Causal Inference via Quantifying Influences

Dan Waxman
Stony Brook University

Stony Brook University

September 13th, 2023
ARI @ OeAW

Causal Inference via Quantifying Influences

Causal Inference via Quantifying Influences

Quantifying Causal Strength

Causal Inference via Quantifying Influences

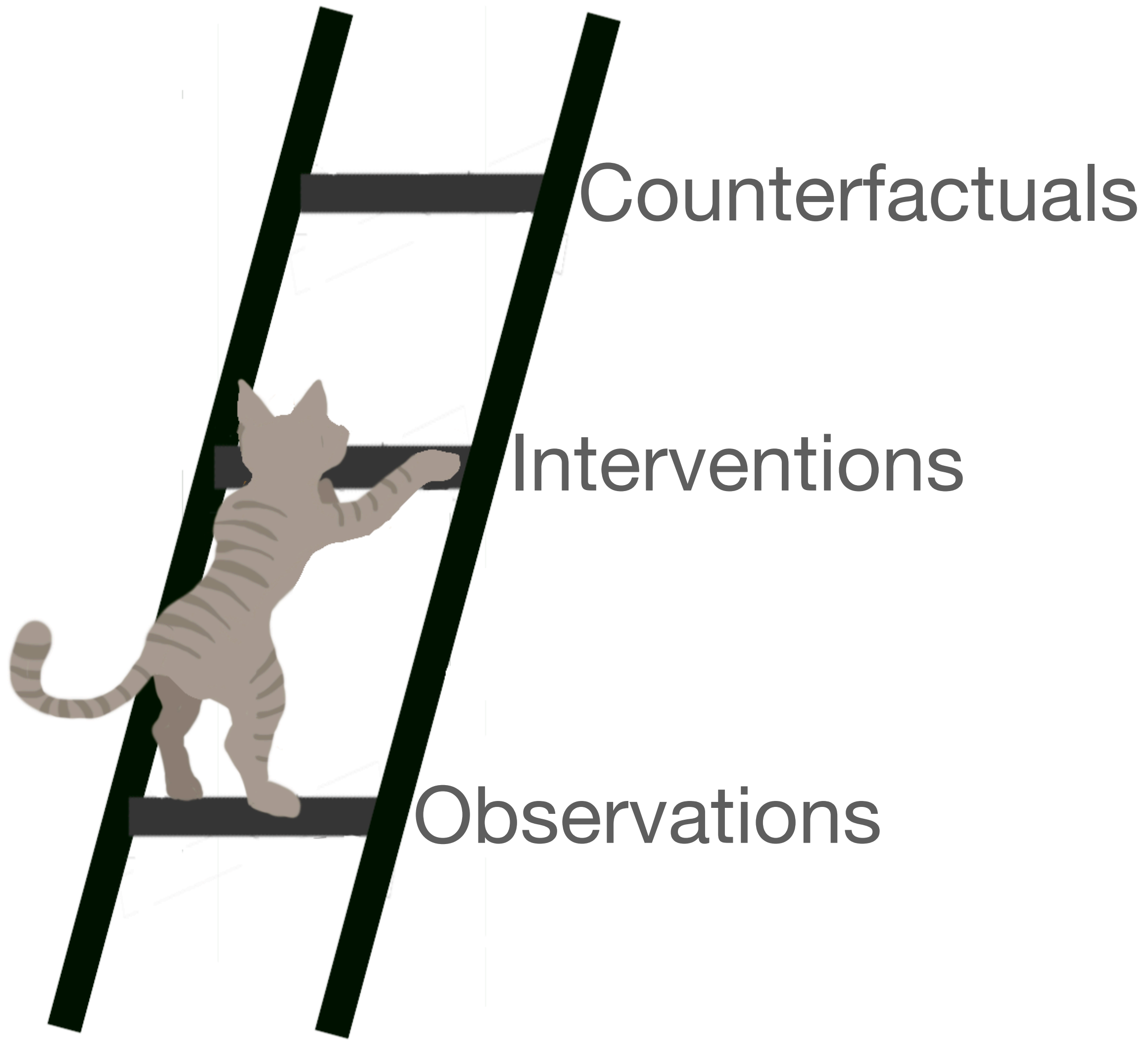
Confounder Detection

Causal Discovery

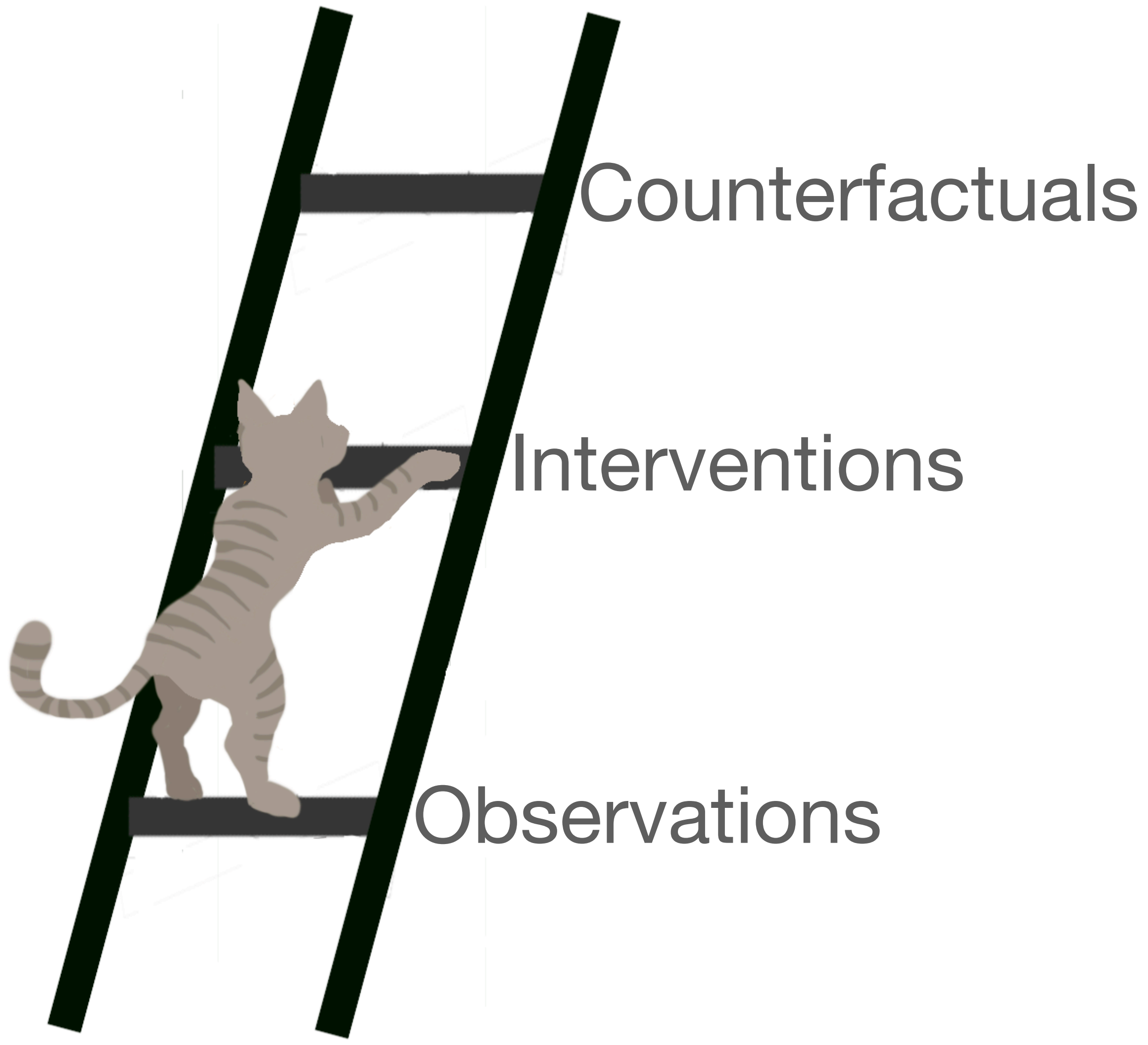
Quantifying Causal Strength

A Crash Course in Causality

Defining Cause and Effect

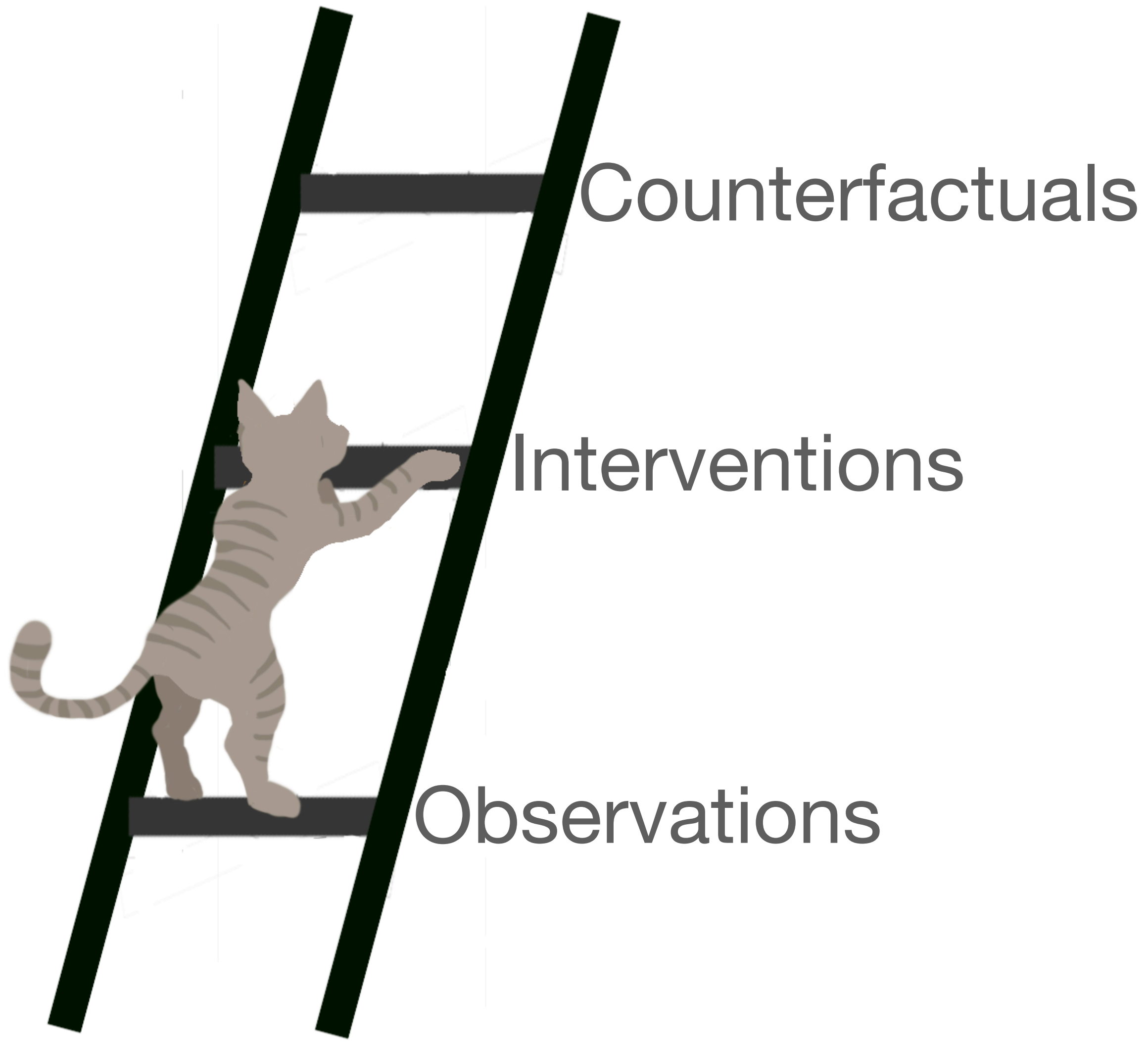


Defining Cause and Effect



Defining causality is difficult. Here are some attempts:

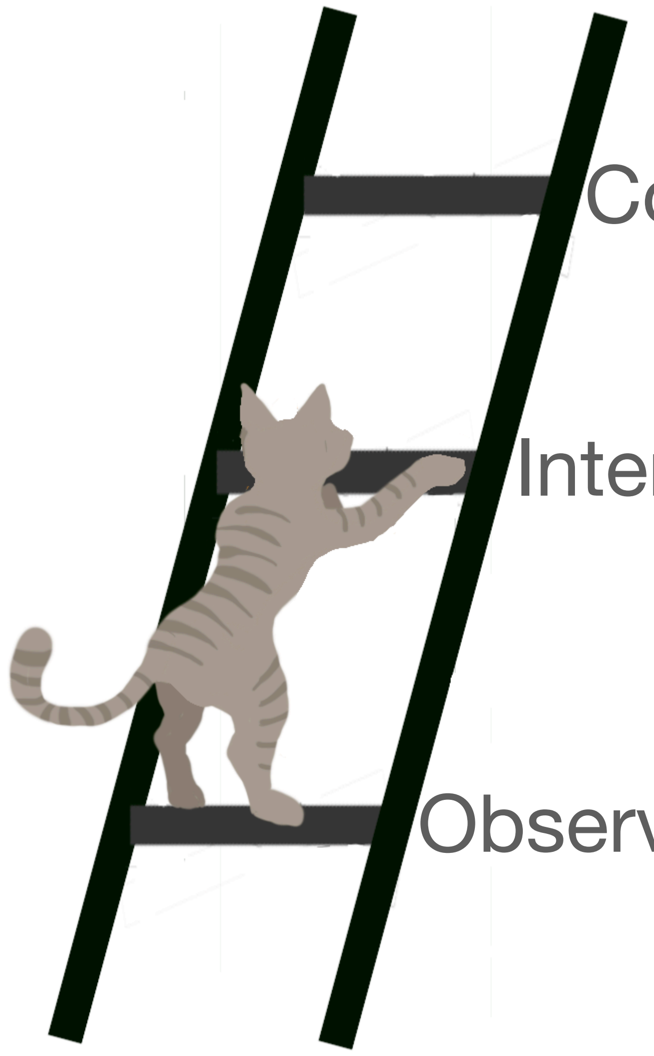
Defining Cause and Effect



Defining causality is difficult. Here are some attempts:

X causes Y if:

Defining Cause and Effect

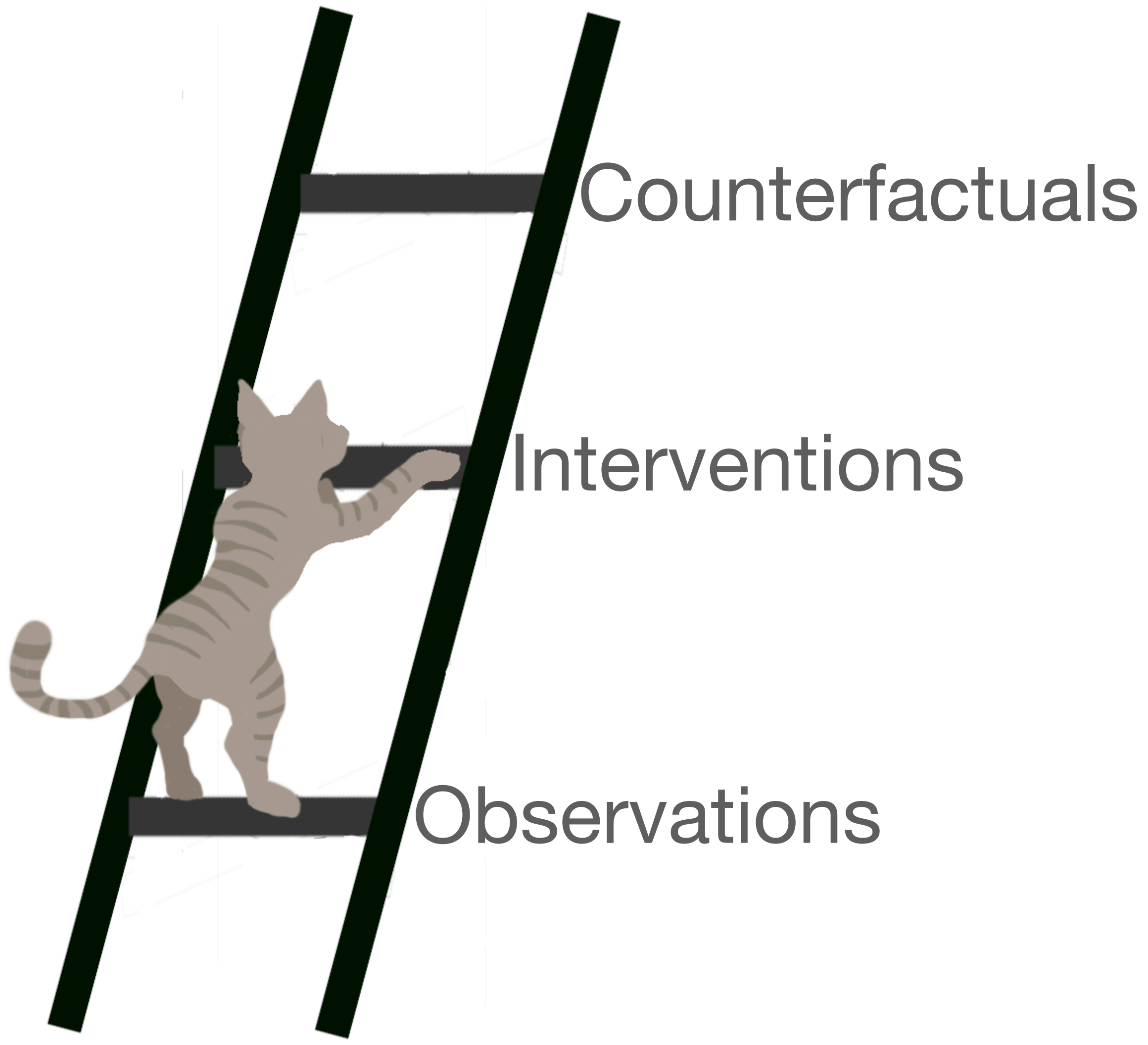


Defining causality is difficult. Here are some attempts:

X causes Y if:

- $\text{Var}[Y_t \mid X_t, Y_{1:t-L}] < \text{Var}[Y_t \mid Y_{1:t-L}]$ [Granger 1969, Wiener 1956]

Defining Cause and Effect

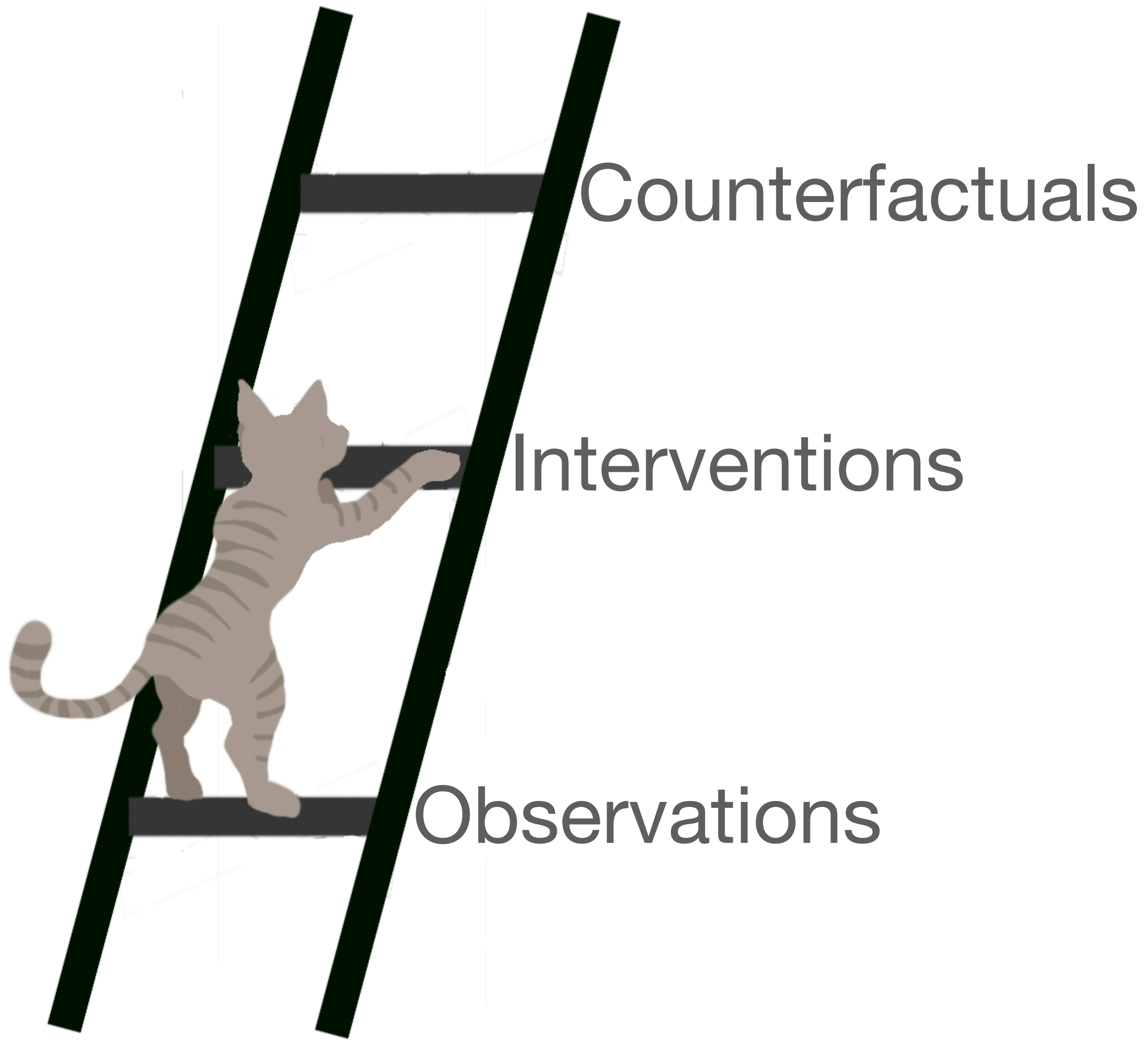


Defining causality is difficult. Here are some attempts:

X causes Y if:

- $\text{Var}[Y_t \mid X_t, Y_{1:t-L}] < \text{Var}[Y_t \mid Y_{1:t-L}]$ [Granger 1969, Wiener 1956]
- $I[Y_t; X_t \mid Y_{1:t-L}] > 0$ [Schrieber 2000]

