# Linear Disentanglement through Action Group Learning

Barthélémy Dang-Nhu, Louis Annabi, Sylvain Argentieri
Sorbonne Université, CNRS, Institut des Systèmes Intelligents et de Robotique, ISIR, F-75005 Paris, France
{dangnhu,annabi,argentieri}@isir.upmc.fr

## 1 Introduction

A causal world model that allows an agent to learn the underlying structure of the environment, rather than just surface-level correlations, is especially important in lifelong learning, where the agent must adapt to out-of-distribution changes. By modeling why things happen (e.g. by capturing causal structures) rather than just what happens (i.e. not relying only on prediction), a causal model enables the agent to make reliable predictions and to adapt quickly to new situations with minimal data. This leads to more robust, transferable learning across evolving tasks and environments [1, 16]. Building a causal world model in an unsupervised setting often requires, in the first place, that the data's distinct generative factors be already identified [19], which can be achieved by acquiring a disentangled representation of the environment, making possible to identify meaningful features that correspond to causal variables. Without this structured representation, inferring causal relations becomes unreliable, as the agent cannot isolate the independent mechanisms generating observations.

Numerous methods have been proposed in recent years to obtain disentangled representations, most of which are based on VAEs [9, 11] or GANs [5]. However, it has been shown that obtaining the correct disentangled representation *in a purely unsupervised manner* is impossible without prior knowledge [12]. In this work, we thus only focus on the disentanglement representation *using self-supervised learning* from sensorimotor interactions.

## 2 Related Work

Higgins et al. [8] propose a formal definition of disentanglement called *Linear Symmetry-Based Disentanglement* (LSBD), based on group theory. Let $X$ denote the set of possible observations of the environment, an encoder is a function $h : X \to Z$ that projects an observation into a latent space. Additionally, we assume the existence of a set of actions $G$, which satisfies the axioms of a group: the existence of an identity element, closure under composition, and the existence of inverses. Furthermore, this group is assumed to decompose into direct factors $G_i$ i.e. $G = G_1 \times \cdots \times G_K$. This group structure defines an action function $\cdot_X : G \times X \to X$, which maps each action $g \in G$ and world state $x \in X$ to a new world state $x' \in X$ after the application of $g$.

**Definition 1.** $h$ is said to be *disentangled* if [8]:

1. There exists an action function $\cdot_Z : G \times Z \to Z$
2. There is equivariance: $\forall g \in G, x \in X, \quad g \cdot_Z h(x) = h(g \cdot_X x)$
3. There exists a decomposition $Z = Z_1 \oplus \ldots \oplus Z_K$ such that $Z_i$ is only affected by $G_i$

Moreover, the representation is said to be *linearly disentangled* if $\cdot_Z$ is linear, i.e., if there exists a group representation $\rho : G \to GL(Z)$ such that $g \cdot_Z z = \rho(g)z$. Several methods have been developed to learn such a disentangled representation [3, 15, 18]. However, they all rely on strong prior knowledge about the direct factors of actions (e.g., direct factors decomposition [3], cyclic groups [15]).

## 3 Contributions

We highlight theoretical and empirical limitations of existing algorithms. Based on these, we extend the LSBD framework by *explicitly* introducing assumptions that are required to ensure that disentanglement is consistently achievable. We then use these hypothesis to propose an algorithm that learns the group decomposition $G = G_1 \times \cdots \times G_K$ from raw actions. Additionally, we propose another algorithm that uses this decomposition to learn a disentanglement representation. Using these two methods combined, we propose the first algorithm for linear disentanglement that is completely agnostic to the action group.

## 4 Method

Our method is decomposed in three steps:

1. We learn a non-disentangled representation *i.e.* a representation satisfying only points 1 and 2 of Definition 1 solely for the purpose of learning an invective morphism $\rho : G \to GL(Z)$.
2. From this representation we learn the decomposition $G = G_1 \times \cdots \times G_K$
3. Thanks to this decomposition we learn disentangled representation

### 4.1 (Step 1) Learn a non-disentangled representation

The encoder $h : X \to Z$ is learned with an adapted version of a Variational Auto-Encoder [10] using a loss $\mathcal{L} = \mathcal{L}_{REC} + \mathcal{L}_{ACT}$. $\mathcal{L}_{REC}$ denotes the classical VAE reconstruction loss, which is completed with $\mathcal{L}_{ACT}$ which ensures that $h(x') = \rho(g)h(x)$ (as stated in point 2 from Definition 1) with $\rho(g) \in \mathbb{R}^{d \times d}$. In practice, $d^2$ parameters are learned for each action $g$.

