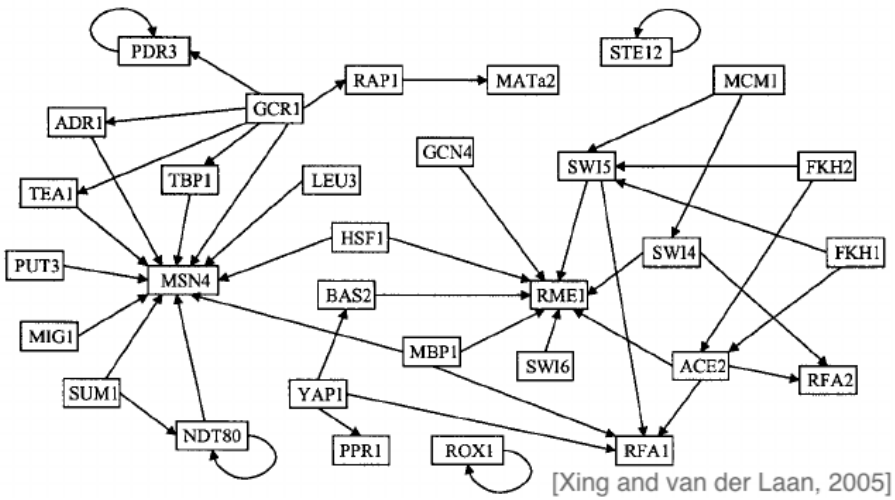
Hidden Variables in Linear Non-Gaussian Causal Models

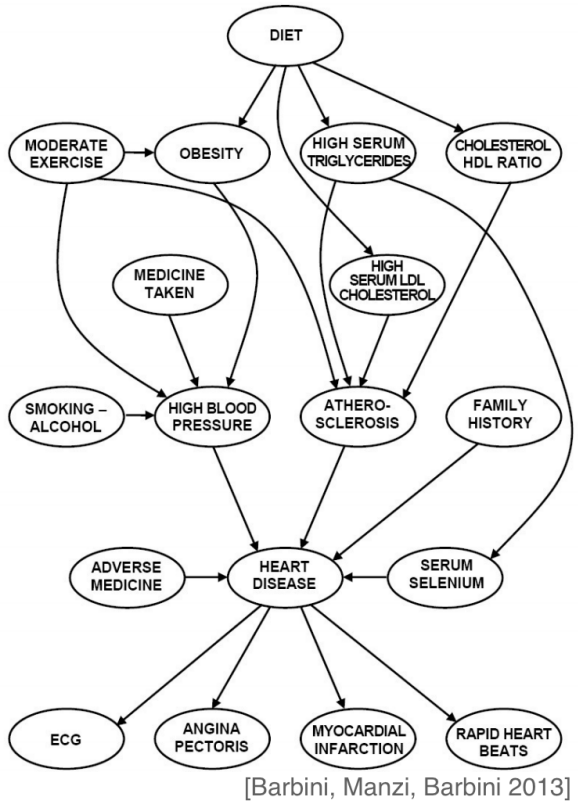
Elina Robeva
The University of British Columbia

May 5, 2021

Causal models



Gene regulatory networks



Disease diagnosis graphs

How can we learn the structure of these graphs from observations?

Structural equation models

Definition

A *structural equation model* consists of a directed acyclic graph (DAG) $G = (V, E)$, and a set of equations between random variables $\{X_v : v \in V\}$:

$$X_v = f_v(X_{\mathrm{pa}(v)}, \varepsilon_v), \ v \in V$$

where $X_{\mathrm{pa}(v)} = (X_u : u \to v \in E)$ and ε_v is noise such that $\{\varepsilon_v : v \in V\}$ are independent noise terms.

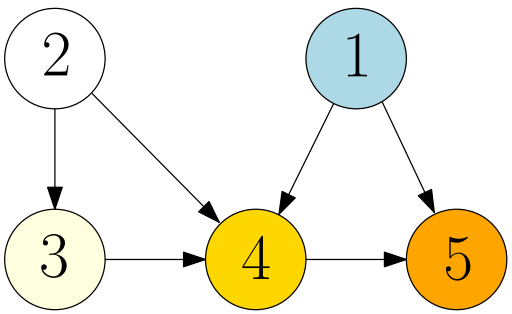
Structural equation models

Definition

A *structural equation model* consists of a directed acyclic graph (DAG) $G = (V, E)$, and a set of equations between random variables $\{X_v : v \in V\}$:

$$X_v = f_v(X_{\mathrm{pa}(v)}, \varepsilon_v), \ \ v \in V$$

where $X_{\mathrm{pa}(v)} = (X_u : u \to v \in E)$ and ε_v is noise such that $\{\varepsilon_v : v \in V\}$ are independent noise terms.



$$\begin{aligned}
X_1 &= f_1(\varepsilon_1) \\
X_2 &= f_2(\varepsilon_2) \\
X_3 &= f_3(X_2, \varepsilon_3) \\
X_4 &= f_4(X_1, X_2, X_3, \varepsilon_4) \\
X_5 &= f_5(X_1, X_4, \varepsilon_5).
\end{aligned}$$

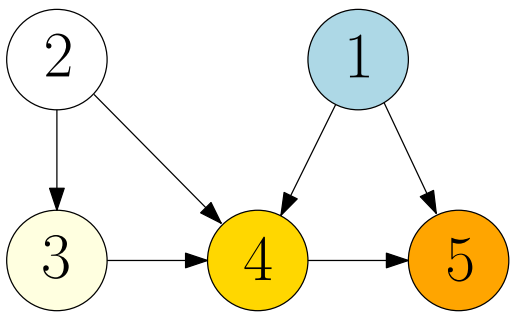
Structural equation models

Definition

A *structural equation model* consists of a directed acyclic graph (DAG) $G = (V, E)$, and a set of equations between random variables $\{X_v : v \in V\}$:

$$X_v = f_v(X_{\mathrm{pa}(v)}, \varepsilon_v), \ \ v \in V$$

where $X_{\mathrm{pa}(v)} = (X_u : u \to v \in E)$ and ε_v is noise such that $\{\varepsilon_v : v \in V\}$ are independent noise terms.



$$\begin{aligned}
X_1 &= f_1(\varepsilon_1) \\
X_2 &= f_2(\varepsilon_2) \\
X_3 &= f_3(X_2, \varepsilon_3) \\
X_4 &= f_4(X_1, X_2, X_3, \varepsilon_4) \\
X_5 &= f_5(X_1, X_4, \varepsilon_5).
\end{aligned}$$

Given samples $X^{(1)}, \ldots, X^{(n)} \in \mathbb{R}^{|V|}$ arising from such a model, can we identify G?

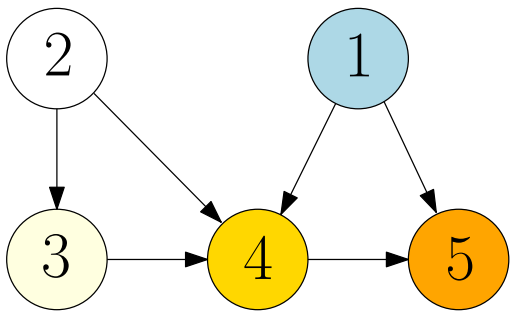
Structural equation models

Definition

A *structural equation model* consists of a directed acyclic graph (DAG) $G = (V, E)$, and a set of equations between random variables $\{X_v : v \in V\}$:

$$X_v = f_v(X_{\mathrm{pa}(v)}, \varepsilon_v), \ \ v \in V$$

where $X_{\mathrm{pa}(v)} = (X_u : u \to v \in E)$ and ε_v is noise such that $\{\varepsilon_v : v \in V\}$ are independent noise terms.

