# Causal Discovery in Linear Non-Gaussian Models

Mathias Drton

Department of Mathematics
Technical University of Munich

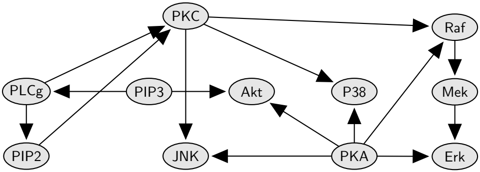
(joint work with Y. Samuel Wang)

# Causal Discovery from Observational Studies

Given: Multivariate i.i.d. sample: $Y^{(1)}, \dots Y^{(n)}$

Goal: Estimate underlying causal relationships. What is possible?

$$\begin{bmatrix} Y_1^{(1)} & \dots & Y_p^{(1)} \\ Y_1^{(2)} & \dots & Y_p^{(2)} \\ \vdots & \dots & \vdots \\ Y_1^{(n)} & \dots & Y_p^{(n)} \end{bmatrix} \quad \Longrightarrow$$



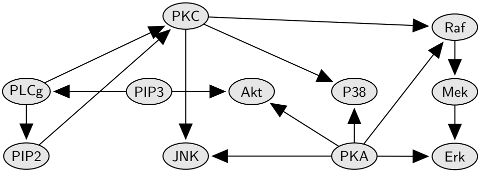

$$\text{PCK} = f(\text{PIP2}, \text{PLCg}, \varepsilon)$$

# Causal Discovery from Observational Studies

**Given:** Multivariate i.i.d. sample: $Y^{(1)}, \dots Y^{(n)}$

**Goal:** Estimate underlying causal relationships. What is possible?

$$\begin{bmatrix} Y_1^{(1)} & \dots & Y_p^{(1)} \\ Y_1^{(2)} & \dots & Y_p^{(2)} \\ \vdots & \dots & \vdots \\ Y_1^{(n)} & \dots & Y_p^{(n)} \end{bmatrix} \Longrightarrow$$




$$\text{PCK} = f(\text{PIP2}, \text{PLCg}, \varepsilon)$$

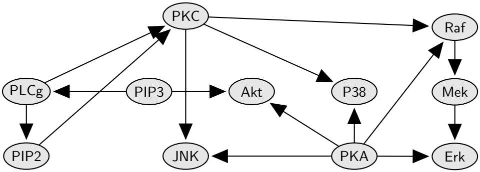
- General, Gaussian, discrete: Markov equivalence

# Causal Discovery from Observational Studies

Given: Multivariate i.i.d. sample: $Y^{(1)}, \dots Y^{(n)}$

Goal: Estimate underlying causal relationships. What is possible?

$$\begin{bmatrix} Y_1^{(1)} & \dots & Y_p^{(1)} \\ Y_1^{(2)} & \dots & Y_p^{(2)} \\ \vdots & \dots & \vdots \\ Y_1^{(n)} & \dots & Y_p^{(n)} \end{bmatrix} \implies$$




$$\text{PCK} = f(\text{PIP2}, \text{PLCg}, \varepsilon)$$

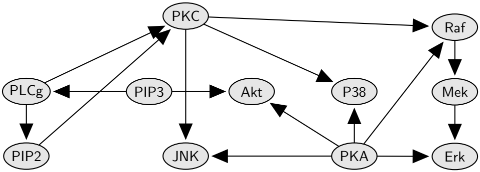
- General, Gaussian, discrete: Markov equivalence
- Unique causal graph under special assumptions such as:

# Causal Discovery from Observational Studies

Given: Multivariate i.i.d. sample: $Y^{(1)}, \ldots Y^{(n)}$

Goal: Estimate underlying causal relationships. What is possible?

$$\begin{bmatrix} Y_1^{(1)} & \ldots & Y_p^{(1)} \\ Y_1^{(2)} & \ldots & Y_p^{(2)} \\ \vdots & \ldots & \vdots \\ Y_1^{(n)} & \ldots & Y_p^{(n)} \end{bmatrix} \Longrightarrow$$




$$\text{PCK} = f(\text{PIP2}, \text{PLCg}, \varepsilon)$$

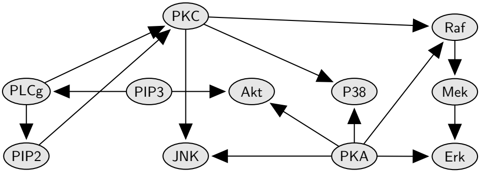
- General, Gaussian, discrete: Markov equivalence
- Unique causal graph under special assumptions such as:
  - Non-linear functional relationships with additive noise

# Causal Discovery from Observational Studies

Given: Multivariate i.i.d. sample: $Y^{(1)}, \ldots Y^{(n)}$

Goal: Estimate underlying causal relationships. What is possible?

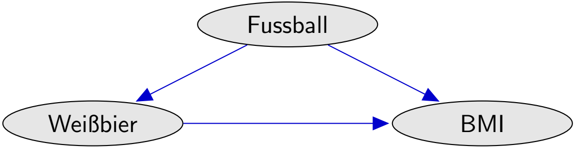
$$\begin{bmatrix} Y_1^{(1)} & \ldots & Y_p^{(1)} \\ Y_1^{(2)} & \ldots & Y_p^{(2)} \\ \vdots & \ldots & \vdots \\ Y_1^{(n)} & \ldots & Y_p^{(n)} \end{bmatrix} \Longrightarrow$$




$$\text{PCK} = f(\text{PIP2}, \text{PLCg}, \varepsilon)$$

- General, Gaussian, discrete: Markov equivalence
- Unique causal graph under special assumptions such as:
  - Non-linear functional relationships with additive noise
  - LiNGAM: Linear functional relationships with non-Gaussian errors (Shimizu, Hoyer, Hyvärinen, Kerminen, . . . )

# Causal Graphs




Directed Graph $G = (V, E_{\rightarrow})$ :

- Nodes correspond to observed variables.
- Edges represent direct causal effects.

Terminology:

- If $v \rightarrow u$, then $v$ is a parent of the child $u$.
- If $v \rightarrow \cdots \rightarrow u$, the $v$ is an ancestor of the descendant $u$.
- Ayclic digraph = directed acyclic graph = DAG

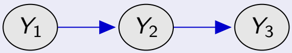
# LiNGAM (Linear Non-Gaussian Acyclic Model)

- Consider $p$-variate observation $Y = (Y_v)_{v \in V}$, so $|V| = p$.
- For convenience, assume $Y$ centered.
- Linear system given by a DAG:

$$Y_v = \sum_{u \in \mathrm{pa}(v)} \beta_{vu} Y_u + \varepsilon_v, \quad v \in V,$$

where the error terms $\varepsilon_v$ are independent and non-Gaussian.

