *and* G_2.

4. Discussion

Limitations. Our work illustrates how current definitions of identifiability and (group-based) disentanglement fail to

capture aspects of the underlying latent space. Despite our proposal on how to reconcile these shortcomings *theoretically,* we do not provide practical means to do so.

Conclusion. Our examples illustrate that disentanglement into independent variables can lead to an inconsistent latent topology. On the other hand, disentanglement according to group structure may capture the structure but admit infinitely many decompositions besides the semantic ones. We defined an equivalence class for both group-structured representations and disentangled group-structured representations and have shown how we may get equivalent group-structured representations but not equivalent disentangled group-structured representations, depending on the structuring group. This specific limitation might be overcome by looking at the statistical structure while learning the group-structured representations.

Acknowledgements

The authors would like to thank Felix Leeb and Nikolay Malkin for fruitful discussions. Wieland Brendel acknowledges financial support via an Emmy Noether Grant funded by the German Research Foundation (DFG) under grant no. BR 6382/1-1. Wieland Brendel is a member of the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645. Hamza Keurti thanks the support of the Max Planck ETH Center for Learning Systems (CLS). Patrik Reizinger thanks the support of the International Max Planck Research School for Intelligent Systems (IMPRS-IS) and acknowledges his membership in the European Laboratory for Learning and Intelligent Systems (ELLIS) PhD program. Patrik Reizinger thanks Michael Kirchhof for insightful correspondence.

References

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, August 2013. ISSN 0162-8828, 2160-9292. doi: 10.1109/tpami.2013.50. URL https://doi.org/10.1109/tpami.2013.50.

Bishop, C. M. *Pattern Recognition and Machine Learning*, volume 4. Springer New York, 2006. doi: 10.1007/978-0-387-45528-0. URL https://doi.org/10.1007/978-0-387-45528-0.

Bronstein, M. M., Bruna, J., Cohen, T., and Veličković, P. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *ArXiv preprint*, abs/2104.13478, 2021.

Chen, T. Q., Li, X., Grosse, R. B., and Duvenaud, D. Isolating Sources of Disentanglement in Variational Autoencoders. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 2615–2625, 2018.

Cohen, T. and Welling, M. Learning the Irreducible Representations of Commutative Lie Groups, May 2014. URL http://arxiv.org/abs/1402.4437. arXiv:1402.4437 [cs].

Eastwood, C. and Williams, C. K. I. A framework for the quantitative evaluation of disentangled representations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Eastwood, C., Kekic, A., and Nicolicioiu, A. L. On the DCI Framework for Evaluating Disentangled Representations: Extensions and Connections to Identifiability. pp. 8, 2022.

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

Hyvärinen, A. and Pajunen, P. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, April 1999. ISSN 0893-6080. doi: 10.1016/s0893-6080(98)00140-3. URL https://doi.org/10.1016/s0893-6080(98)00140-3.

Keurti, H., Pan, H.-R., Besserve, M., Grewe, B. F., and Schölkopf, B. Homomorphism Autoencoder — Learning Group Structured Representations from Interactions. July 2022. URL https://openreview.net/forum?id=9XUM3-KJ50U.

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 2019.

Murphy, K. P. *Machine learning: A probabilistic perspective*. MIT press, 2012.

A. Definitions

We follow the definitions proposed by Higgins et al. (2018) for group structured representations and disentangled group-structured representations.

Definition A.1 (Group Structured Representation)**.** Let $\mathcal{Z}^*$ be the generative factors of the observed space $\mathcal{X}$ through the mapping $b : \mathcal{Z}^* \to \mathcal{X}$, structured by a group G through the action $\cdot : G \times \mathcal{Z}^* \to \mathcal{Z}^*$. A vector representation $\boldsymbol{f}_\theta : \mathcal{X} \to \mathcal{Z}$ is a group-structured representation if it satisfies:

1. There is a (non-trivial) action of G on $\mathcal{Z}$, i.e., $\cdot_{\mathcal{Z}} : G \times \mathcal{Z} \to \mathcal{Z}$.
2. The composition $f = \boldsymbol{f}_\theta \circ b : \mathcal{Z}^* \to \mathcal{Z}$ is equivariant, meaning that transformations of $\mathcal{Z}^*$ are reflected on $\mathcal{Z}$, i.e., $\forall g \in G, z* \in \mathcal{Z}^*, \quad f(g \cdot_{\mathcal{Z}^*} z*) = g \cdot_{\mathcal{Z}} f(z*)$.