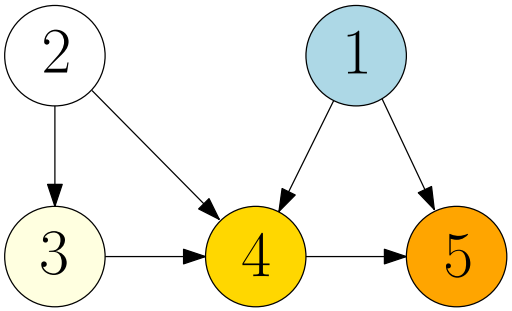


$$\begin{aligned} X_1 &= f_1(\varepsilon_1) \\ X_2 &= f_2(\varepsilon_2) \\ X_3 &= f_3(X_2, \varepsilon_3) \\ X_4 &= f_4(X_1, X_2, X_3, \varepsilon_4) \\ X_5 &= f_5(X_1, X_4, \varepsilon_5). \end{aligned}$$

Given samples $X^{(1)}, \ldots, X^{(n)} \in \mathbb{R}^{|V|}$ arising from such a model, can we identify G?

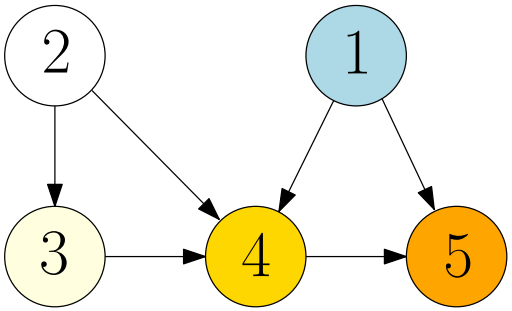
- Linear structural equation models

Linear structural equation models



$$
\begin{aligned}
X_1 &= \varepsilon_1 \\
X_2 &= \varepsilon_2 \\
X_3 &= \lambda_{23} X_2 + \varepsilon_3 \\
X_4 &= \lambda_{14} X_1 + \lambda_{24} X_2 + \lambda_{34} X_3 + \varepsilon_4 \\
X_5 &= \lambda_{15} X_1 + \lambda_{45} X_4 + \varepsilon_5 .
\end{aligned}
$$

Linear structural equation models

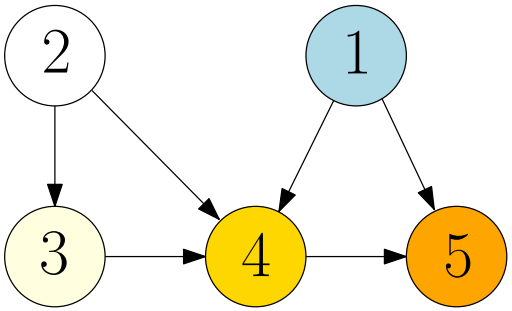


$$
\begin{aligned}
X_1 &= \varepsilon_1 \\
X_2 &= \varepsilon_2 \\
X_3 &= \lambda_{23} X_2 + \varepsilon_3 \\
X_4 &= \lambda_{14} X_1 + \lambda_{24} X_2 + \lambda_{34} X_3 + \varepsilon_4 \\
X_5 &= \lambda_{15} X_1 + \lambda_{45} X_4 + \varepsilon_5 .
\end{aligned}
$$

For a general directed acyclic graph $G = (V, E)$, the *linear structural equation model corresponding to* G consists of the the graph G and the linear equations

$$
X_i = \sum_{j \in \mathrm{pa}(i)} \lambda_{ji} X_j + \varepsilon_i, \qquad \text{where the variables } \{\varepsilon_i\}_{i \in V} \text{ are independent.}
$$

Linear structural equation models



$$
\begin{aligned}
X_1 &= \varepsilon_1 \\
X_2 &= \varepsilon_2 \\
X_3 &= \lambda_{23} X_2 + \varepsilon_3 \\
X_4 &= \lambda_{14} X_1 + \lambda_{24} X_2 + \lambda_{34} X_3 + \varepsilon_4 \\
X_5 &= \lambda_{15} X_1 + \lambda_{45} X_4 + \varepsilon_5 .
\end{aligned}
$$

For a general directed acyclic graph $G = (V, E)$, the *linear structural equation model corresponding to* G consists of the the graph G and the linear equations

$$
X_i = \sum_{j \in \mathrm{pa}(i)} \lambda_{ji} X_j + \varepsilon_i, \qquad \text{where the variables } \{\varepsilon_i\}_{i \in V} \text{ are independent.}
$$

In matrix-vector form

$$
X = \Lambda^T X + \varepsilon.
$$

Equivalently,

$$
X = (I - \Lambda)^{-T} \varepsilon.
$$

Linear Gaussian models

$$X_i = \sum_{j \in \mathsf{pa}(i)} \lambda_{ji} X_j + \epsilon_i, \qquad \text{where } \epsilon \sim \mathcal{N}(\nu, \Omega), \text{ and } \Omega = \mathsf{diag}(\omega_1, \ldots, \omega_n),$$

Linear Gaussian models

$$X_i = \sum_{j \in \mathsf{pa}(i)} \lambda_{ji} X_j + \epsilon_i, \qquad \text{where } \epsilon \sim \mathcal{N}(\nu, \Omega), \text{ and } \Omega = \mathsf{diag}(\omega_1, \ldots, \omega_n),$$

$$X = (I - \Lambda)^{-T} \epsilon.$$

Linear Gaussian models

$$X_i = \sum_{j \in \mathrm{pa}(i)} \lambda_{ji} X_j + \epsilon_i, \qquad \text{where } \epsilon \sim \mathcal{N}(\nu, \Omega), \text{ and } \Omega = \mathrm{diag}(\omega_1, \ldots, \omega_n),$$

$$X = (I - \Lambda)^{-T} \epsilon.$$

Thus, $X \sim \mathcal{N}(\mu, \Sigma)$, where

$$\Sigma = (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}.$$

The set of distributions $\mathcal{M}_G$ arising from a Gaussian linear causal model with DAG $G = (V, E)$ is called the **directed Gaussian graphical model corresponding to G**, and

$$\mathcal{M}_G = \{\Sigma \ : \ \Sigma = (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}, \ \Lambda \in \mathbb{R}^E, \Omega \succ 0 \text{ diagonal}\}.$$

Linear Gaussian models

$$X_i = \sum_{j \in \mathrm{pa}(i)} \lambda_{ji} X_j + \epsilon_i, \qquad \text{where } \epsilon \sim \mathcal{N}(\nu, \Omega), \text{ and } \Omega = \mathrm{diag}(\omega_1, \ldots, \omega_n),$$

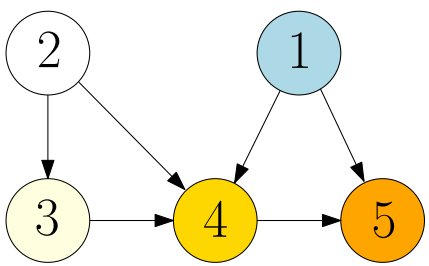
$$X = (I - \Lambda)^{-T} \epsilon.$$

Thus, $X \sim \mathcal{N}(\mu, \Sigma)$, where

$$\Sigma = (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}.$$

The set of distributions $\mathcal{M}_G$ arising from a Gaussian linear causal model with DAG $G = (V, E)$ is called the **directed Gaussian graphical model corresponding to G**, and

$$\mathcal{M}_G = \{\Sigma \ : \ \Sigma = (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}, \ \Lambda \in \mathbb{R}^E, \Omega \succ 0 \text{ diagonal}\}.$$



$$\Sigma = (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}, \text{ where}$$

$$\Lambda = \begin{pmatrix} 0 & 0 & 0 & \lambda_{14} & \lambda_{15} \\ 0 & 0 & \lambda_{23} & \lambda_{24} & 0 \\ 0 & 0 & 0 & \lambda_{34} & 0 \\ 0 & 0 & 0 & 0 & \lambda_{45} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \ \Omega = \begin{pmatrix} \omega_{11} & 0 & 0 & 0 & 0 \\ 0 & \omega_{22} & 0 & 0 & 0 \\ 0 & 0 & \omega_{33} & 0 & 0 \\ 0 & 0 & 0 & \omega_{44} & 0 \\ 0 & 0 & 0 & 0 & \omega_{55} \end{pmatrix}.$$

