Week 9

Wentao Gao

Outline

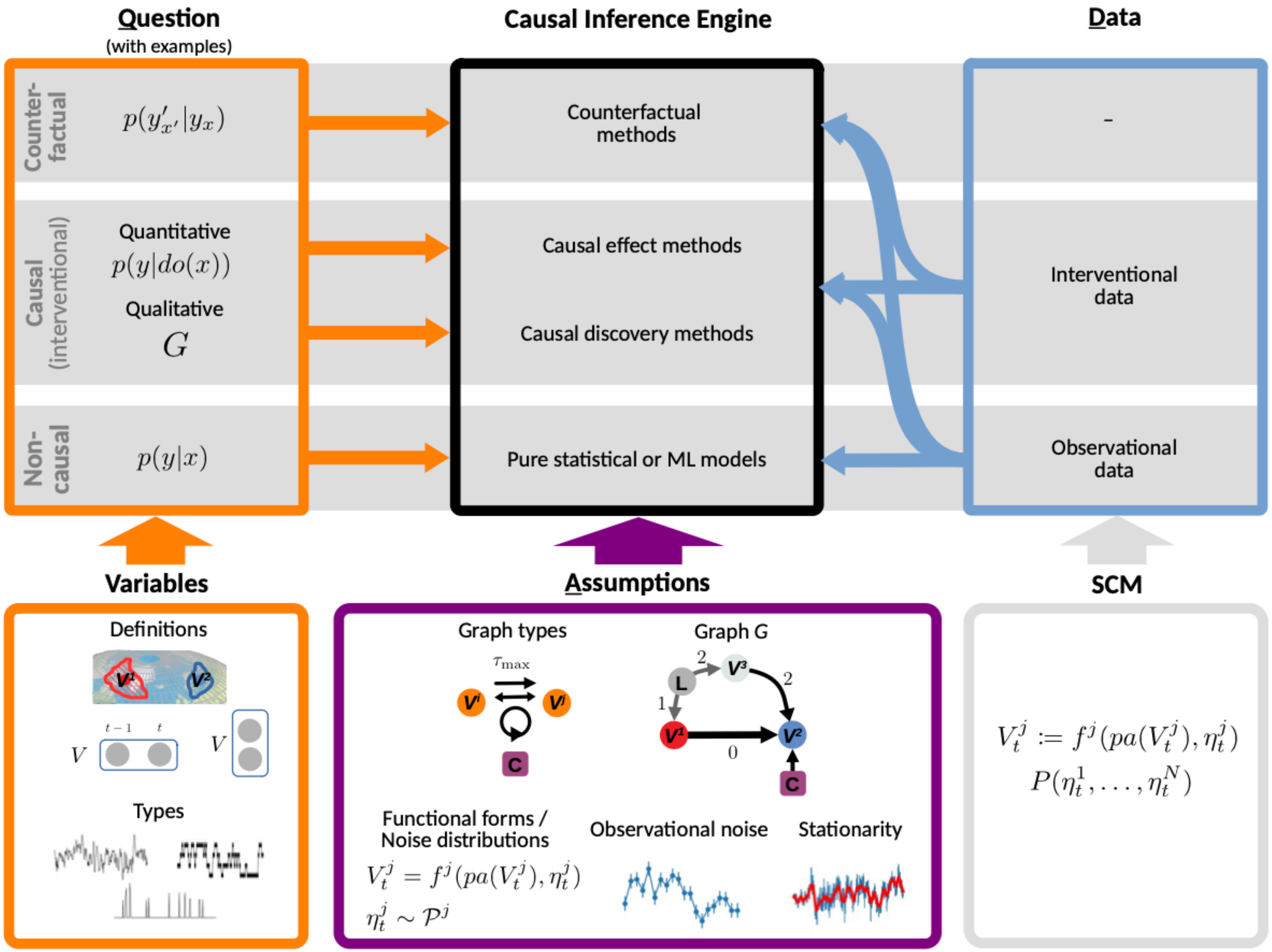
1. Defining causal problem in time series
2. Causal discovery
3. Causal discovery in time series

Defining causal problem in time series

Question-Assumptions-Data template.