**Example**




$$Y_1 = \varepsilon_1,$$
$$Y_2 = \beta_{21} Y_1 + \varepsilon_2,$$
$$Y_3 = \beta_{32} Y_2 + \varepsilon_3.$$

# Non-Gaussianity and Independent Component Analysis

- Let $\varepsilon$ be an $\mathbb{R}^p$-valued random vector with independent components:

$$\varepsilon_1 \perp\!\!\!\perp \varepsilon_2 \perp\!\!\!\perp \ldots \perp\!\!\!\perp \varepsilon_p.$$

# Non-Gaussianity and Independent Component Analysis

- Let $\varepsilon$ be an $\mathbb{R}^p$-valued random vector with independent components:

$$\varepsilon_1 \perp\!\!\!\perp \varepsilon_2 \perp\!\!\!\perp \ldots \perp\!\!\!\perp \varepsilon_p.$$

- ICA Problem (Independent Component Analysis):

  Given an invertible linear transformation $Y = A\varepsilon$, can we recover $A$?

  ... up to permutation and scaling of columns?

# Non-Gaussianity and Independent Component Analysis

- Let $\varepsilon$ be an $\mathbb{R}^p$-valued random vector with independent components:

$$\varepsilon_1 \perp\!\!\!\perp \varepsilon_2 \perp\!\!\!\perp \ldots \perp\!\!\!\perp \varepsilon_p.$$

- ICA Problem (Independent Component Analysis):

  Given an invertible linear transformation $Y = A\varepsilon$, can we recover $A$?

  ...up to permutation and scaling of columns?

- If at least two $\varepsilon_j$ are Gaussian then such recovery is impossible.

# Non-Gaussianity and Independent Component Analysis

- Let $\varepsilon$ be an $\mathbb{R}^p$-valued random vector with independent components:

$$\varepsilon_1 \perp\!\!\!\perp \varepsilon_2 \perp\!\!\!\perp \dots \perp\!\!\!\perp \varepsilon_p.$$

- ICA Problem (Independent Component Analysis):

  Given an invertible linear transformation $Y = A\varepsilon$, can we recover $A$?

  ... up to permutation and scaling of columns?

- If at least two $\varepsilon_j$ are Gaussian then such recovery is impossible.

**Theorem**

*If all (or all but one) $\varepsilon_j$ are non-Gaussian then $A$ can be recovered (up to permutation and scaling).*

# Non-Gaussianity and Independent Component Analysis

- Let $\varepsilon$ be an $\mathbb{R}^p$-valued random vector with independent components:

$$\varepsilon_1 \perp\!\!\!\perp \varepsilon_2 \perp\!\!\!\perp \ldots \perp\!\!\!\perp \varepsilon_p.$$

- ICA Problem (Independent Component Analysis):

  Given an invertible linear transformation $Y = A\varepsilon$, can we recover $A$?

  ...up to permutation and scaling of columns?

- If at least two $\varepsilon_j$ are Gaussian then such recovery is impossible.

**Theorem**

*If all (or all but one) $\varepsilon_j$ are non-Gaussian then $A$ can be recovered (up to permutation and scaling).*

- Practical implementations estimate $W = A^{-1}$ by maximizing "non-Gaussianity" of $WY$.

# ICA-LiNGAM (Shimizu et al., 2006)

- LiNGAM:

$$Y = B\,Y + \varepsilon \iff Y = (I - B)^{-1}\varepsilon$$

with $B$ supported over a DAG.

# ICA-LiNGAM (Shimizu et al., 2006)

- LiNGAM:

$$Y = B\,Y + \varepsilon \iff Y = (I - B)^{-1}\varepsilon$$

with $B$ supported over a DAG.

- ICA yields identifiability of $B$:

1. Find an unmixing/separating matrix $W$, which has to equal $I - B$ up to permutation and scaling of rows.
2. Permute rows of $W$ to have no zero diagonal elements (resolves "up to permutation" as $B$ corresponds to DAG).
3. Scale diagonal elements to unity (resolves "up to scaling").

# ICA-LiNGAM (Shimizu et al., 2006)

- LiNGAM:

$$Y = B\,Y + \varepsilon \iff Y = (I - B)^{-1}\varepsilon$$

  with $B$ supported over a DAG.

- ICA yields identifiability of $B$:

  1. Find an unmixing/separating matrix $W$, which has to equal $I - B$ up to permutation and scaling of rows.
  2. Permute rows of $W$ to have no zero diagonal elements (resolves "up to permutation" as $B$ corresponds to DAG).
  3. Scale diagonal elements to unity (resolves "up to scaling").

- Practice: feasible method but issues (e.g., $\hat{W}$ has all entries nonzero)

# Direct-LiNGAM (Shimizu et al., 2011)

Main Idea:

- Regression residuals are linear combination of the independent errors.
- Source node is characterized by independence from residuals.

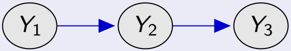
**Theorem (Darmois-Skitovitch)**

*Let $\varepsilon_1, \ldots, \varepsilon_p$ be independent non-degenerate random variables. If $\sum_j a_j \varepsilon_j \perp\!\!\!\perp \sum_j b_j \varepsilon_j$, then*

$$a_j b_j \neq 0 \quad \Longrightarrow \quad \varepsilon_j \sim \textit{Gaussian}.$$

# Direct-LiNGAM (Shimizu et al., 2011)

## Example

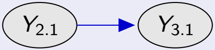



$$
\begin{aligned}
Y_1 &= \varepsilon_1, \\
Y_2 &= \beta_{21} Y_1 + \varepsilon_2, \\
Y_3 &= \beta_{32} Y_2 + \varepsilon_3.
\end{aligned}
$$

Residuals adjusting for $Y_1$ satisfy:

$$
\begin{aligned}
Y_{2.1} &:= Y_2 - \mathbb{E}(Y_2 \mid Y_1) = Y_2 - \beta_{21} Y_1 &&= \varepsilon_2, \\
Y_{3.1} &:= Y_3 - \mathbb{E}(Y_3 \mid Y_1) = Y_3 - \beta_{32}\beta_{21} Y_1 &&= \beta_{32} Y_{2.1} + \varepsilon_3.
\end{aligned}
$$

Observe that $Y_1 \perp\!\!\!\perp (Y_{2.1}, Y_{3.1})$ and



# Direct-LiNGAM Recursion

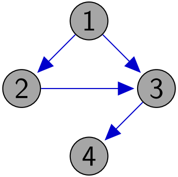
Let $\Theta^{(z)} = (r_1, r_2, \dots, r_z)$ be the set of ordered nodes after step $z$.