Linear Gaussian models

$$X_i = \sum_{j \in \mathrm{pa}(i)} \lambda_{ji} X_j + \epsilon_i, \qquad \text{where } \epsilon \sim \mathcal{N}(\nu, \Omega), \text{ and } \Omega = \mathrm{diag}(\omega_1, \ldots, \omega_n),$$

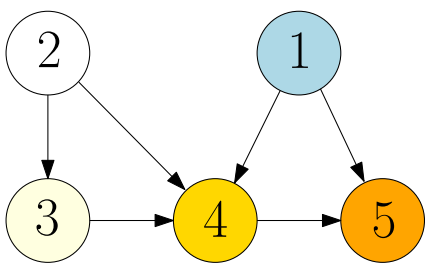
$$X = (I - \Lambda)^{-T} \epsilon.$$

Thus, $X \sim \mathcal{N}(\mu, \Sigma)$, where

$$\Sigma = (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}.$$

The set of distributions $\mathcal{M}_G$ arising from a Gaussian linear causal model with DAG $G = (V, E)$ is called the **directed Gaussian graphical model corresponding to G**, and

$$\mathcal{M}_G = \{\Sigma \ : \ \Sigma = (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}, \ \Lambda \in \mathbb{R}^E, \Omega \succ 0 \text{ diagonal}\}.$$



$$\Sigma = (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1}, \text{ where}$$

$$\Lambda = \begin{pmatrix} 0 & 0 & 0 & \lambda_{14} & \lambda_{15} \\ 0 & 0 & \lambda_{23} & \lambda_{24} & 0 \\ 0 & 0 & 0 & \lambda_{34} & 0 \\ 0 & 0 & 0 & 0 & \lambda_{45} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \ \Omega = \begin{pmatrix} \omega_{11} & 0 & 0 & 0 & 0 \\ 0 & \omega_{22} & 0 & 0 & 0 \\ 0 & 0 & \omega_{33} & 0 & 0 \\ 0 & 0 & 0 & \omega_{44} & 0 \\ 0 & 0 & 0 & 0 & \omega_{55} \end{pmatrix}.$$

- Markov equivalence: $\mathcal{M}_G = \mathcal{M}_H \implies$ cannot identify the graph G uniquely.

Non-Gaussian Linear Structural Equation Models

Given a DAG $G = (V, E)$,

$$X = (I - \Lambda)^{-T}\varepsilon, \quad \text{where the } \{\varepsilon_v\}_{v \in V} \text{ are independent and } \Lambda \in \mathbb{R}^E.$$

Non-Gaussian Linear Structural Equation Models

Given a DAG $G = (V, E)$,

$$X = (I - \Lambda)^{-T}\varepsilon, \quad \text{where the } \{\varepsilon_v\}_{v \in V} \text{ are independent and } \Lambda \in \mathbb{R}^E.$$

- **Independent component analysis:** Given $X = A\varepsilon$, where ε is a vector of independent components, want to recover A,
  ... up to permutation and scaling of its columns.

Non-Gaussian Linear Structural Equation Models

Given a DAG $G = (V, E)$,

$$X = (I - \Lambda)^{-T}\varepsilon, \quad \text{where the } \{\varepsilon_v\}_{v \in V} \text{ are independent and } \Lambda \in \mathbb{R}^E.$$

- **Independent component analysis:** Given $X = A\varepsilon$, where ε is a vector of independent components, want to recover A,
  ... up to permutation and scaling of its columns.
- If at least two ε_j are Gaussian, then recovering A uniquely is impossible.

Non-Gaussian Linear Structural Equation Models

Given a DAG $G = (V, E)$,

$$X = (I - \Lambda)^{-T}\varepsilon, \quad \text{where the } \{\varepsilon_v\}_{v \in V} \text{ are independent and } \Lambda \in \mathbb{R}^E.$$

- **Independent component analysis:** Given $X = A\varepsilon$, where ε is a vector of independent components, want to recover A,
  ... up to permutation and scaling of its columns.
- If at least two ε_j are Gaussian, then recovering A uniquely is impossible.

Theorem (Comon and Jutten, *Handbook of Blind Source Separation*, 2010)

If all (or all but one) ε_j are non-Gaussian, A can be recovered (up to permutation and scaling).

Non-Gaussian Linear Structural Equation Models

Given a DAG $G = (V, E)$,

$$X = (I - \Lambda)^{-T}\varepsilon, \quad \text{where the } \{\varepsilon_v\}_{v \in V} \text{ are independent and } \Lambda \in \mathbb{R}^E.$$

- **Independent component analysis:** Given $X = A\varepsilon$, where ε is a vector of independent components, want to recover A,
  ... up to permutation and scaling of its columns.
- If at least two ε_j are Gaussian, then recovering A uniquely is impossible.

Theorem (Comon and Jutten, *Handbook of Blind Source Separation*, 2010)

If all (or all but one) ε_j are non-Gaussian, A can be recovered (up to permutation and scaling).

- ICA Methods: maximum likelihood estimation, 4th order cumulant tensor decomposition, maximizing |kurtosis| of $A^{-1}X$ (a measure of non-Gaussianity)

Linear Non-Gaussian Acyclic Models (LiNGAM)

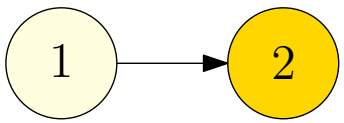
$$X = (I - \Lambda)^{-T} \varepsilon.$$

- Shimizu et al., 2006: LiNGAM; use ICA methods; estimate of $(I - \Lambda)$ has all entries non-zero
- Shimizu et al., 2011: Direct-LiNGAM; a source node is independent from regression residuals; does not work if #observations $<$ #variables (high-dimensions)
- Wang and Drton, 2018: High-dimensional algorithm, exploits *relationships between second and higher order moments of* X

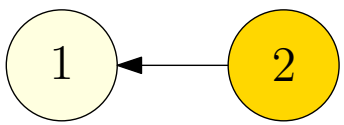
Linear Non-Gaussian Acyclic Models (LiNGAM)

$$X = (I - \Lambda)^{-T}\varepsilon.$$

- Shimizu et al., 2006: LiNGAM; use ICA methods; estimate of $(I - \Lambda)$ has all entries non-zero
- Shimizu et al., 2011: Direct-LiNGAM; a source node is independent from regression residuals; does not work if #observations $<$ #variables (high-dimensions)
- Wang and Drton, 2018: High-dimensional algorithm, exploits *relationships between second and higher order moments of X*



$$\mathbb{E}[X_1X_2]\mathbb{E}[X_1^3] - \mathbb{E}[X_1^2]\mathbb{E}[X_1^2X_2] = 0.$$



$$\mathbb{E}[X_1X_2]\mathbb{E}[X_1^3] - \mathbb{E}[X_1^2]\mathbb{E}[X_1^2X_2] \neq 0$$

generically, in particular, third order moments need to be non-Gaussian.

Looking at higher moments

$$X = (I - \Lambda)^{-T}\varepsilon.$$

Definition

The *linear structural equation model* $\mathcal{M}^{(2,3)}(G)$ of second and third order moments corresponding to a DAG $G = (V, E)$ with $|V| = n$ is defined as

$$\begin{aligned}\mathcal{M}^{(2,3)}(G) = \{(S = (I - \Lambda)^{-T}\Omega^{(2)}(I - \Lambda)^{-1}, \\ T = \Omega^{(3)} \bullet (I - \Lambda)^{-1} \bullet (I - \Lambda)^{-1} \bullet (I - \Lambda)^{-1}) : \\ \Omega^{(2)} \text{ is } n \times n \text{ positive definite diagonal matrix}, \\ \Omega^{(3)} \text{ is } n \times n \times n \text{ diagonal 3-way tensor, and } \Lambda \in \mathbb{R}^E\}.\end{aligned}$$

Here, $\bullet$ denotes the *Tucker product*.

