TECHNISCHE UNIVERSITÄT DARMSTADT

AIML Lab

Winter Semester 2025/26 Lecture

Causality for AI & ML

“Causal Representation Learning”

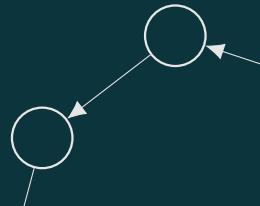
Prof. Dr. Kristian Kersting

Moritz Willig — Today’s speaker

Tim Woydt

Florian Busch

Matej Zečević



Causality in Images

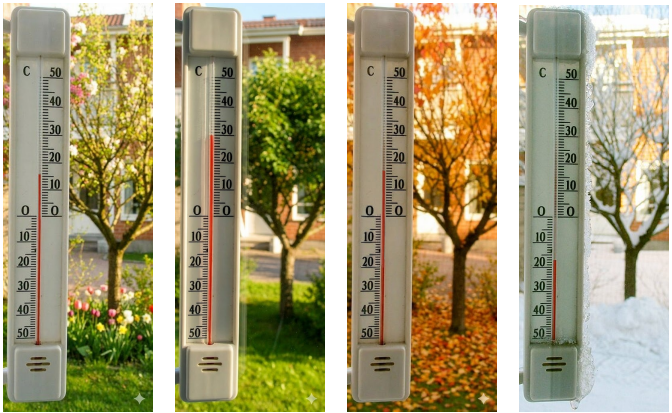
“What is the causality in this image?”

Causality in Images

"What is the causality in this image?"

Causality in Images

"What is the causality in these images?"



rightmost image: https://en.wikipedia.org/wiki/File:Pakkanen.jpg

Causality in Images

Humans have an intuitive ‘high-level’ understanding of the processes depicted in the images.

→ We leverage temporal cues and interventions to disentangle concepts and make sense of the world.

→ We predict underlying dynamics and make inferences about future outcomes.

Causality in Images

Humans have an intuitive ‘high-level’ understanding of the processes depicted in the images.

- → We leverage temporal cues and interventions to disentangle concepts and make sense of the world.
- → We predict underlying dynamics and make inferences about future outcomes.

How can we teach machines to infer the same (causal) dynamics?

Beyond Causal Abstractions

Causal Abstractions commonly match low- and high-level models.

$\rightarrow$ About the *consistency* of variable and intervention mappings.

Beyond Causal Abstractions

Causal Abstractions commonly match low- and high-level models.

$\rightarrow$ About the *consistency* of variable and intervention mappings.

Causal Representation *Learning*:

– Low-level observations - entail no semantically meaningful variables.

Beyond Causal Abstractions

Causal Abstractions commonly match low- and high-level models.
$\rightarrow$ About the *consistency* of variable and intervention mappings.

Causal Representation *Learning*:

- Low-level observations - entail no semantically meaningful variables.
- The high-level causal variables are typically unknown.

Beyond Causal Abstractions

Causal Abstractions commonly match low- and high-level models.
$\rightarrow$ About the *consistency* of variable and intervention mappings.

Causal Representation *Learning*:

- Low-level observations - entail no semantically meaningful variables.
- The high-level causal variables are typically unknown.
- The causal structure is typically unknown (in low- and high-level).

Beyond Causal Abstractions

Causal Abstractions commonly match low- and high-level models.
$\rightarrow$ About the *consistency* of variable and intervention mappings.

Causal Representation *Learning*:

- Low-level observations - entail no semantically meaningful variables.
- The high-level causal variables are typically unknown.
- The causal structure is typically unknown (in low- and high-level).

Task of CRL

Learn the high-level *causal variables* and *structure* from low-level data.

Hierarchy of Causal Tasks

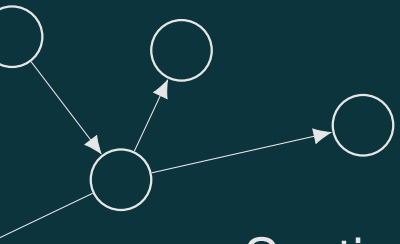
Different tasks in causality operate under different knowledge levels:

Task	Variables Known?	Edges Known?
Effect Identification	✓	✓
Causal Discovery	✓	✗
Causal Representation Learning	✗	✗

Hierarchy of Causal Tasks

Different tasks in causality operate under different knowledge levels:

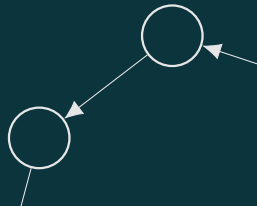
Task	Variables Known?	Edges Known?
Effect Identification	✓	✓
Causal Discovery	✓	✗
Causal Representation Learning	✗	✗
¿Structure Matching?	✗	✓



Section

1

Factors of Variation



Factors of Variation

"What's a cause anyway?"

$\rightarrow$ A constant can never be inferred to be a cause (nor an effect).

Factors of Variation

"What's a cause anyway?"

$\rightarrow$ A constant can never be inferred to be a cause (nor an effect).

We need to show our models enough *variation* to learn the correct causal relations!

Factors of Variation

"What's a cause anyway?"

$\rightarrow$ A constant can never be inferred to be a cause (nor an effect).

We need to show our models enough *variation* to learn the correct causal relations!

What does this mean for the single images shown before?

$\rightarrow$ A model either needs *prior knowledge* or be presented with a *population/series* of images to interpret those.

Factors of Variation

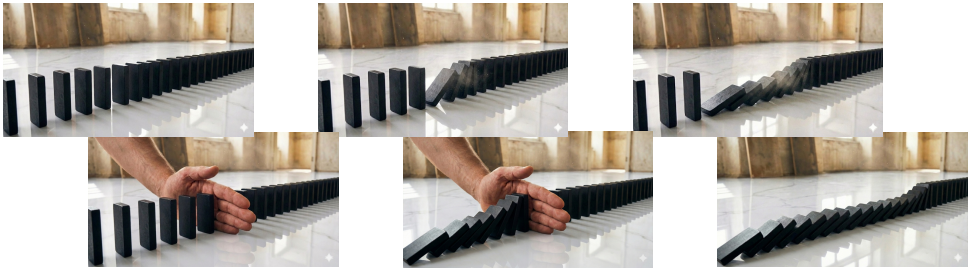
"What's a cause anyway?"

