Causal discovery and weak equivalence of graphs

Søren Wengel Mogensen

Center for Statistics
Department of Finance
Copenhagen Business School

March 21, 2025

This talk in one slide

Causal discovery algorithms learn causal structure (represented by a graph) from observed data. For this, we can leverage constraints (i.e., observable properties that distinguish different causal structures), and conditional independence[1] is one such type of constraint.

We start from the classical case (observations from a random vector), and then continue with the time series case.

[1] For (discrete) random variables X_A, X_B, and X_C, we say that X_A and X_B are *conditionally independent given* X_C if $P(X_A = a, X_B = b | X_C = c) = P(X_A = a | X_C = c) P(X_B = b | X_C = c)$ whenever $P(X_C = c) > 0$.

Before we get started. . .

We're hiring PhD students in AI and statistics.
https://www.cbs.dk/cbs/jobs-paa-cbs/ledige-stillinger/phd-scholarships-in-ai-and-statistics-2025

Before we get started. . .

We're hiring PhD students in AI and statistics.
`https://www.cbs.dk/cbs/jobs-paa-cbs/ledige-stillinger/phd-scholarships-in-ai-and-statistics-2025`

This year (and next), we're hiring assistant/associate professors in statistics/machine learning (first call is coming later this year).

Feel free to reach out if you're interested in learning more (`swm.fi@cbs.dk`).

Before we get started…

We're hiring PhD students in AI and statistics.
https://www.cbs.dk/cbs/jobs-paa-cbs/ledige-stillinger/phd-scholarships-in-ai-and-statistics-2025

This year (and next), we're hiring assistant/associate professors in statistics/machine learning (first call is coming later this year).

Feel free to reach out if you're interested in learning more (swm.fi@cbs.dk).

There is a new Copenhagen-based network for researchers in causal discovery (theory and applications). Feel free to reach out if you're interested in joining (next meetings on April 1 and May 27).

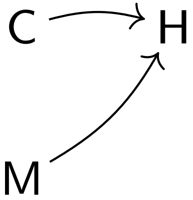
... then we can get started: Conditional independence

Conditional independence of random variables is used widely in quantitative fields of research. We can even use it to distinguish data-generating structures (see, e.g., Spirtes and Zhang [2018]).

... then we can get started: Conditional independence

Conditional independence of random variables is used widely in quantitative fields of research. We can even use it to distinguish data-generating structures (see, e.g., Spirtes and Zhang [2018]).

An example from everyday life (**C**offee, **H**eadache, **M**onday?),



M is exogenous

C is exogenous

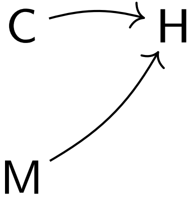
$H = f_H(M, C, \varepsilon_H)$

C and M are (marginally) independent.

... then we can get started: Conditional independence

Conditional independence of random variables is used widely in quantitative fields of research. We can even use it to distinguish data-generating structures (see, e.g., Spirtes and Zhang [2018]).

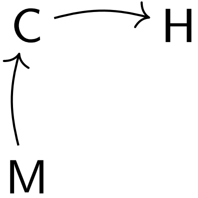
An example from everyday life (**C**offee, **H**eadache, **M**onday?),



M is exogenous

C is exogenous

$H = f_H(M, C, \varepsilon_H)$



M is exogenous

$C = f_C(M, \varepsilon_C)$

$H = f_H(C, \varepsilon_H)$

C and M are (marginally) independent.

H and M are conditionally independent given C.

Structural causal model

The **C**offee-**H**eadache-**M**onday example is a *structural causal model*:

Let $X = (X_1, \dots, X_n)^t$ be a random vector such that

$$\begin{aligned} X_1 &= f_1(X_{pa(1)}, \varepsilon_1) \\ X_2 &= f_2(X_{pa(2)}, \varepsilon_2) \\ &\vdots \\ X_i &= f_i(X_{pa(i)}, \varepsilon_i) \\ &\vdots \\ X_n &= f_n(X_{pa(n)}, \varepsilon_n) \end{aligned}$$

where the ε_i are independent random variables, and $X_{pa(i)}$ is a subset of variables. We make an associated graph with nodes $1, 2, \dots, n$ such that $i \to j$ if $i \in pa(j)$, and we assume this graph to be acyclic (no directed cycle $i \to \dots \to \dots \to i$) in which case it is a *directed acyclic graph* (DAG).

