Probabilistic and Causal Structure Learning

Generative and Deep Learning (GDL)

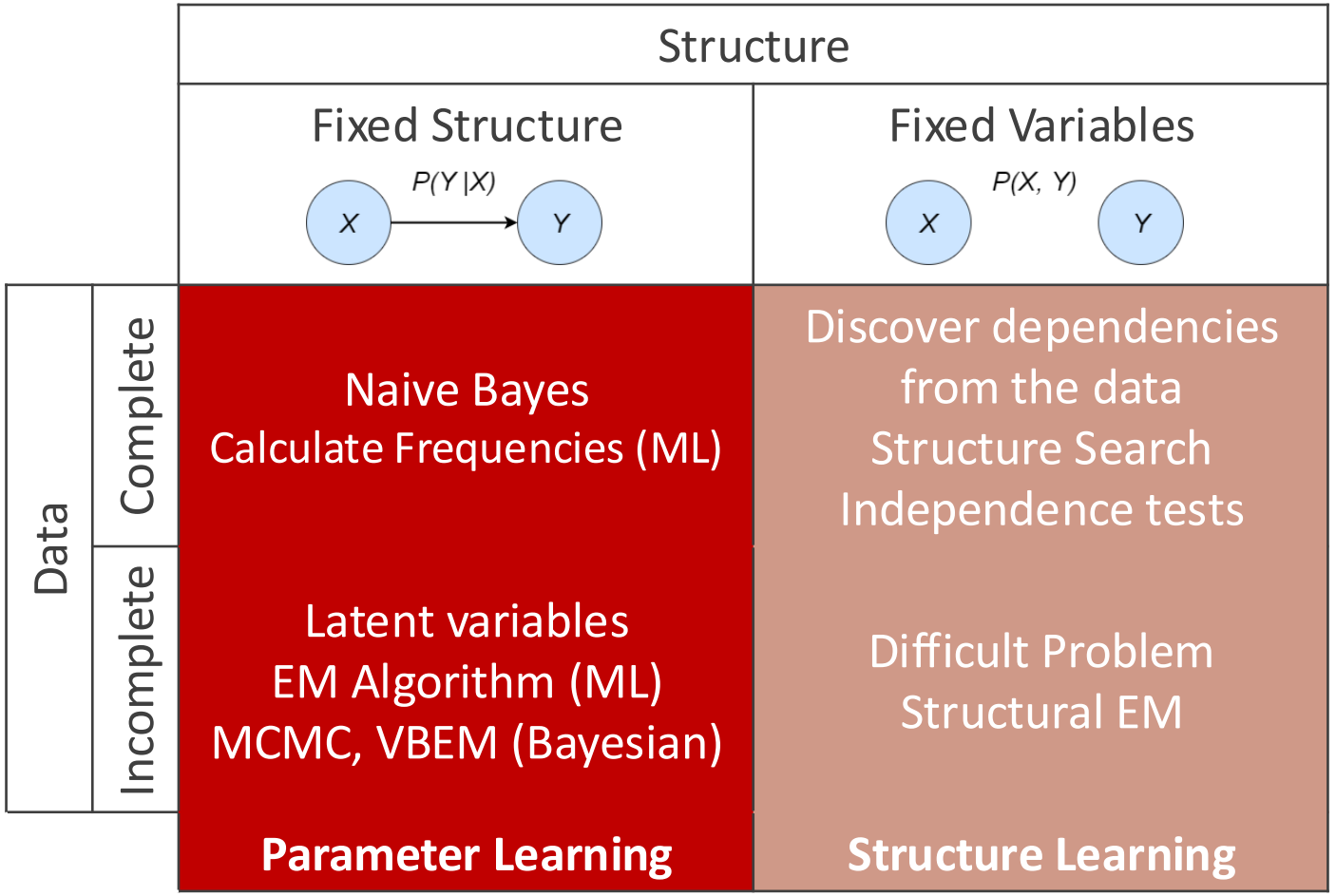
Riccardo Massidda (riccardo.massidda@di.unipi.it)
Davide Bacciu (davide.bacciu@unipi.it)

