Introduction to Causal Inference
&
Identifiability in latent variable models

Overview

Causal inference:

- Causal graphical models
- Interventions (the "do" operator)
- Example: Study of Kidney Stone Treatments
- Backdoor criterion
- The ladder of causation
- Counterfactuals

Identifiability in latent variable models:

- The problem of identifiability in generative models
- Disentanglement
- Independent component analysis (ICA)
- Darmois-Skitovich theorem
- Leveraging temporal dependencies (AMUSE algorithm)
- Nonlinear ICA and its connection to disentanglement

Causal Inference

Causal graphical models (CGM)

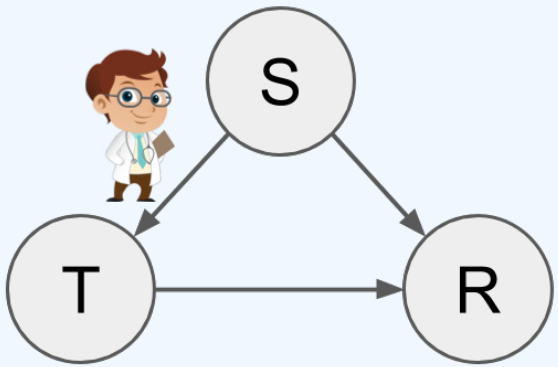
- A causal graphical model (CGM) is a pair $(p, \mathcal{G})$ s.t.
- $\mathcal{G}$ is a **directed acyclic graph** (DAG)
- $p \in \mathcal{L}(\mathcal{G})$, i.e. p factorizes according to $\mathcal{G}$.
- $\mathcal{G}$ describes **causal relationships** between variables, i.e., how the system reacts to **interventions**.

Causal graphical models (CGM)

- A causal graphical model (CGM) is a pair $(p, \mathcal{G})$ s.t.
- $\mathcal{G}$ is a **directed acyclic graph** (DAG)
- $p \in \mathcal{L}(\mathcal{G})$, i.e. p factorizes according to $\mathcal{G}$.
- $\mathcal{G}$ describes **causal relationships** between variables, i.e., how the system reacts to **interventions**.

Example: Kidney stone treatment

T = Treatment $\in \{A, B\}$
S = Stone size $\in \{\text{small}, \text{large}\}$
R = Patient recovered $\in \{0, 1\}$



$$p(S, T, R) = p(S)p(T \mid S)p(R \mid S, T)$$

The "do" operator models the effect of interventions

- Recall $p(x) = \prod_i p(x_i \mid x_{\pi_i^{\mathcal{G}}})$

- Throughout, we will assume **perfect deterministic** interventions.

Definition (The "do" operator)

Given a causal graphical model $(p, \mathcal{G})$,

$$p(x \mid do(x_k := x_k')) := \delta(x_k, x_k') \prod_{i \neq k} p(x_i \mid x_{\pi_i^{\mathcal{G}}}),$$

where $\delta(x_k, x_k') = 1$ when $x_k = x_k'$ and 0 otherwise. Here, x_k is *targeted* by the *intervention*.

- Thus, $p(x \mid do(x_k := x_k'))$ is a "new" joint distribution over X_V.

The "do" operator models the effect of interventions

- Recall $p(x) = \prod_i p(x_i \mid x_{\pi_i^{\mathcal{G}}})$
- Throughout, we will assume **perfect deterministic** interventions.

Definition (The "do" operator)

Given a causal graphical model $(p, \mathcal{G})$,

$$p(x \mid do(x_k := x_k')) := \delta(x_k, x_k') \prod_{i \neq k} p(x_i \mid x_{\pi_i^{\mathcal{G}}}),$$

where $\delta(x_k, x_k') = 1$ when $x_k = x_k'$ and 0 otherwise. Here, x_k is *targeted* by the *intervention*.

- Thus, $p(x \mid do(x_k := x_k'))$ is a "new" joint distribution over X_V.
- Can compute marginals, e.g. $p(x_i | do(x_k := x_k')) = \sum_{x_{V \setminus \{i\}}} p(x | do(x_k := x_k'))$

The "do" operator models the effect of interventions

- Recall $p(x) = \prod_i p(x_i \mid x_{\pi_i^{\mathcal{G}}})$
- Throughout, we will assume **perfect deterministic** interventions.

Definition (The "do" operator)

Given a causal graphical model $(p, \mathcal{G})$,

$$p(x \mid do(x_k := x_k')) := \delta(x_k, x_k') \prod_{i \neq k} p(x_i \mid x_{\pi_i^{\mathcal{G}}}),$$

where $\delta(x_k, x_k') = 1$ when $x_k = x_k'$ and 0 otherwise. Here, x_k is *targeted* by the *intervention*.

- Thus, $p(x \mid do(x_k := x_k'))$ is a "new" joint distribution over X_V.
- Can compute marginals, e.g. $p(x_i | do(x_k := x_k')) = \sum_{x_{V \setminus \{i\}}} p(x | do(x_k := x_k'))$
- ... and conditionals, e.g. $p(x_i | x_j, do(x_k := x_k')) = \frac{p(x_i, x_j | do(x_k := x_k'))}{p(x_j | do(x_k := x_k'))}$

The "do" operator models the effect of interventions

- Recall $p(x) = \prod_i p(x_i \mid x_{\pi_i^{\mathcal{G}}})$
- Throughout, we will assume **perfect deterministic** interventions.

Definition (The "do" operator)

Given a causal graphical model $(p, \mathcal{G})$,

$$p(x \mid do(x_k := x_k')) := \delta(x_k, x_k') \prod_{i \neq k} p(x_i \mid x_{\pi_i^{\mathcal{G}}}),$$

where $\delta(x_k, x_k') = 1$ when $x_k = x_k'$ and 0 otherwise. Here, x_k is *targeted* by the *intervention*.

- Thus, $p(x \mid do(x_k := x_k'))$ is a "new" joint distribution over X_V.
- Can compute marginals, e.g. $p(x_i | do(x_k := x_k')) = \sum_{x_{V \setminus \{i\}}} p(x | do(x_k := x_k'))$
- ... and conditionals, e.g. $p(x_i | x_j, do(x_k := x_k')) = \frac{p(x_i, x_j | do(x_k := x_k'))}{p(x_j | do(x_k := x_k'))}$
- **Truncated factorization:**
  $p(x_{V \setminus \{k\}} \mid do(x_k := x_k')) = \sum_{x_k} \delta(x_k, x_k') \prod_{i \neq k} p(x_i \mid x_{\pi_i^{\mathcal{G}}}) = \prod_{i \neq k} p(x_i \mid x_{\pi_i^{\mathcal{G}}})$.

Conditioning is not the same as doing

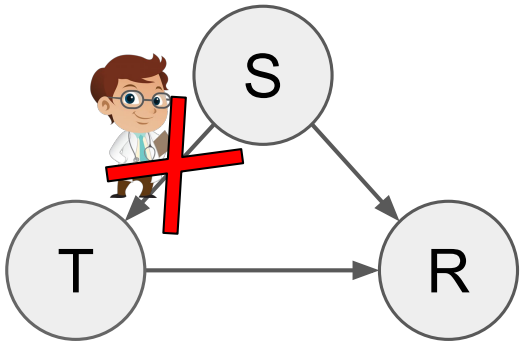
Conditioning is not the same as doing

Consider the simple CGM $X \to Y$

$$p(X|do(Y := Y')) = \cancel{p(Y \mid X)}p(X) \tag{1}$$
$$= p(X) \tag{2}$$
$$\neq p(X \mid Y = Y') \tag{3}$$

The "do" operator

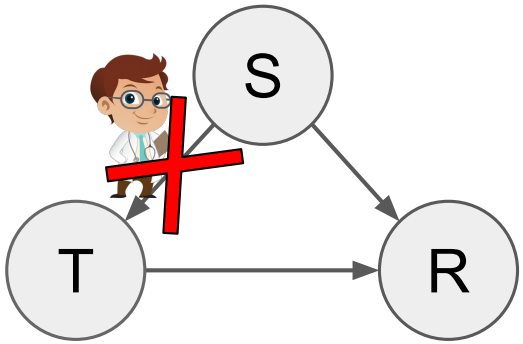
- Back to our example



$$P(S, R \mid do(T = T')) = P(S) \underbrace{\cancel{P(T|S)}}_{\text{The decision of taking treatment } T \text{ does not depend on } S \text{ anymore}} P(R|S, T')$$

The "do" operator

- Back to our example



$$P(S,R \mid do(T=T')) = P(S)\underbrace{\cancel{P(T|S)}}P(R|S,T')$$

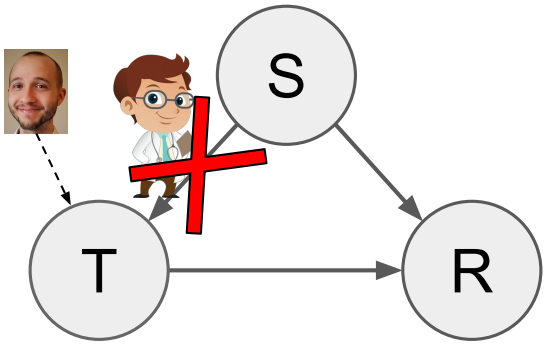
The decision of taking treatment T does not depend on S anymore

- Notice $p(\cdot \mid do(x'_k)) \in \mathcal{L}(\mathcal{G}')$, where $\mathcal{G}'$ is the **mutilated graph**, i.e.

$$\mathcal{G}' = (V, E') \quad E' = \{(i,j) \in E \mid j \neq k\}$$

The "do" operator

- Back to our example



$$P(S,R \mid do(T = T')) = P(S)\underbrace{\cancel{P(T|S)}}P(R|S,T')$$

The decision of taking treatment T does not depend on S anymore

- Notice $p(\cdot \mid do(x_k := x'_k)) \in \mathcal{L}(\mathcal{G}')$, where $\mathcal{G}'$ is the **mutilated graph**, i.e.

