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 \((p, G)\) s.t.

- \(G\) is a directed acyclic graph (DAG)

- \(p \in \mathcal{L}(G)\), i.e. \(p\) factorizes according to \(G\).

- \(G\) 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 \((p, G)\) s.t.

- \(G\) is a **directed acyclic graph** (DAG)

- \(p \in \mathcal{L}(G)\), i.e. \(p\) factorizes according to \(G\).

- \(G\) describes **causal relationships** between variables, i.e., how the system reacts to **interventions**.

**Example: Kidney stone treatment**

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

\[
p(S, T, R) = p(S)p(T \mid S)p(R \mid S, T)
\]
The "do" operator models the effect of interventions

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

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

**Definition (The “do” operator)**

Given a causal graphical model $(p, G)$,

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

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 \( p(x) = \prod_i p(x_i \mid x_{\pi_i^G}) \)

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

**Definition (The “do” operator)**

Given a causal graphical model \((p, G)\),

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

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.

\[
p(x_i \mid do(x_k := x'_k)) = \sum_{x_{V \setminus \{i\}}} p(x \mid do(x_k := x'_k))
\]
The "do" operator models the effect of interventions

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

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

**Definition (The “do” operator)**

Given a causal graphical model $(p, G)$,

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

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. $p(x_i \mid do(x_k := x'_k)) = \sum_{x_{V \setminus \{i\}}} p(x \mid do(x_k := x'_k))$

- ... and conditionals, e.g. $p(x_i \mid x_j, do(x_k := x'_k)) = \frac{p(x_i, x_j \mid do(x_k := x'_k))}{p(x_j \mid do(x_k := x'_k))}$
The "do" operator models the effect of interventions

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

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

**Definition (The “do” operator)**

Given a causal graphical model $(p, G)$,

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

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. $p(x_i \mid do(x_k := x'_k)) = \sum_{x_{V \setminus \{i\}}} p(x \mid do(x_k := x'_k))$

- ... and conditionals, e.g. $p(x_i \mid x_j, do(x_k := x'_k)) = \frac{p(x_i, x_j \mid do(x_k := x'_k))}{p(x_j \mid do(x_k := x'_k))}$

**Truncated factorization:**

$$p(x_{V \setminus \{k\}} \mid do(x_k := x'_k)) = \sum_{x_k} \delta(x_k, x'_k) \prod_{i \neq k} p(x_i \mid x_{\pi_i}) = \prod_{i \neq k} p(x_i \mid x_{\pi_i}).$$
Conditioning is not the same as doing

Consider the simple CGM $X \rightarrow Y$

$$p(X|do(Y := Y')) = p(Y|X)p(X)$$

$$= p(X)$$

$$\neq p(X \mid Y = Y')$$
The "do" operator

- Back to our example

\[
P(S, R \mid do(T = T')) = P(S) P(T \mid S) P(R \mid S, T')
\]

The decision of taking treatment \( T \) does not depend on \( S \) anymore.
The "do" operator

- Back to our example

\[ P(S, R \mid do(T = T')) = P(S) P(T \mid S) P(R \mid S, T') \]

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

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

\[ G' = (V, E') \quad E' = \{(i, j) \in E \mid j \neq k\} \]
The "do" operator

- Back to our example

\[
P(S, R \mid do(T = T')) = P(S) \frac{P(T \mid S)}{P(T \mid S \mid do(T = T'))} P(R \mid S, T')
\]

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 $\in \{\text{small, large}\}$  
$R =$ Patient recovered $\in \{0, 1\}$

**Observations**

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

**Perfect intervention**

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

**Imperfect intervention**

- $p(S)p(T \mid S)p(R \mid S, T)$
Different types of interventions

Intervening on the treatment $T$

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

**Observations**

$p(S)p(T | S)p(R | S, T)$

**Perfect intervention**

$p(S)p(T)p(R | S, T)$

**Imperfect intervention**

$p(S)p(T | S)p(R | S, T)$

Definition presented previously is a perfect intervention with $\bar{p}(T) := \delta(T, T')$. It is sometimes called a **perfect deterministic intervention**.
Why should I care!?! (Kidney Stone Treatment)

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

\[p(S)p(T \mid S)p(R \mid S, T)\]

(Example taken from *Element of Causal Inference* by Peters et al. p111)
Why should I care!?! (Kidney Stone Treatment)

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

\[ p(S)p(T \mid S)p(R \mid S, T) \]

(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...