Problem defining in Causal inference

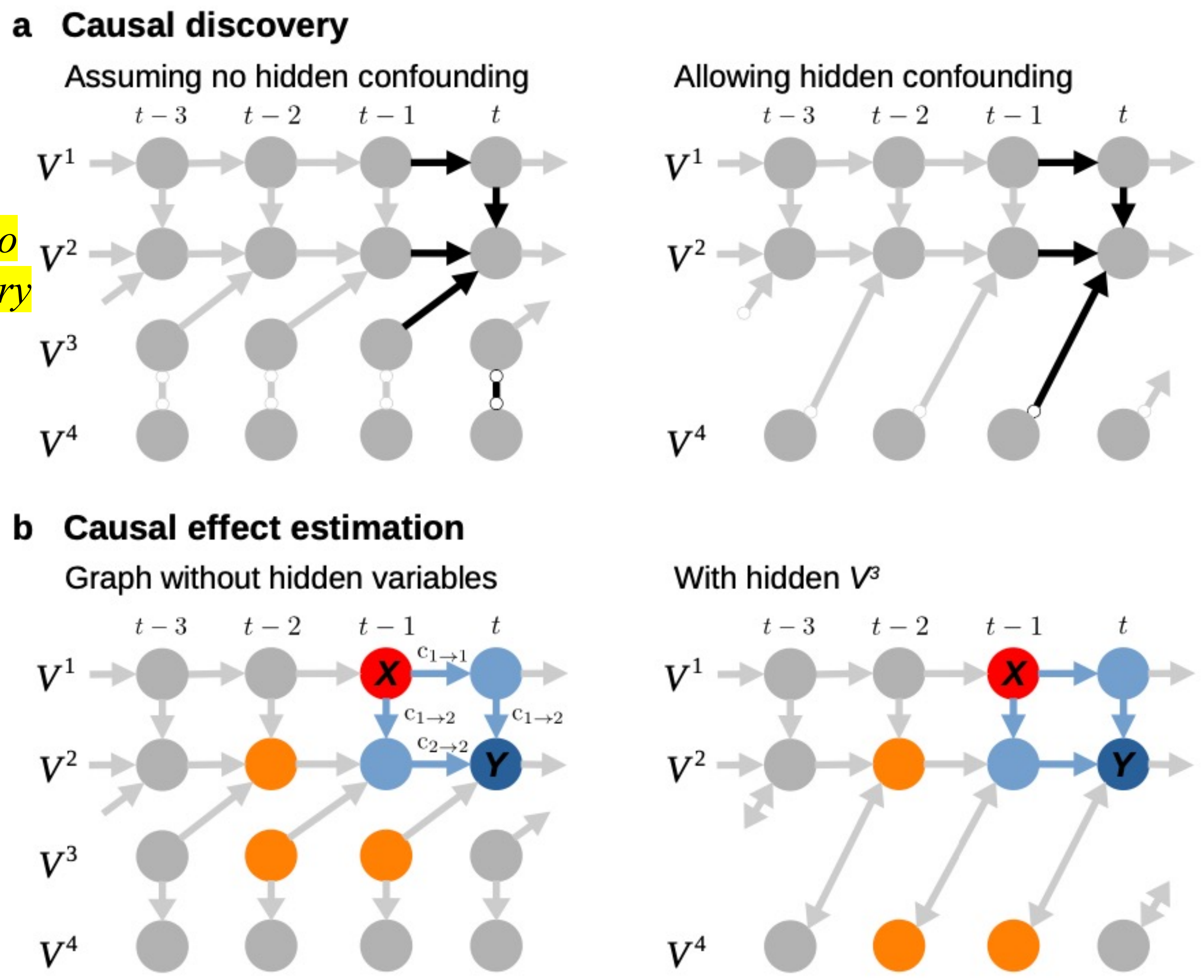
Interventional level

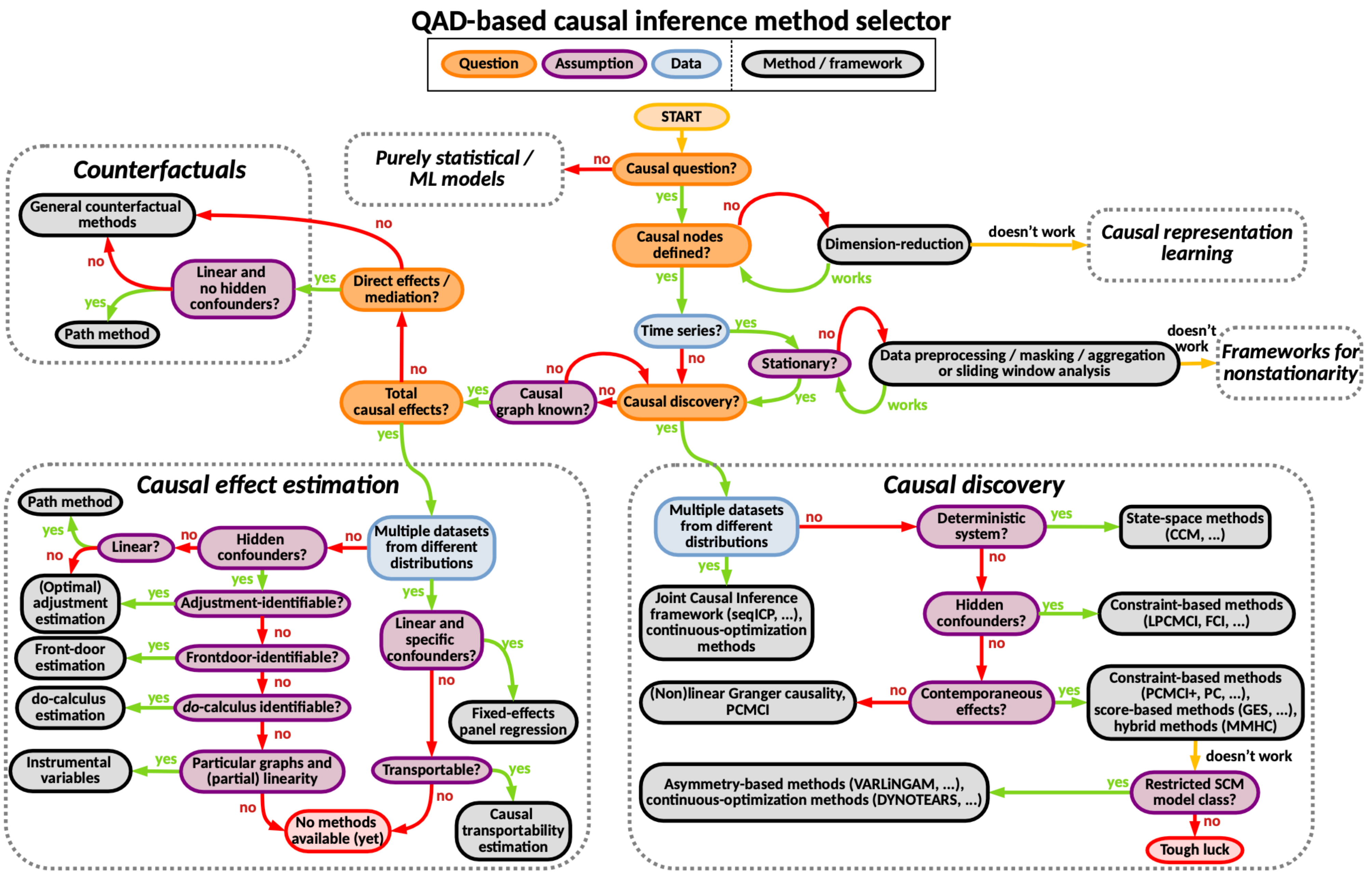


The problem transferred to Causality can be simply devided into two kinds of problem.

Based on the climate time series data, trying to build up the causal graph with causal discovery method

Based on the causal graph and related Assumption, Using adjustment formula to calculate the causal effect.