→ A constant can never be inferred to be a cause (nor an effect).

We need to show our models enough *variation* to learn the correct causal relations!

What does this mean for the single images shown before?

→ A model either needs *prior knowledge* or be presented with a *population/series* of images to interpret those.



Entanglement from a Causal Perspective

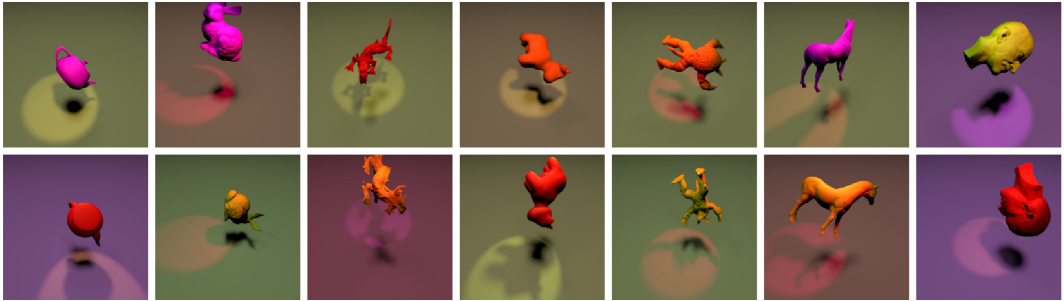
Assume that images consist of **content** (e.g., “this is a dog”) and **style** variability like camera position, lighting, background, . . .

Content is defined as the set of features that are invariant to the specific augmentations applied.

Style is defined as the set of features that are variant (changed) by the augmentations.

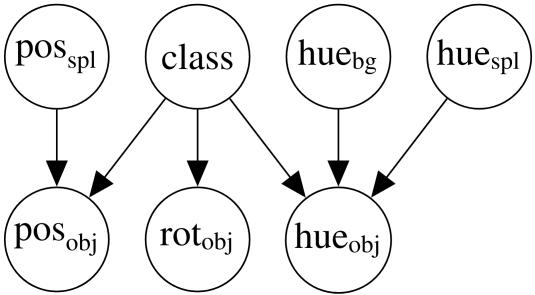
Standard unsupervised learning (like VAEs) cannot guarantee that one part of the learned vector is “content” and the other is “style”.

Causal3DIdent



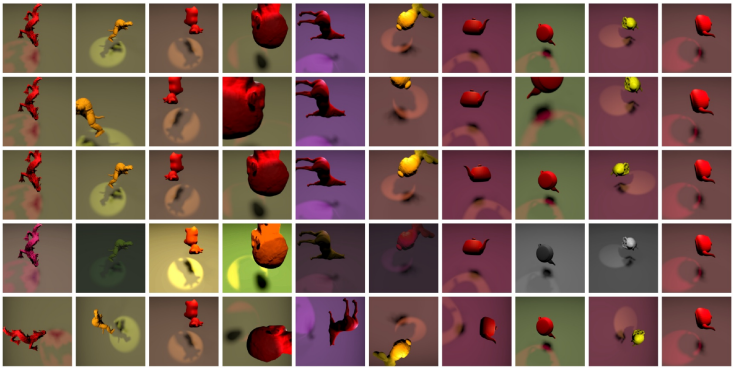
Von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M. and Locatello, F., 2021. Self-supervised learning with data augmentations provably isolates content from style. Advances in neural information processing systems, 34, pp.16451-16467.

Causal3DIdent



Von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M. and Locatello, F., 2021. Self-supervised learning with data augmentations provably isolates content from style. Advances in neural information processing systems, 34, pp.16451-16467.

Causal3DIdent - Augmentations

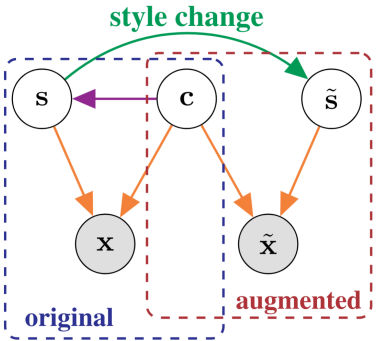


Augmentations (per row): original images, small random crop (+ random flip), large random crop (+ random flip), color distortion (jitter & drop), random rotation.

Von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M. and Locatello, F., 2021. Self-supervised learning with data augmentations provably isolates content from style. Advances in neural information processing systems, 34, pp.16451-16467.

Content-Style Separation

Can we guarantee disentanglement between content and style?



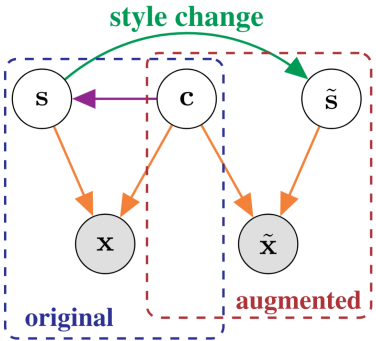
Von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M. and Locatello, F., 2021. Self-supervised learning with data augmentations provably isolates content from style. Advances in neural information processing systems, 34, pp.16451-16467.

Content-Style Separation

Can we guarantee disentanglement between content and style?

(1) Create pairs of images: $(\mathbf{x}, \tilde{\mathbf{x}})$ with shared content $\mathbf{c}$ but different styles $\mathbf{s}, \tilde{\mathbf{s}}$.

(2) Train with generative self-supervised learning (SSL), e.g. an VAE with contrastive learning.



Von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M. and Locatello, F., 2021. Self-supervised learning with data augmentations provably isolates content from style. Advances in neural information processing systems, 34, pp.16451-16467.

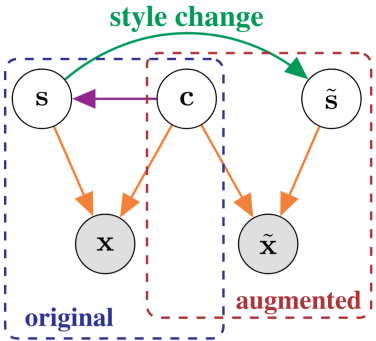
Content-Style Separation

Can we guarantee disentanglement between content and style?