UNIVERSITÀ DI PISA

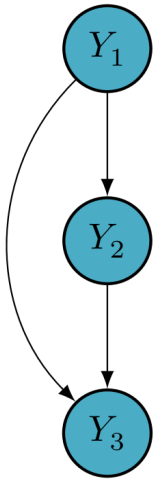
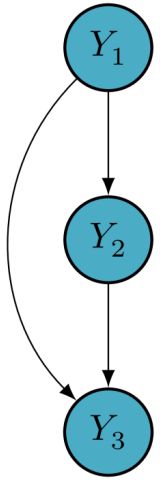
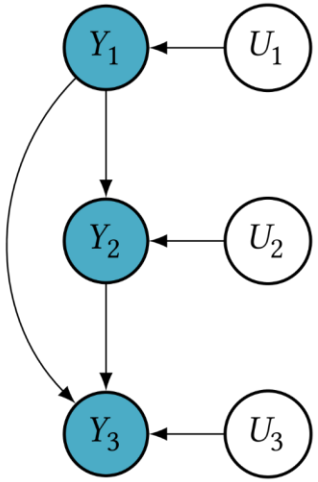
Probabilistic and Causal Learning

- Bayesian Networks (Tuesday 24th)
- d-separation, Markov blankets (Wednesday 25th)
- Graphical Causal Models (Thursday 26th)
- Probabilistic and Causal Structure Learning (Tuesday 3rd, **today**!)
  - Constraint-Based Methods (**PC**, FCI)
  - Score-Based Methods (GES)
  - Parametric Assumptions (Additive Noise Models)

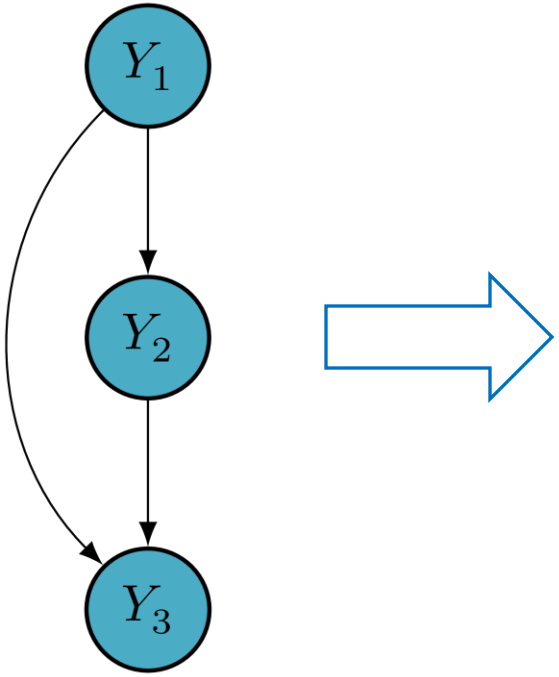
Learning from Structured Data



Probabilistic and Causal Models (a short recap)

model:	**Bayesian Network**	**Causal Bayesian Network**	**Structural Causal Model**
			
encodes:	Conditional probabilities	Conditional probabilities w/ causal order	Deterministic causal relations w/ exogenous noise
queries:	$P(Y \mid X)$	$P(Y \mid \mathrm{do}(X))$	$P(Y' \mid \mathrm{do}(X'), X, Y)$
	conditional/observational	interventional	counterfactual

Data Generating Process



Y_1	Y_2	Y_3
0.192	-0.123	0.456
0.789	-0.456	0.123
⋮	⋮	⋮
0.321	-0.654	0.789

Observations are given for a set of **fixed random variables** whose network structure is not specified.

