ADVANCED TOPICS IN MACHINE LEARNING

Learning Disentangled and Structured representations:
A Causal Perspective

Emanuele Marconato♣,♠
♣DISI, University of Trento, ♠DI, University of Pisa

December 14, 2022

Tabel of Contents:

Main Message:

- [ ] The importance of the interventional formulation

Motivation

What is the world?

What is the world?

It's made of things, atoms, information.

What is the world?

It's made of things, atoms, information.

Humans are incredibly good at understanding information in coarser ways, denoting objects with names/symbols, and abstracting them.

What is the world?

It's made of things, atoms, information.

Humans are incredibly good at understanding information in coarser ways, denoting objects with names/symbols, and abstracting them.

We shift to semantic content when communicating.

What is the world?

It's made of things, atoms, information.

Humans are incredibly good at understanding information in coarser ways, denoting objects with names/symbols, and abstracting them.

We shift to semantic content when communicating.

When describing the world we provide models of it.

Levels of modelization

Model	Predict in i.i.d. setting	Predict under distr. shift/intervention	Answer counter-factual questions	Obtain physical insight	Learn from data
Mechanistic/physical	yes	yes	yes	yes	?
Structural causal	yes	yes	yes	?	?
Causal graphical	yes	yes	no	?	?
Statistical	yes	no	no	no	yes

Figure 1: Credits: Towards Causal Representation Learning - Schölkopf et al. (2021)

Levels of modelization

Model	Predict in i.i.d. setting	Predict under distr. shift/intervention	Answer counter-factual questions	Obtain physical insight	Learn from data
Mechanistic/physical	yes	yes	yes	yes	?
Structural causal	yes	yes	yes	?	?
Causal graphical	yes	yes	no	?	?
Statistical	yes	no	no	no	yes

Figure 1: Credits: Towards Causal Representation Learning - Schölkopf et al. (2021)

- **Statistical:** associations like $p(\mathbf{X}, \mathbf{Y}) \rightarrow p(\mathbf{Y}|\mathbf{X}) \cdot p(\mathbf{X})$

Levels of modelization

Model	Predict in i.i.d. setting	Predict under distr. shift/intervention	Answer counter-factual questions	Obtain physical insight	Learn from data
Mechanistic/physical	yes	yes	yes	yes	?
Structural causal	yes	yes	yes	?	?
Causal graphical	yes	yes	no	?	?
Statistical	yes	no	no	no	yes

Figure 1: Credits: Towards Causal Representation Learning - Schölkopf et al. (2021)

- **Statistical:** associations like $p(\mathbf{X}, \mathbf{Y}) \rightarrow p(\mathbf{Y}|\mathbf{X}) \cdot p(\mathbf{X})$
- **Causal Graphical:** causal decomposition $p(X_1, ..., X_n) = \prod_i p(X_i|\mathbf{PA}_i)$

Levels of modelization

Model	Predict in i.i.d. setting	Predict under distr. shift/intervention	Answer counter-factual questions	Obtain physical insight	Learn from data
Mechanistic/physical	yes	yes	yes	yes	?
Structural causal	yes	yes	yes	?	?
Causal graphical	yes	yes	no	?	?
Statistical	yes	no	no	no	yes

Figure 1: Credits: Towards Causal Representation Learning - Schölkopf et al. (2021)

- **Statistical:** associations like $p(\mathbf{X}, \mathbf{Y}) \rightarrow p(\mathbf{Y}|\mathbf{X}) \cdot p(\mathbf{X})$
- **Causal Graphical:** causal decomposition $p(X_1, ..., X_n) = \prod_i p(X_i|\mathbf{PA}_i)$
- **Structural Causal:** Structural Causal Models (SCMs) $X_i \leftarrow f_i(\mathbf{PA}_i; U_i)$ where $U_i \perp\!\!\!\perp U_j$

Levels of modelization

Model	Predict in i.i.d. setting	Predict under distr. shift/intervention	Answer counter-factual questions	Obtain physical insight	Learn from data
Mechanistic/physical	yes	yes	yes	yes	?
Structural causal	yes	yes	yes	?	?
Causal graphical	yes	yes	no	?	?
Statistical	yes	no	no	no	yes

Figure 1: Credits: Towards Causal Representation Learning - Schölkopf et al. (2021)

- **Statistical:** associations like $p(\mathbf{X}, \mathbf{Y}) \to p(\mathbf{Y}|\mathbf{X}) \cdot p(\mathbf{X})$
- **Causal Graphical:** causal decomposition $p(X_1, ..., X_n) = \prod_i p(X_i|\mathbf{PA}_i)$
- **Structural Causal:** Structural Causal Models (SCMs) $X_i \leftarrow f_i(\mathbf{PA}_i; U_i)$ where $U_i \perp\!\!\!\perp U_j$
- **Physical:** differential equations $i\hbar\partial_t\psi = \hat{H}\,\psi$

Levels of Reality

In many cases, we can refer to a finer and to a coarser level of representation

$$T_b \to T_t$$

T_b : bottom theory, fine-grained, referred to low-level objects
T_t : top theory, coarse-grained, associated with high-level entities
De Haro, *Towards a theory of emergence for the physical sciences* (2019)

Example: Ising Model of Ferromagnets

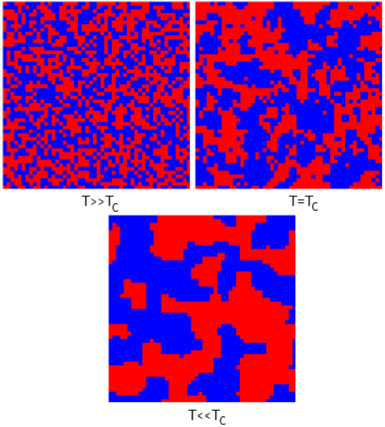


Figure 2: T_c denotes a coarse-level parameter of control of low-level configurations.