Theorem (Améndola, Drton, Grosdos, Homs-Pons, and R., 2021+)

The set of second and third order moments (T, S) of a linear non-Gaussian causal model corresponding to a tree DAG are precisely the ones that satisfy certain quadratic binomials which arise as the 2×2 minors of certain matrices constructed from the DAG.

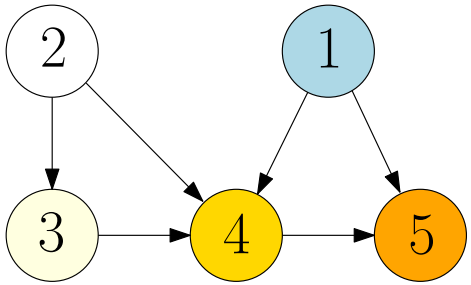
- $s_{ij} = 0$ for all $i, j \in V$ for which there is no 2-trek between i and j;
- $t_{ijk} = 0$ for all $i, j, k \in V$ for which there is no 2-trek between i, j, k;
- the 2×2 minors of the matrix A_{ij} are 0 whenever there is a path from i to j, where

$$A_{ij} = \begin{bmatrix} s_{ik_1} & \cdots & s_{ik_r} & t_{i\ell_1 m_1} & \cdots & t_{i\ell_q m_q} \\ s_{jk_1} & \cdots & s_{jk_r} & t_{j\ell_1 m_1} & \cdots & t_{j\ell_q m_q} \end{bmatrix},$$

where

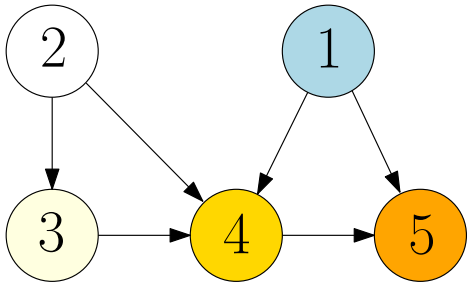
  - $k_1, \ldots, k_r$ are all vertices such that $\text{top}(i, k_a) = \text{top}(j, k_a)$ and
  - $(l_1, m_1), \ldots, (l_q, m_q)$ are all pairs of vertices such that $\text{top}(i, l_b, m_b) = \text{top}(j, l_b, m_b)$.

Introducing hidden variables

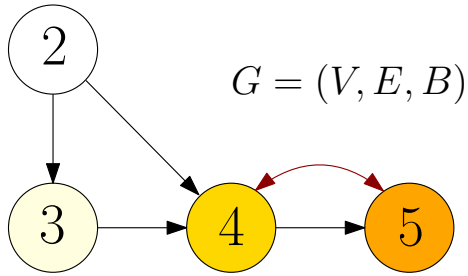


$$X_1 = \epsilon_1$$
$$X_2 = \epsilon_2$$
$$X_3 = \lambda_{23} X_2 + \epsilon_3$$
$$X_4 = \lambda_{14} X_1 + \lambda_{24} X_2 + \lambda_{34} X_3 + \epsilon_4$$
$$X_5 = \lambda_{15} X_1 + \lambda_{45} X_4 + \epsilon_5$$

Introducing hidden variables

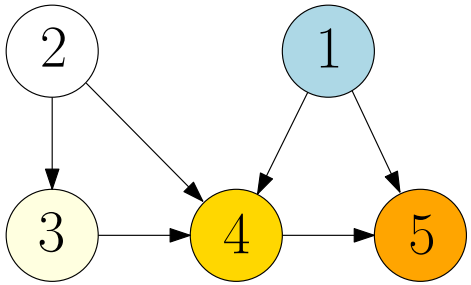


$$X_1 = \epsilon_1$$
$$X_2 = \epsilon_2$$
$$X_3 = \lambda_{23} X_2 + \epsilon_3$$
$$X_4 = \lambda_{14} X_1 + \lambda_{24} X_2 + \lambda_{34} X_3 + \epsilon_4$$
$$X_5 = \lambda_{15} X_1 + \lambda_{45} X_4 + \epsilon_5$$



$$X_2 = \epsilon_2$$
$$X_3 = \lambda_{23} X_2 + \epsilon_3$$
$$X_4 = \lambda_{24} X_2 + \lambda_{34} X_3 + \tilde{\epsilon}_4$$
$$X_5 = \lambda_{45} X_4 + \tilde{\epsilon}_5$$

Introducing hidden variables

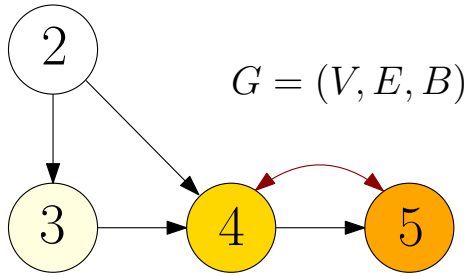


$$
\begin{aligned}
X_1 &= \epsilon_1 \\
X_2 &= \epsilon_2 \\
X_3 &= \lambda_{23} X_2 + \epsilon_3 \\
X_4 &= \lambda_{14} X_1 + \lambda_{24} X_2 + \lambda_{34} X_3 + \epsilon_4 \\
X_5 &= \lambda_{15} X_1 + \lambda_{45} X_4 + \epsilon_5
\end{aligned}
$$

$$
\Sigma = (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1},
$$

$$
\text{where } \Lambda = \begin{pmatrix} 0 & 0 & 0 & \lambda_{14} & \lambda_{15} \\ 0 & 0 & \lambda_{23} & \lambda_{24} & 0 \\ 0 & 0 & 0 & \lambda_{34} & 0 \\ 0 & 0 & 0 & 0 & \lambda_{45} \\ 0 & 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}^E,
$$

$$
\Omega = \begin{pmatrix} \omega_{11} & 0 & 0 & 0 & 0 \\ 0 & \omega_{22} & 0 & 0 & 0 \\ 0 & 0 & \omega_{33} & 0 & 0 \\ 0 & 0 & 0 & \omega_{44} & 0 \\ 0 & 0 & 0 & 0 & \omega_{55} \end{pmatrix} \in \text{PD}.
$$



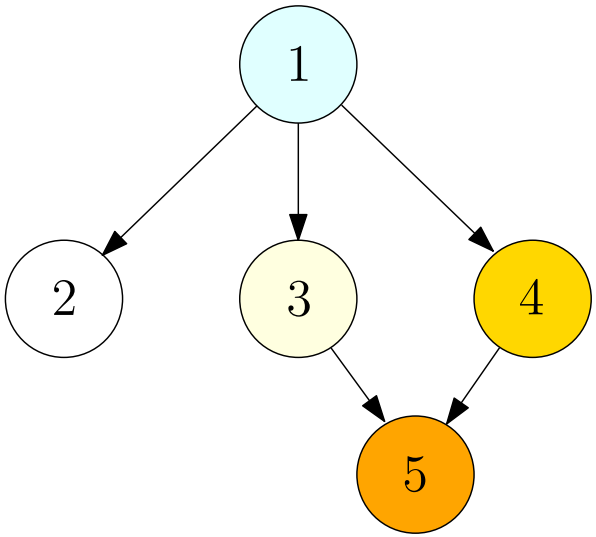
$$
\begin{aligned}
X_2 &= \epsilon_2 \\
X_3 &= \lambda_{23} X_2 + \epsilon_3 \\
X_4 &= \lambda_{24} X_2 + \lambda_{34} X_3 + \tilde{\epsilon}_4 \\
X_5 &= \lambda_{45} X_4 + \tilde{\epsilon}_5
\end{aligned}
$$

$$
\Sigma = (I - \Lambda)^{-T} \Omega (I - \Lambda)^{-1},
$$

$$
\text{where } \Lambda = \begin{pmatrix} 0 & \lambda_{23} & \lambda_{24} & 0 \\ 0 & 0 & \lambda_{34} & 0 \\ 0 & 0 & 0 & \lambda_{45} \\ 0 & 0 & 0 & 0 \end{pmatrix} \in \mathbb{R}^E,
$$