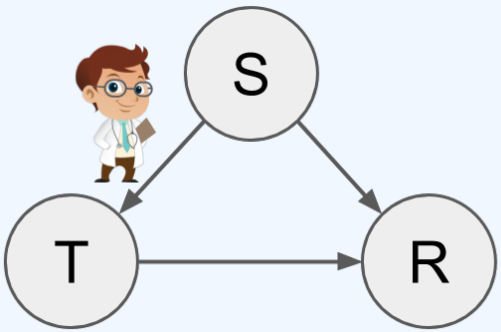
$$\mathcal{G}' = (V, E') \quad E' = \{(i,j) \in E \mid j \neq k\}$$

Different types of interventions

Intervening on the treatment T

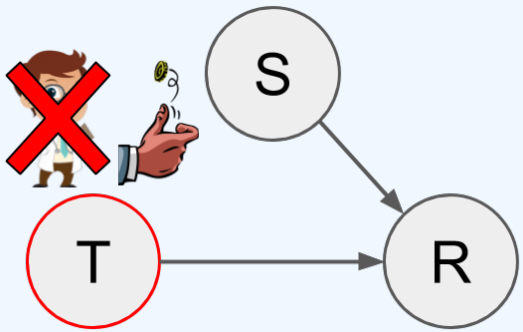
T = Treatment $\in \{A, B\}$
S = Stone size $\in \{\text{small}, \text{large}\}$
R = Patient recovered $\in \{0, 1\}$

Observations



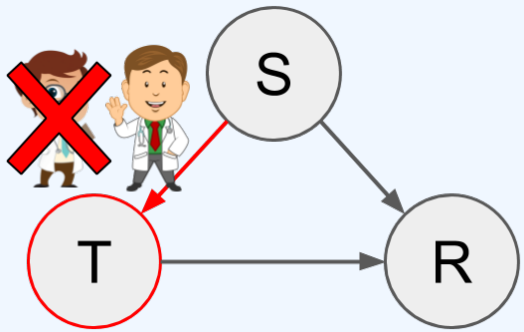
$p(S)p(T \mid S)p(R \mid S, T)$

Perfect intervention



$p(S)\tilde{p}(T)p(R \mid S, T)$

Imperfect intervention



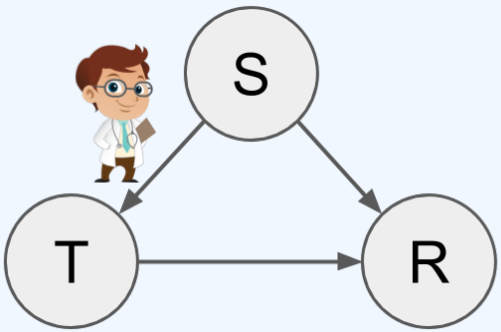
$p(S)\tilde{p}(T \mid S)p(R \mid S, T)$

Different types of interventions

Intervening on the treatment T

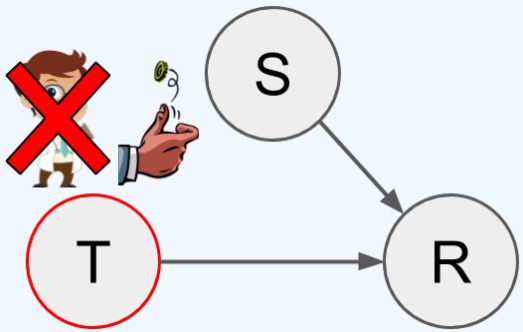
T = Treatment $\in \{A, B\}$
S = Stone size $\in \{\text{small}, \text{large}\}$
R = Patient recovered $\in \{0, 1\}$

Observations



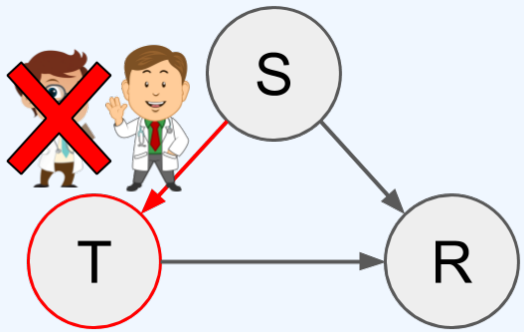
$p(S)p(T \mid S)p(R \mid S, T)$

Perfect intervention



$p(S)\tilde{p}(T)p(R \mid S, T)$

Imperfect intervention

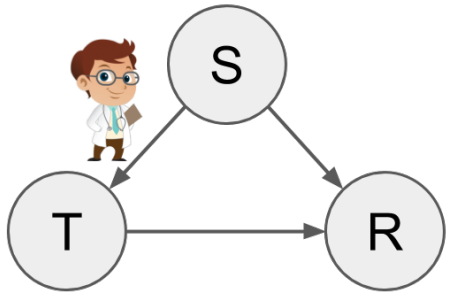


$p(S)\tilde{p}(T \mid S)p(R \mid S, T)$

Definition presented previously is a perfect intervention with $\tilde{p}(T) := \delta(T, T')$.
It is sometimes called a **perfect deterministic intervention**.

Why should I care!?! (Kidney Stone Treatment)

T = Treatment $\in \{A, B\}$
S = Stone size $\in \{\text{small}, \text{large}\}$
R = Patient recovered $\in \{0, 1\}$



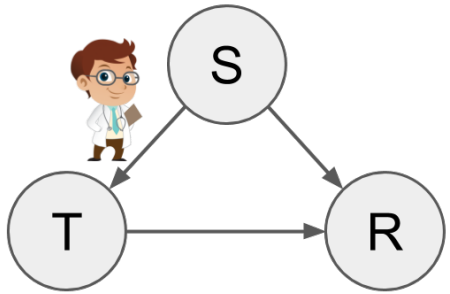
$$p(S)p(T \mid S)p(R \mid S, T)$$

	Overall	Patients with small stones	Patients with large stones
Treatment a: Open surgery	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment b: Percutaneous nephrolithotomy	**83%** (289/350)	87% (234/270)	69% (55/80)

(Example taken from *Element of Causal Inference* by Peters et al. p111)

Why should I care!?! (Kidney Stone Treatment)

T = Treatment $\in \{A, B\}$
S = Stone size $\in \{\text{small}, \text{large}\}$
R = Patient recovered $\in \{0, 1\}$



$$p(S)p(T \mid S)p(R \mid S, T)$$

	Overall	Patients with small stones	Patients with large stones
Treatment a: Open surgery	78% (273/350)	**93%** (81/87)	**73%** (192/263)
Treatment b: Percutaneous nephrolithotomy	**83%** (289/350)	87% (234/270)	69% (55/80)

(Example taken from *Element of Causal Inference* by Peters et al. p111)

Known as **Simpson's Paradox**

Why should I care!?! (Kidney Stone Treatment)

Pay attention to these two questions...

Why should I care!?! (Kidney Stone Treatment)

Pay attention to these two questions...

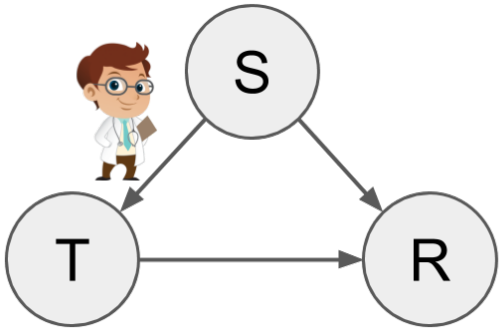
1- What is your chance of recovery knowing that the doctor gave you treatment A?

2- What is your chance of recovery if you decide to take treatment A?

(In both cases, assume you don't know the size of your stone)

Why should I care!?! (Kidney Stone Treatment)

T = Treatment $\in \{A, B\}$
Z = Stone size $\in \{\text{small}, \text{large}\}$
R = Patient recovered $\in \{0, 1\}$

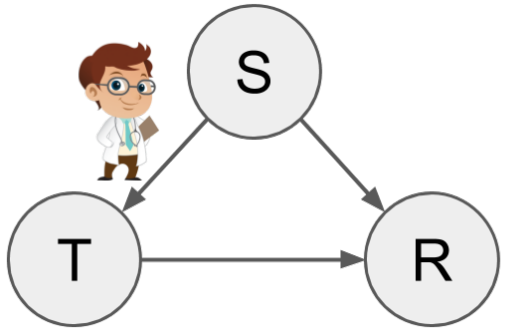


What is your chance of recovery knowing that the doctor gave you treatment A?

- Compute $P(R = 1 \mid T = A)$! (we know how to do that :D)
- Knowing that your doctor gave you treatment A tells you that you probably have a large kidney stone ... $P(S = \mathit{large} | T = A) = 0.75$
- ... which reduces your chance of recovery
  $P(R = 1 | T = A, S = \mathit{large}) = 0.73 < 0.93 = P(R = 1 | T = A, S = \mathit{small})$

Why should I care!?! (Kidney Stone Treatment)

T = Treatment $\in \{A, B\}$
Z = Stone size $\in \{\text{small}, \text{large}\}$
R = Patient recovered $\in \{0, 1\}$



What is your chance of recovery knowing that the doctor gave you treatment A?

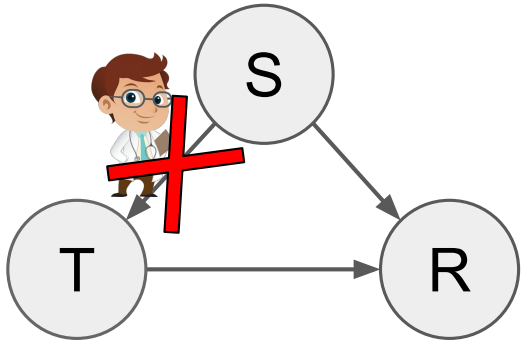
- Compute $P(R = 1 \mid T = A)$! (we know how to do that :D)
- Knowing that your doctor gave you treatment A tells you that you probably have a large kidney stone ... $P(S = large|T = A) = 0.75$
- ... which reduces your chance of recovery
  $P(R = 1|T = A, S = large) = 0.73 < 0.93 = P(R = 1|T = A, S = small)$

What is your chance of recovery if you decide to take treatment A?

- $P(R = 1 \mid do(T = A))$
- Your really don't know anything about your kidney stone

Why should I care!?! (Kidney Stone Treatment)

T = Treatment $\in \{A, B\}$
S = Stone size $\in \{\text{small}, \text{large}\}$
R = Patient recovered $\in \{0, 1\}$



$$P(S, R \mid do(T)) = P(S)\underbrace{\cancel{P(T|S)}}P(R|S,T)$$

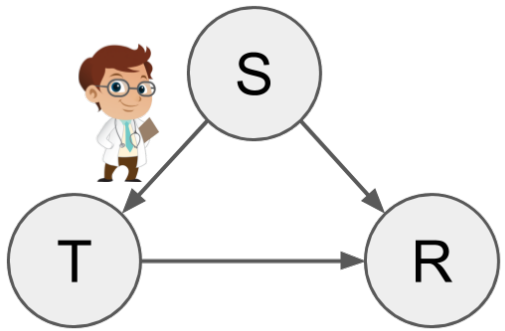
The decision of taking treatment T does not depend on S anymore

Then simply marginalize as usual:

$$P(R = 1|do(T = A)) = \sum_S P(R = 1, S|do(T = A))$$
$$= \sum_S P(R = 1|S, T = A)P(S) = 0,832$$

Why should I care!?! (Kidney Stone Treatment)

T = Treatment $\in \{A, B\}$
S = Stone size $\in \{\text{small}, \text{large}\}$
R = Patient recovered $\in \{0, 1\}$



What is your chance of recovery knowing that the doctor gave you treatment A?

$P(R = 1|T = A) = 0,78$ $\quad$ $P(R = 1|T = B) =$ **0,83**

What is your chance of recovery if you decide to take treatment A?