(1) Create pairs of images: $(\mathbf{x}, \tilde{\mathbf{x}})$ with shared content $\mathbf{c}$ but different styles $\mathbf{s}, \tilde{\mathbf{s}}$.

(2) Train with generative self-supervised learning (SSL), e.g. an VAE with contrastive learning.

Intuition: The model learns that 'content' is whatever remains constant when applying augmentations.



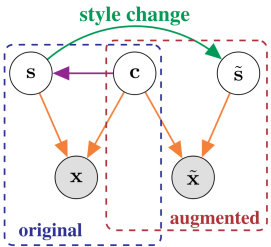
Von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M. and Locatello, F., 2021. Self-supervised learning with data augmentations provably isolates content from style. Advances in neural information processing systems, 34, pp.16451-16467.

Content-Style Separation

InfoNCE contrastive loss:

"Pull together *positive pairs* $\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}'_i$, push apart *negative pairs* $\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}'_j$."

$$\mathcal{L}_{\text{InfoNCE}}(\mathbf{g}; \tau, K) = \mathbb{E}_{\{\mathbf{x}_i\}_{i=1}^K \sim p_{\mathbf{x}}} \left[-\sum_{i=1}^{K} \log \frac{\exp\{\text{sim}(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}'_i)/\tau\}}{\sum_{j=1}^{K} \exp\{\text{sim}(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}'_j)/\tau\}} \right]$$



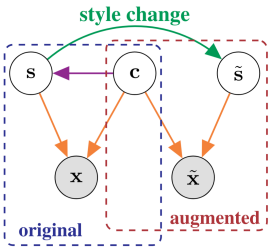
Von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M. and Locatello, F., 2021. Self-supervised learning with data augmentations provably isolates content from style. Advances in neural information processing systems, 34, pp.16451-16467.

Content-Style Separation

InfoNCE contrastive loss:

"Pull together *positive pairs* $\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}'_i$, push apart *negative pairs* $\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}'_j$."

$$\mathcal{L}_{\text{InfoNCE}}(\mathbf{g}; \tau, K) = \mathbb{E}_{\{\mathbf{x}_i\}_{i=1}^K \sim p_{\mathbf{x}}} \left[-\sum_{i=1}^{K} \log \frac{\exp\{\text{sim}(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}'_i)/\tau\}}{\sum_{j=1}^{K} \exp\{\text{sim}(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}'_j)/\tau\}} \right]$$



Training Loss: $\mathcal{L}_{\text{AlignMaxEnt}}(\mathbf{g}) := \mathbb{E}_{(\mathbf{x},\tilde{\mathbf{x}})\sim p_{\mathbf{x},\tilde{\mathbf{x}}}} \left[\|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\tilde{\mathbf{x}})\|_2^2 \right] - H(\mathbf{g}(\mathbf{x}))$

where H is the differential entropy.

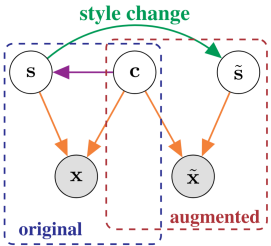
Von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M. and Locatello, F., 2021. Self-supervised learning with data augmentations provably isolates content from style. Advances in neural information processing systems, 34, pp.16451-16467.

Content-Style Separation

InfoNCE contrastive loss:

"Pull together *positive pairs* $\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}'_i$, push apart *negative pairs* $\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}'_j$."

$$\mathcal{L}_{\text{InfoNCE}}(\mathbf{g}; \tau, K) = \mathbb{E}_{\{\mathbf{x}_i\}_{i=1}^K \sim p_{\mathbf{x}}} \left[-\sum_{i=1}^{K} \log \frac{\exp\{\text{sim}(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}'_i)/\tau\}}{\sum_{j=1}^{K} \exp\{\text{sim}(\tilde{\mathbf{z}}_i, \tilde{\mathbf{z}}'_j)/\tau\}} \right]$$



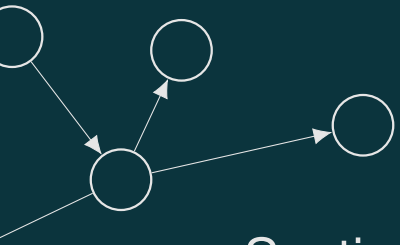
Training Loss: $\mathcal{L}_{\text{AlignMaxEnt}}(\mathbf{g}) := \mathbb{E}_{(\mathbf{x},\tilde{\mathbf{x}}) \sim p_{\mathbf{x},\tilde{\mathbf{x}}}} \left[\|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\tilde{\mathbf{x}})\|_2^2 \right] - H(\mathbf{g}(\mathbf{x}))$

where H is the differential entropy.

[Thm. 4.4, Kügelgen et al.]

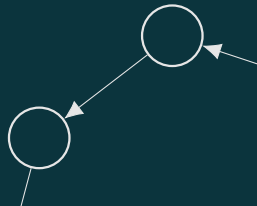
Contrastive Learning with InfoNCE (e.g., SimCLR) asymptotically isolates content from style.

Von Kügelgen, J., Sharma, Y., Gresele, L., Brendel, W., Schölkopf, B., Besserve, M. and Locatello, F., 2021. Self-supervised learning with data augmentations provably isolates content from style. Advances in neural information processing systems, 34, pp.16451-16467.



Section 2

Causal Representation Learning



Latent Causal Process Discovery

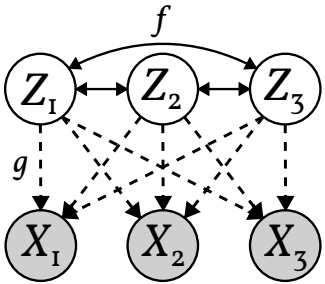
Task of CRL: Recover the latent factors $\mathbf{z} \in \mathbb{R}^n$ (and graph) from an observed signal $\mathbf{x} \in \mathbb{R}^k$.