Defining Cause and Effect

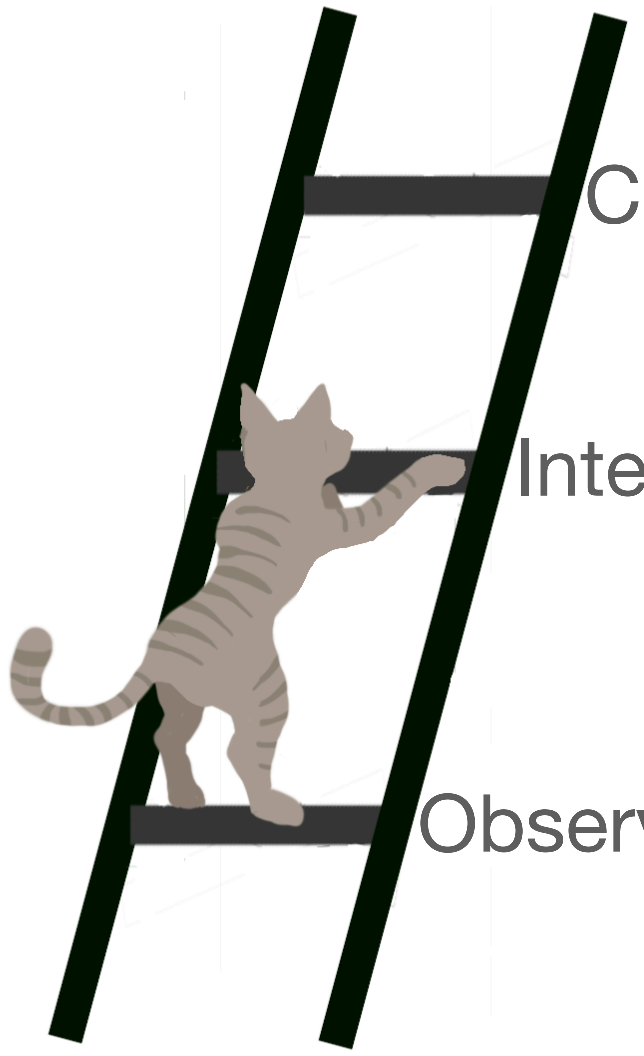


Defining causality is difficult. Here are some attempts:

X causes Y if:

- Var$[Y_t \mid X_t, Y_{1:t-L}] <$ Var$[Y_t \mid Y_{1:t-L}]$ [Granger 1969, Wiener 1956]
- $I[Y_t; X_t \mid Y_{1:t-L}] > 0$ [Schrieber 2000]
- If we intervene on X, then Y changes [Pearl 2009, among others]

Defining Cause and Effect

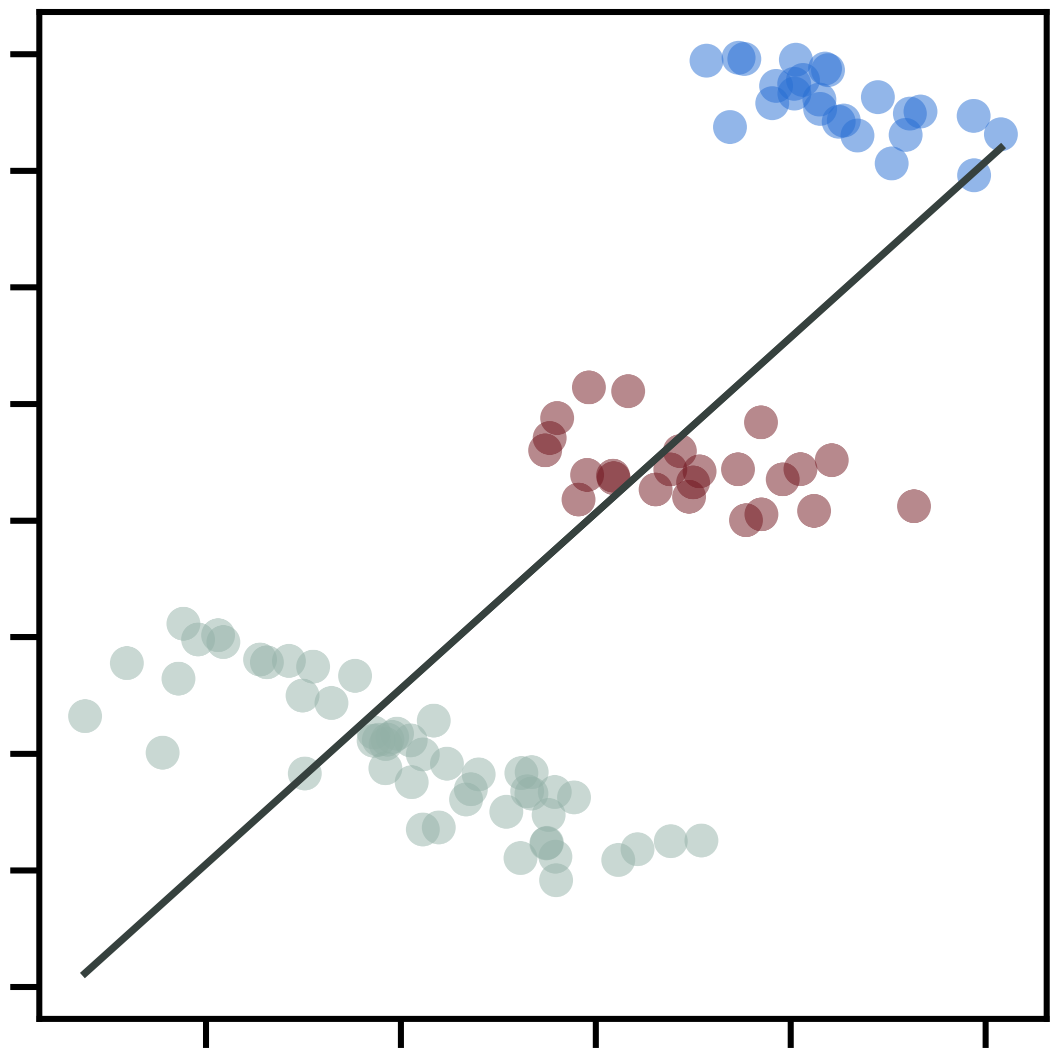


Defining causality is difficult. Here are some attempts:

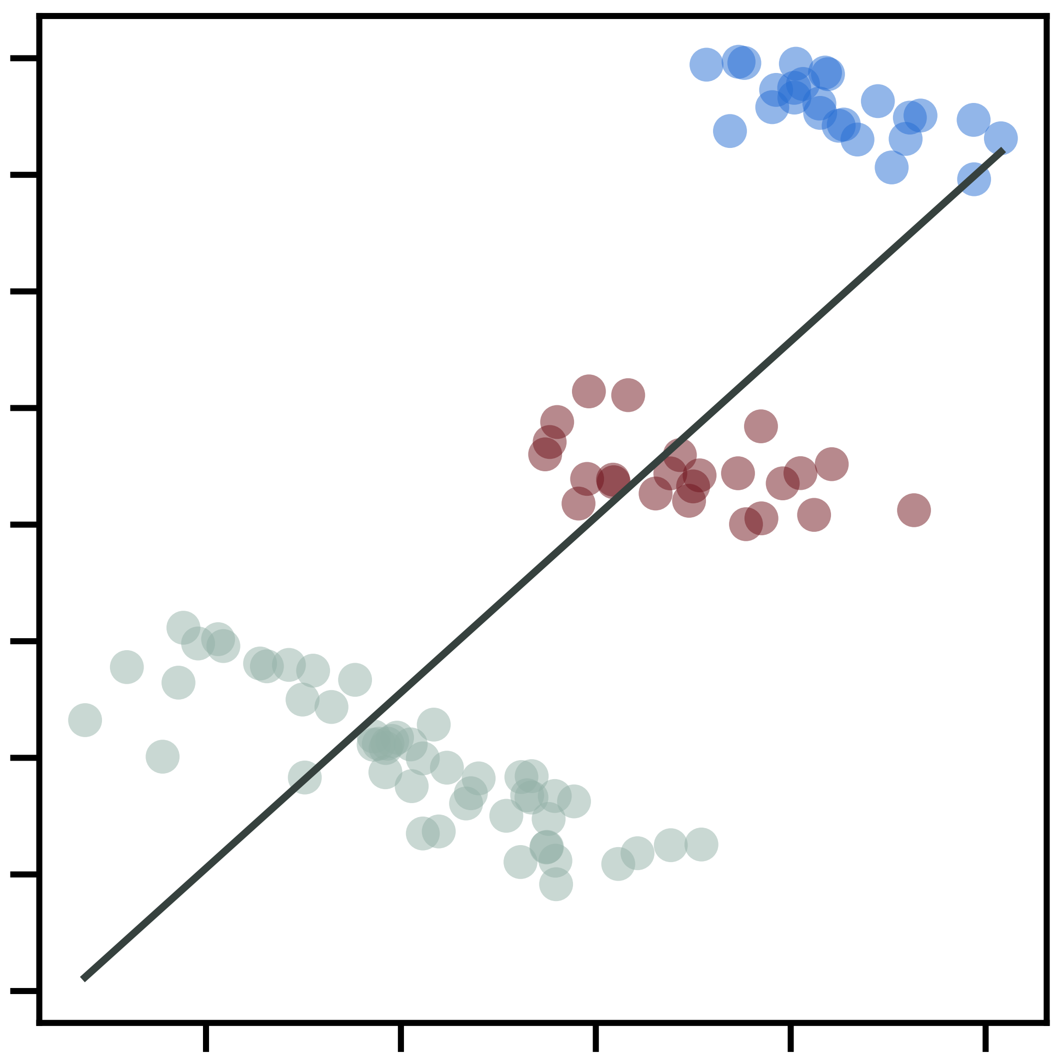
X causes Y if:

- $\mathrm{Var}[Y_t \mid X_t, Y_{1:t-L}] < \mathrm{Var}[Y_t \mid Y_{1:t-L}]$ [Granger 1969, Wiener 1956]
- $I[Y_t; X_t \mid Y_{1:t-L}] > 0$ [Schrieber 2000]
- If we intervene on X, then Y changes [Pearl 2009, among others]

Interventions are a Necessary Concept

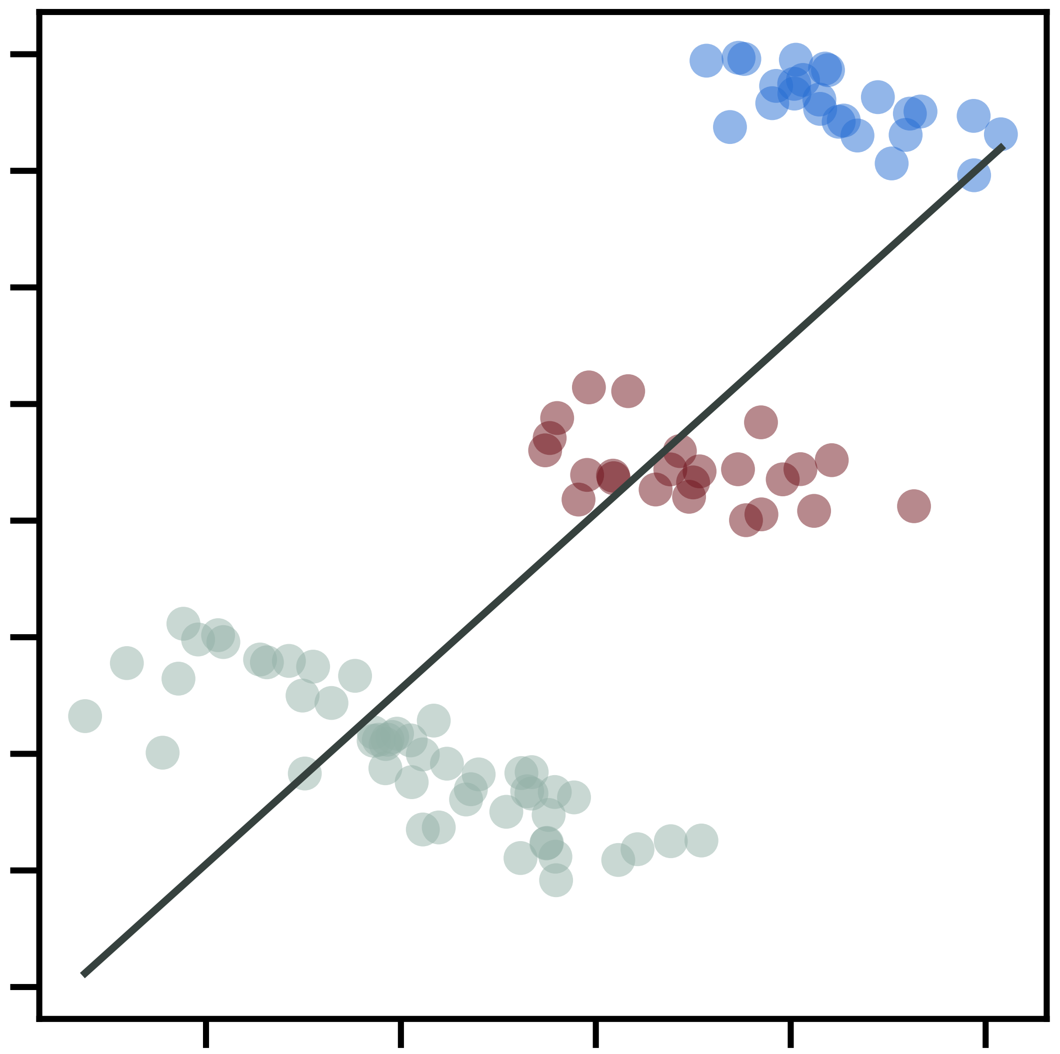


Interventions are a Necessary Concept



Intervention is stronger than any purely statistical concept

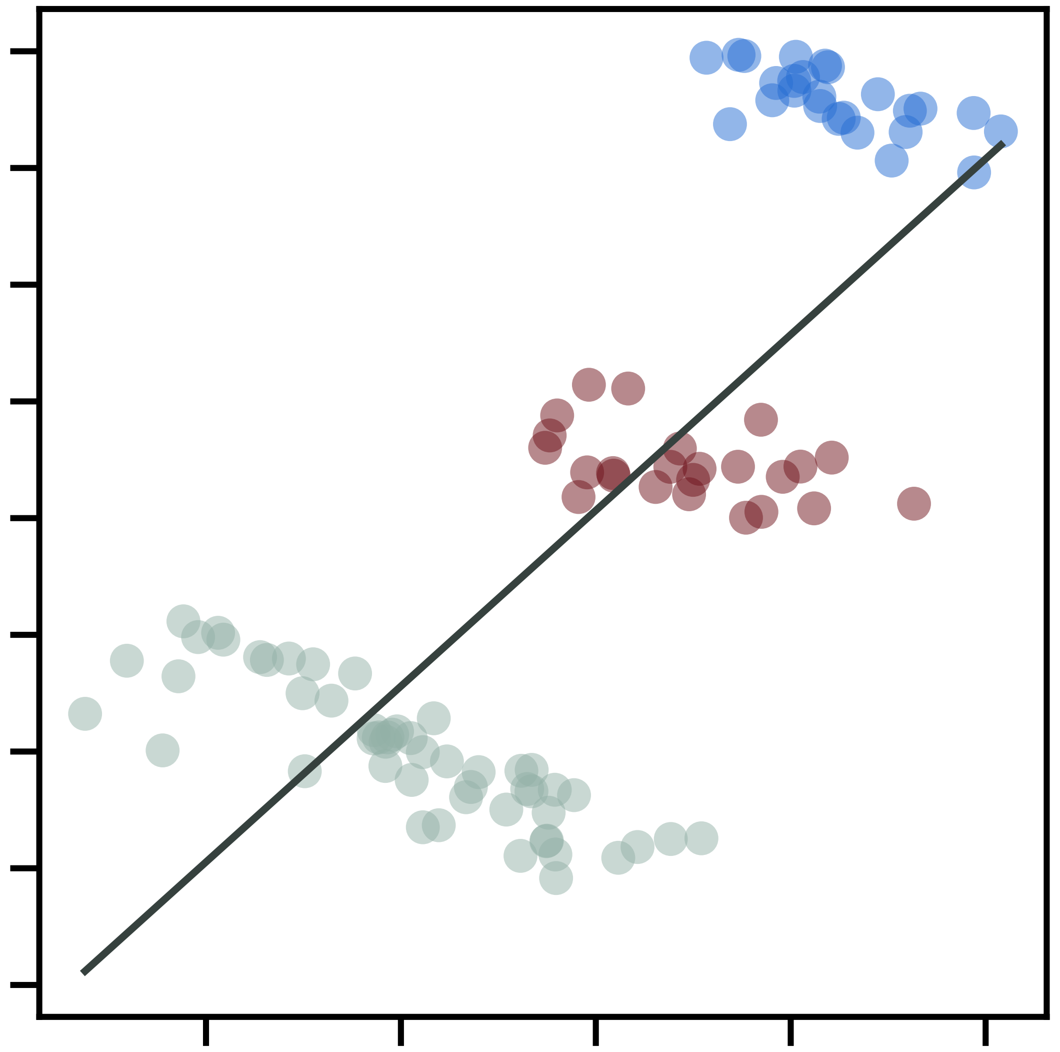
Interventions are a Necessary Concept



Intervention is stronger than any purely statistical concept

The notation for an intervention is $\mathrm{do}(X = x)$

Interventions are a Necessary Concept

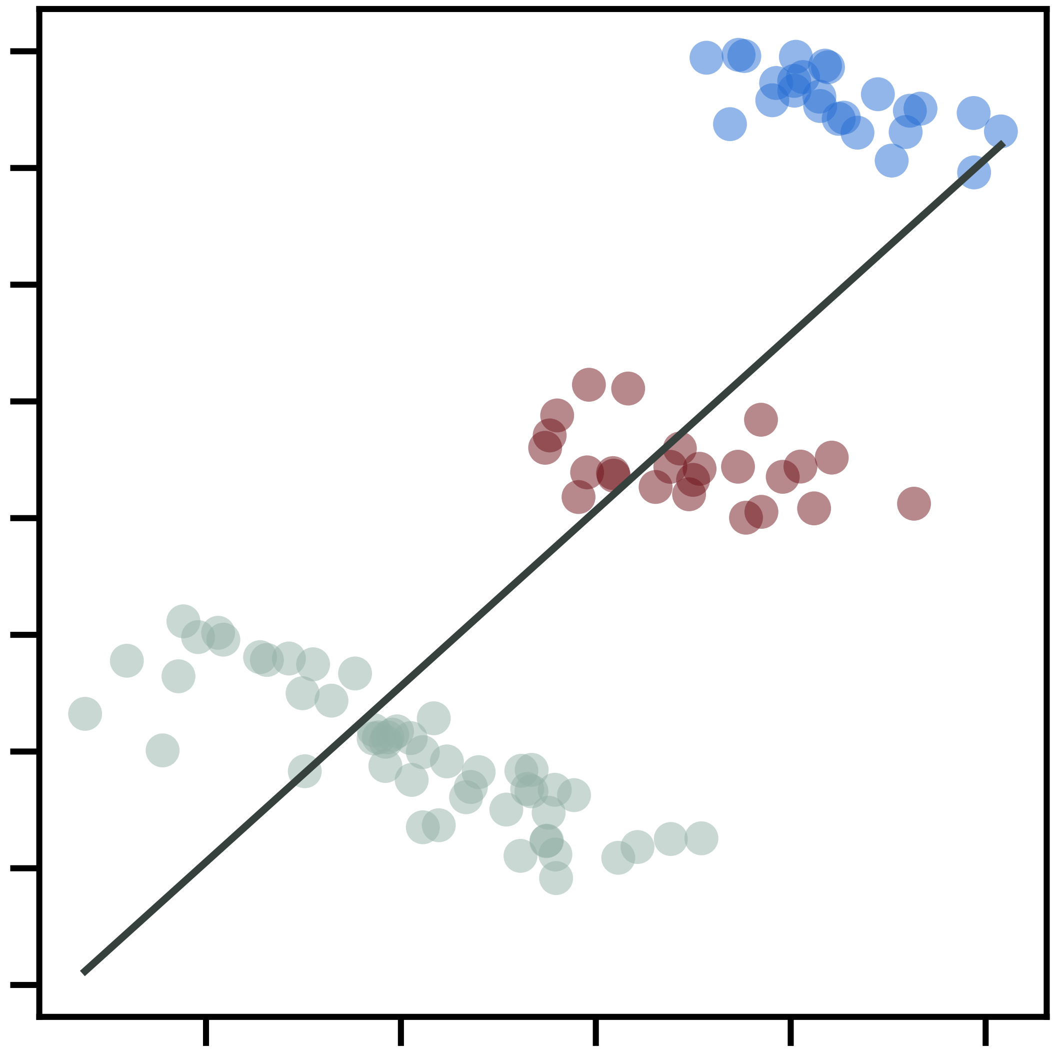


Intervention is stronger than any purely statistical concept

The notation for an intervention is $\mathrm{do}(X = x)$

Consider Simpson's paradox

Interventions are a Necessary Concept



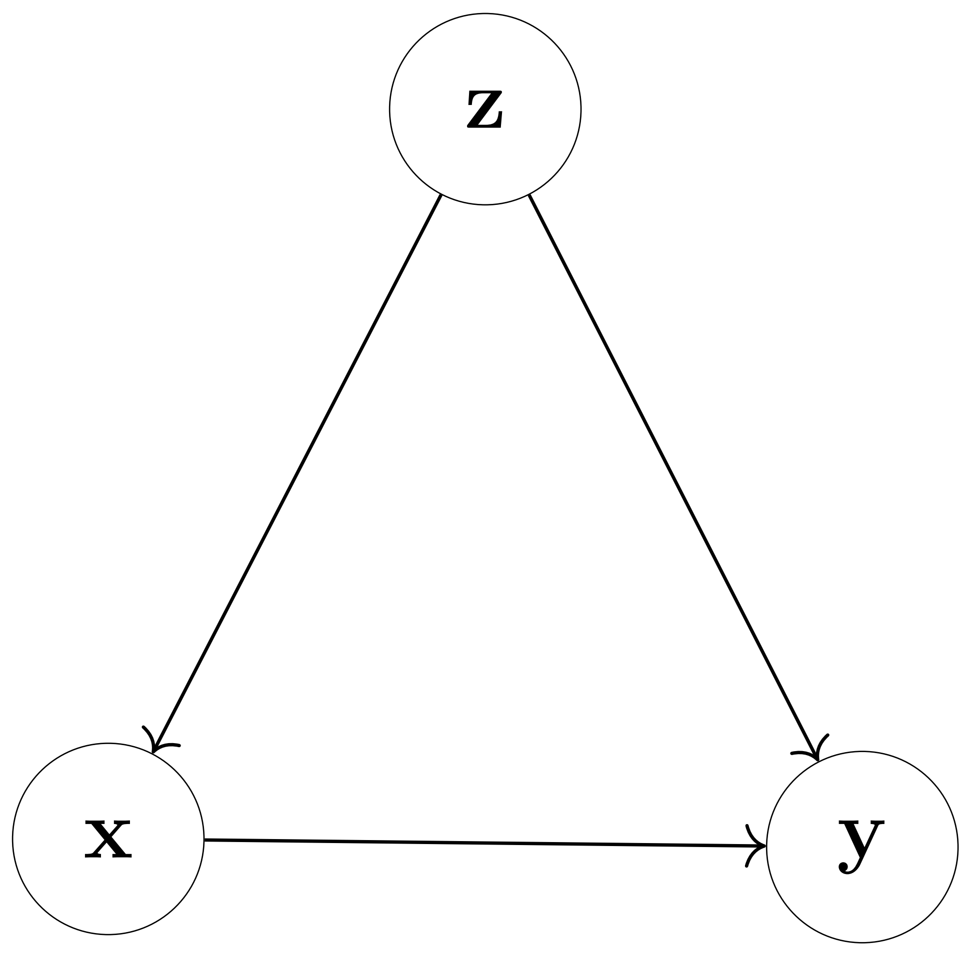
Intervention is stronger than any purely statistical concept

