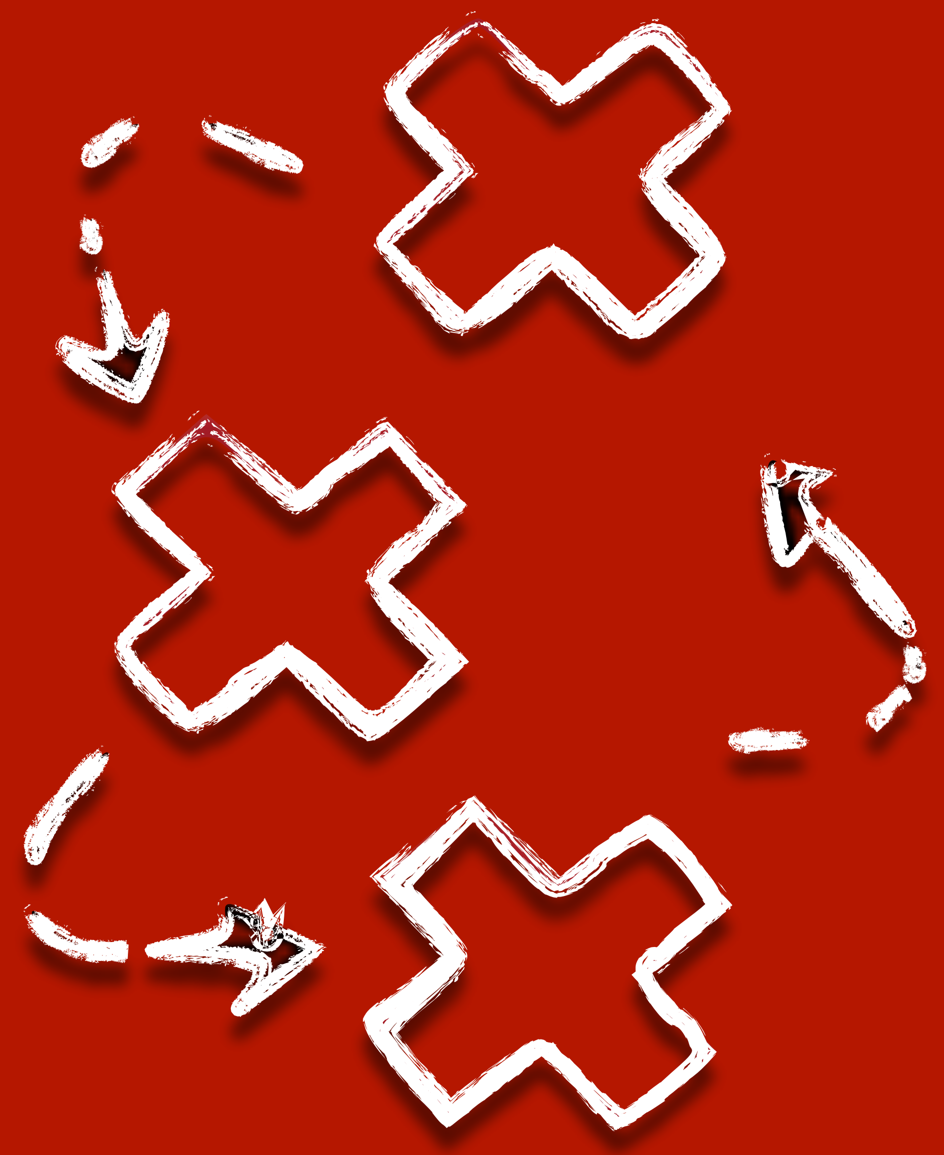


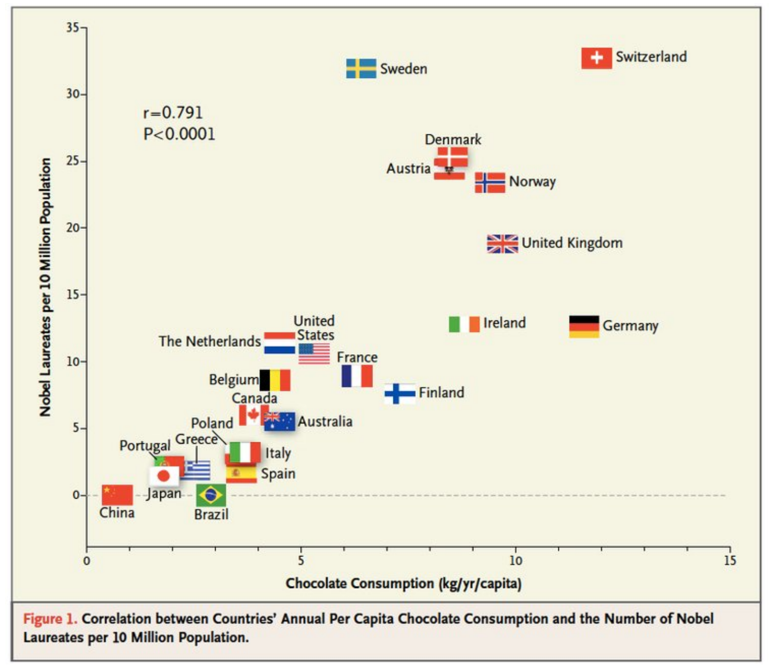
Deep learning 2: Causality & DL

2.2: Neural causal discovery

Lecturer: Sara Magliacane

UvA - Spring 2022

Causal discovery (structure learning) - simplest setting

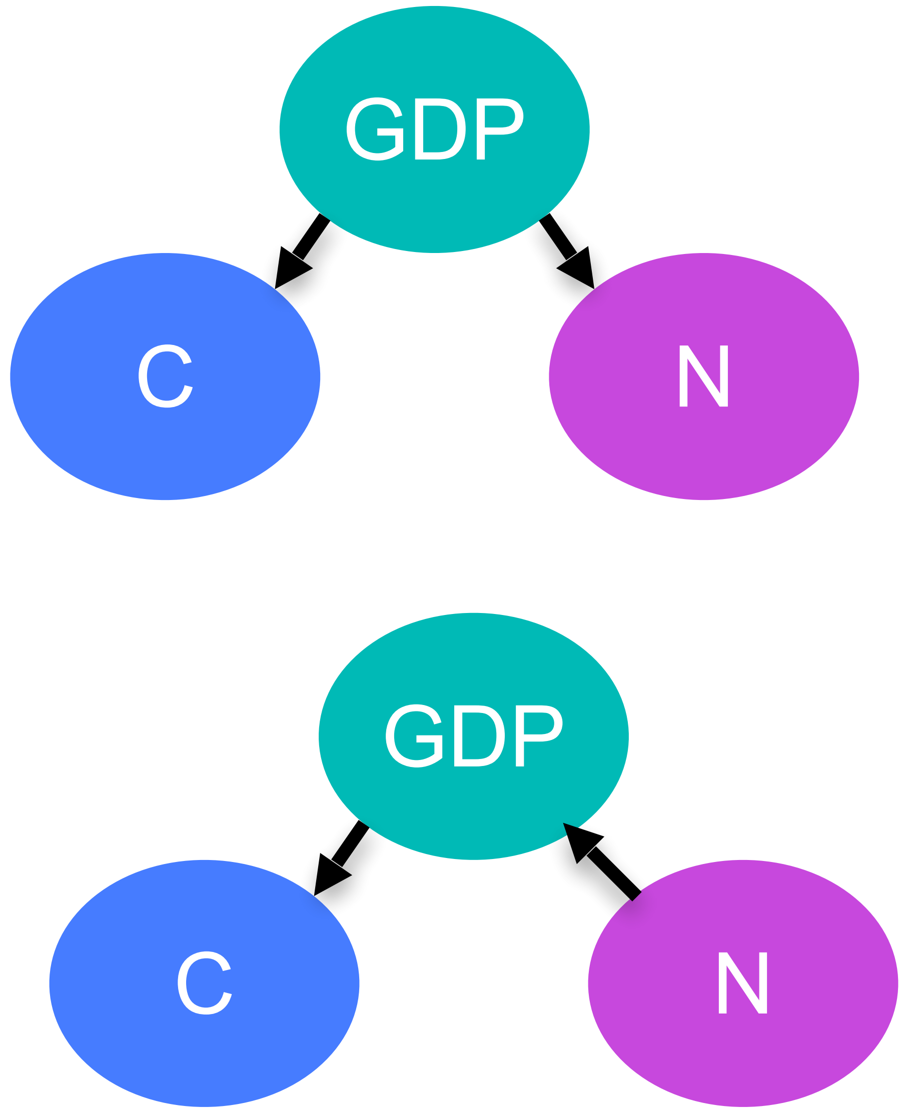


C	N	GDP
4.5	5	33k
12	30	86k
10	20	46k
….	…	…

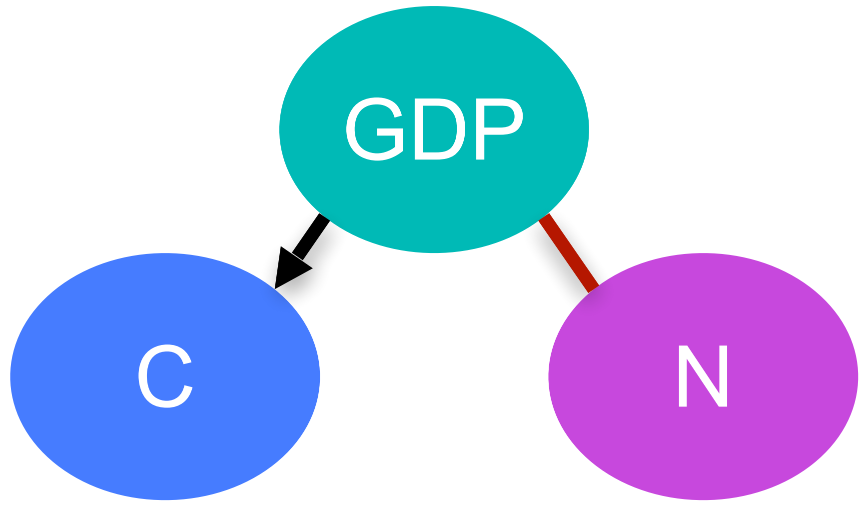
Observational data

$$C \nrightarrow GDP$$

[Optional] Background knowledge



Sets of graphs that fit the data and background knowledge



Summary graph

Causal discovery simplified overview

Constraint-based causal discovery

- Conditional independence tests
- Observational data
- Output: MEC
- SGS, PC, FCI

Score-based causal discovery

- Penalised likelihood
- Observational data
- Output: MEC
- GES, MMHC

Restricted models

- Nonlinear additive noise, Linear Non-Gaussianity
- Observational data
- Output: DAG
- RESIT, LINGAM

Interventional causal discovery / causal invariance

- Observational and Interventional data
- Output: parents of Y
- ICP, GIES, JCI

Common assumptions

- If P is **Markov and faithful** to G, then for any disjoint $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$:

$$\mathbf{A} \perp_G \mathbf{B} \,|\, \mathbf{C} \iff X_{\mathbf{A}} \perp\!\!\!\perp_P X_{\mathbf{B}} \,|\, X_{\mathbf{C}}$$

Common assumptions

- If P is **Markov and faithful** to G, then for any disjoint $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$:

$$\mathbf{A} \perp_G \mathbf{B} \,|\, \mathbf{C} \iff X_{\mathbf{A}} \perp\!\!\!\perp_P X_{\mathbf{B}} \,|\, X_{\mathbf{C}}$$

$$\begin{cases} x_1 = \varepsilon_1 \\ x_2 = 3 \cdot x_1 + \varepsilon_2 \\ x_3 = x_2 - 3x_1 + \varepsilon_3 \end{cases}$$