QAD-based causal inference method selector
Question
Assumption
Data
Method / framework
START
Causal question?
no
Purely statistical / ML models
yes
Causal nodes defined?
no
Dimension-reduction
doesn't work
Causal representation learning
works
yes
Time series?
yes
Stationary?
no
Data preprocessing / masking / aggregation or sliding window analysis
doesn't work
Frameworks for nonstationarity
works
yes
no
Causal discovery?
no
Causal graph known?
no
yes
Total causal effects?
no
Direct effects / mediation?
no
yes
Counterfactuals
General counterfactual methods
no
Linear and no hidden confounders?
yes
Path method
yes
yes
Causal effect estimation
Path method
yes
Linear?
no
no
Hidden confounders?
no
Multiple datasets from different distributions
yes
(Optimal) adjustment estimation
yes
Adjustment-identifiable?
no
Front-door estimation
yes
Frontdoor-identifiable?
no
do-calculus estimation
yes
do-calculus identifiable?
no
Instrumental variables
yes
Particular graphs and (partial) linearity
no
No methods available (yet)
yes
Linear and specific confounders?
yes
Fixed-effects panel regression
no
Transportable?
yes
Causal transportability estimation
no
Causal discovery
Multiple datasets from different distributions
no
Deterministic system?
yes
State-space methods (CCM, ...)
yes
Joint Causal Inference framework (seqICP, ...), continuous-optimization methods
no
Hidden confounders?
yes
Constraint-based methods (LPCMCI, FCI, ...)
no
(Non)linear Granger causality, PCMCI
no
Contemporaneous effects?
yes
Constraint-based methods (PCMCI+, PC, ...), score-based methods (GES, ...), hybrid methods (MMHC)
doesn't work
Asymmetry-based methods (VARLiNGAM, ...), continuous-optimization methods (DYNOTEARS, ...)
yes
Restricted SCM model class?
no
Tough luck

Causal Discovery from Observational Data

Markov Assumption

Markov assumption tells us if variables are d-separated in the graph G, then they are independent in the distribution P

$$X \perp\!\!\!\perp_G Y \mid Z \implies X \perp\!\!\!\perp_P Y \mid Z$$

However, going from independencies in the distribution P to d-separations in the graph G isn't something that the Markov assumption gives us, what we need is converse of Markov Assumption

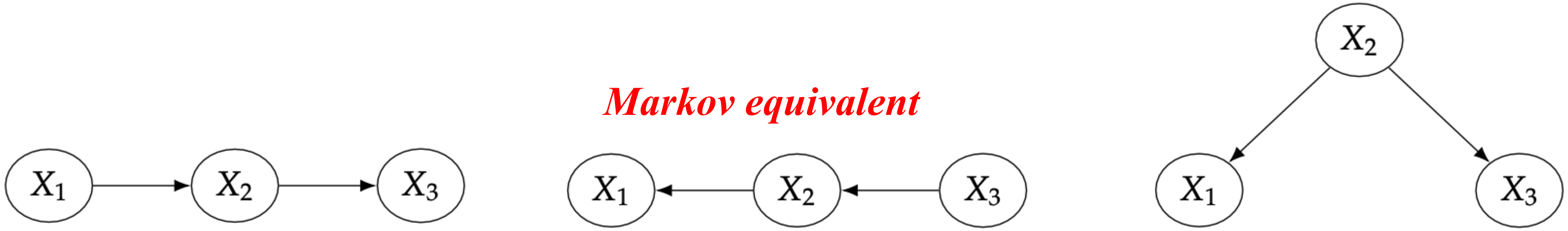
Assumption 11.1 (Faithfulness)

$$X \perp\!\!\!\perp_G Y \mid Z \impliedby X \perp\!\!\!\perp_P Y \mid Z \tag{11.1}$$

In addition to faithfulness, many methods also assume that there are no unobserved confounders, which is known as *causal sufficiency*.

Assumption 11.2 (Causal Sufficiency) *There are no unobserved confounders of any of the variables in the graph.*

Under the Markov, faithfulness, causal sufficiency, and acyclicity assumptions, we can partially identify the causal graph.



(a) Chain directed to the right

(b) Chain directed to the left

(c) Fork

Figure 11.2: Three Markov equivalent graphs

Different graphs correspond to the same set of independencies.

But for collider, it is its own Markov equivalence class

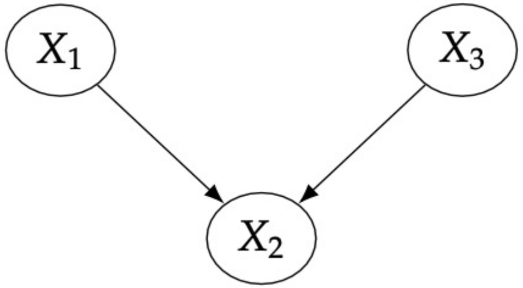


Figure 11.3: Immoralities are in their own Markov equivalence class.

How do we distinguish the chain and fork structure?

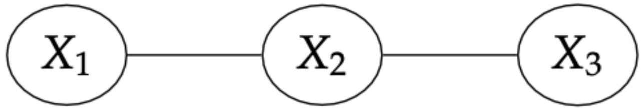


Figure 11.4: Chain/fork skeleton.

We can distinguish these by their skeleton

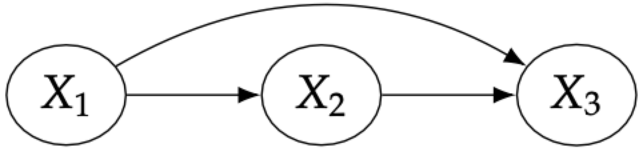


Figure 11.5: Complete graph.

To recap, we've pointed out two structural qualities that we can use to distinguish graphs from each other:

1. Immoralities
2. Skeleton

Proposition 11.1 (Markov Equivalence via Immoral Skeletons) *Two graphs are Markov equivalent if and only if they have the same skeleton and same immoralities.*

This means, we cannot directly get the causal graph like A → B or B → A,

However, We can get its skeleton like A – B. this is known as the *essential graph* or *CPDAG* (Completed Partially Directed Acyclic Graph).

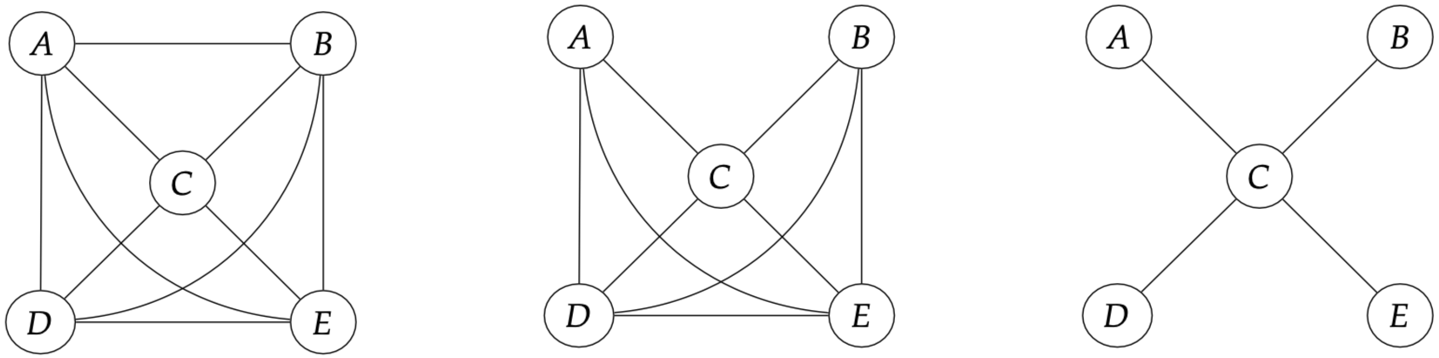
One popular algorithm for learning the essential graph is the PC algorithm.