The notation for an intervention is $\mathrm{do}(X = x)$

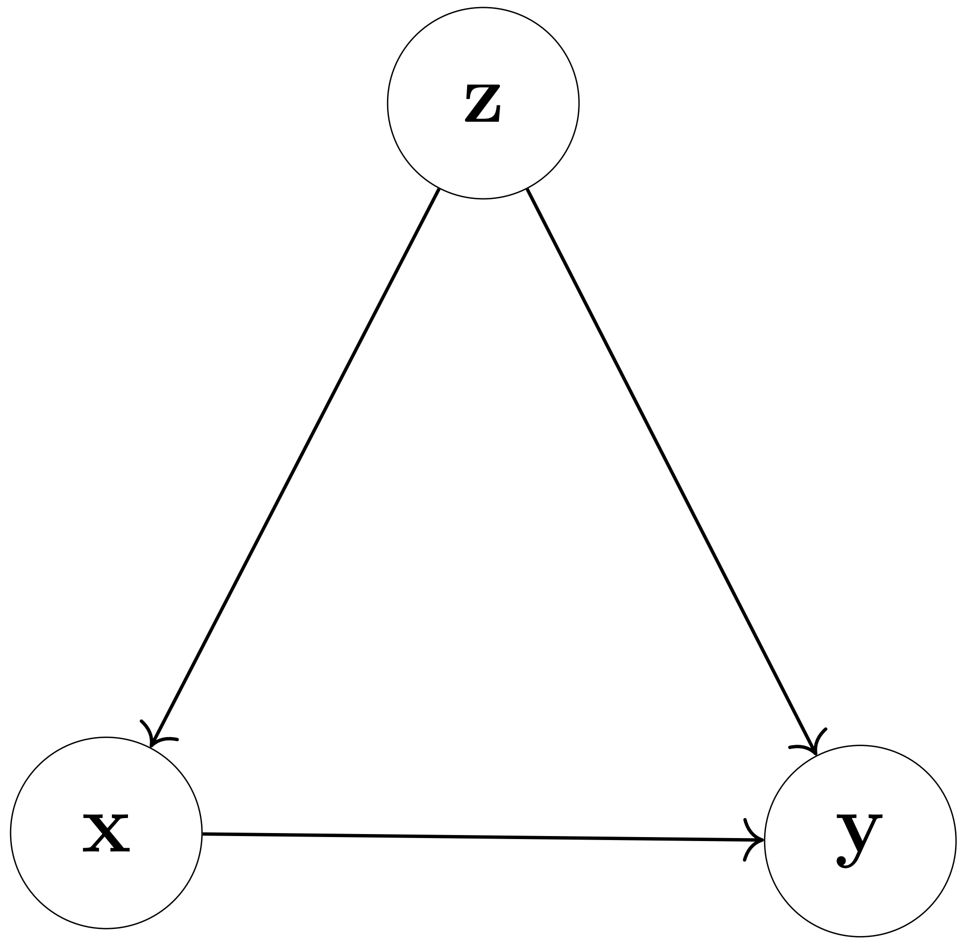
Consider Simpson's paradox

Statistically, X is positively correlated with Y; but the intervention $\mathrm{do(X = x)}$ decreases Y

Modeling Cause and Effect

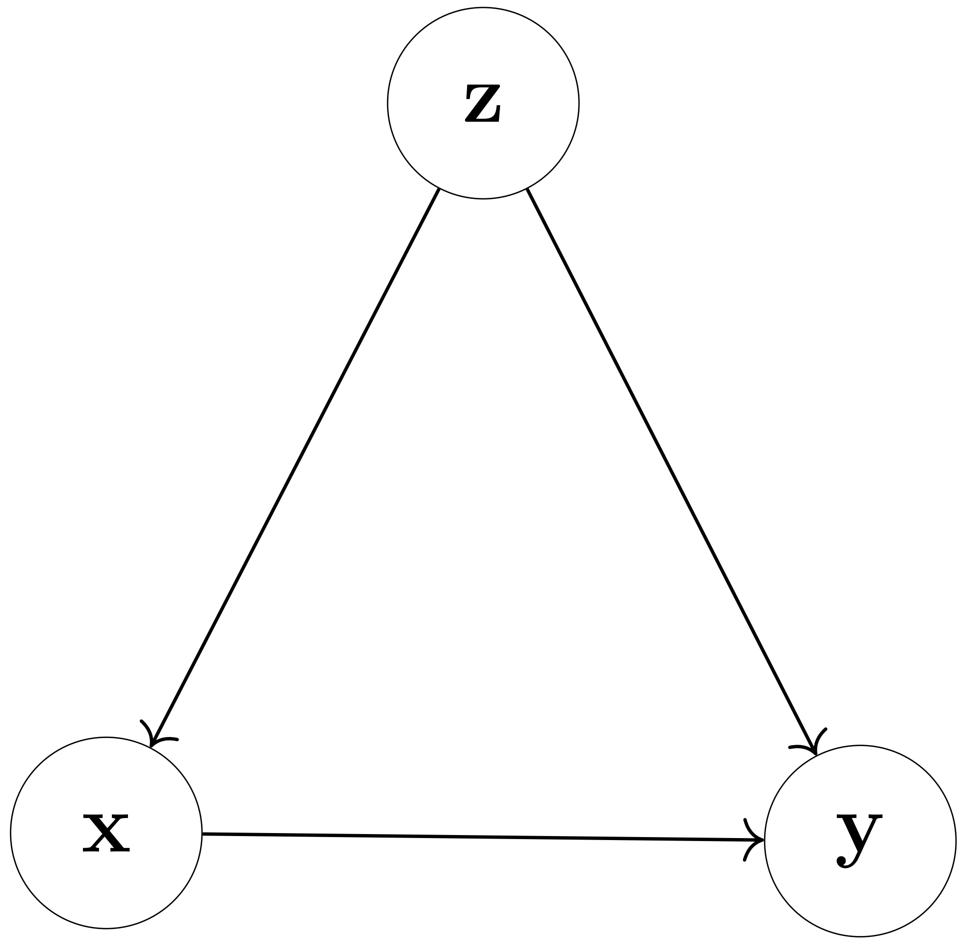


Modeling Cause and Effect



To model Pearl's causality, we use *Bayesian networks*

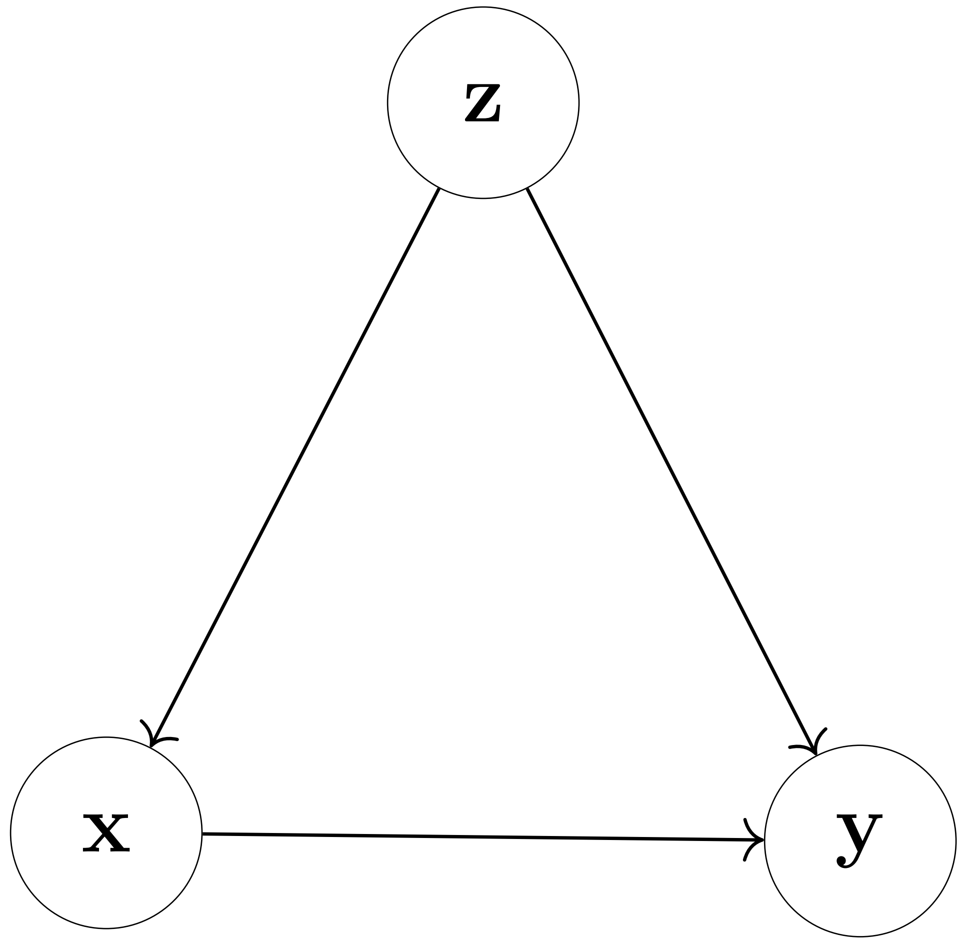
Modeling Cause and Effect



To model Pearl’s causality, we use *Bayesian networks*

These imply joint probabilities + data-generating factorization

Modeling Cause and Effect

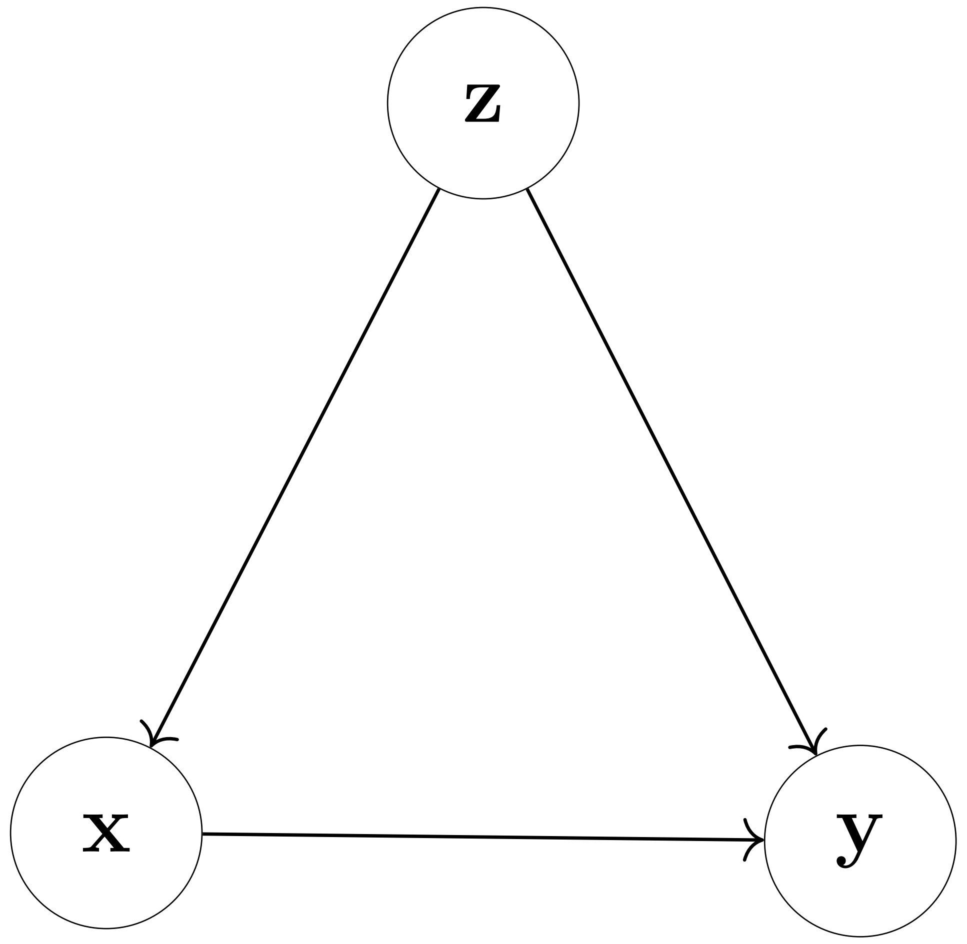


To model Pearl's causality, we use *Bayesian networks*

These imply joint probabilities + data-generating factorization

$$p(x, y, z) = p(y \mid x, z) \ \times \ p(x \mid z) \ \times \ p(z)$$

Modeling Cause and Effect



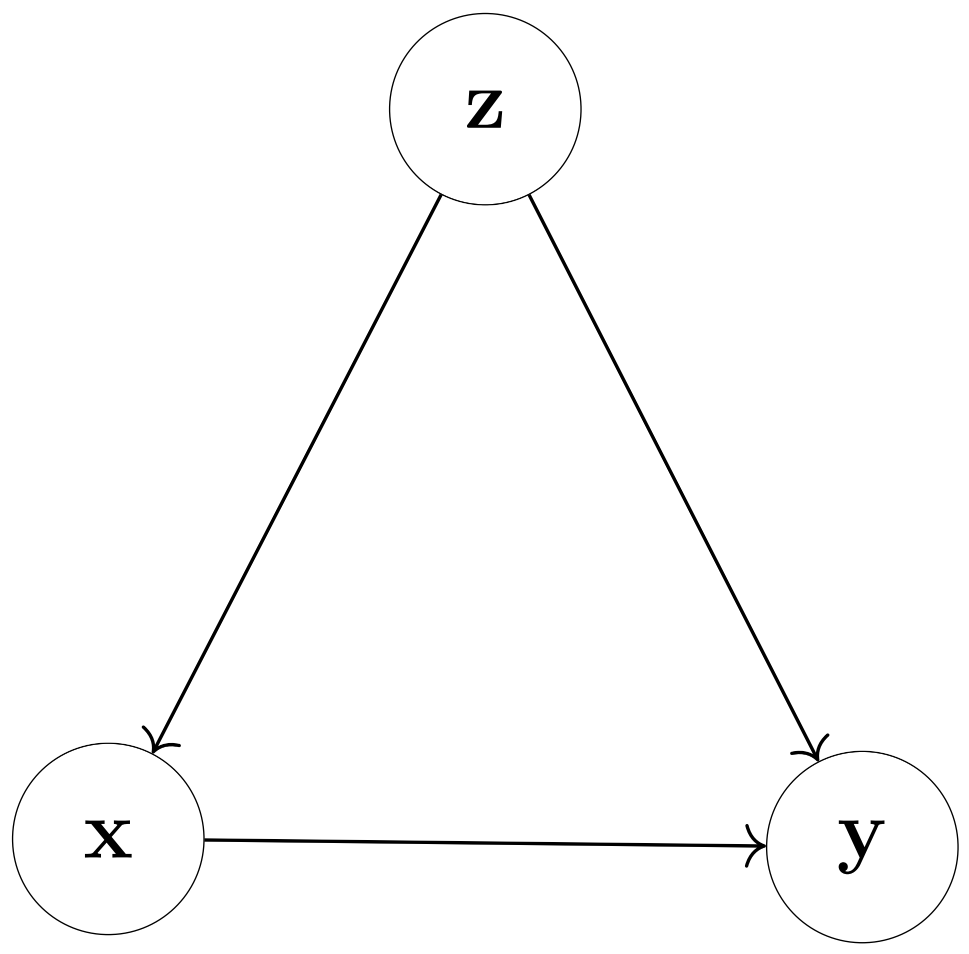
To model Pearl's causality, we use *Bayesian networks*

These imply joint probabilities + data-generating factorization

$$p(x, y, z) = p(y \mid x, z) \ \times \ p(x \mid z) \ \times \ p(z)$$

"Y is caused by X and Z"

Modeling Cause and Effect

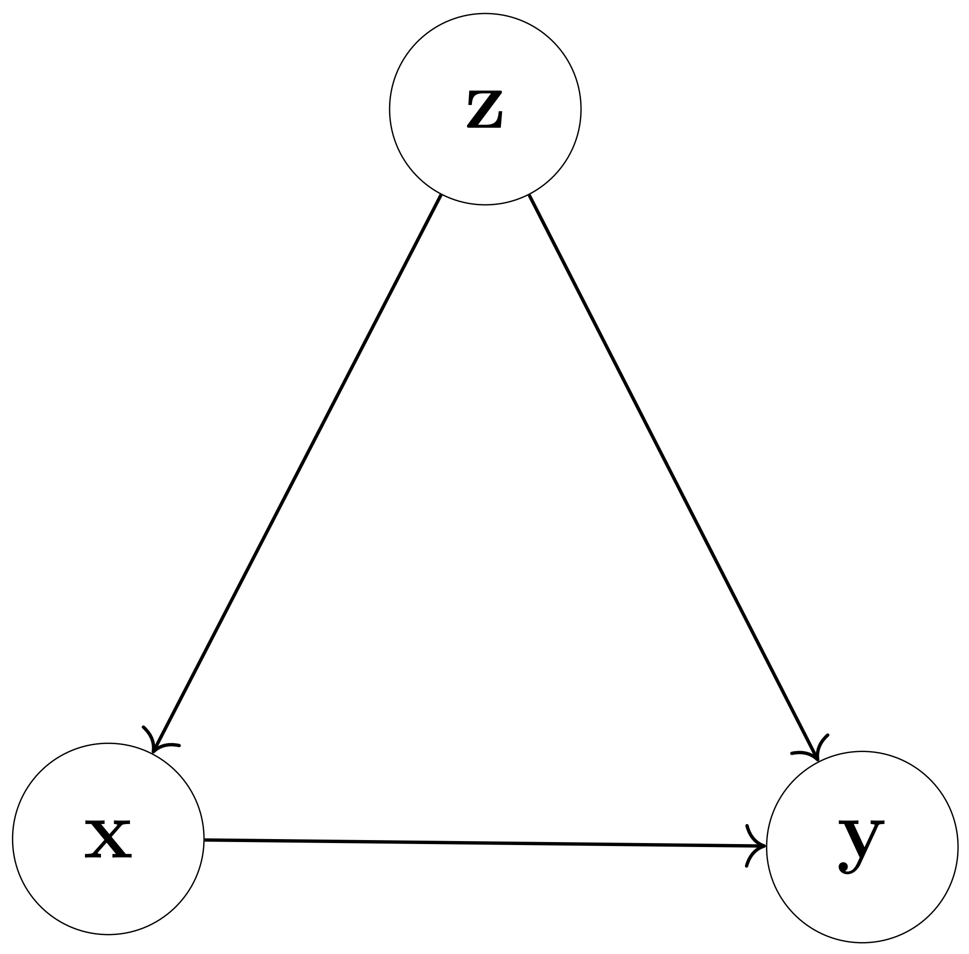


To model Pearl's causality, we use *Bayesian networks*

These imply joint probabilities + data-generating factorization

$$p(x, y, z) = \underbrace{p(y \mid x, z)}_{\text{“}Y\text{ is caused by }X\text{ and }Z\text{”}} \times \underbrace{p(x \mid z)}_{\text{“}X\text{ is caused by }Z\text{”}} \times p(z)$$

Modeling Cause and Effect



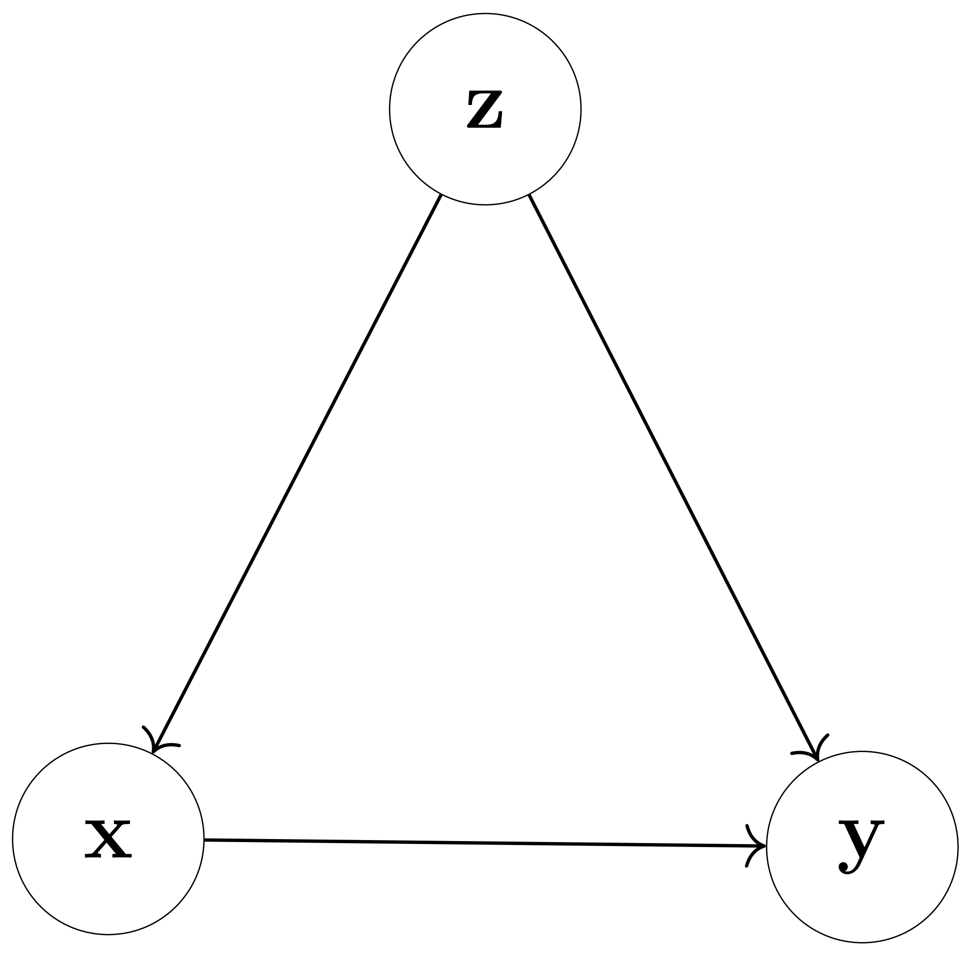
To model Pearl's causality, we use *Bayesian networks*