$P(R = 1|do(T = A)) =$ **0,832** $\quad$ $P(R = 1|do(T = B)) = 0,782$

Why should I care!?! (Kidney Stone Treatment)

- Again, conditioning is not the same as doing!

$$P(R=1|do(T=A)) = \sum_S P(R=1|S,T=A)P(S)$$
$$P(R=1|T=A) = \sum_S P(R=1|S,T=A)P(S \mid T=A)$$

Why should I care!?! (Kidney Stone Treatment)

- What just happened? We showed

$$\underbrace{P(R=1|do(T=A))}_{\text{Never observed data from } p(T,S,R \mid do(T=A))} = \underbrace{\sum_S P(R=1|S,T=A)P(S)}_{\text{...Yet I can estimate the query, since there is no "do" here :D}}$$

Why should I care!?! (Kidney Stone Treatment)

- What just happened? We showed

$$\underbrace{P(R=1|do(T=A))}_{\text{Never observed data from } p(T,S,R \mid do(T=A))} = \underbrace{\sum_S P(R=1|S,T=A)P(S)}_{\text{...Yet I can estimate the query, since there is no "do" here :D}}$$

- Formally, this means $p(R=1 \mid do(T=A))$ is **identifiable from** $p(R,T,S)$ and $\mathcal{G}$ (our computations *critically* relied on the causal graph).

Why should I care!?! (Kidney Stone Treatment)

- What just happened? We showed

$$\underbrace{P(R=1|do(T=A))}_{\text{Never observed data from } p(T,S,R \mid do(T=A))} = \underbrace{\sum_S P(R=1|S,T=A)P(S)}_{\text{...Yet I can estimate the query, since there is no "do" here :D}}$$

- Formally, this means $p(R=1 \mid do(T=A))$ is **identifiable from** $p(R,T,S)$ and $\mathcal{G}$ (our computations *critically* relied on the causal graph).

- Turns out what we just did is an instance of the **backdoor criterion**...

Backdoor criterion

Theorem (Backdoor criterion)

$p(x_i \mid do(x_k)) = \sum_{x_S} p(x_i \mid x_k, x_S) p(x_S)$ *if*

1. S *contains no descendants of* x_k*, and*
2. S *blocks all paths from* x_i *to* x_k *entering* x_k *from "the backdoor", i.e. such that* $x_k \leftarrow \ldots x_i$

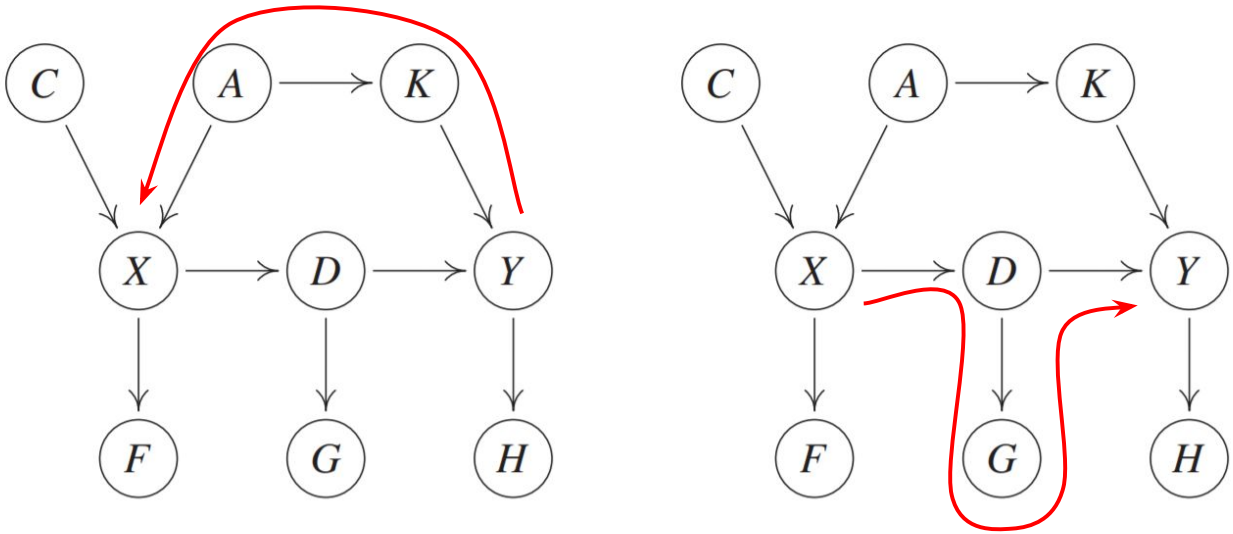
Backdoor criterion

Theorem (Backdoor criterion)

$p(x_i \mid do(x_k)) = \sum_{x_S} p(x_i \mid x_k, x_S) p(x_S)$ *if*

1. S *contains no descendants of* x_k*, and*
2. S *blocks all paths from* x_i *to* x_k *entering* x_k *from "the backdoor", i.e. such that* $x_k \leftarrow \ldots x_i$

Say we want to compute $p(y|do(x))$:



Left path: Only backdoor path. Blocked by $S = \{K\}$. **Right path:** Why we cannot include a descendant of X in S.

Backdoor criterion

Can all identifiable queries $p(x_i \mid do(x_k))$ be expressed with the backdoor criterion?

Backdoor criterion

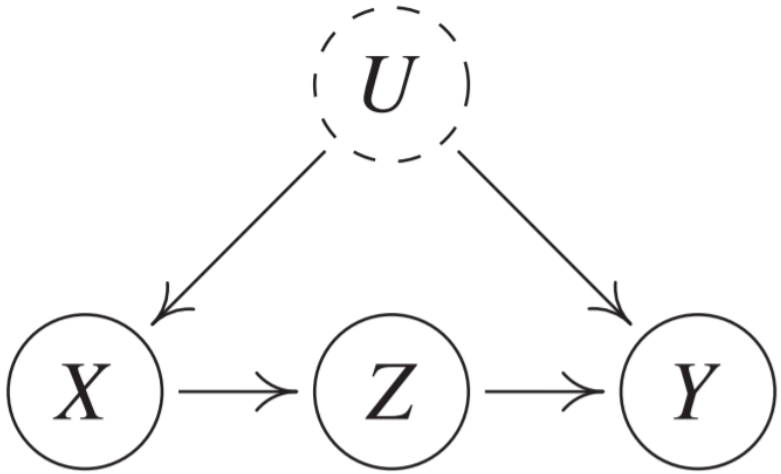
Can all identifiable queries $p(x_i \mid do(x_k))$ be expressed with the backdoor criterion?

Answer: No!

Backdoor criterion

Can all identifiable queries $p(x_i \mid do(x_k))$ be expressed with the backdoor criterion?

Answer: No!



- Since U is unobserved, we cannot apply the backdoor criterion...
- Turns out we can nevertheless identify $p(y|do(x))$ from $p(X, Z, Y)$ using the **front-door criterion**. Look it up!

Do-calculus

- Do-calculus is a set of **three rules** that can be applied to transform an interventional query (including a "do") into an observational expression (without any "do").
- Not enough time to present them...
- All identifiable queries can be found by a subsequent application of these rules, i.e. the rules are **complete**.

The ladder of causation

You now know about the first two steps of Pearl's "ladder of causation".

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in Y?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing Intervening	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

Fig. 1. The Causal Hierarchy. Questions at level i can only be answered if information from level i or higher is available.

Taken from "The Seven Tools of Causal Inference with Reflections on Machine Learning" by Judea Pearl

Counterfactual

You need **structural causal models (SCM)**. Let $\mathcal{G}$ be a DAG:

$$X_1 := f_1(X_{\pi_1^{\mathcal{G}}}, N_1) \tag{4}$$

$$X_2 := f_2(X_{\pi_2^{\mathcal{G}}}, N_2) \tag{5}$$

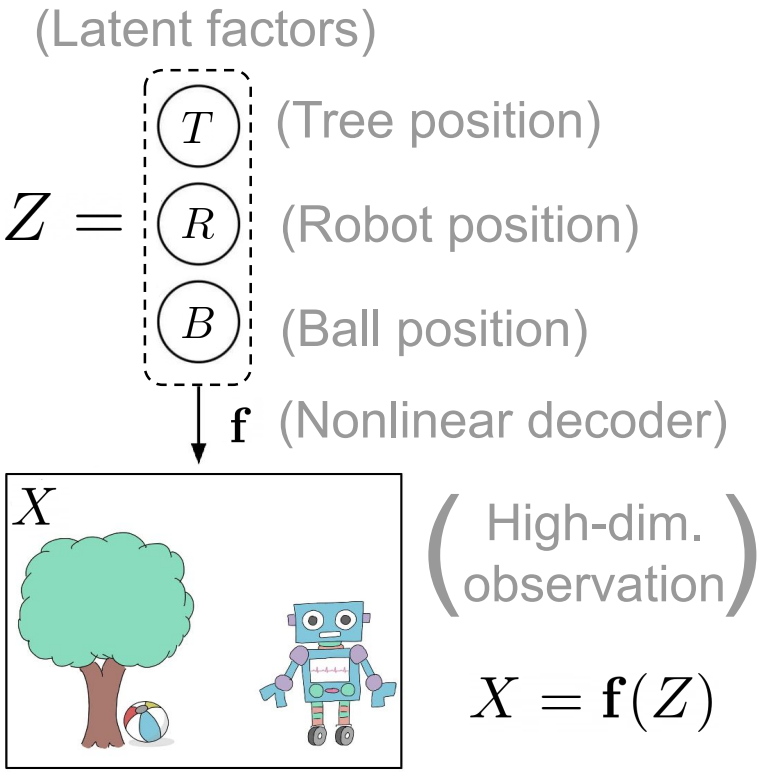
$$\dots \tag{6}$$

$$X_d := f_d(X_{\pi_d^{\mathcal{G}}}, N_d) \tag{7}$$

- This induces an **observational** distribution
- Can define **interventions** as well
- Can define **counterfactual** statements (not possible with a causal graphical model). See Section 6.4 in ECI.

