UAI 2023

UAI 2023 Workshop

The History and Development of Search Methods for Causal Structure

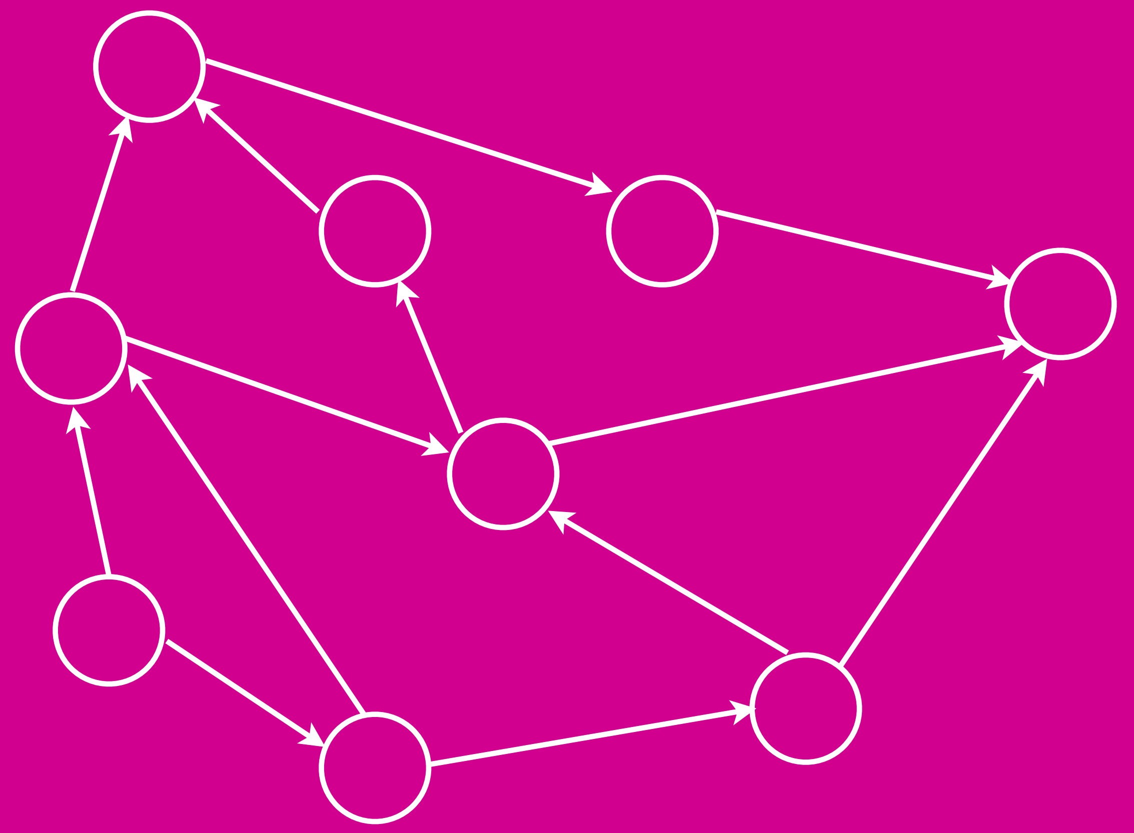
Google

MOHAMED BIN ZAYED UNIVERSITY OF ARTIFICIAL INTELLIGENCE

Carnegie Mellon University

JPMORGAN CHASE & CO.

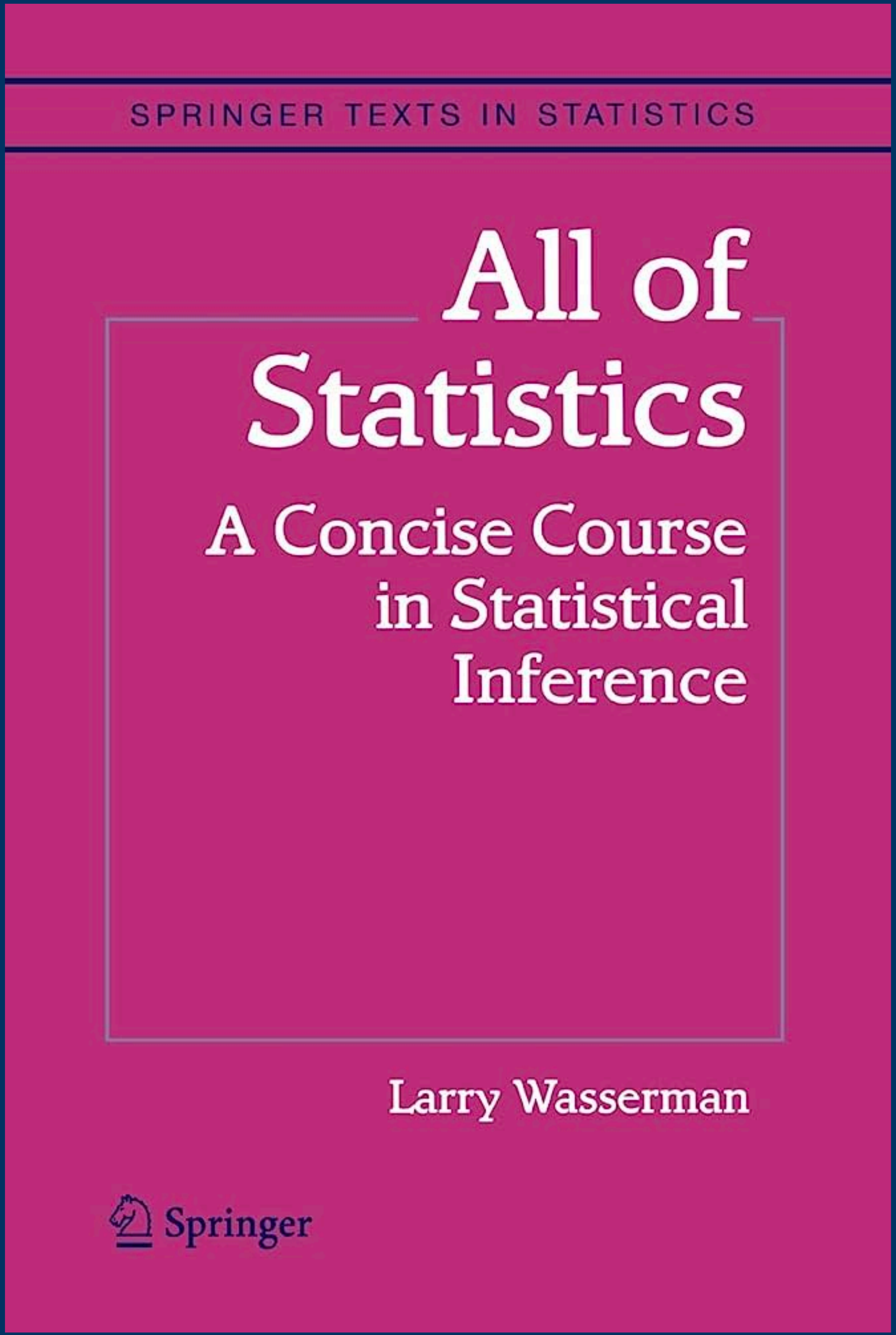
IBM Research

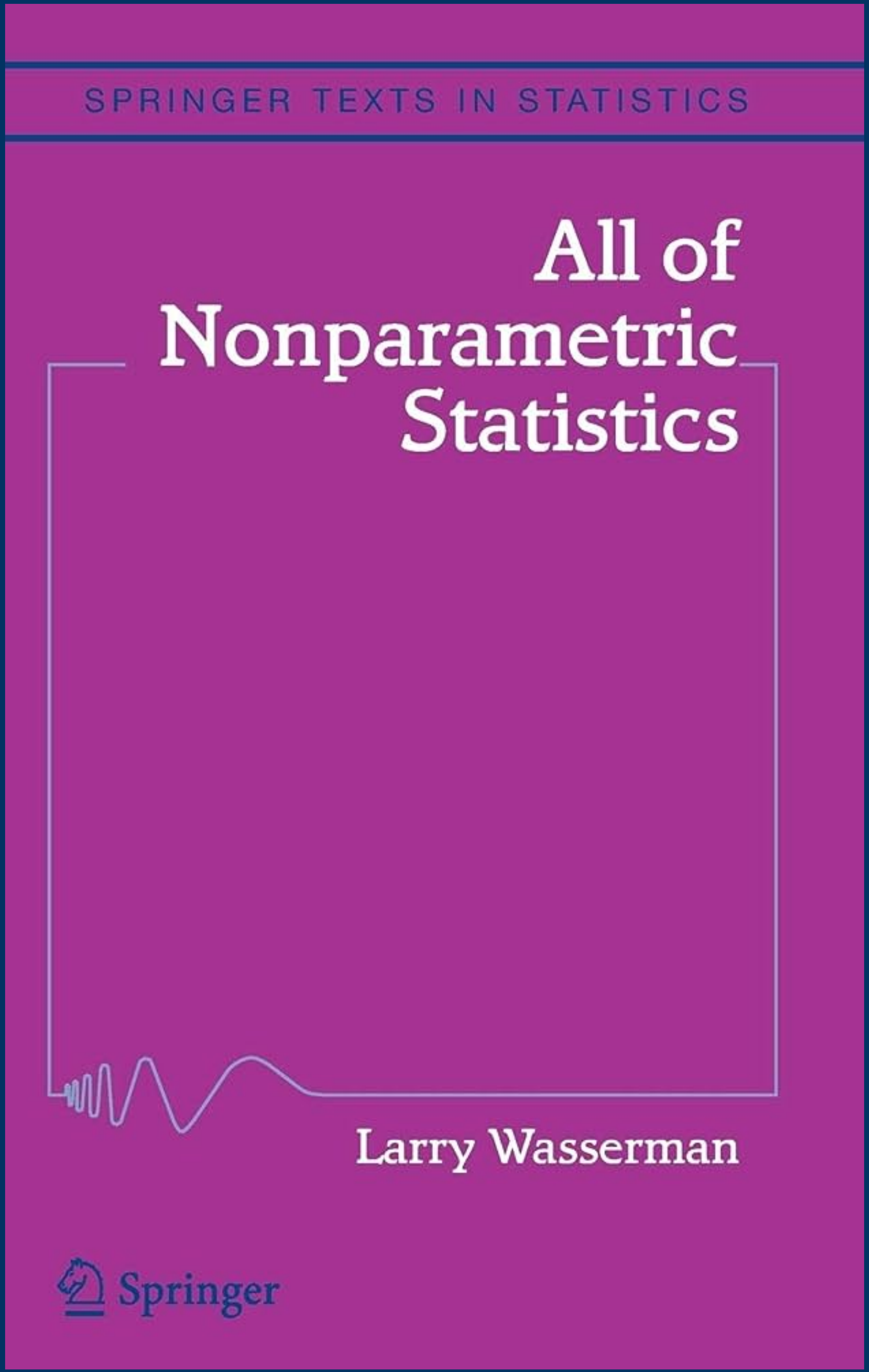


All of© Causal Discovery — now

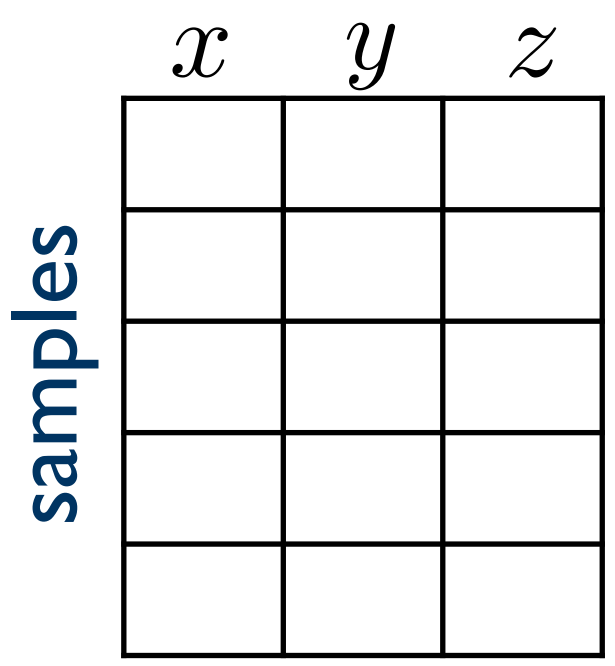
Frederick Eberhardt

Caltech


SPRINGER TEXTS IN STATISTICS
All of Statistics
A Concise Course in Statistical Inference
Larry Wasserman
Springer


SPRINGER TEXTS IN STATISTICS
All of Nonparametric Statistics
Larry Wasserman
Springer

Causal Discovery (from observational data)



data sample

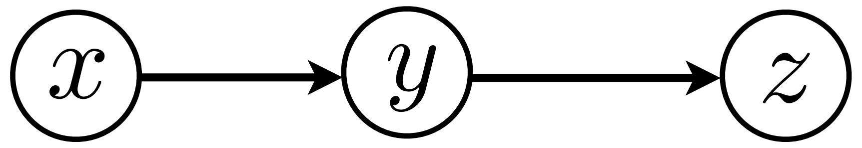
inference algorithm

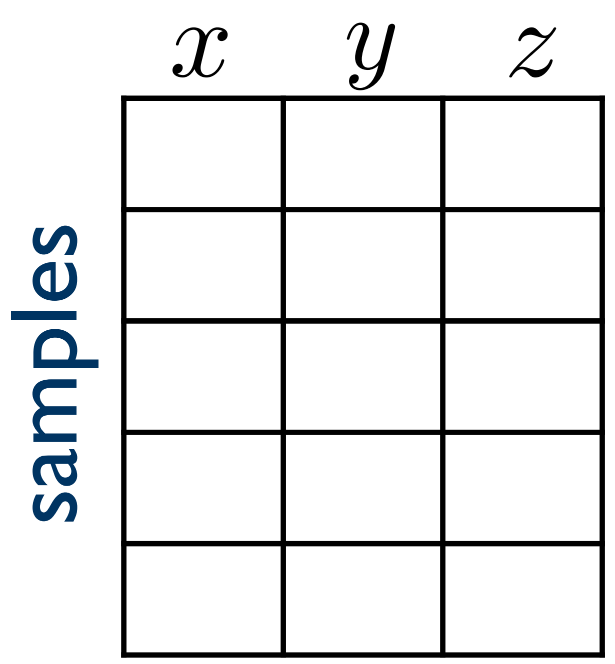
$$x \rightarrow y \rightarrow z$$

causal structure

Causal Discovery (from observational data)