$\varepsilon_1, \varepsilon_2, \varepsilon_3 \sim N(0,1)$

$$x_3 = 3 \cdot x_1 + \varepsilon_2 - 3x_1 + \varepsilon_3$$
$$= \varepsilon_2 + \varepsilon_3$$

$X_3 \perp\!\!\!\perp X_1 \qquad X_3 \not\perp_d X_1$

CANCELLING PATHS

Common assumptions

- If P is **Markov and faithful** to G, then for any disjoint $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$:

$$\mathbf{A} \perp_G \mathbf{B} \,|\, \mathbf{C} \iff X_{\mathbf{A}} \perp\!\!\!\perp_P X_{\mathbf{B}} \,|\, X_{\mathbf{C}}$$

- **Causal sufficiency** - no latent confounders (common causes), no selection bias

Common assumptions

- If P is **Markov and faithful** to G, then for any disjoint $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$:

$$\mathbf{A} \perp_G \mathbf{B} \,|\, \mathbf{C} \iff X_{\mathbf{A}} \perp\!\!\!\perp_P X_{\mathbf{B}} \,|\, X_{\mathbf{C}}$$

- **Causal sufficiency** - no latent confounders (common causes), no selection bias

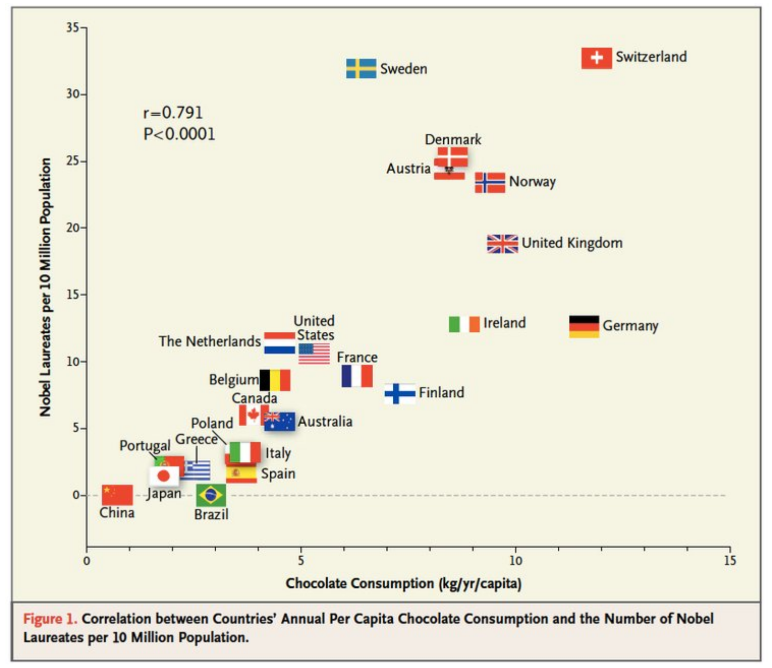
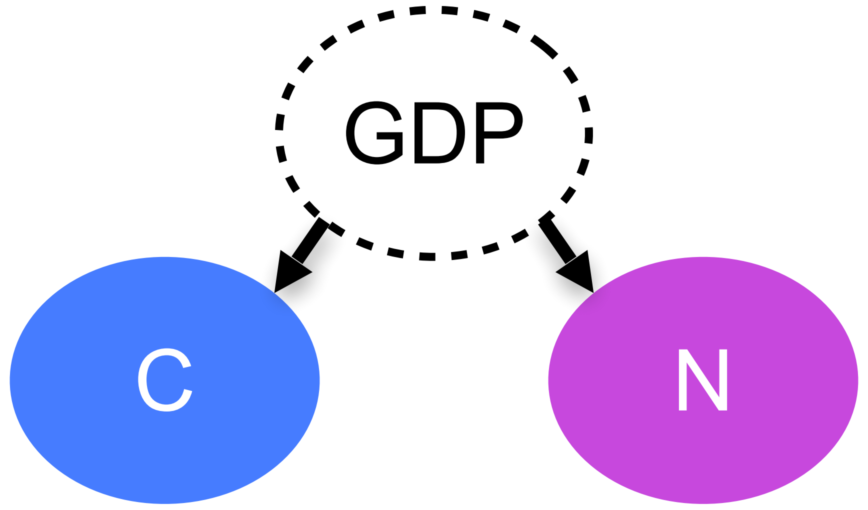


Figure 1. Correlation between Countries' Annual Per Capita Chocolate Consumption and the Number of Nobel Laureates per 10 Million Population.



Common assumptions

- If P is **Markov and faithful** to G, then for any disjoint $\mathbf{A}, \mathbf{B}, \mathbf{C} \subseteq \mathbf{V}$:

$$\mathbf{A} \perp_G \mathbf{B} \,|\, \mathbf{C} \iff X_{\mathbf{A}} \perp\!\!\!\perp_P X_{\mathbf{B}} \,|\, X_{\mathbf{C}}$$

- **Causal sufficiency** - no latent confounders (common causes), no selection bias

- **Acyclicity** - the underlying graph is acyclic

- Cycles + causal insufficiency: sigma separation, Joint Causal Inference

Causal discovery simplified overview

Constraint-based causal discovery

- Conditional independence tests

- Observational data
- Output: MEC

- SGS, PC, FCI

Score-based causal discovery

- Penalised likelihood

- Observational data
- Output: MEC

- GES, MMHC

Restricted models

- Nonlinear additive noise, Linear Non-Gaussianity

- Observational data
- Output: DAG

- RESIT, LINGAM

Interventional causal discovery / causal invariance

- Observational and Interventional data
- Output: parents of Y

- ICP, GIES, JCI

Constraint-based causal discovery

- If P is **Markov and faithful** to G, $\mathbf{A} \perp_G \mathbf{B} \,|\, \mathbf{C} \iff X_{\mathbf{A}} \perp\!\!\!\perp_P X_{\mathbf{B}} \,|\, X_{\mathbf{C}}$
- **In a nutshell:** we perform a set of conditional independence tests on the data and use them to constrain the possible graphs using d-separation