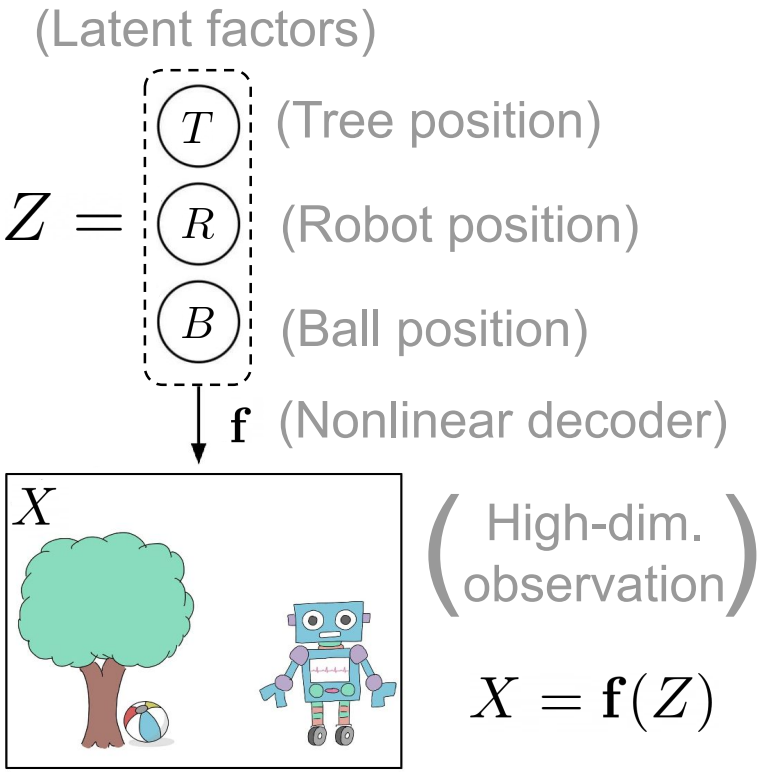
Identifiability in latent variable models

Disentanglement



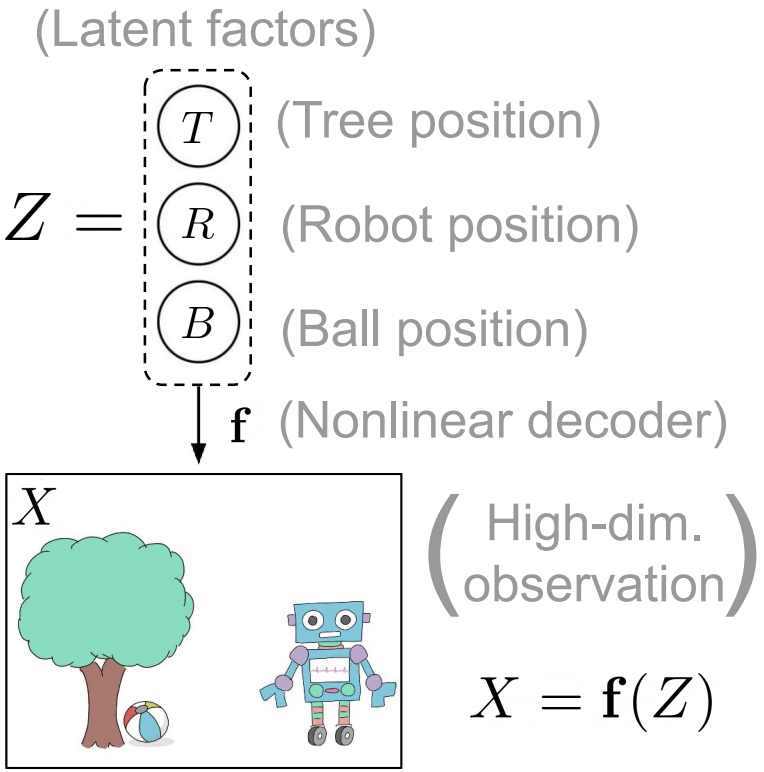
- Disentanglement is about recovering *natural factors of variations* from $p(X)$.

Disentanglement



- Disentanglement is about recovering *natural factors of variations* from $p(X)$.
- But can't we just learn a latent variable model using EM or a variational autoencoder (VAE)?

Disentanglement



- Disentanglement is about recovering *natural factors of variations* from $p(X)$.
- But can't we just learn a latent variable model using EM or a variational autoencoder (VAE)?
- Typically not as simple... One has to keep in mind the problem of *identifiability*.

The general problem of identifiability for generative models

Consider the following simple generative model:

$$Z \sim \mathbb{P}_Z, \; X := \mathbf{f}(Z) \implies \mathbb{P}_X$$

Consider this other model:

$$\hat{Z} := UZ, \; \hat{X} := \underbrace{\mathbf{f}(U^{-1}}_{\hat{\mathbf{f}}} \hat{Z}) \implies \mathbb{P}_{\hat{X}}$$

The general problem of identifiability for generative models

Consider the following simple generative model:

$$Z \sim \mathbb{P}_Z, \quad X := \mathbf{f}(Z) \implies \mathbb{P}_X$$

Consider this other model:

$$\hat{Z} := UZ, \quad \hat{X} := \underbrace{\mathbf{f}(U^{-1}}_{\hat{\mathbf{f}}}\,\hat{Z}) \implies \mathbb{P}_{\hat{X}}$$

$\mathbb{P}_X = \mathbb{P}_{\hat{X}}$

Both models represent the same distribution over X...

... but their representations can be drastically different

The general problem of identifiability for generative models

Consider the following simple generative model:

$$Z \sim \mathbb{P}_Z, \quad X := \mathbf{f}(Z) \implies \mathbb{P}_X$$

Consider this other model:

$$\hat{Z} := UZ, \quad \hat{X} := \underbrace{\mathbf{f}(U^{-1}}_{\hat{\mathbf{f}}}\hat{Z}) \implies \mathbb{P}_{\hat{X}}$$

$\mathbb{P}_X = \mathbb{P}_{\hat{X}}$

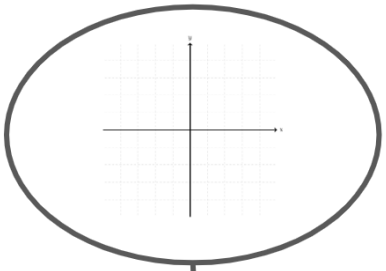
Both models represent the same distribution over X...

... but their representations can be drastically different

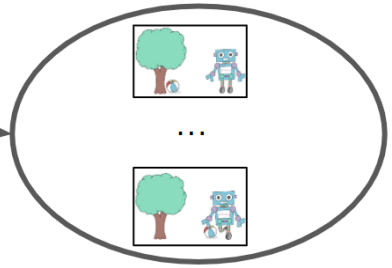
This poses a problem for **interpretability**!

What is disentanglement?

