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.,

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

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

LVM Latent Variable Model

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.