In many cases, we can refer to a finer and to a coarser level of representation

$$T_b \to T_t$$

T_b : bottom theory, fine-grained, referred to low-level objects
T_t : top theory, coarse-grained, associated with high-level entities
De Haro, *Towards a theory of emergence for the physical sciences* (2019)

Our case study:

1. We consider **high-level entities** which originate **a lower-level representation**

Example: Ising Model of Ferromagnets

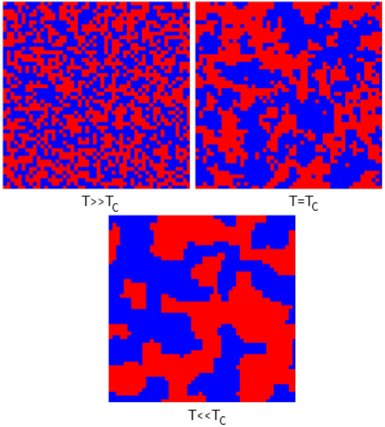


Figure 2: T_c denotes a coarse-level parameter of control of low-level configurations.

Levels of Reality

In many cases, we can refer to a finer and to a coarser level of representation

$$T_b \rightarrow T_t$$

T_b : bottom theory, fine-grained, referred to low-level objects
T_t : top theory, coarse-grained, associated with high-level entities

De Haro, *Towards a theory of emergence for the physical sciences* (2019)

Our case study:

1. We consider **high-level entities** which originate **a lower-level representation**
2. We require that such a map **exists** and can be **inferred**

Example: Ising Model of Ferromagnets

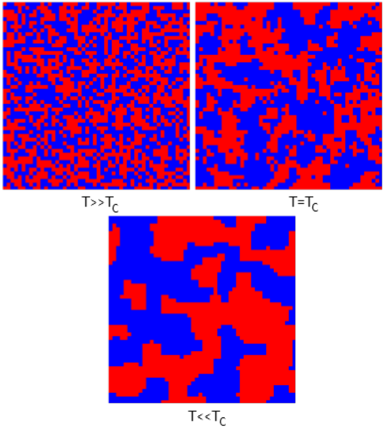


Figure 2: T_c denotes a coarse-level parameter of control of low-level configurations.

Levels of Reality

In many cases, we can refer to a finer and to a coarser level of representation

$$T_b \to T_t$$

T_b : bottom theory, fine-grained, referred to low-level objects
T_t : top theory, coarse-grained, associated with high-level entities
De Haro, *Towards a theory of emergence for the physical sciences* (2019)

Our case study:

1. We consider **high-level entities** which originate **a lower-level representation**
2. We require that such a map **exists** and can be **inferred**
3. We try to **learn from data high-level entities/representations**, but in cases where we have control

Example: Ising Model of Ferromagnets

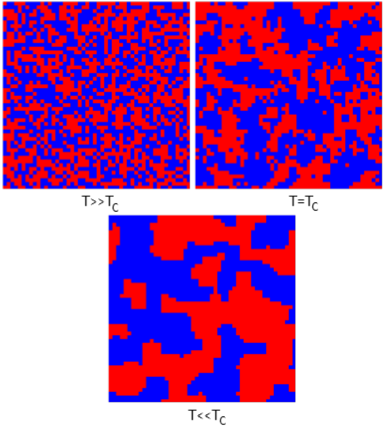


Figure 2: T_c denotes a coarse-level parameter of control of low-level configurations.

Generative Models

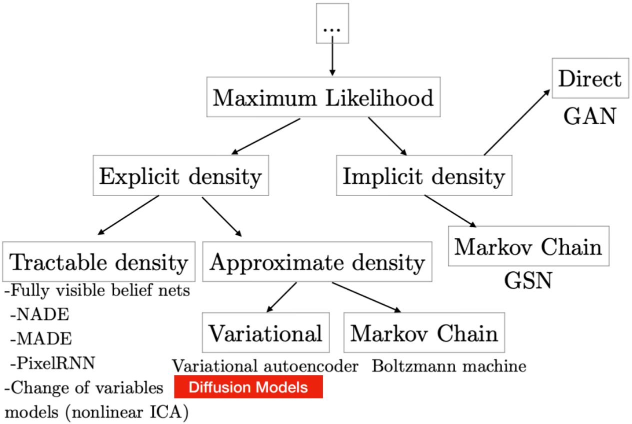


Figure 3: Credits: Dall'Asen N., SML-Journal Club presentation.

Our focus: Variational Autoencoders (VAEs)

VAEs learn a generative model through **latent variables**. This method follows from lower-bounding the log-likelihood of the observed data and introducing variational inference.

Kingma and Welling, Autoencoding Variational Bayes (2014)

VAEs learn a generative model through **latent variables**. This method follows from lower-bounding the log-likelihood of the observed data and introducing variational inference.

Kingma and Welling, Autoencoding Variational Bayes (2014)

$$p(\mathbf{x}) = \int p^*(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z} = \mathbb{E}_{p(\mathbf{z})}[p^*(\mathbf{x}|\mathbf{z})] \quad \text{where } \mathbf{x} \in \mathbb{R}^D \text{ and } \mathbf{z} \in \mathbb{R}^k$$

VAEs learn a generative model through **latent variables**. This method follows from lower-bounding the log-likelihood of the observed data and introducing variational inference.

Kingma and Welling, Autoencoding Variational Bayes (2014)

$$p(\mathbf{x}) = \int p^*(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z} = \mathbb{E}_{p(\mathbf{z})}[p^*(\mathbf{x}|\mathbf{z})] \quad \text{where } \mathbf{x} \in \mathbb{R}^D \text{ and } \mathbf{z} \in \mathbb{R}^k$$

ELBO from likelihood:

$$\log p_\theta(\mathbf{x}) = \log \int p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z}$$

VAEs learn a generative model through **latent variables**. This method follows from lower-bounding the log-likelihood of the observed data and introducing variational inference.

Kingma and Welling, Autoencoding Variational Bayes (2014)

$$p(\mathbf{x}) = \int p^*(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z} = \mathbb{E}_{p(\mathbf{z})}[p^*(\mathbf{x}|\mathbf{z})] \quad \text{where } \mathbf{x} \in \mathbb{R}^D \text{ and } \mathbf{z} \in \mathbb{R}^k$$

ELBO from likelihood:

$$\begin{aligned} \log p_\theta(\mathbf{x}) &= \log \int p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z} \\ &= \log \int p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})\frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})}\mathrm{d}\mathbf{z} \end{aligned}$$

VAEs learn a generative model through **latent variables**. This method follows from lower-bounding the log-likelihood of the observed data and introducing variational inference.