Structural causal model

The **C**offee-**H**eadache-**M**onday example is a *structural causal model*:

Let $X = (X_1, \ldots, X_n)^t$ be a random vector such that

$$\begin{aligned}
X_1 &= f_1(X_{pa(1)}, \varepsilon_1) \\
X_2 &= f_2(X_{pa(2)}, \varepsilon_2) \\
&\vdots \\
X_i &= a \quad \cancel{X_i = f_i(X_{pa(i)}, \varepsilon_i)} \\
&\vdots \\
X_n &= f_n(X_{pa(n)}, \varepsilon_n)
\end{aligned}$$

We assume that its *stable* under interventions, i.e., an intervention, $do(X_i = a)$, changes a single equation and leaves the other equations unchanged [Pearl, 2009].

The global Markov property

d-separation is a graphical concept via which the graph of a structural causal model implies a set of conditional independence constraints.

For disjoint node sets $A, B, C \subseteq \{1, 2, \ldots, n\}$ and a graph D on nodes $\{1, 2, \ldots, n\}$, there are straightforward algorithms to decide d-separation, i.e., if A and B are d-separated given C.

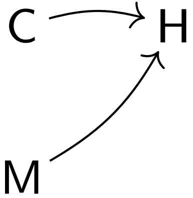
Theorem (Pearl [2009])

Let D be the graph associated with a structural causal model, and let A, B, and C be disjoint subsets of its node set.

If A and B are d-separated given C, then X_A and X_B are conditionally independent given X_C.

Revisiting the example

We see that the DAGs represent the dependence structure of the structural equations.

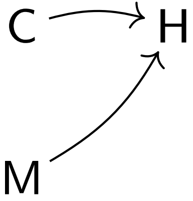


$$
\begin{aligned}
M &= f_M(\varepsilon_M)\\
C &= f_C(\varepsilon_C)\\
H &= f_H(M, C, \varepsilon_H)
\end{aligned}
$$

In this graph, C and M are *d*-separated by the empty set, implying that C and M are (marginally) independent.

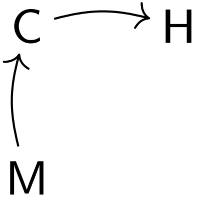
Revisiting the example

We see that the DAGs represent the dependence structure of the structural equations.



$$
\begin{aligned}
M &= f_M(\varepsilon_M) \\
C &= f_C(\varepsilon_C) \\
H &= f_H(M, C, \varepsilon_H)
\end{aligned}
$$

In this graph, C and M are d-separated by the empty set, implying that C and M are (marginally) independent.



$$
\begin{aligned}
M &= f_M(\varepsilon_M) \\
C &= f_C(M, \varepsilon_C) \\
H &= f_H(C, \varepsilon_H)
\end{aligned}
$$

In this graph, H and M are d-separated given C, implying that H and M are conditionally independent given C.

Causal discovery

Causal discovery aims to learn the underlying graph from observed data. One approach is to conduct statistical tests of conditional independence and map the associated *p*-values to a graph.

Causal discovery

Causal discovery aims to learn the underlying graph from observed data. One approach is to conduct statistical tests of conditional independence and map the associated p-values to a graph.

Many classical algorithms are *adaptive* in the sense that testing is done sequentially and a single test result may rule out certain graphs, see, e.g., Spirtes and Zhang [2018].

Causal discovery

Causal discovery aims to learn the underlying graph from observed data. One approach is to conduct statistical tests of conditional independence and map the associated *p*-values to a graph.

Many classical algorithms are *adaptive* in the sense that testing is done sequentially and a single test result may rule out certain graphs, see, e.g., Spirtes and Zhang [2018].

Another class of algorithms frame this learning problem as an optimization [Eberhardt et al., 2024]

$$\min_{D \text{ is a DAG}} g(D, \mathbb{P})$$

where g is a function measuring the discrepancy between the encoded independences and the observed p-values, $\mathbb{P}$ (i.e., $\mathbb{P}$ is a set of p-values from testing 'is X_i and X_j conditionally independent given X_C?').

Markov equivalence

This naturally leads to the question of *Markov equivalence*. We say that DAGs D_1 and D_2 (on a common node set) are *Markov equivalent* if for all disjoint node sets A, B, and C it holds that A and B are d-separated given C in D_1 if and only if A and B are d-separated given C in D_2.