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)
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)

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

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 ... \( P(S = \text{large} \mid T = A) = 0.75 \)

- ... which reduces your chance of recovery
  \[ P(R = 1 \mid T = A, S = \text{large}) = 0.73 < 0.93 = P(R = 1 \mid T = A, S = \text{small}) \]
Why should I care?! (Kidney Stone Treatment)

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

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 ... $P(S = \text{large} \mid T = A) = 0.75$

- ... which reduces your chance of recovery
  $P(R = 1 \mid T = A, S = \text{large}) = 0.73 < 0.93 = P(R = 1 \mid T = A, S = \text{small})$

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)

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

\[
P(S, R \mid do(T)) = P(S) P(T \mid S) P(R \mid S, T)
\]

The decision of taking treatment \( T \) does not depend on \( S \) anymore.

Then simply marginalize as usual:

\[
P(R = 1 \mid do(T = A)) = \sum_S P(R = 1, S \mid do(T = A))
\]

\[
= \sum_S P(R = 1 \mid S, T = A) P(S) = 0.832
\]
Why should I care!?! (Kidney Stone Treatment)

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

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

\[ P(R = 1|T = A) = 0.78 \]
\[ P(R = 1|T = B) = 0.83 \]

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

\[ P(R = 1|do(T = A)) = 0.832 \]
\[ P(R = 1|do(T = B)) = 0.782 \]
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

\[ P(R = 1 | \text{do}(T = A)) = \sum_S P(R = 1 | S, T = A)P(S) \]

Never observed data from \( p(T, S, R | \text{do}(T = A)) \)

...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

\[
P(R = 1 | \text{do}(T = A)) = \sum_S P(R = 1 | S, T = A) P(S)
\]

Never observed data from \( p(T, S, R | \text{do}(T = A)) \)

...Yet I can estimate the query, since there is no "do" here :D

Formally, this means \( p(R = 1 \mid \text{do}(T = A)) \) is **identifiable from** \( p(R, T, S) \) and \( G \)

(our computations critically relied on the causal graph).
What just happened? We showed

\[
P(R = 1|do(T = A)) = \sum_S P(R = 1|S, T = A)P(S)
\]

Never observed data from \(p(T, S, R | do(T = A))\)

...Yet I can estimate the query, since there is no "do" here :D

Formally, this means \(p(R = 1 | do(T = A))\) is identifiable from \(p(R, T, S)\) and \(G\) (our computations critically relied on the causal graph).

Turns out what we just did is an instance of the backdoor criterion...
Theorem (Backdoor criterion)

\[ p(x_i \mid do(x_k)) = \sum_{x_S} p(x_i \mid x_k, x_S)p(x_S) \quad \text{if} \]

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 \( x_k \leftarrow \ldots x_i \).
Theorem (Backdoor criterion)

\[ p(x_i \mid do(x_k)) = \sum_{x_S} p(x_i \mid x_k, x_S)p(x_S) \text{ if} \]

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 \( p(y \mid do(x)) \):

**Left path:** Only backdoor path. Blocked by \( S = \{K\} \).

**Right path:** Why we cannot include a descendant of \( X \) in \( S \).
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 $p(y \mid do(x))$ from $p(X, Z, Y)$ using the front-door criterion. Look it up!
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 \( p(y \mid do(x)) \) from \( p(X, Z, Y) \) using the front-door criterion. Look it up!
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 $p(y \mid do(x))$ from $p(X, Z, Y)$ using the **front-door criterion**. Look it up!
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**.
You now know about the first two steps of Pearl’s "ladder of causation".