Kingma and Welling, Autoencoding Variational Bayes (2014)

$$p(\mathbf{x}) = \int p^*(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z} = \mathbb{E}_{p(\mathbf{z})}[p^*(\mathbf{x}|\mathbf{z})] \quad \text{where } \mathbf{x} \in \mathbb{R}^D \text{ and } \mathbf{z} \in \mathbb{R}^k$$

ELBO from likelihood:

$$\begin{aligned}
\log p_\theta(\mathbf{x}) &= \log \int p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z} \\
&= \log \int p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})\frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})}\mathrm{d}\mathbf{z} \\
&\geq \int q_\phi(\mathbf{z}|\mathbf{x}) \log p_\theta(\mathbf{x}|\mathbf{z})\frac{p(\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})}\mathrm{d}\mathbf{z}
\end{aligned}$$

VAEs learn a generative model through **latent variables**. This method follows from lower-bounding the log-likelihood of the observed data and introducing variational inference.

Kingma and Welling, Autoencoding Variational Bayes (2014)

$$p(\mathbf{x}) = \int p^*(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z} = \mathbb{E}_{p(\mathbf{z})}[p^*(\mathbf{x}|\mathbf{z})] \quad \text{where } \mathbf{x} \in \mathbb{R}^D \text{ and } \mathbf{z} \in \mathbb{R}^k$$

ELBO from likelihood:

$$\begin{aligned}
\log p_\theta(\mathbf{x}) &= \log \int p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z} \\
&= \log \int p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})\frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})}\mathrm{d}\mathbf{z} \\
&\geq \int q_\phi(\mathbf{z}|\mathbf{x}) \log p_\theta(\mathbf{x}|\mathbf{z})\frac{p(\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})}\mathrm{d}\mathbf{z} \\
&= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z}))
\end{aligned}$$

VAEs learn a generative model through **latent variables**. This method follows from lower-bounding the log-likelihood of the observed data and introducing variational inference.

Kingma and Welling, Autoencoding Variational Bayes (2014)

$$p(\mathbf{x}) = \int p^*(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z} = \mathbb{E}_{p(\mathbf{z})}[p^*(\mathbf{x}|\mathbf{z})] \quad \text{where } \mathbf{x} \in \mathbb{R}^D \text{ and } \mathbf{z} \in \mathbb{R}^k$$

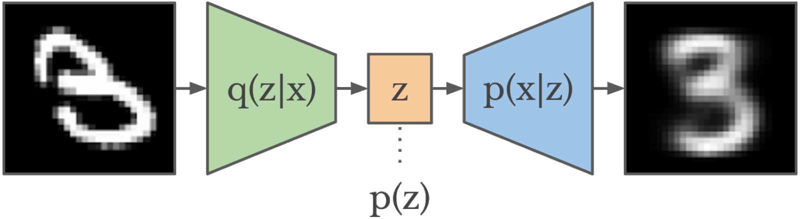
ELBO from likelihood:

$$\begin{aligned} \log p_\theta(\mathbf{x}) &= \log \int p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})\mathrm{d}\mathbf{z} \\ &= \log \int p_\theta(\mathbf{x}|\mathbf{z})p(\mathbf{z})\frac{q_\phi(\mathbf{z}|\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})}\mathrm{d}\mathbf{z} \\ &\geq \int q_\phi(\mathbf{z}|\mathbf{x}) \log p_\theta(\mathbf{x}|\mathbf{z})\frac{p(\mathbf{x})}{q_\phi(\mathbf{z}|\mathbf{x})}\mathrm{d}\mathbf{z} \\ &= \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[p_\theta(\mathbf{x}|\mathbf{z})] - \mathrm{KL}(q_\phi(\mathbf{z}|\mathbf{x})||p(\mathbf{z})) \end{aligned}$$

where $p_\theta(\mathbf{x}|\mathbf{z})$ is our generative ansatz, $q_\phi(\mathbf{z}|\mathbf{x})$ is the approximate posterior, and $p(\mathbf{z})$ is the prior for the model. Learning parameters θ and ϕ.

$$\text{factorization}: \quad p_\theta(\mathbf{x}|\mathbf{z}) = \frac{q_\phi(\mathbf{z}|\mathbf{x})p(\mathbf{x})}{p(\mathbf{x})} \qquad \text{variational}: \quad q_\phi(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\mathbf{z}|\mu, \sigma)$$

VAEs in Deep Learning



☐ We can sample from $p(\mathbf{z})$ and create new examples.

☐ ELBO only lower-bounds the log-likelihood, but it has good properties when $\mathbf{z} \in \mathbb{R}^D$

Reizinger: Embracing the gap: VAEs perform independent mechanisms analysis (2022)

☐ Endless number of variants:

- β-VAEs
- Info-VAEs
- Total-Correlation VAEs
- DIP-VAEs
- Regularized-AEs
- Factor-VAE
- HVAEs
- JL1-VAEs
- ...

Disentangled mechanisms

For each datum $\mathbf{x}$, we can associate a set of elements $\mathbf{g}$ (even stochastic) which describe it in an approximate way.

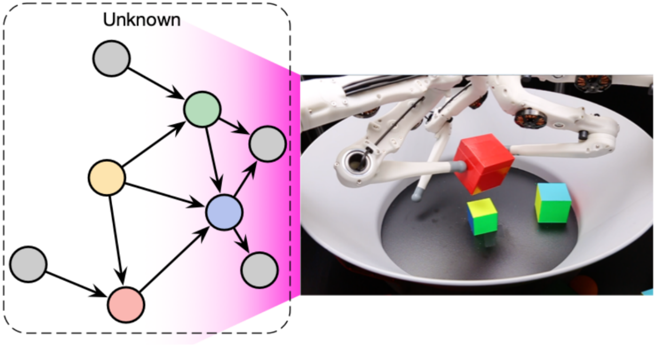
Figure 4: Example of a **datum** to which we associate a sets of **concepts** which describe it.

$$\text{binding}: \quad i : \mathbf{X} \to \mathbf{G}$$

Achille and Soatto, On the Learnability of Physical Concepts: Can a Neural Network Understand What's Real? (2022).

2. The generative mechanism

In simple cases, **all possible variations** on $\mathbf{X}$ can be reconducted to changes on $\mathbf{G}$ + noise. E.g., synthetic datasets, robotic systems, virtual world, etc.



$$\text{generative process}: \quad \mathbf{g}: (\mathbf{G}, \mathbf{N}) \rightarrow \mathbf{X}$$

where $\mathbf{N}$ is a noise term (or *nuisance*). $\mathbf{G}$ are called **generative factors**.