Markov equivalence

This naturally leads to the question of *Markov equivalence*. We say that DAGs D_1 and D_2 (on a common node set) are *Markov equivalent* if for all disjoint node sets A, B, and C it holds that A and B are d-separated given C in D_1 if and only if A and B are d-separated given C in D_2.

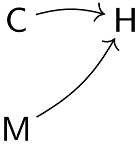
So for causal discovery to output a complete solution, we need to characterize the Markov equivalence classes of graphs.

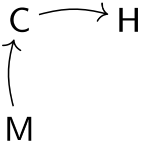
Markov equivalence

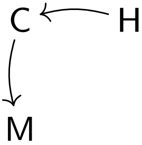
This naturally leads to the question of *Markov equivalence*. We say that DAGs D_1 and D_2 (on a common node set) are *Markov equivalent* if for all disjoint node sets A, B, and C it holds that A and B are d-separated given C in D_1 if and only if A and B are d-separated given C in D_2.

So for causal discovery to output a complete solution, we need to characterize the Markov equivalence classes of graphs.

A classical result [Pearl, 2009] shows that D_1 and D_2 are Markov equivalent if and only if they have the same *skeleton* (undirected graph after transforming $\rightarrow$ to —) and the same *unshielded colliders* (triples $i \rightarrow k \leftarrow j$ such that there is no edge between i and j). The 1st graph is not Markov equivalent with the 2nd (different skeleton and unshielded colliders), while the 2nd and 3rd are equivalent.







Weak Markov equivalence

We stated the causal discovery problem as an optimization,

$$\min_{D \text{ is a DAG}} g(D, \mathbb{P})$$

The set $\mathbb{P}$ contains p-values from testing 'is X_i and X_j conditionally independent given X_C?'. In graphs of moderate size, there is a very large number of such tests as C can be any subset of $V \setminus \{i, j\}$.

Weak Markov equivalence

We stated the causal discovery problem as an optimization,

$$\min_{D \text{ is a DAG}} g(D, \mathbb{P})$$

The set $\mathbb{P}$ contains p-values from testing 'is X_i and X_j conditionally independent given X_C?'. In graphs of moderate size, there is a very large number of such tests as C can be any subset of $V \setminus \{i, j\}$.

Moreover, statistical tests with large conditioning sets are expected to have low statistical power. This leads to the idea of *weak Markov equivalence*.

We say that DAGs D_1 and D_2 (on a common node set) are *k-weakly Markov equivalent* if for all disjoint node sets A, B, and C **such that** $|C| \leq k$ it holds that A and B are d-separated given C in D_1 if and only if A and B are d-separated given C in D_2.

Weak Markov equivalence

If $[D]$ is the set of DAGs that are Markov equivalent with D, and $[D]_k$ is the set of DAGs that are k-weakly Markov equivalent with D, then it follows directly that $[D] \subseteq [D]_k$. This means that a weak equivalence class is less informative, however, the associated learning problem is more feasible.

Weak Markov equivalence

If $[D]$ is the set of DAGs that are Markov equivalent with D, and $[D]_k$ is the set of DAGs that are k-weakly Markov equivalent with D, then it follows directly that $[D] \subseteq [D]_k$. This means that a weak equivalence class is less informative, however, the associated learning problem is more feasible.

One can characterize weak equivalence of DAGs similarly to Markov equivalence which facilitates causal discovery of weak equivalence classes [Kocaoglu, 2024, Mogensen, 2025b].

From random vectors to stochastic processes

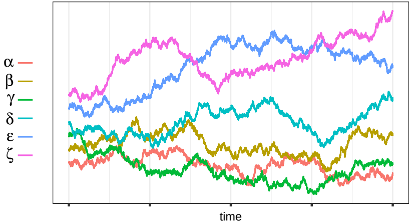
We will now make the following substitutions,

- random vector $\mapsto$ (multivariate) stochastic process,
- nodes represent random variables $\mapsto$ nodes represent coordinate processes,
- conditional independence $\mapsto$ (conditional) Granger (non)causality.

From random vectors to stochastic processes

We observe data from a multivariate time series in discrete time, $X = (X_t^\alpha)_{t\in\mathbb{Z},\alpha\in V}$, where V is the index set of the *coordinate processes* of X.

Example data where $V = \{\alpha,\beta,\gamma,\delta,\varepsilon,\zeta\} \simeq \{1,2,3,4,5,6\}, n = 6$.