These imply joint probabilities + data-generating factorization

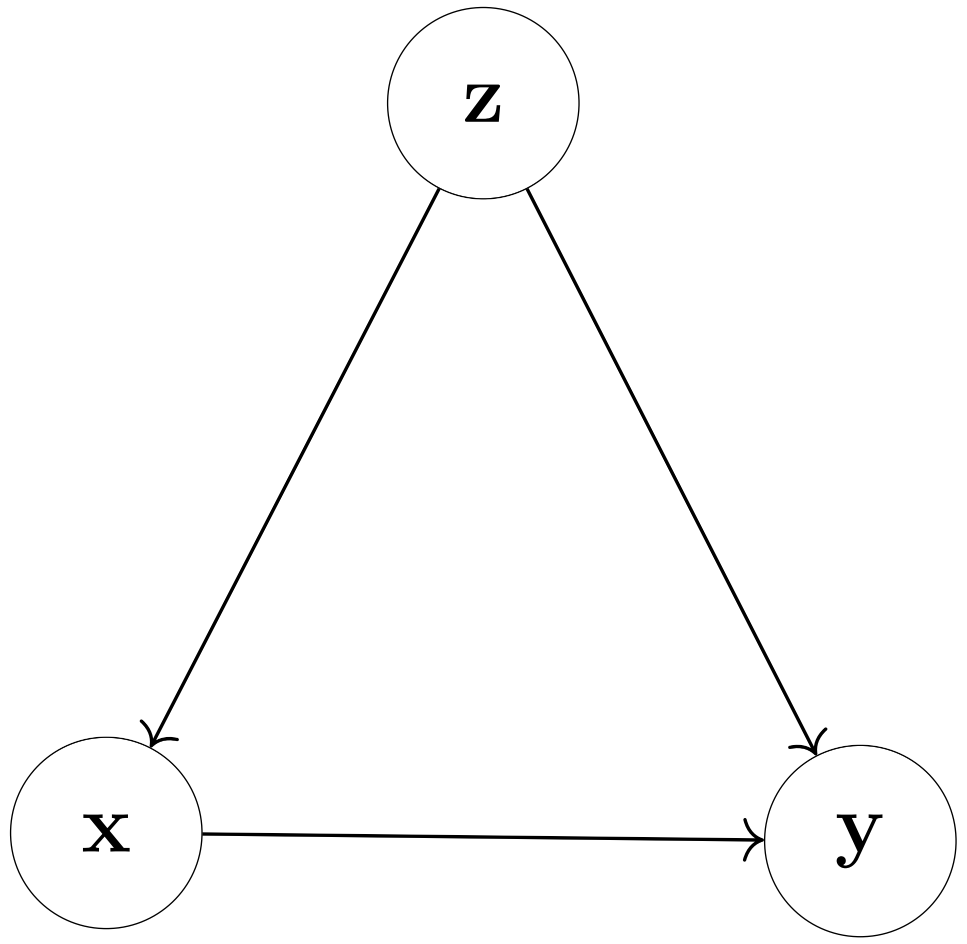
$$p(x, y, z) = p(y \,|\, x, z) \;\times\; p(x \,|\, z) \;\times\; p(z)$$

"Y is caused by X and Z" "X is caused by Z" "Z is exogenous"

Modeling Cause and Effect

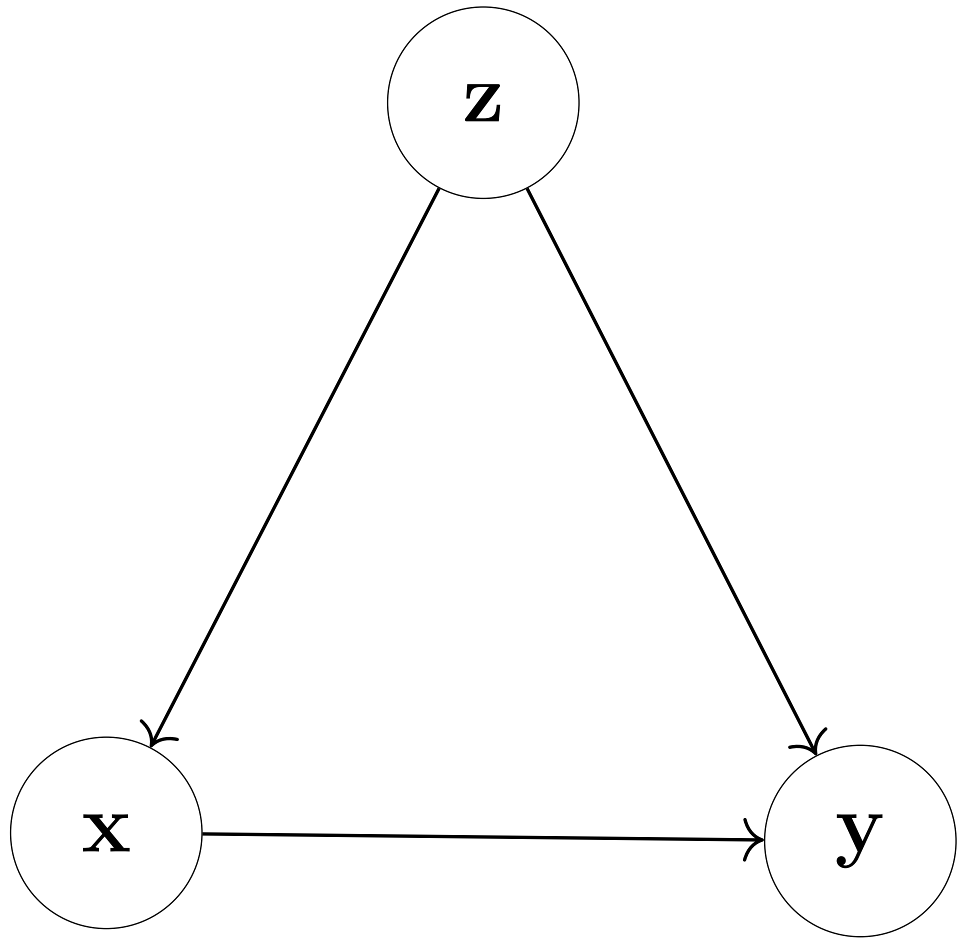


Modeling Cause and Effect



By reparameterizing, we can always rewrite the Bayesian network functionally

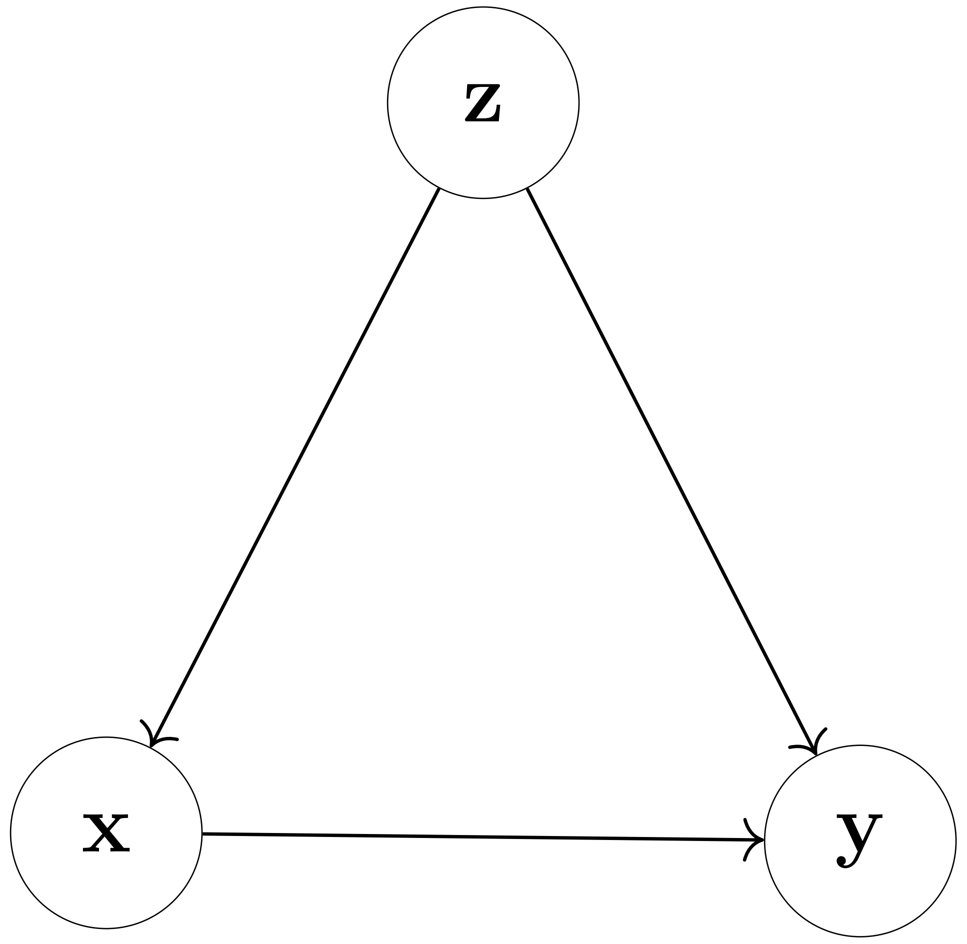
Modeling Cause and Effect



By reparameterizing, we can always rewrite the Bayesian network functionally

This is called the *structural causal model (SCM),* and consists of a *causal graph* + *structural equations*

Modeling Cause and Effect

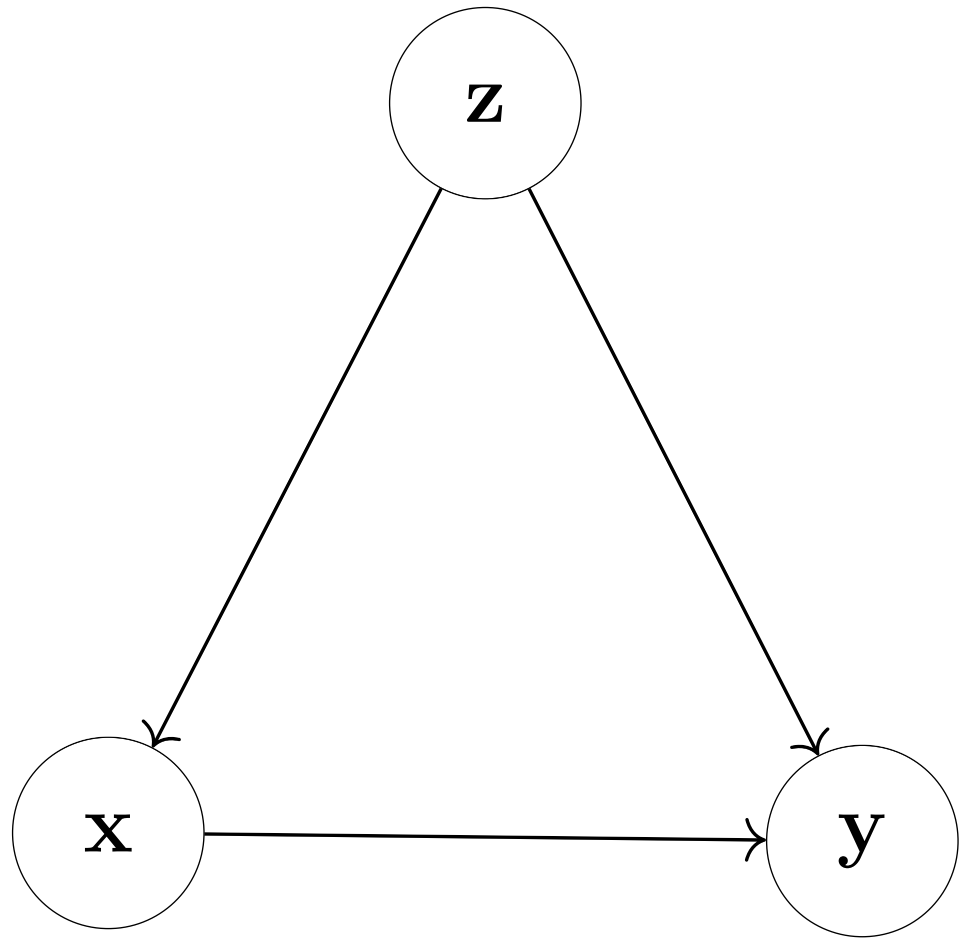


By reparameterizing, we can always rewrite the Bayesian network functionally

This is called the *structural causal model (SCM),* and consists of a *causal graph + structural equations*

$$X := f(\mathrm{Pa}(X), \varepsilon)$$

Modeling Cause and Effect



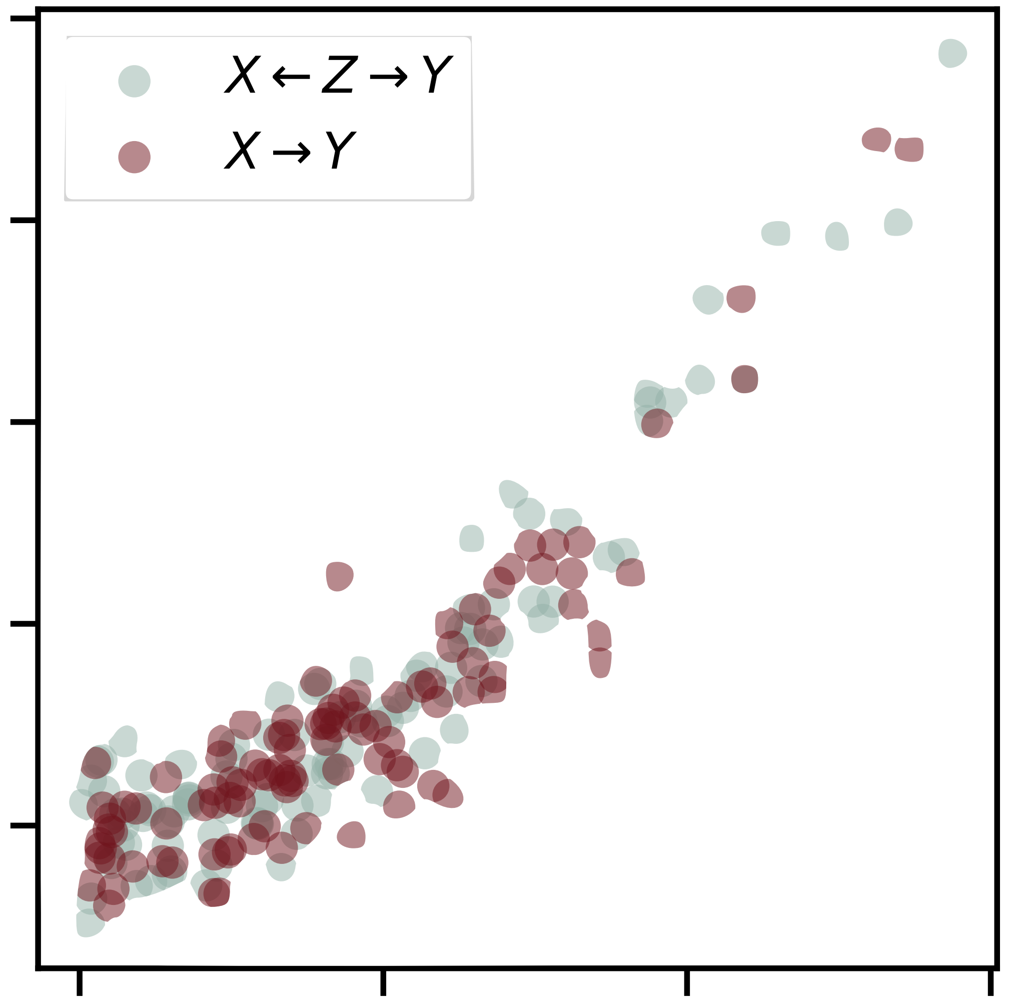
By reparameterizing, we can always rewrite the Bayesian network functionally

This is called the *structural causal model (SCM),* and consists of a *causal graph* + *structural equations*

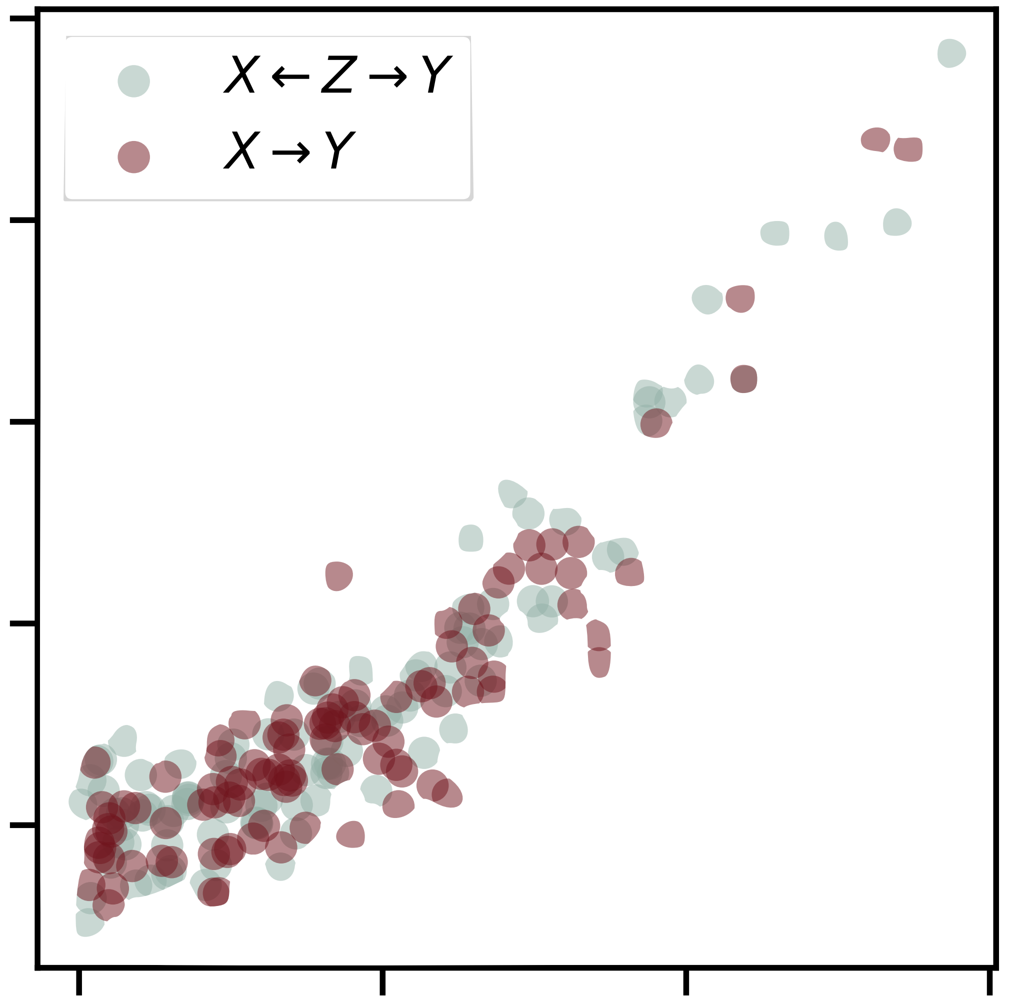
$$X := f(\mathrm{Pa}(X), \varepsilon)$$

"X is a function of its parents (causes)
and a noise term"

We Need Interventions or Assumptions To Discover Causality

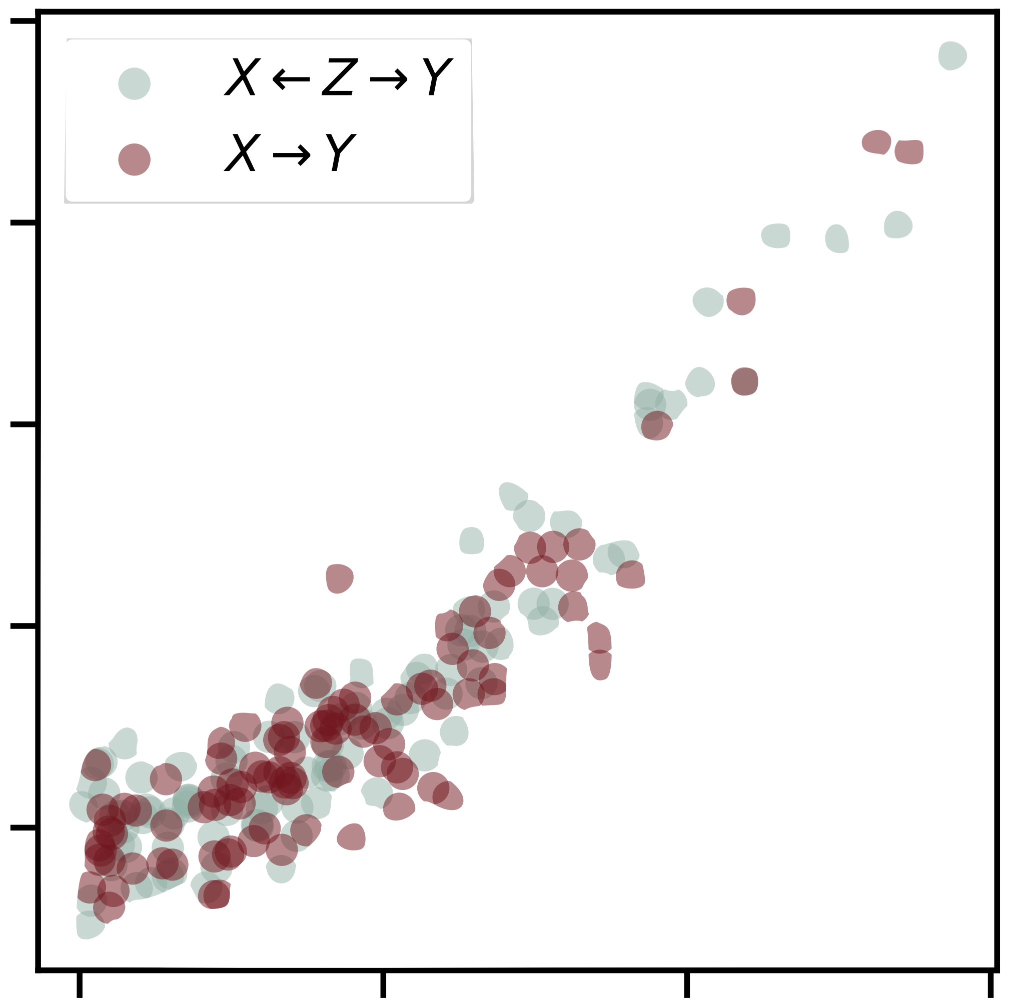


We Need Interventions or Assumptions To Discover Causality



Causal discovery is ill-posed for observational data

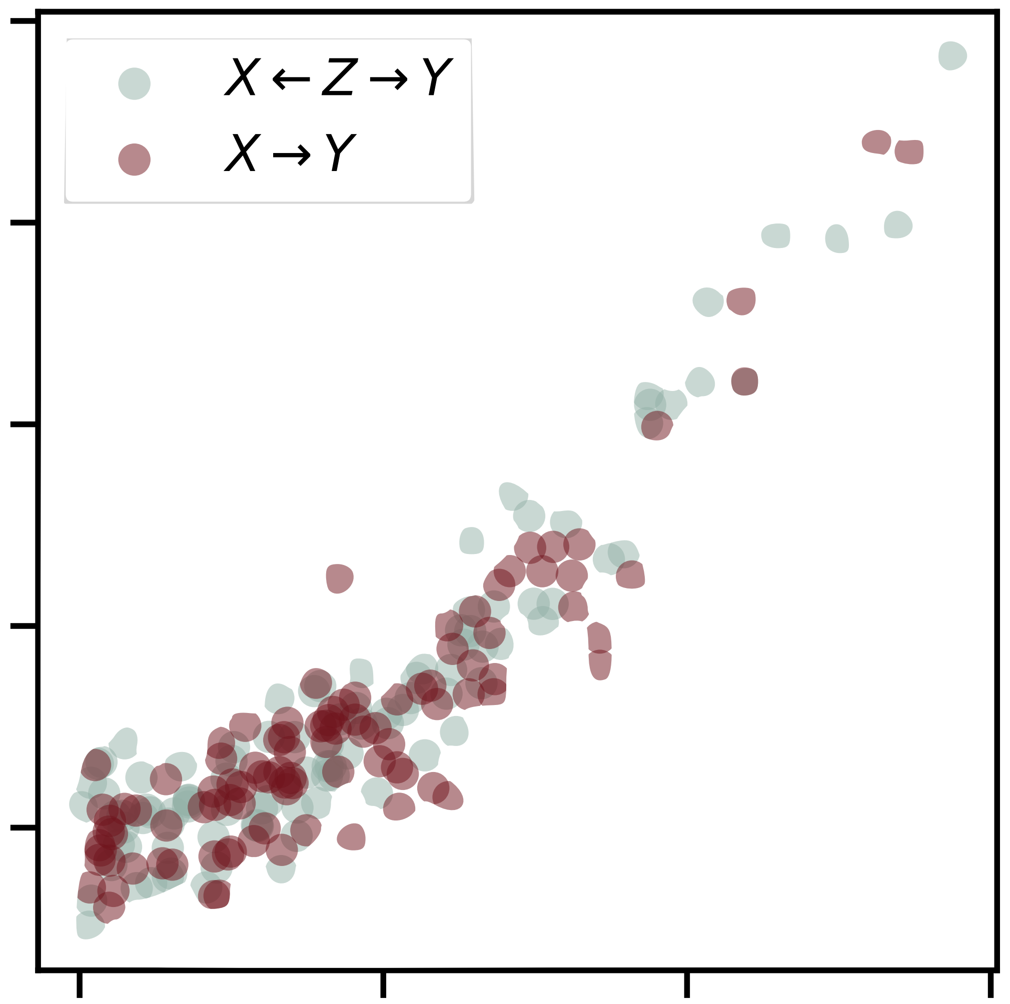
We Need Interventions or Assumptions To Discover Causality