Assume a **latent process**:

$$\mathbf{z}_i := f_i(\{\mathbf{z}_j \mid \mathbf{z}_j \in \text{pa}(\mathbf{z}_i)\}, \epsilon_i)$$

and a **mixing function**:

$$\mathbf{x}_i := g_i(\mathbf{z})$$



Latent Causal Process Discovery

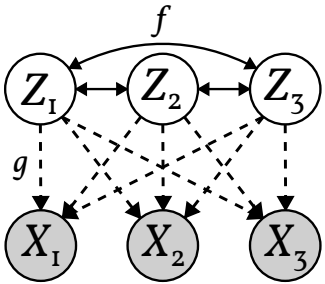
Task of CRL: Recover the latent factors $\mathbf{z} \in \mathbb{R}^n$ (and graph) from an observed signal $\mathbf{x} \in \mathbb{R}^k$.

Assume a **latent process**:

$$\mathbf{z}_i := f_i(\{\mathbf{z}_j \mid \mathbf{z}_j \in \text{pa}(\mathbf{z}_i)\}, \epsilon_i)$$

and a **mixing function**:

$$\mathbf{x}_i := g_i(\mathbf{z})$$



In the simplest form g is an (invertible) linear mixing function, $\mathbf{A} \in R^{k \times n}$:

$$\mathbf{x} = \mathbf{A}\,\mathbf{z}$$

The general case is underdetermined $\rightarrow$ The latent structure is unidentifiable!

Latent Causal Process Discovery - Mixing

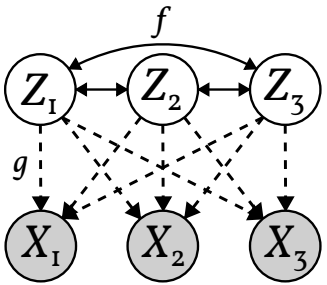
Task of CRL: Recover the latent factors $\mathbf{z} \in \mathbb{R}^n$ (and graph) from an observed signal $\mathbf{x} \in \mathbb{R}^k$.

Assume a **latent process**:

$$\mathbf{z}_i := f_i(\{\mathbf{z}_j \mid \mathbf{z}_j \in \text{pa}(\mathbf{z}_i)\}, \epsilon_i)$$

and a **mixing function**:

$$\mathbf{x}_i := g_i(\mathbf{z})$$



Additional assumptions needed:

- Independent & non-Gaussian latents ($\rightarrow$ ICA)
- Interventional data
- Additional structural assumptions (pure children, time series, …)

Squires, C., Seigal, A., Bhate, S.S. and Uhler, C., 2023, July. Linear causal disentanglement via interventions. In International conference on machine learning (pp. 32540-32560). PMLR.

Temporal Causal Representation Learning

Task of BSS: Recover the latent factors $\mathbf{z} \in \mathbb{R}^n$ (and graph) from an observed signal $\mathbf{x} \in \mathbb{R}^k$ in a timeseries.

Assume an underlying process:

Latent process: $\mathbf{z}_{it} := f_i(\text{pa}(\mathbf{z}_{it}), \epsilon_{it})$
where $\text{pa}(\mathbf{z}_{it}) \subseteq \mathbf{z}$

Mixing function: $\mathbf{x}_{it} := g_i(\mathbf{z}_t)$

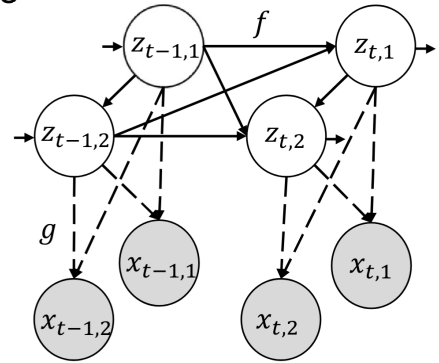
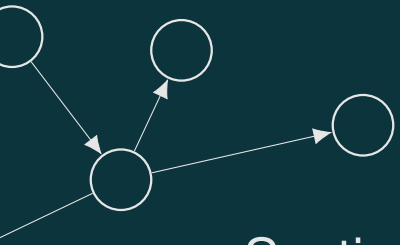
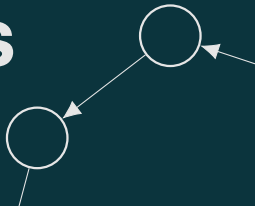


Figure adapted from: Li, Z., Shen, Y., Zheng, K., Cai, R., Song, X., Gong, M., Chen, G. and Zhang, K., On the Identification of Temporal Causal Representation with Instantaneous Dependence. In The Thirteenth International Conference on Learning Representations.



Section 3

Identifiability under Interventions



Linear Mixing with Interventions

Linear Latent Causal Process: $Z = A_k Z + \Omega_k^{1/2} \epsilon$
per context $k \in [1, \ldots, d]$

Linear Mixing Function: $X = GZ$

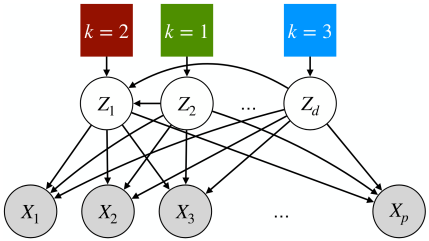
Assumptions:

- $G \in \mathbb{R}^{p \times d}$ is full column rank.
- There are d different contexts with perfect interventions $i_1, \ldots, i_d$.

Task: Recover G (and therefore Z).

Solving for Z gives: $Z = B_k^{-1} \epsilon$ with $B_k = \Omega_k^{-1/2}(I_d - A_k)$

Identifiable up to permutation, so we make B_k upper triangular.



Squires, C., Seigal, A., Bhate, S.S. and Uhler, C., 2023, July. Linear causal disentanglement via interventions. In International conference on machine learning (pp. 32540-32560). PMLR.

Linear Mixing with Interventions

Generating process:

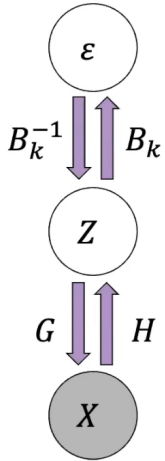
$Z := B_k^{-1}\epsilon$

$X := GZ$

Backward Process:

$Z = HX$ (with $H = G^{\dagger}$ and † is the pseudo-inverse.)