Most results in this talk also hold for continuous-time processes, e.g., diffusions and point processes. For simplicity, we will stick to the (discrete-time) time series.

Structural causal model

We now assume a *dynamical* structural causal model,

$$X_t^i = \mathbf{F}_i(\bar{X}_{t-1}, N_t)$$

where N_t are i.i.d. random vectors with independent entries and $\bar{X}_{t-1} = \{\ldots, X_{t-3}, X_{t-2}, X_{t-1}\}$.

Structural causal model

We now assume a *dynamical* structural causal model,

$$X_t^i = \mathbf{F}_i(\bar{X}_{t-1}, N_t)$$

where N_t are i.i.d. random vectors with independent entries and $\bar{X}_{t-1} = \{\ldots, X_{t-3}, X_{t-2}, X_{t-1}\}$.

From this representation, we define a directed graph $\mathcal{D} = (V, E)$ such that for each $i, j \in V$, the edge $i \rightarrow j$ is in E if and only if $\mathbf{F}_j$ depends on $\bar{X}_{i,t-1}$. We say that $\mathcal{D}$ is the *causal graph* of the stochastic process X_t.

Structural causal model

We now assume a *dynamical* structural causal model,

$$X_t^i = \mathbf{F}_i(\bar{X}_{t-1}, N_t)$$

where N_t are i.i.d. random vectors with independent entries and $\bar{X}_{t-1} = \{\ldots, X_{t-3}, X_{t-2}, X_{t-1}\}$.

From this representation, we define a directed graph $\mathcal{D} = (V, E)$ such that for each $i, j \in V$, the edge $i \to j$ is in E if and only if $\mathbf{F}_j$ depends on $\bar{X}_{i,t-1}$. We say that $\mathcal{D}$ is the *causal graph* of the stochastic process X_t.

We now need a testable constraint in this model class.

Granger (non)causality

If $D \subseteq V$, let $X_t^D = \{X_t^d : d \in D\}$ and $X_{<t}^D = \{X_s^d : d \in D, s < t\}$. If $D = \{d\}$, then $X_t^D = X_t^d$ and $X_{<t}^D = \{\ldots, X_{t-3}^d, X_{t-2}^d, X_{t-1}^d\}$.

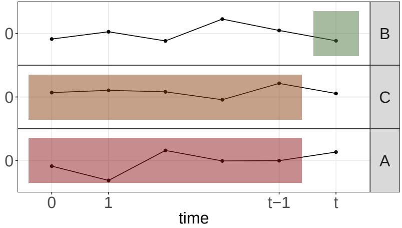
Definition (Granger (non)causality, Granger [1969], Eichler and Didelez [2010])

We say that A is Granger noncausal for B given C, and write $A \not\rightarrow B \mid C$, if for all $t \in \mathbb{Z}$,

$$X_{<t}^A \perp X_t^B \mid X_{<t}^C.$$

Granger causality is analogous to local independence [Schweder, 1970, Aalen, 1987] in continuous-time processes.

Note that Granger (non)causality is not symmetric, i.e., $A \not\rightarrow B \mid C$ does not imply $B \not\rightarrow A \mid C$.



Granger (non)causality

If $D \subseteq V$, let $X_t^D = \{X_t^d : d \in D\}$ and $X_{<t}^D = \{X_s^d : d \in D, s < t\}$. If $D = \{d\}$, then $X_t^D = X_t^d$ and $X_{<t}^D = \{\ldots, X_{t-3}^d, X_{t-2}^d, X_{t-1}^d\}$.

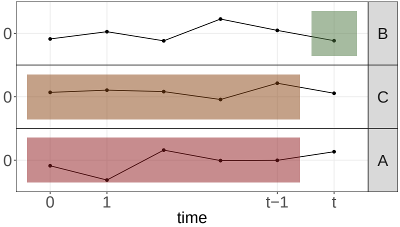
Definition (Granger (non)causality, Granger [1969], Eichler and Didelez [2010])

We say that A is Granger noncausal for B given C, and write $A \not\to B \mid C$, if for all $t \in \mathbb{Z}$,

$$X_{<t}^A \perp X_t^B \mid X_{<t}^C.$$

Granger causality is analogous to local independence [Schweder, 1970, Aalen, 1987] in continuous-time processes.