The PC Algorithm

PC starts with a complete undirected graph and then trims it down and orients edges via three step

1. Identify the skeleton.
2. Identify immoralities and orient them.
3. Orient qualifying edges that are incident on colliders.

1. Identify the skeleton



(a) Complete undirected graph that we start with

(b) Undirected graph that remains after removing $X - Y$ edges where $X \perp\!\!\!\perp Y$

(c) Undirected graph that remains after removing $X - Y$ edges where $X \perp\!\!\!\perp Y \mid Z$

Figure 11.7: Illustration of the process of step 1 of PC, where we start with the complete graph (left) and remove edges until we've identified the skeleton of the graph (right), given that the true graph is the one in Figure 11.6.

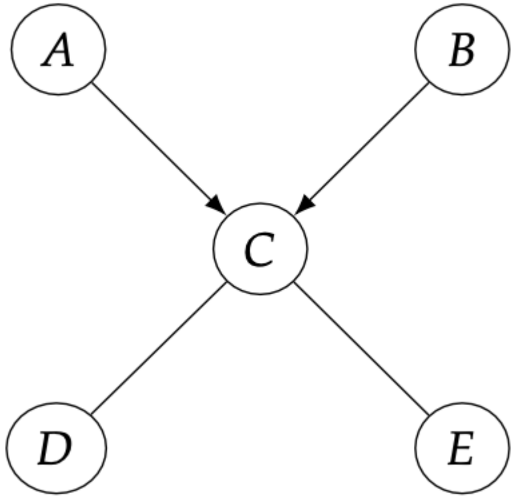


Figure 11.8: Graph from PC after we've oriented the immoralities.

2. Identifying the Immoralities

[3] This is called *orientation propagation*.

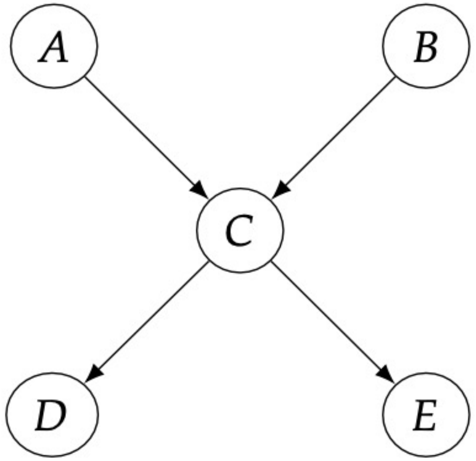


Figure 11.9: Graph from PC after we've oriented edges that would form immoralities if they were oriented in the other (incorrect) direction.

3. Orienting Qualifying Edges Incident on Colliders

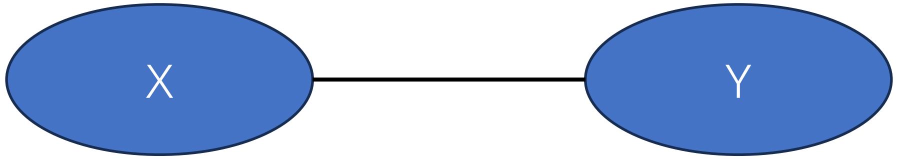
More Algorithm

The FCI (Fast Causal Inference) algorithm. — No Need assuming causal sufficiency

The CCD algorithm — No Need assuming acyclicity

SAT-based causal discovery — No Need assuming causal sufficiency and acyclicity

The best thing the conditional independence test can do is identify the skeleton. If we need to get the causal graph, we need to think about more.



Proposition 11.3 (Non-Identifiability of Two-Node Graphs) *For every joint distribution $P(x, y)$ on two real-valued random variables, there is an SCM in either direction that generates data consistent with $P(x, y)$.*

Mathematically, there exists a function f_Y such that

$$Y = f_Y(X, U_Y), \quad X \perp\!\!\!\perp U_Y \tag{11.6}$$

and there exists a function f_X such that

$$X = f_X(Y, U_X), \quad Y \perp\!\!\!\perp U_X \tag{11.7}$$

where U_Y and U_X are real-valued random variables.

Linear Non-Gaussian Noise

Assumption 11.3 (Linear Non-Gaussian) *All structural equations (causal mechanisms that generate the data) are of the following form:*

$$Y := f(X) + U \tag{11.8}$$

where f is a linear function, $X \perp\!\!\!\perp U$, and U is distributed as a non-Gaussian random variable.

Then, in this linear non-Gaussian setting, we can identify which of graphs $X \to Y$ and $X \leftarrow Y$ is the true causal graph.

Theorem 11.4 (Identifiability in Linear Non-Gaussian Setting) *In the linear non-Gaussian setting, if the true SCM is*

$$Y := f(X) + U\,, \quad X \perp\!\!\!\perp U\,, \tag{11.9}$$

then, there does not exist an SCM in the reverse direction

$$X := g(Y) + \tilde{U}\,, \quad Y \perp\!\!\!\perp \tilde{U}\,, \tag{11.10}$$

that can generate data consistent with $P(x, y)$.