**Algorithm 1** Select an ordering

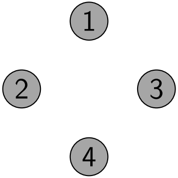
```
Θ^(0) = ∅; Y^(0) = Y
for z = 0, ..., p − 1 do
    Identify a source r ∉ Θ^(z) using Y^(z)
    for v ∉ Θ^(z) ∪ {r} do
        Y_v^(z+1) = Y_v^(z) − β̂_vr Y_r^(z)
    end for
    Update Θ^(z+1) = Append(Θ^(z), r).
end for
return Θ^(p) as an ordering
Prune ancestors which are not parents
```

$$\Theta^{(0)} = \emptyset$$

(a) "True" Graph of $Y^{(z)}$




(b) Estimated Graph



# Direct-LiNGAM Recursion

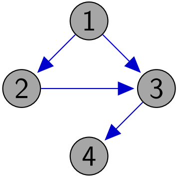
Let $\Theta^{(z)} = (r_1, r_2, \ldots, r_z)$ be the set of ordered nodes after step $z$.

**Algorithm 1** Select an ordering

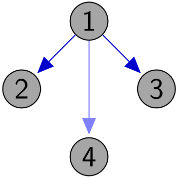
$\Theta^{(0)} = \emptyset$; $Y^{(0)} = Y$
**for** $z = 0, \ldots, p-1$ **do**
    Identify a source $r \notin \Theta^{(z)}$ using $Y^{(z)}$
    **for** $v \notin \Theta^{(z)} \cup \{r\}$ **do**
        $Y_v^{(z+1)} = Y_v^{(z)} - \hat{\beta}_{vr} Y_r^{(z)}$
    **end for**
    Update $\Theta^{(z+1)} = \text{Append}(\Theta^{(z)}, r)$.
**end for**
**return** $\Theta^{(p)}$ as an ordering
Prune ancestors which are not parents

$$\Theta^{(1)} = (1)$$

(a) “True” Graph of $Y^{(z)}$




(b) Estimated Graph



# Direct-LiNGAM Recursion

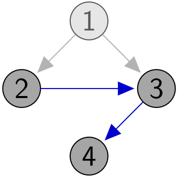
Let $\Theta^{(z)} = (r_1, r_2, \ldots, r_z)$ be the set of ordered nodes after step $z$.

**Algorithm 1** Select an ordering

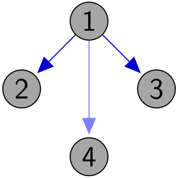
```
Θ^(0) = ∅; Y^(0) = Y
for z = 0, ..., p − 1 do
    Identify a source r ∉ Θ^(z) using Y^(z)
    for v ∉ Θ^(z) ∪ {r} do
        Y_v^(z+1) = Y_v^(z) − β̂_vr Y_r^(z)
    end for
    Update Θ^(z+1) = Append(Θ^(z), r).
end for
return Θ^(p) as an ordering
Prune ancestors which are not parents
```

$$\Theta^{(1)} = (1)$$

(a) “True” Graph of $Y^{(z)}$




(b) Estimated Graph



# Direct-LiNGAM Recursion

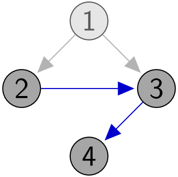
Let $\Theta^{(z)} = (r_1, r_2, \ldots, r_z)$ be the set of ordered nodes after step $z$.

**Algorithm 1** Select an ordering

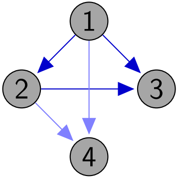
$\Theta^{(0)} = \emptyset$; $Y^{(0)} = Y$
**for** $z = 0, \ldots, p-1$ **do**
    Identify a source $r \notin \Theta^{(z)}$ using $Y^{(z)}$
    **for** $v \notin \Theta^{(z)} \cup \{r\}$ **do**
        $Y_v^{(z+1)} = Y_v^{(z)} - \hat{\beta}_{vr} Y_r^{(z)}$
    **end for**
    Update $\Theta^{(z+1)} = \text{Append}(\Theta^{(z)}, r)$.
**end for**
**return** $\Theta^{(p)}$ as an ordering
Prune ancestors which are not parents

$$\Theta^{(2)} = (1, 2)$$

(a) "True" Graph of $Y^{(z)}$




(b) Estimated Graph



# Direct-LiNGAM Recursion

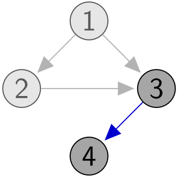
Let $\Theta^{(z)} = (r_1, r_2, \ldots, r_z)$ be the set of ordered nodes after step $z$.

**Algorithm 1** Select an ordering

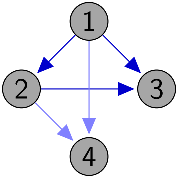
```
Θ^(0) = ∅; Y^(0) = Y
for z = 0, ..., p − 1 do
    Identify a source r ∉ Θ^(z) using Y^(z)
    for v ∉ Θ^(z) ∪ {r} do
        Y_v^(z+1) = Y_v^(z) − β̂_vr Y_r^(z)
    end for
    Update Θ^(z+1) = Append(Θ^(z), r).
end for
return Θ^(p) as an ordering
Prune ancestors which are not parents
```

$$\Theta^{(2)} = (1, 2)$$

(a) "True" Graph of $Y^{(z)}$




(b) Estimated Graph



# Direct-LiNGAM Recursion

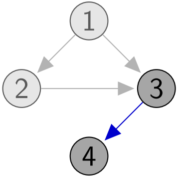
Let $\Theta^{(z)} = (r_1, r_2, \ldots, r_z)$ be the set of ordered nodes after step $z$.

**Algorithm 1** Select an ordering

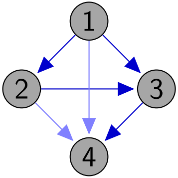
$\Theta^{(0)} = \emptyset$; $Y^{(0)} = Y$
**for** $z = 0, \ldots, p-1$ **do**
    Identify a source $r \notin \Theta^{(z)}$ using $Y^{(z)}$
    **for** $v \notin \Theta^{(z)} \cup \{r\}$ **do**
        $Y_v^{(z+1)} = Y_v^{(z)} - \hat{\beta}_{vr} Y_r^{(z)}$
    **end for**
    Update $\Theta^{(z+1)} = \text{Append}(\Theta^{(z)}, r)$.
**end for**
**return** $\Theta^{(p)}$ as an ordering
Prune ancestors which are not parents

$$\Theta^{(3)} = (1, 2, 3, 4)$$

(a) "True" Graph of $Y^{(z)}$




(b) Estimated Graph



# Two Problems

1. **High-dimensional DAGs**
   - Allow for #variables $= p > n =$ # observations.
   - Assuming sparsity.
   - Existing methods of Shimizu et al. (2006, 2011) and Hyvärinen and Smith (2013) not applicable.