(Ground-truth) $\mathcal{Z} = \mathbb{R}^{d_z}$

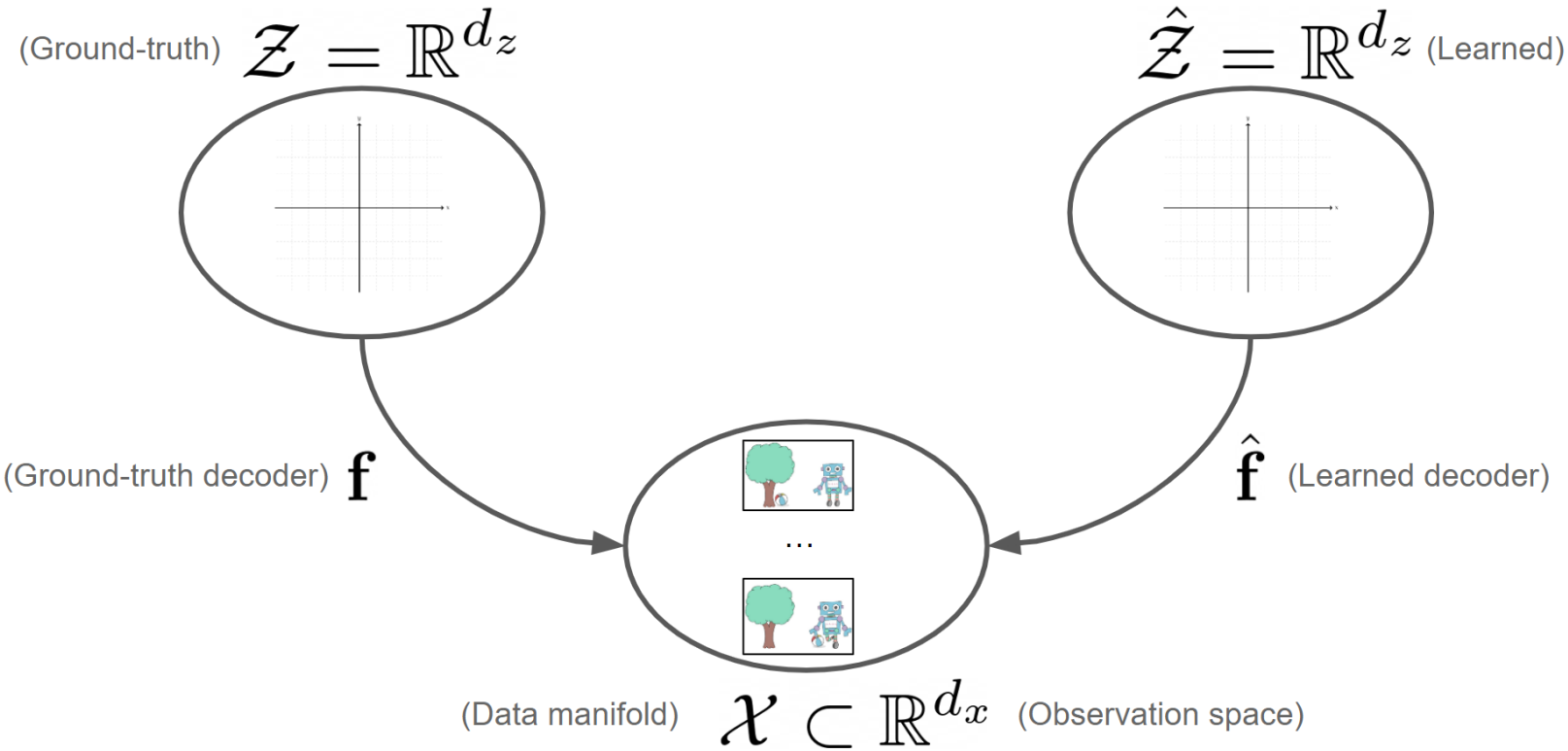


(Ground-truth decoder) $\mathbf{f}$

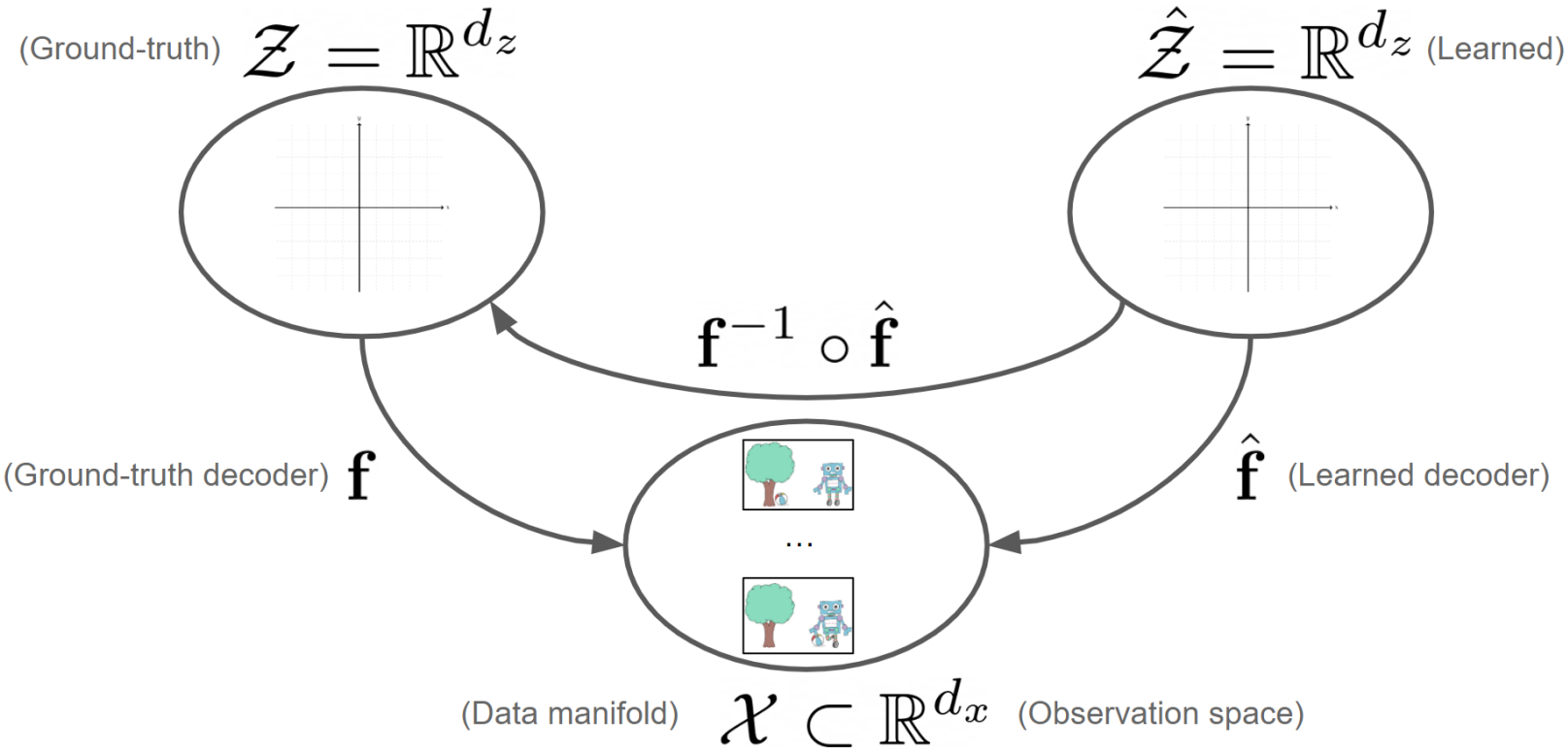


(Data manifold) $\mathcal{X} \subset \mathbb{R}^{d_x}$ (Observation space)

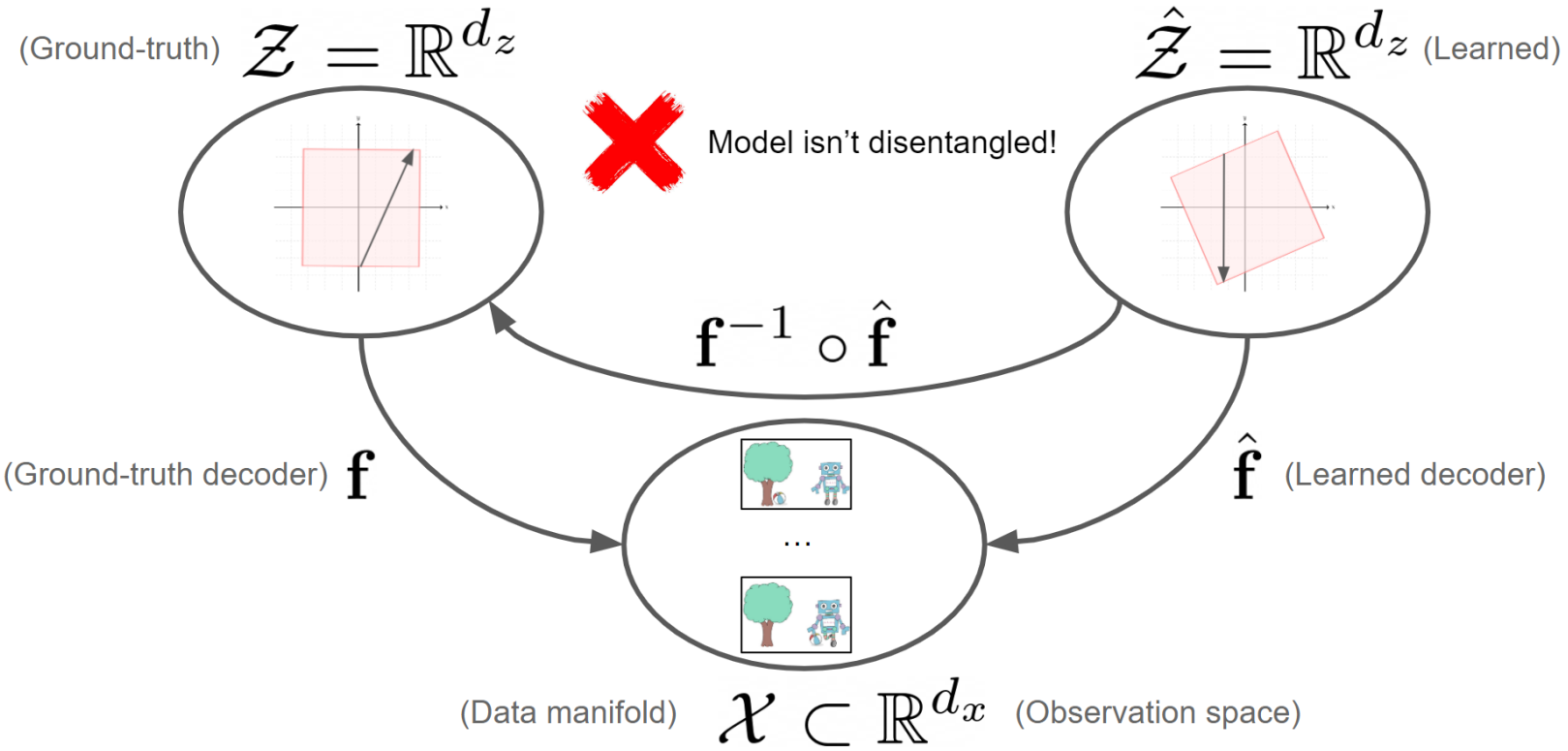
What is disentanglement?



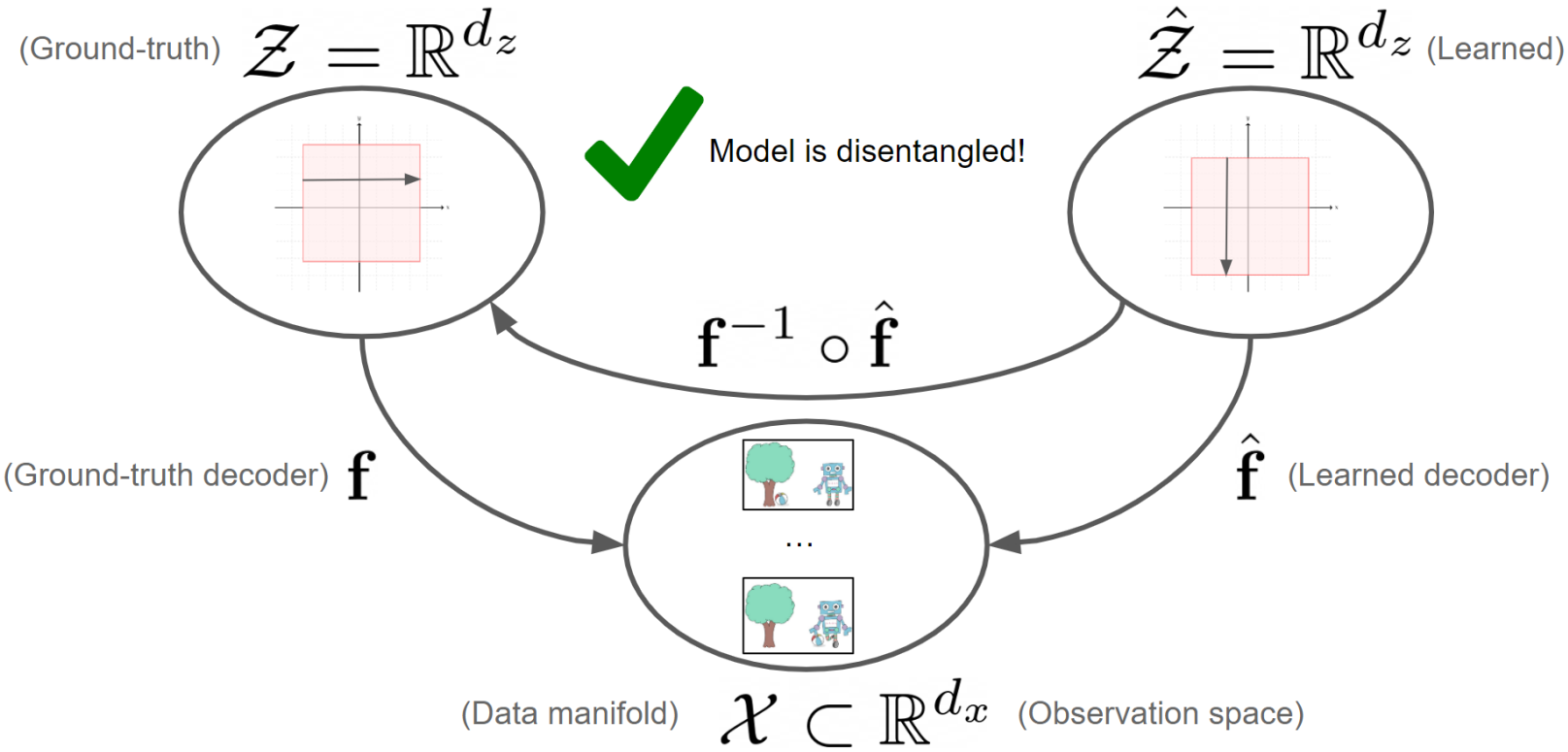
What is disentanglement?



What is disentanglement?



What is disentanglement?



Illustrating unidentifiability: Factor analysis

Representation in factor analysis is unidentifiable

Factor analysis model:

$$z \sim \mathcal{N}(0, I_k) \qquad x = Wz + \mu + \epsilon \qquad W \in \mathbb{R}^{d \times k} \qquad \epsilon \sim \mathcal{N}(0, D) \qquad \epsilon \perp\!\!\!\perp z$$

We can specify a model with a different representation z, but expressing the same marginal over x:

$$\hat{z} := Uz \ (U \text{ orthogonal}) \implies \hat{z} \sim \mathcal{N}(0, I_k)$$

$$\begin{aligned} \hat{W} := WU^\top \implies \hat{x} &= \hat{W}\hat{z} + \mu + \epsilon && (8) \\ &= WU^\top Uz + \mu + \epsilon && (9) \\ &= Wz + \mu + \epsilon = x && (10) \end{aligned}$$

Illustrating unidentifiability: Factor analysis

Representation in factor analysis is unidentifiable

Factor analysis model:

$$z \sim \mathcal{N}(0, I_k) \qquad x = Wz + \mu + \epsilon \qquad W \in \mathbb{R}^{d \times k} \qquad \epsilon \sim \mathcal{N}(0, D) \qquad \epsilon \perp\!\!\!\perp z$$

We can specify a model with a different representation z, but expressing the same marginal over x:

$$\hat{z} := Uz \ (U \text{ orthogonal}) \implies \hat{z} \sim \mathcal{N}(0, I_k)$$