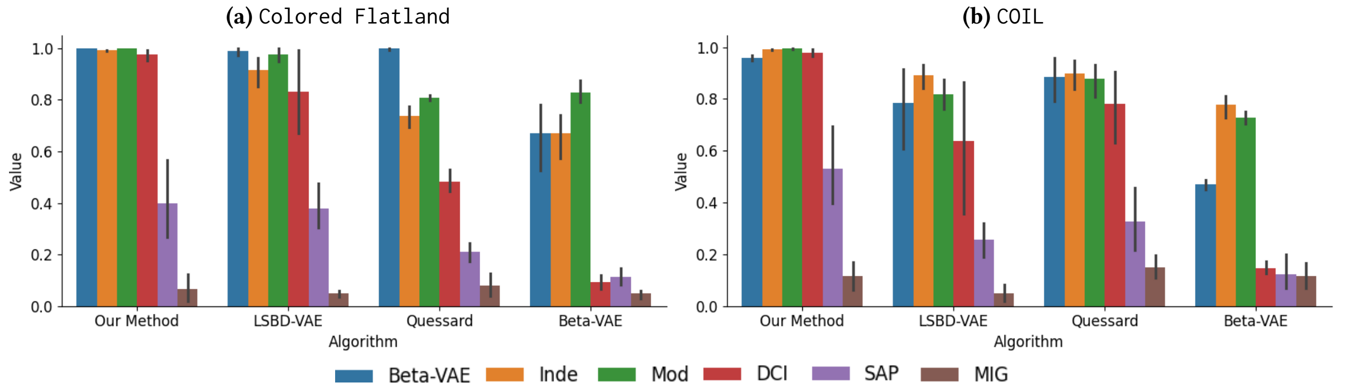

**Figure 1.** Six disentangled metrics computed for 4 different approaches. (Left) ColoredFlatland dataset. (Right) COIL dataset. Our approach exhibits almost systematically better disentangled metrics.

### 4.2 (Step 2) Learn the group decomposition

We argue that it is impossible to guarantee that the correct disentanglement is learned without additional assumptions. To address this issue, we introduce the following additional assumptions:

1. Only a subset of whole group action $\mathcal{G} \subset G$ is available for the agent;
2. The available actions are disentangled: each action in $\mathcal{G}$ belongs to a unique direct factor $G_i$
3. For each available action pair of a same direct factor $G_i$, there is another action from one to another.

We show that previous methods [3, 15, 18] make similar *implicit* hypotheses for the first two assumptions. Thanks to three hypotheses, we derive a criterion to determine whether two actions belong to the same direct factor. This criterion relies on the representation learned in the previous part.

### 4.3 (Step 3) Learn a disentangled representation

Learning a disentangled representation with respect to the LSBD framework is equivalent to learning a representation where all $\rho(g)$ are the identity matrix except for one block on the diagonal. Additionally we want actions of the same direct factor to learn the same specific block of their matrix, thus justifying Step 2 of the approach. To that end, we add a learnable mask to build matrices with such properties. Thanks to a continuous relaxation on those masks, this disentangled learning only needs one additional loss compared to the non-disentangled learning of Step 1.

## 5 Results

We first show that our group decomposition in Step 2 works perfectly well on different environments. Then we compare our Step 3 to state of the art algorithms for linear disentanglement: LSBD-VAE [18] and Quessard [15]; we also added $\beta$-VAE [7] for a purely unsupervised baseline. For that purpose, we used two environments:

- ColoredFlatland, an extension of Flatland [2] generating a disk moving along the x/y axes of a black background, but enriched with colors;
- COIL [13] where different objects can be rotated; permutations have been added to the original data.

As shown in Figure 1, our 3 steps approach outperforms these previous works in all of the disentanglement metrics proposed in [4, 6, 7, 11, 14, 17]. But one question remains: for a given learned model, how to evaluate the disentanglement without the ground-truth features ? We experimentally show that our method guarantees that low training loss implies necessarily maximised disentanglement metrics.

## 6 Conclusion and future work

In conclusion, our method outperforms existing approaches but has several limitations. The first one lies in the use of multiple successive training phases, which incurs additional computational cost as Step 3 is learned from scratch because we didn't find so far a way to re-use the representation obtained in Step 1. The second limitation is that Step 2 involves a hard-clustering of actions in direct factors, which can lead to poor learning performances in Step 3 if the discovered decomposition is incorrect.