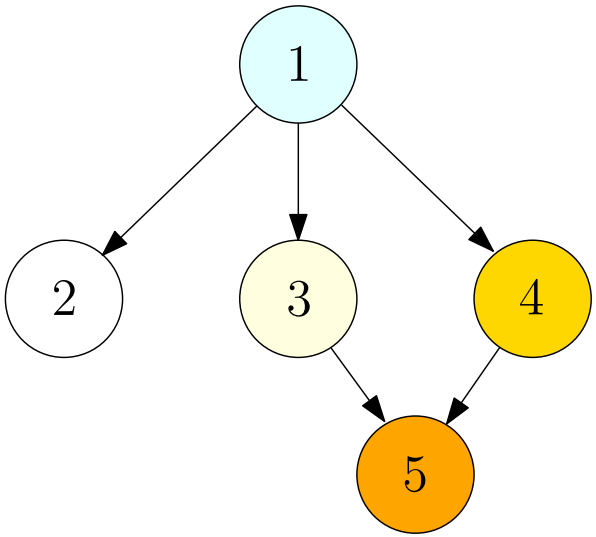
$$
\Omega = \begin{pmatrix} \omega_{22} & 0 & 0 & 0 \\ 0 & \omega_{33} & 0 & 0 \\ 0 & 0 & \omega_{44} & \omega_{45} \\ 0 & 0 & \omega_{45} & \omega_{55} \end{pmatrix} \in \text{PD}^B.
$$

Introducing hidden variables

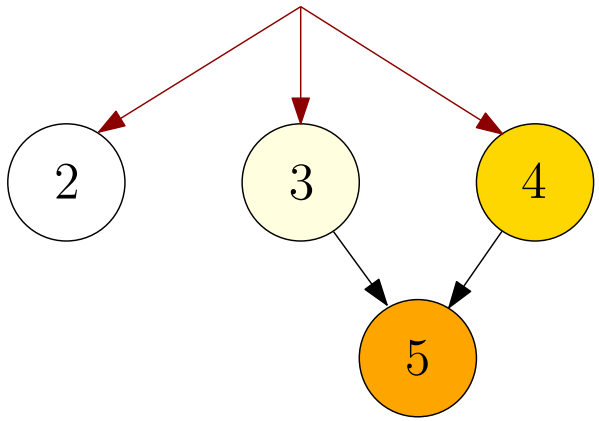


$$X_1 = \varepsilon_1$$
$$X_2 = \lambda_{12} X_1 + \varepsilon_2$$
$$X_3 = \lambda_{13} X_1 + \varepsilon_3$$
$$X_4 = \lambda_{14} X_1 + \varepsilon_4$$
$$X_5 = \lambda_{35} X_3 + \lambda_{45} X_4 + \varepsilon_5.$$

Introducing hidden variables



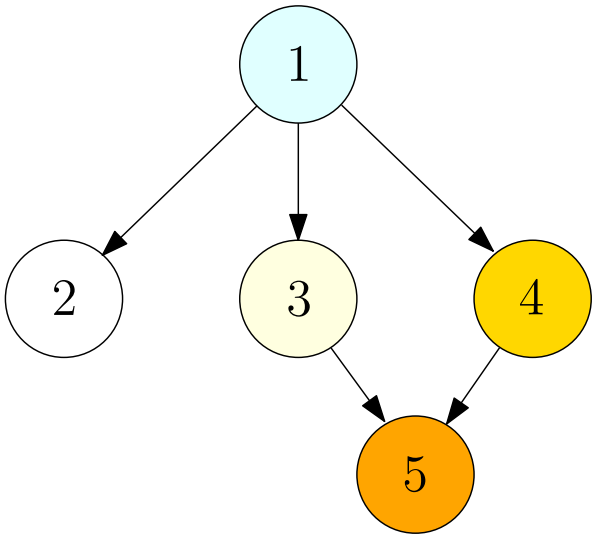
$$X_1 = \varepsilon_1$$
$$X_2 = \lambda_{12} X_1 + \varepsilon_2$$
$$X_3 = \lambda_{13} X_1 + \varepsilon_3$$
$$X_4 = \lambda_{14} X_1 + \varepsilon_4$$
$$X_5 = \lambda_{35} X_3 + \lambda_{45} X_4 + \varepsilon_5.$$

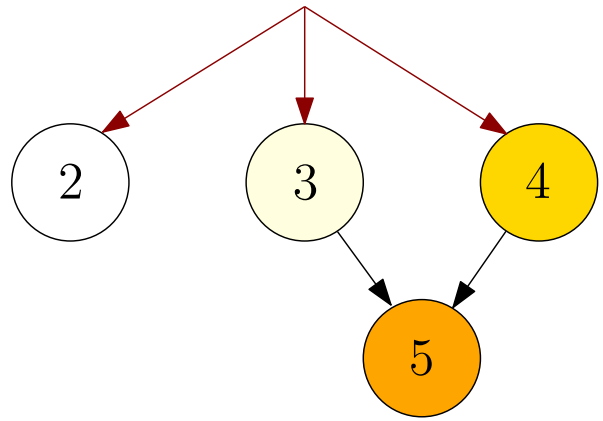


$$X_2 = \tilde{\varepsilon}_2$$
$$X_3 = \tilde{\varepsilon}_3$$
$$X_4 = \tilde{\varepsilon}_4$$
$$X_5 = \lambda_{35} X_3 + \lambda_{45} X_4 + \varepsilon_5,$$

where $\varepsilon_5 \perp\!\!\!\perp \tilde{\varepsilon}_2, \tilde{\varepsilon}_3, \tilde{\varepsilon}_4$.

Introducing hidden variables





$$X_1 = \varepsilon_1$$
$$X_2 = \lambda_{12} X_1 + \varepsilon_2$$
$$X_3 = \lambda_{13} X_1 + \varepsilon_3$$
$$X_4 = \lambda_{14} X_1 + \varepsilon_4$$
$$X_5 = \lambda_{35} X_3 + \lambda_{45} X_4 + \varepsilon_5.$$

$$X_2 = \tilde{\varepsilon}_2$$
$$X_3 = \tilde{\varepsilon}_3$$
$$X_4 = \tilde{\varepsilon}_4$$
$$X_5 = \lambda_{35} X_3 + \lambda_{45} X_4 + \varepsilon_5,$$

where $\varepsilon_5 \perp\!\!\!\perp \tilde{\varepsilon}_2, \tilde{\varepsilon}_3, \tilde{\varepsilon}_4$.

The new graph $G = (V, E, H)$ has directed edges E and *multi-directed edges* H.