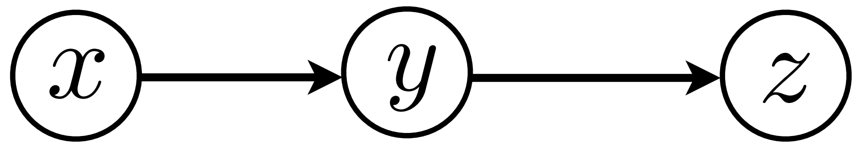
truth
(unknown)





data sample

inference algorithm



causal structure

Causal Discovery (from observational data)

assumptions
- causal Markov
- causal faithfulness
- causal sufficiency
- acyclicity
- …

truth
(unknown)

$$x \to y \to z$$

samples

x	y	z

data sample

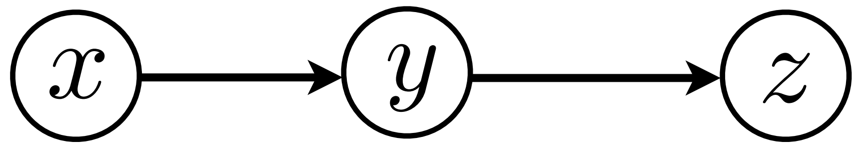
inference algorithm

$$x \to y \to z$$

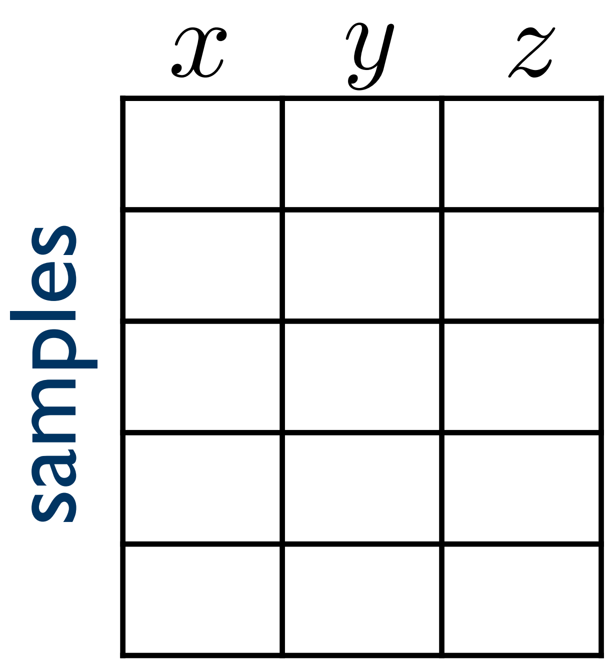
causal structure

Causal Discovery (from observational data)

truth
(unknown)

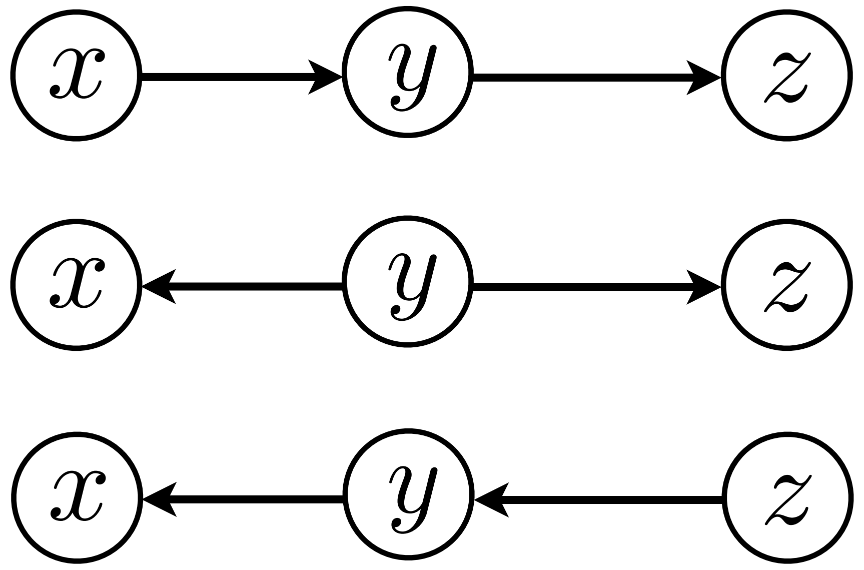


assumptions
- causal Markov
- causal faithfulness
- causal sufficiency
- acyclicity
- …



data sample

inference algorithm

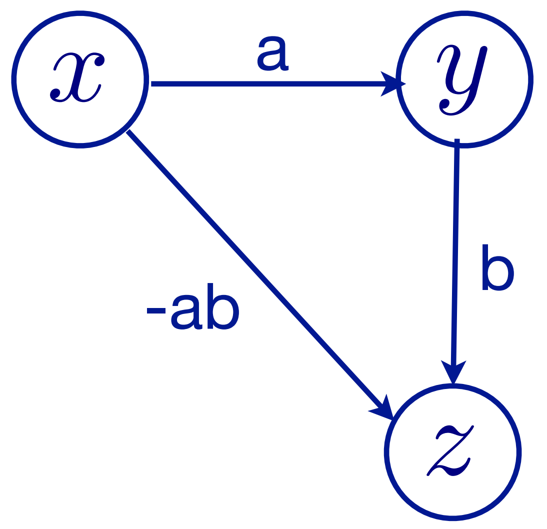


equivalence class of causal structures

Assumptions

- Markov Conditon: (conditional) probabilistic dependence implies (conditional) d-connection

- Faithfulness Conditon: (conditional) probabilistic independence implies (conditional) d-separation

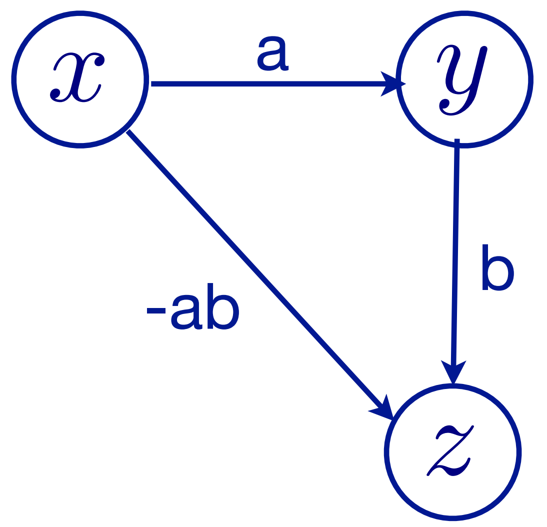


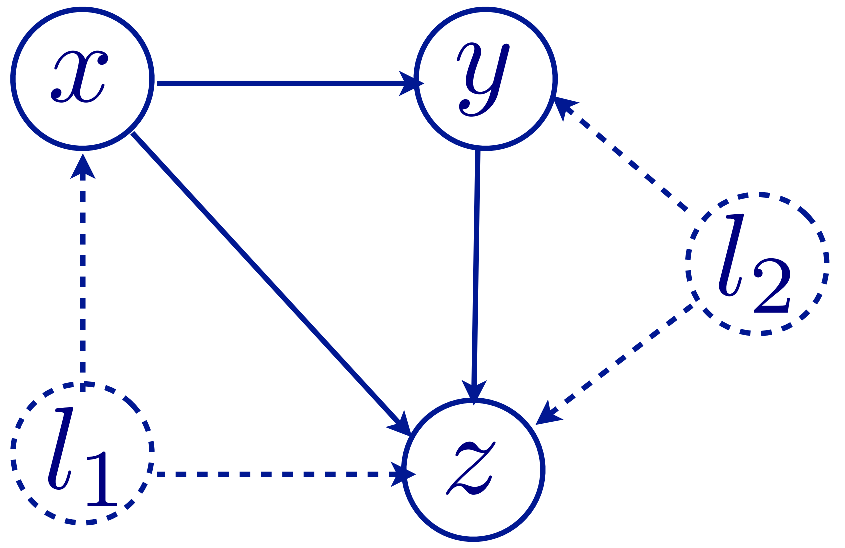
Assumptions

- Markov Conditon: (conditional) probabilistic dependence implies (conditional) d-connection

- Faithfulness Conditon: (conditional) probabilistic independence implies (conditional) d-separation

- Causal Sufficiency: there are no unmeasured common causes of two or more measured variables.





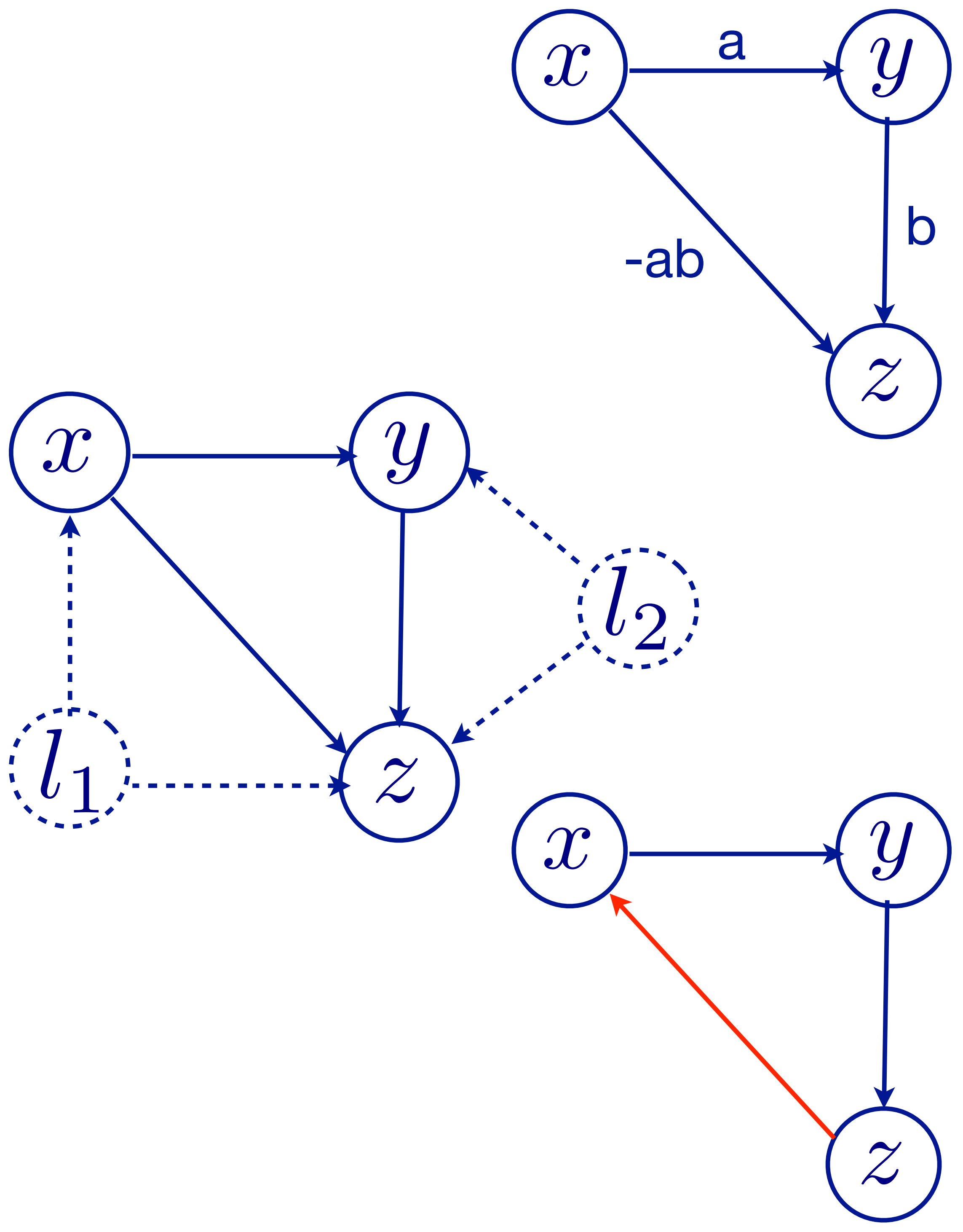
Assumptions

- Markov Conditon: (conditional) probabilistic dependence implies (conditional) d-connection

- Faithfulness Conditon: (conditional) probabilistic independence implies (conditional) d-separation

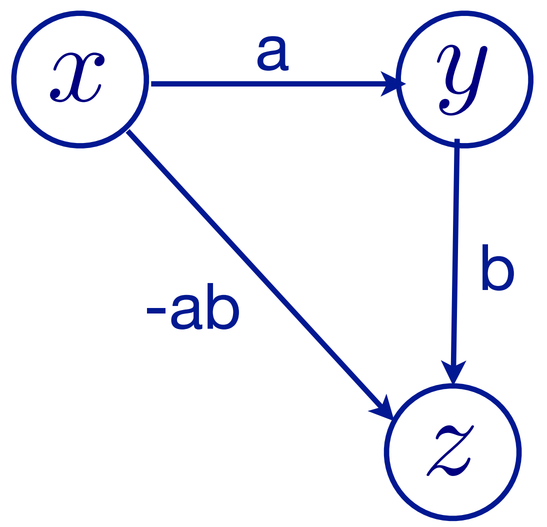
- Causal Sufficiency: there are no unmeasured common causes of two or more measured variables.