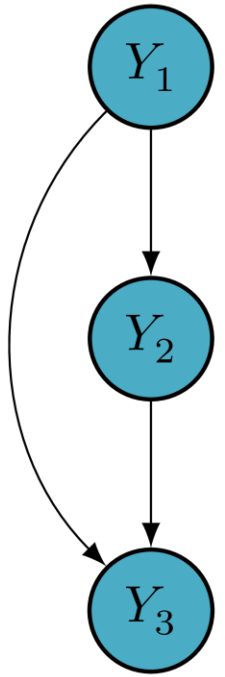
Structure Learning

Y_1	Y_2	Y_3
0.192	-0.123	0.456
0.789	-0.456	0.123
⋮	⋮	⋮
0.321	-0.654	0.789

Structure Learning Algorithm

Assumptions, e.g,

- Markov Property
- Faithfulness
- Causal Sufficiency
- Acyclicity
- …



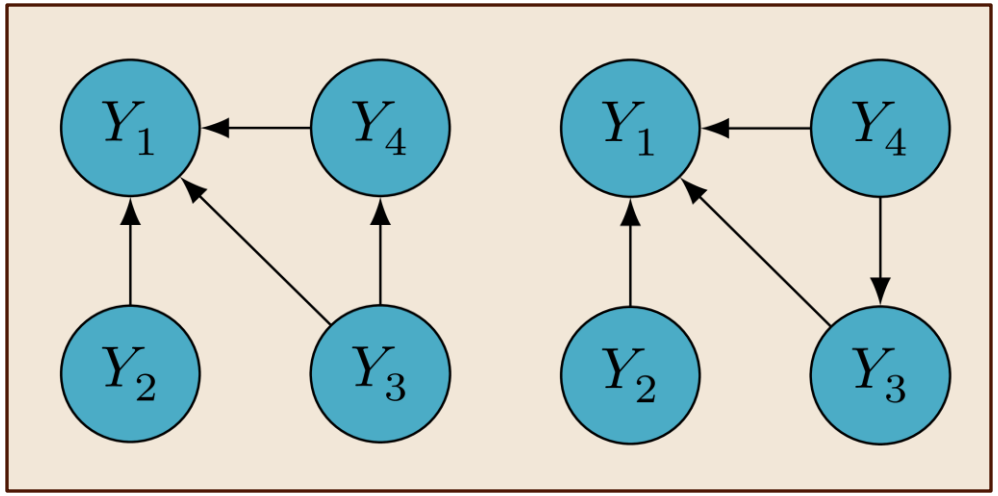
Structure Finding Approaches

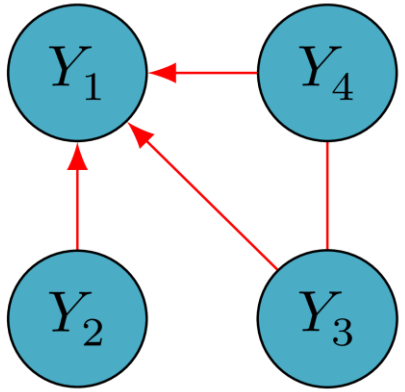
- Constraint Based
  - Use **tests of conditional independence**
  - Constrain the network
- Search and Score-based
  - **Model selection** approach
  - Search in the space of the graphs
- Parametric Identifiability

Markov Equivalence Class

- A **Markov Equivalence Class** (MEC) is a set of DAGs encoding the same set of conditional independences.
- Two DAGs are **Markov equivalent** if and only if they have the same **skeleton** and the same set of colliders (**v-structures**).
- We encode MECs using partially completed DAGs (**CPDAGs**).

Markov equivalence class





CPDAG representing the MEC

Constraint-Based Methods

- We can reconstruct the Markov Equivalence Class by iteratively performing **conditional independence testing** (χ2-test, KCI-test, Fisher z-test, G-square test, …).
- The Spirtes, Glymour, and Scheines (**SGS**) and the Peter and Clark (**PC**) algorithms are the fundamental constraint-based discovery methods.
- These algorithms assume the **Markov Condition**, **Faithfulness**, and **Causal Sufficiency**.

SGS Algorithm: Skeleton

- Two variables X and Y are adjacent in the **skeleton** if they are **always** conditionally **dependent**, i.e., there exists no separating set without X and Y.