2. **Latent variables (Bow-free Acyclic Path Diagrams)**
   - Allow for certain types of unobserved confounding
   - Existing methods involve difficult overcomplete ICA computations/require prior knowledge (Hoyer et al., 2008; Shimizu and Bollen, 2014) or may return inconclusive results (Entner and Hoyer, 2010; Tashiro et al., 2014)

Causal Discovery in High-Dimensional Settings

https://arxiv.org/abs/1803.11273

# Direct-LiNGAM Approach

- Problem in a high-dimensional setting:
  - Adjusting by all prior variables propagates error proportional to $p$.
  - Residuals are uninformative/zero if $p > n$.

# Direct-LiNGAM Approach

- Problem in a high-dimensional setting:
  - Adjusting by all prior variables propagates error proportional to $p$.
  - Residuals are uninformative/zero if $p > n$.

- Solution: Only adjust by smallest set necessary.

# Direct-LiNGAM Approach

- Problem in a high-dimensional setting:
  - Adjusting by all prior variables propagates error proportional to $p$.
  - Residuals are uninformative/zero if $p > n$.
- Solution: Only adjust by smallest set necessary.
- Need parameter/statistic to determine causal direction while adjusting for possible confounding.
- Selecting a source should be computationally inexpensive.

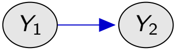
# Help from Non-Gaussianity? Looking at 3rd Moments...

Consider the polynomial: $\tau_{p \to c} = \mathbb{E}(Y_p^2 Y_c)\mathbb{E}(Y_p^2) - \mathbb{E}(Y_p^3)\mathbb{E}(Y_p Y_c)$

# Help from Non-Gaussianity? Looking at 3rd Moments...

Consider the polynomial: $\tau_{p \to c} = \mathbb{E}(Y_p^2 Y_c)\mathbb{E}(Y_p^2) - \mathbb{E}(Y_p^3)\mathbb{E}(Y_p Y_c)$

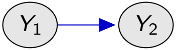
Causal graph:



# Help from Non-Gaussianity? Looking at 3rd Moments...

Consider the polynomial: $\tau_{p \to c} = \mathbb{E}(Y_p^2 Y_c)\mathbb{E}(Y_p^2) - \mathbb{E}(Y_p^3)\mathbb{E}(Y_p Y_c)$

Causal graph:




$$\frac{\mathbb{E}(Y_1 Y_2)}{\mathbb{E}(Y_1^2)} = \frac{\mathbb{E}\left[\varepsilon_1(\beta_{21}\varepsilon_1 + \varepsilon_2)\right]}{\mathbb{E}(\varepsilon_1^2)} = \beta_{21}$$

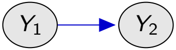
$$\frac{\mathbb{E}(Y_1^2 Y_2)}{\mathbb{E}(Y_1^3)} = \frac{\mathbb{E}\left[\varepsilon_1^2(\beta_{21}\varepsilon_1 + \varepsilon_2)\right]}{\mathbb{E}\left(\varepsilon_1^3\right)} = \beta_{21}$$

It follows that $\tau_{1 \to 2} \equiv 0$.

# Help from Non-Gaussianity? Looking at 3rd Moments...

Consider the polynomial: $\tau_{p \to c} = \mathbb{E}(Y_p^2 Y_c)\mathbb{E}(Y_p^2) - \mathbb{E}(Y_p^3)\mathbb{E}(Y_p Y_c)$

Causal graph:

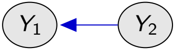



$$\frac{\mathbb{E}(Y_1 Y_2)}{\mathbb{E}(Y_1^2)} = \frac{\mathbb{E}\left[\varepsilon_1(\beta_{21}\varepsilon_1 + \varepsilon_2)\right]}{\mathbb{E}(\varepsilon_1^2)} = \beta_{21}$$

$$\frac{\mathbb{E}(Y_1^2 Y_2)}{\mathbb{E}(Y_1^3)} = \frac{\mathbb{E}\left[\varepsilon_1^2(\beta_{21}\varepsilon_1 + \varepsilon_2)\right]}{\mathbb{E}\left(\varepsilon_1^3\right)} = \beta_{21}$$

It follows that $\tau_{1 \to 2} \equiv 0$.

Causal graph:




$$\frac{\mathbb{E}(Y_1 Y_2)}{\mathbb{E}(Y_1^2)} = \frac{\beta_{12}\mathbb{E}(\varepsilon_2^2)}{\beta_{12}^2\mathbb{E}(\varepsilon_2^2) + \mathbb{E}(\varepsilon_1^2)}$$

$$\frac{\mathbb{E}(Y_1^2 Y_2)}{\mathbb{E}(Y_1^3)} = \frac{\beta_{12}^2\mathbb{E}(\varepsilon_2^3)}{\beta_{12}^3\mathbb{E}(\varepsilon_2^3) + \mathbb{E}(\varepsilon_1^3)}$$

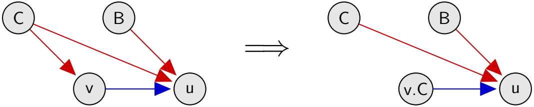
Now, $\tau_{1 \to 2} \not\equiv 0$.

($\neq 0$ generically, in particular, 3rd moments need to be non-Gaussian).

# Moment Relation

For $u \neq v$, $C \subseteq V \setminus \{u, v\}$, and residual $Y_{v.C} = Y_v - \mathbb{E}(Y_v \mid Y_C)$:

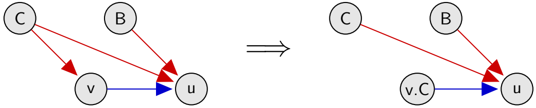
$$\tau^{(K)}_{v.C \to u} := \mathbb{E}\left(Y_{v.C}^{K-1} Y_u\right) \mathbb{E}\left(Y_{v.C}^2\right) - \mathbb{E}\left(Y_{v.C}^K\right) \mathbb{E}\left(Y_{v.C} Y_u\right)$$



# Moment Relation

For $u \neq v$, $C \subseteq V \setminus \{u, v\}$, and residual $Y_{v.C} = Y_v - \mathbb{E}(Y_v \mid Y_C)$:

$$\boxed{\tau^{(K)}_{v.C \to u} := \mathbb{E}\left(Y_{v.C}^{K-1} Y_u\right) \mathbb{E}\left(Y_{v.C}^2\right) - \mathbb{E}\left(Y_{v.C}^K\right) \mathbb{E}\left(Y_{v.C} Y_u\right)}$$




(i) If $u \notin \mathrm{pa}(v)$, then

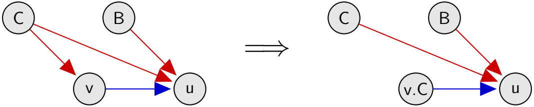
$$\min_C |\tau^{(K)}_{v.C \to u}| = 0.$$

Achieved for $C = \mathrm{pa}(v)$. If $|\mathrm{pa}(v)| \leq J$, testing $|C| \leq J$ enough.