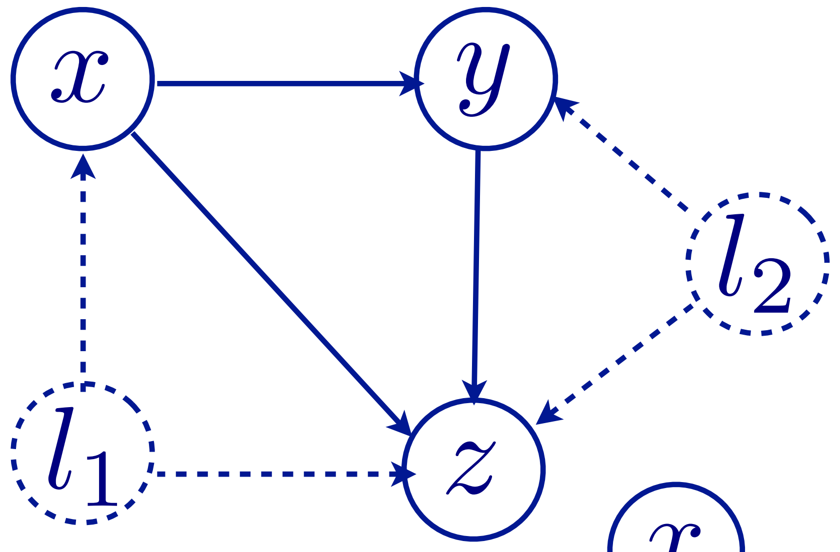
- Acyclicity: the causal structure contains no cycles

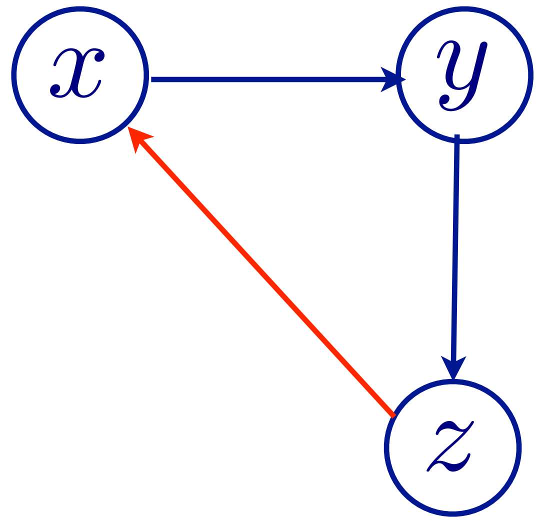


Assumptions

- Markov Conditon: (conditional) probabilistic dependence implies (conditional) d-connection
- Faithfulness Conditon: (conditional) probabilistic independence implies (conditional) d-separation
- Causal Sufficiency: there are no unmeasured common causes of two or more measured variables.
- Acyclicity: the causal structure contains no cycles
- Parametric assumption: the causal relation is described by a particular functional form.







$$y = f(pa(y)) + \epsilon_y$$

Where I ended in 2013:

https://www.youtube.com/watch?v=PpY7Slo57XQ&t=2098s

assumption/ algorithm	PC / GES	FCI	CCD	LiNGaM	lvLiNGaM	cyclic LiNGaM	non-linear additive noise	SAT
Markov	✓	✓	✓	✓	✓	✓	✓	✓
faithfulness	✓	✓	✓	✗	✓	~	minimality	✓
causal sufficiency	✓	✗	✓	✓	✗	✓	✓	✗
acyclicity	✓	✓	✗*	✓	✓	✗	✓	✗*
parametric assumption	✗	✗	✗	linear non-Gaussian	linear non-Gaussian	linear non-Gaussian	non-linear additive noise	✗
output	Markov equivalence	PAG	PAG	unique DAG	set of DAGs	set of graphs	unique DAG	query based
application	wide use	some?	none	fMRI	requires too much data	fMRI	starting	in development

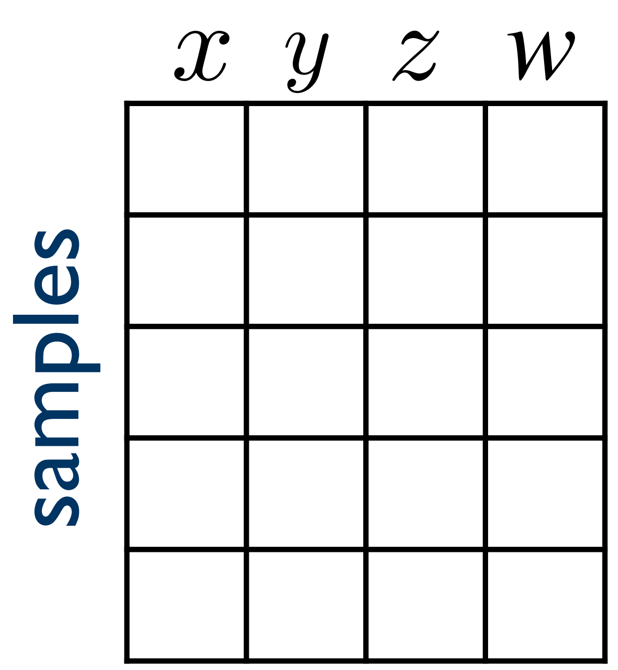
~ special case
* care needs to be taken how cyclicity is modeled

Exploiting the independence structure

assumption/ algorithm	PC / GES	FCI	CCD	LiNGaM	lvLiNGaM	cyclic LiNGaM	non-linear additive noise	SAT
Markov	✓	✓	✓	✓	✓	✓	✓	✓
faithfulness	✓	✓	✓	✗	✓	~	minimality	✓
causal sufficiency	✓	✗	✓	✓	✗	✓	✓	✗
acyclicity	✓	✓	✗*	✓	✓	✗	✓	✗*
parametric assumption	✗	✗	✗	linear non-Gaussian	linear non-Gaussian	linear non-Gaussian	non-linear additive noise	✗
output	Markov equivalence	PAG	PAG	unique DAG	set of DAGs	set of graphs	unique DAG	query based
application	wide use	some?	none	fMRI	requires too much data	fMRI	starting	in development

~ special case
* care needs to be taken how cyclicity is modeled

Exploiting the independence structure

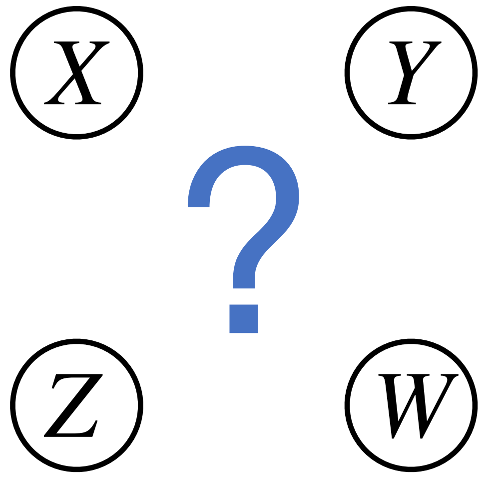


assumptions
- ...

$$
\begin{array}{ll}
X \perp\!\!\!\perp Y & X \not\!\perp\!\!\!\perp Y|Z \\
X \not\!\perp\!\!\!\perp Z & X \not\!\perp\!\!\!\perp Y|W \\
X \not\!\perp\!\!\!\perp W & X \not\!\perp\!\!\!\perp Z|Y \\
Y \not\!\perp\!\!\!\perp Z & X \not\!\perp\!\!\!\perp Z|W \\
Y \not\!\perp\!\!\!\perp W & X \not\!\perp\!\!\!\perp W|Y \\
Z \not\!\perp\!\!\!\perp W & X \perp\!\!\!\perp W|Z
\end{array}
$$

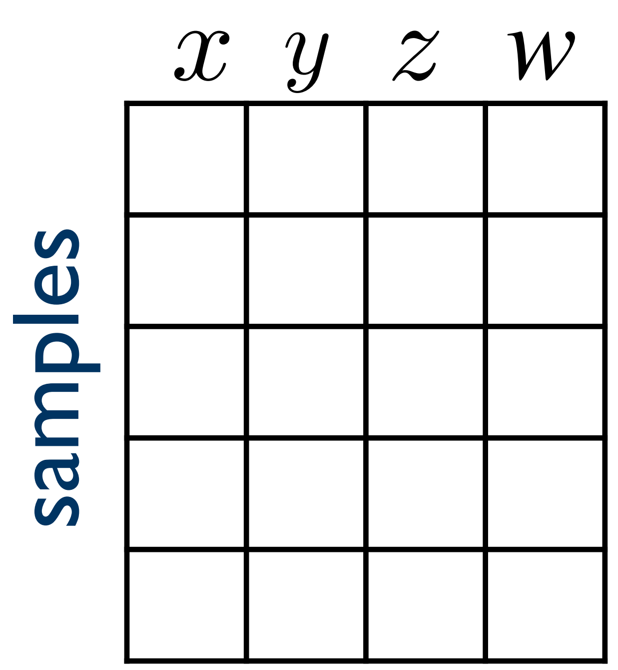
(in)dependence test results or local scores

inference algorithm



equivalence class of causal structures

How do indep-structure-based algos differ?



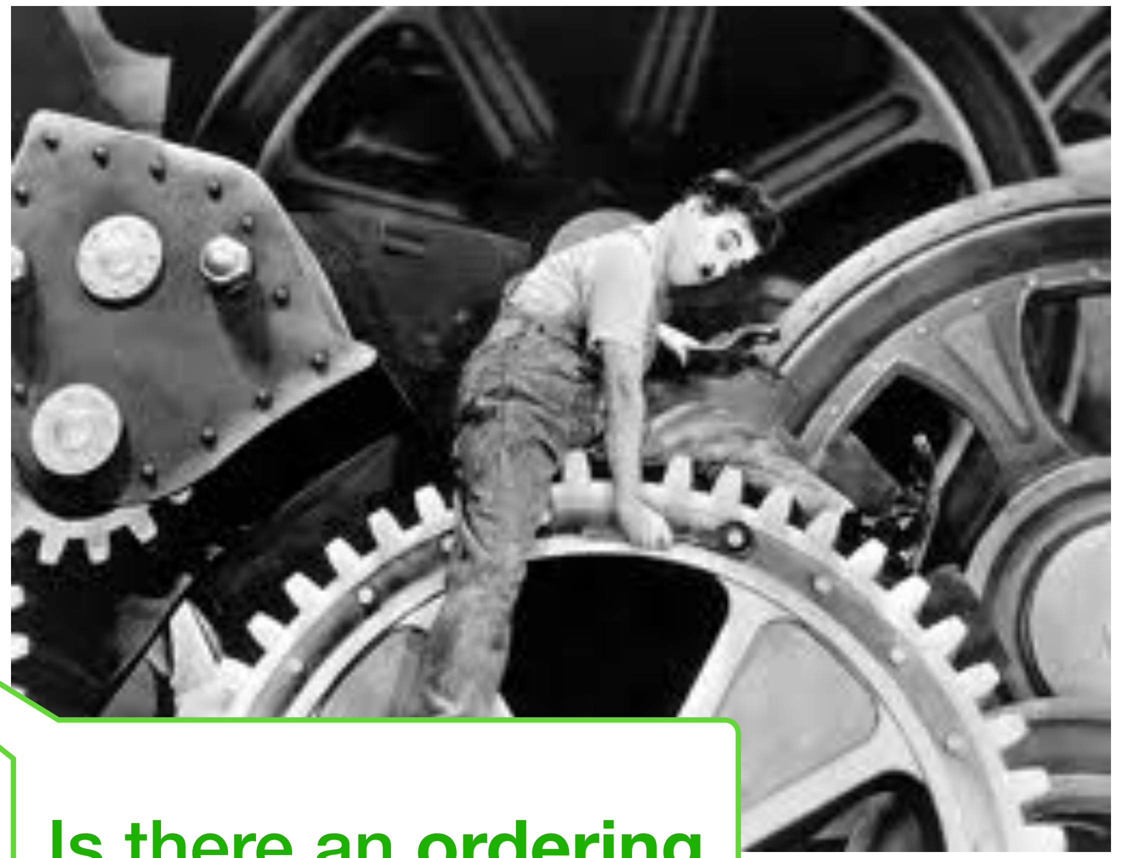
Are **scores** or **independence test** constraints used?

$X \perp\!\!\!\perp Y$ | $X \not\!\perp\!\!\!\perp Y | Z$
$X \not\!\perp\!\!\!\perp Z$ | $X \not\!\perp\!\!\!\perp Y | W$
$X \not\!\perp\!\!\!\perp W$ | $X \not\!\perp\!\!\!\perp Z | Y$
$Y \not\!\perp\!\!\!\perp Z$ | $X \not\!\perp\!\!\!\perp Z | W$
$Y \not\!\perp\!\!\!\perp W$ | $X \not\!\perp\!\!\!\perp W | Y$
$Z \not\!\perp\!\!\!\perp W$ | $X \perp\!\!\!\perp W | Z$

(in)dependence test results or local scores

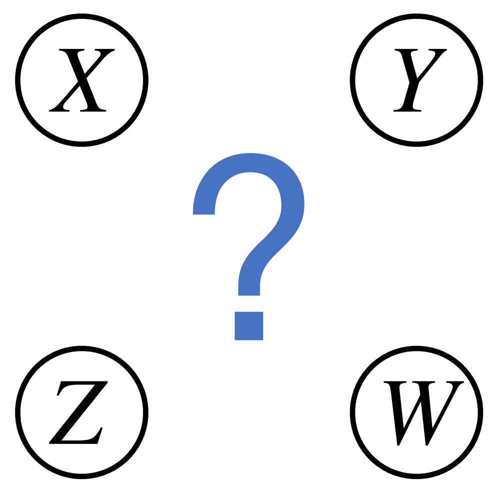
assumptions
- …

What are the model space **assumptions?**



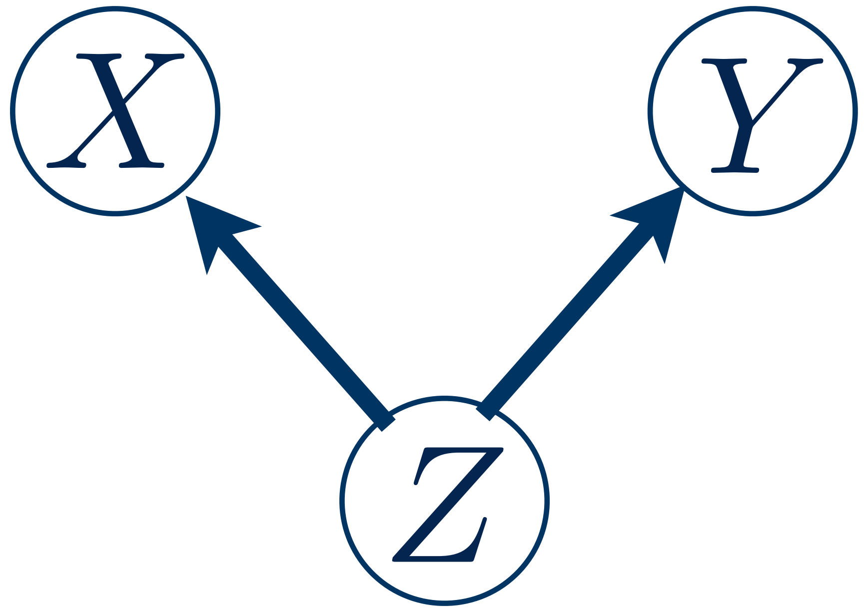
Is there an **ordering** to how constraints are processed?

rithm



equivalence class of causal structures

Conflicted Constraints



$X \not\perp\!\!\!\perp Y$

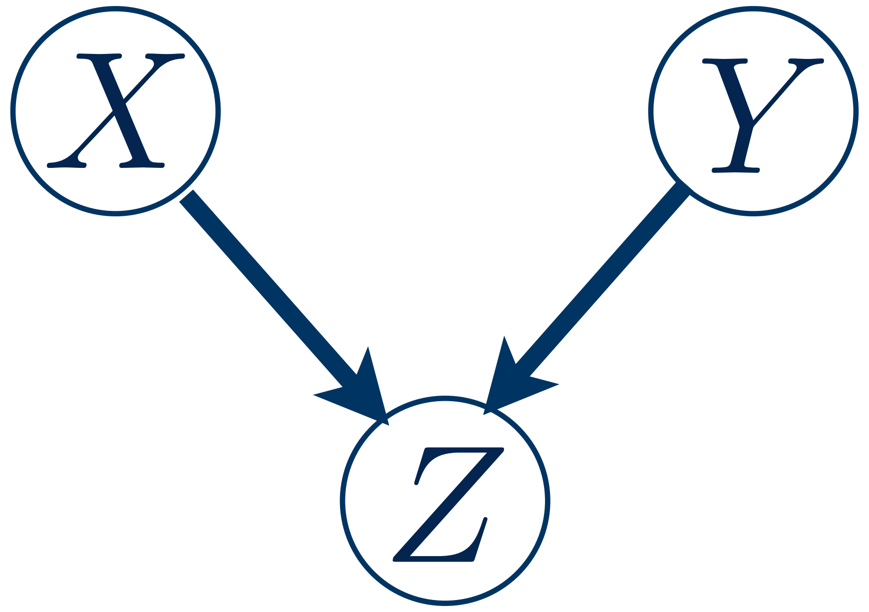
$$X \perp\!\!\!\perp Y$$

$$X \not\perp\!\!\!\perp Z$$

$$Y \not\perp\!\!\!\perp Z$$

$$X \perp\!\!\!\perp Y \,|\, Z$$

$X \not\perp\!\!\!\perp Y|Z$



(in)dependence
constraints

Discovery guarantees

- Completeness: Given the true (conditional) independence and dependence relations [the algorithm] identifies all there is to discover about the true underlying graph, namely, its Markov equivalence class.