Constraint-based causal discovery

- If P is **Markov and faithful** to G, $\mathbf{A} \perp_G \mathbf{B} \,|\, \mathbf{C} \iff X_\mathbf{A} \perp\!\!\!\perp_P X_\mathbf{B} \,|\, X_\mathbf{C}$

- **In a nutshell:** we perform a set of conditional independence tests on the data and use them to constrain the possible graphs using d-separation

- In general, we can narrow down the possible graphs only up to their **Markov equivalence class (MEC)**
  - Sets of graphs with the same d-separation statements

Constraint-based causal discovery

- If P is **Markov and faithful** to G, $\mathbf{A} \perp_G \mathbf{B} \mid \mathbf{C} \iff X_{\mathbf{A}} \perp\!\!\!\perp_P X_{\mathbf{B}} \mid X_{\mathbf{C}}$

- **In a nutshell:** we perform a set of conditional independence tests on the data and use them to constrain the possible graphs using d-separation

- In general, we can narrow down the possible graphs only up to their **Markov equivalence class (MEC)**
  - Sets of graphs with the same d-separation statements

- We can represent all the graphs in a MEC with a **summary graph**

Markov equivalence example

$X \not\perp\!\!\!\perp Y \quad X \perp\!\!\!\perp Y \mid Z$

$X - Z - Y$

$X \rightarrow Z \rightarrow Y$

$X \not\perp_d Y$ ✓

$X \perp_d Y \mid Z$ ✓

$X \leftarrow Z \rightarrow Y$

$X \not\perp_d Y$ ✓

$X \perp_d Y \mid Z$ ✓

$X \leftarrow Z \leftarrow Y$

$X \not\perp_d Y$ ✓

$X \perp_d Y \mid Z)$ ✓

$X \rightarrow Z \leftarrow Y$

$X \perp_d Y$ ×

$X \not\perp_d Y \mid Z$ ×

Markov equivalence class and CPDAGs

- We can represent the skeleton and the orientations (edge marks) all DAGs in a MEC with a **Complete Partially Directed Acyclic Graph (CPDAG):**
  - $i \to j$ if all DAGs in the MEC have $i \to j$
  - $i - j$ if some DAGs in the MEC have $i \to j$ and others have $j \to i$

$X \to Z \to Y$ $\quad$ $X \not\perp_d Y$ ✓ $\quad$ $X \perp_d Y \mid Z$ ✓

$X \leftarrow Z \to Y$ $\quad$ $X \not\perp_d Y$ ✓ $\quad$ $X \perp_d Y \mid Z$ ✓

$X \leftarrow Z \leftarrow Y$ $\quad$ $X \not\perp_d Y$ ✓ $\quad$ $X \perp_d Y \mid Z)$ ✓

$X - Z - Y$

SGS algorithm (Spirtes, Glymour, Scheines)

- Assuming P is Markov and faithful to an unknown graph G

- We can estimate a CPDAG from samples of P in three steps:
  1. Determine the **skeleton**

     if $\exists \mathbf{S}$ s.t. $X_i \perp\!\!\!\perp X_j \mid X_{\mathbf{S}}$, then $i \not- j$

  2. Determine the **v-structures**

     if $i - j, j - k$, $i \not- k$ and $\nexists \mathbf{S}$ s.t. $X_i \perp\!\!\!\perp X_k \mid X_j \cup X_{\mathbf{S}}$, then $i \rightarrow j \leftarrow k$

  3. Direct as many remaining edges as possible using "acyclicity" and "no new v-structures"

PC algorithm (Peter Spirtes, Clark Glymour)

- Assuming P is Markov and faithful to an unknown graph G

- We can estimate a CPDAG from samples of P in three steps:
  1. Determine the **skeleton in an optimised way**
     - Use the nodes that are adjacent, $\mathrm{Adj}(i)$ or $\mathrm{Adj}(j)$ in U at a given iteration (*superset of the parents*)
  2. Determine the **v-structures**
  3. Direct as many remaining edges as possible

     - https://www.researchgate.net/publication/242448131_Causation_Prediction_and_Search

Causal discovery simplified overview

Constraint-based causal discovery

- Conditional independence tests

- Observational data
- Output: MEC

- SGS, PC, FCI

Score-based causal discovery

- Penalised likelihood

- Observational data
- Output: MEC

- GES, MMHC

Restricted models

- Nonlinear additive noise, Linear Non-Gaussianity

- Observational data
- Output: DAG

- RESIT, LINGAM

Interventional causal discovery / causal invariance

- Observational and Interventional data
- Output: parents of Y

- ICP, GIES, JCI

Score-based causal discovery

- **Score-based causal discovery:** find the graph that maximises a **score** $S(G, D)$ (fit of graph G on data D)

Score-based causal discovery

- **Score-based causal discovery:** find the graph that maximises a **score** $S(G, D)$ (fit of graph G on data D)

- Typically we use **BIC (Bayesian information criterion)**

number of edges + const.

$$BIC(D, G) := 2 \cdot \log p(D \mid G, \theta^{MLE}) - \log(n) \cdot \#\text{parameters}$$

number of data points

Score-based causal discovery

- **Score-based causal discovery:** find the graph that maximises a **score** $S(G, D)$ (fit of graph G on data D)

- Typically we use **BIC (Bayesian information criterion)**

number of edges + const.

$$BIC(D, G) := 2 \cdot \log\, p(D \mid G, \theta^{MLE}) - \log(n) \cdot \#\text{parameters}$$

number of data points

  - **Score equivalence:** all DAGs in a MEC get the same score
  - **Decomposable:** we can decompose the score as the sum of the contributions for each variable and its parents
  - **Local consistency**

Number of DAGs with n nodes ?