Require: Dataset of observed variables $\mathcal{D}$ over variables V
Ensure: Markov Equivalence Class as CPDAG $\mathcal{G}$

$\mathcal{G} \leftarrow$ Fully connected CPDAG over V.
for all Pairs (X, Y) in V **do**
  for all $Z \subseteq V \setminus \{X, Y\}$ **do**
    if $X \perp Y \mid Z$ **then**
      Prune $X - Y$ in $\mathcal{G}$.
    end if
  end for
end for

SGS Algorithm: v-structures

- If two variables X and Y are **not** adjacent in the skeleton and there exists a third variable W that
  - It is adjacent to both X and Y, and
  - It is not a member of any separating set:
- We found a **collider**!

for all Triplets (X, W, Y) s.t.
  $X - W - Y$ in $\mathcal{G}$, and
  $X - Y$ **not** in $\mathcal{G}$ **do**
  if W is not in **any** separating set of X and Y **then**
    Orient $X \rightarrow W \leftarrow Y$ as a collider.
  end if
end for

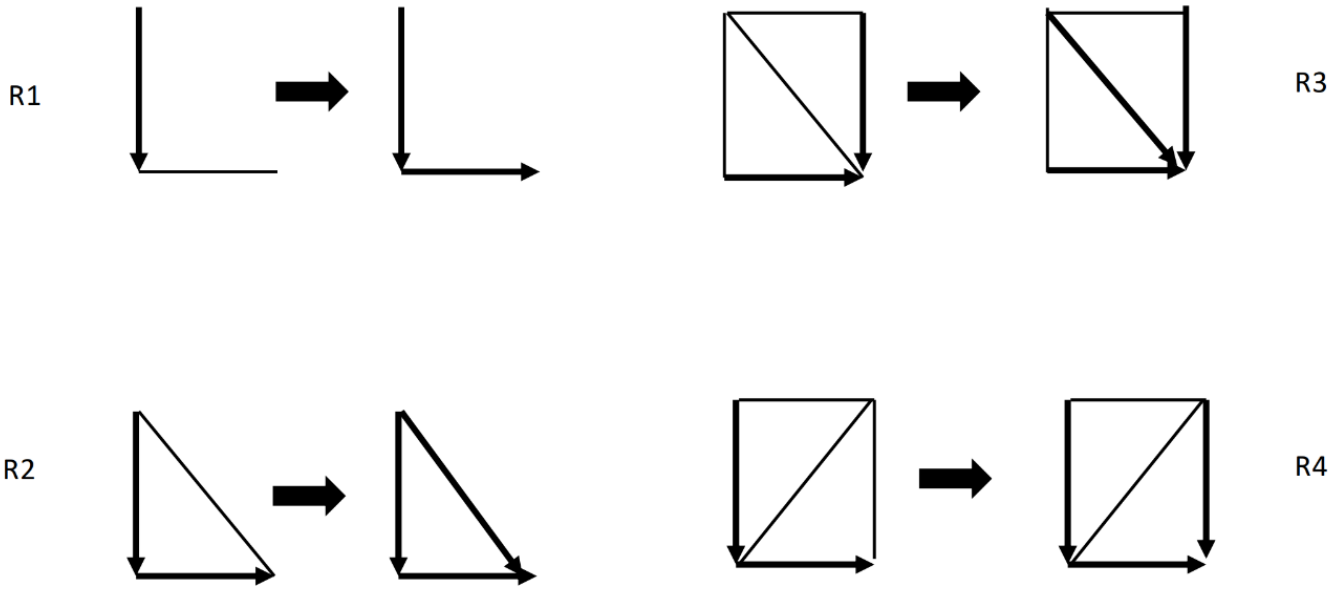
SGS Algorithm: Additional Orientations

- By **avoiding** the introduction of **new colliders** and **cycles**, we can further orient other edges.
- Still, we might have some **unoriented edges** and return a Completed Partially Directed Acyclic Graph (**CPDAG**).

while Further edge orientations are possible **do**
  for all Triplets (X, Y, Z) s.t.
    $X \to Y$ and $Y - Z$ in $\mathcal{G}$ **do**
    Orient $Y \to Z$.
  end for
  for all Pairs (X, Y) in $\mathcal{G}$ **do**
    if X is an ancestor of Y and $X - Y$ **then**
      Orient $X \to Y$.
    end if
  end for
end while
return The CPDAG of the Markov Equivalence Class.

