Causal Models: Representation and Learning

INTELLIGENT SYSTEMS FOR PATTERN RECOGNITION (ISPR)

RICCARDO MASSIDDA, DAVIDE BACCIU – DIPARTIMENTO DI INFORMATICA - UNIVERSITA' DI PISA

RICCARDO.MASSIDDA@DI.UNIPI.IT DAVIDE.BACCIU@UNIPI.IT

Probabilistic and Causal Learning

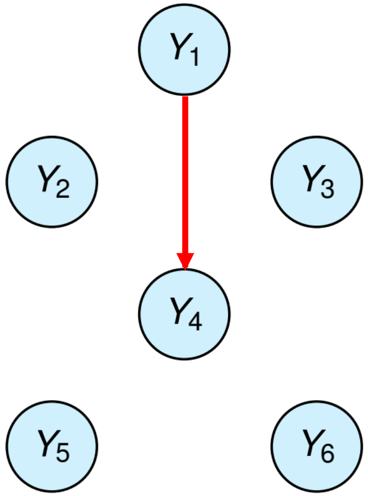
- Bayesian Networks (Tuesday 4th)
- Bayesian Networks (Thursday 6th)
- Graphical Causal Models (Tuesday 11th)
- Structure Learning and Causal Discovery (Wednesday 12th, **today**!)
  - Constraint-Based Methods (**PC**, FCI)
  - Score-Based Methods (GES)
  - Parametric Assumptions (Additive Noise Models)

Learning with Bayesian Networks

		Structure	
		Fixed Structure $P(Y \mid X)$ $X \rightarrow Y$	Fixed Variables $P(X, Y)$ $X \quad Y$
Data	Complete	Naive Bayes Calculate Frequencies (ML)	Discover dependencies from the data Structure Search Independence tests
	Incomplete	Latent variables EM Algorithm (ML) MCMC, VBEM (Bayesian)	Difficult Problem Structural EM
		Parameter Learning	**Structure Learning**

The Structure Learning Problem

Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
1	2	1	0	3	4
4	0	0	0	1	2
...	...	...	...	...	...
...	...	...	...	...	...
0	0	1	3	2	1



- Observations are given for a set of fixed random variables
- Network structure is not specified
  - Determine which arcs exist in the network (causal relationships ⇒ causal discovery)
  - Compute Bayesian network parameters (conditional probability tables) or SCM parameters (structural functions)
- Determining the graph entails
  - Deciding on arc presence
  - Directing edges

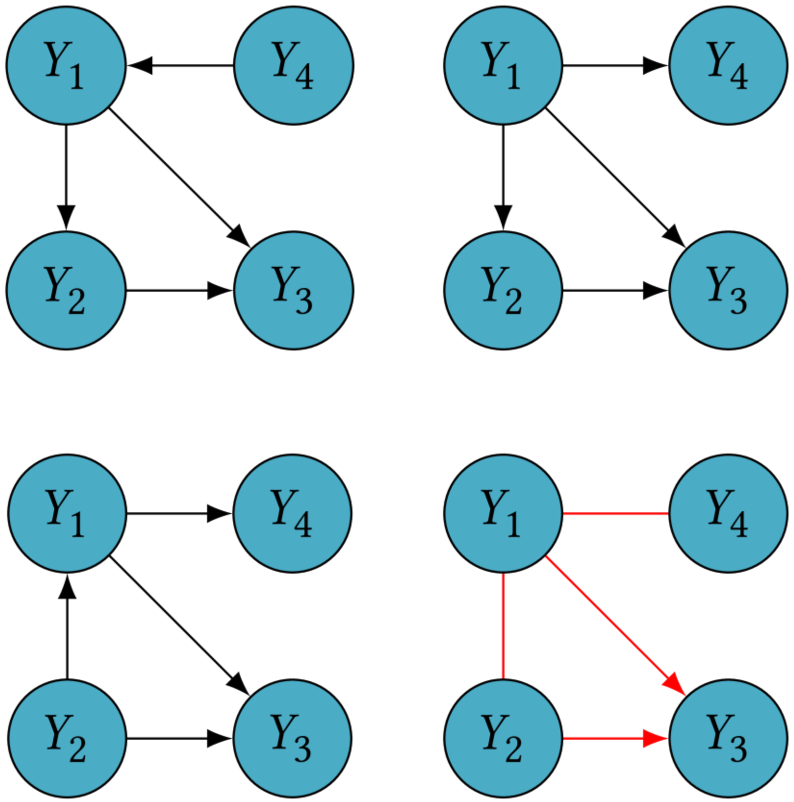
Structure Finding Approaches

- Constraint Based
  - Use **tests of conditional independence**
  - Constrain the network
- Search and Score
  - **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**).



Constraint-Based Methods

Constraint-based methods require:

- **Faithfulness**, i.e., all conditional independences are represented from the distribution are represented in the graph.
- **Causal Sufficiency**, i.e., all confounders are observed and there is not selection bias.

Constraint-Based Methods

- We can reconstruct the Markov Equivalence Class by iteratively performing **conditional independence testing** ($\chi 2$-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.