Causal discovery is ill-posed for observational data

Non-parametrically, no reason to favor the “true” factorization

We Need Interventions or Assumptions To Discover Causality

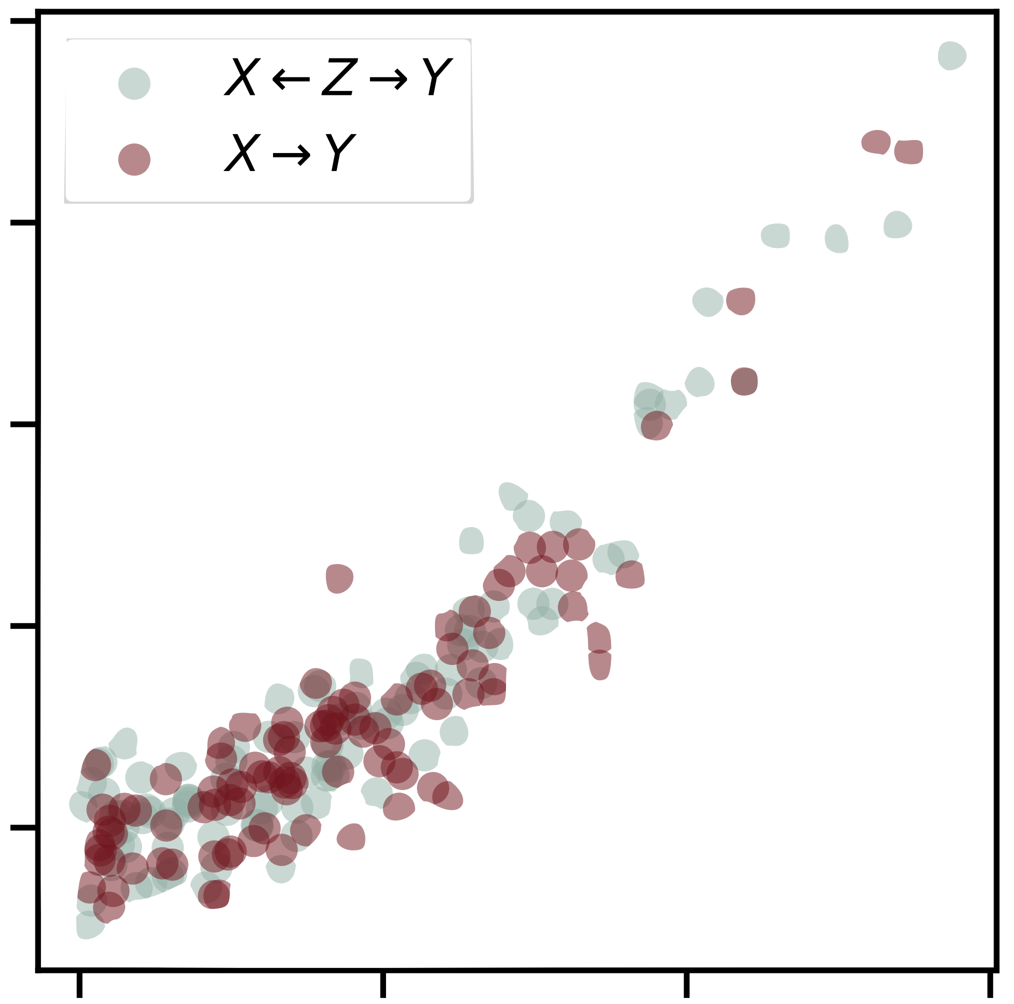


Causal discovery is ill-posed for observational data

Non-parametrically, no reason to favor the “true” factorization

Two (mutually-inclusive) options:

We Need Interventions or Assumptions To Discover Causality



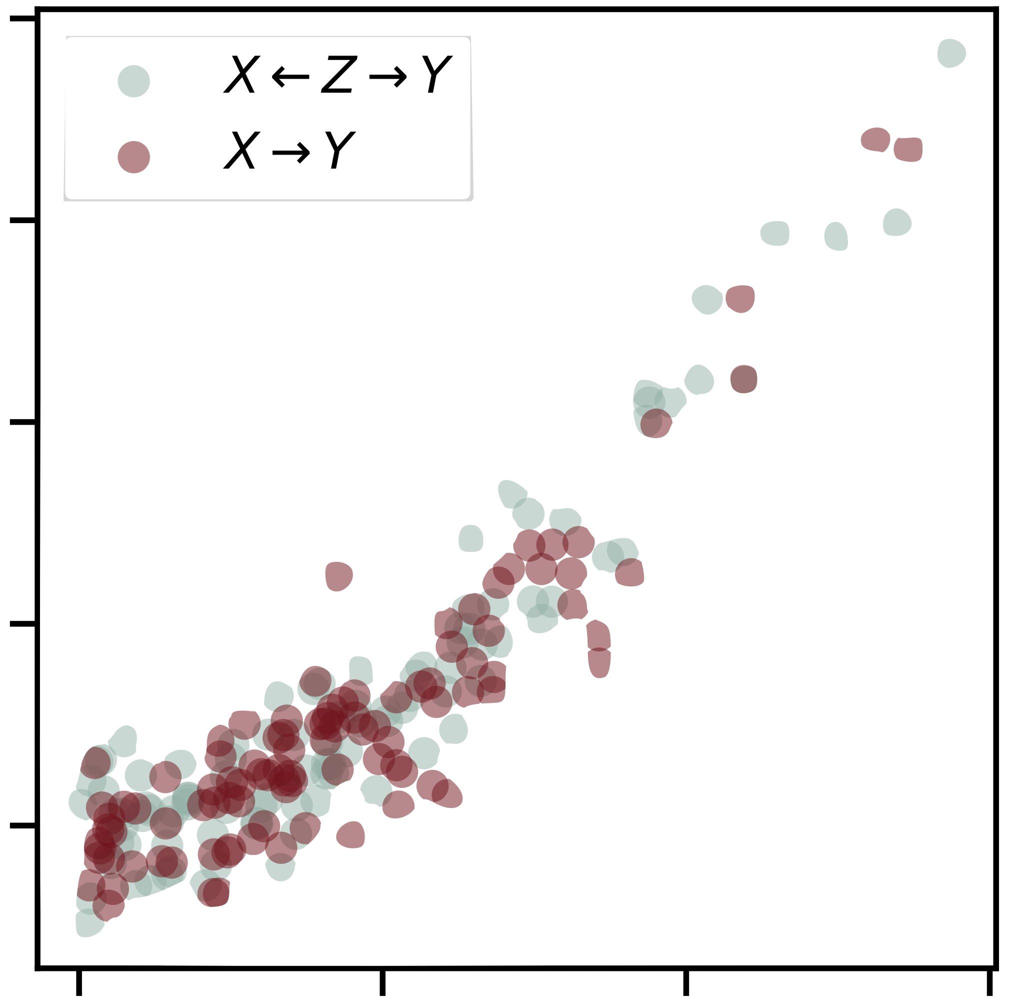
Causal discovery is ill-posed for observational data

Non-parametrically, no reason to favor the “true” factorization

Two (mutually-inclusive) options:

- Intervene on the system

We Need Interventions or Assumptions To Discover Causality



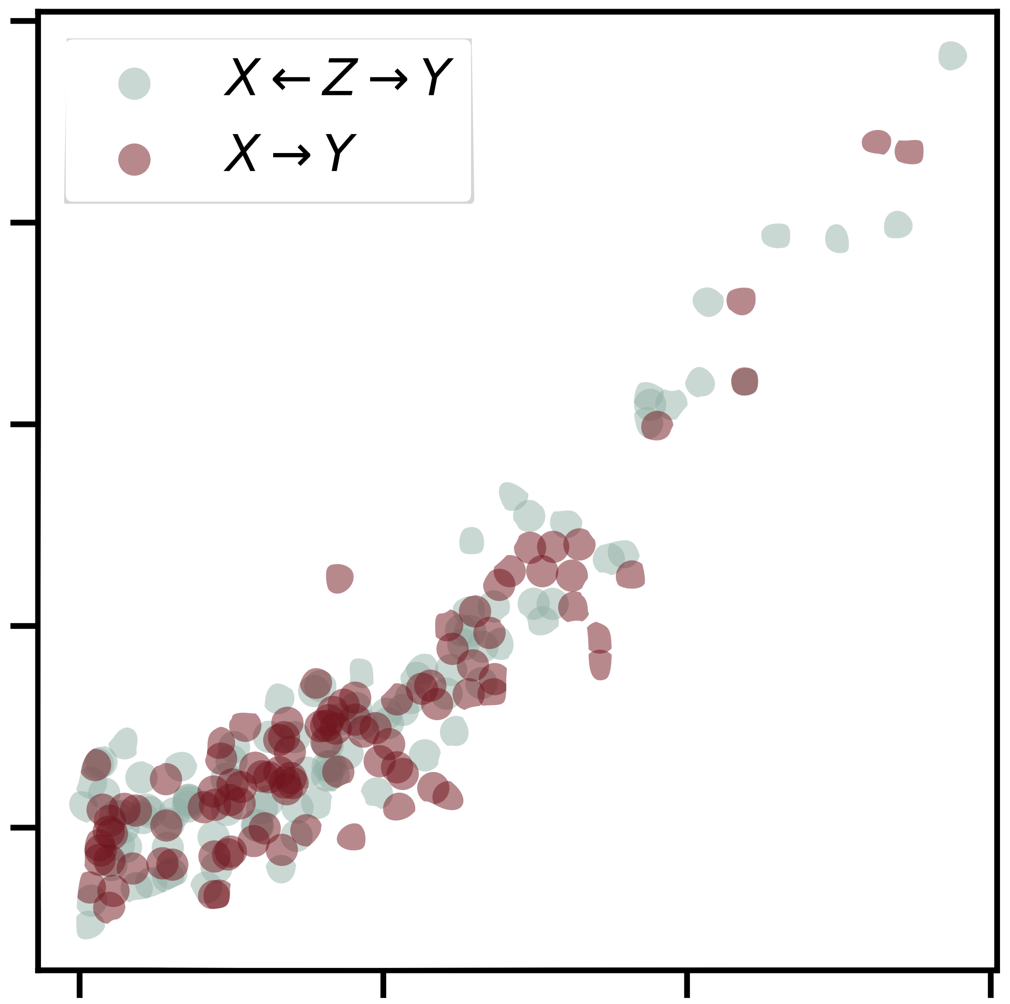
Causal discovery is ill-posed for observational data

Non-parametrically, no reason to favor the “true” factorization

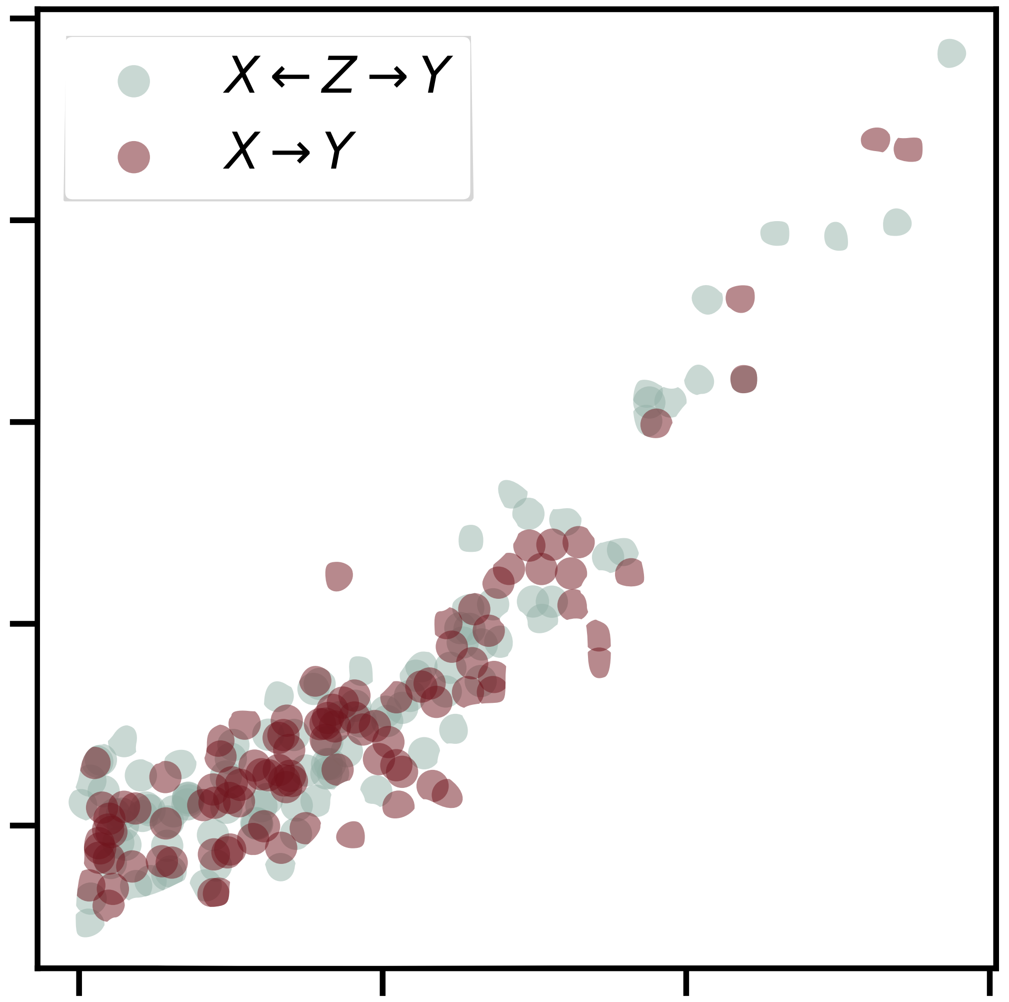
Two (mutually-inclusive) options:

- Intervene on the system
- Make assumptions about the data-generating process

We Need Interventions or Assumptions To Discover Causality

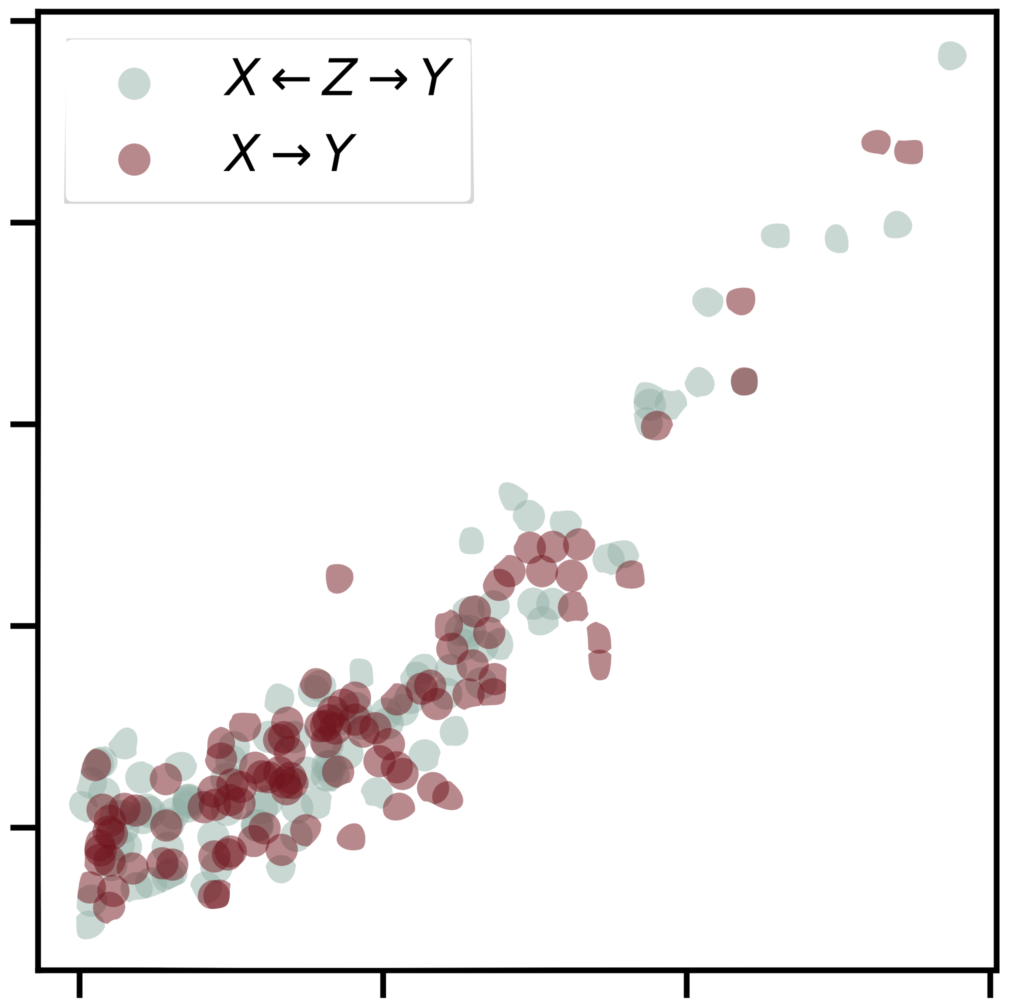


We Need Interventions or Assumptions To Discover Causality



Example assumptions:

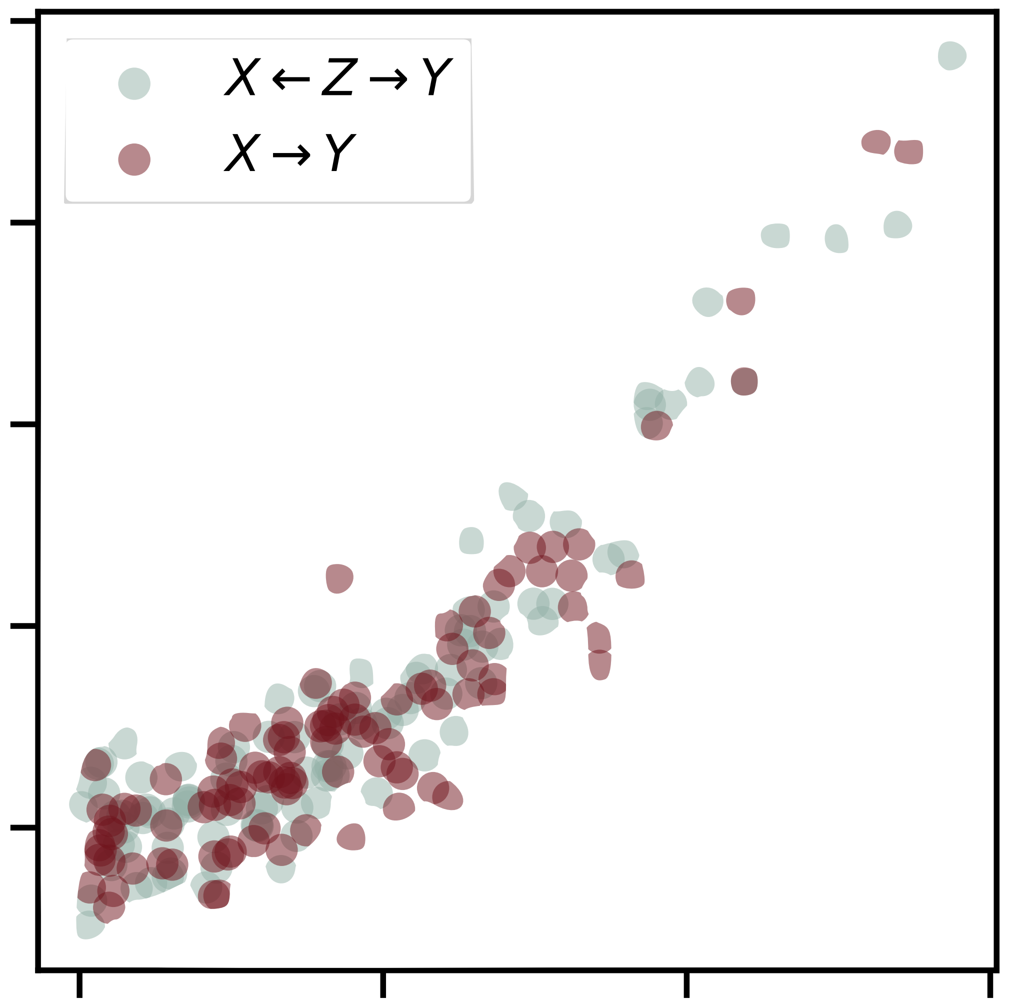
We Need Interventions or Assumptions To Discover Causality



Example assumptions:

- Lack of causations $\implies$ statistical independence (Causal Markov)