Definition A.2 (Disentangled Group Structured Representation)**.** The group-structured representation is disentangled with regard to the group decomposition $G = G_1 \times ... \times G_n$ if it satisfies this additional condition:

3. Z can be written as a product of spaces $Z = Z_1 \times ... \times Z_n$ or as a direct sum of subspaces $Z = Z_1 \oplus ... \oplus Z_n$ such that each subgroup G_i acts non trivially on Z_i and acts trivially on Z_j for $j \neq i$.

Definition A.3 (Strong Identifiability (Khemakhem et al., 2020b))**.** Given a parameter class Θ, when the feature extractors $\boldsymbol{f}_{\theta_1}, \boldsymbol{f}_{\theta_2} : \mathcal{X} \to \mathcal{Z}$ produce latent representations $\boldsymbol{z}_1 = \boldsymbol{f}_{\theta_1}(\boldsymbol{x}), \boldsymbol{z}_2 = \boldsymbol{f}_{\theta_2}(\boldsymbol{x})$ that are equivalent up to scaled permutations and offsets c for all $\theta_1, \theta_2 \in \Theta$, i.e.,

$$\theta_1 \sim \theta_2 \iff z = \boldsymbol{f}_{\theta_1}(\boldsymbol{x}) = \mathbf{D}\mathbf{P}\boldsymbol{f}_{\theta_2}(\boldsymbol{x}) + c, \tag{1}$$

where $\mathbf{D}$ is a diagonal and $\mathbf{P}$ a permutation matrix. Then θ_1, θ_2 fulfill an *equivalence* relationship.

Definition A.4 (Weak Identifiability (Khemakhem et al., 2020b))**.** Given a parameter class Θ, when the feature extractors $\boldsymbol{f}_{\theta_1}, \boldsymbol{f}_{\theta_2} : \mathcal{X} \to \mathcal{Z}$ produce latent representations $\boldsymbol{z}_1 = \boldsymbol{f}_{\theta_1}(\boldsymbol{x}), \boldsymbol{z}_2 = \boldsymbol{f}_{\theta_2}(\boldsymbol{x})$ that are equivalent up to matrix multiplications and offsets c for all $\theta_1, \theta_2 \in \Theta$, i.e.,

$$\theta_1 \sim \theta_2 \iff \boldsymbol{z} = \boldsymbol{f}_{\theta_1}(\boldsymbol{x}) = \mathbf{A}\boldsymbol{f}_{\theta_2}(\boldsymbol{x}) + c, \tag{2}$$

where $\operatorname{rank}(\mathbf{A}) \geq \min(\dim \mathcal{Z}; \dim \mathcal{X})$. Then θ_1, θ_2 fulfill an *equivalence* relationship.

Definition A.5 (Identifiability up to elementwise nonlinearities (Hyvärinen & Morioka, 2017))**.** Given a parameter class Θ, when the feature extractors $\boldsymbol{f}_{\theta_1}, \boldsymbol{f}_{\theta_2} : \mathcal{X} \to \mathcal{Z}$ produce latent representations $\boldsymbol{z}_1 = \boldsymbol{f}_{\theta_1}(\boldsymbol{x}), \boldsymbol{z}_2 = \boldsymbol{f}_{\theta_2}(\boldsymbol{x})$ that are equivalent up to elementwise nonlinearities, matrix multiplications and offsets c for all $\theta_1, \theta_2 \in \Theta$, i.e.,

$$\theta_1 \sim \theta_2 \iff \boldsymbol{z} = \boldsymbol{f}_{\theta_1}(\boldsymbol{x}) = \mathbf{A}\sigma\left[\boldsymbol{f}_{\theta_2}(\boldsymbol{x})\right] + c, \tag{3}$$

where $\operatorname{rank}(\mathbf{A}) \geq \min(\dim \mathcal{Z}; \dim \mathcal{X})$ and σ denotes an elementwise nonlinear transformation. Then θ_1, θ_2 fulfill an *equivalence* relationship.