Here, we have now disentangled the different features of the representation, the next step is to uncover the causal structure between the latent factors. This involves identifying which variables influence others and how interventions affect outcomes. With disentangled features serving as candidate causal variables, the agent can now explore and analyze their relationships –through observational data, interventions, or counterfactual reasoning– to build a structured causal world model that supports robust generalization and reasoning across tasks.

# References

[1] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A Meta-Transfer Objective for Learning to Disentangle Causal Mechanisms, February 2019. URL http://arxiv.org/abs/1901.10912. arXiv:1901.10912 [cs].

[2] Hugo Caselles-Dupré, Louis Annabi, Oksana Hagen, Michael Garcia-Ortiz, and David Filliat. Flatland: a lightweight first-person 2-d environment for reinforcement learning, 2018. URL https://arxiv.org/abs/1809.00510.

[3] Hugo Caselles-Dupré, Michael Garcia-Ortiz, and David Filliat. Symmetry-Based Disentangled Representation Learning requires Interaction with Environments, September 2019. URL http://arxiv.org/abs/1904.00243. arXiv:1904.00243 [cs].

[4] Ricky T. Q. Chen, Xuechen Li, Roger Grosse, and David Duvenaud. Isolating Sources of Disentanglement in Variational Autoencoders, April 2019. URL http://arxiv.org/abs/1802.04942. arXiv:1802.04942 [cs].

[5] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. InfoGAN: Interpretable Representation Learning by Information Maximizing Generative Adversarial Nets, June 2016. URL http://arxiv.org/abs/1606.03657. arXiv:1606.03657 [cs].

[6] Cian Eastwood and Christopher K. I. Williams. A Framework for the Quantitative Evaluation of Disentangled Representations. February 2018. URL https://openreview.net/forum?id=By-7dz-AZ.

[7] 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. URL https://openreview.net/forum?id=Sy2fzU9gl.

[8] Irina Higgins, David Amos, David Pfau, Sebastien Racaniere, Loic Matthey, Danilo Rezende, and Alexander Lerchner. Towards a Definition of Disentangled Representations, December 2018. URL http://arxiv.org/abs/1812.02230. arXiv:1812.02230 [cs].

[9] Hyunjik Kim and Andriy Mnih. Disentangling by Factorising, July 2019. URL http://arxiv.org/abs/1802.05983. arXiv:1802.05983 [stat].

[10] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

[11] Abhishek Kumar, Prasanna Sattigeri, and Avinash Balakrishnan. Variational Inference of Disentangled Latent Concepts from Unlabeled Observations, December 2018. URL http://arxiv.org/abs/1711.00848. arXiv:1711.00848 [cs].

[12] Francesco Locatello, Stefan Bauer, Mario Lucic, Gunnar Rätsch, Sylvain Gelly, Bernhard Schölkopf, and Olivier Bachem. Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations, June 2019. URL http://arxiv.org/abs/1811.12359. arXiv:1811.12359 [cs].

[13] Samer A. Nene, Shree K. Nayar, and Hiroshi Murase. Columbia object image library (coil-20). Technical Report CUCS-005-96, Department of Computer Science, Columbia University, February 1996.

[14] Matthew Painter, Jonathon Hare, and Adam Prugel-Bennett. Linear Disentangled Representations and Unsupervised Action Estimation, December 2020. URL http://arxiv.org/abs/2008.07922. arXiv:2008.07922 [cs].

[15] Robin Quessard, Thomas Barrett, and William Clements. Learning Disentangled Representations and Group Structure of Dynamical Environments. In *Advances in Neural Information Processing Systems*, volume 33, pp. 19727–19737. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/e449b9317dad920c0dd5ad0a2a2d5e49-Abstract.html.

[16] Jonathan Richens and Tom Everitt. Robust agents learn causal world models. In *The Twelfth International Conference on Learning Representations*, 2024.

[17] Karl Ridgeway and Michael C. Mozer. Learning Deep Disentangled Embeddings with the F-Statistic Loss, May 2018. URL http://arxiv.org/abs/1802.05312. arXiv:1802.05312 [cs].

[18] Loek Tonnaer, Luis A. Pérez Rey, Vlado Menkovski, Mike Holenderski, and Jacobus W. Portegies. Quantifying and Learning Linear Symmetry-Based Disentanglement, June 2022. URL http://arxiv.org/abs/2011.06070. arXiv:2011.06070 [cs].

[19] Alessio Zanga and Fabio Stella. A Survey on Causal Discovery: Theory and Practice, May 2023. URL http://arxiv.org/abs/2305.10032. arXiv:2305.10032 [cs].