In Time Series

1. Causal discovery
2. Causal effect estimation

1. Causal discovery

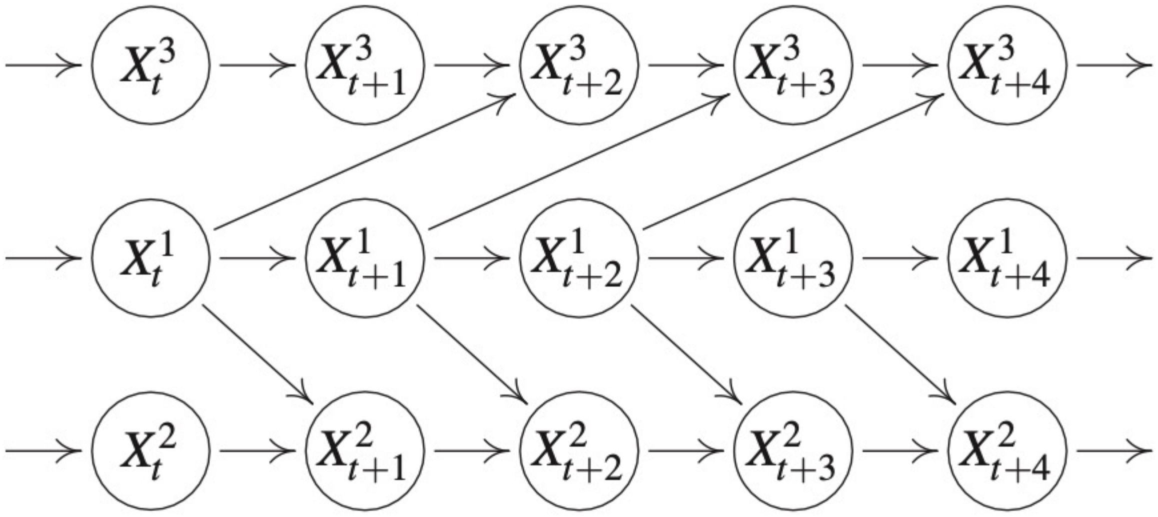


Figure 10.1: Example of a time series with no instantaneous effects.

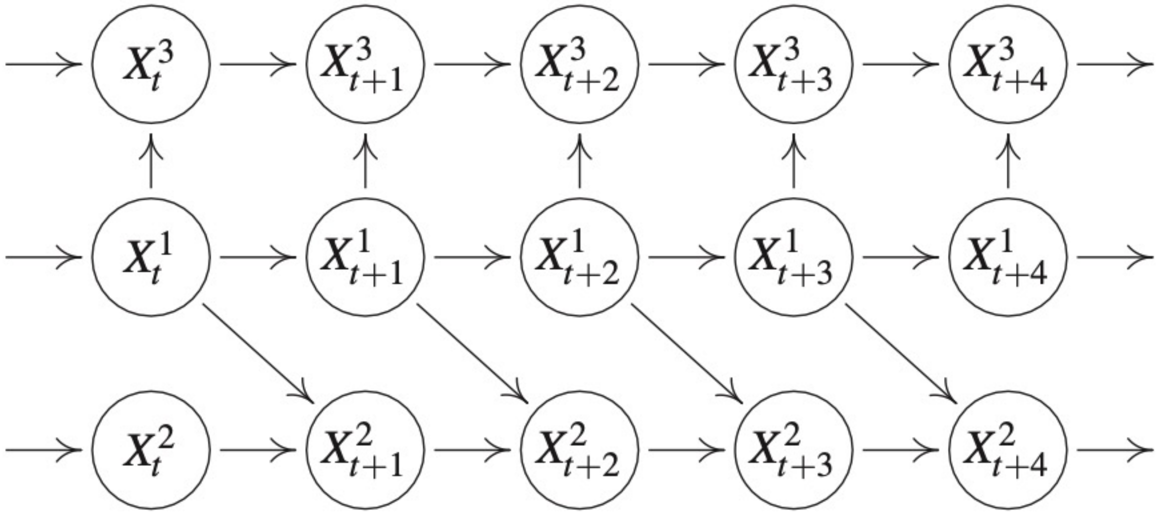
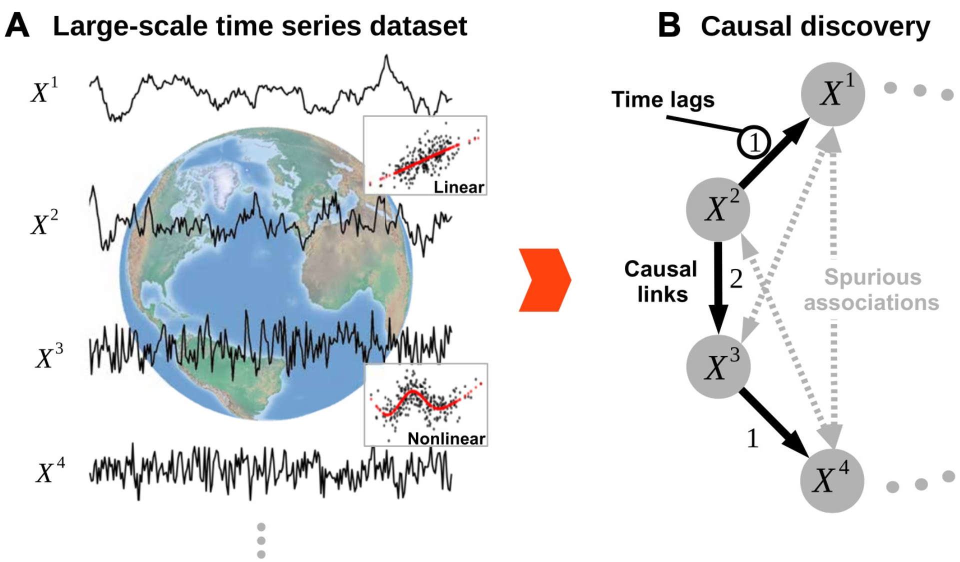


Figure 10.2: Example of a time series with instantaneous effects.