A003024 as a simple table

n	a(n)
0	1
1	1
2	3
3	25
4	543
5	29281
6	3781503
7	1138779265
8	783702329343
9	1213442454842881
10	4175098976430598143
11	31603459396418917607425
12	521939651343829405020504063
13	18676600744432035186664816926721
14	1439428141044398334941790719839535103

https://oeis.org/A003024

Greedy Equivalence Search (GES)

- Optimise **Bayesian Information Criterion (BIC)** **greedily** (take best scoring neighbour iteratively)

- Reduce search space by **searching over CPDAGs** instead of DAGs

- BIC is score-equivalent, so DAGs in same Markov equivalence class (so represented by same CPDAG) have the same score *(so you can pick any)*

- It can be shown that searching over CPDAGs with local consistency allows us to find the **global optimum** *(in large sample limit)*

Greedy Equivalence Search (GES)

1. Start with empty CPDAG
2. Add edges one by one until local maxima in BIC
3. Remove edges one by one until local maxima in BIC

decomposability helps us
to recompute only one factor

Phase 1 neighbours ε^+:

Given a a starting equivalence class ε, another class ε' is in the neighbours ε^+ if there exists a DAG $G \in \varepsilon$, such that adding an edge to G results in $G' \in \varepsilon'$

Phase 2 neighbours ε^-: same with removing an edge

Differentiable causal discovery

- Observational data - linear - NOTEARS
- Observational data - nonlinear - DAG-GNN
- Observational + interventional data - ENCO

Differentiable score-based methods: NOTEARS

Non-combinatorial **O**ptimization via **T**race **E**xponential and **A**ugmented lag**R**angian for **S**tructure learning

- **Linear SCM for $\mathbf{X} = (X_1, \ldots, X_d)$:**

$$X_i = w_i^T \cdot \mathbf{X} + Z_i \text{ for } i = 1,\ldots, d$$

$$W = [w_1 \,|\, w_2 \,|\, \ldots \,|\, w_d]$$

https://arxiv.org/pdf/1803.01422.pdf

Differentiable score-based methods: NOTEARS

Non-combinatorial **O**ptimization via **T**race **E**xponential and **A**ugmented lag**R**angian for **S**tructure learning

- **Linear SCM for** $\mathbf{X} = (X_1, \ldots, X_d)$:

$$X_i = w_i^T \cdot \mathbf{X} + Z_i \text{ for } i = 1, \ldots, d$$

$$W = [w_1 \,|\, w_2 \,|\, \ldots \,|\, w_d]$$

$$\min_W \ \ell(W; X) \quad \text{s.t.} \ G(W) \in DAG \qquad \overset{?}{\Longleftrightarrow} \qquad \min_W \ \ell(W; X) \quad \text{s.t.} \ h(W) = 0$$

(combinatorial 😱) (smooth 😎)

(possibly with a simple gradient)

https://arxiv.org/pdf/1803.01422.pdf

Differentiable score-based methods: NOTEARS

Non-combinatorial **O**ptimization via **T**race **E**xponential and **A**ugmented lag**R**angian for **S**tructure learning

- **Linear SCM for $\mathbf{X} = (X_1, \ldots, X_d)$:**

$$X_i = w_i^T \cdot \mathbf{X} + Z_i \text{ for } i = 1,\ldots,d$$

$$W = [w_1 \,|\, w_2 \,|\, \ldots \,|\, w_d]$$

$$\frac{1}{2n}\|\mathbf{X} - \mathbf{X}W\|_F^2 + \lambda\|W\|_1.$$

$$\min_W \ \ell(W; X) \quad \text{s.t. } G(W) \in \mathsf{DAG} \qquad \overset{?}{\Longleftrightarrow} \qquad \min_W \ \ell(W; X) \quad \text{s.t. } h(W) = 0$$

(possibly with a simple gradient)

(combinatorial 😱) (smooth 😎)

https://arxiv.org/pdf/1803.01422.pdf

Differentiable score-based methods: NOTEARS

- **Linear SCM for $\mathbf{X} = (X_1, \ldots, X_d)$:**

$$X_i = w_i^T \cdot \mathbf{X} + Z_i \text{ for } i = 1,\ldots,d;\ W = [w_1 \,|\, w_2 \,|\, \ldots \,|\, w_d]$$

$$\min_W \ \ell(W;X) \quad \text{s.t.} \quad G(W) \in DAG \qquad \overset{?}{\Longleftrightarrow} \qquad \min_W \ \ell(W;X) \quad \text{s.t.} \quad h(W) = 0$$

(combinatorial 😱) (smooth 😎)

(but non convex)

$$h(W) = \operatorname{tr}\left(e^{W \circ W}\right) - d = 0$$

Extension to nonlinear case: https://arxiv.org/abs/1909.13189

Differentiable score-based methods: DAG-GNN

$$X = A^T X + Z \qquad X = (I - A^T)^{-1} Z$$

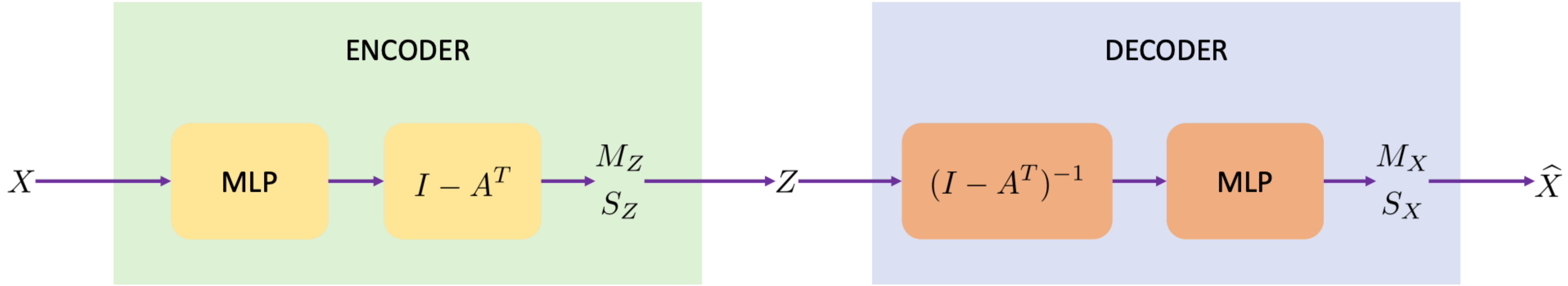
$$X = f_2((I - A^T)^{-1} f_1(Z)) \qquad Z = f_4((I - A^T) f_3(X)),$$