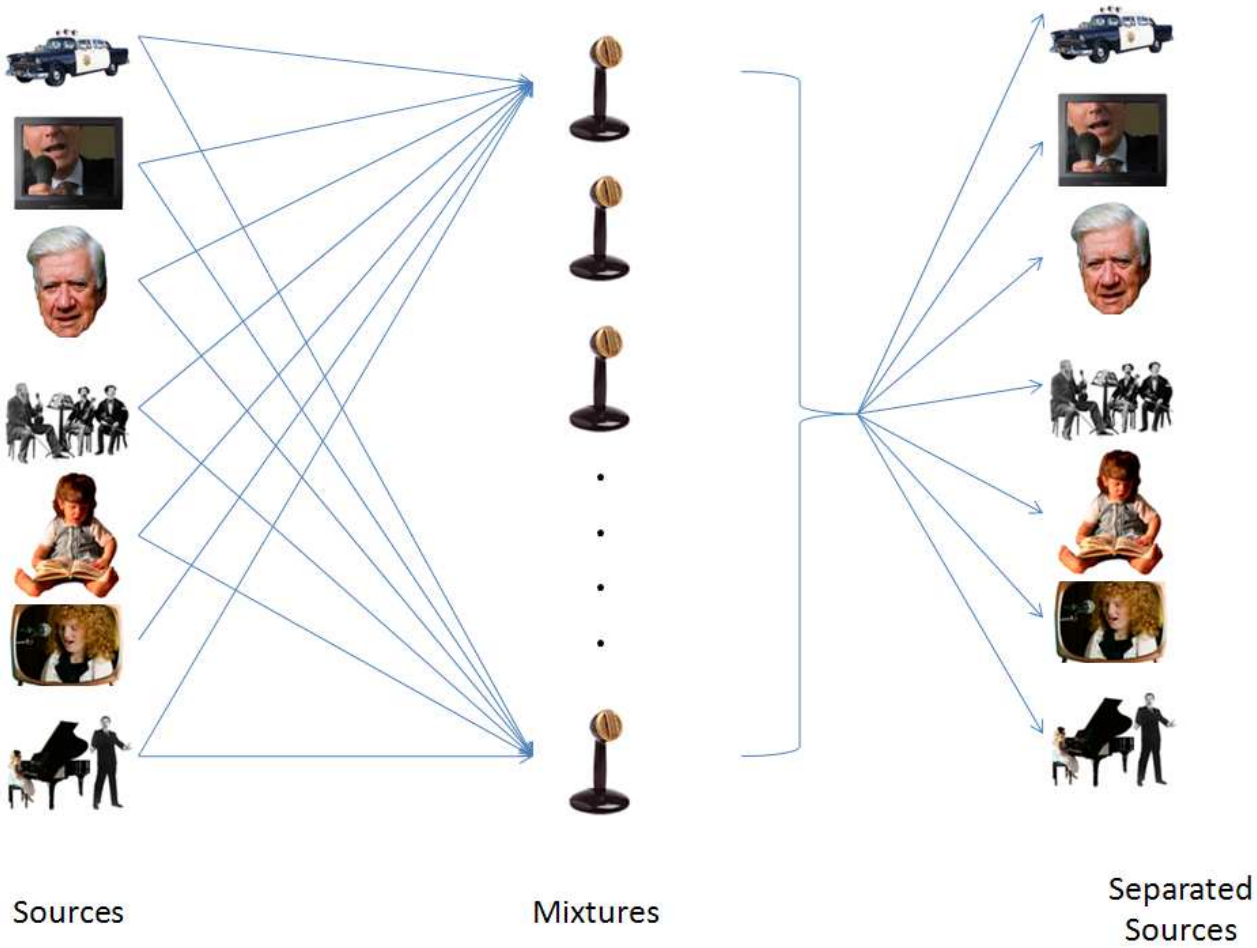
$$\begin{aligned} \hat{W} := WU^\top \implies \hat{x} &= \hat{W}\hat{z} + \mu + \epsilon && (8) \\ &= WU^\top Uz + \mu + \epsilon && (9) \\ &= Wz + \mu + \epsilon = x && (10) \end{aligned}$$

Both models have different representations $\mathbb{E}[z \mid x]$ (one is a linear transformation of the other):

$$\begin{aligned} \mathbb{E}[\hat{z} \mid \hat{x}] &= \hat{W}^\top (\hat{W}\hat{W}^\top + D)^{-1}(\hat{x} - \mu) && \text{[From class on FA]} && (11) \\ &= UW^\top (WW^\top + D)^{-1}(x - \mu) && && (12) \\ &= U\mathbb{E}[z \mid x] && && (13) \end{aligned}$$

Blind source separation

- Unidentifiability is a problem if we want to recover the "ground-truth latent factors"!



Source: https://onionesquereality.wordpress.com/2010/01/30/blind-source-separation-in-magnetic-resonance-images/

Independent component analysis (ICA)

- Is there any hope of recovering the original latents?

Independent component analysis (ICA)

- Is there any hope of recovering the original latents?
- Yes! If the latent variables are **mutually independent** and **Non-Gaussian**.

Independent component analysis (ICA)

- Is there any hope of recovering the original latents?
- Yes! If the latent variables are **mutually independent** and **Non-Gaussian**.

Theorem (Identifiability of linear ICA (Comon, 1992))

Suppose $x = Wz$ where $W \in \mathbb{R}^{d \times d}$ is invertible and where z is a random d-dimensional vector (non-constant) with mutually independent components with at most one Gaussian component. Let $A \in \mathbb{R}^{d \times d}$ be an invertible matrix such that $y := Ax$ has mutually independent components. Then $y = PDz$ where P is permutation matrix and D is an invertible diagonal matrix.

P. Comon. Independent component analysis. Higher-Order Statistics, 1992.

Independent component analysis (ICA)

- Is there any hope of recovering the original latents?
- Yes! If the latent variables are **mutually independent** and **Non-Gaussian**.

Theorem (Identifiability of linear ICA (Comon, 1992))

Suppose $x = Wz$ where $W \in \mathbb{R}^{d \times d}$ is invertible and where z is a random d-dimensional vector (non-constant) with mutually independent components with at most one Gaussian component. Let $A \in \mathbb{R}^{d \times d}$ be an invertible matrix such that $y := Ax$ has mutually independent components. Then $y = PDz$ where P is permutation matrix and D is an invertible diagonal matrix.

P. Comon. Independent component analysis. Higher-Order Statistics, 1992.

- Note that we can recover the latent factors only **up to permutation and scaling**.

Independent component analysis (ICA)

- Is there any hope of recovering the original latents?
- Yes! If the latent variables are **mutually independent** and **Non-Gaussian**.

Theorem (Identifiability of linear ICA (Comon, 1992))

Suppose $x = Wz$ where $W \in \mathbb{R}^{d \times d}$ is invertible and where z is a random d-dimensional vector (non-constant) with mutually independent components with at most one Gaussian component. Let $A \in \mathbb{R}^{d \times d}$ be an invertible matrix such that $y := Ax$ has mutually independent components. Then $y = PDz$ where P is permutation matrix and D is an invertible diagonal matrix.

P. Comon. Independent component analysis. Higher-Order Statistics, 1992.

- Note that we can recover the latent factors only **up to permutation and scaling**.
- Theorem suggests the following: Find a linear transformation of your data A such that the transformed data $y := Ax$ have mutually independent components.

Independent component analysis (ICA)

- Is there any hope of recovering the original latents?
- Yes! If the latent variables are **mutually independent** and **Non-Gaussian**.

Theorem (Identifiability of linear ICA (Comon, 1992))

Suppose $x = Wz$ where $W \in \mathbb{R}^{d \times d}$ is invertible and where z is a random d-dimensional vector (non-constant) with mutually independent components with at most one Gaussian component. Let $A \in \mathbb{R}^{d \times d}$ be an invertible matrix such that $y := Ax$ has mutually independent components. Then $y = PDz$ where P is permutation matrix and D is an invertible diagonal matrix.

P. Comon. Independent component analysis. Higher-Order Statistics, 1992.

- Note that we can recover the latent factors only **up to permutation and scaling**.
- Theorem suggests the following: Find a linear transformation of your data A such that the transformed data $y := Ax$ have mutually independent components.
- Many methods exist to acheive this: Maximizing non-gaussianity, MLE, minimizing mutual information ...etc.

Darmois-Skitovich theorem

Can prove identifiability of linear ICA via the Darmois-Skitovich theorem:

Theorem (Darmois (1953); Skitivic (1953))

Let $x_j, j = 1, \ldots, n$ with $n \geq 2$ be mutually independent random variables and let α_j, β_j be constants. Let

$$y_1 := \sum_{j=1}^{n} \alpha_j x_j \qquad y_2 := \sum_{j=1}^{n} \beta_j x_j \tag{14}$$

be two independent random variables. Then, whenever $\alpha_j \beta_j \neq 0$, the variable x_j is either constant or Gaussian.

- For a recent treatment of these ideas, see Pavan & Miranda (2018).

G. Darmois. Analyse générale des liaisons stochastiques: etude particulière de l'analyse factorielle lineaire. Revue de l'Institut International de Statistique, 1953.

V. P. Skitivic. On a property of the normal distribution. Izvestiya Akademii Nauk SSSR. Seriya Matematicheskaya, 1953.

F. R. M. Pavan and M. D. Miranda. On the darmois-skitovich theorem and spatial independence in blind source separation. Journal of Communication and Information Systems, 2018.

Independent component analysis (ICA)

Theorem (Identifiability of linear ICA (Comon, 1992))

Suppose $x = Wz$ where $W \in \mathbb{R}^{d \times d}$ is invertible and where z is a random d-dimensional vector (non-constant) with mutually independent components with at most one Gaussian component. Let $A \in \mathbb{R}^{d \times d}$ be an invertible matrix such that $y := Ax$ has mutually independent components. Then $y = PDz$ where P is permutation matrix and D is an invertible diagonal matrix.

P. Comon. Independent component analysis. Higher-Order Statistics, 1992.

- ICA amounts to finding a linear transformation A such that $y := Ax$ has mutually independent component.
- As a first step, start by making the features **decorrelated** (whitening).

Whitening a.k.a. "half ICA"

- Let's find a matrix V such that $cov(Vx) = I$.

Whitening a.k.a. "half ICA"

- Let's find a matrix V such that $cov(Vx) = I$.
- Eigen decomposition of covariance: $cov(x) = U\Lambda U^\top$, with orthogonal U
  (Symmetric $\implies$ exists an orthogonal basis of eigenvectors)
  (Positive definite $\implies$ eigenvalues are positive)

Whitening a.k.a. "half ICA"

- Let's find a matrix V such that $cov(Vx) = I$.
- Eigen decomposition of covariance: $cov(x) = U\Lambda U^\top$, with orthogonal U
  (Symmetric $\implies$ exists an orthogonal basis of eigenvectors)
  (Positive definite $\implies$ eigenvalues are positive)
- By taking $V := \Lambda^{-\frac{1}{2}} U^\top$, we get

$$cov(Vx) = Vcov(x)V^\top \tag{15}$$
$$= \Lambda^{-1/2} U^\top U \Lambda U^\top U \Lambda^{-1/2} \tag{16}$$
$$= \Lambda^{-1/2} \Lambda \Lambda^{-1/2} = I \tag{17}$$

Whitening a.k.a. "half ICA"

- Let's find a matrix V such that $cov(Vx) = I$.
- Eigen decomposition of covariance: $cov(x) = U\Lambda U^\top$, with orthogonal U
  (Symmetric $\implies$ exists an orthogonal basis of eigenvectors)
  (Positive definite $\implies$ eigenvalues are positive)
- By taking $V := \Lambda^{-\frac{1}{2}} U^\top$, we get

$$cov(Vx) = Vcov(x)V^\top \tag{15}$$
$$= \Lambda^{-1/2} U^\top U \Lambda U^\top U \Lambda^{-1/2} \tag{16}$$
$$= \Lambda^{-1/2} \Lambda \Lambda^{-1/2} = I \tag{17}$$

- We denote the whitened data by $\bar{x} := Vx$.
- Exercise: Show that, for any orthogonal matrix A, $cov(A\bar{x}) = I$.