PCMCI



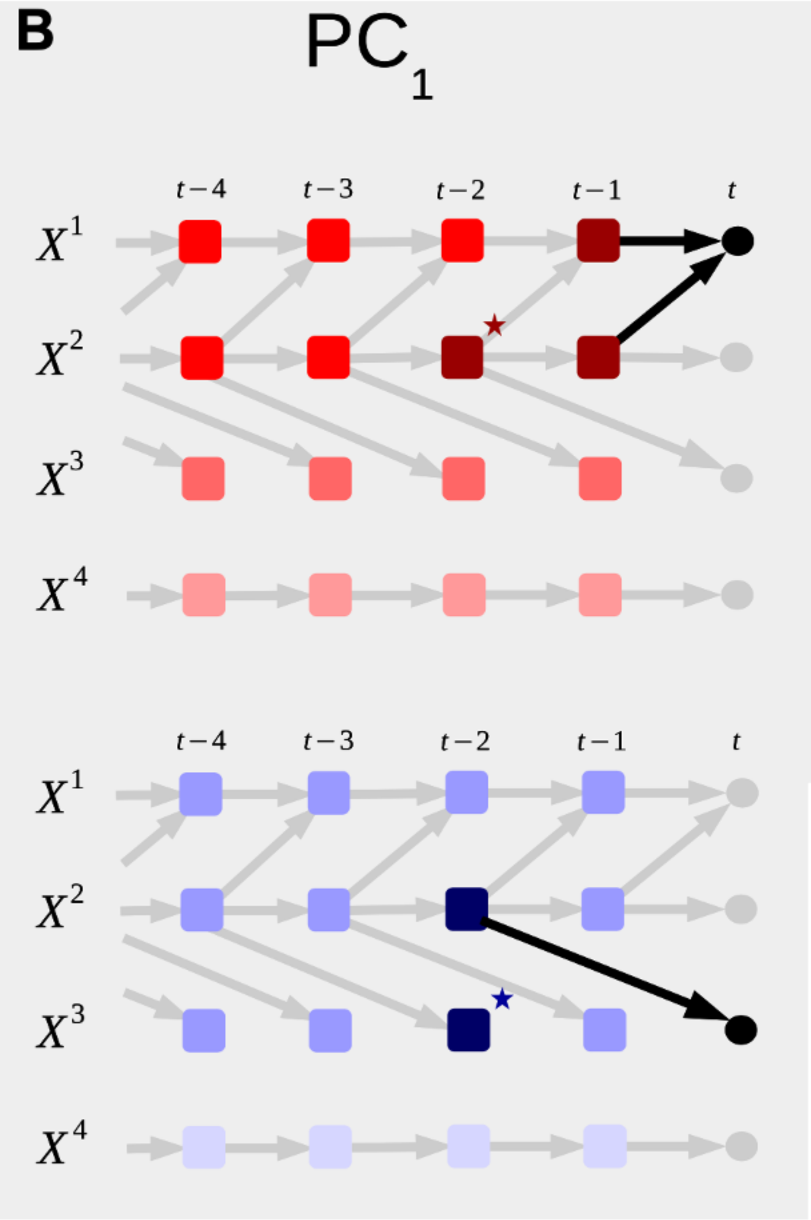
Consider an underlying time-dependent system

$$X_t^j = f_j(\mathcal{P}(X_t^j), \eta_t^j)$$

(1) PC1 condition selection to identify relevant conditions for all included time series variables

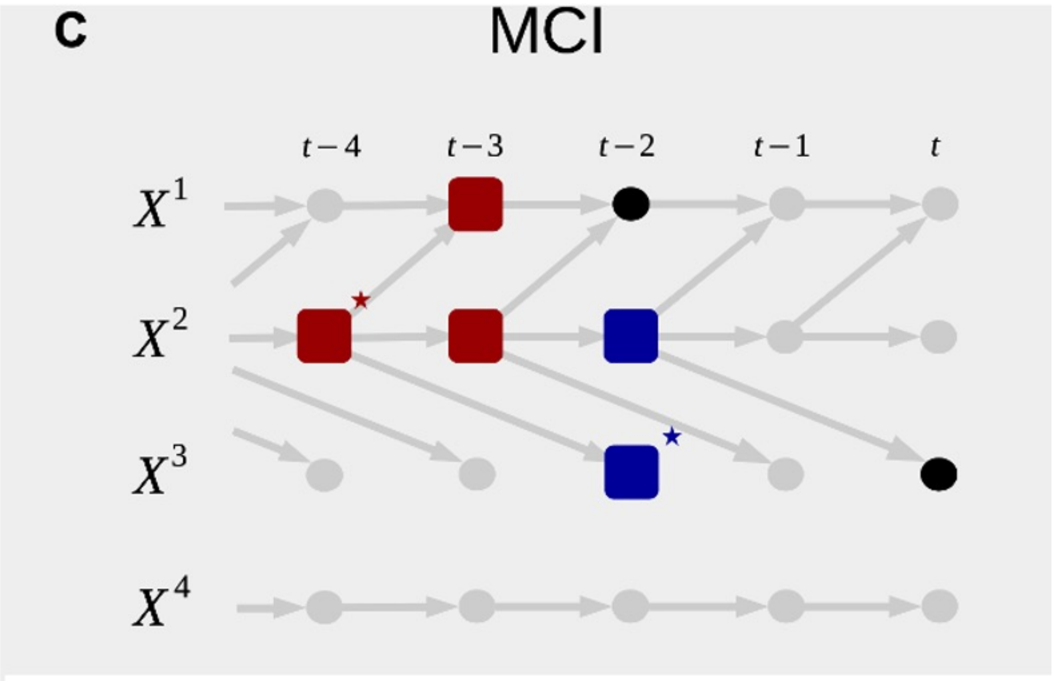
(2) the momentary conditional independence (MCI) test to test whether $X_{t-\tau}^i \rightarrow X_t^j$ with

$$\text{MCI:} \quad X_{t-\tau}^i \perp\!\!\!\perp X_t^j \mid \widehat{\mathscr{P}}(X_t^j) \setminus \{X_{t-\tau}^i\}, \widehat{\mathscr{P}}(X_{t-\tau}^i)\,.$$



Let V be the set of all variables, and X be the target variable.

1.Initialization:
V'= {variables in V that are not unconditionally independent of X}
2.For each iteration *i*:
a. Identify i = variable in V' with the strongest dependency on X from the prior iteration.
b. Update V'= V' - {variables in V' that are independent of X given Yi}