Note that Granger (non)causality is not symmetric, i.e., $A \not\to B \mid C$ does not imply $B \not\to A \mid C$.



The global Markov property

δ-/μ-separation is a graphical separation criterion, analogous to *d*-separation in DAGs (and *m*-separation in marginalizations of DAGs) [Didelez, 2008, Mogensen and Hansen, 2020].

The global Markov property

δ-/μ-separation is a graphical separation criterion, analogous to *d*-separation in DAGs (and *m*-separation in marginalizations of DAGs) [Didelez, 2008, Mogensen and Hansen, 2020].

If B is μ-separated from A given C in the graph $\mathcal{D}$, then we write $A \perp_\mu B \mid C \; [\mathcal{D}]$. μ-separation is not symmetric. We let $\mathcal{I}(\mathcal{D}) = \{(A, B, C) : A \perp_\mu B \mid C \; [\mathcal{D}]\}$ denote the set of μ-separations implied by a graph, $\mathcal{D}$.

The global Markov property

δ-/μ-separation is a graphical separation criterion, analogous to *d*-separation in DAGs (and *m*-separation in marginalizations of DAGs) [Didelez, 2008, Mogensen and Hansen, 2020].

If B is μ-separated from A given C in the graph $\mathcal{D}$, then we write $A \perp_\mu B \mid C \ [\mathcal{D}]$. μ-separation is not symmetric. We let $\mathcal{I}(\mathcal{D}) = \{(A, B, C) : A \perp_\mu B \mid C \ [\mathcal{D}]\}$ denote the set of μ-separations implied by a graph, $\mathcal{D}$.

Theorem (Eichler [2007], Eichler and Didelez [2010] and Mogensen and Hansen [2020] (supplementary material))

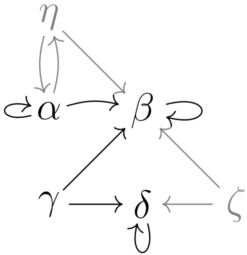
Let X be a multivariate time series and let $\mathcal{D}$ be its Granger-causal graph. Let $A, B, C \subseteq V$. Under regularity conditions,

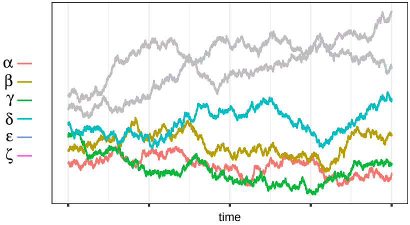
$$A \perp_\mu B \mid C \ [\mathcal{D}] \Rightarrow A \not\to B \mid C,$$

equivalently $\mathcal{I}(\mathcal{D}) \subseteq \mathcal{I} = \{(A, B, C) : (A \not\to B \mid C) \text{ in } P_X\}$.

Marginalization

Often it is only reasonable to assume that we observe a subset of the processes in the stochastic system, $O \subseteq V$.





No assumptions are made about the number of unobserved processes or their connections.

Graphical marginalization

We would like a graph on nodes $O \subseteq V$ that expresses the separations in $\mathcal{D} = (V, E)$, i.e., a graph $\mathcal{G} = (O, F)$ such that for all $A, B, C \subseteq O$

$$A \perp_\mu B \mid C \ [\mathcal{D}] \Leftrightarrow A \perp_\mu B \mid C \ [\mathcal{G}],$$

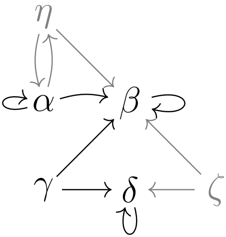
and a procedure to construct $\mathcal{G}$ from $\mathcal{D}$. We can use

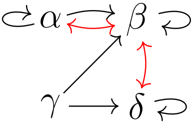
- the class of *directed mixed graphs* (DMGs), and
- *latent projection* [Verma and Pearl, 1991, Richardson et al., 2023, Mogensen and Hansen, 2020].

In a DMG two nodes α and β can be joined by any subset of edges $\{\alpha \rightarrow \beta, \beta \rightarrow \alpha, \alpha \leftrightarrow \beta\}$.

Graphical marginalization

Let $O = \{\alpha, \beta, \gamma, \delta\}$ below. The *marginal* (over O) independence models are equal for the two DMGs.