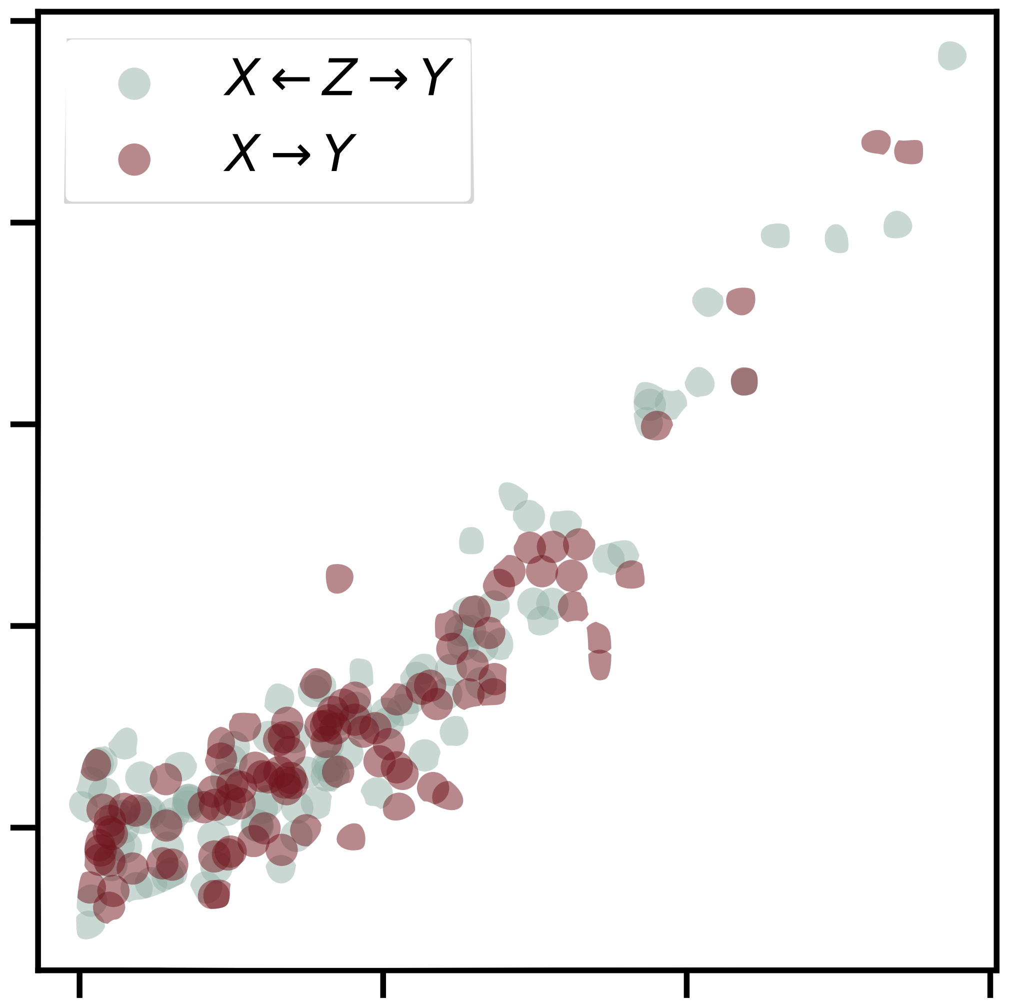
We Need Interventions or Assumptions To Discover Causality



Example assumptions:

- Lack of causations $\Longrightarrow$ statistical independence (Causal Markov)
- Statistical independence $\Longrightarrow$ lack of causation (Causal Faithfulness)

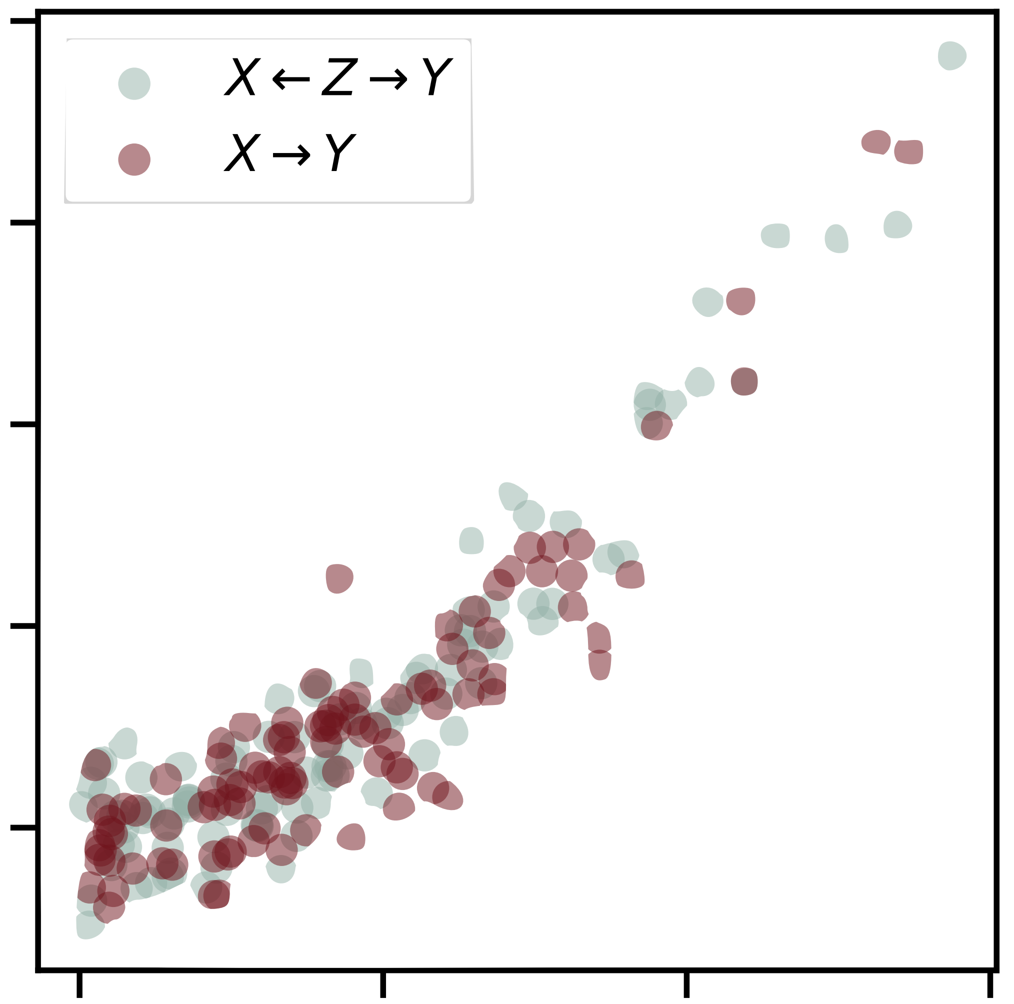
We Need Interventions or Assumptions To Discover Causality



Example assumptions:

- Lack of causations $\implies$ statistical independence (Causal Markov)
- Statistical independence $\implies$ lack of causation (Causal Faithfulness)
- All relevant variables are observed (Causal Sufficiency)

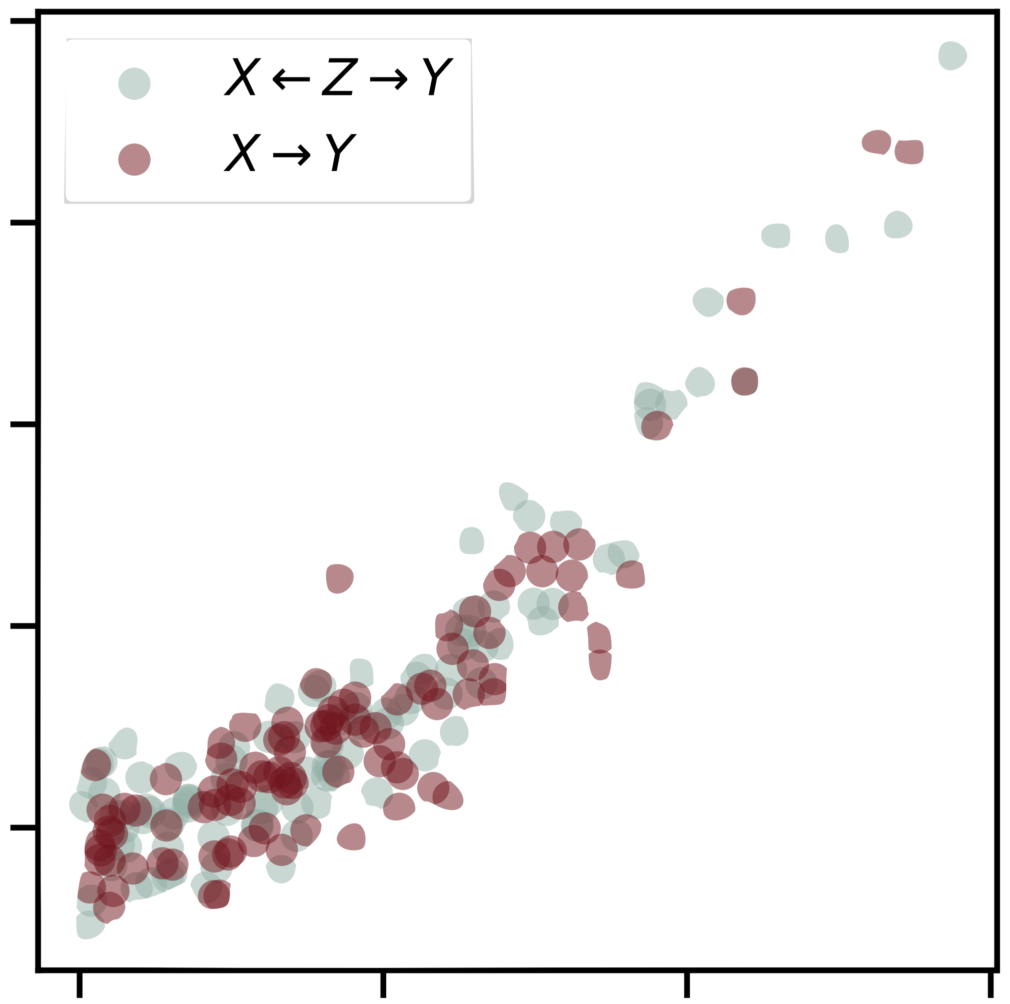
We Need Interventions or Assumptions To Discover Causality



Example assumptions:

- Lack of causations $\implies$ statistical independence (Causal Markov)
- Statistical independence $\implies$ lack of causation (Causal Faithfulness)
- All relevant variables are observed (Causal Sufficiency)
- True causal model maximizes a score function (Causal Identifiability)

We Need Interventions or Assumptions To Discover Causality



Example assumptions:

- Lack of causations $\implies$ statistical independence (Causal Markov)
- Statistical independence $\implies$ lack of causation (Causal Faithfulness)
- All relevant variables are observed (Causal Sufficiency)
- True causal model maximizes a score function (Causal Identifiability)

All of these assumptions break symmetry

An Outline of Today

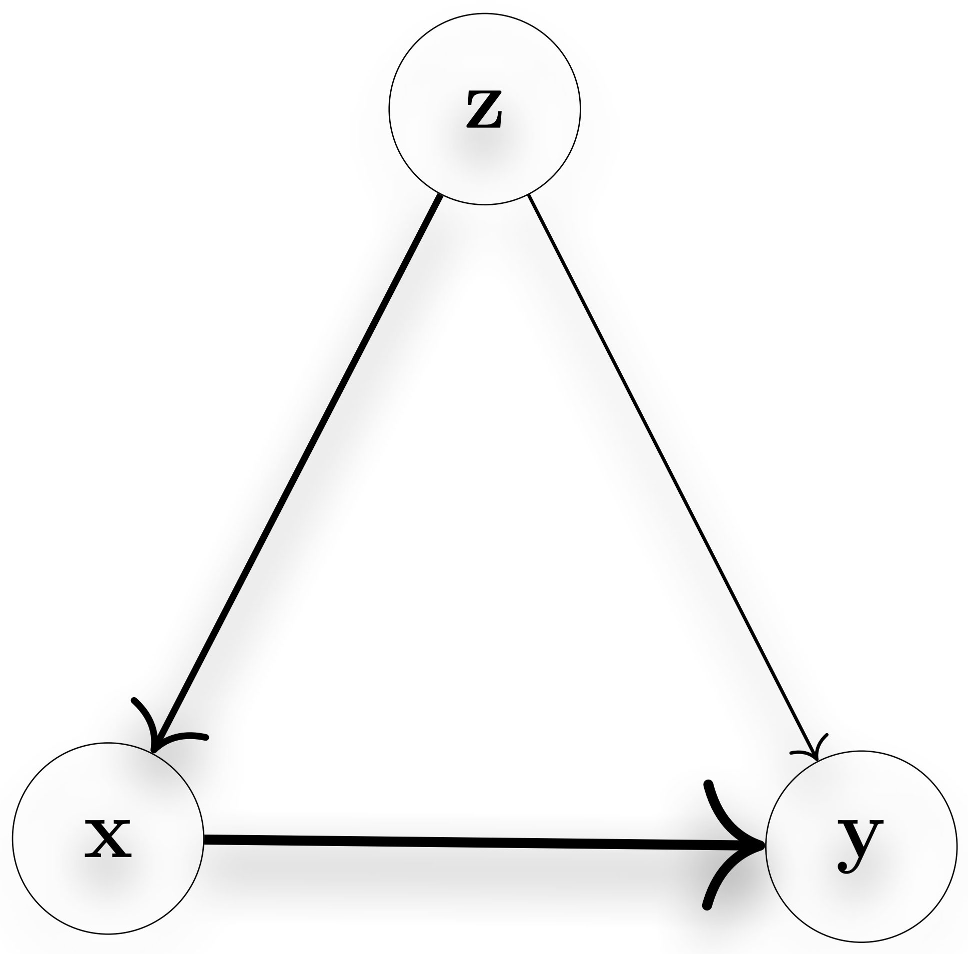
An Outline of Today

Using these notions from causal inference, we'll talk about

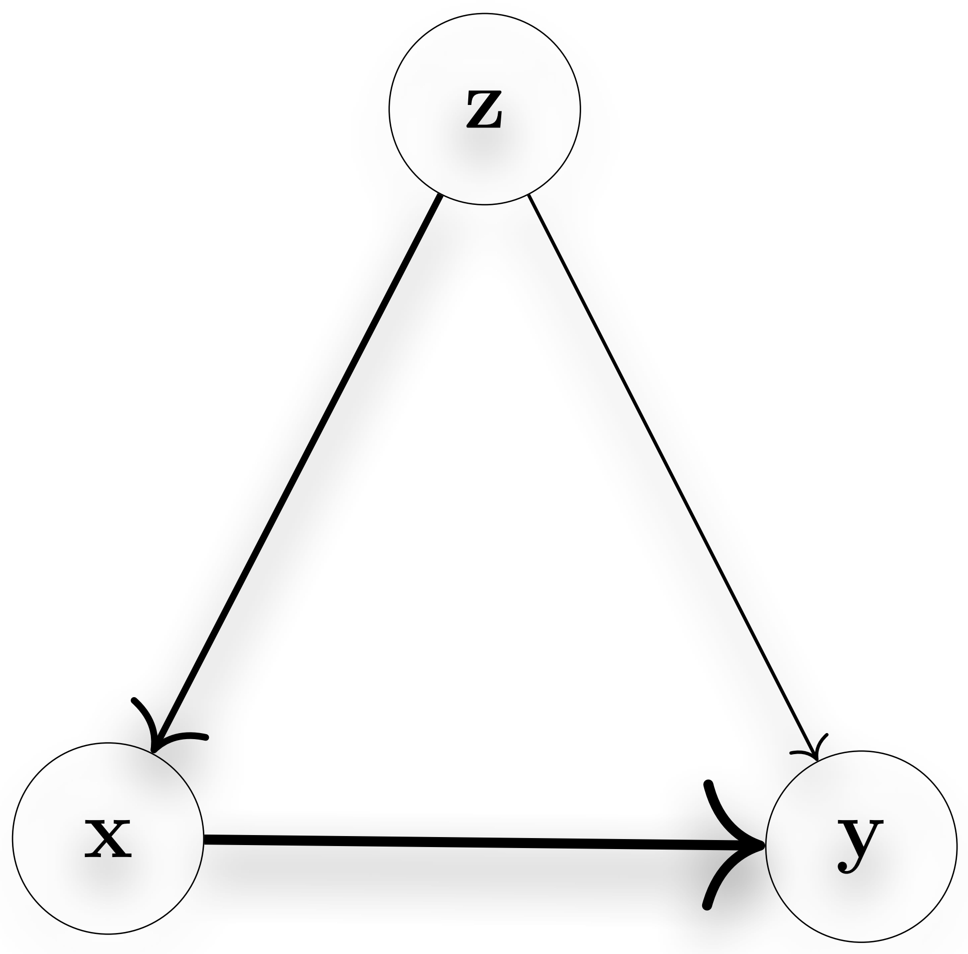
- How to quantify causal strengths
- How to find "hidden common causes" using causal strength
- Using causal strength to discovery causal relationships from observational data
- Some concluding thoughts

Quantifying Causal Strength

Why Causal Strength?

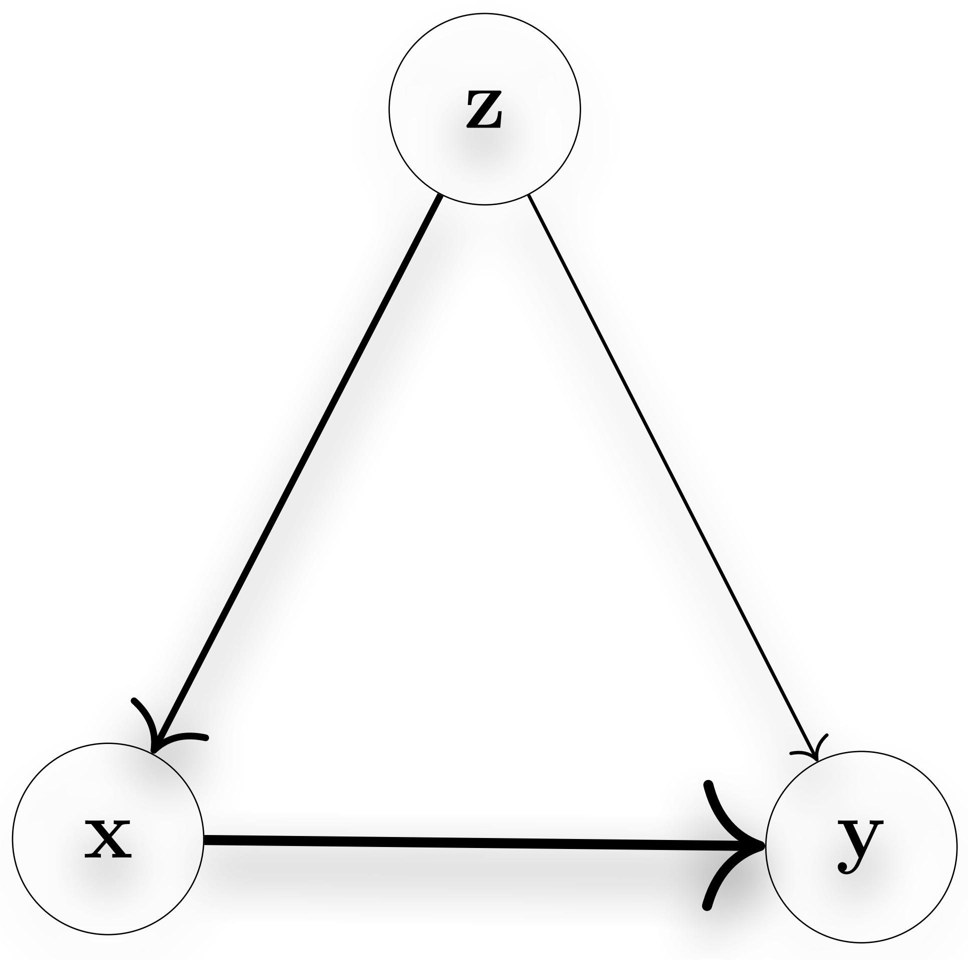


Why Causal Strength?



Practically, we don't only care that a causal relation exists

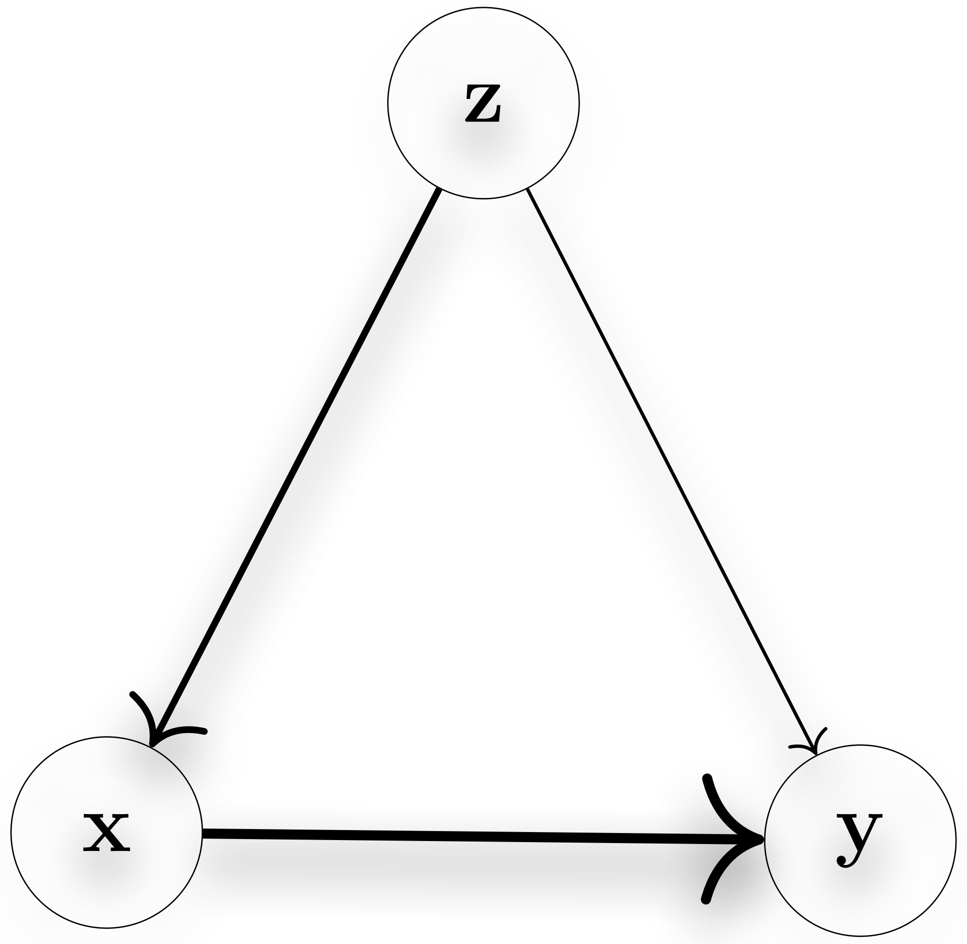
Why Causal Strength?