Meek Rules

- The orientations of the SGS algorithm (R1, R2) are generalized by the **Meek** rules to **avoid** indirectly introducing **new v-structures**.
- They still do **not guarantee a DAG** but decrease the MEC.



PC Algorithm: Skeleton

- Instead of checking all possible separating sets, as in SGS, the **PC** algorithm considers **separating sets** of **increasing size**.
- Same worst case of SGS, much **better on average**!

$\mathcal{G} \leftarrow$ Fully connected CPDAG over V.
$K = 0$
while $K \leq |V|$ **do**
  for all Pairs (X, Y) in $\mathcal{G}$ **do**
    $A = \{Z \mid X - Z \text{ in } \mathcal{G}\} \setminus \{Y\}$
    for all $Z \subseteq A, |Z| \leq K$ **do**
      if $X \perp Y \mid Z$ **then**
        Prune $X - Y$ in $\mathcal{G}$.
      end if
    end for
  end for
  $K \leftarrow K + 1$
end while

Testing Strategy

- Choice of the **testing order** is fundamental to avoid **super-exponential complexity**.
- **Level-wise Testing**
  - Tests are performed on conditioning sets of increasing size (PC Algorithm)
- **Node-wise Testing**
  - Tests are performed one edge at the time, exhausting all possible conditioning sets (TPDA Algorithm)
- Another common heuristic is to choose from the neighborhood of the nodes incident on the edge.
- If only a particular relation X→Y is of interest, we can discover a subgraph containing X, Y, and their adjustment set (**Local Causal Discovery**)

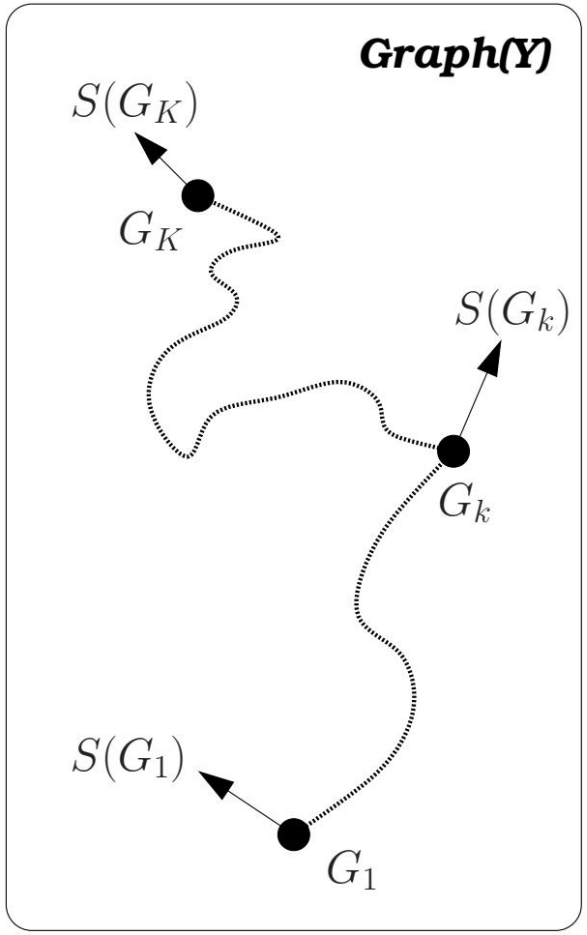
Constraint-Based Methods

- Is the CPDAG produced by a constraint-based method **enough**?
- **Probabilistic** Queries $\mathrm{P}(\mathrm{Y}|\mathrm{X})$