Markov equivalence

Let $\mathcal{G}_1 = (V, E_1)$ and $\mathcal{G}_2 = (V, E_2)$ be DMGs. We say that $\mathcal{G}_1$ and $\mathcal{G}_2$ are *Markov equivalent* if for all $A, B, C \subseteq V$,

$$A \perp_\mu B \mid C \; [\mathcal{G}_1] \Leftrightarrow A \perp_\mu B \mid C \; [\mathcal{G}_2].$$

We use $[\mathcal{G}_1]$ to denote the Markov equivalence class of $\mathcal{G}_1$. What can we say about the Markov equivalence classes?

Complexity of μ-separation DMGs

Theorem (Mogensen [2025a])

Deciding Markov equivalence of DMGs is coNP-complete.

The theorem also holds under certain sparsity constraints on the graphs.

This also implies that no polynomial-time algorithm which takes an independence model as input and which is correct in the oracle case can output the greatest element of the corresponding Markov equivalence class.

Weak equivalence

Let $V = \{1, 2, \ldots, n\}$.

Definition (Weak equivalence, Mogensen [2025a])

Let $k = 0, 1, 2, \ldots, n$. We say that DMGs $\mathcal{G}_1 = (V, E_1)$ and $\mathcal{G}_2 = (V, E_2)$ are *k-weakly equivalent* if for all $\alpha, \beta \in V$ and $C \subseteq V$ such that $|C| \leq k$

$$(\alpha, \beta, C) \in \mathcal{I}(\mathcal{G}_1) \Leftrightarrow (\alpha, \beta, C) \in \mathcal{I}(\mathcal{G}_2).$$

One can also define more general weak equivalences of DMGs [Mogensen, 2025a]. n-weak equivalence is the same as Markov equivalence.

Let $k_1 \leq k_2$. If $\mathcal{G}_1$ and $\mathcal{G}_2$ are k_2-weakly equivalent, then they are also k_1-weakly equivalent.

Weak equivalence, greatest element

Let $[\mathcal{G}]_k$ denote the k-weak equivalence class of $\mathcal{G}$. Weak equivalence classes contain a greatest element, just like Markov equivalence classes. This means that we can again use a DMEG as a representation of a weak equivalence class.

Theorem (Mogensen [2025a])

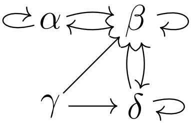
The set $[\mathcal{G}]_k$ contains a greatest element.

The theorem also holds under more general weak equivalences [Mogensen, 2025a].

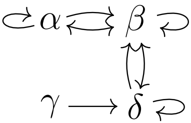
For a fixed V, we can think of k as a parameter controlling the granularity of the graphical modeling (smaller k gives larger equivalence classes). We can visualize this hierarchy by listing all pairs $(\mathcal{G}, k)$ such that $\mathcal{G}$ is the greatest element of $[\mathcal{G}]_k$ and connect the pairs $(\mathcal{G}, k)$ and $(\mathcal{G}^-, k-1)$ if $\mathcal{G} \in [\mathcal{G}^-]_{k-1}$. This gives us a *forest*.