<table>
<thead>
<tr>
<th>Level (Symbol)</th>
<th>Typical Activity</th>
<th>Typical Questions</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Association $P(y</td>
<td>x)$</td>
<td>Seeing</td>
<td>What is? How would seeing $X$ change my belief in $Y$?</td>
</tr>
<tr>
<td>2. Intervention $P(y</td>
<td>do(x), z)$</td>
<td>Doing Intervening</td>
<td>What if? What if I do $X$?</td>
</tr>
<tr>
<td>3. Counterfactuals $P(y_x</td>
<td>x', y')$</td>
<td>Imagining, Retrospection</td>
<td>Why? Was it $X$ that caused $Y$? What if I had acted differently?</td>
</tr>
</tbody>
</table>

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

\[
X_1 := f_1(X_{\pi_1}, N_1) \hspace{1cm} (4) \\
X_2 := f_2(X_{\pi_2}, N_2) \hspace{1cm} (5) \\
... \hspace{1cm} (6) \\
X_d := f_d(X_{\pi_d}, N_d) \hspace{1cm} (7)
\]

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

\[ Z \sim P_Z, \quad X := f(Z) \quad \Rightarrow \quad P_X \]

Consider this other model:

\[ \hat{Z} := UZ, \quad \hat{X} := f(U^{-1}\hat{Z}) \quad \Rightarrow \quad P_{\hat{X}} \]
The general problem of identifiability for generative models

Consider the following simple generative model:

\[ Z \sim P_Z, \quad X := f(Z) \quad \implies \quad P_X \]

Consider this other model:

\[ \hat{Z} := UZ, \quad \hat{X} := f(U^{-1} \hat{Z}) \quad \implies \quad P_{\hat{X}} \]

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:

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

Consider this other model:

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

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) $f$

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

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

(Ground-truth decoder) \( f \)

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

\( \cdots \)

(Learned) \( \hat{\mathcal{Z}} = \mathbb{R}^{d_z} \)

(Learned decoder) \( \hat{f} \)
What is disentanglement?

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

(Learned) \( \hat{\mathcal{Z}} = \mathbb{R}^{d_z} \)

(Ground-truth decoder) \( f \)

(Learned decoder) \( \hat{f} \)

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

(Observation space)
What is disentanglement?

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

Model isn’t disentangled!

(Learned) \( \hat{\mathcal{Z}} = \mathbb{R}^{d_z} \)

(Ground-truth decoder) \( f \)

(Learned decoder) \( \hat{f} \)

(Data manifold) \( \mathcal{X} \subset \mathbb{R}^{d_x} \) (Observation space)
What is disentanglement?

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

Model is disentangled!

(Learned) \( \hat{\mathcal{Z}} = \mathbb{R}^{d_z} \)

(Ground-truth decoder) \( f \)

(Learned decoder) \( \hat{f} \)

(Data manifold) \( \mathcal{X} \subset \mathbb{R}^{d_x} \) (Observation space)
Representing unidentifiability: Factor analysis

**Factor analysis model:**

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

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

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

\[ \hat{W} := WU^\top \quad \implies \hat{x} = \hat{W}\hat{z} + \mu + \epsilon \quad (8) \]

\[ = WU^\top Uz + \mu + \epsilon \quad (9) \]

\[ = Wz + \mu + \epsilon = x \quad (10) \]
Illustrating unidentifiability: Factor analysis

Representation in factor analysis is unidentifiable

Factor analysis model:

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

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

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

\[ \hat{W} := WU^\top \implies \hat{x} = \hat{W}\hat{z} + \mu + \epsilon \]

\[ = WU^\top Uz + \mu + \epsilon \]

\[ = Wz + \mu + \epsilon = x \]

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

\[ \mathbb{E}[\hat{z} \mid \hat{x}] = \hat{W}^\top (\hat{W}\hat{W}^\top + D)^{-1}(\hat{x} - \mu) \]

[From class on FA]

\[ = UW^\top (WW^\top + D)^{-1}(x - \mu) \]

\[ = U\mathbb{E}[z \mid x] \]
Unidentifiability is a problem if we want to recover the "ground-truth latent factors"!