# Moment Relation

For $u \neq v$, $C \subseteq V \setminus \{u, v\}$, and residual $Y_{v.C} = Y_v - \mathbb{E}(Y_v \mid Y_C)$:

$$\boxed{\tau^{(K)}_{v.C \to u} := \mathbb{E}\left(Y^{K-1}_{v.C} Y_u\right) \mathbb{E}\left(Y^2_{v.C}\right) - \mathbb{E}\left(Y^K_{v.C}\right) \mathbb{E}\left(Y_{v.C} Y_u\right)}$$




(i) If $u \notin \mathrm{pa}(v)$, then

$$\min_C |\tau^{(K)}_{v.C \to u}| = 0.$$

Achieved for $C = \mathrm{pa}(v)$. If $|\mathrm{pa}(v)| \leq J$, testing $|C| \leq J$ enough.

(ii) If $u \in \mathrm{pa}(v)$, then generically over sets $C \subseteq V \setminus (\mathrm{de}(v) \cup \{v, u\})$

$$\min_C |\tau^{(K)}_{v.C \to u}| > 0.$$

# Using in Direct-LiNGAM recursion

- Given a set of already 'ordered nodes'.
- Find source $v$ in subgraph of 'unordered nodes' by

$$\max_{u} \min_{C} |\tau^{(K)}_{v.C \to u}| \ = \ 0.$$

  where $u \in$ 'unordered' and $|C| \leq J$ subset of 'ordered'.

- Add $v$ to 'ordered nodes'.
- In practice take $v$ with smallest 'max-min'.

# Modified Direct-LiNGAM

- Concentration inequalities for sample moments give:

  *Under 'strong parental faithfulness', for log-concave errors and DAG of in-degree J, modified Direct-LiNGAM is consistent if*

  $$\frac{\log(p) J^{5/2}}{n^{1/(2K)}} \rightarrow 0.$$

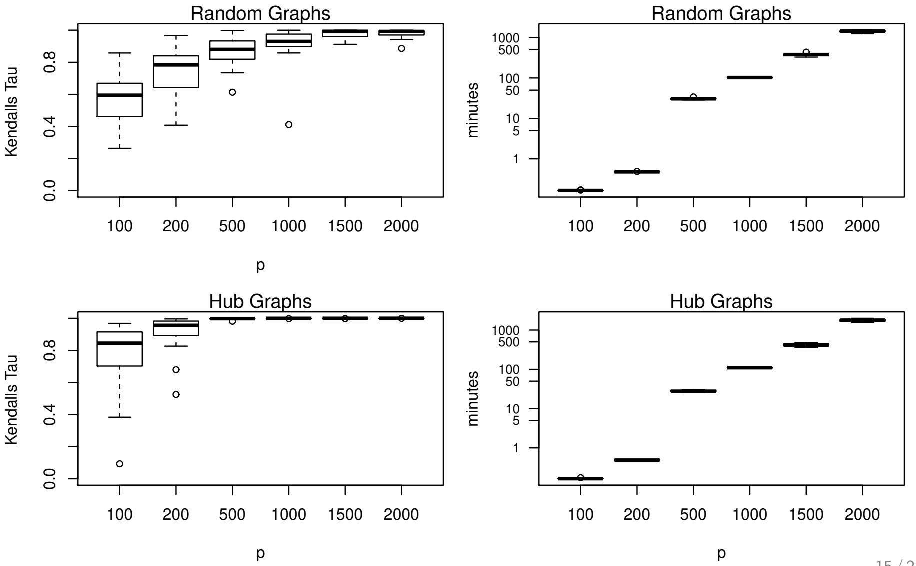
  Parental faithfulness: Total effect between parent and child does not vanish when adjusting on non-descendants.

- Computation:
  - Testing restricted subsets becomes computationally demanding:

    $$|\{C : C \subseteq V_1, |C| = J\}| = O(|V_1|^J)$$

  - Pruning:
    Record when moment relations indicate that node is ancestor but not parent of $v \in$ 'unordered'.

# Illustration



# Causal Discovery with Unobserved Confounding

. . . 2020

# Capturing Unobserved Confounding

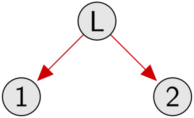



Figure: Children of a common unobserved parent

# Capturing Unobserved Confounding


Figure: Children of a common unobserved parent

# Capturing Unobserved Confounding


Figure: Children of a common unobserved parent

- Mixed graph $G = (V, E_{\to}, E_{\leftrightarrow})$.
- Non-Gaussian Linear Model:

$$Y_v = \sum_{u \in \mathrm{pa}(v)} \beta_{vu} Y_u + \varepsilon_v, \quad v \in V,$$

  with $\mathbb{E}(\varepsilon_v \varepsilon_u) = \omega_{vu} \neq 0$ only if $u = v$ or $u \leftrightarrow v \in E_{\leftrightarrow}$ (siblings).
- Continue to assume that $E_{\to}$ is acyclic.
- In which settings might we be able to infer the underlying graph $G$?

# Existing Work

**Gaussian or conditional independence based methods**:

- Constraint testing[1] and greedy methods[2] for maximal ancestral graphs
- Greedy search[3] for bow-free acyclic path diagrams (BAPs)

**Explicitly non-Gaussian**:

- Overcomplete ICA[4]
- Bayesian specification[5]
- Conservative Direct-LiNGAM approach[6]
- ParceLiNGAM[7] (still use independence of residuals from regression)

---

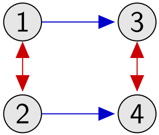
[1] Richardson and Spirtes (2002),Colombo et al. (2012),Claassen et al. (2013)
[2] Triantafillou and Tsamardinos (2016)
[3] Nowzohour et al. (2017)
[4] Hoyer et al. (2008)
[5] Shimizu and Bollen (2014)
[6] Entner and Hoyer (2010)
[7] Tashiro et al. (2014)

# Ancestral Graphs

- ParceLiNGAM applies Direct-LiNGAM (locating sources) and its "dual" (locating sinks) to all subsets of variables.
- Amounts to checks of

$$Y_{v.C} = Y_v - \mathbb{E}(Y_v \mid Y_C) \perp\!\!\!\perp Y_C, \quad v \in V,\ C \subseteq V \setminus \{v\}.$$

- ParceLiNGAM is sound: returns a partial ordering that extends to a topological ordering of the mixed graph $G$.
- Example:




**Theorem**

*ParceLiNGAM recovers a topological ordering of $G$ iff $G$ ancestral.*

# What's special about Ancestral Graphs?

- A graph is ancestral if it does not contain semi-directed cycles of form

$$v \leftrightarrow w \to \cdots \to v.$$

**Theorem**

*(i) The graph $G$ is ancestral if and only if*

$$\mathbb{E}(Y_v \mid Y_{\mathrm{pa}(v)}) = \sum_{c \in \mathrm{pa}(v)} \beta_{vc} Y_c \qquad \textit{for all nodes } v.$$

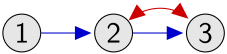
*(ii) The graph $G$ is ancestral if and only if*

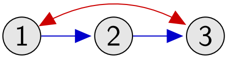
$$\Big[\varepsilon_v = Y_v - \sum_{u \in \mathrm{pa}(v)} \beta_{vu} Y_u\Big] \perp\!\!\!\perp \Big[Y_{\mathrm{pa}(v)} = f(\varepsilon_{\mathrm{an}(v)})\Big] \qquad \textit{for all nodes } v.$$

- Regression residual $Y_v - \mathbb{E}(Y_v \mid Y_{\mathrm{pa}(v)}) = \varepsilon_v$ independent of $Y_{\mathrm{pa}(v)}$.