B. Background

Let $\boldsymbol{f}_\theta : \mathcal{X} \to \mathcal{Z}$ be a feature extractor (encoder) parametrized by $\theta \in \Theta$, where $\mathcal{X} \subseteq \mathbb{R}^D, \mathcal{Z} \subseteq \mathbb{R}^d$ are the observation and latent spaces. $\mathbf{A} \in GL(d), c \in \mathbb{R}^d, \mathbf{D} = \operatorname{diag}(D_1, \ldots, D_d) : D_i \neq 0$.

Group theory. A group G structures the space $\mathcal{S} \in \{\mathcal{X}, \mathcal{Z}\}$ through a group action $\cdot : G \times \mathcal{S} \to \mathcal{S}$, associating an invertible transformation of $\mathcal{S}$ to every group element $g \in G$. The induced map is a group homomorphism. E.g., given the orientation of a 2D image by a scalar phase, it can be changed via scalar addition modulo the rotation period in $\mathcal{Z}$, or by a rotation matrix in $\mathcal{X}$. The structure of the latent space and the symmetry group is expressed via decomposition, i.e., $\mathcal{Z} = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_k$ and $G = G_1 \times \cdots \times G_k$, where only the subgroup G_i affects the subspace $\mathcal{Z}_i$ via the action $\cdot_i : G \times \mathcal{Z}_i \to \mathcal{Z}_i$ ($k \leq d$)—the dimensionality of $\mathcal{Z}_i$ and that of the action's representation of G_i can have *different* dimensions. E.g., the cyclic, scalar representation of color cannot be expressed with a one-dimensional linear transformation. Among symmetry relationships, *equivariance* has a distinguished role, i.e., when $\boldsymbol{f}_\theta(g \cdot \boldsymbol{x}) = g \cdot \boldsymbol{f}_\theta(\boldsymbol{x})$ holds.

Disentanglement. Inspired by Weyl's principle from physics (Kanatani, 2011), an equivariance-based notion of *disentanglement* was first proposed by Cohen & Welling (2014), followed by Higgins et al. (2018). **??** deems a representation disentangled w.r.t. a decomposition of G if the representation also decomposes into independent subspaces $\mathcal{Z}_i$ that are only affected by G_i. **??** depends on the group decomposition into subgroups. I.e., disentangled representations are non-unique since the "true decomposition" is nontrivial. For the subgroups' dimensionality is not prescribed, the representation granularity and the bases of $\mathcal{Z}_i$ can be arbitrary.

Identifiability. Identifiability attempts to construct model classes with theoretical guarantees for reconstructing the latent factors (up to indeterminacies, such as scalings, permutations, or elementwise transformations). This is impossible without additional assumptions (Hyvärinen & Pajunen, 1999) restricting the data distribution (Guo et al., 2022; Hyvärinen & Morioka, 2017; Khemakhem et al., 2020a; Morioka et al., 2021; Hyvärinen & Morioka, 2016) or the function class (Gresele et al., 2021). A factorizing joint latent distribution $p(\boldsymbol{z}) = \prod_i p(z_i)$ over $\mathcal{Z}$ is central to identifiability, with recent work relying on auxiliary variables $\mathbf{u}$ that introduce conditional independence (Khemakhem et al., 2020a). Furthermore, $\boldsymbol{f}$ is assumed to be *at least* injective (Khemakhem et al., 2020a); most works assume bijectivity (Hyvärinen & Morioka, 2017; 2016; Zhang & Hyvarinen, 2012; Hyvärinen et al., 2019) since they assume $\dim \mathcal{X} = \dim \mathcal{Z}$. Appx. A summarizes the notions of identifiability—with the common denominator that $\forall \theta_1, \theta_2 \in \Theta$ the marginals $p_{\theta_1}(\boldsymbol{x}), p_{\theta_2}(\boldsymbol{x})$ are equivalent; expressed as $\theta_1 \sim \theta_2$. However, the feature extractors $\boldsymbol{f}_{\theta_i}$ map $\boldsymbol{x}$ to an equivalent $\boldsymbol{z}$ up to a certain *equivalence class*, including *invertible transformations*: $\mathbf{DP}\boldsymbol{z} + c$ with permutation matrix $\mathbf{P}$ for *strong*; $\mathbf{A}\boldsymbol{z} + c$ for *weak identifiability*. Hyvärinen & Morioka (2017; 2016) include elementwise (monotonous) (non)linear transformations (denoted as σ), i.e., $\mathbf{A}\sigma[\boldsymbol{z}] + c$. Alternatively, the parameters θ_1, θ_2 are equivalent if they parametrize feature extractors that (or, equivalently, the representation they produce) equal up to specific transformations.