$\epsilon = B_k Z$



Squires, C., Seigal, A., Bhate, S.S. and Uhler, C., 2023, July. Linear causal disentanglement via interventions. In International conference on machine learning (pp. 32540-32560). PMLR.

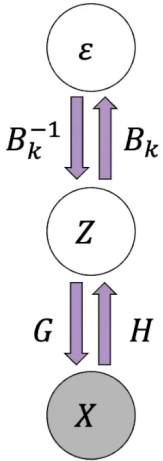
Figure: https://www.youtube.com/watch?v=ukhEEapKM-E

Linear Mixing with Interventions

Generating process:
$Z := B_k^{-1}\epsilon$
$X := GZ$

Backward Process:
$Z = HX$ (with $H = G^\dagger$ and † is the pseudo-inverse.)
$\epsilon = B_k Z$

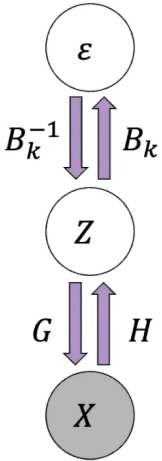
$\text{Cov}(\epsilon)^{-1} = I_d$
$\text{Cov}_k(Z)^{-1} = B_k^T B_k$
$\Theta_k = \text{Cov}_k(X)^\dagger = H^T B_k^T B_k H$



Squires, C., Seigal, A., Bhate, S.S. and Uhler, C., 2023, July. Linear causal disentanglement via interventions. In International conference on machine learning (pp. 32540-32560). PMLR.
Figure: https://www.youtube.com/watch?v=ukhEEapKM-E

Linear Mixing with Interventions

$\text{Cov}(\epsilon)^{-1} = I_d$
$\text{Cov}_k(Z)^{-1} = B_k^T B_k$
$\Theta_k = \text{Cov}_k(X)^\dagger = H^T B_k^T B_k H$

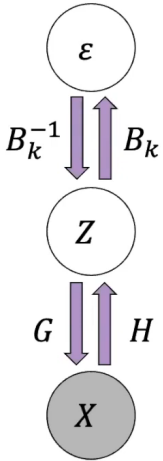


Squires, C., Seigal, A., Bhate, S.S. and Uhler, C., 2023, July. Linear causal disentanglement via interventions. In International conference on machine learning (pp. 32540-32560). PMLR.
Figure: https://www.youtube.com/watch?v=ukhEEapKM-E

Linear Mixing with Interventions

$\text{Cov}(\epsilon)^{-1} = I_d$
$\text{Cov}_k(Z)^{-1} = B_k^T B_k$
$\Theta_k = \text{Cov}_k(X)^\dagger = H^T B_k^T B_k H$

Input: $\Theta_0, \ldots, \Theta_K$
Task: Recover H and identify all $B_0, \ldots, B_K$.



Squires, C., Seigal, A., Bhate, S.S. and Uhler, C., 2023, July. Linear causal disentanglement via interventions. In International conference on machine learning (pp. 32540-32560). PMLR.
Figure: https://www.youtube.com/watch?v=ukhEEapKM-E

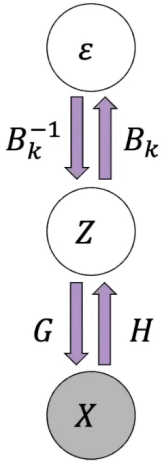
Linear Mixing with Interventions

$\text{Cov}(\epsilon)^{-1} = I_d$
$\text{Cov}_k(Z)^{-1} = B_k^T B_k$
$\Theta_k = \text{Cov}_k(X)^\dagger = H^T B_k^T B_k H$

Input: $\Theta_0, \ldots, \Theta_K$
Task: Recover H and identify all $B_0, \ldots, B_K$.

Squires et al., 2023

One intervention per latent node is sufficient (and in the worst case **necessary**) to recover $H = G^\dagger$ and $B_0, \ldots, B_K$.

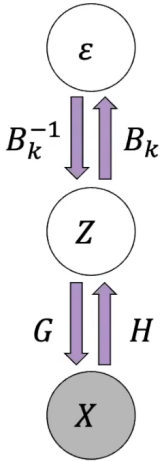


Squires, C., Seigal, A., Bhate, S.S. and Uhler, C., 2023, July. Linear causal disentanglement via interventions. In International conference on machine learning (pp. 32540-32560). PMLR.
Figure: https://www.youtube.com/watch?v=ukhEEapKM-E

Linear Mixing with Interventions

$\text{Cov}(\epsilon)^{-1} = I_d$
$\text{Cov}_k(Z)^{-1} = B_k^T B_k$
$\Theta_k = \text{Cov}_k(X)^\dagger = H^T B_k^T B_k H$

Input: $\Theta_0, \ldots, \Theta_K$
Task: Recover H and all $B_0, \ldots, B_K$.



Squires, C., Seigal, A., Bhate, S.S. and Uhler, C., 2023, July. Linear causal disentanglement via interventions. In International conference on machine learning (pp. 32540-32560). PMLR.
Figure: https://www.youtube.com/watch?v=ukhEEapKM-E

Linear Mixing with Interventions

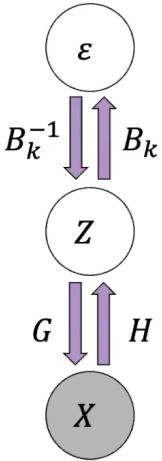
$\mathrm{Cov}(\epsilon)^{-1} = I_d$
$\mathrm{Cov}_k(Z)^{-1} = B_k^T B_k$
$\Theta_k = \mathrm{Cov}_k(X)^\dagger = H^T B_k^T B_k H$

Input: $\Theta_0, \ldots, \Theta_K$
Task: Recover H and all $B_0, \ldots, B_K$.

1) $\Theta_k - \Theta_0 = (H^T B_k^T \mathbf{e}_{i_k})^{\otimes 2} - (H^T B_0^T \mathbf{e}_{i_k})^{\otimes 2}$ is of rank 2, *except* for source nodes i_k!