# Bow-Free Acyclic Graphs

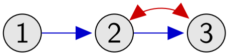
- Bow-free: At most one edge between any pair of nodes

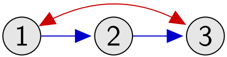


# Bow-Free Acyclic Graphs

- Bow-free: At most one edge between any pair of nodes



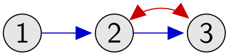

- Complications exemplified (top right):

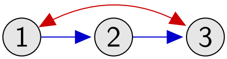
$$\mathbb{E}(Y_3 | Y_2) = \frac{\beta_{32}(\beta_{21}^2\omega_{11} + \omega_{22}) + \beta_{21}\omega_{13}}{\beta_{12}^2\omega_{11} + \omega_{22}} \neq \beta_{32}$$

$$\left[\varepsilon_3 = Y_2 - \beta_{32} Y_3\right] \not\perp\!\!\!\perp \left[Y_2 = \varepsilon_2 + \beta_{21}\varepsilon_1\right]$$

# Bow-Free Acyclic Graphs

- Bow-free: At most one edge between any pair of nodes




- Complications exemplified (top right):

$$\mathbb{E}(Y_3|Y_2) = \frac{\beta_{32}(\beta_{21}^2\omega_{11} + \omega_{22}) + \beta_{21}\omega_{13}}{\beta_{12}^2\omega_{11} + \omega_{22}} \neq \beta_{32}$$

$$\left[\varepsilon_3 = Y_2 - \beta_{32}Y_3\right] \not\perp\!\!\!\perp \left[Y_2 = \varepsilon_2 + \beta_{21}\varepsilon_1\right]$$

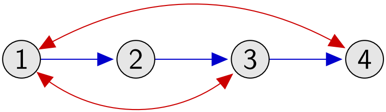
- Bow-free acyclic graphs can be recovered :

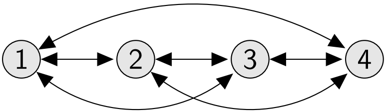
  i) $(\beta_{vu})$ generically identifiable from Var$(Y)$.

  In fact, each $\beta_{v,\mathrm{pa}(v)}$ identifiable from Var$(Y_{\mathrm{an}(v)})$ and $\mathbb{E}(Y_v \mid Y_{\mathrm{an}(v)})$.

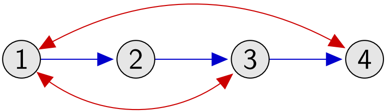
  ii) $\varepsilon_v \perp\!\!\!\perp \varepsilon_{\mathrm{pa}(v)}$.

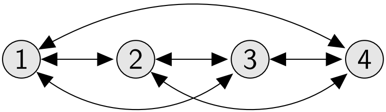
# Algorithm in an Example



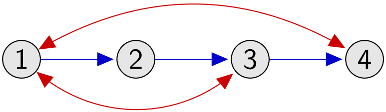
# Algorithm in an Example

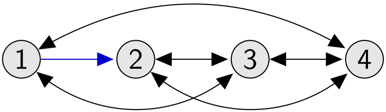



(a) First test independence of regression residuals: $Y_{v.C} \perp\!\!\!\perp Y_C$

# Algorithm in an Example

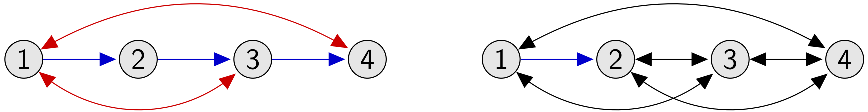



(a) First test independence of regression residuals: $Y_{v.C} \perp\!\!\!\perp Y_C$

- Only $Y_{2.1} \perp\!\!\!\perp Y_1$: parent/ancestor relation $1 \to 2$ <u>and</u> $\beta_{21}$;
- Adjust $Y_2$ to $\bar{Y}_2 = Y_2 - \beta_{21} Y_1 = \varepsilon_2$.

# Algorithm in an Example

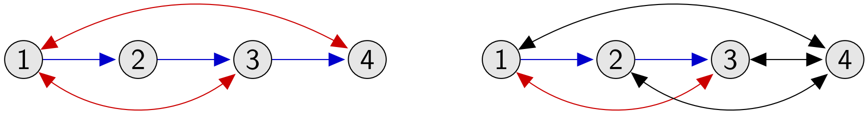



(a) First test independence of regression residuals: $Y_{v.C} \perp\!\!\!\perp Y_C$

- Only $Y_{2.1} \perp\!\!\!\perp Y_1$: parent/ancestor relation $1 \to 2$ and $\beta_{21}$;
- Adjust $Y_2$ to $\bar{Y}_2 = Y_2 - \beta_{21} Y_1 = \varepsilon_2$.

(b) Test again with adjusted observations and estimates of $\beta_{vu}$:

# Algorithm in an Example



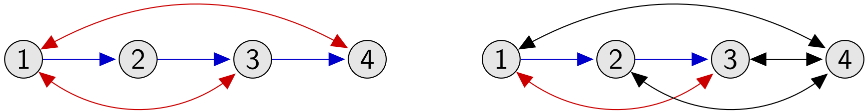

(a) First test independence of regression residuals: $Y_{v.C} \perp\!\!\!\perp Y_C$

- Only $Y_{2.1} \perp\!\!\!\perp Y_1$: parent/ancestor relation $1 \to 2$ and $\beta_{21}$;
- Adjust $Y_2$ to $\bar{Y}_2 = Y_2 - \beta_{21} Y_1 = \varepsilon_2$.

(b) Test again with adjusted observations and estimates of $\beta_{vu}$:

- Discovery of $2 \to 3$ and $\beta_{32}$.
  Form $\bar{Y}_3 = Y_3 - \beta_{32} Y_2 = \varepsilon_3$, and find $\bar{Y}_3 \perp\!\!\!\perp \bar{Y}_2 = \varepsilon_2$;
- Also, $1 \notin \mathrm{pa}(3)$ as even after correct adjustment we have $\varepsilon_1 \not\!\perp\!\!\!\perp \varepsilon_3$.