Causal Discovery Over Three Variables

Equivalence Classes of Causal Models Over Three Variables

Discovery guarantees

- Completeness: Given the true (conditional) independence and dependence relations [the algorithm] identifies all there is to discover about the true underlying graph, namely, its Markov equivalence class.

For the causally sufficient, acyclic case, simulations suggest that on average there are about 4-5 DAGs per Markov equivalence class, i.e. that the underdetermination is independent of the number of variables (Gillispie & Perlman, 2002; He et al., 2015; Radhakrishnan et al, 2018).

Discovery guarantees

- Completeness: Given the true (conditional) independence and dependence relations [the algorithm] identifies all there is to discover about the true underlying graph, namely, its Markov equivalence class.

- Statistical guaranteee: point-wise consistency, i.e. as sample size tends to infinity, the Markov equivalence class of the true graph can be identified

Discovery guarantees

- Completeness: Given the true (conditional) independence and dependence relations [the algorithm] identifies all there is to discover about the true underlying graph, namely, its Markov equivalence class.

- Statistical guaranteee: point-wise consistency, i.e. as sample size tends to infinity, the Markov equivalence class of the true graph can be identified

Very weak convergence guarantee

Discovery guarantees

- Completeness: Given the true (conditional) independence and dependence relations [the algorithm] identifies all there is to discover about the true underlying graph, namely, its Markov equivalence class.

- Statistical guaranteee: point-wise consistency, i.e. as sample size tends to infinity, the Markov equivalence class of the true graph can be identified

Very weak convergence guarantee

Robins et al (2003): Tough luck, this is as good as it gets (for any method) given the set of assumptions.

Discovery guarantees

- Completeness: Given the true (conditional) independence and dependence relations [the algorithm] identifies all there is to discover about the true underlying graph, namely, its Markov equivalence class.

- Statistical guaranteee: point-wise consistency, i.e. as sample size tends to infinity, the Markov equivalence class of the true graph can be identified

Very weak convergence guarantee

true graph

close to unfaithful graph

$X \perp\!\!\!\perp Y$

Robins et al (2003): Tough luck, this is as good as it gets (for any method) given the set of assumptions.

Using faithfulness

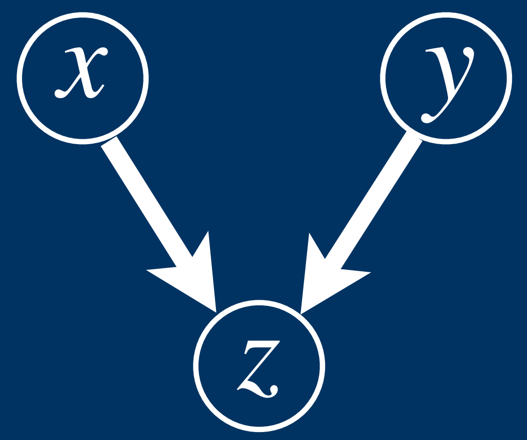
assumption/ algorithm	PC / GES	FCI	CCD	LiNGaM	lvLiNGaM	cyclic LiNGaM	non-linear additive noise	SAT
Markov	✓	✓	✓	✓	✓	✓	✓	✓
faithfulness	✓	✓	✓	✗	✓	~	minimality	✓♣
causal sufficiency	✓	✗	✓	✓	✗	✓	✓	✗
acyclicity	✓	✓	✗*	✓	✓	✗	✓	✗*
parametric assumption	✗	✗	✗	linear non-Gaussian	linear non-Gaussian	linear non-Gaussian	non-linear additive noise	✗
output	Markov equivalence	PAG	PAG	unique DAG	set of DAGs	set of graphs	unique DAG	query based

~ special case
* care needs to be taken how cyclicity is modeled
✤ there are approaches that weaken faithfulness

Search for the sparsest permutation

Raskutti & Uhler, 2013/18; Solus et al. 2017

DAG G



Associate a DAG with each permutation π and distribution $\mathscr{P}$:

$$\pi_i \to \pi_j \in G \iff i < j \text{ and } \pi_i \not\perp\!\!\!\perp \pi_j \mid \{\pi_1, \ldots, \pi_j\} \setminus \{\pi_i, \pi_j\}$$

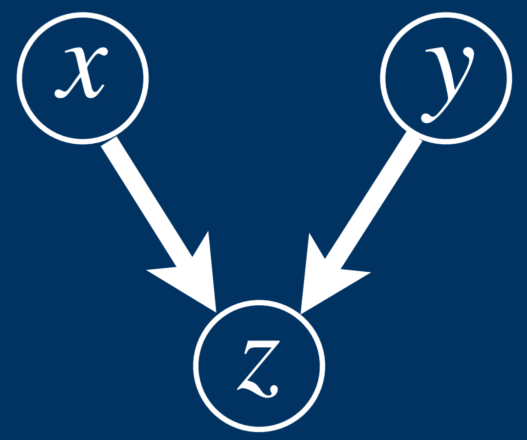
Permutation

$$\pi = xyz$$

Search for the sparsest permutation

Raskutti & Uhler, 2013/18; Solus et al. 2017

DAG G



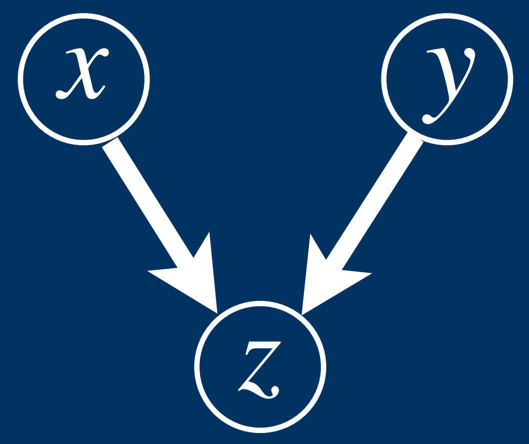
Associate a DAG with each permutation π and distribution $\mathscr{P}$:

$$\pi_i \rightarrow \pi_j \in G \iff i < j \text{ and } \pi_i \not\perp\!\!\!\perp \pi_j \,|\, \{\pi_1, \ldots, \pi_j\} \backslash \{\pi_i, \pi_j\}$$

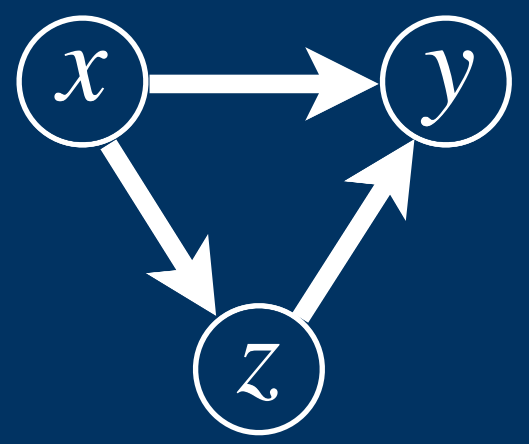
Permutation

$\pi = xyz$

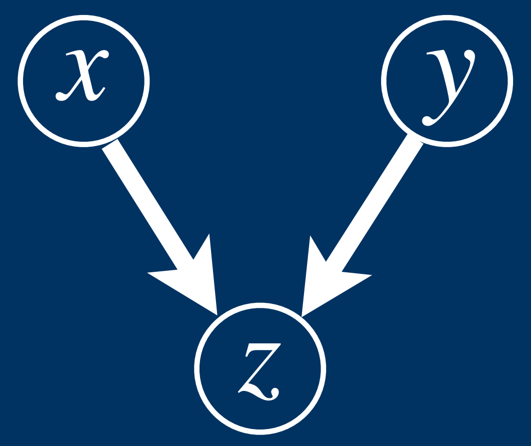
$\pi = xyz$



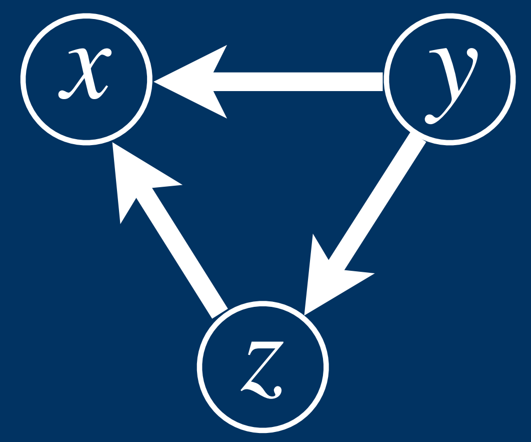
$\pi = xzy$



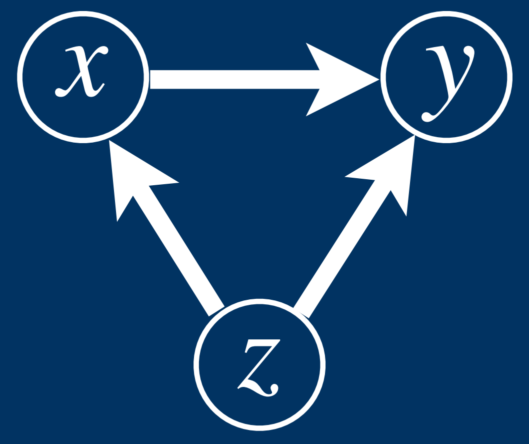
$\pi = yxz$



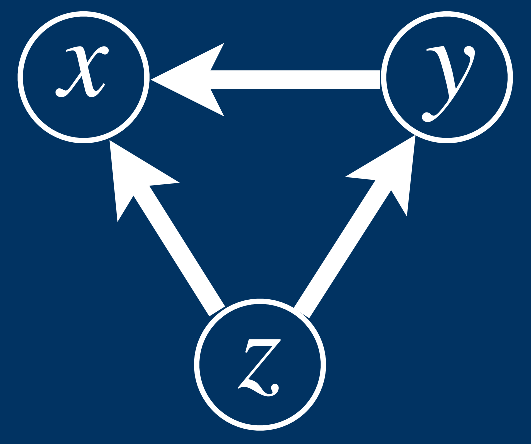
$\pi = yzx$



$\pi = zxy$



$\pi = zyx$

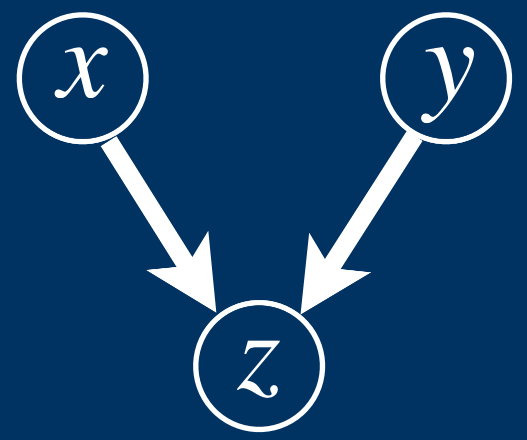


$$\text{Maximize score}(G, \mathscr{P}) = \begin{cases} -|G| \text{ if } G \text{ is Markov to } \mathscr{P} \\ -\infty \text{ otherwise} \end{cases}$$

Search for the sparsest permutation

Raskutti & Uhler, 2013/18; Solus et al. 2017

DAG G



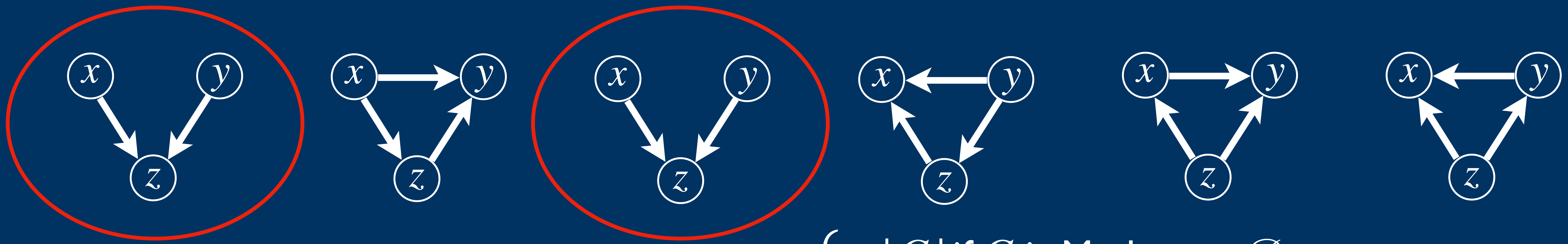
Associate a DAG with each permutation π and distribution $\mathscr{P}$:

$$\pi_i \to \pi_j \in G \iff i < j \text{ and } \pi_i \not\perp\!\!\!\perp \pi_j \mid \{\pi_1, \ldots, \pi_j\} \setminus \{\pi_i, \pi_j\}$$

Permutation

$\pi = xyz$

$\pi = xyz$ $\quad$ $\pi = xzy$ $\quad$ $\pi = yxz$ $\quad$ $\pi = yzx$ $\quad$ $\pi = zxy$ $\quad$ $\pi = zyx$



$$\text{Maximize score}(G, \mathscr{P}) = \begin{cases} -|G| \text{ if } G \text{ is Markov to } \mathscr{P} \\ -\infty \text{ otherwise} \end{cases}$$

Sparse Permutation Search

- Completeness: Given the true (conditional) independence and dependence relations (greedy) sparse permutation search identifies all there is to discover about the true underlying graph, namely, its Markov equivalence class.

Sparse Permutation Search

- Completeness: Given the true (conditional) independence and dependence relations (greedy) sparse permutation search identifies all there is to discover about the true underlying graph, namely, its Markov equivalence class.