Independent Mechanisms

- We ground our construction on a Causal Perspective - Schölkopf et al. (2021)
  We look at DAGs: $p(\mathbf{G}) = \prod_i p(G_i|\mathbf{PA}_i)$

The decomposition of a DAG implies a structure of statistical independence among variables ($i \neq j$):

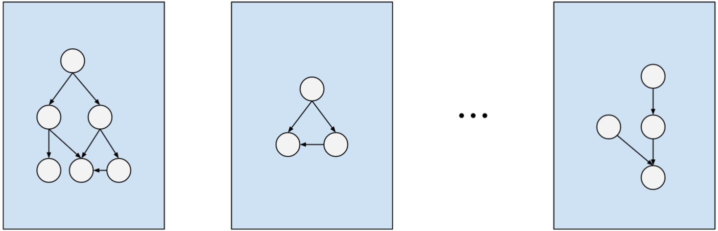
$$P(G_i|\mathbf{PA}_i) \perp\!\!\!\perp P(G_j|\mathbf{PA}_j)$$

Independent Mechanisms

- We ground our construction on a Causal Perspective - Schölkopf et al. (2021)
  We look at DAGs: $p(\mathbf{G}) = \prod_i p(G_i|\mathbf{PA}_i)$

The decomposition of a DAG implies a structure of statistical independence among variables ($i \neq j$):

$$P(G_i|\mathbf{PA}_i) \perp\!\!\!\perp P(G_j|\mathbf{PA}_j)$$



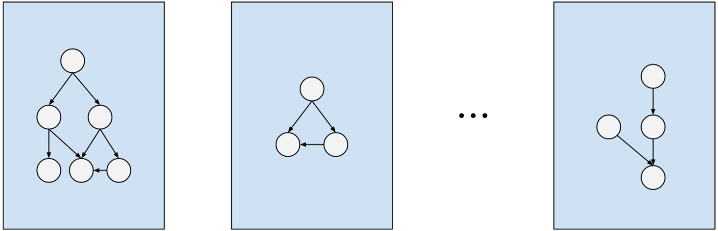
1. **no influence:** changing one mechanism $P(G_i|\mathbf{PA}_i)$ does not change other mechanisms $P(G_j|\mathbf{PA}_j)$;

Independent Mechanisms

- We ground our construction on a Causal Perspective - Schölkopf et al. (2021)
  We look at DAGs: $p(\mathbf{G}) = \prod_i p(G_i|\mathbf{PA}_i)$

The decomposition of a DAG implies a structure of statistical independence among variables ($i \neq j$):

$$P(G_i|\mathbf{PA}_i) \perp\!\!\!\perp P(G_j|\mathbf{PA}_j)$$



1. **no influence:** changing one mechanism $P(G_i|\mathbf{PA}_i)$ does not change other mechanisms $P(G_j|\mathbf{PA}_j)$;
2. **no information:** knowing some other mechanisms $P(G_i|\mathbf{PA}_i)$ does not give us information about a mechanism $P(G_j|\mathbf{PA}_j)$.

Disentangled Mechanisms

- We refer to the simpler, non-trivial case of **single disentangled generative factors**.

- We refer to the simpler, non-trivial case of **single disentangled generative factors**.
- It is represented as a set of **independent factors** $\mathbf{G} = (G_1, ..., G_K)$

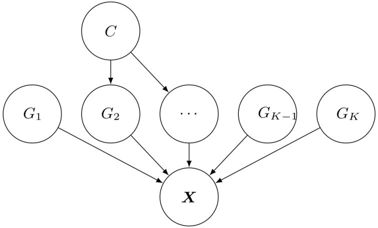
Disentangled Mechanisms

- We refer to the simpler, non-trivial case of **single disentangled generative factors**.
- It is represented as a set of **independent factors** $\mathbf{G} = (G_1, ..., G_K)$
- We also assume that exist confounders $\mathbf{C} = (C_1, ..., C_L)$ which allow for statistical dependencies on $\mathbf{G}$

Disentangled Mechanisms

- We refer to the simpler, non-trivial case of **single disentangled generative factors**.
- It is represented as a set of **independent factors** $\mathbf{G} = (G_1, ..., G_K)$
- We also assume that exist confounders $\mathbf{C} = (C_1, ..., C_L)$ which allow for statistical dependencies on $\mathbf{G}$

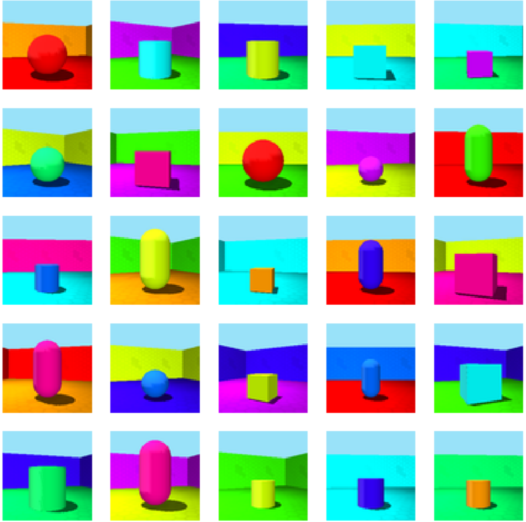
$$\text{generative process}: \quad \mathbf{C} \to \mathbf{G} \to X$$



Credits: Suter *et al.*, Robustly Disentangled Causal Mechanisms (2019)

What are the disentangled factors?

☐ 3D-Shapes dataset.



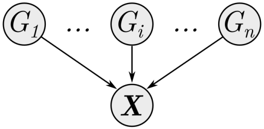
G_1= **floor hue**: 10 values linearly spaced in [0, 1]
G_2= **wall hue**: 10 values linearly spaced in [0, 1]
G_3 = **object hue**: 10 values linearly spaced in [0, 1]
G_4 = **scale**: 8 values linearly spaced in [0, 1]
G_5 = **shape**: 4 values in [0, 1, 2, 3]
G_6 = **orientation**: 15 values linearly spaced in [-30, 30]

Formal Model: Disentangled Causal Mechanism

- ☐ Several **generative factors** $\mathbf{G} = (G_1, \ldots, G_K)$

Structural causal model (SCM), adapted from Suter *et al.* (2019).