Learning LiNGAMs with hidden variables from observational data

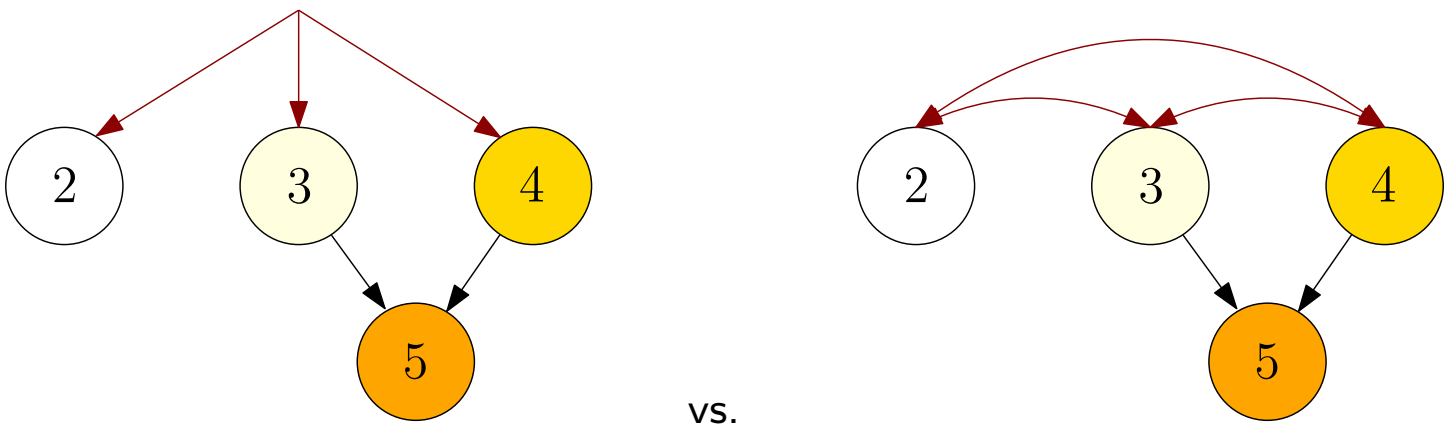
Existing methods for learning $G = (V, E, H)$ either

- Use ICA methods (Hoyer et al., 2008), which don't guarantee convergence to a global optimum, OR
- Only learn a graph $G = (V, E, B)$ with directed and *bidirected* edges (ParcelLiNGAM, Tashiro et al., 2014, Wang and Drton, 2020).

Learning LiNGAMs with hidden variables from observational data

Existing methods for learning $G = (V, E, H)$ either

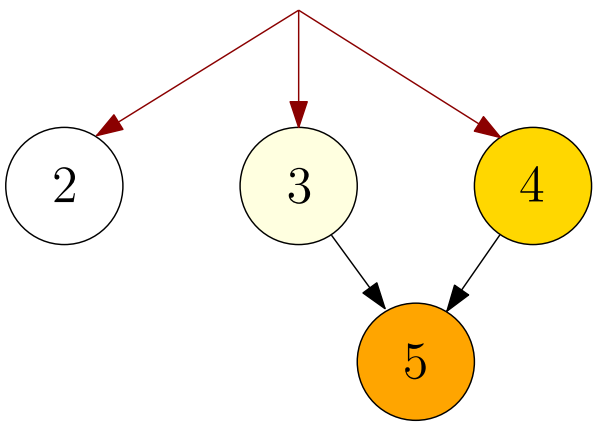
- Use ICA methods (Hoyer et al., 2008), which don't guarantee convergence to a global optimum, OR
- Only learn a graph $G = (V, E, B)$ with directed and *bidirected* edges (ParcelLiNGAM, Tashiro et al., 2014, Wang and Drton, 2020).



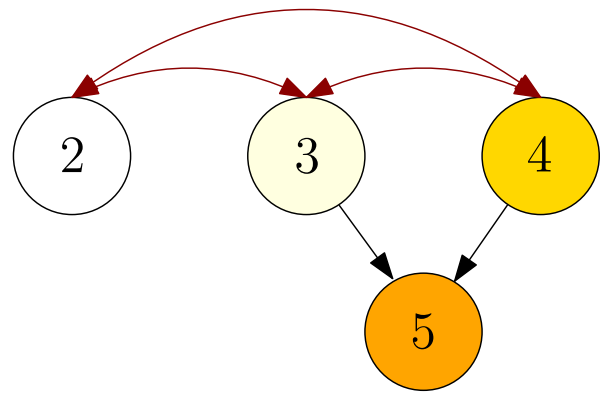
Learning LiNGAMs with hidden variables from observational data

Existing methods for learning $G = (V, E, H)$ either

- Use ICA methods (Hoyer et al., 2008), which don't guarantee convergence to a global optimum, OR
- Only learn a graph $G = (V, E, B)$ with directed and *bidirected* edges (ParcelLiNGAM, Tashiro et al., 2014, Wang and Drton, 2020).



vs.



- (Liu, Robeva, Wang, 2020): Learn $G = (V, E, H)$, where H has multidirected edges; G is a bow-free acyclic graph; use high-order cumulant information (Robeva, Seby, 2020)

Vanishing of cumulants

- For a zero-mean random vector $X = (X_1, \ldots, X_d)$, its *k-th order cumulant* is an $d \times \cdots \times d$ (k times) tensor $C^{(k)}$ whose entries can be obtained from the moments of X, e.g. for $k = 4$:

$$C^{(4)}_{i_1,i_2,i_3,i_4} = \mathbb{E}[X_{i_1}X_{i_2}X_{i_3}X_{i_4}] - \mathbb{E}[X_{i_1}X_{i_2}]\mathbb{E}[X_{i_3}X_{i_4}] - \mathbb{E}[X_{i_1}X_{i_3}]\mathbb{E}[X_{i_2}X_{i_4}] - \mathbb{E}[X_{i_1}X_{i_4}]\mathbb{E}[X_{i_2}X_{i_3}].$$

Vanishing of cumulants

- For a zero-mean random vector $X = (X_1, \ldots, X_d)$, its *k-th order cumulant* is an $d \times \cdots \times d$ (k times) tensor $C^{(k)}$ whose entries can be obtained from the moments of X, e.g. for $k = 4$:

$$C^{(4)}_{i_1,i_2,i_3,i_4} = \mathbb{E}[X_{i_1}X_{i_2}X_{i_3}X_{i_4}] - \mathbb{E}[X_{i_1}X_{i_2}]\mathbb{E}[X_{i_3}X_{i_4}] - \mathbb{E}[X_{i_1}X_{i_3}]\mathbb{E}[X_{i_2}X_{i_4}] - \mathbb{E}[X_{i_1}X_{i_4}]\mathbb{E}[X_{i_2}X_{i_3}].$$

Theorem (Robeva and Seby, 2020)

If X comes from a linear non-Gaussian acyclic model with graph $G = (V, E, H)$ and X has cumulants $C^{(k)}$, then

$$C^{(k)}_{i_1,\ldots,i_k} = 0$$

if and only if there is no k-trek between the vertices $i_1, \ldots, i_k$ in G.

Vanishing of cumulants

- For a zero-mean random vector $X = (X_1, \ldots, X_d)$, its *k-th order cumulant* is an $d \times \cdots \times d$ (k times) tensor $C^{(k)}$ whose entries can be obtained from the moments of X, e.g. for $k = 4$:

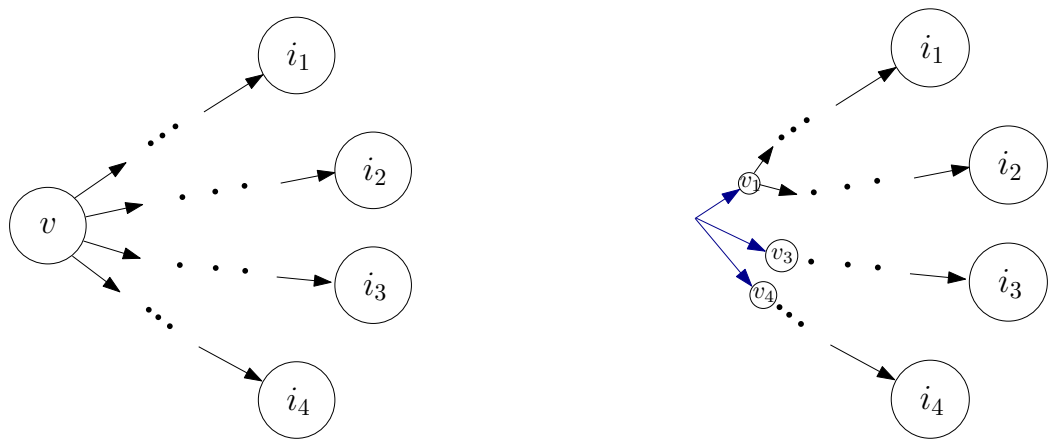
$$C^{(4)}_{i_1,i_2,i_3,i_4} = \mathbb{E}[X_{i_1}X_{i_2}X_{i_3}X_{i_4}] - \mathbb{E}[X_{i_1}X_{i_2}]\mathbb{E}[X_{i_3}X_{i_4}] - \mathbb{E}[X_{i_1}X_{i_3}]\mathbb{E}[X_{i_2}X_{i_4}] - \mathbb{E}[X_{i_1}X_{i_4}]\mathbb{E}[X_{i_2}X_{i_3}].$$