https://arxiv.org/pdf/1904.10098.pdf

Differentiable score-based methods: DAG-GNN

$$X = A^T X + Z \qquad X = (I - A^T)^{-1} Z$$

$$X = f_2((I - A^T)^{-1} f_1(Z)) \qquad Z = f_4((I - A^T) f_3(X)),$$

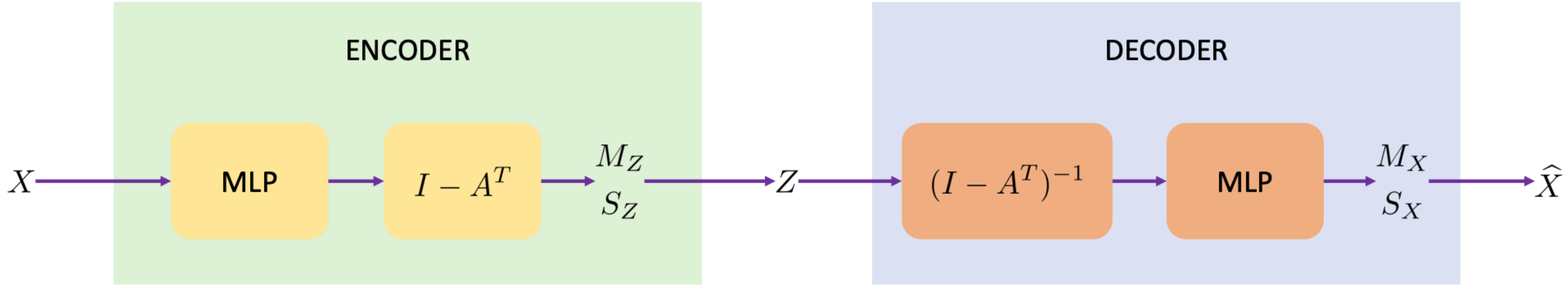


https://arxiv.org/pdf/1904.10098.pdf

Differentiable score-based methods: DAG-GNN

$$X = A^T X + Z \qquad X = (I - A^T)^{-1} Z$$

$$X = f_2((I - A^T)^{-1} f_1(Z)) \qquad Z = f_4((I - A^T) f_3(X)),$$



Max ELBO + equality constraint: for given $\alpha > 0$, $\mathrm{tr}[(\mathrm{I} + \alpha \cdot \mathrm{A} \odot \mathrm{A})^{\mathrm{d}}] - d = 0$

https://arxiv.org/pdf/1904.10098.pdf

Differentiable score-based methods: ENCO

Efficient Neural Causal Discovery without Acyclicity Constraints

- **Assumption:** Single node interventional data (perfect) for all variables

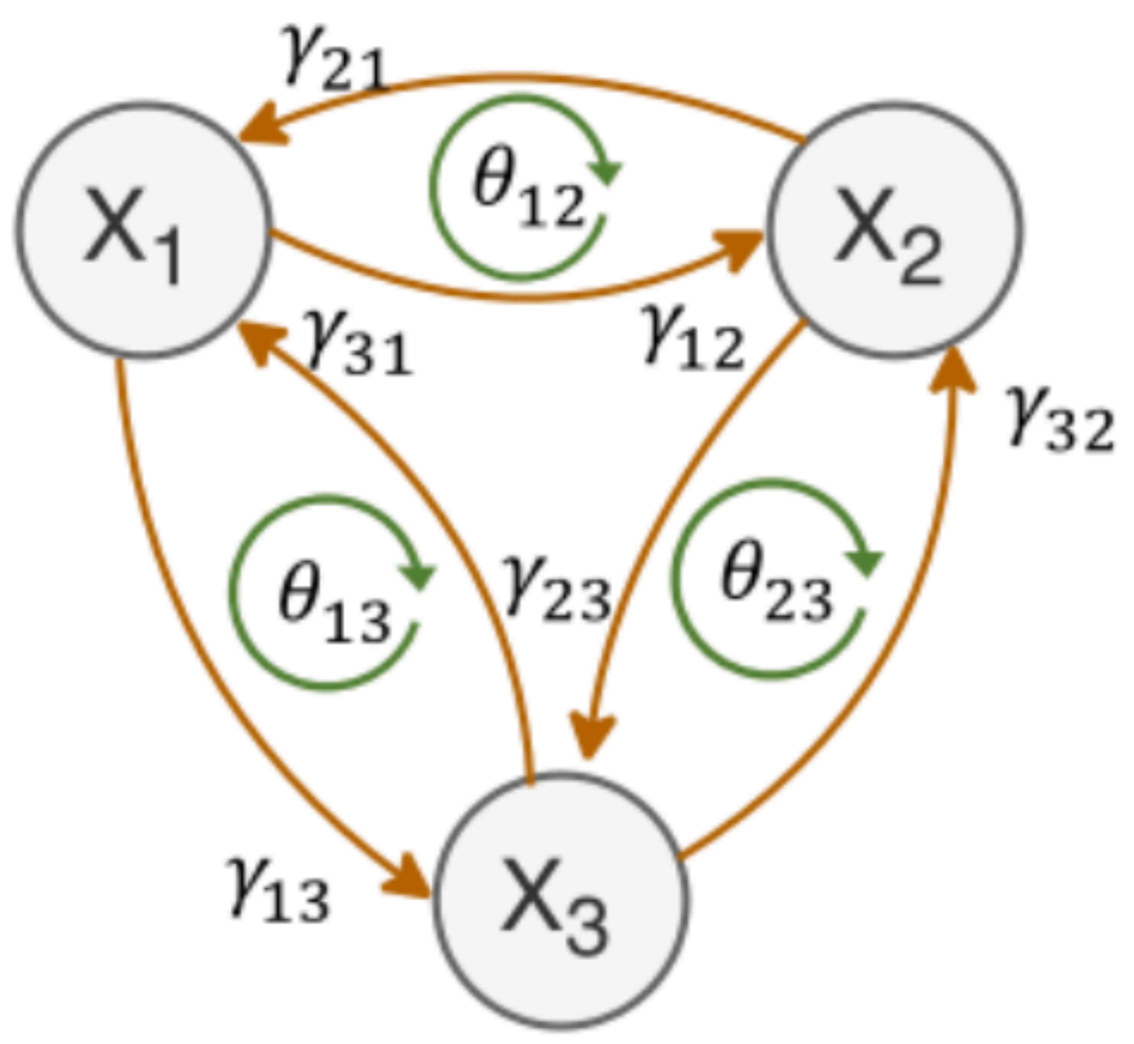
Differentiable score-based methods: ENCO