- Several **generative factors** $\mathbf{G} = (G_1, \ldots, G_K)$
- They jointly give rise to a **datum** $\mathbf{X}$



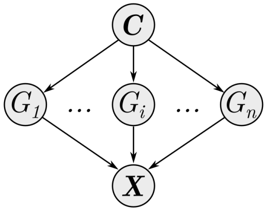
Structural causal model (SCM), adapted from Suter *et al.* (2019).

Formal Model: Disentangled Causal Mechanism

☐ Several **generative factors** $\mathbf{G} = (G_1, \ldots, G_K)$

☐ They jointly give rise to a **datum** $\mathbf{X}$

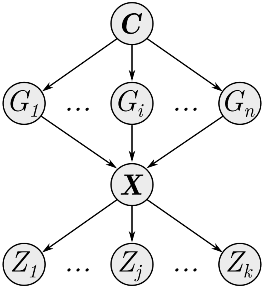
☐ Factors $\mathbf{G}$ may be correlated because of **confounds** $\mathbf{C}$, but are **disentangled** in the sense that **they can be independently manipulated**



Structural causal model (SCM), adapted from Suter *et al.* (2019).

Formal Model: Disentangled Causal Mechanism

- ☐ Several generative factors $\mathbf{G} = (G_1, \ldots, G_K)$
- ☐ They jointly give rise to a datum $\mathbf{X}$
- ☐ Factors $\mathbf{G}$ may be correlated because of confounds $\mathbf{C}$, but are disentangled in the sense that **they can be independently manipulated**
- ☐ Model acquires latent factors $Z_1, \ldots, Z_k$

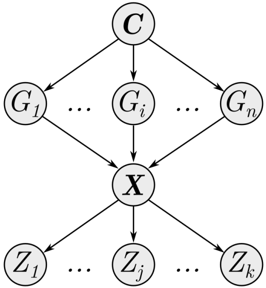


Structural causal model (SCM), adapted from Suter *et al.* (2019).

Formal Model: Disentangled Causal Mechanism

- Several **generative factors** $\mathbf{G} = (G_1, \dots, G_K)$
- They jointly give rise to a **datum** $\mathbf{X}$
- Factors $\mathbf{G}$ may be correlated because of **confounds** $\mathbf{C}$, but are **disentangled** in the sense that **they can be independently manipulated**
- Model acquires **latent factors** $Z_1, \dots, Z_k$
- SCM formulation:

$$\mathbf{C} \leftarrow \mathbf{N}_c$$
$$G_i \leftarrow f_i(\mathbf{PA}_i^C, N_i), \quad \mathbf{PA}_i^C \subset \{C_1, \dots, C_L\},\ i = 1, \dots, K$$
$$\mathbf{X} \leftarrow g(\mathbf{G}, N_x)$$
$$Z_j \leftarrow e_j(\mathbf{X}, (N_z)_j)$$



Structural causal model (SCM), adapted from Suter *et al.* (2019).

Some Properties of the Disentangled Causal Mechanism

Proposition 1 (from Suter *et al.* (2019)): *A disentangled causal process fulfills the following properties:*

Some Properties of the Disentangled Causal Mechanism

Proposition 1 (from Suter *et al.* (2019)): *A disentangled causal process fulfills the following properties:*

1. $p(\mathbf{x}|\mathbf{g})$ describes a causal mechanism invariant to changes in the distribution of $p(g_i)$

Some Properties of the Disentangled Causal Mechanism

Proposition 1 (from Suter *et al.* (2019)): *A disentangled causal process fulfills the following properties:*

1. $p(\mathbf{x}|\mathbf{g})$ describes a causal mechanism invariant to changes in the distribution of $p(g_i)$
2. In general, the latent factors can be dependent

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

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

$$G_i \perp\!\!\!\perp G_j | \mathbf{C} \; \forall i \neq j$$

Some Properties of the Disentangled Causal Mechanism

Proposition 1 (from Suter *et al.* (2019)): *A disentangled causal process fulfills the following properties:*

1. $p(\mathbf{x}|\mathbf{g})$ describes a causal mechanism invariant to changes in the distribution of $p(g_i)$
2. In general, the latent factors can be dependent

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

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

$$G_i \perp\!\!\!\perp G_j | \mathbf{C} \; \forall i \neq j$$

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

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

Some Properties of the Disentangled Causal Mechanism

Proposition 1 (from Suter *et al.* (2019)): *A disentangled causal process fulfills the following properties:*

1. $p(\mathbf{x}|\mathbf{g})$ describes a causal mechanism invariant to changes in the distribution of $p(g_i)$
2. In general, the latent factors can be dependent

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

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

$$G_i \perp\!\!\!\perp G_j|\mathbf{C} \; \forall i \neq j$$

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

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

4. The remaining components of $\mathbf{G}$, i.e. $\mathbf{G}_{-j}$, are a **valid adjustment set** to estimate interventional effects from G_j to $\mathbf{X}$ based on observational data, i.e.,

$$p(\mathbf{x}|\mathrm{do}(G_j \leftarrow g_j^{\triangle})) = \int p(\mathbf{x}|g_j^{\triangle}, \mathbf{g}_{-j})p(\mathbf{g}_{-j})\mathrm{d}\mathbf{g}_{-j}$$

Some Properties of the Disentangled Causal Mechanism

Proposition 1 (from Suter *et al.* (2019)): *A disentangled causal process fulfills the following properties:*

1. $p(\mathbf{x}|\mathbf{g})$ describes a causal mechanism invariant to changes in the distribution of $p(g_i)$
2. In general, the latent factors can be dependent

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

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

$$G_i \perp\!\!\!\perp G_j|\mathbf{C} \ \forall i \neq j$$

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

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

4. The remaining components of $\mathbf{G}$, i.e. $\mathbf{G}_{-j}$, are a **valid adjustment set** to estimate interventional effects from G_j to $\mathbf{X}$ based on observational data, i.e.,

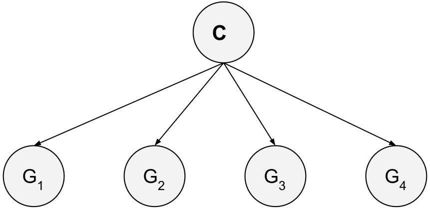
$$p(\mathbf{x}|\mathrm{do}(G_j \leftarrow g_j^{\triangle})) = \int p(\mathbf{x}|g_j^{\triangle}, \mathbf{g}_{-j})p(\mathbf{g}_{-j})\mathrm{d}\mathbf{g}_{-j}$$