Theorem (Robeva and Seby, 2020)

If X comes from a linear non-Gaussian acyclic model with graph $G = (V, E, H)$ and X has cumulants $C^{(k)}$, then

$$C^{(k)}_{i_1,\ldots,i_k} = 0$$

if and only if there is no k-trek between the vertices $i_1, \ldots, i_k$ in G.



k-treks

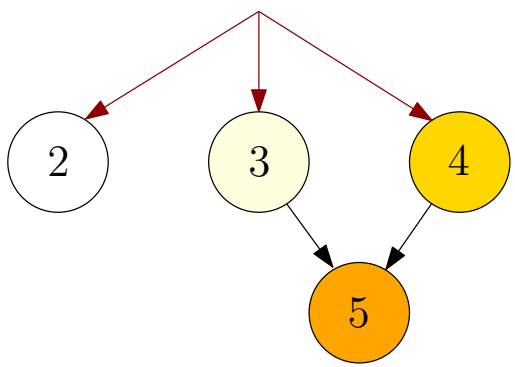
Theorem (Robeva and Seby, 2020)

If X comes from a linear non-Gaussian causal model with graph $G = (V, E, H)$ and X has cumulants $C^{(k)}$, then

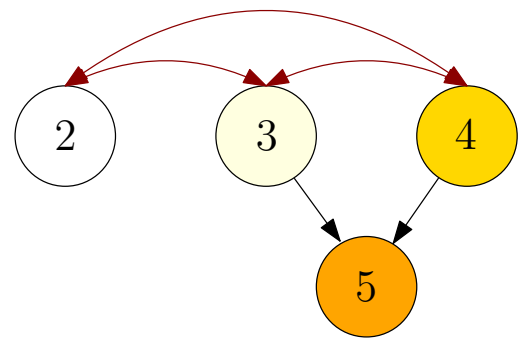
$$C^{(k)}_{i_1,\ldots,i_k} = 0$$

if and only if there is no k-trek between the vertices $i_1, \ldots, i_k$ in G.

- Thus, we can distinguish:



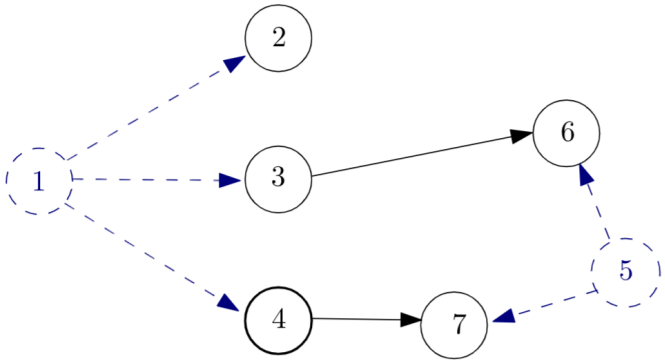
vs.



$C^{(3)}_{2,3,4} \neq 0$ $\qquad$ $C^{(3)}_{2,3,4} = 0.$

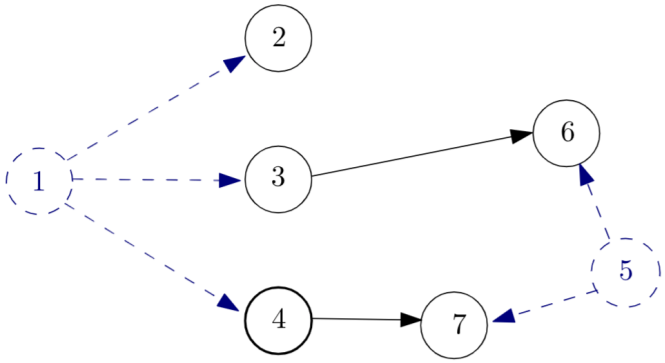
- Get and algorithm to learn $G = (V, E, H)$ based on high-order cumulants.

Learning $G = (V, E, H)$ [Liu, Robeva, Wang, 2020]

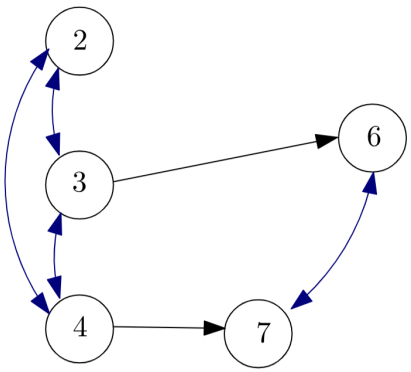


1. Obtain samples $Y = (Y^{(1)}, \ldots, Y^{(N)})$ from LiNGAM with unknown $G = (V, E, H)$

Learning $G = (V, E, H)$ [Liu, Robeva, Wang, 2020]

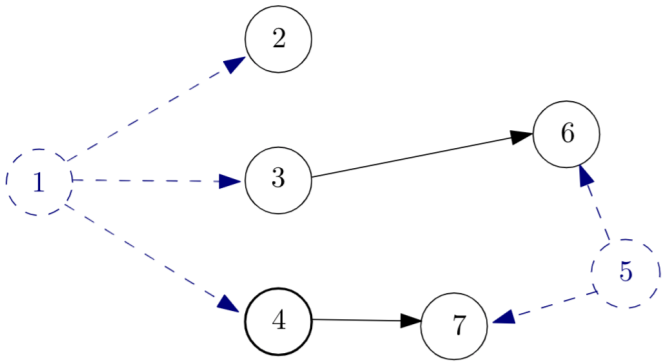


1. Obtain samples $Y = (Y^{(1)}, \ldots, Y^{(N)})$ from LiNGAM with unknown $G = (V, E, H)$

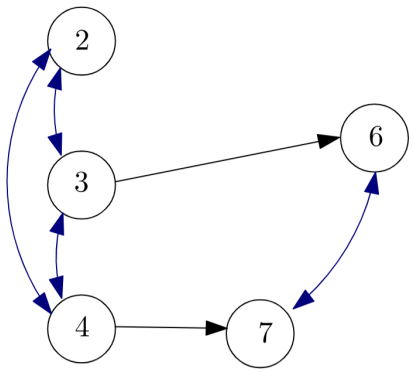


2. Learn a LiNGAM with graph (V, E, B) with bidirected edges and coefficient matrix Λ, e.g. using [Wang and Drton, 2020]

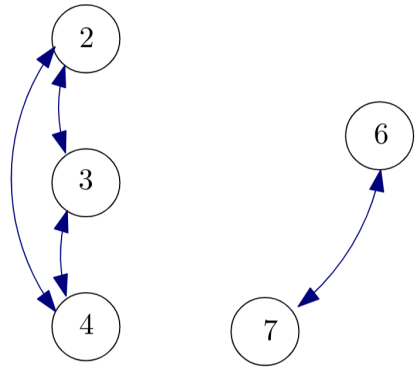
Learning $G = (V, E, H)$ [Liu, Robeva, Wang, 2020]