Efficient Neural Causal Discovery without Acyclicity Constraints

- **Assumption:** Single node interventional data (perfect) for all variables
- Central idea: learn distributions $p(X_1 \mid \dots)$ from observational data, test generalization to interventional data
- Parametrize graph with edge existence and orientation parameters
  - Probability of an edge: $\sigma\left(\gamma_{ij}\right) \cdot \sigma\left(\theta_{ij}\right)$, with $\theta_{ij} = -\theta_{ji}$
- Benefits of two-variable parameterisation:

  ⇒ More control over gradient updates

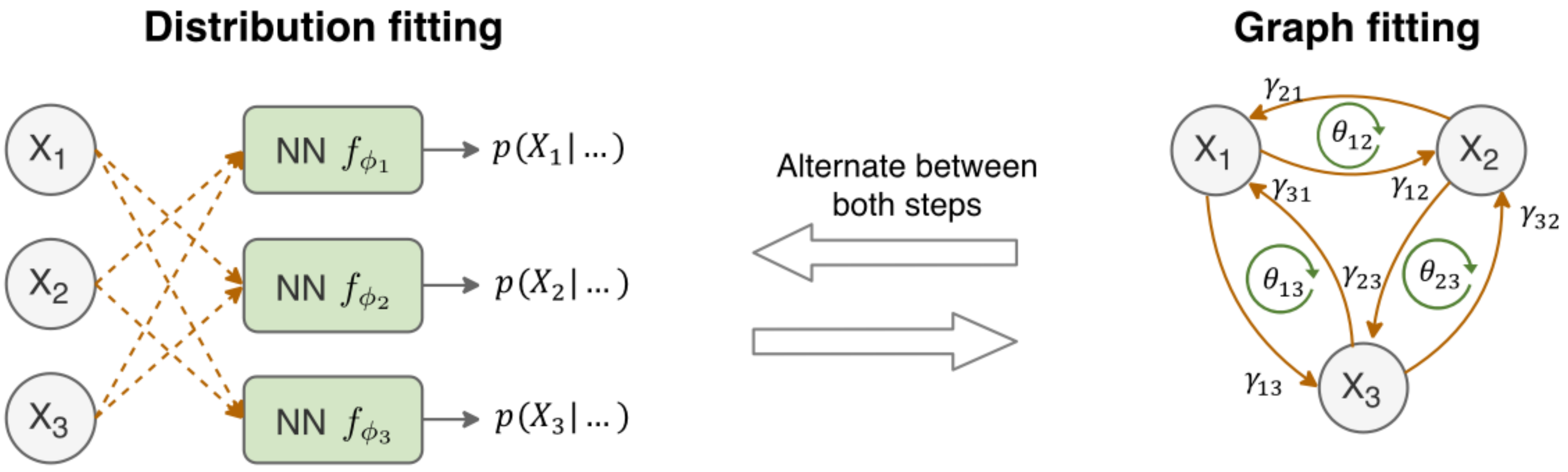
  ⇒ No constraint or regularization for acyclicity needed!