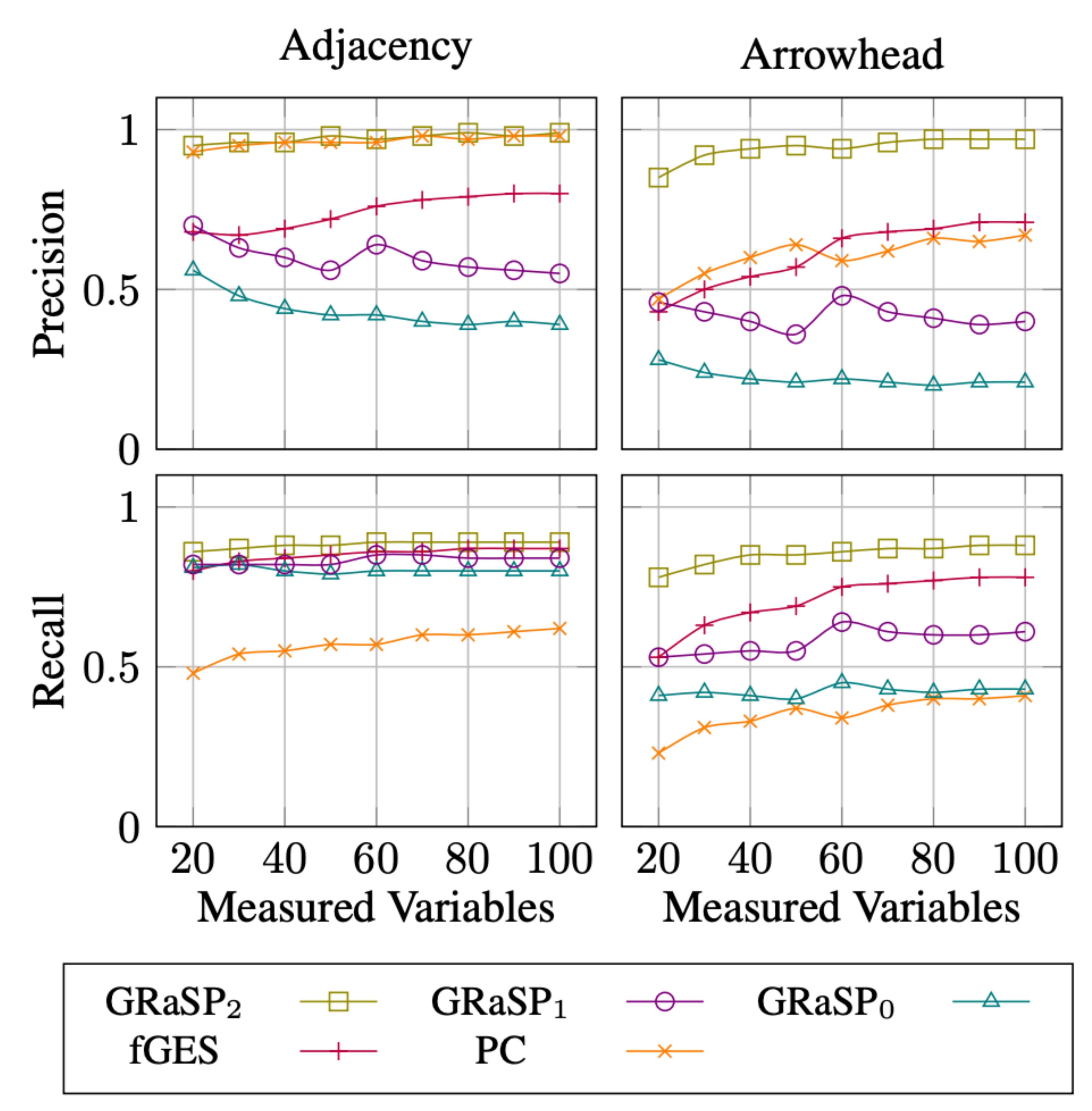
- Statistical guarantee:
  - Point-wise consistency with an assumption strictly weaker than faithfulness ("unique frugality" / "sparsest Markov representation")
  - Uniform consistency with a slight strengthening of faithfulness

Sparse Permutation Search

- Completeness: Given the true (conditional) independence and dependence relations (greedy) sparse permutation search identifies all there is to discover about the true underlying graph, namely, its Markov equivalence class.

- Statistical guarantee:
  - Point-wise consistency with an assumption strictly weaker than faithfulness ("unique frugality" / "sparsest Markov representation")
  - Uniform consistency with a slight strengthening of faithfulness

- Accuracy and Scalability



Lam et al., 2022

Causal Discovery

assumption/ algorithm	PC / GES	sparse permutation search	FCI	CCD	LiNGaM	lvLiNGaM	cyclic LiNGaM	non-linear additive noise	SAT
Markov	✓	✓	✓	✓	✓	✓	✓	✓	✓
faithfulness	✓	u-frugality	✓	✓	✗	✓	~	minimality	✓✣
causal sufficiency	✓	✓	✗	✓	✓	✗	✓	✓	✗
acyclicity	✓	✓	✓	✗*	✓	✓	✗	✓	✗*
parametric assumption	✗	✗	✗	✗	linear non-Gaussian	linear non-Gaussian	linear non-Gaussian	non-linear additive noise	✗
output	Markov equivalence class		PAG	PAG	unique DAG	set of DAGs	set of graphs	unique DAG	query based

~ special case
* care needs to be taken how cyclicity is modeled
✣ there are approaches that weaken faithfulness

Exploiting the parametric form

assumption/ algorithm	PC / GES	sparse permutation search	FCI	CCD	LiNGaM	lvLiNGaM	cyclic LiNGaM	non-linear additive noise	SAT
Markov	✓	✓	✓	✓	✓	✓	✓	✓	✓
faithfulness	✓	u-frugality	✓	✓	✗	✓	~	minimality	✓♣
causal sufficiency	✓	✓	✗	✓	✓	✗	✓	✓	✗
acyclicity	✓	✓	✓	✗*	✓	✓	✗	✓	✗*
parametric assumption	✗	✗	✗	✗	linear non-Gaussian	linear non-Gaussian	linear non-Gaussian	non-linear additive noise	✗
output	Markov equivalence class		PAG	PAG	unique DAG	set of DAGs	set of graphs	unique DAG	query based

~ special case
* care needs to be taken how cyclicity is modeled
✤ there are approaches that weaken faithfulness

Linear Non-Gaussian Models (LinGaM)

- Linear causal relations:

$$x_i = \sum_{x_j \in Pa(x_i)} \beta_{ij} x_j + \epsilon_i$$

- Assumptions:
  - causal Markov
  - causal sufficiency
  - acyclicity

Linear Non-Gaussian Models (LinGaM)

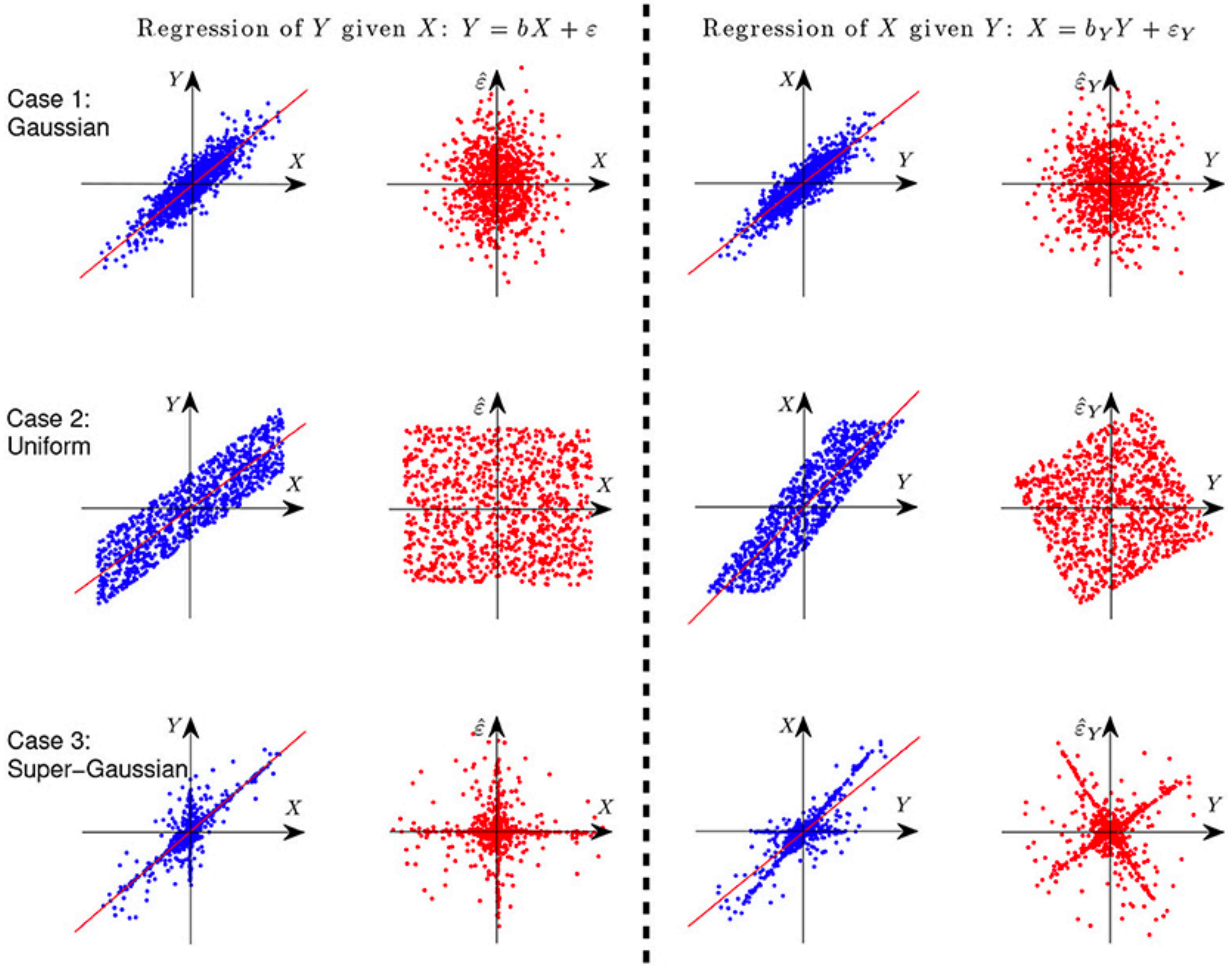
- Linear causal relations:

$$x_i = \sum_{x_j \in Pa(x_i)} \beta_{ij} x_j + \epsilon_i$$

- Assumptions:
  - causal Markov
  - causal sufficiency
  - acyclicity

- If $\epsilon_i \sim$ non-Gaussian and independent, then the true graph is **uniquely identifiable** from the joint distribution.

Shimizu et al, 2006



(correct) forward model

$x \perp\!\!\!\perp \hat{\epsilon}_y$

(false) backward model

$y \not\!\perp\!\!\!\perp \hat{\epsilon}_x$

(figure from Glymour et al. 2019)

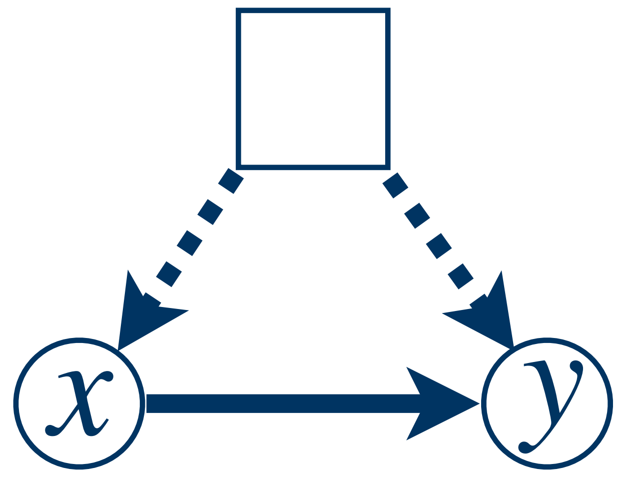
Linear Non-Gaussian Models (LinGaM)

- Linear causal relations:

$$x_i = \sum_{x_j \in Pa(x_i)} \beta_{ij} x_j + \epsilon_i$$

- Assumptions:
  - causal Markov
  - ~~causal sufficiency~~
  - acyclicity

Confounding



- The residual of a linear regression of the effect on the cause will be dependent with the cause IFF there is confounding of the cause and effect.

$$x \not\perp\!\!\!\perp \hat{\epsilon}_y$$

Tashiro et al, 2014

Independent Noise

Linear Non-Gaussian (Lingam):

- forwards model $y = ax + \epsilon_y \quad x \perp\!\!\!\perp \epsilon_y$
- backwards model. $x = by + \tilde{\epsilon}_x \quad y \not\!\perp\!\!\!\perp \tilde{\epsilon}_x$

Confounding in Lingam

- Unconfounded forwards model $x \perp\!\!\!\perp \epsilon_y$
- Confounded forwards model $y \not\!\perp\!\!\!\perp \tilde{\epsilon}_x$

Independent Noise

Linear Non-Gaussian (Lingam):

- forwards model $y = ax + \epsilon_y$ $x \perp\!\!\!\perp \epsilon_y$
- backwards model. $x = by + \tilde{\epsilon}_x$ $y \not\!\perp\!\!\!\perp \tilde{\epsilon}_x$

Confounding in Lingam

- Unconfounded forwards model $x \perp\!\!\!\perp \epsilon_y$
- Confounded forwards model $y \not\!\perp\!\!\!\perp \tilde{\epsilon}_x$

GIN-condition in Lingam with latents (Xie et al 2020):

- Two variable sets $\mathbf{Y}, \mathbf{Z}$ satisfy GIN iff $E_{\mathbf{Y}||\mathbf{Z}} \perp\!\!\!\perp \mathbf{Z}$, where $E_{\mathbf{Y}||\mathbf{Z}}$ is a "cleaned up" version of $\mathbf{Y}$.

➡ Satisfaction of GIN permits remarkable discovery of latent variable structure

Independent Noise

Is there an underlying **motivation or justification** why an independence between cause and noise on the effect is desirable?

Linear Non-Gaussian (Lingam):

- forwards model $y = ax + \epsilon_y$ $\quad x \perp\!\!\!\perp \epsilon_y$
- backwards model. $x = by + \tilde{\epsilon}_x$ $\quad y \not\perp\!\!\!\perp \tilde{\epsilon}_x$

Confounding in Lingam

- Unconfounded forwards model $x \perp\!\!\!\perp \epsilon_y$
- Confounded forwards model $y \not\perp\!\!\!\perp \tilde{\epsilon}_x$

GIN-condition in Lingam with latents (Xie et al 2020):

- Two variable sets $\mathbf{Y}, \mathbf{Z}$ satisfy GIN iff $E_{\mathbf{Y}||\mathbf{Z}} \perp\!\!\!\perp \mathbf{Z}$, where $E_{\mathbf{Y}||\mathbf{Z}}$ is a "cleaned up" version of $\mathbf{Y}$.