Whitening a.k.a. "half ICA"

- Let's find a matrix V such that $cov(Vx) = I$.
- Eigen decomposition of covariance: $cov(x) = U\Lambda U^\top$, with orthogonal U
  (Symmetric $\implies$ exists an orthogonal basis of eigenvectors)
  (Positive definite $\implies$ eigenvalues are positive)
- By taking $V := \Lambda^{-\frac{1}{2}} U^\top$, we get

$$cov(Vx) = Vcov(x)V^\top \tag{15}$$
$$= \Lambda^{-1/2} U^\top U \Lambda U^\top U \Lambda^{-1/2} \tag{16}$$
$$= \Lambda^{-1/2} \Lambda \Lambda^{-1/2} = I \tag{17}$$

- We denote the whitened data by $\bar{x} := Vx$.
- Exercise: Show that, for any orthogonal matrix A, $cov(A\bar{x}) = I$.
- Recall that independence implies zero covariance, but that the converse is false!
- So to perform ICA, we need to go one step further and find the orthogonal matrix A that makes the latents independent.

Objectives to perform ICA

Most algorithms to perform ICA first whiten the data ($\bar{x} = Vx$) and then search for an orthogonal matrix A that optimizes one of these objectives.

- MLE: Choose a model class for the distribution of the latents $p_z(z) = \prod_{j=1}^{d} p_j(z_j)$ (common choice is Laplacian, to induce sparsity) and maximize log-likelihood:

$$\sum_{i=1}^{n} \log p(\bar{x}^{(i)}; A) = \frac{1}{n} \sum_{i=1}^{n} \log p_z(A\bar{x}^{(i)}) + \underbrace{\log |\det A|}_{=0}$$

Objectives to perform ICA

Most algorithms to perform ICA first whiten the data ($\bar{x} = Vx$) and then search for an orthogonal matrix A that optimizes one of these objectives.

- MLE: Choose a model class for the distribution of the latents $p_z(z) = \prod_{j=1}^{d} p_j(z_j)$ (common choice is Laplacian, to induce sparsity) and maximize log-likelihood:

$$\sum_{i=1}^{n} \log p(\bar{x}^{(i)}; A) = \frac{1}{n} \sum_{i=1}^{n} \log p_z(A\bar{x}^{(i)}) + \underbrace{\log |\det A|}_{=0}$$

- Maximizing non-gaussianity via kurtosis (Related to fourth-moment $\mathbb{E}[y_j^4]$). Gaussian distribution has kurtosis = 0.

Objectives to perform ICA

Most algorithms to perform ICA first whiten the data ($\bar{x} = Vx$) and then search for an orthogonal matrix A that optimizes one of these objectives.

- MLE: Choose a model class for the distribution of the latents $p_z(z) = \prod_{j=1}^{d} p_j(z_j)$ (common choice is Laplacian, to induce sparsity) and maximize log-likelihood:

$$\sum_{i=1}^{n} \log p(\bar{x}^{(i)}; A) = \frac{1}{n} \sum_{i=1}^{n} \log p_z(A\bar{x}^{(i)}) + \underbrace{\log |\det A|}_{=0}$$

- Maximizing non-gaussianity via kurtosis (Related to fourth-moment $\mathbb{E}[y_j^4]$). Gaussian distribution has kurtosis = 0.
- Minimizing mutual information between the components of $y := A\bar{x}$.
- See Hyvarinen et al. (2001) for more details!

A. Hyvärinen, J. Karhunen, and E. Oja. Independent Component Analysis. Wiley, 2001.

ICA via temporal dependencies (AMUSE algorithm, Tong et al., (1990))

- Instead of leveraging higher-order statistics, can we leverage temporal correlations?

ICA via temporal dependencies (AMUSE algorithm, Tong et al., (1990))

- Instead of leveraging higher-order statistics, can we leverage temporal correlations?
- Assume the sequence of latents $\{z_t\}_t$ forms a "wide-sense stationary process" i.e.
  - Expectation $\mathbb{E}[z_t]$ does not depend on t (and equals 0)
  - Covariance matrix $cov(z_t)$ does not depend on t
  - Lagged covariance matrices $cov(z_t, z_{t-\tau})$ do not dependent on t (but can depend on τ)

ICA via temporal dependencies (AMUSE algorithm, Tong et al., (1990))

- Instead of leveraging higher-order statistics, can we leverage temporal correlations?
- Assume the sequence of latents $\{z_t\}_t$ forms a "wide-sense stationary process" i.e.
  - Expectation $\mathbb{E}[z_t]$ does not depend on t (and equals 0)
  - Covariance matrix $cov(z_t)$ does not depend on t
  - Lagged covariance matrices $cov(z_t, z_{t-\tau})$ do not dependent on t (but can depend on τ)
- We assume the components are decorrelated. Formally $cov(z_t) = I$ and $cov(z_t, z_{t-\tau}) = D_\tau$, where D_τ is diagonal.

ICA via temporal dependencies (AMUSE algorithm, Tong et al., (1990))

- Instead of leveraging higher-order statistics, can we leverage temporal correlations?
- Assume the sequence of latents $\{z_t\}_t$ forms a "wide-sense stationary process" i.e.
  - Expectation $\mathbb{E}[z_t]$ does not depend on t (and equals 0)
  - Covariance matrix $cov(z_t)$ does not depend on t
  - Lagged covariance matrices $cov(z_t, z_{t-\tau})$ do not dependent on t (but can depend on τ)
- We assume the components are decorrelated. Formally $cov(z_t) = I$ and $cov(z_t, z_{t-\tau}) = D_\tau$, where D_τ is diagonal.
- $x_t = Wz_t$

ICA via temporal dependencies (AMUSE algorithm, Tong et al., (1990))

- Instead of leveraging higher-order statistics, can we leverage temporal correlations?
- Assume the sequence of latents $\{z_t\}_t$ forms a "wide-sense stationary process" i.e.
  - Expectation $\mathbb{E}[z_t]$ does not depend on t (and equals 0)
  - Covariance matrix $cov(z_t)$ does not depend on t
  - Lagged covariance matrices $cov(z_t, z_{t-\tau})$ do not dependent on t (but can depend on τ)
- We assume the components are decorrelated. Formally $cov(z_t) = I$ and $cov(z_t, z_{t-\tau}) = D_\tau$, where D_τ is diagonal.
- $x_t = Wz_t$
- Note that $cov(x_t) = Wcov(z_t)W^\top = WW^\top$

ICA via temporal dependencies (AMUSE algorithm, Tong et al., (1990))

$$cov(z_t) = I \quad cov(z_t, z_{t-\tau}) = D_\tau \text{ (diagonal)} \quad x_t = Wz_t \quad cov(x_t) = WW^\top$$

- Start by whitening the data:

$$cov(x_t) = WW^\top = U\Lambda U^\top$$

$$\bar{x}_t := \Lambda^{-1/2} U^\top x_t = \underbrace{\Lambda^{-1/2} U^\top W}_{\bar{W}:=} z_t$$

ICA via temporal dependencies (AMUSE algorithm, Tong et al., (1990))

$$cov(z_t) = I \quad cov(z_t, z_{t-\tau}) = D_\tau \text{ (diagonal)} \quad x_t = Wz_t \quad cov(x_t) = WW^\top$$

- Start by whitening the data:

$$cov(x_t) = WW^\top = U\Lambda U^\top$$

$$\bar{x}_t := \Lambda^{-1/2} U^\top x_t = \underbrace{\Lambda^{-1/2} U^\top W}_{\bar{W}:=} z_t$$

- We would like to recover $\bar{W}$ up to permutation of its columns, since with it, we can infer the latents associated to an observation x by doing $\bar{W}\bar{x}$.

ICA via temporal dependencies (AMUSE algorithm, Tong et al., (1990))

$$cov(z_t) = I \quad cov(z_t, z_{t-\tau}) = D_\tau \text{ (diagonal)} \quad x_t = Wz_t \quad cov(x_t) = WW^\top$$

- Start by whitening the data:

$$cov(x_t) = WW^\top = U\Lambda U^\top$$

$$\bar{x}_t := \Lambda^{-1/2} U^\top x_t = \underbrace{\Lambda^{-1/2} U^\top W}_{\bar{W}:=} z_t$$

- We would like to recover $\bar{W}$ up to permutation of its columns, since with it, we can infer the latents associated to an observation x by doing $\bar{W}\bar{x}$.

- Turns out $\bar{W}$ is orthogonal:

$$\bar{W}\bar{W}^\top = \Lambda^{-1/2} U^\top WW^\top U\Lambda^{-1/2} = \Lambda^{-1/2} U^\top U\Lambda U^\top U\Lambda^{-1/2} = I$$

ICA via temporal dependencies (AMUSE algorithm, Tong et al., (1990))

$$cov(z_t) = I \quad cov(z_t, z_{t-\tau}) = D_\tau \text{ (diagonal)} \quad x_t = Wz_t \quad cov(x_t) = WW^\top$$

$$\bar{x}_t := \Lambda^{-1/2} U^\top x_t \text{ (Whitened } x_t) \quad \bar{x}_t = \bar{W} z_t \quad \bar{W}\bar{W}^\top = I$$

ICA via temporal dependencies (AMUSE algorithm, Tong et al., (1990))

$$cov(z_t) = I \quad cov(z_t, z_{t-\tau}) = D_\tau \text{ (diagonal)} \quad x_t = Wz_t \quad cov(x_t) = WW^\top$$

$$\bar{x}_t := \Lambda^{-1/2} U^\top x_t \text{ (Whitened } x_t) \quad \bar{x}_t = \bar{W} z_t \quad \bar{W}\bar{W}^\top = I$$

- Consider the lagged covariance between $\bar{x}_t$ and $\bar{x}_{t-\tau}$, which can be estimated empirically!

$$\begin{aligned} cov(\bar{x}_t, \bar{x}_{t-\tau}) &= \mathbb{E}[\bar{x}_t \bar{x}_{t-\tau}^\top] && (18) \\ &= \mathbb{E}[\bar{W} z_t z_{t-\tau}^\top \bar{W}^\top] && (19) \\ &= \bar{W} cov(z_t, z_{t-\tau}) \bar{W}^\top && (20) \\ &= \bar{W} D_\tau \bar{W}^\top && (21) \end{aligned}$$

ICA via temporal dependencies (AMUSE algorithm, Tong et al., (1990))

$$cov(z_t) = I \quad cov(z_t, z_{t-\tau}) = D_\tau \text{ (diagonal)} \quad x_t = Wz_t \quad cov(x_t) = WW^\top$$

$$\bar{x}_t := \Lambda^{-1/2} U^\top x_t \text{ (Whitened } x_t) \quad \bar{x}_t = \bar{W} z_t \quad \bar{W}\bar{W}^\top = I$$

- Consider the lagged covariance between $\bar{x}_t$ and $\bar{x}_{t-\tau}$, which can be estimated empirically!

$$\begin{aligned} cov(\bar{x}_t, \bar{x}_{t-\tau}) &= \mathbb{E}[\bar{x}_t \bar{x}_{t-\tau}^\top] && (18) \\ &= \mathbb{E}[\bar{W} z_t z_{t-\tau}^\top \bar{W}^\top] && (19) \\ &= \bar{W} cov(z_t, z_{t-\tau}) \bar{W}^\top && (20) \\ &= \bar{W} D_\tau \bar{W}^\top && (21) \end{aligned}$$

- How cool! The matrix $\bar{W}$ appears in an eigendecomposition of $cov(\bar{x}_t, \bar{x}_{t-\tau})$!