Practically, we don't only care that a causal relation exists

The strength of a relationship is also important

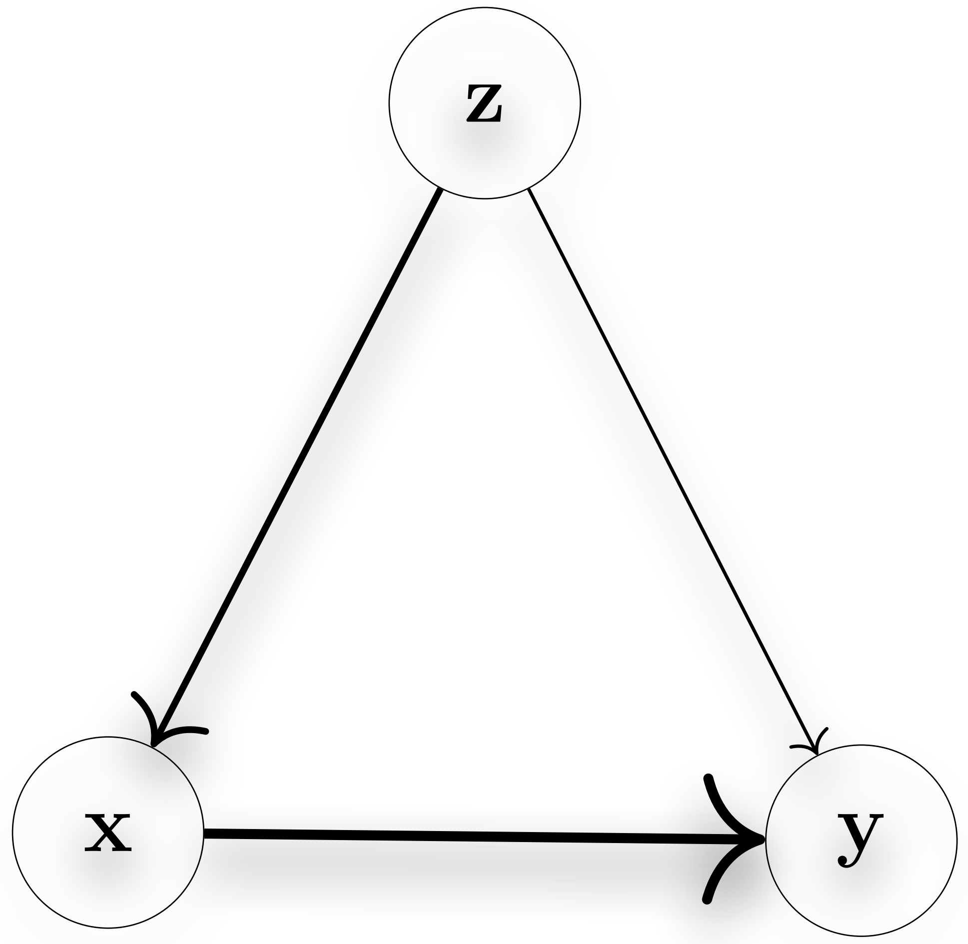
Why Causal Strength?



Practically, we don't only care that a causal relation exists

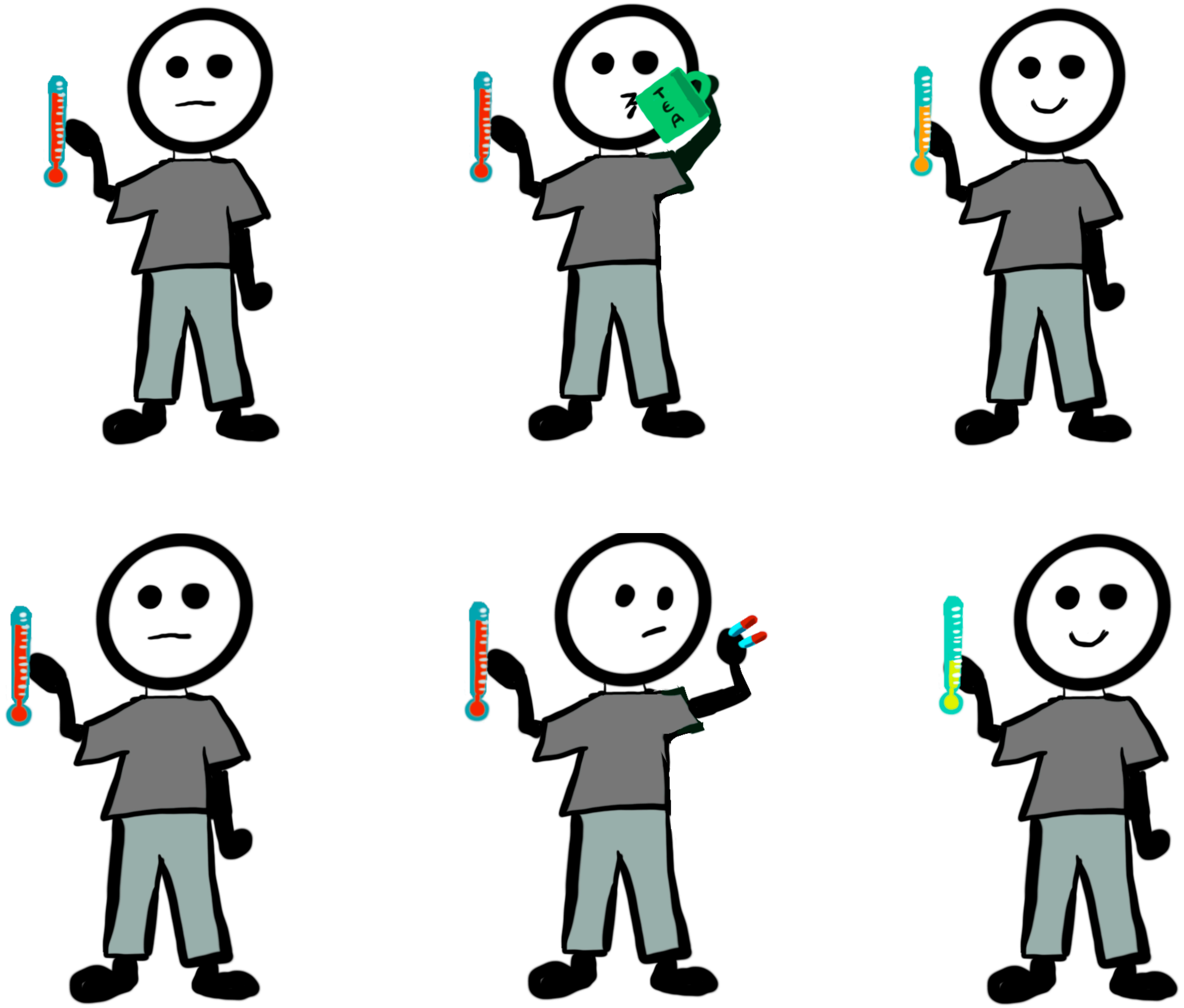
The strength of a relationship is also important

Why Causal Strength?



Practically, we don't only care that a causal relation exists

The strength of a relationship is also important



Defining Causal Strength

Defining Causal Strength

Just as before, definitions are difficult

Defining Causal Strength

Just as before, definitions are difficult

There are many intuitive statistical/information-theoretic attempts

Defining Causal Strength

Just as before, definitions are difficult

There are many intuitive statistical/information-theoretic attempts

For example:

Defining Causal Strength

Just as before, definitions are difficult

There are many intuitive statistical/information-theoretic attempts

For example:

- Fraction of the variance of X_j which is controlled by X_i [ANOVA]

Defining Causal Strength

Just as before, definitions are difficult

There are many intuitive statistical/information-theoretic attempts

For example:

- Fraction of the variance of X_j which is controlled by X_i [ANOVA]
- Kullback-Leibler divergence of marginal vs. conditional distribution [Information Flow]

Defining Causal Strength

Defining Causal Strength

Some interventional attempts:

Defining Causal Strength

Some interventional attempts:

- Kullback-Leibler divergence of interventional distributions with/without specified edge [Janzing et al., AOS 2013]

Defining Causal Strength

Some interventional attempts:

- Kullback-Leibler divergence of interventional distributions with/without specified edge [Janzing et al., AOS 2013]
- Differentiating the expected value under $\mathrm{do}(X_i = x_i)$:

$$\frac{\partial}{\partial x_i}\mathbb{E}[X_j \,|\, \mathrm{do}(X_i = x_i)]$$ [Average Causal Effect]

Causal Effects in Linear Models

Causal Effects in Linear Models

The linear case is instructive for us

$$X_j = \sum_{X_i \in \mathrm{Pa}(X_j)} \beta_{i \to j} X_i + \varepsilon$$

Causal Effects in Linear Models

The linear case is instructive for us

$$X_j = \sum_{X_i \in \mathrm{Pa}(X_j)} \beta_{i \to j} X_i + \varepsilon$$

Then, it's intuitive to link $\beta_{i \to j}$ to "causal strength"

Causal Effects in Linear Models

The linear case is instructive for us

$$X_j = \sum_{X_i \in \mathrm{Pa}(X_j)} \beta_{i \to j} X_i + \varepsilon$$

Then, it's intuitive to link $\beta_{i \to j}$ to "causal strength"

This aligns with the average causal effect, is similar to ANOVA, and more

Causal Effects in Linear Models

The linear case is instructive for us

$$X_j = \sum_{X_i \in \mathrm{Pa}(X_j)} \beta_{i \to j} X_i + \varepsilon$$

Then, it's intuitive to link $\beta_{i \to j}$ to "causal strength"

This aligns with the average causal effect, is similar to ANOVA, and more

How do we move to nonlinear case?

Generalizing to Nonlinear Effects

Generalizing to Nonlinear Effects

A widely used measure is the average causal effect (ACE):

$$\text{ACE}_{X_i \to X_j} \triangleq \frac{\partial}{\partial x_i} \mathbb{E}[X_j \,|\, \mathrm{do}(X_i = x_i)]$$

Generalizing to Nonlinear Effects

A widely used measure is the average causal effect (ACE):

$$\mathrm{ACE}_{X_i \to X_j} \triangleq \frac{\partial}{\partial x_i} \mathbb{E}[X_j \,|\, \mathrm{do}(X_i = x_i)]$$

Our alternative is the differential causal effect (DCE) [Butler et al., SPL 2022]:

$$X_j = f(X_1, \ldots, X_N, \varepsilon)$$

$$\mathrm{DCE}_{X_i \to X_j}(x_i) \triangleq \left[\frac{\partial}{\partial X_i} f(X_1, \ldots, X_N, \varepsilon) \right]_{X_i = x_i}$$

Generalizing to Nonlinear Effects

A widely used measure is the average causal effect (ACE):

$$\mathrm{ACE}_{X_i \to X_j} \triangleq \frac{\partial}{\partial x_i} \mathbb{E}[X_j \,|\, \mathrm{do}(X_i = x_i)]$$

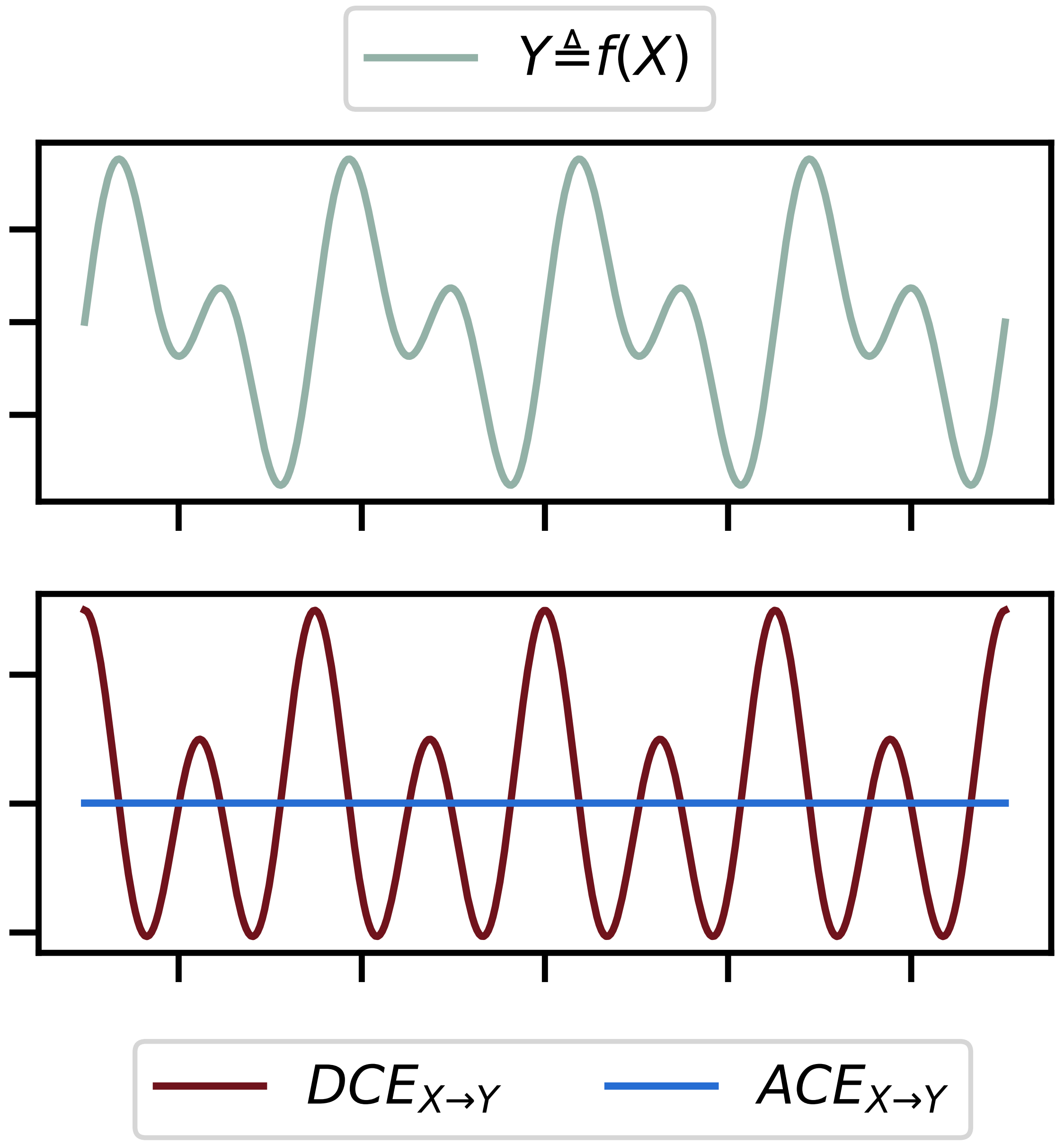
Our alternative is the differential causal effect (DCE) [Butler et al., SPL 2022]:

$$X_j = f(X_1, \ldots, X_N, \varepsilon)$$

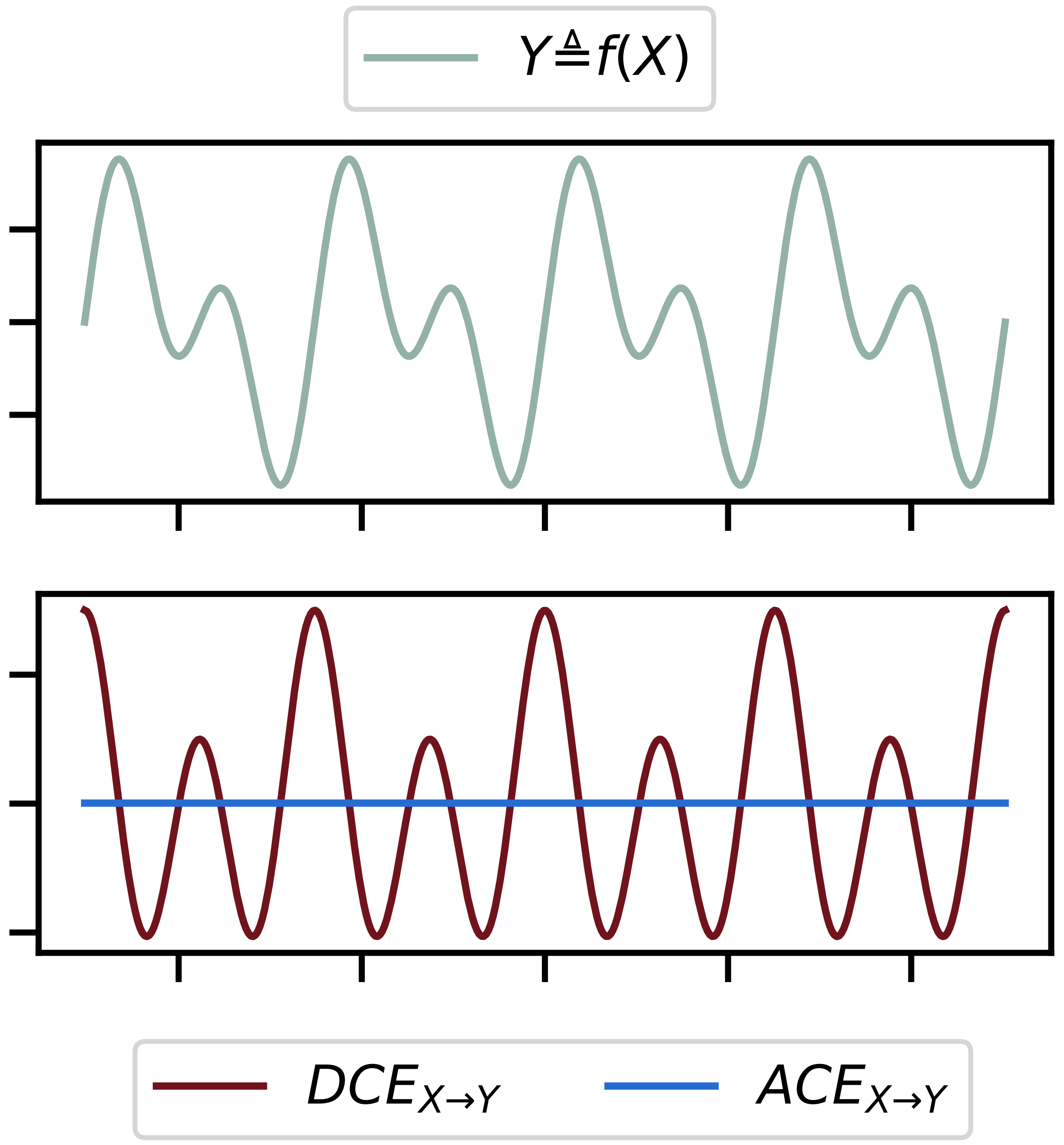
$$\mathrm{DCE}_{X_i \to X_j}(x_i) \triangleq \left[\frac{\partial}{\partial X_i} f(X_1, \ldots, X_N, \varepsilon) \right]_{X_i = x_i}$$