2) Remove found node and repeat.

   In later iterations, it is necessary to keep track of the removed source nodes.



Squires, C., Seigal, A., Bhate, S.S. and Uhler, C., 2023, July. Linear causal disentanglement via interventions. In International conference on machine learning (pp. 32540-32560). PMLR.
Figure: https://www.youtube.com/watch?v=ukhEEapKM-E

Temporal Causal Representation Learning

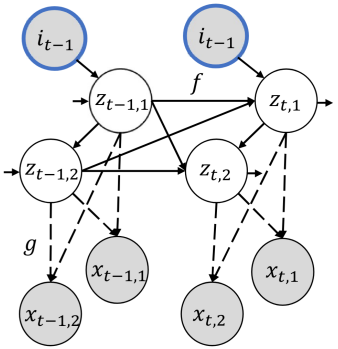
Task of CRL: Recover the latent factors $\mathbf{z} \in \mathbb{R}^n$ (and graph) from an observed signal $\mathbf{x} \in \mathbb{R}^k$ in a timeseries under interventions.

Assume an underlying process:

Latent process: $\mathbf{z}_{it} := f_i(\text{pa}(\mathbf{z}_{it}), \epsilon_{it})$
where $\text{pa}(\mathbf{z}_{it}) \subseteq \mathbf{z}$

Mixing function: $\mathbf{x}_{it} := g_i(\mathbf{z}_t)$

Interventions: $I^t = \{i_1^t, \dots\}$



iCRITIS [1]

Figure: Li, Z., Shen, Y., Zheng, K., Cai, R., Song, X., Gong, M., Chen, G. and Zhang, K., On the Identification of Temporal Causal Representation with Instantaneous Dependence. In The Thirteenth International Conference on Learning Representations.

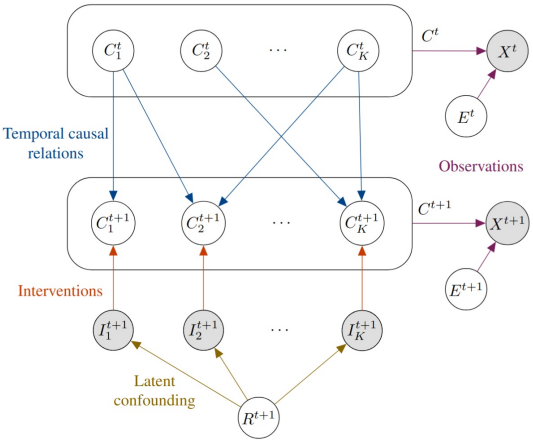
[1] Lippe, P., Magliacane, S., Löwe, S., Asano, Y.M., Cohen, T. and Gavves, E., Causal Representation Learning for Instantaneous and Temporal Effects in Interactive Systems. In The Eleventh International Conference on Learning Representations.

CITRIS: Causal Identifiability from Temporal Intervened Sequences

Temporal Intervened Sequences (TRIS):

Assumptions:

1. Same causal graph between all timesteps (*stationary graph*).
2. No instantaneous causal effects.
3. Agent takes actions R^{t+1} in the environment.



Lippe, P., Magliacane, S., Löwe, S., Asano, Y.M., Cohen, T. and Gavves, E., Causal Representation Learning for Instantaneous and Temporal Effects in Interactive Systems. In The Eleventh International Conference on Learning Representations.

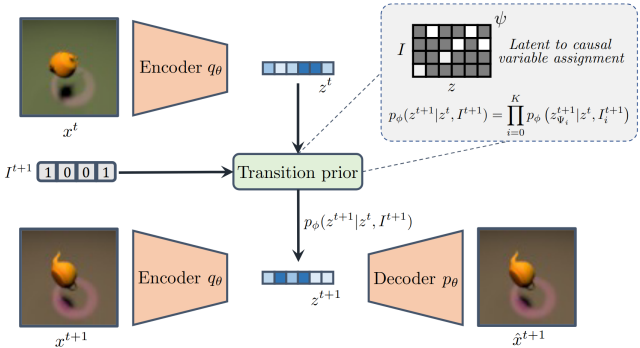
CITRIS: Causal Identifiability from Temporal Intervened Sequences

Loss Term:

1. Reconstruction loss.
2. Transition dynamics between time steps.

Assumes intervention targets are observed!

Assumes multi-valued variables. Requires mapping $\psi : I \to C$.



Lippe, P., Magliacane, S., Löwe, S., Asano, Y.M., Cohen, T. and Gavves, E., Causal Representation Learning for Instantaneous and Temporal Effects in Interactive Systems. In The Eleventh International Conference on Learning Representations.

BISCUIT - Causal Representation Learning from Binary Interactions

Sometimes, we might take actions in an environment, but do not know the exact set of affected variables.

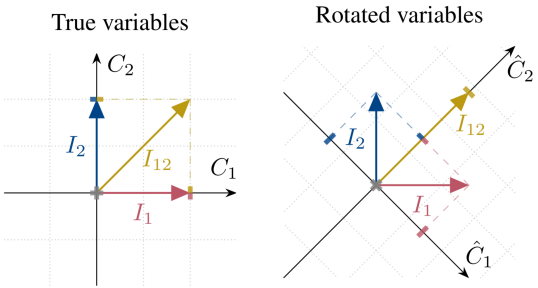
‘Can we get rid of observed intervention target assumption?’

Lippe, P., Magliacane, S., Löwe, S., Asano, Y.M., Cohen, T. and Gavves, E., 2023, July. Biscuit: Causal representation learning from binary interactions. In Uncertainty in Artificial Intelligence (pp. 1263-1273). PMLR.

BISCUIT - Causal Representation Learning from Binary Interactions

Binary interaction information gives identifiability!

- Arrows indicate interactions.
- Ticks show resulting variable means.



Lippe, P., Magliacane, S., Löwe, S., Asano, Y.M., Cohen, T. and Gavves, E., 2023, July. Biscuit: Causal representation learning from binary interactions. In Uncertainty in Artificial Intelligence (pp. 1263-1273). PMLR.