An equivalence class and its DMEG

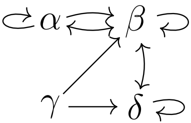
①



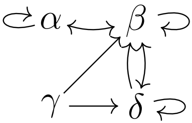
②



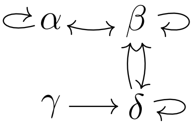
③



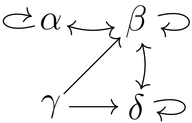
④



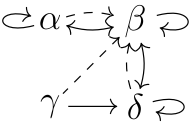
⑤



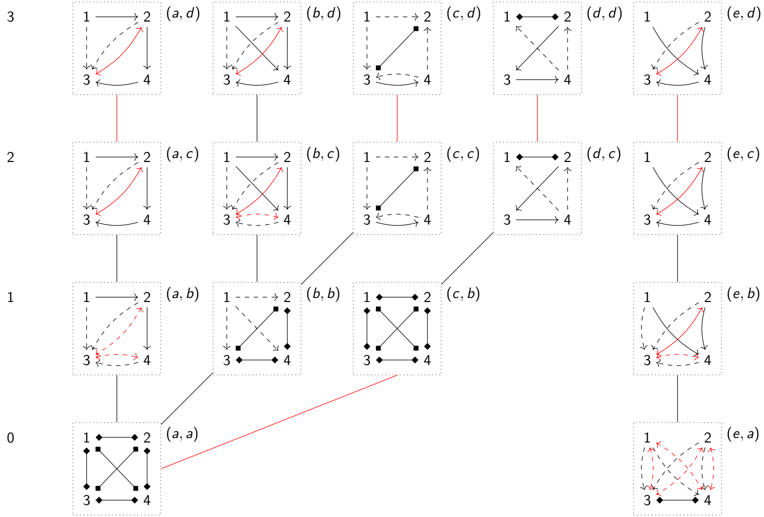
⑥



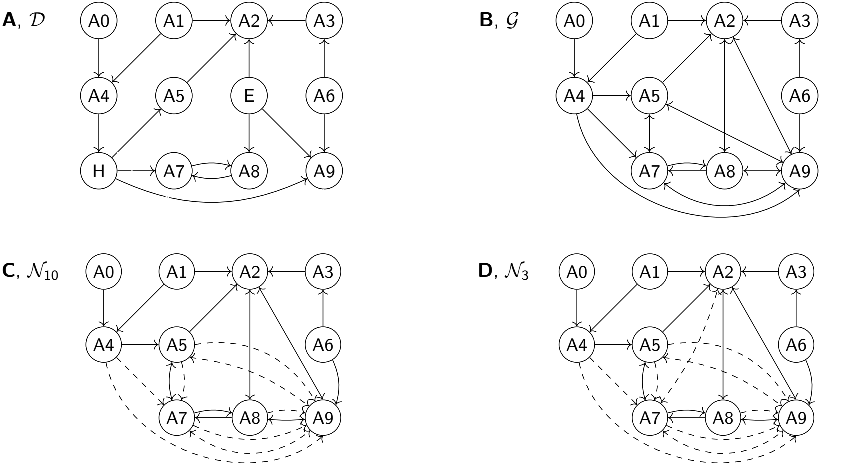
⑦



A part of the hierarchy on four nodes



Weak equivalence, another example



Summary

- In DAG-based causal models, causal discovery often uses tests of conditional independence.
- Some DAGs encode the same set of conditional independences, leading to Markov equivalence classes.
- Statistical and computational considerations lead to the consideration of weak equivalence.
- In time series, Granger (non)causality tests can be used to learn Markov equivalence classes of causal graphs.
- This leads to computationally hard problems, so learning a weak equivalence class is more feasible.

Research sponsored by VILLUM FONDEN (research grant 13358) and Independent Research Fund Denmark (research grant 0164-00023B).

Thank you for listening!

References I

Odd O. Aalen. Dynamic modelling and causality. *Scandinavian Actuarial Journal*, 1987(3-4):177–190, 1987.

Vanessa Didelez. Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):245–264, 2008.

Frederick Eberhardt, Nur Kaynar, and Auyon Siddiq. Discovering causal models with optimization: Confounders, cycles, and instrument validity. *Management Science*, 2024.

Michael Eichler. Granger causality and path diagrams for multivariate time series. *Journal of Econometrics*, 137:334–353, 2007.

Michael Eichler and Vanessa Didelez. On Granger causality and the effect of interventions in time series. *Lifetime Data Analysis*, 16(1):3–32, 2010.

Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica*, 37(3):424–438, 1969.

References II

Murat Kocaoglu. Characterization and learning of causal graphs with small conditioning sets. *Advances in Neural Information Processing Systems*, 36, 2024.

Søren Wengel Mogensen. Weak equivalence of local independence graphs. *Bernoulli*, 2025a. (to appear).

Søren Wengel Mogensen. Weak equivalence of maximal ancestral graphs, 2025b.

Søren Wengel Mogensen and Niels Richard Hansen. Markov equivalence of marginalized local independence graphs. *The Annals of Statistics*, 48(1): 539–559, 2020.

Judea Pearl. *Causality*. Cambridge University Press, 2009.

Thomas S Richardson, Robin J Evans, James M Robins, and Ilya Shpitser. Nested Markov properties for acyclic directed mixed graphs. *The Annals of Statistics*, 51(1):334–361, 2023.

Tore Schweder. Composable Markov processes. *Journal of Applied Probability*, 7 (2):400–410, 1970.

References III

Peter Spirtes and Kun Zhang. Search for causal models. In *Handbook of Graphical Models*, pages 439–470. CRC Press, 2018.

Thomas Verma and Judea Pearl. Equivalence and synthesis of causal models. Technical Report R-150, Computer Science Department, University of California, Los Angeles, 1991.