  ⇒ We can take *any* graph in the MEC and use it! 🥳
- **Interventional** $\mathrm{P}(\mathrm{Y}|\mathrm{do}(\mathrm{X}))$ or **counterfactual** $\mathrm{P}(\mathrm{Y}|\mathrm{do}(\mathrm{X}), \mathrm{Y}')$ queries

  ⇒ We need further knowledge to orient undirected edges. 💀
- Given the graph, we need to **choose** the **distribution** families, for BNs/CBNs, or the **mechanisms**, for SCMs, and **learn the parameters**.
- The **more** the **variables**, the **more CI tests** needed. 💸

Search & Score



- Search the space *Graph*(**Y**) of graphs G_k that can be built on the random variables $\boldsymbol{Y} = Y_1, \dots, Y_N$
- Score each structure by $S(G_k)$
- Return the highest scoring graph G^*
- Two fundamental aspects
  - Scoring function
  - Search strategy

Scoring Function

- Fundamental properties
  - **Consistency**: same score for graphs in the same equivalence class
  - **Decomposability**: can be locally computed
- Approaches
  - **Information theoretic**: based on data likelihood plus some model-complexity penalization terms (AIC, BIC, MDL, …)
  - **Bayesian**: score the structures using a graph posterior (likelihood + proper prior choice)

Bayesian Information Criterion

$$S(\mathcal{G}; \mathbf{D}) = \log p_{\hat{\theta}}(\mathbf{D} \mid \mathcal{G}) - \frac{k}{2} \log n$$

The Bayesian Information Criterion (**BIC**) is a **consistent** and **local** score on a dataset with n data points and k parameters.

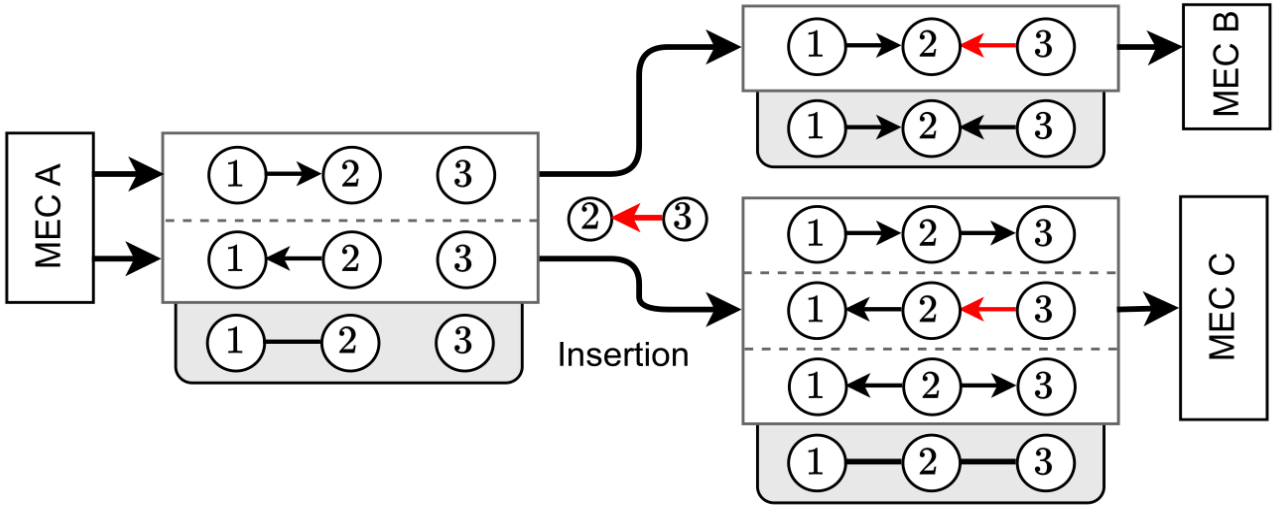
Search Strategy

- Finding maximal scoring structures is NP complete (Chickering, 2002)
- Constrain search strategy
  - Starting from a candidate structure modify iteratively by local operations (edge/node addition or deletion)
  - Each operation has a cost
  - Cost optimization problem: greedy hill-climbing, simulated annealing, …
- Constrain search space
  - Known node order – Can reduce the search space to the parents of each node (Markov Blanket)
  - Search in the space of structure equivalence classes (GES algorithm)
  - Search in the space of node orderings (Friedman and Koller, 2003)