# Algorithm in an Example

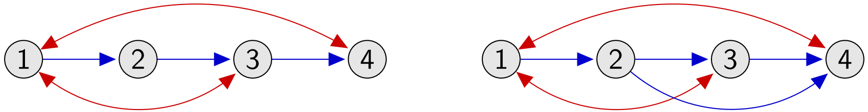



(a) First test independence of regression residuals: $Y_{v.C} \perp\!\!\!\perp Y_C$
- Only $Y_{2.1} \perp\!\!\!\perp Y_1$: parent/ancestor relation $1 \to 2$ <u>and</u> $\beta_{21}$;
- Adjust $Y_2$ to $\bar{Y}_2 = Y_2 - \beta_{21} Y_1 = \varepsilon_2$.

(b) Test again with adjusted observations and estimates of $\beta_{vu}$:
- Discovery of $2 \to 3$ <u>and</u> $\beta_{32}$.
  Form $\bar{Y}_3 = Y_3 - \beta_{32} Y_2 = \varepsilon_3$, and find $\bar{Y}_3 \perp\!\!\!\perp \bar{Y}_2 = \varepsilon_2$;
- Also, $1 \notin \mathrm{pa}(3)$ as even after correct adjustment we have $\varepsilon_1 \not\!\perp\!\!\!\perp \varepsilon_3$.

(c) Repeat:

# Algorithm in an Example




(a) First test independence of regression residuals: $Y_{v.C} \perp\!\!\!\perp Y_C$

- Only $Y_{2.1} \perp\!\!\!\perp Y_1$: parent/ancestor relation $1 \to 2$ <u>and</u> $\beta_{21}$;
- Adjust $Y_2$ to $\bar{Y}_2 = Y_2 - \beta_{21} Y_1 = \varepsilon_2$.

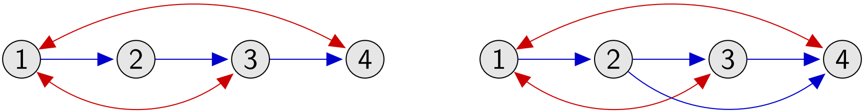
(b) Test again with adjusted observations and estimates of $\beta_{vu}$:

- Discovery of $2 \to 3$ <u>and</u> $\beta_{32}$.
  Form $\bar{Y}_3 = Y_3 - \beta_{32} Y_2 = \varepsilon_3$, and find $\bar{Y}_3 \perp\!\!\!\perp \bar{Y}_2 = \varepsilon_2$;
- Also, $1 \notin \mathrm{pa}(3)$ as even after correct adjustment we have $\varepsilon_1 \not\perp\!\!\!\perp \varepsilon_3$.

(c) Repeat:

- Infer $2 \to 4$ and $3 \to 4$, so discover $\mathrm{pa}(4) \subset \{2, 3\}$ and $\{2, 3\} \cap \mathsf{sib}(4) = \emptyset$. Discover $1 \notin \mathrm{pa}(4)$.

# Algorithm in an Example




(a) First test independence of regression residuals: $Y_{v.C} \perp\!\!\!\perp Y_C$

- Only $Y_{2.1} \perp\!\!\!\perp Y_1$: parent/ancestor relation $1 \to 2$ and $\beta_{21}$;
- Adjust $Y_2$ to $\bar{Y}_2 = Y_2 - \beta_{21} Y_1 = \varepsilon_2$.

(b) Test again with adjusted observations and estimates of $\beta_{vu}$:

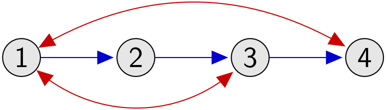
- Discovery of $2 \to 3$ and $\beta_{32}$.
  Form $\bar{Y}_3 = Y_3 - \beta_{32} Y_2 = \varepsilon_3$, and find $\bar{Y}_3 \perp\!\!\!\perp \bar{Y}_2 = \varepsilon_2$;
- Also, $1 \notin \mathrm{pa}(3)$ as even after correct adjustment we have $\varepsilon_1 \not\perp\!\!\!\perp \varepsilon_3$.

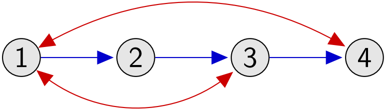
(c) Repeat:

- Infer $2 \to 4$ and $3 \to 4$, so discover $\mathrm{pa}(4) \subset \{2, 3\}$ and $\{2, 3\} \cap \mathrm{sib}(4) = \emptyset$. Discover $1 \notin \mathrm{pa}(4)$.

(d) Prune $\to$ edges accounting for dependence induced by $\leftrightarrow$.

# Algorithm in an Example




(a) First test independence of regression residuals: $Y_{v.C} \perp\!\!\!\perp Y_C$
- Only $Y_{2.1} \perp\!\!\!\perp Y_1$: parent/ancestor relation $1 \to 2$ <u>and</u> $\beta_{21}$;
- Adjust $Y_2$ to $\bar{Y}_2 = Y_2 - \beta_{21} Y_1 = \varepsilon_2$.

(b) Test again with adjusted observations and estimates of $\beta_{vu}$:
- Discovery of $2 \to 3$ <u>and</u> $\beta_{32}$.
  Form $\bar{Y}_3 = Y_3 - \beta_{32} Y_2 = \varepsilon_3$, and find $\bar{Y}_3 \perp\!\!\!\perp \bar{Y}_2 = \varepsilon_2$;
- Also, $1 \notin \mathrm{pa}(3)$ as even after correct adjustment we have $\varepsilon_1 \not\!\perp\!\!\!\perp \varepsilon_3$.

(c) Repeat:
- Infer $2 \to 4$ and $3 \to 4$, so discover $\mathrm{pa}(4) \subset \{2, 3\}$ and $\{2, 3\} \cap \mathrm{sib}(4) = \emptyset$. Discover $1 \notin \mathrm{pa}(4)$.

(d) Prune $\to$ edges accounting for dependence induced by $\leftrightarrow$.

# Simulations: Maximal Ancestral Graphs

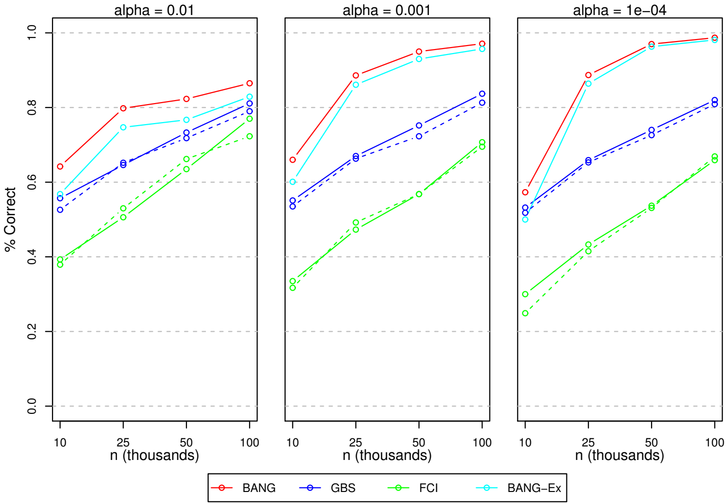



Figure: 1000 Random MAGs with $p = 5$. Solid lines are log-normal errors; dotted lines are Gaussian errors.