https://arxiv.org/abs/2107.10483, slides by Phillip Lippe

Differentiable score-based methods: ENCO

Efficient Neural Causal Discovery without Acyclicity Constraints



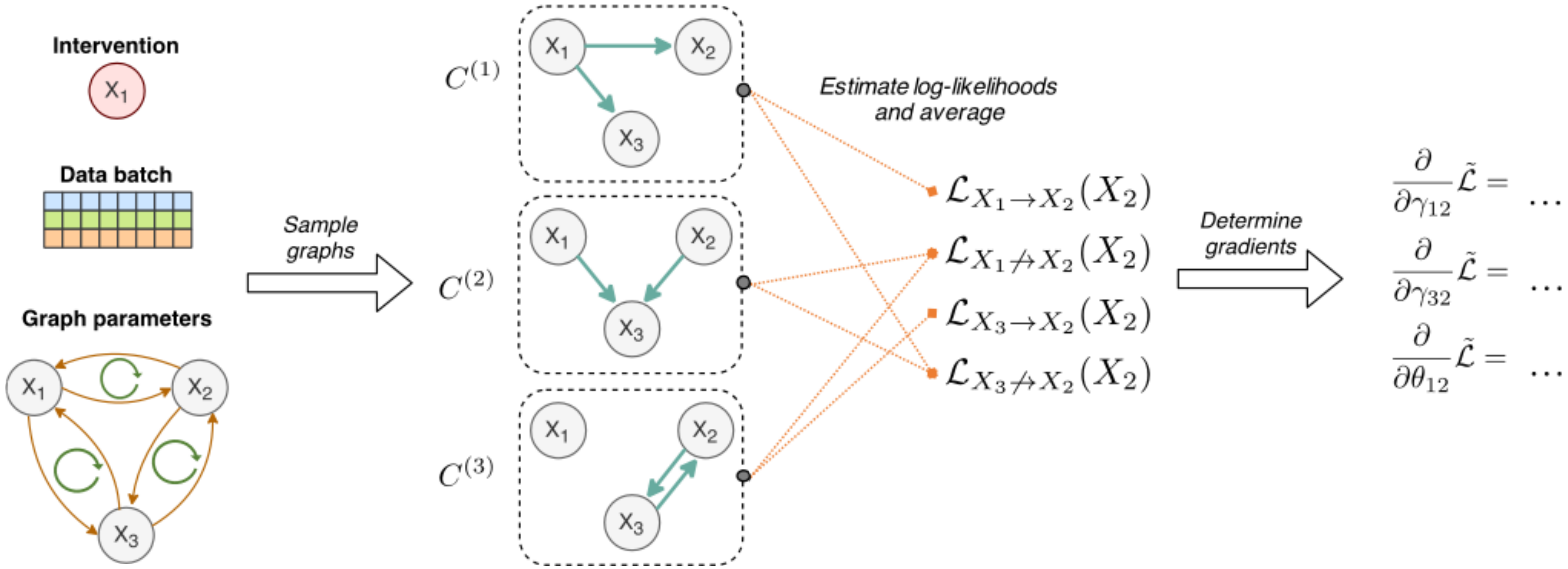
Learn neural networks by fitting conditional distributions on observational data

Learn edge and orientation parameters based on fitted distributions

https://arxiv.org/abs/2107.10483, slides by Phillip Lippe

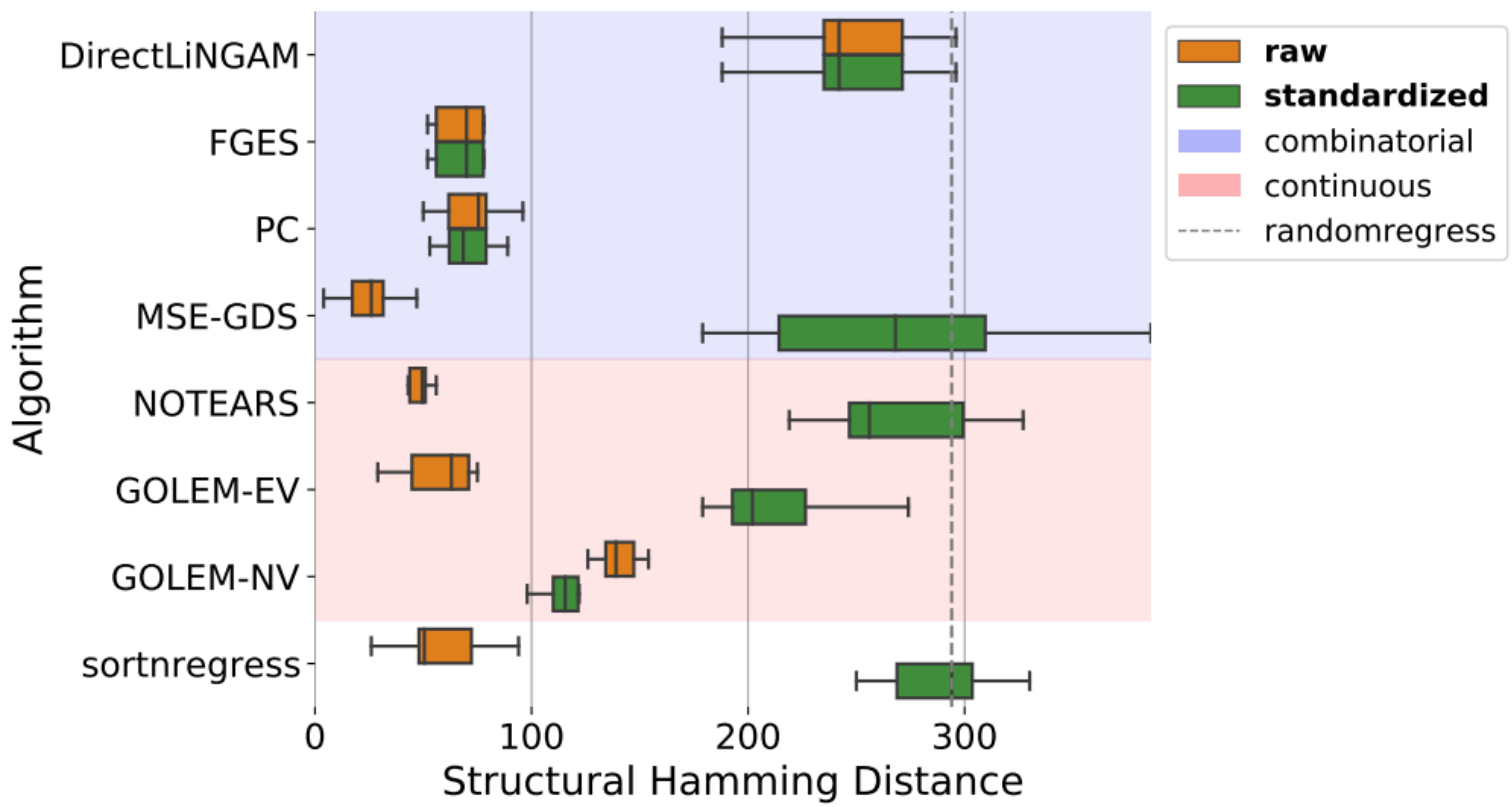
Differentiable score-based methods: ENCO

Efficient Neural Causal Discovery without Acyclicity Constraints



https://arxiv.org/abs/2107.10483, slides by Phillip Lippe

Side note: Varsortability



https://arxiv.org/pdf/2102.13647.pdf