➡ Satisfaction of GIN permits remarkable discovery of latent variable structure

Independent Noise

Linear Non-Gaussian (Lingam):

- forwards model $y = ax + \epsilon_y \quad x \perp\!\!\!\perp \epsilon_y$
- backwards model. $x = by + \tilde{\epsilon}_x \quad y \not\!\perp\!\!\!\perp \tilde{\epsilon}_x$

Confounding in Lingam

- Unconfounded forwards model $x \perp\!\!\!\perp \epsilon_y$
- Confounded forwards model $y \not\!\perp\!\!\!\perp \tilde{\epsilon}_x$

GIN-condition in Lingam with latents (Xie et al 2020):

- Two variable sets $\mathbf{Y}, \mathbf{Z}$ satisfy GIN iff $E_{\mathbf{Y}||\mathbf{Z}} \perp\!\!\!\perp \mathbf{Z}$, where $E_{\mathbf{Y}||\mathbf{Z}}$ is a "cleaned up" version of $\mathbf{Y}$.

➡ Satisfaction of GIN permits remarkable discovery of latent variable structure

Is there an underlying motivation or justification why an independence between cause and noise on the effect is desirable?

It clearly is not generally satisfied: **heteroskedastic** noise can arise from an interactive effect between the cause and noise

Independent Noise

Linear Non-Gaussian (Lingam):

- forwards model $y = ax + \epsilon_y \quad x \perp\!\!\!\perp \epsilon_y$
- backwards model. $x = by + \tilde{\epsilon}_x \quad y \not\perp\!\!\!\perp \tilde{\epsilon}_x$

Confounding in Lingam

- Unconfounded forwards model $x \perp\!\!\!\perp \epsilon_y$
- Confounded forwards model $y \not\perp\!\!\!\perp \tilde{\epsilon}_x$

GIN-condition in Lingam with latents (Xie et al 2020):

- Two variable sets $\mathbf{Y}, \mathbf{Z}$ satisfy GIN iff $E_{\mathbf{Y}||\mathbf{Z}} \perp\!\!\!\perp \mathbf{Z}$, where $E_{\mathbf{Y}||\mathbf{Z}}$ is a "cleaned up" version of $\mathbf{Y}$.

➡ Satisfaction of GIN permits remarkable discovery of latent variable structure

Is there an underlying motivation or justification why an independence between cause and noise on the effect is desirable?

It clearly is not generally satisfied: heteroskedastic noise can arise from an interactive effect between the cause and noise

But that violates the functional assumption of the Lingam model.

Independent Noise

Linear Non-Gaussian (Lingam):

- forwards model $y = ax + \epsilon_y \quad x \perp\!\!\!\perp \epsilon_y$
- backwards model. $x = by + \tilde{\epsilon}_x \quad y \not\perp\!\!\!\perp \tilde{\epsilon}_x$

Confounding in Lingam

- Unconfounded forwards model $x \perp\!\!\!\perp \epsilon_y$
- Confounded forwards model $y \not\perp\!\!\!\perp \tilde{\epsilon}_x$

GIN-condition in Lingam with latents (Xie et al 2020):

- Two variable sets $\mathbf{Y}, \mathbf{Z}$ satisfy GIN iff $E_{\mathbf{Y}||\mathbf{Z}} \perp\!\!\!\perp \mathbf{Z}$, where $E_{\mathbf{Y}||\mathbf{Z}}$ is a "cleaned up" version of $\mathbf{Y}$.

➡ Satisfaction of GIN permits remarkable discovery of latent variable structure

Is there an underlying motivation or justification why an independence between cause and noise on the effect is desirable?

It clearly is not generally satisfied: heteroskedastic noise can arise from an interactive effect between the cause and noise

But that violates the functional assumption of the Lingam model.

Suggestion: Searching for the independence between cause and noise is, within the Lingam model, an application of the **Principle of Independent Mechanisms.**

see Besserve et al, 2018, Sec 4.3

Principle of Independent Mechanisms

- The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other. (Peters et al. 2017, Janzing et al. 2008)

P(X) is "uninformative" of P(Y|X)

Principle of Independent Mechanisms

- The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other. (Peters et al. 2017, Janzing et al. 2008)

P(X) is "uninformative" of P(Y|X)

Lingam:

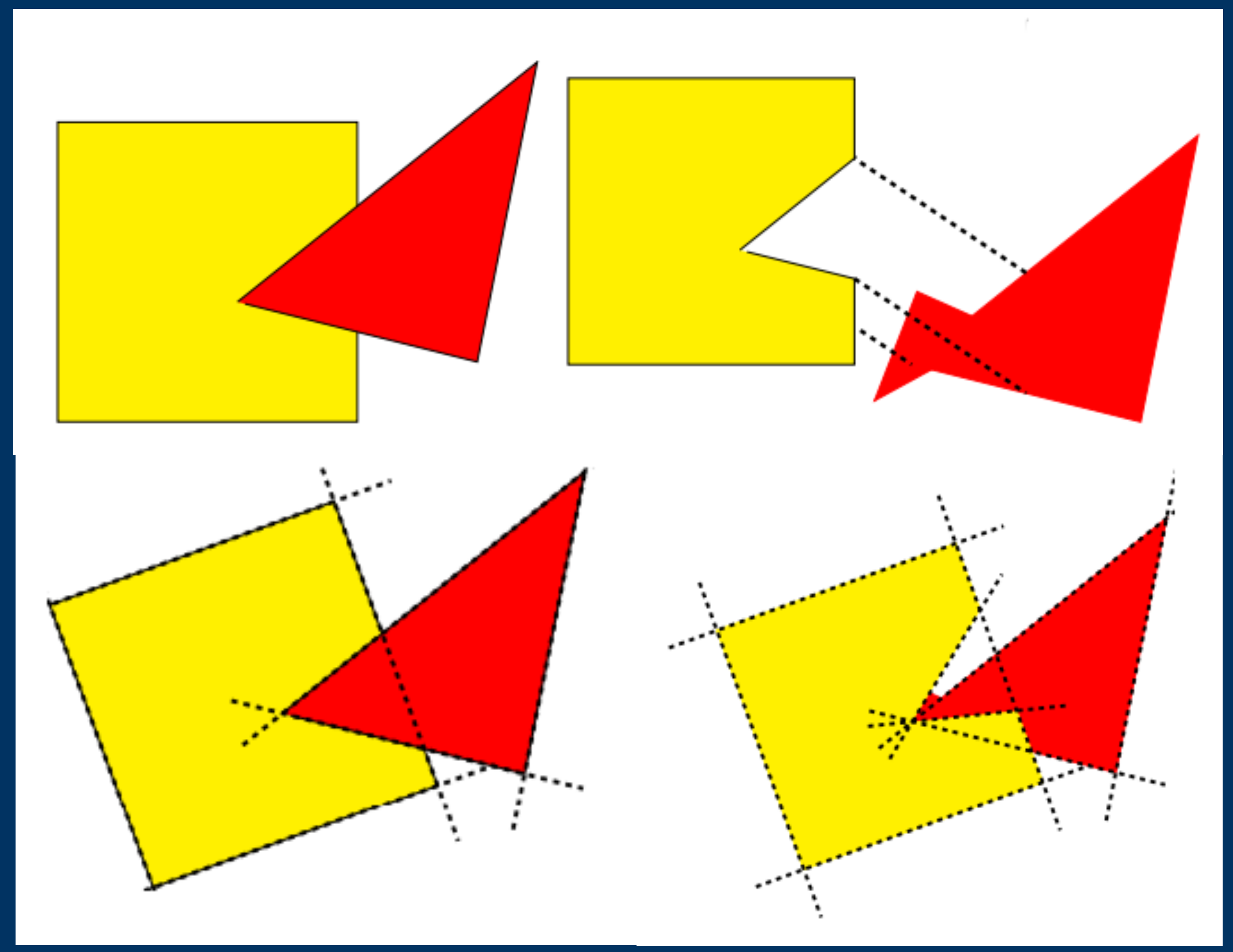
- In the Lingam model, assessing whether P(X) is informative about P(Y|X) amounts to assessing whether P(X) is informative about P(ϵ)

Principle of Independent Mechanisms

- The causal generative process of a system's variables is composed of autonomous modules that do not inform or influence each other. (Peters et al. 2017, Janzing et al. 2008)

P(X) is "uninformative" of P(Y|X)

How to assess PIM for more general model classes:



- Group Invariance for Causal Discovery (Besserve et al 2018)

  ➡ Use generic group transformations of X to assess whether the observed relation between P(X) and P(Y|X) is expected
- Independent Mechanism Analysis (Gresele et al 2022)

Approaches using Independent Mechanisms

- Inferential power: Extraordinary results on what can and cannot be identified, including about latent causal structure.

Using rather strong assumptions about the functional form.

Approaches using Independent Mechanisms

- Inferential power: Extraordinary results on what can and cannot be identified, including about latent causal structure.

Using rather strong assumptions about the functional form.

- Potential for generalizations: The connection to approaches based on the principle of independent mechanisms raises the hope that maybe the strong parametric assumptions can be made much more generic.

Likely to be computationally very intensive and it remains unclear what sorts of statistical guarantees may be forthcoming.

Approaches using Independent Mechanisms

- Inferential power: Extraordinary results on what can and cannot be identified, including about latent causal structure.

Using rather strong assumptions about the functional form.

- Potential for generalizations: The connection to approaches based on the principle of independent mechanisms raises the hope that maybe the strong parametric assumptions can be made much more generic.

Likely to be computationally very intensive and it remains unclear what sorts of statistical guarantees may be forthcoming.

- Criticisms of PIM apply perhaps more broadly: Mechanisms that have been subject to evolutionary pressures, are unlikely to exhibit the independence required by PIM; presumably a similar argument applies for social settings.

If the search for independent noise in the Lingam setting is an application of PIM, then these concerns may carry over to Lingam-based methods.

ML joins the discovery game: NOTEARS

ML joins the discovery game: NOTEARS

assumption/ algorithm	PC / GES	sparse permutation search	FCI	CCD	LiNGaM	lvLiNGaM	cyclic LiNGaM	non-linear additive noise	SAT
Markov	✓	✓	✓	✓	✓	✓	✓	✓	✓
faithfulness	✓	u-frugality	✓				~	minimality	✓♣
causal sufficiency	✓	✓	✗	✓	✓	✗	✓	✓	✗
acyclicity	✓	✓	✓	✗*	✓	✓	✗	✓	✗*
parametric assumption	✗	✗	✗	✗	linear non-Gaussian	linear non-Gaussian	linear non-Gaussian	non-linear additive noise	✗
output	Markov equivalence class		PAG	PAG	unique DAG	set of DAGs	set of graphs	unique DAG	query based

In its original form, NOTEARS was designed as a method for DAG search.

~ special case
* care needs to be taken how cyclicity is modeled
✤ there are approaches that weaken faithfulness

NOTEARS

Zheng et al, 2018

consistent score

(weighted) adjacency matrix

Previous methods: $\min_{W \in \mathbb{R}^{d \times d}} S(W)$ subject to G(W) being a DAG

combinatorial optimization

NOTEARS

Zheng et al, 2018

consistent score

(weighted) adjacency matrix

differentiable function that is 0 iff W represents a DAG

NOTEARS: $\min_{W \in \mathbb{R}^{d \times d}} S(W)$ subject to h(W) =0

continuous optimization

$$\Longleftrightarrow$$

Previous methods: $\min_{W \in \mathbb{R}^{d \times d}} S(W)$ subject to G(W) being a DAG

combinatorial optimization

NOTEARS

Zheng et al, 2018

consistent score

(weighted) adjacency matrix

differentiable function that is 0 iff W represents a DAG

NOTEARS: $\min_{W \in \mathbb{R}^{d \times d}} S(W)$ subject to h(W) =0

continuous optimization

In the linear case: $h(W) = tr(e^{W \circ W}) - d$

matrix exponential of Hadamard product

NOTEARS

Zheng et al, 2018

consistent score

(weighted) adjacency matrix

differentiable function that is 0 iff W represents a DAG

NOTEARS: $\min_{W \in \mathbb{R}^{d \times d}} S(W)$ subject to h(W) =0

continuous optimization

In the linear case: $h(W) = tr(e^{W \circ W}) - d$

matrix exponential of Hadamard product

Why does this function have a gradient towards being a DAG?

NOTEARS

Zheng et al, 2018

consistent score

(weighted) adjacency matrix

differentiable function that is 0 iff W represents a DAG

NOTEARS: $\min_{W \in \mathbb{R}^{d \times d}} S(W)$ subject to h(W) =0

continuous optimization

In the linear case: $h(W) = tr(e^{W \circ W}) - d$

matrix exponential of Hadamard product

Why does this function have a gradient towards being a DAG?

- Matrix exponential e^B is a geometric series of ever higher B^k
- In a linear system $\mathbf{x} = B\mathbf{x} + \epsilon, B^k$ represents the paths of length k
- The trace sums the weighted paths from a node to itself
- The Hadamard product ensures that the sum is over positive quantities

NOTEARS

Zheng et al, 2018

consistent score

(weighted) adjacency matrix

differentiable function that is 0 iff W represents a DAG

NOTEARS: $\min_{W \in \mathbb{R}^{d \times d}} S(W)$ subject to h(W) =0

continuous optimization

In the linear case: $h(W) = tr(e^{W \circ W}) - d$

matrix exponential of Hadamard product

Derivative is simple ➡ OPTIMIZE!

Non-convex, so use your tricks!

ML joins the discovery game: NOTEARS

assumption/ algorithm	PC / GES	sparse permutation search	NOTEARS	FCI	CCD	LiNGaM	lvLiNGaM	cyclic LiNGaM	non-linear additive noise	SAT
Markov	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
faithfulness	✓	u-frugality	✓?	✓	✓	✗	✓	~	minimality	✓✣
causal sufficiency	✓	✓	✓	✗	✓	✓	✗	✓	✓	✗
acyclicity	✓	✓	✓	✓	✗*	✓	✓	✗	✓	✗*
parametric assumption	✗	✗	?	✗	✗	linear non-Gaussian	linear non-Gaussian	linear non-Gaussian	non-linear additive noise	✗
output	Markov equivalence class		DAG, but…	PAG	PAG	unique DAG	set of DAGs	set of graphs	unique DAG	query based

~ special case
* care needs to be taken how cyclicity is modeled
✣ there are approaches that weaken faithfulness

ML joins the discovery game: NOTEARS

assumption/ algorithm	PC / GES	sparse permutation search	NOTEARS	FCI	CCD	LiNGaM	lvLiNGaM	cyclic LiNGaM	non-linear additive noise	SAT
Markov	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
faithfulness	✓	u-frugality	✓?	✓	✓	✗	✓	~	minimality	✓✣
causal sufficiency	✓	✓	✓	✗					✓	✗
acyclicity	✓	✓	✓	✓					✓	✗*
parametric assumption	✗	✗	?	✗	✗	non-Gaussian	non-Gaussian	non-Gaussian	non-linear additive noise	✗
output	Markov equivalence class		DAG, but…	PAG	PAG	unique DAG	set of DAGs	set of graphs	unique DAG	query based

The optimization constraint requires a model parameterization, and there are several for linear models. But of course other methods also need a score or type of independence test.