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 $\mathcal{G}$

```
1: 𝒢 ← Fully connected CPDAG over V.
2: for all Pairs (X, Y) in V do
3:     for all Z ⊆ V \ {X, Y} do
4:         if X ⊥ Y | Z then
5:             Prune X − Y in 𝒢.
6:         end if
7:     end for
8: end for
```

UNIVERSITÀ DI PISA

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

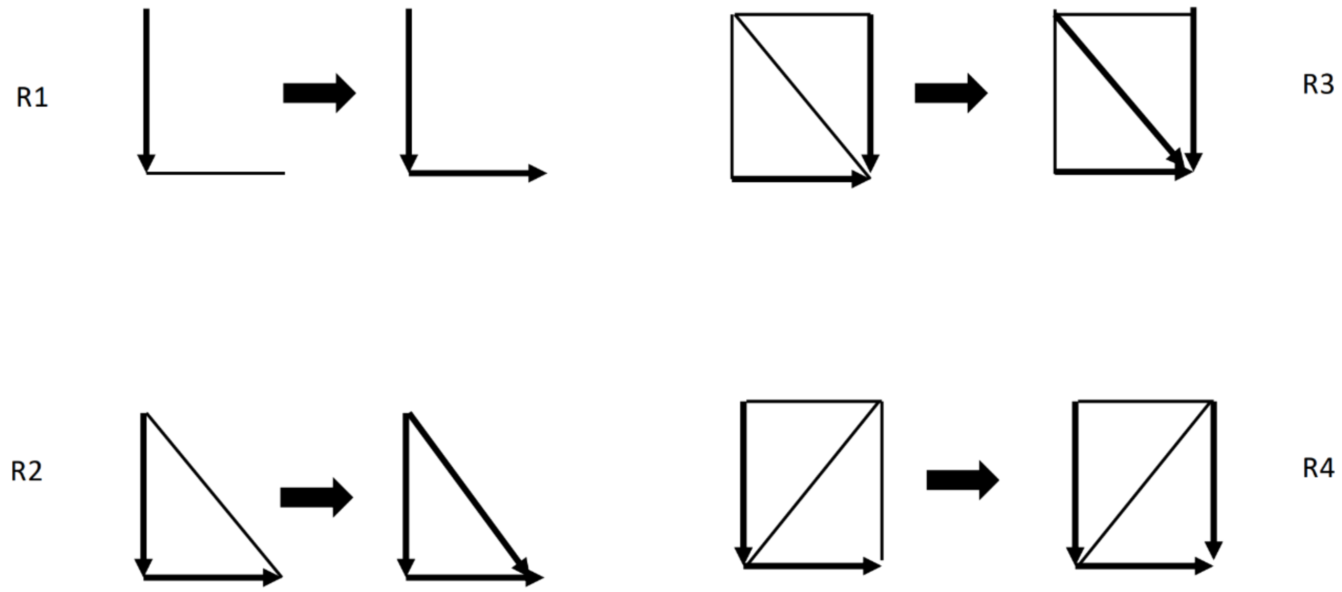
SGS Algorithm: Additional Orientations

- 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.
        $X \rightarrow Y$ and $Y - Z$ in $\mathcal{G}$ **do**
        Orient $Y \rightarrow Z$.
    end for
    for all Pairs (X, Y) in $\mathcal{G}$ **do**
        if X is an ancestor of Y and $X - Y$ **then**
            Orient $X \rightarrow Y$.
        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.



Testing Strategy

- Choice of the testing order is fundamental for avoiding a super-exponential complexity
- Level-wise testing
  - Tests $I(X_i, X_j | Z)$ are performed in order of increasing size of the conditioning set Z (starting from empty Z)
  - PC algorithm (Spirtes, 1995)
- Node-wise testing
  - Tests are performed on a single edge at the time, exhausting independence checks on all conditioning variables
  - TPDA Algorithm
- Nodes that enter Z are chosen in the neighborhood of X_i and X_j

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

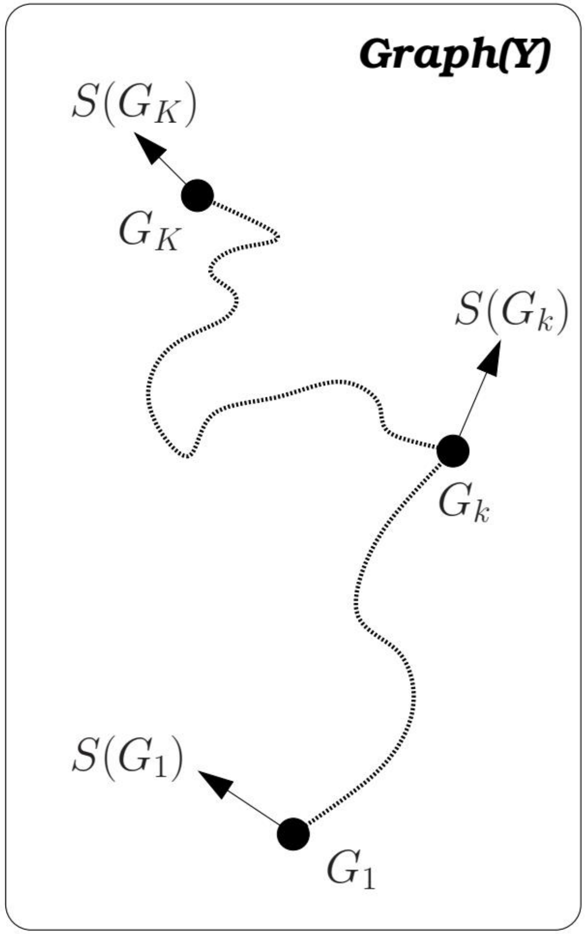
Constraint-Based Methods

- Is the CPDAG produced by a constraint-based method **enough**?