BISCUIT - Causal Representation Learning from Binary Interactions

Binary interaction information gives identifiability!

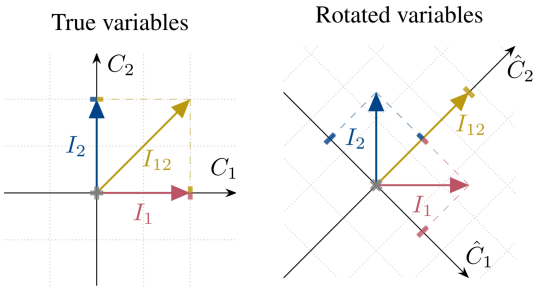
Arrows indicate interactions.

Ticks show resulting variable means.

If variables are entangled:

1) Actions will alter multiple latent variables.

2) Outcome of the joint intervention no longer matches the outcome of the individual interventions.



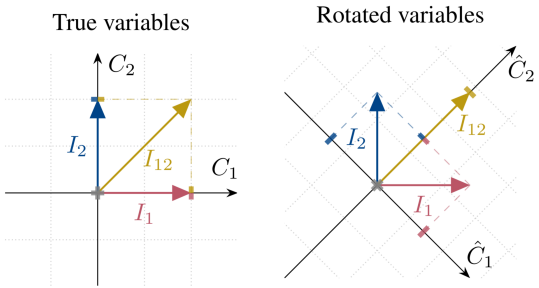
Lippe, P., Magliacane, S., Löwe, S., Asano, Y.M., Cohen, T. and Gavves, E., 2023, July. Biscuit: Causal representation learning from binary interactions. In Uncertainty in Artificial Intelligence (pp. 1263-1273). PMLR.

BISCUIT - Causal Representation Learning from Binary Interactions

Binary interaction information gives identifiability!

Arrows indicate interactions.

Ticks show resulting variable means.

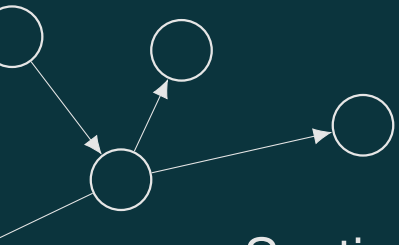


Required Assumptions:

1) Any two variables are not always intervened jointly at the same time.
2) Any two variables are sometimes intervened jointly.
3) Mechanisms vary over time or through interaction.

Under optimal intervention design, requires $\log_2 K + 2$ actions.

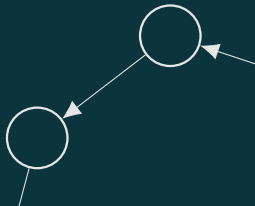
Lippe, P., Magliacane, S., Löwe, S., Asano, Y.M., Cohen, T. and Gavves, E., 2023, July. Biscuit: Causal representation learning from binary interactions. In Uncertainty in Artificial Intelligence (pp. 1263-1273). PMLR.



Section

4

Identifiability under Sufficient Variation



Temporal Causal Representation Learning

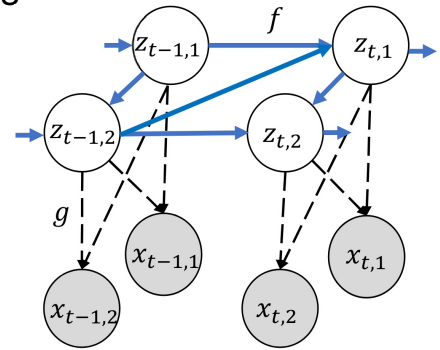
Task of CRL: Recover the latent factors $\mathbf{z} \in \mathbb{R}^n$ from an observed signal $\mathbf{x} \in \mathbb{R}^k$.

Assume an underlying process:

Latent process: $\mathbf{z}_{it} := f_i(\text{pa}(\mathbf{z}_{it}), \epsilon_{it})$
where $\text{pa}(\mathbf{z}_{it}) \subseteq \mathbf{z}$

Mixing function: $\mathbf{x}_{it} := g_i(\mathbf{z}_t)$

Sparse Latent Process: *(blue edges)*



IDOL [1]

[1] Li, Z., Shen, Y., Zheng, K., Cai, R., Song, X., Gong, M., Chen, G. and Zhang, K., On the Identification of Temporal Causal Representation with Instantaneous Dependence. In The Thirteenth International Conference on Learning Representations.

Temporal Causal Representation Learning

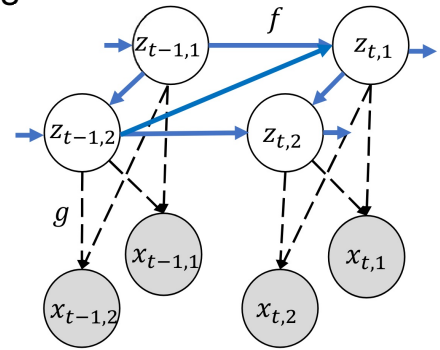
Task of CRL: Recover the latent factors $\mathbf{z} \in \mathbb{R}^n$ from an observed signal $\mathbf{x} \in \mathbb{R}^k$.

Assume an underlying process:

Latent process: $\mathbf{z}_{it} := f_i(\text{pa}(\mathbf{z}_{it}), \epsilon_{it})$
where $\text{pa}(\mathbf{z}_{it}) \subseteq \mathbf{z}$

Mixing function: $\mathbf{x}_{it} := g_i(\mathbf{z}_t)$

Sparse Latent Process: *(blue edges)*



IDOL [1]

Idea: Restrict data in a way that makes the structure become identifiable.

1. Enforce independencies of the shown graph.
2. Require further variability to make the system have a unique solution.

[1] Li, Z., Shen, Y., Zheng, K., Cai, R., Song, X., Gong, M., Chen, G. and Zhang, K., On the Identification of Temporal Causal Representation with Instantaneous Dependence. In The Thirteenth International Conference on Learning Representations.

Sufficient Variability

$\mathbf{c}_t \triangleq \{\mathbf{z}_{t-1}, \mathbf{z}_t\}$ are the variables of two consecutive timesteps.