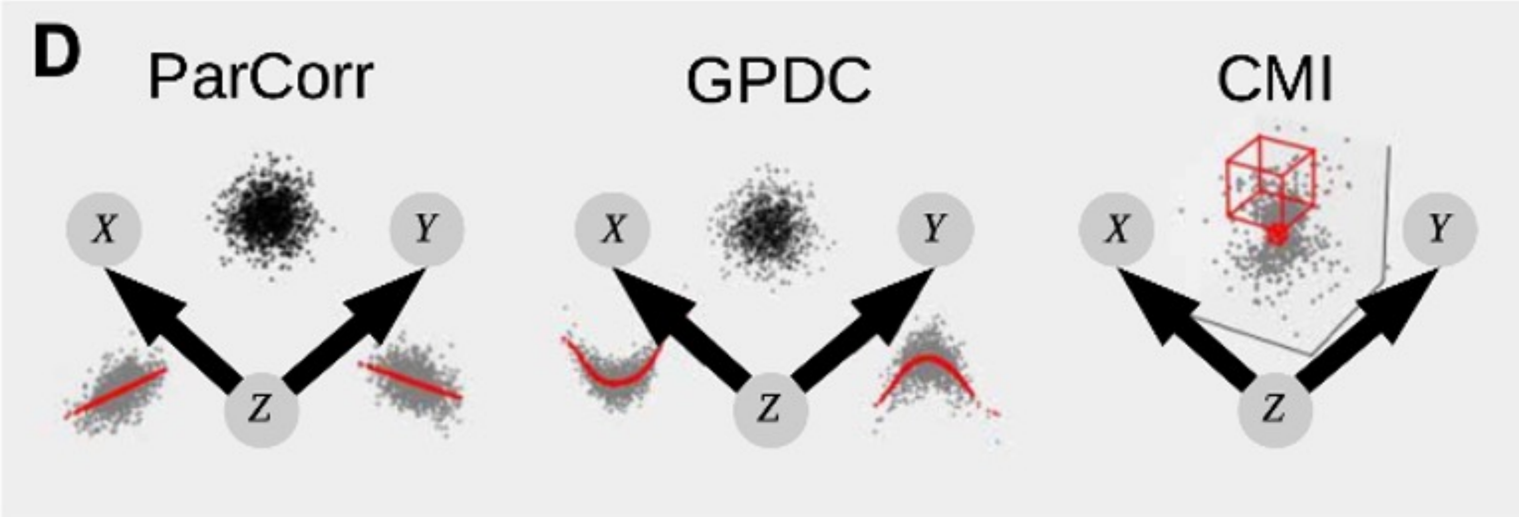
Momentary Conditional Independence(MCI)

1. Conditional independence
2. Considering Time Step
3. Autocorrelation

$$\text{MCI} : X^i_{t-\tau} \not\perp\!\!\!\perp X^j_t \mid \widehat{\mathcal{P}}\left(X^j_t\right) \setminus \left\{X^i_{t-\tau}\right\}, \widehat{\mathcal{P}}\left(X^i_{t-\tau}\right)$$

These low-dimensional conditions are then used in the MCI conditional independence test:

For testing $X^1_{t-2} \rightarrow X^3_t$, the conditions ˆP (X^3_t) (blue boxes) are sufficient to establish conditional independence, while the additional conditions on the parents ˆP (X^1_{t-2}) (red boxes) account for autocorrelation and make MCI an estimator of causal strength.



•The gray scatter plots depict regressions of X and Y based on Z.
•The black scatter plots represent the residuals of these regressions.
•The red cubes in the context of CMI symbolize the k-nearest neighbor test, which operates adaptively with the data without requiring an additivity assumption.

Linear vs. Nonlinear:
•**ParCorr (Partial Correlation)**: A linear independence test that *assumes linear additive noise models*.
•**GPDC (Generalized Partial Directed Coherence)**: A nonlinear test, but it makes an *assumption of additivity, not delving deeper into other nonlinear relationships*.
•**CMI (Conditional Mutual Information)**: Another nonlinear test. This one employs a data-adaptive*, model-free k-nearest neighbor technique*.

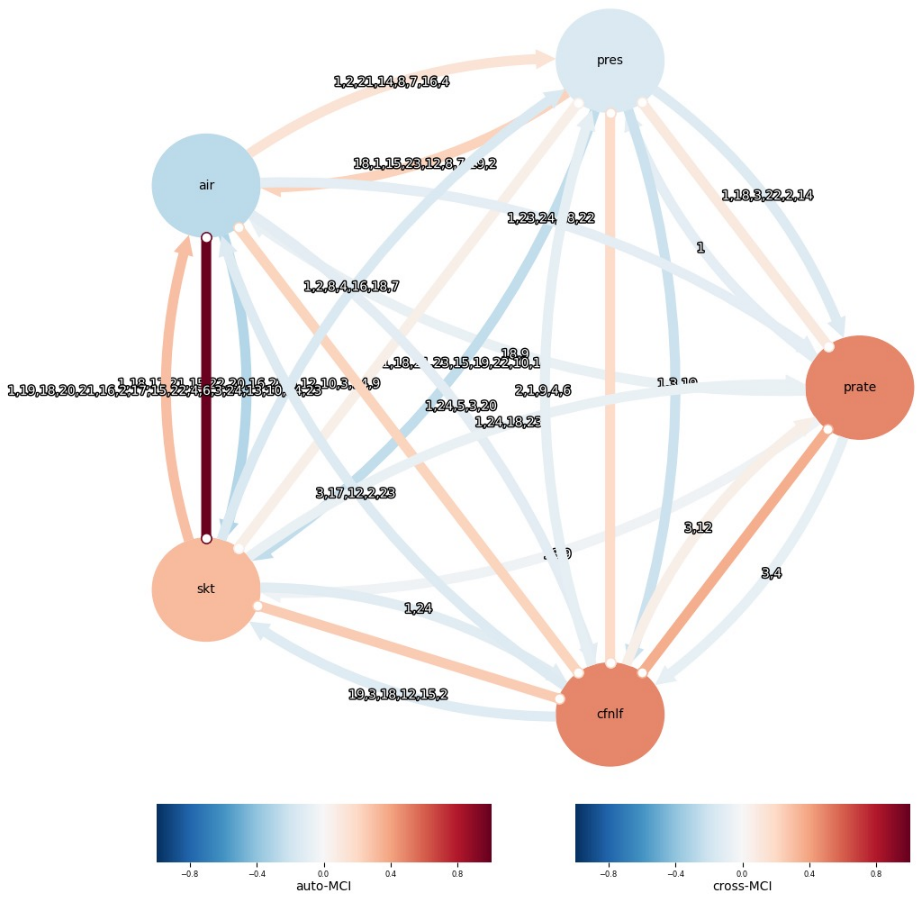
PCMCI+

Separate Skeleton Edge Removal into Two Phases:

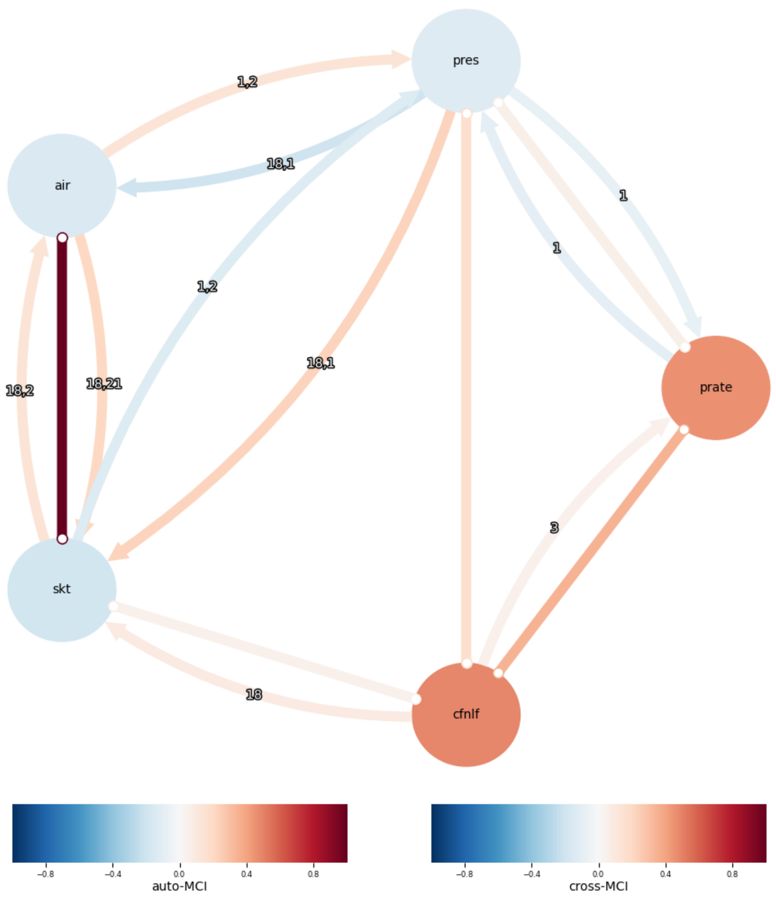
1. Lagged Conditioning Phase
2. Contemporaneous Conditioning Phase

Adaptive Testing Strategy: Instead of applying a uniform test to all possible links, PCMCI+ adaptively selects the most appropriate CI test for each link based on the characteristics of the data. This is crucial because, in complex datasets, not all relationships are of the same type; some might be linear, while others could be nonlinear.

PCMCI


pres
air
prate
skt
cfnlf
1,2,21,14,8,7,16,4
1,18,3,22,2,14
1
1,2,8,4,16,18,7
3,17,12,2,23
3,12
3,4
1,24
19,3,18,12,15,2
auto-MCI
cross-MCI
−0.8
−0.4
0.0
0.4
0.8

PCMCI+


pres
air
prate
skt
cfnlf
1,2
18,1
1
1
1,2
18,1
18,2
18,21
3
18
auto-MCI
cross-MCI
−0.8
−0.4
0.0
0.4
0.8

2. Causal effect estimation

A main goal of causal effect estimation is to estimate the total causal effect (often just referred to as causal effect) of a set of variables $X = \{X_1, \ldots, X_{NX}\}$ on another variable Y .

$$\Delta_{\mathbf{X}\to Y}(\mathbf{x}', \mathbf{x}) = \mathbb{E}[Y \mid do(\mathbf{X} = \mathbf{x}')] - \mathbb{E}[Y \mid do(\mathbf{X} = \mathbf{x})] \ ,$$

as the difference in the expected value of Y when setting X by intervention to x' as opposed to x.

Covariate adjustment.

Linear causal effect estimation

Covariate adjustment.

Formula articulates how the distribution of *Y* is associated with *Z* for a given value of *X*. Specifically, it encapsulates the potential causal effect of *X* on *Y* when factoring in the values of *Z*.

$$p(y \mid do(\mathbf{X} = \mathbf{x})) = \int p(y \mid \mathbf{x}, \mathbf{z}) p(\mathbf{z}) d\mathbf{z} \,.$$

Formula demonstrates how the expected (or average) value of *Y* for a given value of *X* is associated with *Z*.

$$\mathbb{E}\left[Y \mid do(\mathbf{X} = \mathbf{x})\right] = \mathbb{E}_{\mathbf{z}}\left[\mathbb{E}\left[Y \mid \mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}\right]\right] \,.$$

Linear causal effect estimation and the path method

Linear causal effect estimation

Linear causal effect estimation and the path method. For linear models $\Delta_{\mathbf{X}\to Y}(\mathbf{x}', \mathbf{x})$ in eq. (3) reduces to

$$\Delta_{\mathbf{X}\to Y}(\mathbf{x}', \mathbf{x}) = \Delta\mathbf{x} \cdot \vec{\delta}, \tag{6}$$

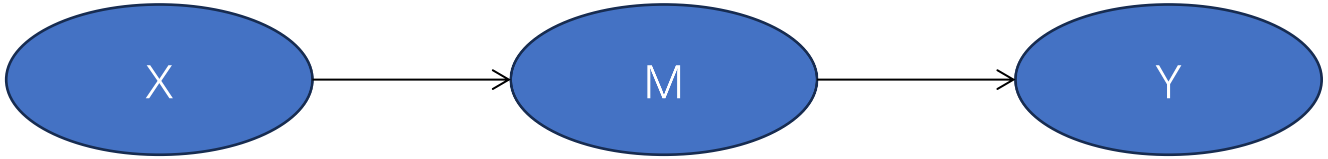
where $\vec{\delta} = (\delta_1, \ldots, \delta_{N_X})$ and δ_k is the controlled direct effect of $X_k \in \mathbf{X}$ on Y relative to $\mathbf{X} \setminus \{X_k\}$ and can be estimated as the regression parameter of $X_k \in \mathbf{X}$ in the linear regression of Y on $\mathbf{X} \cup \mathbf{Z}$.

Path method

The path method offers an efficient estimation approach for linear models and graphs that lack hidden variables. In linear causal mechanisms, each edge from *Vi* to *Vj* possesses a weight, often termed as link coefficient. The weight of each causal path is the product of all edge weights along that path, and *δk* is the sum of path weights across all appropriate causal paths.

However, to employ this method, all variables on the relevant paths and all parents of those variables must be observed.

Linear mediation analysis



Beyond quantifying the total causal effect of X on Y, one might be interested in the causal pathways — the mechanisms through which this effect is transmitted. This leads to questions like: how significant is the portion of the causal effect of X on Y that passes (or does not pass) through M, where M is a given set of mediators.

For linear models, the effect that transits through M is the sum-product of edge weights on proper causal paths from X to Y that pass through M. In the same vein, the effect that doesn't go through M can be estimated.