~ special case
* care needs to be taken how cyclicity is modeled
✣ there are approaches that weaken faithfulness

ML joins the discovery game: NOTEARS

assumption/ algorithm	PC / GES	sparse permutation search	NOTEARS	FCI	CCD	LiNGaM	lvLiNGaM	cyclic LiNGaM	non-linear additive noise	SAT
Markov	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
faithfulness	✓	u-frugality	✓?	✓	✓	✗	✓	~	minimality	✓✣
causal sufficiency	✓	✓	✓	✗					✓	✗
acyclicity	✓	✓	✓	✓					✓	✗*
parametric assumption	✗	✗	?	✗	✗	linear non-Gaussian	linear non-Gaussian	linear non-Gaussian	non-linear additive noise	✗
output	Markov equivalence class		DAG, but…	PAC	P…			…of …phs	unique DAG	query based

The optimization constraint requires a model parameterization, and there are several for linear models. But of course other methods also need a score or type of independence test.

NOTEARS returns a DAG, but the results are still limited to Markov equivalence.

~ special case
* care needs to be taken how cyclicity is modeled
✣ there are approaches that weaken faithfulness

NOTEARS and its variants

Change the score.

NOTEARS: $\min_{W \in \mathbb{R}^{d \times d}} S(W)$ subject to h(W) =0

Change the constraint that describes acyclicity

In the linear case: $h(W) = tr(e^{W \circ W}) - d$

Derivative is simple ➡ OPTIMIZE!

Change how the optimization is done.

NOTEARS and its variants

Change the score.

NOTEARS: $$\min_{W \in \mathbb{R}^{d \times d}} S(W) \quad \text{subject to } h(W) = 0$$

Change the constraint that describes acyclicity

In the linear case: $h(W) = tr(e^{W \circ W}) - d$

Derivative is simple ➡ OPTIMIZE!

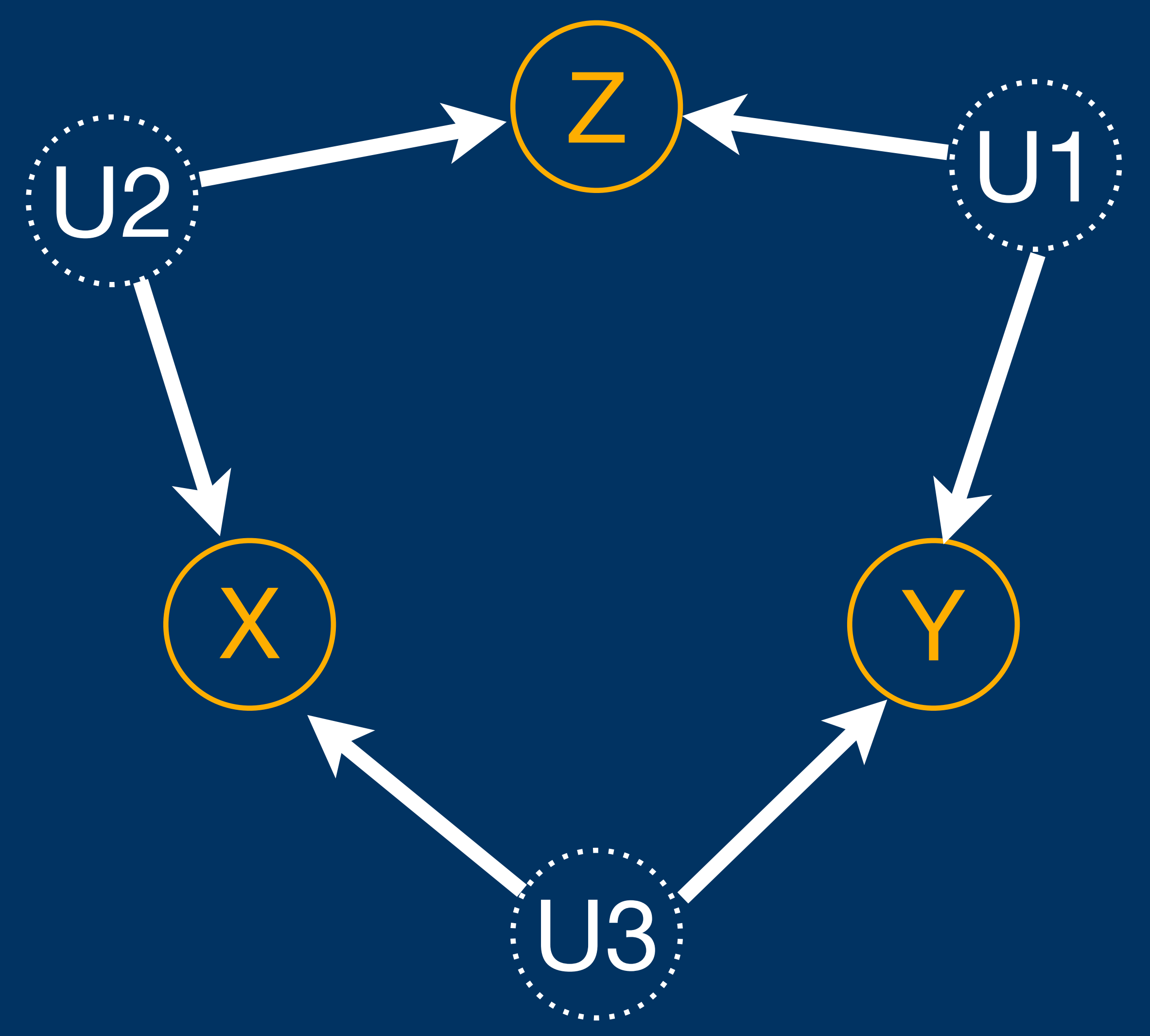
Change how the optimization is done.

Method	Year	Data	Acycl.	Interv.	Output
CMS [152]	2014	low	-	no	Bi
NO TEARS [267]	2018	low	yes	no	DAG
CGNN [75]	2018	low	yes	no	DAG
Graphite [83]	2019	low/medium	no	no	UG
SAM [122]	2019	low/medium	yes	no	DAG
DAG-GNN [262]	2019	low	yes	no	DAG
GAE [177]	2019	low	yes	no	DAG
NO BEARS [142]	2019	low/medium/high	yes	no	DAG
Meta-Transfer [19]	2019	Bi	yes	yes	Bi
DEAR [214]	2020	high	yes	no	-
CAN [167]	2020	low/medium/high	yes	no	DAG
NO FEARS [251]	2020	low	yes	no	DAG
GOLEM [176]	2020	low	yes	no	DAG
ABIC [20]	2020	low	yes	no	ADMG/PAG
DYNOTEARS [178]	2020	low	yes	no	SVAR
SDI [124]	2020	low	yes	yes	DAG
AEQ [64]	2020	Bi	-	no	direction
RL-BIC [272]	2020	low	yes	no	DAG
CRN [125]	2020	low	yes	yes	DAG
ACD [151]	2020	low	Granger	no	time-series DAG
V-CDN [145]	2020	high	Granger	no	time-series DAG
CASTLE (reg.) [138]	2020	low/medium	yes	no	DAG
GranDAG [139]	2020	low	yes	no	DAG
MaskedNN [175]	2020	low	yes	no	DAG
CausalVAE [257]	2020	high	yes	yes	DAG
CAREFL [126]	2020	low	yes	no	DAG / Bi
Varando [244]	2020	low	yes	no	DAG
NO TEARS+ [268]	2020	low	yes	no	DAG
ICL [250]	2020	low	yes	no	DAG
LEAST [271]	2020	low/medium/high	yes	no	DAG

Continuous optimization-based approaches to causal discovery (Vowels et al. 2021)

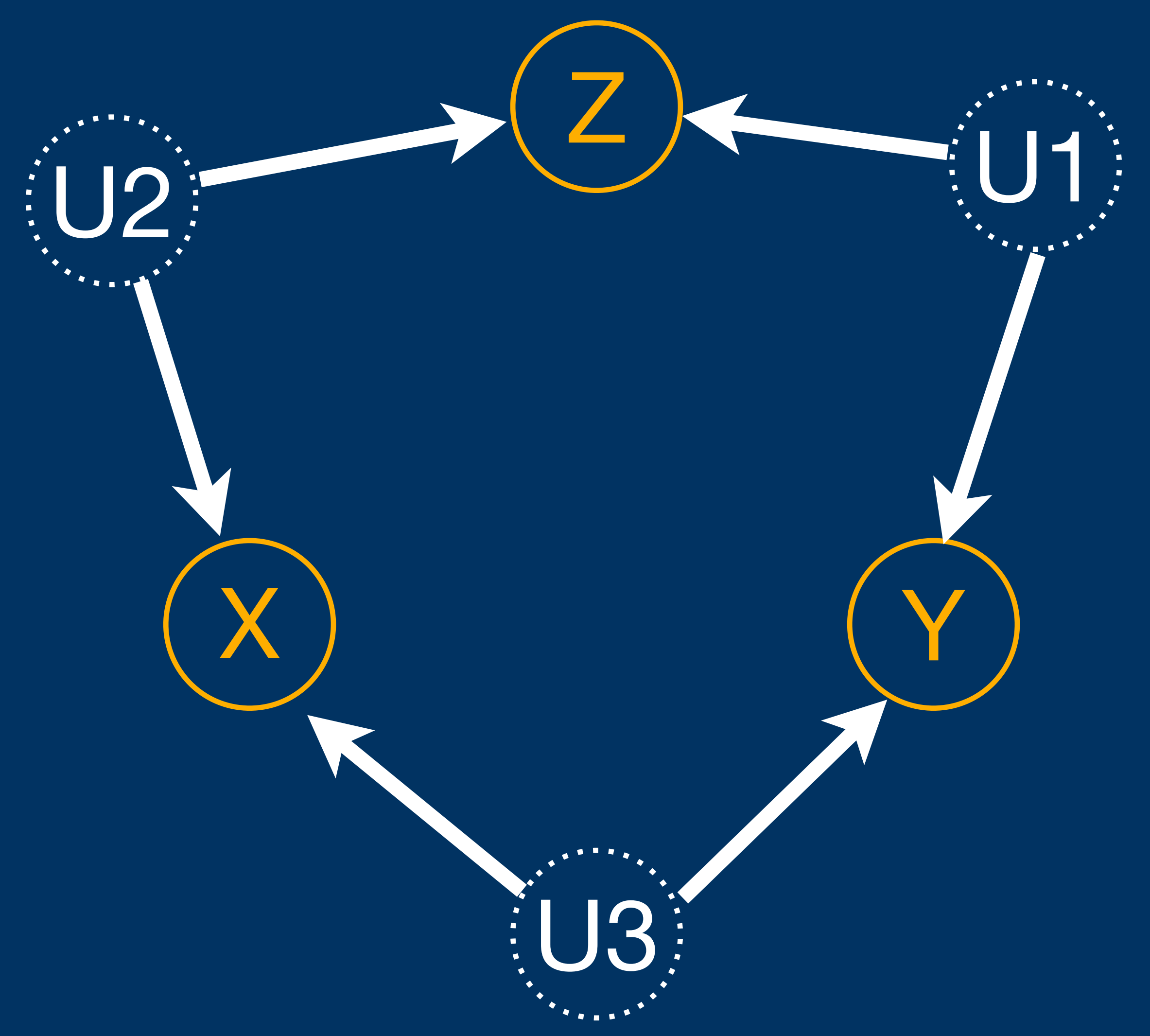
Inflation (for discrete finite probability spaces)

Wolfe, 2017; Navascués & Wolfe 2020



Inflation (for discrete finite probability spaces)

Wolfe, 2017; Navascués & Wolfe 2020

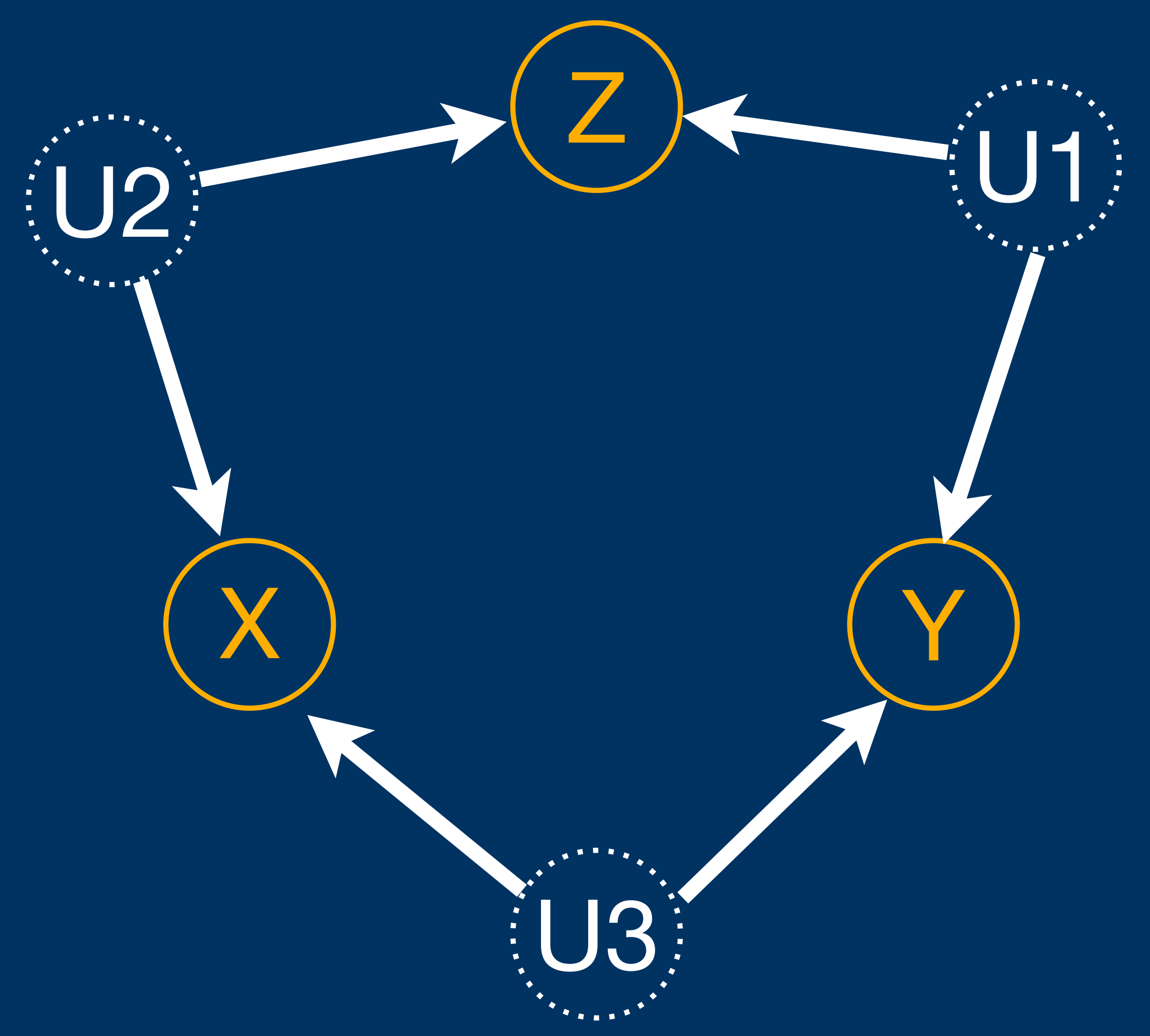


compatible?

$$P(X, Y, Z)$$

Inflation (for discrete finite probability spaces)

Wolfe, 2017; Navascués & Wolfe 2020



compatible?