Is there any hope of recovering the original latents?

**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.


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 achieve this: Maximizing non-gaussianity, MLE, minimizing mutual information ... etc.
Independent component analysis (ICA)

- Is there any hope of recovering the original latents?

- Yes! If the latent variables are **mutually independent** and **Non-Gaussian**.
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.

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.


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.


- 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.
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.

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 achieve this: Maximizing non-gaussianity, MLE, minimizing mutual information ...etc.
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 $\alpha_j, \beta_j$ be constants. Let

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

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).

---


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.


- 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 \textit{decorrelated} (whitening).
Whitening a.k.a. "half ICA"

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

Eigen decomposition of covariance:

$\text{cov}(x) = U \Lambda U^\top$, with orthogonal $U$ ($\text{Symmetric} \Rightarrow \text{exists an orthogonal basis of eigenvectors}$) ($\text{Positive definite} \Rightarrow \text{eigenvalues are positive}$)

By taking $V = \Lambda^{-1/2} U^\top$, we get

$\text{cov}(Vx) = V \text{cov}(x) V^\top$

$= \Lambda^{-1/2} U^\top U \Lambda U^\top U \Lambda^{-1/2}$

$= \Lambda^{-1/2} \Lambda \Lambda^{-1/2}$

$= I$  \hspace{1cm} (17)

We denote the whitened data by $\bar{x}$:

$\bar{x} = Vx$.

Exercise: Show that, for any orthogonal matrix $A$, $\text{cov}(A\bar{x}) = I$.

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.
Whitening a.k.a. "half ICA"

- Let’s find a matrix $V$ such that $\text{cov}(Vx) = I$.

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

Eigen decomposition of covariance: $\text{cov}(x) = U\Lambda U^\top$, 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

\[
\text{cov}(Vx) = V\text{cov}(x)V^\top
\]
\[
= \Lambda^{-1/2} U^\top U\Lambda U^\top U\Lambda^{-1/2}
\]
\[
= \Lambda^{-1/2} \Lambda \Lambda^{-1/2} = I
\]
Whitening a.k.a. "half ICA"

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

- Eigen decomposition of covariance: $\text{cov}(x) = U\Lambda U^\top$, with orthogonal $U$
  (Symmetric $\iff$ exists an orthogonal basis of eigenvectors)
  (Positive definite $\iff$ eigenvalues are positive)

- By taking $V := \Lambda^{-\frac{1}{2}} U^\top$, we get
  \[ \text{cov}(Vx) = V\text{cov}(x)V^\top = \Lambda^{-1/2} U^\top U\Lambda U^\top U\Lambda^{-1/2} = \Lambda^{-1/2} \Lambda \Lambda^{-1/2} = I \]

- We denote the whitened data by $\bar{x} := Vx$.

- Exercise: Show that, for any orthogonal matrix $A$, $\text{cov}(A\bar{x}) = I$. 

Whitening a.k.a. "half ICA"

- Let’s find a matrix $V$ such that $\text{cov}(Vx) = I$.

- Eigen decomposition of covariance: $\text{cov}(x) = U\Lambda U^\top$, 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
  \[
  \text{cov}(Vx) = V\text{cov}(x)V^\top = \Lambda^{-1/2} U^\top U\Lambda U^\top U\Lambda^{-1/2} = \Lambda^{-1/2} \Lambda \Lambda^{-1/2} = I
  \]

- We denote the whitened data by $\bar{x} := Vx$.

- Exercise: Show that, for any orthogonal matrix $A$, $\text{cov}(A\bar{x}) = I$.

- 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 (\( \tilde{x} = Vx \)) 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:

\[
\sum_{i=1}^n \log p(\tilde{x}^{(i)}; A) = \frac{1}{n} \sum_{i=1}^n \log p_z(A\tilde{x}^{(i)}) + \log |\det A| = 0
\]

Maximizing non-gaussianity via kurtosis (Related to fourth-moment \( E[y^4_j] \)). Gaussian distribution has kurtosis = 0.