Sufficient Variability Constraint (for $m \in [1, \dots, n]$):

$$w(m) = \left(\frac{\partial^3 \log p(\mathbf{c}_t|\mathbf{z}_{t-2})}{\partial c_{t,1}^2 \partial z_{t-2,m}}, \dots, \frac{\partial^3 \log p(\mathbf{c}_t|\mathbf{z}_{t-2})}{\partial c_{t,2n}^2 \partial z_{t-2,m}} \right) \oplus \left(\frac{\partial^2 \log p(\mathbf{c}_t|\mathbf{z}_{t-2})}{\partial c_{t,1} \partial z_{t-2,m}}, \dots, \frac{\partial^2 \log p(\mathbf{c}_t|\mathbf{z}_{t-2})}{\partial c_{t,2n} \partial z_{t-2,m}} \right) \oplus \left(\frac{\partial^3 \log p(\mathbf{c}_t|\mathbf{z}_{t-2})}{\partial c_{t,i} \partial c_{t,j} \partial z_{t-2,m}} \right)_{(i,j) \in \mathcal{E}(\mathcal{M}_{\mathbf{c}_t})}$$

$(i,j) \in \mathcal{E}(\mathcal{M}_{\mathbf{c}_t})$ indicates all direct edges (i,j) in $\mathcal{M}_{\mathbf{c}_t}$.

For $m \in [1, \dots, n]$, there exist $4n + |M_{\mathbf{c}_t}|$ different values of $\mathbf{z}_{t-2,m}$, such that the $4n + |M_{\mathbf{c}_t}|$ values of vector functions $w(m)$ are linearly independent.

Li, Z., Shen, Y., Zheng, K., Cai, R., Song, X., Gong, M., Chen, G. and Zhang, K., On the Identification of Temporal Causal Representation with Instantaneous Dependence. In The Thirteenth International Conference on Learning Representations.

Sufficient Variability - High-Level Idea

If sufficient variability holds, then for any two different entries $\hat{c}_{t,k}, \hat{c}_{t,l}$ of $\hat{\mathbf{c}}_t \in \mathbb{R}^{2n}$ that are **not adjacent** in the Markov network $\mathcal{M}_{\hat{\mathbf{c}}_t}$ over estimated $\hat{\mathbf{c}}_t$,

(i) Each ground-truth latent variable $c_{t,i}$ of $\mathbf{c}_t \in \mathbb{R}^{2n}$ is a function of at most one of $\hat{c}_k$ and $\hat{c}_l$,

(ii) For each pair of ground-truth latent variables $c_{t,i}$ and $c_{t,j}$ of $\mathbf{c}_t \in \mathbb{R}^{2n}$ that are **adjacent** in $\mathcal{M}_{\mathbf{c}_t}$ over $\mathbf{c}_t$, they can not be a function of $\hat{c}_{t,k}$ and $\hat{c}_{t,l}$ respectively.

The first two groups constrain structure of the network (Lin et al., 1997):

$$c_{t,i} \perp c_{t,j} \mid \mathbf{c}_t \setminus \{c_{t,i}, c_{t,j}\} \text{ implies } \frac{\partial^2 \log p(\mathbf{c}_t)}{\partial c_{t,i} \partial c_{t,j}} = 0$$

Li, Z., Shen, Y., Zheng, K., Cai, R., Song, X., Gong, M., Chen, G. and Zhang, K., On the Identification of Temporal Causal Representation with Instantaneous Dependence. In The Thirteenth International Conference on Learning Representations.

Lin, J., 1997. Factorizing multivariate function classes. Advances in neural information processing systems, 10.

Sufficient Variability - High-Level Idea

If sufficient variability holds, then for any two different entries $\hat{c}_{t,k}, \hat{c}_{t,l}$ of $\hat{\mathbf{c}}_t \in \mathbb{R}^{2n}$ that are **not adjacent** in the Markov network $\mathcal{M}_{\hat{\mathbf{c}}_t}$ over estimated $\hat{\mathbf{c}}_t$,

(i) Each ground-truth latent variable $c_{t,i}$ of $\mathbf{c}_t \in \mathbb{R}^{2n}$ is a function of at most one of $\hat{c}_k$ and $\hat{c}_l$,

(ii) For each pair of ground-truth latent variables $c_{t,i}$ and $c_{t,j}$ of $\mathbf{c}_t \in \mathbb{R}^{2n}$ that are **adjacent** in $\mathcal{M}_{\mathbf{c}_t}$ over $\mathbf{c}_t$, they can not be a function of $\hat{c}_{t,k}$ and $\hat{c}_{t,l}$ respectively.

Particularly, the constraint enforces a unique solution, whenever

$$\frac{\partial c_{t,i}}{\partial \hat{c}_{t,k}} \cdot \frac{\partial c_{t,i}}{\partial \hat{c}_{t,l}} = 0, \quad \frac{\partial c_{t,i}}{\partial \hat{c}_{t,k}} \cdot \frac{\partial c_{t,j}}{\partial \hat{c}_{t,l}} = 0, \quad \frac{\partial^2 c_{t,i}}{\partial c_{t,k} \partial c_{t,l}} = 0$$

The first two terms correspond to statements (i) and (ii).

Li, Z., Shen, Y., Zheng, K., Cai, R., Song, X., Gong, M., Chen, G. and Zhang, K., On the Identification of Temporal Causal Representation with Instantaneous Dependence. In The Thirteenth International Conference on Learning Representations.

Next: Large Scale Models

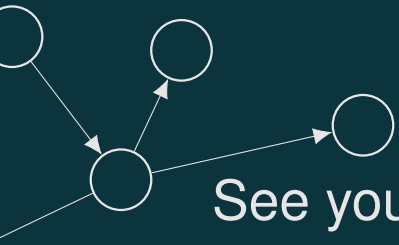
We considered training small, task-specific models.

Training was limited by the availability of data

$\rightarrow$ Does large-scale data contain enough variation to disentangle factors?

$\rightarrow$ Can foundation models (e.g., LLMs) 'evolve' a causal understanding?

Next lecture…



See you next week!