Useful representations. The usefulness of a representation is not well-defined: identifiability defines it via independence and a relation to the ground truth, disentanglement via semantic meaning and symmetries. Achille & Soatto (2018) postulate sufficiency, minimality, invariance, and disentanglement to call a representation optimal. Eastwood & Williams (2018) use disentanglement, completeness, and informativeness. Cohen & Welling (2014) and Higgins et al. (2018) advocate for group-based structure. The plethora of metrics measuring disentanglement makes it especially hard to navigate the literature. To add insult to injury, the word disentanglement is overloaded several times, and the metrics measure distinct though often correlated propeties (Locatello et al., 2019; Sepliarskaia et al., 2021; Eastwood & Williams, 2018; Higgins et al., 2018).

C. Related work

Identifiability reasons about the true Data Generating Process (DGP), whereas disentanglement takes a more empirical approach and measures the performance of (heuristic) methods such as β-Variational Autoencoder (VAE) (Higgins et al., 2017), TCVAE (Chen et al., 2018), FactorVAE (Kim & Mnih, 2018) with a set of diverse metrics (for comparison, see (Locatello et al., 2019)). Thus, despite a conceptual connection was already present in the seminal work of Bengio et al. (2013), the two communities largely developed independently; metrics, such as Mean Correlation Coefficient (MCC) (Hyvärinen & Morioka, 2016) started to appear in the disentanglement literature, although proposed for identifiability. The group-theoretic formalization of disentanglement is a recent development (Cohen & Welling, 2014; Higgins et al., 2017; 2022; Bronstein et al., 2021) and was leveraged for different problems (Cohen et al., 2019; Keurti et al., 2022). Until recently, there was no formal connection between the two notions. The first such result known to the authors is (Eastwood et al., 2022), which proves a connection between optimizing the DCI disentanglement score (Eastwood & Williams, 2018) and identifiability up to permutation and sign. Ahuja et al. (2022) describe the identifiability indeterminacies for a specific model from the perspective of the equivariances of the mechanisms mapping $\mathcal{Z} \to \mathcal{X}$.

D. Notation

Acronyms

DCI Disentanglement Completeness Informativeness score
DGP Data Generating Process
MCC Mean Correlation Coefficient
MIG Mutual Information Gap
VAE Variational Autoencoder

Nomenclature

G symmetry group
$\mathbf{u}$ auxiliary variable vector
$\mathcal{S}$ hypersphere
Ker kernel space
$\boldsymbol{f}$ encoder map $\mathcal{X} \to \mathcal{Z}$
g group element

Algebra

$\mathbf{D}$ diagonal matrix
$\mathbf{P}$ permutation matrix

Latents

$\boldsymbol{z}$ latent vector
$\mathcal{Z}$ latents
d dimensionality of the latent space $\mathcal{Z}$
z latent single component

Observations

D dimensionality of the observation space $\mathcal{X}$

$\boldsymbol{x}$ observation vector

$\mathcal{X}$ observation space

References

Achille, A. and Soatto, S. Emergence of Invariance and Disentanglement in Deep Representations. In *2018 Information Theory and Applications Workshop (ITA)*, pp. 1–9, San Diego, CA, February 2018. IEEE. ISBN 978-1-72810-124-8. doi: 10.1109/ITA.2018.8503149. URL https://ieeexplore.ieee.org/document/8503149/.

Ahuja, K., Hartford, J., and Bengio, Y. Properties from mechanisms: an equivariance perspective on identifiable representation learning. March 2022. URL https://openreview.net/forum?id=g5ynW-jMq4M.

Bengio, Y., Courville, A., and Vincent, P. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(8):1798–1828, August 2013. ISSN 0162-8828, 2160-9292. doi: 10.1109/tpami.2013.50. URL https://doi.org/10.1109/tpami.2013.50.

Bronstein, M. M., Bruna, J., Cohen, T., and Veličković, P. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *ArXiv preprint*, abs/2104.13478, 2021.

Chen, T. Q., Li, X., Grosse, R. B., and Duvenaud, D. Isolating Sources of Disentanglement in Variational Autoencoders. In Bengio, S., Wallach, H. M., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pp. 2615–2625, 2018.