5. If there is no confounding, conditioning is sufficient to obtain the post-interventional distribution of $\mathbf{X}$:

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

A remark on do-calculus

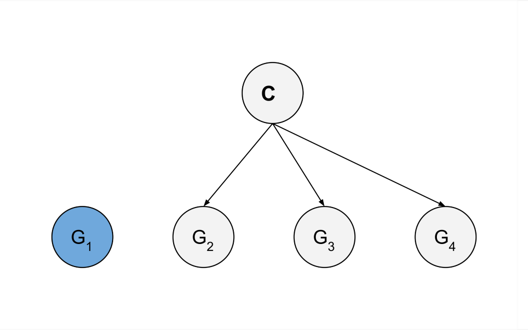
$$p(\mathbf{G}_{-j}, \mathbf{C} | \mathrm{do}(G_j \leftarrow g_j)) \neq p(\mathbf{G}_{-j}, \mathbf{C} | g_j)$$



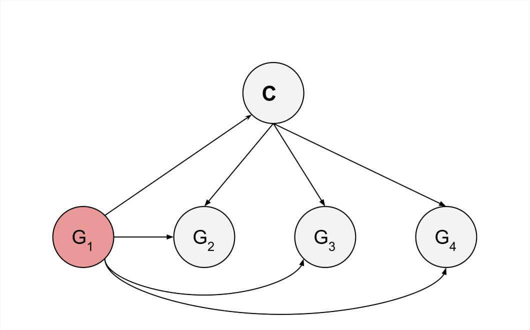
A remark on do-calculus

$$p(\mathbf{G}_{-j}, \mathbf{C} | \mathrm{do}(G_j \leftarrow g_j)) \neq p(\mathbf{G}_{-j}, \mathbf{C} | g_j)$$

Intervened



Conditioned



Disentangled Representations

Disentangled Representations

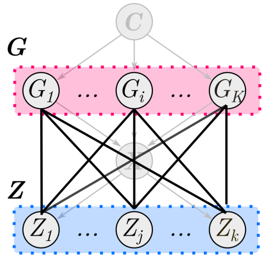
We define the interventional effect of a *group* of generative factors $\mathbf{G}_I$ on the implied latent space encodings $\mathbf{Z}_J$ with proxy posterior $q_\phi(\mathbf{z}|\mathbf{x})$ from a VAE (or variant), where $I \subset \{1, \ldots, K\}$ and $J \subset \{1, \ldots, k\}$ as:

$$p(\mathbf{z}_J|\mathrm{do}(\mathbf{G}_I \leftarrow \mathbf{G}_I^{\triangle})) = \int q_\phi(\mathbf{z}_J|\mathbf{x})p(\mathbf{x}|\mathrm{do}(\mathbf{G}_J \leftarrow \mathbf{g}_J^{\triangle}))\mathrm{d}\mathbf{x}$$

Meaning of a disentangled representation:

- Variations of a single latent factor Z_j depends on at most one generative factor G_i variations:

$$Z_j \leftarrow \alpha_j(g_{\pi(j)}, N_j)$$



Entangled representations.

Disentangled Representations

We define the interventional effect of a *group* of generative factors $\mathbf{G}_I$ on the implied latent space encodings $\mathbf{Z}_J$ with proxy posterior $q_\phi(\mathbf{z}|\mathbf{x})$ from a VAE (or variant), where $I \subset \{1, \ldots, K\}$ and $J \subset \{1, \ldots, k\}$ as:

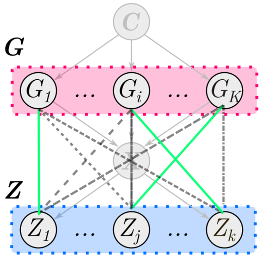
$$p(\mathbf{z}_J|\mathrm{do}(\mathbf{G}_I \leftarrow \mathbf{G}_I^{\triangle})) = \int q_\phi(\mathbf{z}_J|\mathbf{x})p(\mathbf{x}|\mathrm{do}(\mathbf{G}_J \leftarrow \mathbf{g}_J^{\triangle}))\mathrm{d}\mathbf{x}$$

Meaning of a disentangled representation:

- Variations of a single latent factor Z_j depends on at most one generative factor G_i variations:

$$Z_j \leftarrow \alpha_j(g_{\pi(j)}, N_j)$$

- There can be different copies of the same generative factor G_i, but disentanglement still holds.



Disentangled representations.

Disentangled Representations

We define the interventional effect of a *group* of generative factors $\mathbf{G}_I$ on the implied latent space encodings $\mathbf{Z}_J$ with proxy posterior $q_\phi(\mathbf{z}|\mathbf{x})$ from a VAE (or variant), where $I \subset \{1,\dots,K\}$ and $J \subset \{1,\dots,k\}$ as:

$$p(\mathbf{z}_J|\mathrm{do}(\mathbf{G}_I \leftarrow \mathbf{G}_I^{\triangle})) = \int q_\phi(\mathbf{z}_J|\mathbf{x})p(\mathbf{x}|\mathrm{do}(\mathbf{G}_J \leftarrow \mathbf{g}_J^{\triangle}))\mathrm{d}\mathbf{x}$$

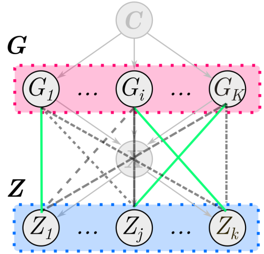
Meaning of a disentangled representation:

- Variations of a single latent factor Z_j depends on at most one generative factor G_i variations:

$$Z_j \leftarrow \alpha_j(g_{\pi(j)}, N_j)$$

- There can be different copies of the same generative factor G_i, but disentanglement still holds.

where α_j is a general (non-linear) function for $j = 1,\dots,d$,
$\pi : \{1,\dots,d\} \to \{1,\dots,K\} \cup \varnothing$ an element-wise correspondence,
$\alpha_j(g_\varnothing, N_j) = \alpha_j(N_j)$,
and N_j are independent noise terms.



Disentangled representations.

There have been many proposals to measure it, but none of them is optimal

Do and Tran, Theory and Evaluation for Learning Disentangled Representations (2020);

Carbonneau *et al.*, Measuring Disentanglement: A Review of Metrics (2022).