Minimizing mutual information between the components of \( y = Ax \): see Hyvarinen et al. (2001) for more details!

Objectives to perform ICA

Most algorithms to perform ICA first whiten the data ($\tilde{x} = Vx$) 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:

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

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

See Hyvarinen et al. (2001) for more details!

Objectives to perform ICA

Most algorithms to perform ICA first whiten the data ($\tilde{x} = Vx$) 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:

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

  $$= 0$$

- **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 $y := A\tilde{x}$.

- **See Hyvarinen et al. (2001) for more details!**

Instead of leveraging higher-order statistics, can we leverage temporal correlations?
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 \( \text{cov}(z_t) \) does not depend on \( t \)
- Lagged covariance matrices \( \text{cov}(z_t, z_{t-\tau}) \) do not dependent on \( t \) (but can depend on \( \tau \))
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 \( \text{cov}(z_t) \) does not depend on \( t \)
  
  - Lagged covariance matrices \( \text{cov}(z_t, z_{t-\tau}) \) do not dependent on \( t \) (but can depend on \( \tau \))

- We assume the components are decorrelated. Formally \( \text{cov}(z_t) = I \) and \( \text{cov}(z_t, z_{t-\tau}) = D_\tau \), where \( D_\tau \) is diagonal.
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 \( \text{cov}(z_t) \) does not depend on \( t \)
- Lagged covariance matrices \( \text{cov}(z_t, z_{t-\tau}) \) do not dependent on \( t \) (but can depend on \( \tau \))

We assume the components are decorrelated. Formally \( \text{cov}(z_t) = I \) and \( \text{cov}(z_t, z_{t-\tau}) = D_\tau \), where \( D_\tau \) is diagonal.

\[ x_t = Wz_t \]
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 $\text{cov}(z_t)$ does not depend on $t$
- Lagged covariance matrices $\text{cov}(z_t, z_{t-\tau})$ do not dependent on $t$ (but can depend on $\tau$)

We assume the components are decorrelated. Formally $\text{cov}(z_t) = I$ and $\text{cov}(z_t, z_{t-\tau}) = D_\tau$, where $D_\tau$ is diagonal.

$x_t = Wz_t$

Note that $\text{cov}(x_t) = W\text{cov}(z_t)W^\top = WW^\top$
ICA via temporal dependencies (AMUSE algorithm, Tong et al., (1990))

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

- Start by whitening the data:

\[
cov(x_t) = WW^T = U\Lambda U^T
\]

\[
\bar{x}_t := \Lambda^{-1/2}U^T x_t = \Lambda^{-1/2}U^T W z_t
\]
ICA via temporal dependencies (AMUSE algorithm, Tong et al., (1990))

\[
\begin{align*}
\text{cov}(z_t) &= I \\
\text{cov}(z_t, z_{t-\tau}) &= D_\tau \quad \text{(diagonal)} \\
x_t &= Wz_t \\
\text{cov}(x_t) &= WW^\top
\end{align*}
\]

- Start by whitening the data:

\[
\text{cov}(x_t) = WW^\top = U\Lambda U^\top
\]

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

- 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))

\[
\begin{align*}
\text{cov}(z_t) &= I \\
\text{cov}(z_t, z_{t-\tau}) &= D_\tau \text{ (diagonal)} \\
x_t &= Wz_t \\
\text{cov}(x_t) &= WW^T
\end{align*}
\]

- Start by whitening the data:

\[
\text{cov}(x_t) = WW^T = U\Lambda U^T
\]