Cohen, T. and Welling, M. Learning the Irreducible Representations of Commutative Lie Groups, May 2014. URL http://arxiv.org/abs/1402.4437. arXiv:1402.4437 [cs].

Cohen, T., Weiler, M., Kicanaoglu, B., and Welling, M. Gauge equivariant convolutional networks and the icosahedral CNN. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2019.

Eastwood, C. and Williams, C. K. I. A framework for the quantitative evaluation of disentangled representations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

Eastwood, C., Kekic, A., and Nicolicioiu, A. L. On the DCI Framework for Evaluating Disentangled Representations: Extensions and Connections to Identifiability. pp. 8, 2022.

Gresele, L., von Kügelgen, J., Stimper, V., Schölkopf, B., and Besserve, M. Independent mechanisms analysis, a new concept? In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*, pp. 28233–28248. Curran Associates, Inc., December 2021. URL https://proceedings.neurips.cc/paper/2021/file/edc27f139c3b4e4bb29d1cdbc45663f9-Paper.pdf.

Guo, S., Tóth, V., Schölkopf, B., and Huszár, F. Causal de finetti: On the identification of invariant causal structure in exchangeable data. *ArXiv preprint*, abs/2203.15756, 2022.

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., and Lerchner, A. β-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017.

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

Higgins, I., Racaniere, S., and Rezende, D. Symmetry-Based Representations for Artificial and Biological General Intelligence. *Frontiers in Computational Neuroscience*, 16, 2022. ISSN 1662-5188.

Hyvärinen, A. and Morioka, H. Unsupervised Feature Extraction by Time-Contrastive Learning and Nonlinear ICA. In Lee, D. D., Sugiyama, M., von Luxburg, U., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pp. 3765–3773, 2016.

Hyvärinen, A. and Morioka, H. Nonlinear ICA of temporally dependent stationary sources. In Singh, A. and Zhu, X. J. (eds.), *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pp. 460–469. PMLR, 2017.

Hyvärinen, A. and Pajunen, P. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, April 1999. ISSN 0893-6080. doi: 10.1016/s0893-6080(98)00140-3. URL https://doi.org/10.1016/s0893-6080(98)00140-3.

Hyvärinen, A., Sasaki, H., and Turner, R. E. Nonlinear ICA using auxiliary variables and generalized contrastive learning. In Chaudhuri, K. and Sugiyama, M. (eds.), *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pp. 859–868. PMLR, 2019.

Kanatani, K. *Group-theoretical methods in image understanding*. Springer Series in Information Sciences. Springer, Berlin, Germany, October 2011.

Keurti, H., Pan, H.-R., Besserve, M., Grewe, B. F., and Schölkopf, B. Homomorphism Autoencoder — Learning Group Structured Representations from Interactions. July 2022. URL https://openreview.net/forum?id=9XUM3-KJ50U.

Khemakhem, I., Kingma, D. P., Monti, R. P., and Hyvärinen, A. Variational Autoencoders and Nonlinear ICA: A Unifying Framework. In Chiappa, S. and Calandra, R. (eds.), *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2207–2217. PMLR, 2020a.

Khemakhem, I., Monti, R. P., Kingma, D. P., and Hyvärinen, A. ICE}-{BeeM: Identifiable conditional energy-based deep models based on nonlinear ICA. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M.-F., and Lin, H.-T. (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020b.

Kim, H. and Mnih, A. Disentangling by factorising. In Dy, J. G. and Krause, A. (eds.), *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2654–2663. PMLR, 2018.

Locatello, F., Bauer, S., Lucic, M., Rätsch, G., Gelly, S., Schölkopf, B., and Bachem, O. Challenging common assumptions in the unsupervised learning of disentangled representations. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4114–4124. PMLR, 2019.

Morioka, H., Hälvä, H., and Hyvärinen, A. Independent innovation analysis for nonlinear vector autoregressive process. In Banerjee, A. and Fukumizu, K. (eds.), *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1549–1557. PMLR, 2021.

Sepliarskaia, A., Kiseleva, J., and de Rijke, M. How to Not Measure Disentanglement, March 2021. URL http://arxiv.org/abs/1910.05587. arXiv:1910.05587 [cs, stat].

Zhang, K. and Hyvarinen, A. On the Identifiability of the Post-Nonlinear Causal Model. *arXiv:1205.2599 [cs, stat]*, 2012. arXiv: 1205.2599.