ICA via temporal dependencies (AMUSE algorithm, Tong et al., (1990))

$$cov(z_t) = I \quad cov(z_t, z_{t-\tau}) = D_\tau \text{ (diagonal)} \quad x_t = Wz_t \quad cov(x_t) = WW^\top$$

$$\bar{x}_t := \Lambda^{-1/2} U^\top x_t \text{ (Whitened } x_t\text{)} \quad \bar{x}_t = \bar{W} z_t \quad \bar{W}\bar{W}^\top = I$$

- Consider the lagged covariance between $\bar{x}_t$ and $\bar{x}_{t-\tau}$, which can be estimated empirically!

$$\begin{aligned} cov(\bar{x}_t, \bar{x}_{t-\tau}) &= \mathbb{E}[\bar{x}_t \bar{x}_{t-\tau}^\top] && (18) \\ &= \mathbb{E}[\bar{W} z_t z_{t-\tau}^\top \bar{W}^\top] && (19) \\ &= \bar{W} cov(z_t, z_{t-\tau}) \bar{W}^\top && (20) \\ &= \bar{W} D_\tau \bar{W}^\top && (21) \end{aligned}$$

- How cool! The matrix $\bar{W}$ appears in an eigendecomposition of $cov(\bar{x}_t, \bar{x}_{t-\tau})$!
- But is this decomposition unique up to permutation and rescaling? If the entries of D_τ are all distinct, then yes! (Because each eigenspace is one-dimensional)
- This means we can estimate $\bar{W}$ by diagonalizing $cov(\bar{x}_t, \bar{x}_{t-\tau})$

ICA via temporal dependencies (AMUSE algorithm, Tong et al., (1990))

Practical consideration:

- In practice, the empirical $cov(\bar{x}_t, \bar{x}_{t-\tau})$ is not symmetric, and thus we can't find an orthogonal basis of eigenvectors.

- AMUSE algorithm uses a trick to symmetrize it (Tong et al., 1990).

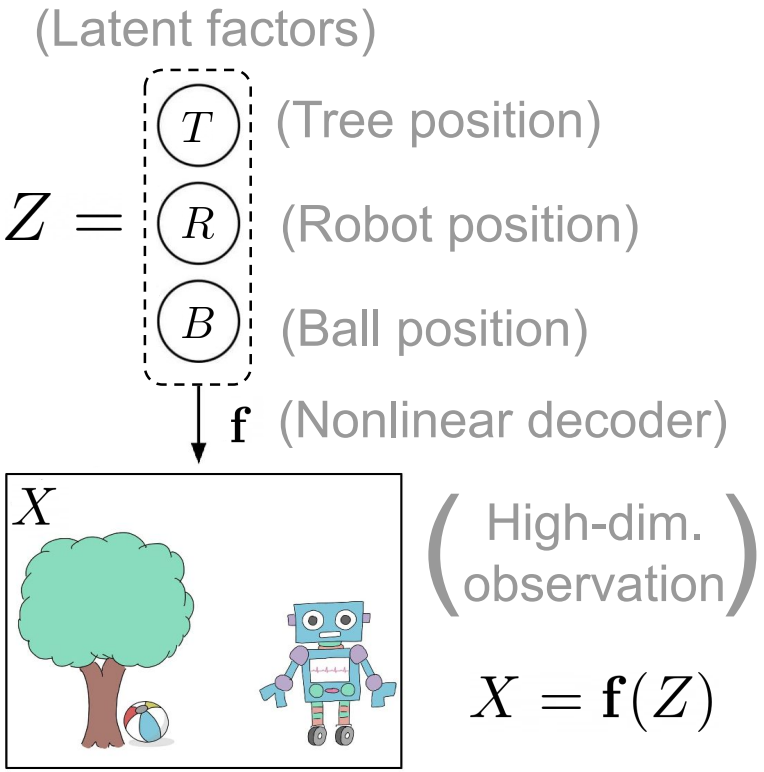
ICA via temporal dependencies (AMUSE algorithm, Tong et al., (1990))

Practical consideration:

- In practice, the empirical $cov(\bar{x}_t, \bar{x}_{t-\tau})$ is not symmetric, and thus we can't find an orthogonal basis of eigenvectors.
- AMUSE algorithm uses a trick to symmetrize it (Tong et al., 1990).
- Can leverage multiple time lags via **simultaneous diagonalization**.

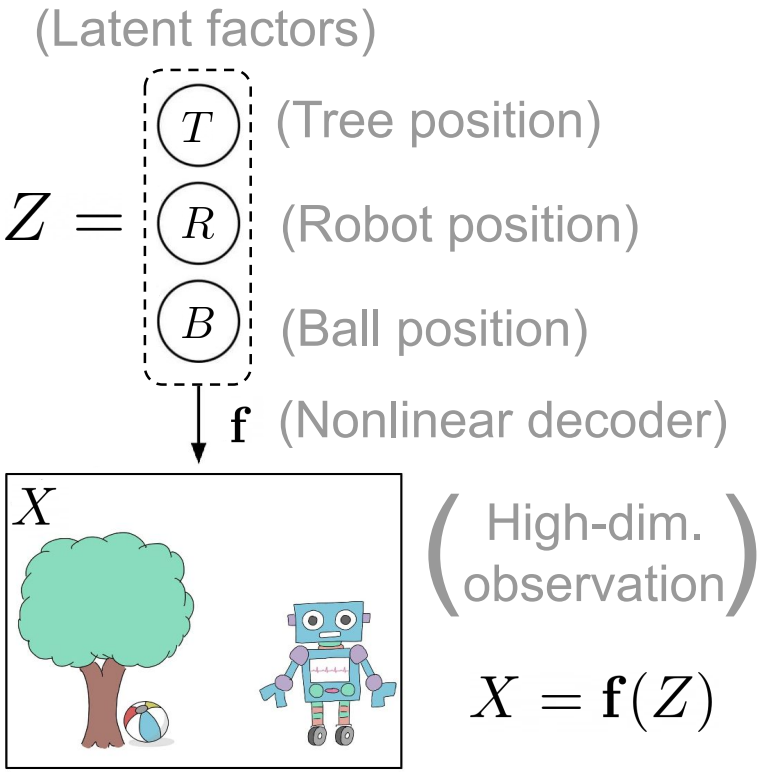
L. Tong, V.C. Soon, Y.F. Huang, and R. Liu. Amuse: a new blind identification algorithm. In IEEE International Symposium on Circuits and Systems, 1990.

Back to initial motivation...



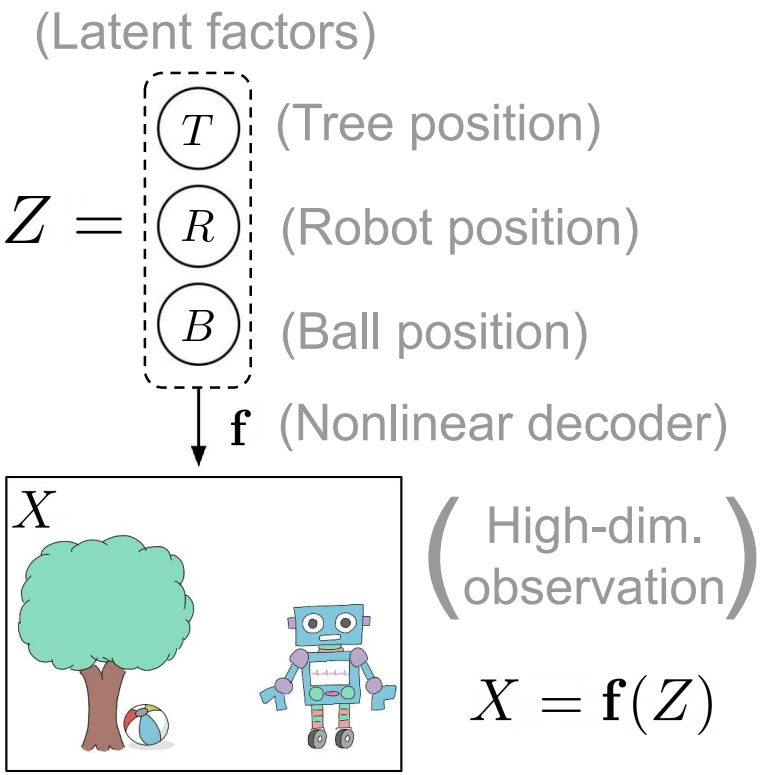
- For more involved application, the "linear decoder" assumption does not hold...

Back to initial motivation...



- For more involved application, the "linear decoder" assumption does not hold...
- Can we prove identifiability for nonlinear decoder?

Back to initial motivation...



- For more involved application, the "linear decoder" assumption does not hold...
- Can we prove identifiability for nonlinear decoder?
- It turns out independence and non-gaussianity of the latents are **insufficient** in that case (Hyvarinen & Pajunen, 1999)
- We need stronger assumptions...

A. Hyvarinen and P. Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. Neural Networks, 1999.

Identifiability results for Nonlinear ICA (far from exhaustive list)

- Leveraging contrastive learning and (diagonal) temporal dependencies

  A. Hyvarinen and H. Morioka. **Nonlinear ICA of Temporally Dependent Stationary Sources.** In Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 2017.

- Leveraging VAE's and non-stationarity of the sources

  I. Khemakhem, D. Kingma, R. Monti, and A. Hyvarinen. **Variational autoencoders and nonlinear ICA: A unifying framework.** In Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics, 2020.

- Leveraging sparse temporal dependencies (not necessarily diagonal) and interventions on the latents

  S. Lachapelle, P. Rodriguez Lopez, Y. Sharma, K. E. Everett, R. Le Priol, A. Lacoste, and S. Lacoste-Julien. **Disentanglement via mechanism sparsity regularization: A new principle for nonlinear ICA.** In First Conference on Causal Learning and Reasoning, 2022.