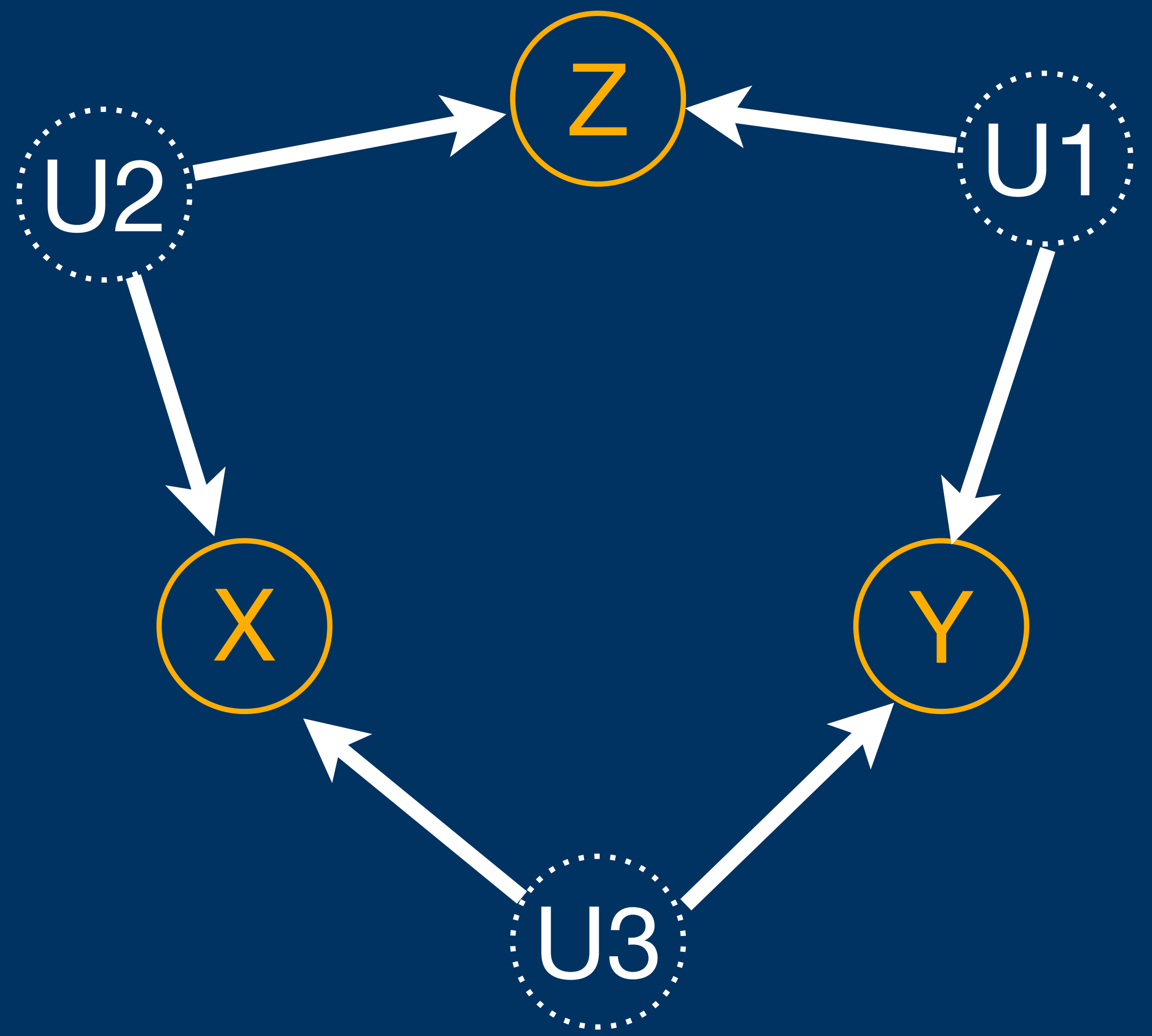
$$P(X, Y, Z)$$

$$P(X, Y, Z) = \sum_{U_1, U_2, U_3} P(U_1)P(U_2)P(U_3)P(X \mid U_2, U_3)P(Y \mid U_1, U_3)P(Z \mid U_1, U_2)$$

If the observed variables have finite cardinality, then the distributions P compatible with G form a semi-algebraic set.

Inflation (for discrete finite probability spaces)

Wolfe, 2017; Navascués & Wolfe 2020



compatible?

$$P(X, Y, Z)$$

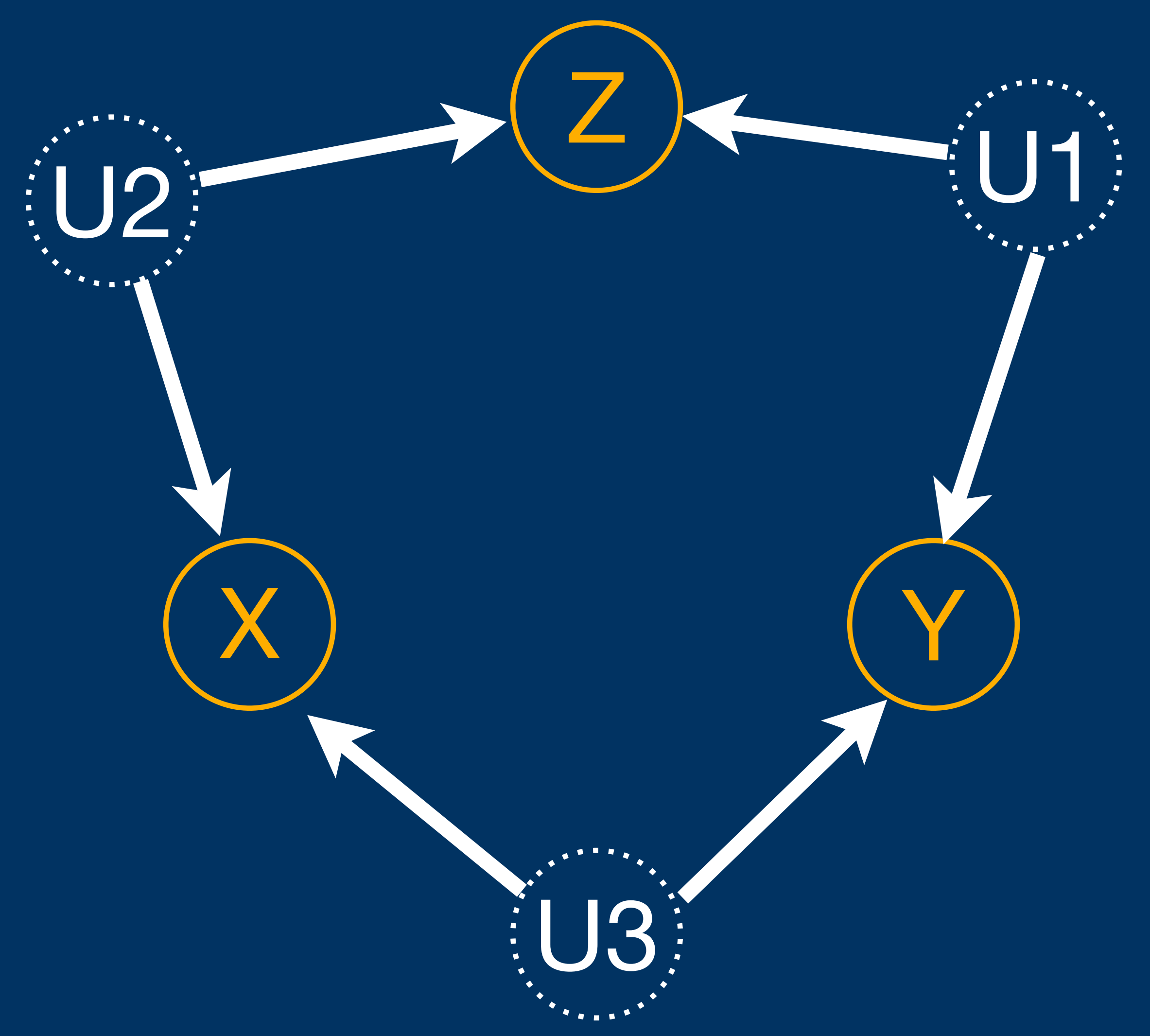
$$P(X, Y, Z) = \sum_{U_1, U_2, U_3} P(U_1)P(U_2)P(U_3)P(X \mid U_2, U_3)P(Y \mid U_1, U_3)P(Z \mid U_1, U_2)$$

If the observed variables have finite cardinality, then the distributions P compatible with G form a semi-algebraic set.

It follows that the set of distributions can be characterized by a finite set of polynomial inequalities.

Inflation (for discrete finite probability spaces)

Wolfe, 2017; Navascués & Wolfe 2020



compatible?

$$P(X, Y, Z)$$

$$P(X, Y, Z) = \sum_{U_1, U_2, U_3} P(U_1)P(U_2)P(U_3)P(X \mid U_2, U_3)P(Y \mid U_1, U_3)P(Z \mid U_1, U_2)$$

If the observed variables have finite cardinality, then the distributions P compatible with G form a semi-algebraic set.

It follows that the set of distributions can be characterized by a finite set of polynomial inequalities.

➡ **Inflation** is a technique that iteratively identifies **all** these constraints.

Inflation

- Include inequality constraints in causal discovery
- Technique for testing latent variable models
- Potential to advance causal discovery in the categorical setting.
- Important connections to questions in quantum mechanics.

Inflation

- Include inequality constraints in causal discovery
- Technique for testing latent variable models
- Potential to advance causal discovery in the categorical setting.
- Important connections to questions in quantum mechanics.

- I did not say it was efficient.
- Interesting questions about how to test for the inequalities in practice.

Comments

Causal discovery needs:

- contributions to address foundational challenges, such as reliable and fast non-parametric conditional independence tests
- Well-maintained code bases that are easily manipulable
- More users who actually apply the methods to a real scientific problem and publish the results in that scientific discipline

A huge shout-out to the **pcalg** group at ETH and the **Tetrad** group at CMU.

References

- Gillispie, Steven B., and Michael D. Perlman. "The size distribution for Markov equivalence classes of acyclic digraph models." *Artificial Intelligence* 141.1-2 (2002): 137-155.
- He, Yangbo, Jinzhu Jia, and Bin Yu. "Counting and exploring sizes of Markov equivalence classes of directed acyclic graphs." The Journal of Machine Learning Research 16.1 (2015): 2589-2609.
- Radhakrishnan, Adityanarayanan, Liam Solus, and Caroline Uhler. "Counting Markov equivalence classes for DAG models on trees." Discrete Applied Mathematics 244 (2018): 170-185.
- Robins, J. M., Scheines, R., Spirtes, P., & Wasserman, L. (2003). Uniform consistency in causal inference. Biometrika, 90(3), 491-515.
- Raskutti, Garvesh, and Caroline Uhler. "Learning directed acyclic graph models based on sparsest permutations." Stat 7.1 (2018): e183.
- Solus, Liam, Yuhao Wang, and Caroline Uhler. "Consistency guarantees for greedy permutation-based causal inference algorithms." Biometrika 108.4 (2021): 795-814.
- Lam, Wai-Yin, Bryan Andrews, and Joseph Ramsey. "Greedy relaxations of the sparsest permutation algorithm." Uncertainty in Artificial Intelligence. PMLR, 2022.
- Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., & Jordan, M. (2006). A linear non-Gaussian acyclic model for causal discovery. Journal of Machine Learning Research, 7(10).
- Glymour, Clark, Kun Zhang, and Peter Spirtes. "Review of causal discovery methods based on graphical models." Frontiers in genetics 10 (2019): 524.
- Tashiro, T., Shimizu, S., Hyvärinen, A., & Washio, T. (2014). ParceLiNGAM: A causal ordering method robust against latent confounders. Neural computation, 26(1), 57-83.
- Xie, F., Cai, R., Huang, B., Glymour, C., Hao, Z., & Zhang, K. (2020). Generalized independent noise condition for estimating latent variable causal graphs. Advances in neural information processing systems, 33, 14891-14902.
- Besserve, M., Shajarisales, N., Schölkopf, B., & Janzing, D. (2018, March). Group invariance principles for causal generative models. In International Conference on Artificial Intelligence and Statistics (pp. 557-565). PMLR.
- Peters, Jonas, Dominik Janzing, and Bernhard Schölkopf. Elements of causal inference: foundations and learning algorithms. The MIT Press, 2017.
- Janzing, Dominik, and Bernhard Schölkopf. "Causal inference using the algorithmic Markov condition." IEEE Transactions on Information Theory 56.10 (2010): 5168-5194.
- Gresele, L., Von Kügelgen, J., Stimper, V., Schölkopf, B., & Besserve, M. (2021). Independent mechanism analysis, a new concept?. Advances in neural information processing systems, 34, 28233-28248.
- Zheng, X., Aragam, B., Ravikumar, P. K., & Xing, E. P. (2018). Dags with no tears: Continuous optimization for structure learning. Advances in neural information processing systems, 31.
- Vowels, Matthew J., Necati Cihan Camgoz, and Richard Bowden. "D'ya like dags? a survey on structure learning and causal discovery." ACM Computing Surveys 55.4 (2022): 1-36
- Wolfe, Elie, Robert W. Spekkens, and Tobias Fritz. "The inflation technique for causal inference with latent variables." Journal of Causal Inference 7.2 (2019): 20170020.
- Navascués, Miguel, and Elie Wolfe. "The inflation technique completely solves the causal compatibility problem." Journal of Causal Inference 8.1 (2020): 70-91.

Other resources:
- Simons Institute Causality program bootcamp: https://simons.berkeley.edu/workshops/causality-boot-camp/videos#simons-tabs (note especially the causal discovery tutorials by Daniel Malinsky)

Thank you!