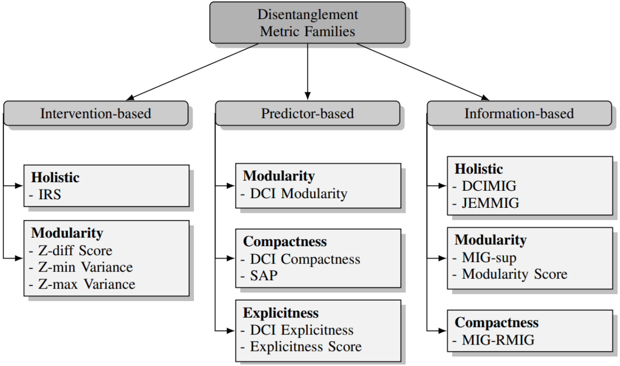


Figure 5: Taxonomy of (some) known metrics.

How to measure Disentanglement of the representations

A quick look at IRS (**Interventional Robustness Score**), from Suter *et al.* (2019):

$$PIDA(l|i,j) := d\Big(\mathbb{E}[z_l\,|\mathrm{do}(G_i \leftarrow g_i)],\, \mathbb{E}[z_l\,|\mathrm{do}(G_i \leftarrow g_i), \mathrm{do}(G_j \leftarrow g_j)]\Big)$$

How to measure Disentanglement of the representations

A quick look at IRS (**Interventional Robustness Score**), from Suter *et al.* (2019):

$$PIDA(l|i,j) := d\Big(\mathbb{E}[z_l\,|\mathrm{do}(G_i \leftarrow g_i)],\, \mathbb{E}[z_l\,|\mathrm{do}(G_i \leftarrow g_i), \mathrm{do}(G_j \leftarrow g_j)]\Big)$$

and when:

$$PIDA \to 0 \;\; \forall l \implies IRS \to 0$$

Learning Disentangled Representations

Can we learn disentangled representations in unsupervised settings? **No**, (i) without implicit bias or (ii) without supervision.

Locatello *et al.*, Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations (2019)

Can we learn disentangled representations in unsupervised settings? **No**, (i) without implicit bias or (ii) without supervision.

Locatello *et al.*, Challenging Common Assumptions in the Unsupervised Learning of Disentangled Representations (2019)

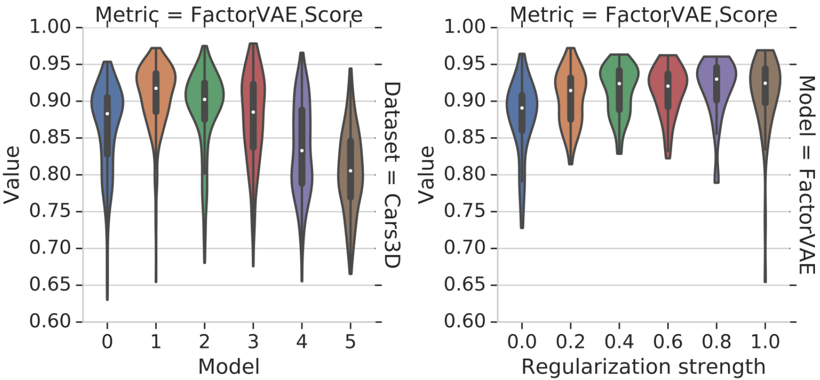


Figure 6: Drastical variations of the obtained disentangled (a) upon changing the VAE variant and (b) the regularization strength.

In turn, we can include (weak)-supervision, such as:

Learning Disentangled Representations

In turn, we can include (weak)-supervision, such as:

- **Generative Factors supervision**, only small amounts are sufficient to achieve better-disentangled representations;

In turn, we can include (weak)-supervision, such as:

- **Generative Factors supervision**, only small amounts are sufficient to achieve better-disentangled representations;
- **Match pairing**, saying on couples (x, x') which generative factors coincide;

In turn, we can include (weak)-supervision, such as:

- **Generative Factors supervision**, only small amounts are sufficient to achieve better-disentangled representations;
- **Match pairing**, saying on couples (x, x') which generative factors coincide;
- **Rank pairing**, saying for a couple (x, x') the order relation, such as $(g_i > g_i') =$True.

In turn, we can include (weak)-supervision, such as:

- **Generative Factors supervision**, only small amounts are sufficient to achieve better-disentangled representations;
- **Match pairing**, saying on couples (x, x') which generative factors coincide;
- **Rank pairing**, saying for a couple (x, x') the order relation, such as $(g_i > g_i') =$True.
- **Transferring properties**, changing in a datum x some factors based on x', and matching the reconstruction.

In turn, we can include (weak)-supervision, such as:

- **Generative Factors supervision**, only small amounts are sufficient to achieve better-disentangled representations;
- **Match pairing**, saying on couples (x, x') which generative factors coincide;
- **Rank pairing**, saying for a couple (x, x') the order relation, such as $(g_i > g_i') =$True.
- **Transferring properties**, changing in a datum x some factors based on x', and matching the reconstruction.

Hungry for Theorems? Check Shu *et al.*, Weakly Supervised Disentanglement with Guarantees (2020).

Other formulations

Formal definitions of disentangled representations:

- ☑ Causal Disentanglement
- ☐ Identifiability in Non-linear Independent Component Analysis (ICA)
- ☐ Group-based Disentanglement

Identifiability in ICA

Identifiability $\implies$ retrieving the independent component generating the input

Identifiability in ICA

Identifiability $\implies$ retrieving the independent component generating the input

Definition 1.(Identifiability) *Independent component analysis in* $(\mathcal{F}, \mathcal{P})$ *is identifiable up to* $\mathcal{S}$ *if for functions* $f, f' \in \mathcal{F}$ *and distributions* $\mathbb{P}, \mathbb{P}' \in \mathcal{P}$ *the relation*

$$f(s) =^{\mathcal{D}} f'(s') \quad \text{where } s \sim \mathbb{P} \text{ and } s' \sim \mathbb{P}'$$

implies that there is $h \in \mathcal{S}$ *that* $h = f'^{-1} \circ f$ *on the support of* $\mathbb{P}$.

Buchholz *et al.*, Function Classes for Identifiable Nonlinear Independent Component Analysis (2022).