- **Probabilistic** Queries $P(Y|X)$
  $\Rightarrow$ We can take *any* graph in the MEC and use it! 🥳
- **Interventional** $P(Y|do(X))$ or **counterfactual** $P(Y|do(X), Y')$ queries
  $\Rightarrow$ 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**.

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)

$$\log P(D|G) \approx \sum_{D} \sum_{X} \log \tilde{P}(x|\boldsymbol{pa}(x)) + \log P(G)$$

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)

Hybrid Models

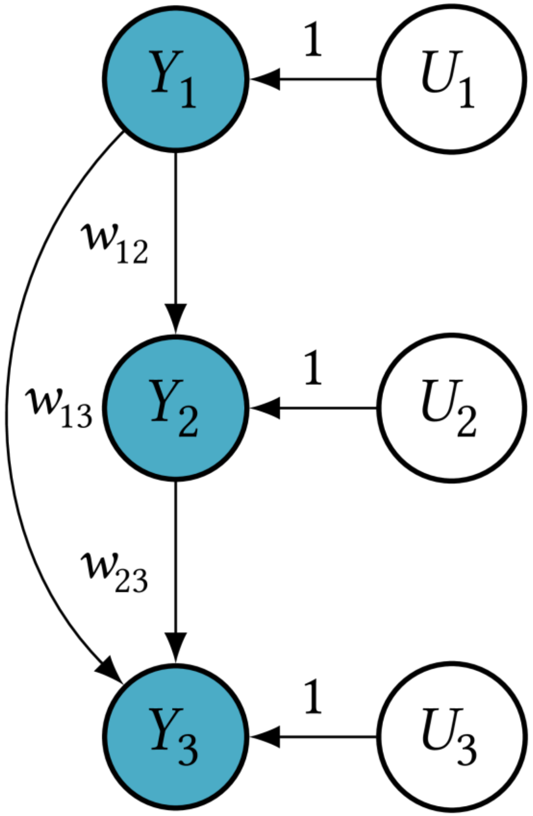
- Multi-stage algorithms combining previous approaches
- Independence tests to find a sub-optimal skeleton (good starting point)
- Search and score starting from the skeleton
  - Skeleton refinement
  - Edge orientation
- Max-Min Hill Climbing (MMHC) model
  - Optimized constraint-based approach to reconstruct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

Linear Additive Noise Model

A Linear Additive Noise Model (**ANM**) is a structural causal model where the **functional mechanisms** are **linear** and the noise is additive.

Formally, given a matrix $W \in \mathbb{R}^{n \times n}$,

$$Y_j = \sum_{Y_i \in \mathrm{pa}(Y_j)} w_{ij} Y_i + U_j$$



Identifiability of Linear ANMs

The **identifiability** of a linear Additive Noise Model from data strongly depends on the **distribution** of the **noise** terms.

- Gaussian w/ Equal Variance ⇒ Yes!
  - Not so common in real-world applications.
- Gaussian w/o Equal Variance ⇒ No!
- Non-Gaussian Noise ⇒ Yes!
  - ICA-LiNGAM, DirectLingam, PairwiseLingam...

Common Identifiability Results for SCMs

Type of structural assignment		Condition on funct.	DAG identif.
(General) SCM:	$X_j := f_j(X_{\mathbf{PA}_j}, N_j)$	—	✗
ANM:	$X_j := f_j(X_{\mathbf{PA}_j}) + N_j$	nonlinear	✓
CAM:	$X_j := \sum_{k \in \mathbf{PA}_j} f_{jk}(X_k) + N_j$	nonlinear	✓
Linear Gaussian:	$X_j := \sum_{k \in \mathbf{PA}_j} \beta_{jk} X_k + N_j$	linear	✗
Lin. G., eq. error var.:	$X_j := \sum_{k \in \mathbf{PA}_j} \beta_{jk} X_k + N_j$	linear	✓

From **"Elements of Causal Inference"** by Peters et al.

Take Home Messages

- Directed graphical models
  - Represent asymmetric (causal) relationships between RV and conditional probabilities in compact way
  - Difficult to assess conditional independence (v-structures)
  - Ok for prior knowledge and interpretation
- Undirected graphical models
  - Represent bi-directional relationships (e.g. constraints)
  - Factorization in terms of generic potential functions (not probabilities)
  - Easy to assess conditional independence, but difficult to interpret
  - Serious computational issues due to normalization factor
- Structure learning to infer multivariate causal relationships from data