# Simulations: BAPs

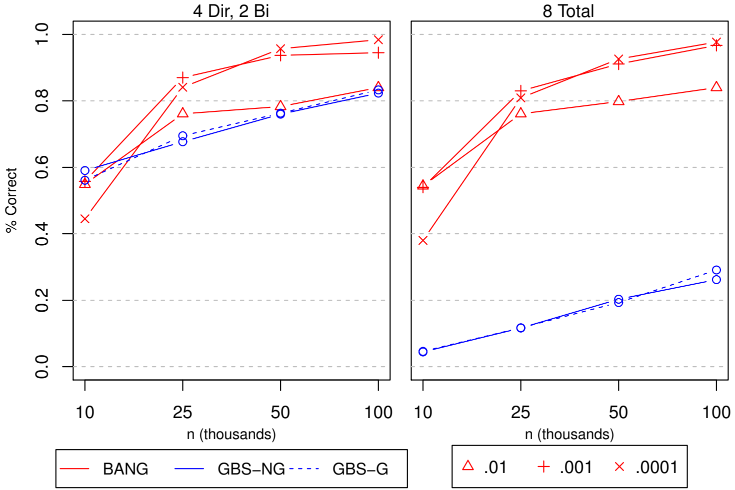
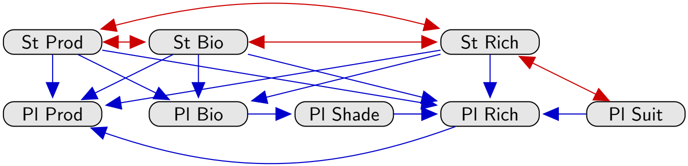
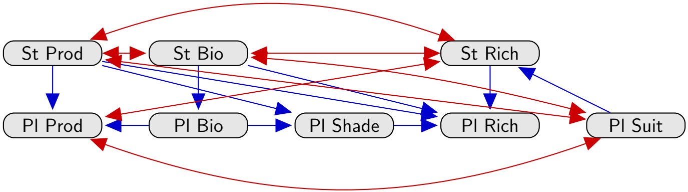



Figure: 1000 Random BAPs with $p = 5$. Solid lines are log-normal errors; dotted lines are Gaussian errors.

# Data Example: Ecology Data from Grace et al. (2016)




(a) BAP representation of plot specific model from Grace et al. (2016).




(b) Discovered model matches 16 out of 28 edges. Probability of 16 or more edges by random guessing is .002.

To the organizers, a big:

# THANK YOU!

# References I

Claassen, T., Mooij, J. M., and Heskes, T. (2013). Learning sparse causal models is not np-hard. In Nicholson, A. and Smyth, P., editors, *Proceedings of the Twenty-Ninth Conference on Uncertainty in Artificial Intelligence, UAI 2013, Bellevue, WA, USA, August 11-15, 2013*. AUAI Press.

Colombo, D., Maathuis, M. H., Kalisch, M., and Richardson, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *Ann. Statist.*, 40(1):294–321.

Entner, D. and Hoyer, P. O. (2010). Discovering unconfounded causal relationships using linear non-gaussian models. In Onada, T., Bekki, D., and McCready, E., editors, *New Frontiers in Artificial Intelligence - JSAI-isAI 2010 Workshops, LENLS, JURISIN, AMBN, ISS, Tokyo, Japan, November 18-19, 2010, Revised Selected Papers*, volume 6797 of *Lecture Notes in Computer Science*, pages 181–195. Springer.

Grace, J. B., Anderson, T. M., Seabloom, E. W., Borer, E. T., Adler, P. B., Harpole, W. S., Hautier, Y., Hillebrand, H., Lind, E. M., Pärtel, M., et al. (2016). Integrative modelling reveals mechanisms linking productivity and plant species richness. *Nature*, 529(7586):390–393.

# References II


Hoyer, P. O., Shimizu, S., Kerminen, A. J., and Palviainen, M. (2008). Estimation of causal effects using linear non-Gaussian causal models with hidden variables. *International Journal of Approximate Reasoning*, 49(2):362 – 378. Special Section on Probabilistic Rough Sets and Special Section on PGM06.

Hyvärinen, A. and Smith, S. M. (2013). Pairwise likelihood ratios for estimation of non-Gaussian structural equation models. *J. Mach. Learn. Res.*, 14:111–152.

Nowzohour, C., Maathuis, M. H., Evans, R. J., and Bühlmann, P. (2017). Distributional equivalence and structure learning for bow-free acyclic path diagrams. *Electron. J. Stat.*, 11(2):5342–5374.

Richardson, T. and Spirtes, P. (2002). Ancestral graph Markov models. *Ann. Statist.*, 30(4):962–1030.

Shimizu, S. and Bollen, K. (2014). Bayesian estimation of causal direction in acyclic structural equation models with individual-specific confounder variables and non-Gaussian distributions. *Journal of Machine Learning Research*, 15(1):2629–2652.

Shimizu, S., Hoyer, P. O., Hyvärinen, A., and Kerminen, A. (2006). A linear non-Gaussian acyclic model for causal discovery. *J. Mach. Learn. Res.*, 7:2003–2030.

# References III


Shimizu, S., Inazumi, T., Sogawa, Y., Hyvärinen, A., Kawahara, Y., Washio, T., Hoyer, P. O., and Bollen, K. (2011). DirectLiNGAM: a direct method for learning a linear non-Gaussian structural equation model. *J. Mach. Learn. Res.*, 12:1225–1248.

Tashiro, T., Shimizu, S., Hyvärinen, A., and Washio, T. (2014). ParceLiNGAM: a causal ordering method robust against latent confounders. *Neural Comput.*, 26(1):57–83.

Triantafillou, S. and Tsamardinos, I. (2016). Score-based vs constraint-based causal learning in the presence of confounders. In Eberhardt, F., Bareinboim, E., Maathuis, M. H., Mooij, J. M., and Silva, R., editors, *Proceedings of the UAI 2016 Workshop on Causation: Foundation to Application co-located with the 32nd Conference on Uncertainty in Artificial Intelligence (UAI 2016), Jersey City, USA, June 29, 2016.*, volume 1792 of *CEUR Workshop Proceedings*, pages 59–67. CEUR-WS.org.