GES Algorithm

The **Greedy Equivalent Search** builds on three main operations:

1. **Insertion** to add an edge leading to a different MEC
2. **Deletion** to remove an edge leading to a different MEC
3. **Reversal** to flip an edge leading to a different MEC



GES Algorithm

- GES greedily tries all possible insertions, all possible deletions, and all possible reversals.
- Using the **BIC** score, GES **ensures to find the MEC** using only insertions and deletions.
- …but in practice reversal helps with limited samples.

Algorithm 1: Greedy Equivalence Search (GES)

Input: Data $\mathbf{D} \in \mathbb{R}^{n \times d}$, score function S
Define: $\delta_{\mathbf{D},M}(O) = S(\text{Apply}(O, M); \mathbf{D}) - S(M; \mathbf{D})$
Output: MEC of G^*
$M \leftarrow \{([d], \varnothing)\}$ // Empty graph's MEC
$\mathcal{I} \leftarrow$ get all insertions valid for M
while $|\mathcal{I}| > 0$ **do**
  $O^* \leftarrow \arg\max_{I \in \mathcal{I}} \{\delta_{\mathbf{D},M}(I)\}$ // Get best insertion
  if $\delta_{\mathbf{D},M}(O^*) \leq 0$ **then break**
  $M \leftarrow \text{Apply}(O^*, M)$ // Apply best insertion
  $\mathcal{I} \leftarrow$ get all insertions valid for M
$\mathcal{D} \leftarrow$ get all deletions valid for M
while $|\mathcal{D}| > 0$ **do**
  $O^* \leftarrow \arg\max_{D \in \mathcal{D}} \{\delta_{\mathbf{D},M}(D)\}$ // Get best deletion
  if $\delta_{\mathbf{D},M}(O^*) \leq 0$ **then break**
  $M \leftarrow \text{Apply}(O^*, M)$ // Apply best deletion
  $\mathcal{D} \leftarrow$ get all deletions valid for M
/* (Optional) 3rd phase like above but with reversals */
return M

Hybrid Models

- Multi-stage algorithms combining previous approaches
- Independence tests to find a sub-optimal skeleton (good starting point)
- Search and score starting from the skeleton
  - Skeleton refinement
  - Edge orientation
- Max-Min Hill Climbing (MMHC) model
  - Optimized constraint-based approach to reconstruct the skeleton (Max-Min Parents and Children)
  - Use the candidate parents in the skeleton to run a search and score approach

Why structure learning only returns the MEC?

Pearl & Geiger 1988; Meek 1995

For heteroskedastic **linear Gaussian** and **multinomial** causal relations, an algorithm that identifies the **Markov Equivalence Class** of the true model is **complete**.

adapted from Prof. Eberhardt talk on the state of Causal Discovery

Summary of Identifiability Results for SCMs

Type of structural assignment		Noise	DAG identif.
General SCM:	$Y_j := f_j(Y_{\mathbf{PA}_j}, U_j)$	–	×
Nonlinear ANM:	$Y_j := f_j(Y_{\mathbf{PA}_j}) + U_j$	–	✓
Nonlinear CAM:	$Y_j := \sum_{i \in \mathbf{PA}_j} f_{ij}(Y_i) + U_j$	–	✓
Linear	$Y_j := \sum_{i \in \mathbf{PA}_j} w_{ij} Y_i + U_j$	Heteroskedastic Gaussian	×
		Homoskedastic Gaussian	✓
		non-Gaussian	✓

table adapted from **"Elements of Causal Inference"** by Peters et al.

Take Home Messages

- The main problem in structure learning is the **lack of identifiability**.
- **Assumptions** are needed to **restrict the search space** of DAGs.
- Directed models are powerful and interpretable tools when combined with **prior knowledge**.