Identifiability in ICA

Identifiability $\implies$ retrieving the independent component generating the input

Definition 1.(Identifiability) *Independent component analysis in* $(\mathcal{F}, \mathcal{P})$ *is identifiable up to* $\mathcal{S}$ *if for functions* $f, f' \in \mathcal{F}$ *and distributions* $\mathbb{P}, \mathbb{P}' \in \mathcal{P}$ *the relation*

$$f(s) =^{\mathcal{D}} f'(s') \quad \text{where } s \sim \mathbb{P} \text{ and } s' \sim \mathbb{P}'$$

implies that there is $h \in \mathcal{S}$ *that* $h = f'^{-1} \circ f$ *on the support of* $\mathbb{P}$*.*

Buchholz *et al.*, Function Classes for Identifiable Nonlinear Independent Component Analysis (2022).

Causal Disentanglement and Identifiability in non-linear ICA have been reconciled:

- Theorem 11 in Wang and Jordan, Desiderata for Representation Learning: a Causal Perspective (2021). Identifiability up to permutations $h \in \mathcal{S}_{perm}$.

Group-based Disentanglement

There exist a product group $\mathbb{G} = \mathbb{G}_1 \times \ldots \times \mathbb{G}_K$ acting on $\mathbf{G}$. Condition for disentanglement:

Group-based Disentanglement

There exist a product group $\mathbb{G} = \mathbb{G}_1 \times \ldots \times \mathbb{G}_K$ acting on $\mathbf{G}$. Condition for disentanglement:

- The learned map implicitly defines a group $\mathbb{H}$ acting on the representation $\mathbf{Z}$

Group-based Disentanglement

There exist a product group $\mathbb{G} = \mathbb{G}_1 \times \ldots \times \mathbb{G}_K$ acting on $\mathbf{G}$. Condition for disentanglement:

- The learned map implicitly defines a group $\mathbb{H}$ acting on the representation $\mathbf{Z}$
- The map $e \circ g : \mathbf{G} \to \mathbf{Z}$ is **equivariant** between the actions on $\mathbf{G}$ and $\mathbf{Z}$, and

Group-based Disentanglement

There exist a product group $\mathbb{G} = \mathbb{G}_1 \times \ldots \times \mathbb{G}_K$ acting on $\mathbf{G}$. Condition for disentanglement:

- The learned map implicitly defines a group $\mathbb{H}$ acting on the representation $\mathbf{Z}$
- The map $e \circ g : \mathbf{G} \to \mathbf{Z}$ is **equivariant** between the actions on $\mathbf{G}$ and $\mathbf{Z}$, and
- There is a decomposition $\mathbf{Z} = Z_1 \oplus \ldots \oplus Z_d$ such that each Z_i is fixed by the action of all $\mathbb{G}_k$, $k \neq j$ and affected only by $\mathbb{G}_j$.

Higgins *et al.*, Towards a Definition of Disentangled Representations (2018).

Group-based Disentanglement

There exist a product group $\mathbb{G} = \mathbb{G}_1 \times \ldots \times \mathbb{G}_K$ acting on $\mathbf{G}$. Condition for disentanglement:

- The learned map implicitly defines a group $\mathbb{H}$ acting on the representation $\mathbf{Z}$
- The map $e \circ g : \mathbf{G} \to \mathbf{Z}$ is **equivariant** between the actions on $\mathbf{G}$ and $\mathbf{Z}$, and
- There is a decomposition $\mathbf{Z} = Z_1 \oplus \ldots \oplus Z_d$ such that each Z_i is fixed by the action of all $\mathbb{G}_k$, $k \neq j$ and affected only by $\mathbb{G}_j$.

Higgins *et al.*, Towards a Definition of Disentangled Representations (2018).

☐ The group acting on $\mathbf{G}_i$ can be complicated.
☐ There is no statistical notion in this formulation (yet).

Interpretability of the representations

We proposed a definition of **Interpretability** as **alignment** between generative factors and the representations:

- Variations of a single latent factor Z_j depends on at most one generative factor G_i variations:

$$Z_j \leftarrow \alpha_j(g_{\pi(j)}) + N_j$$

Interpretability of the representations

We proposed a definition of **Interpretability** as **alignment** between generative factors and the representations:

- ☐ Variations of a single latent factor Z_j depends on at most one generative factor G_i variations:

$$Z_j \leftarrow \alpha_j(g_{\pi(j)}) + N_j$$

- The map α is monotonic

Interpretability of the representations

We proposed a definition of **Interpretability** as **alignment** between generative factors and the representations:

- Variations of a single latent factor Z_j depends on at most one generative factor G_i variations:

$$Z_j \leftarrow \alpha_j(g_{\pi(j)}) + N_j$$

- The map α is monotonic

where α_j is a **monotonic** function for $j = 1, \dots, d$,
$\pi : \{1, \dots, d\} \to \{1, \dots, K\} \cup \varnothing$ an element-wise correspondence,
$\alpha_j(g_\varnothing) = 0$
and N_j are independent noise terms.

Marconato, Passerini, and Teso, Glancenets: Interpretable, Leak-proof Concept-based Models

Interpretability of the representations

We proposed a definition of **Interpretability** as **alignment** between generative factors and the representations:

- ☐ Variations of a single latent factor Z_j depends on at most one generative factor G_i variations:

$$Z_j \leftarrow \alpha_j(g_{\pi(j)}) + N_j$$

- ■ The map α is monotonic

where α_j is a **monotonic** function for $j = 1, \ldots, d$,
$\pi : \{1, \ldots, d\} \to \{1, \ldots, K\} \cup \varnothing$ an element-wise correspondence,
$\alpha_j(g_\varnothing) = 0$
and N_j are independent noise terms.

Marconato, Passerini, and Teso, Glancenets: Interpretable, Leak-proof Concept-based Models

Identifiability (up to permutations) $\implies$ **Alignment** $\implies$ **Disentanglement**

New directions

- Disentanglement in OOD scenarios: (1) combinatiorial generalization and (2) concept leakage.
- Disentanglement in Real-World scenarios: ViT and stuff like that.
- Learning Causal Mechanisms: integration of interventions in learning.
- Equivariance in representations: Geometric Deep Learning.

Thank you for the attention!

Interested in a thesis?

- Project works in this field
- Connection between **causal** and **group-based** disentanglement
- Unsupervised discovery of **concepts** through Neuro-Symbolic integration

emanuele.marconato@unitn.it