1. Obtain samples $Y = (Y^{(1)}, \ldots, Y^{(N)})$ from LiNGAM with unknown $G = (V, E, H)$

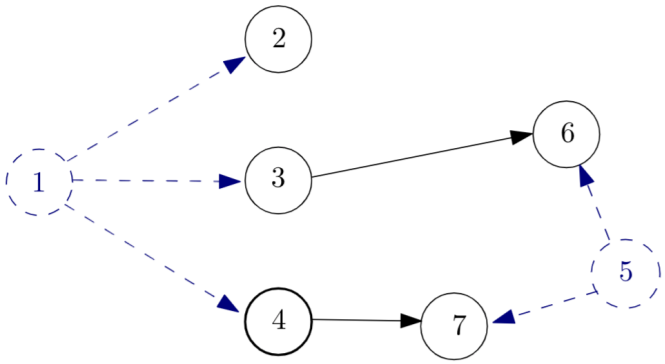


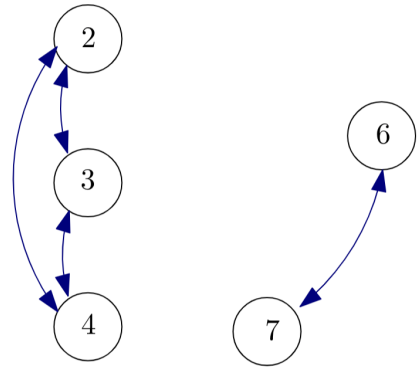
2. Learn a LiNGAM with graph (V, E, B) with bidirected edges and coefficient matrix Λ, e.g. using [Wang and Drton, 2020]



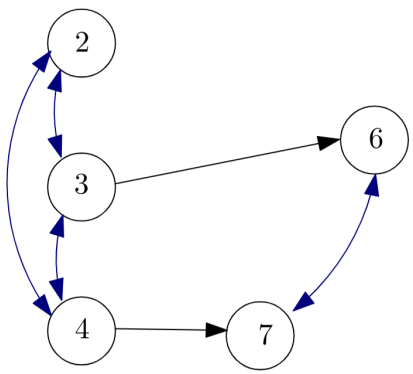
3. "Remove" directed edges E via $X = Y - \Lambda^T Y$.

Learning $G = (V, E, H)$ [Liu, Robeva, Wang, 2020]



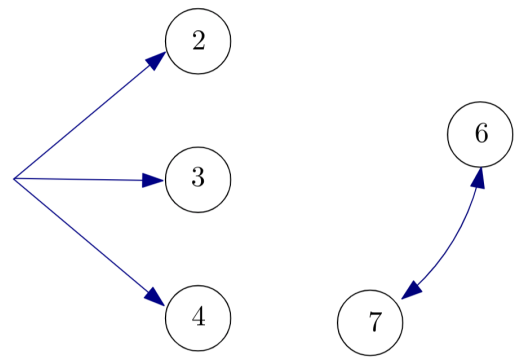


1. Obtain samples $Y = (Y^{(1)}, \ldots, Y^{(N)})$ from LiNGAM with unknown $G = (V, E, H)$



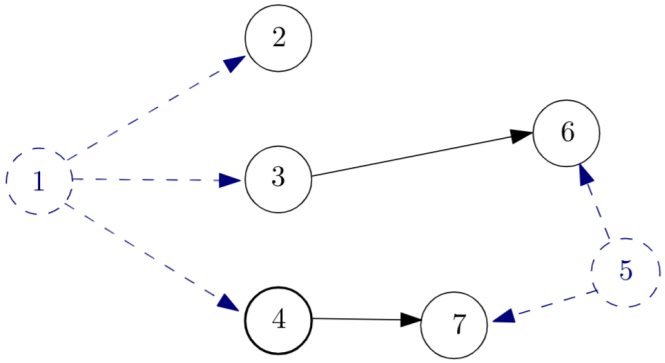
2. Learn a LiNGAM with graph (V, E, B) with bidirected edges and coefficient matrix Λ, e.g. using [Wang and Drton, 2020]

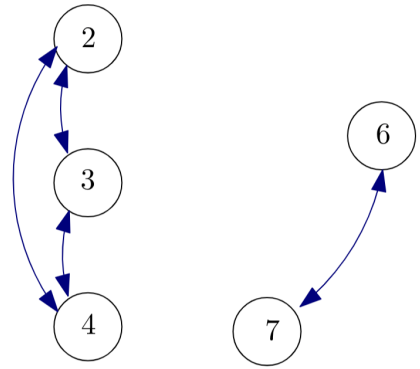
3. "Remove" directed edges E via $X = Y - \Lambda^T Y$.



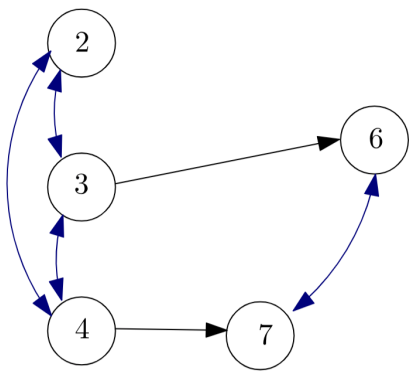
4. Identify the multidirected edges H by "merging" some of the bidirected edges in graph $(V, \emptyset, B)$ using the cumulants of X.

Learning $G = (V, E, H)$ [Liu, Robeva, Wang, 2020]



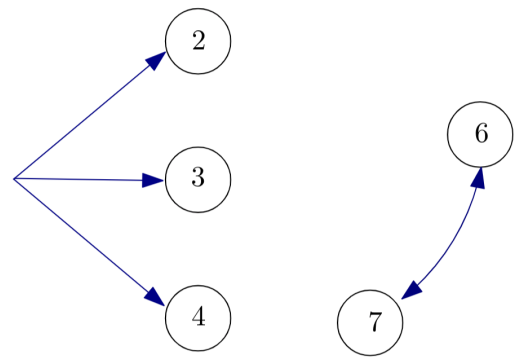


1. Obtain samples $Y = (Y^{(1)}, \ldots, Y^{(N)})$ from LiNGAM with unknown $G = (V, E, H)$



2. Learn a LiNGAM with graph (V, E, B) with bidirected edges and coefficient matrix Λ, e.g. using [Wang and Drton, 2020]

3. "Remove" directed edges E via $X = Y - \Lambda^T Y$.



4. Identify the multidirected edges H by "merging" some of the bidirected edges in graph $(V, \emptyset, B)$ using the cumulants of X.

5. Combine to obtain $G = (V, E, H)$.

More algebraic constraints

- (Robeva, Seby, 2020): Characterize vanishing of *determinants of subtensors* of k-th cumulant tensor $C^{(k)}$ in a LiNGAM with graph $G = (V, E, H)$;

$$\det(C^{(k)}_{A_1,\ldots,A_k}) = 0$$

if and only if every system of k-treks between $A_1, \ldots, A_k$ has a sided intersection.

Here:

$$\det(T) = \sum_{\sigma_2,\ldots,\sigma_k \in \mathfrak{S}(d)} \operatorname{sign}(\sigma_2) \cdots \operatorname{sign}(\sigma_k) \prod_{i=1}^{d} T_{i,\sigma_2(i),\ldots,\sigma_k(i)}$$

is the combinatorial hyperdeterminant.

More algebraic constraints

- (Robeva, Seby, 2020): Characterize vanishing of *determinants of subtensors* of k-th cumulant tensor $C^{(k)}$ in a LiNGAM with graph $G = (V, E, H)$;

$$\det(C^{(k)}_{A_1,\ldots,A_k}) = 0$$

if and only if every system of k-treks between $A_1, \ldots, A_k$ has a sided intersection.

Here:

$$\det(T) = \sum_{\sigma_2,\ldots,\sigma_k \in \mathfrak{S}(d)} \operatorname{sign}(\sigma_2) \cdots \operatorname{sign}(\sigma_k) \prod_{i=1}^{d} T_{i,\sigma_2(i),\ldots,\sigma_k(i)}$$

is the combinatorial hyperdeterminant.

- Can we learn such relationships in the case of both cycles and hidden variables?

Thank you!


C. Améndola, M. Drton, A. Grosdos, R. Homs-Pons, and E. Robeva. *Third-order moment varieties of non-Gaussian graphical models*. In preparation.


Y. Liu, E. Robeva, and H. Wang. *Learning Linear Non-Gaussian Graphical Models with Multidirected Edges*. Submitted (2020)


E. Robeva and J.B. Seby. *Multi-trek Separation in Linear Structural Equation Models*. SIAM Journal on Applied Algebra and Geometry (SIAGA) (2021)