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

- 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:

\[
\bar{W}\bar{W}^T = \Lambda^{-1/2}U^T WW^T U\Lambda^{-1/2} = \Lambda^{-1/2}U^T U\Lambda U^T U\Lambda^{-1/2} = I
\]
ICA via temporal dependencies (AMUSE algorithm, Tong et al., (1990))

\[
\begin{align*}
\text{cov}(z_t) &= I \\
\text{cov}(z_t, z_{t-\tau}) &= D_{\tau} \text{ (diagonal)} \\
x_t &= Wz_t \\
\text{cov}(x_t) &= WW^T \\
\bar{x}_t := \Lambda^{-1/2}U^T x_t \text{ (Whitened } x_t) \\
\bar{x}_t &= \bar{W}z_t \\
\bar{W}\bar{W}^T &= I
\end{align*}
\]
\[ \text{cov}(z_t) = I \quad \text{cov}(z_t, z_{t-\tau}) = D_\tau \text{ (diagonal)} \quad x_t = Wz_t \quad \text{cov}(x_t) = WW^\top \]

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

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

\[ \text{cov}(\bar{x}_t, \bar{x}_{t-\tau}) = \mathbb{E}[\bar{x}_t\bar{x}_{t-\tau}^\top] \quad (18) \]
\[ = \mathbb{E}[\bar{W}z_tz_{t-\tau}^\top \bar{W}^\top] \quad (19) \]
\[ = \bar{W}\text{cov}(z_t, z_{t-\tau})\bar{W}^\top \quad (20) \]
\[ = \bar{W}D_\tau\bar{W}^\top \quad (21) \]
ICA via temporal dependencies (AMUSE algorithm, Tong et al., (1990))

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

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

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

\[
\begin{align*}
\text{cov}(\bar{x}_t, \bar{x}_{t-\tau}) &= \mathbb{E}[\bar{x}_t\bar{x}_{t-\tau}^\top] \\
&= \mathbb{E}[\bar{W}z_tz_{t-\tau}^\top \bar{W}^\top] \\
&= \bar{W}\text{cov}(z_t, z_{t-\tau})\bar{W}^\top \\
&= \bar{W}D_\tau\bar{W}^\top
\end{align*}
\]

- How cool! The matrix \( \bar{W} \) appears in an eigendecomposition of \( \text{cov}(\bar{x}_t, \bar{x}_{t-\tau}) \)!
ICA via temporal dependencies (AMUSE algorithm, Tong et al., (1990))

\[
\begin{align*}
\text{cov}(z_t) &= I \\
\text{cov}(z_t, z_{t-\tau}) &= D_\tau \text{ (diagonal)} \\
x_t &= Wz_t \\
\text{cov}(x_t) &= WW^T
\end{align*}
\]

\[
\bar{x}_t := \Lambda^{-1/2} U^T x_t \text{ (Whitened } x_t) \\
\bar{x}_t = \bar{W}z_t \\
\bar{W}\bar{W}^T = I
\]

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

\[
\begin{align*}
\text{cov}(\bar{x}_t, \bar{x}_{t-\tau}) &= \mathbb{E}[\bar{x}_t\bar{x}_{t-\tau}^T] \\
&= \mathbb{E}[\bar{W}z_t z_{t-\tau}^T \bar{W}^T] \\
&= \bar{W} \text{cov}(z_t, z_{t-\tau}) \bar{W}^T \\
&= \bar{W} D_\tau \bar{W}^T
\end{align*}
\]

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

- But is this decomposition unique up to permutation and rescaling? If the entries of \(D_\tau\) are all distinct, then yes! (Because each eigenspace is one-dimensional)

- This means we can estimate \(\bar{W}\) by diagonalizing \(\text{cov}(\bar{x}_t, \bar{x}_{t-\tau})\)
Practical consideration:

- In practice, the empirical $\text{cov}(\tilde{x}_t, \tilde{x}_{t-\tau})$ 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 $\text{cov}(\tilde{x}_t, \tilde{x}_{t-\tau})$ 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.

Back to initial motivation...

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

\[ Z = \begin{pmatrix} T \\ R \\ B \end{pmatrix} \]

\[ X = f(Z) \]
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...

Identifiability results for Nonlinear ICA (far from exhaustive list)

- **Leveraging contrastive learning and (diagonal) temporal dependencies**

- **Leveraging VAE’s and non-stationarity of the sources**

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