By differentiating under the integral sign, the ACE is the average DCE

The Average Causal Effect Has Issues



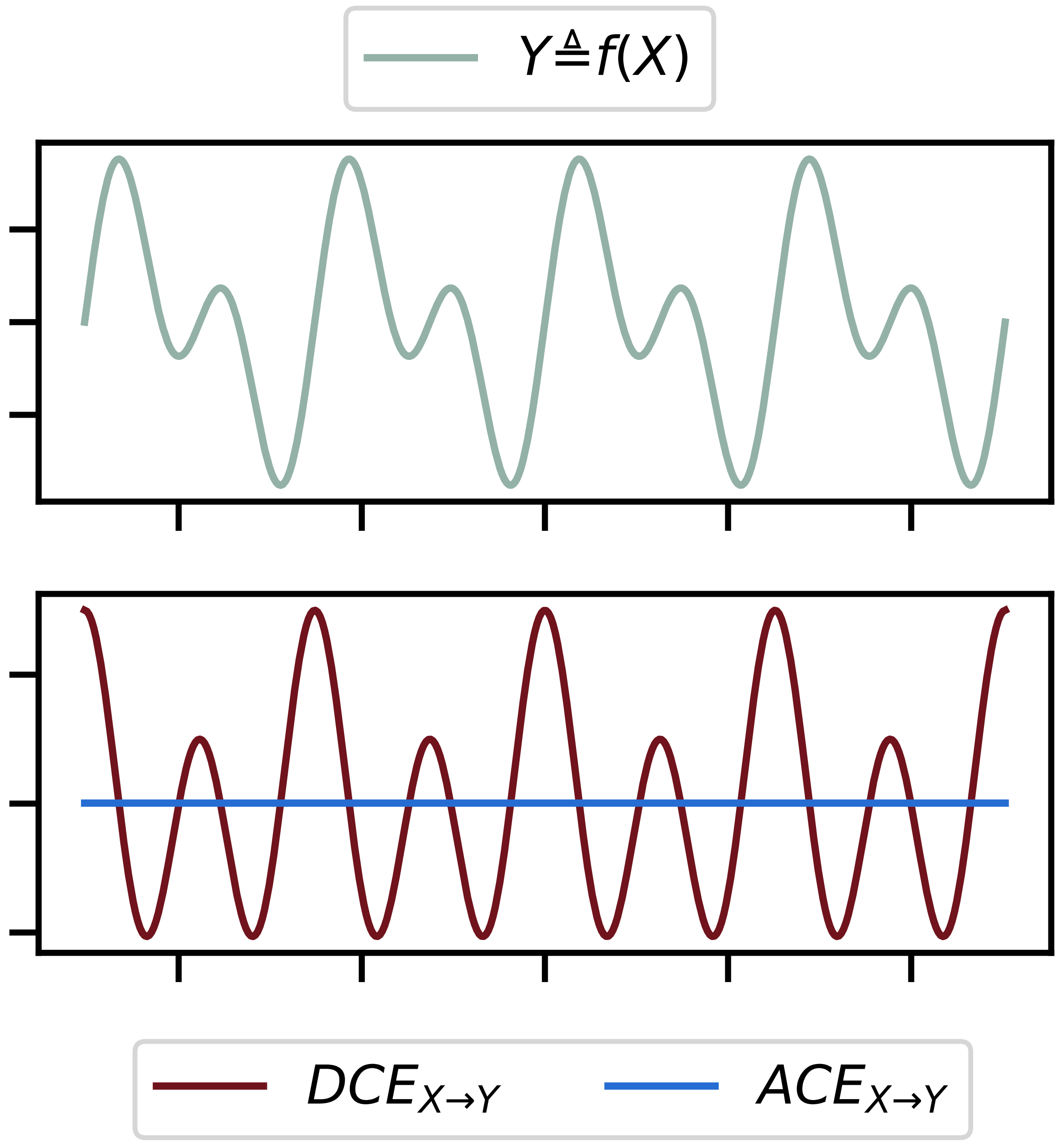
The Average Causal Effect Has Issues



For strength $\mathcal{S}_{X_i \to X_j}$, it's desirable that

$$\mathcal{S}_{X_i \to X_j} \neq 0 \iff X_i \to X_j$$

The Average Causal Effect Has Issues

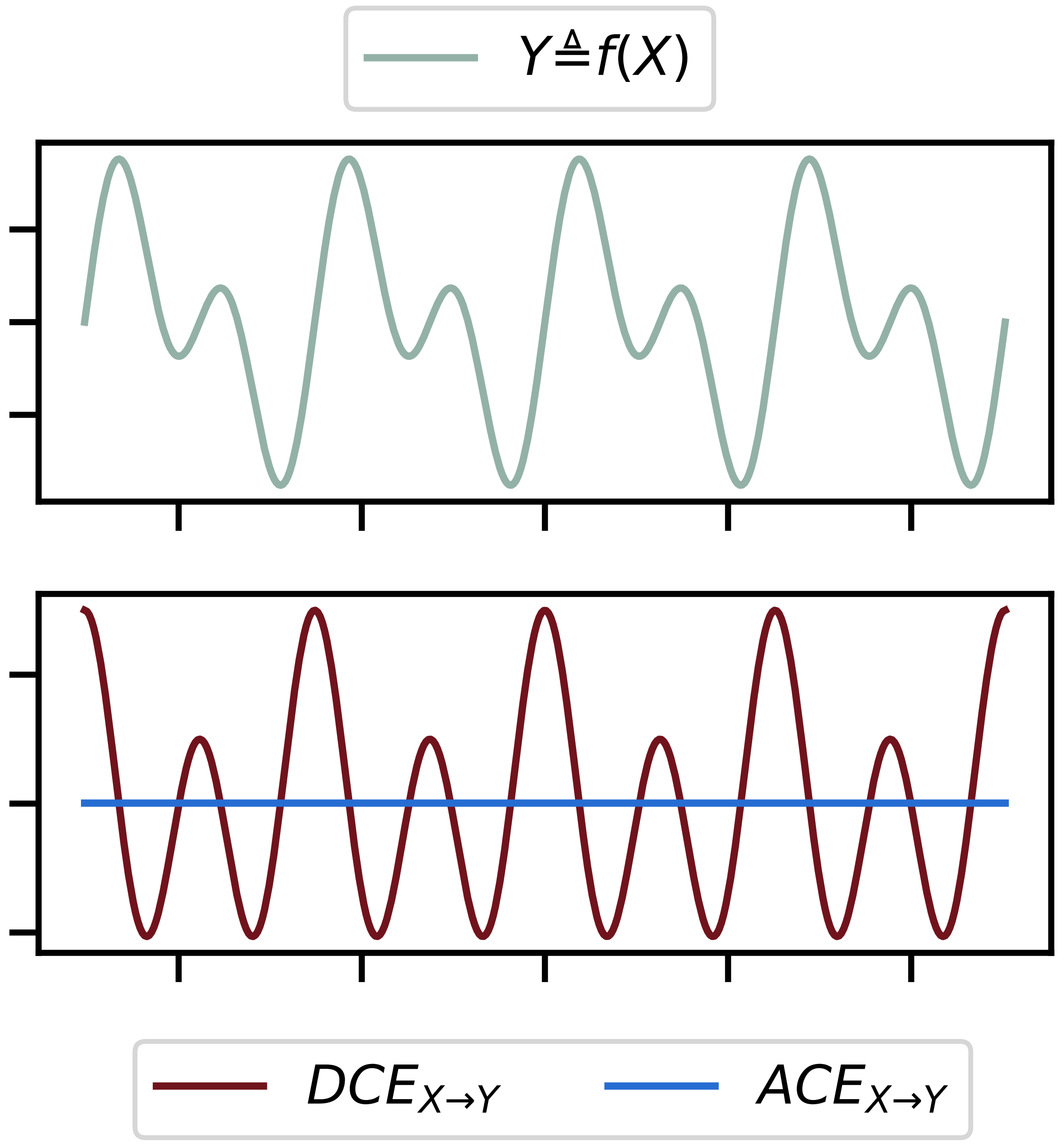


For strength $\mathcal{S}_{X_i \to X_j}$, it's desirable that

$$\mathcal{S}_{X_i \to X_j} \neq 0 \iff X_i \to X_j$$

The ACE fails this test!

The Average Causal Effect Has Issues



For strength $\mathcal{S}_{X_i \to X_j}$, it's desirable that

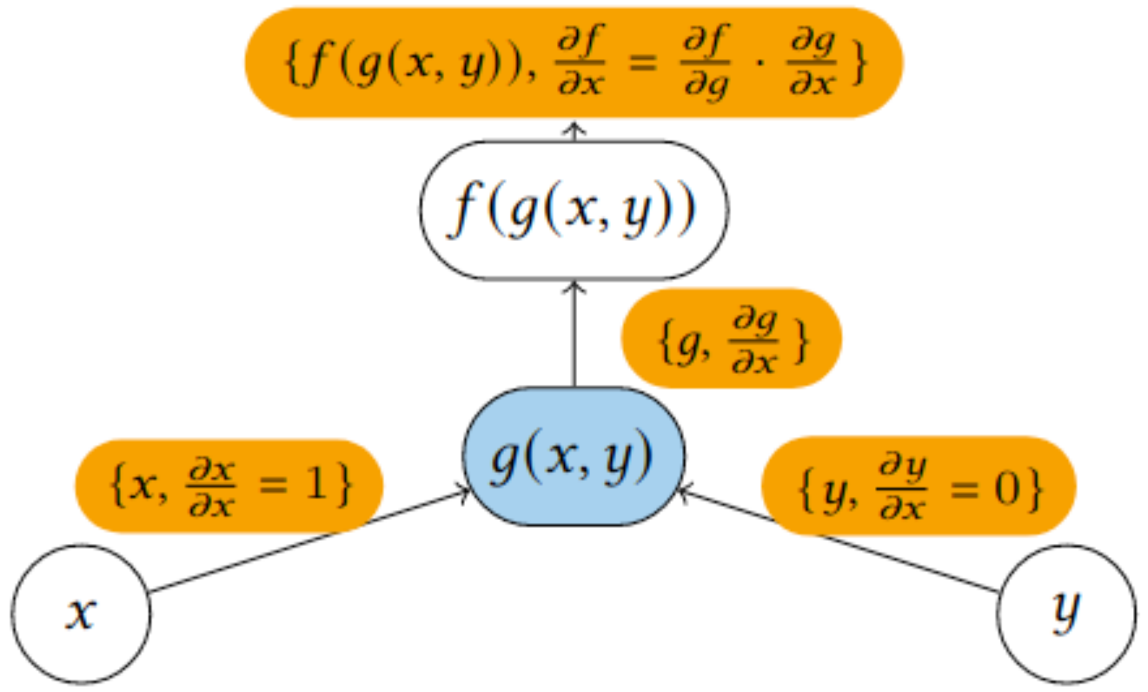
$$\mathcal{S}_{X_i \to X_j} \neq 0 \iff X_i \to X_j$$

The ACE fails this test!

This is the case whenever the DCE is zero-mean, i.e.

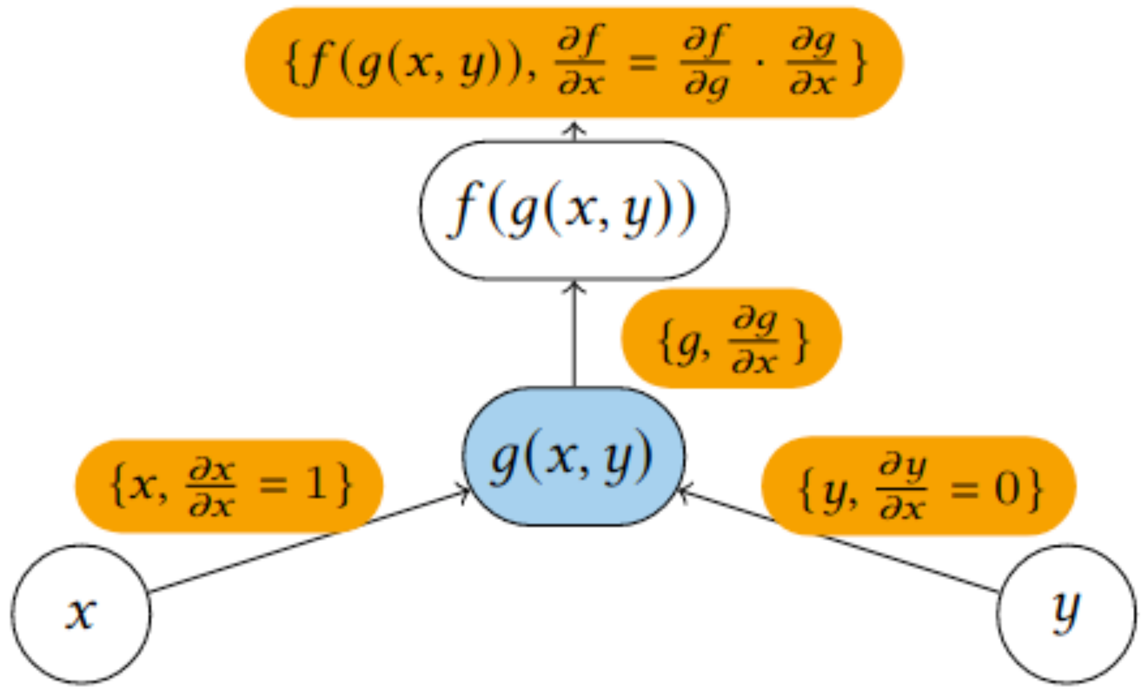
$$\mathbb{E}\left[\mathrm{DCE}_{X_i \to X_j}\right] = 0$$

Calculating the DCE is Easy



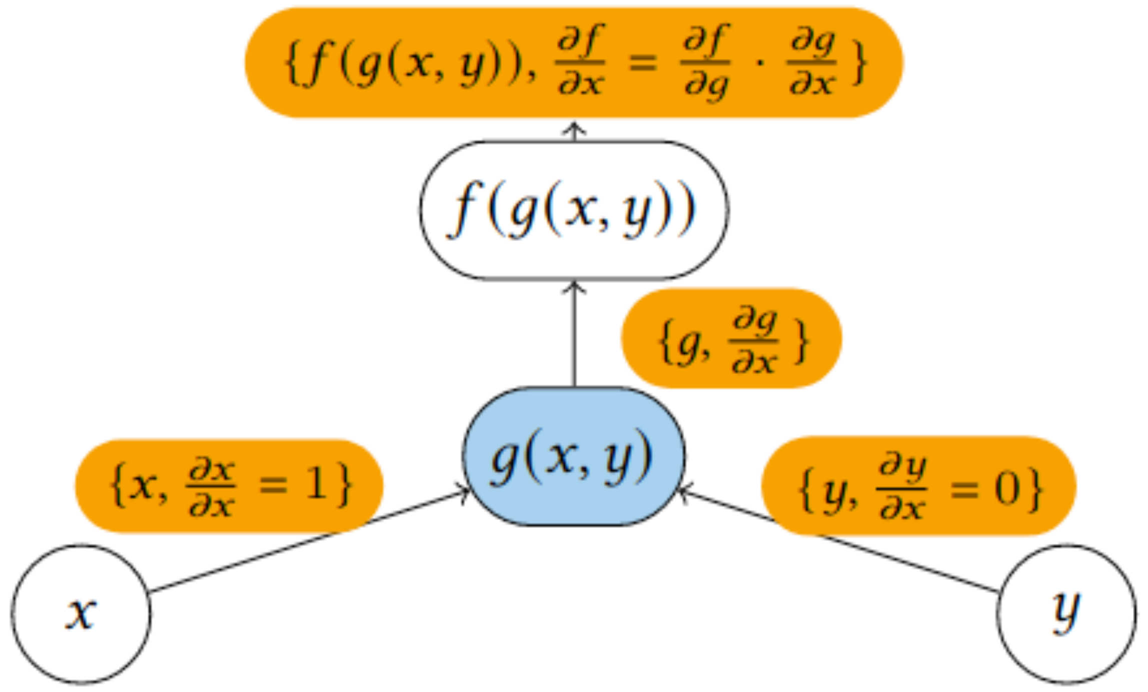
“ForwardAD.png” by MaxEmanuel, Wikimedia Commons, CC-BY-SA-4.0

Calculating the DCE is Easy



How we estimate $\mathrm{DCE}_{X_i \to X_j}$ depends on how we estimate f_j

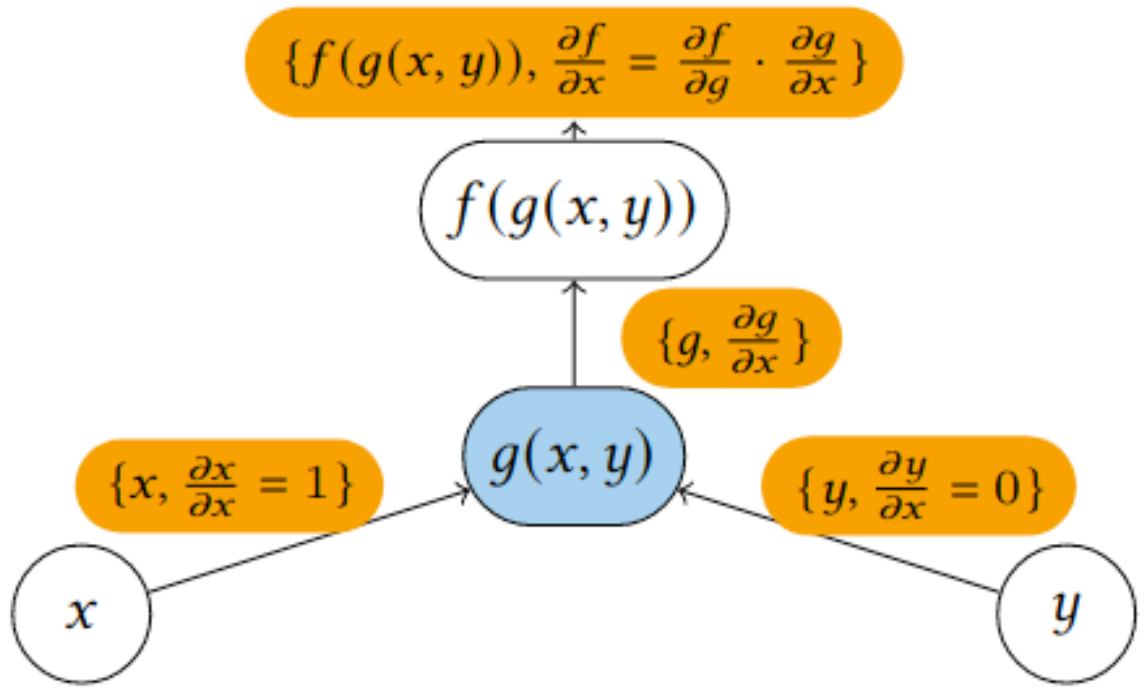
Calculating the DCE is Easy



How we estimate $\mathrm{DCE}_{X_i \to X_j}$ depends on how we estimate f_j

For many estimators, this is available in closed-form

Calculating the DCE is Easy



“ForwardAD.png” by MaxEmanuel, Wikimedia Commons, CC-BY-SA-4.0

How we estimate $\mathrm{DCE}_{X_i \to X_j}$ depends on how we estimate f_j

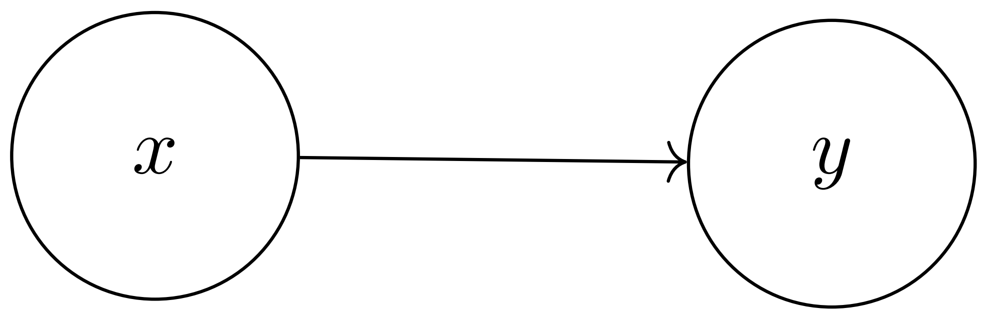
For many estimators, this is available in closed-form

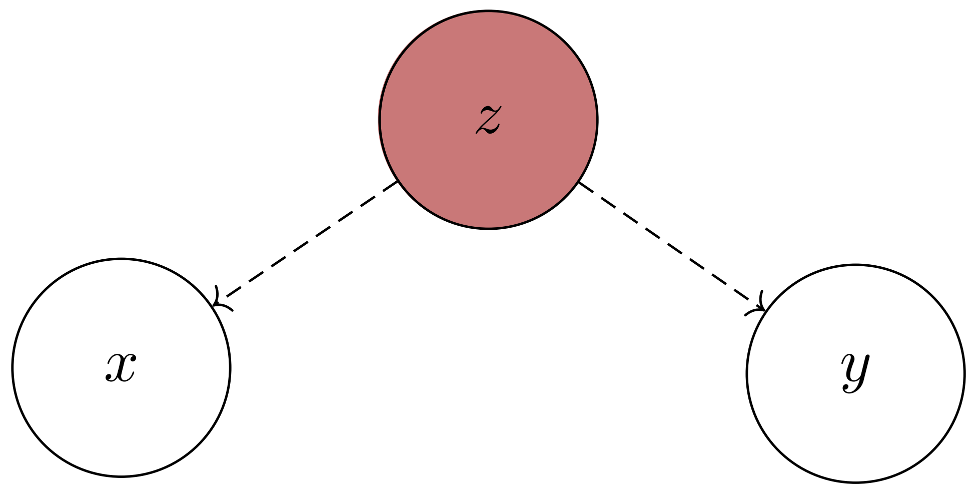
Even more generally, automatic differentiation makes this easy enough

Confounder Detection

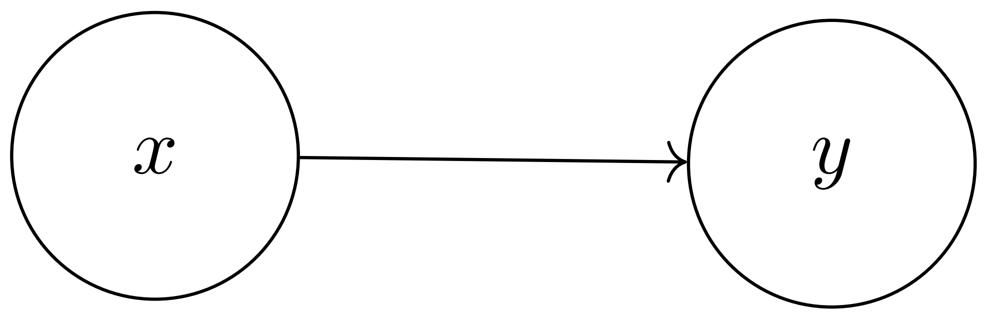
Y. Liu, C, Cui, D. Waxman, K. Butler, and P. M. Djurić,
"Detecting confounders in multivariate time
series using strength of causation,"
Proceedings of the 31st European Signal
Processing Conference, Helsinki, Finland, 2023.

Confounders Are Everywhere

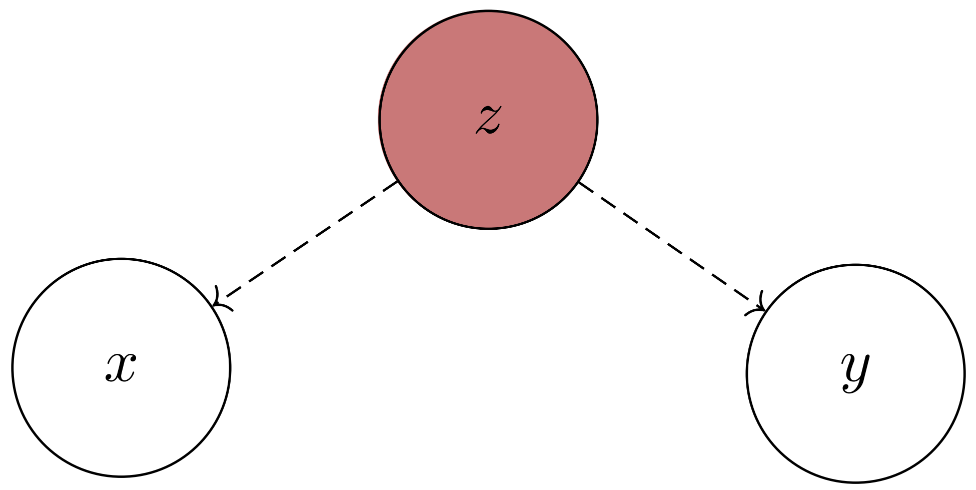




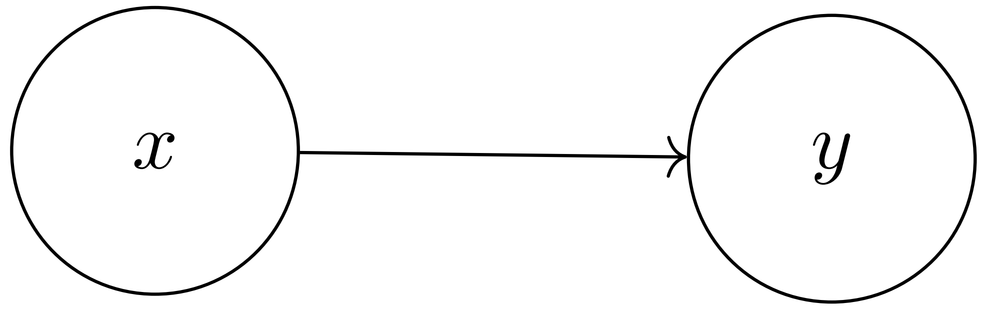
Confounders Are Everywhere



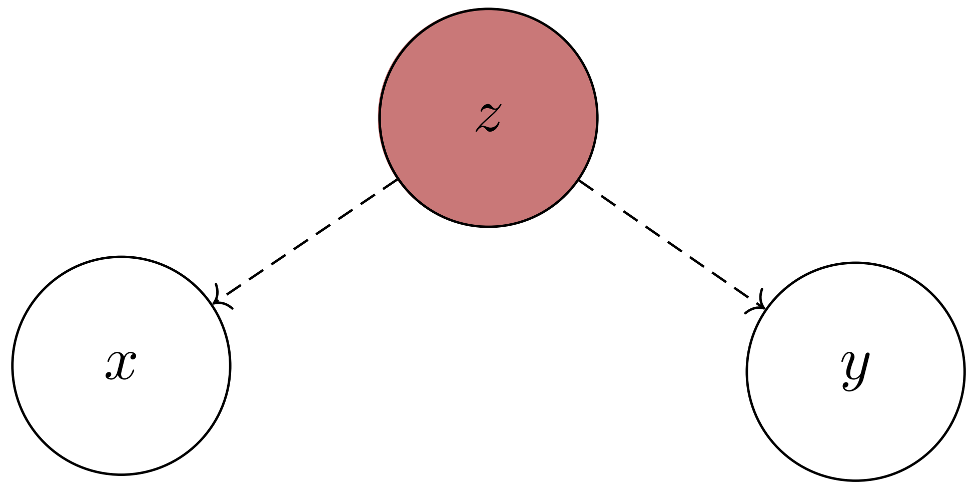
Definition: if there exists a variable Z causing at least two other variables, it is known as a confounder



Confounders Are Everywhere

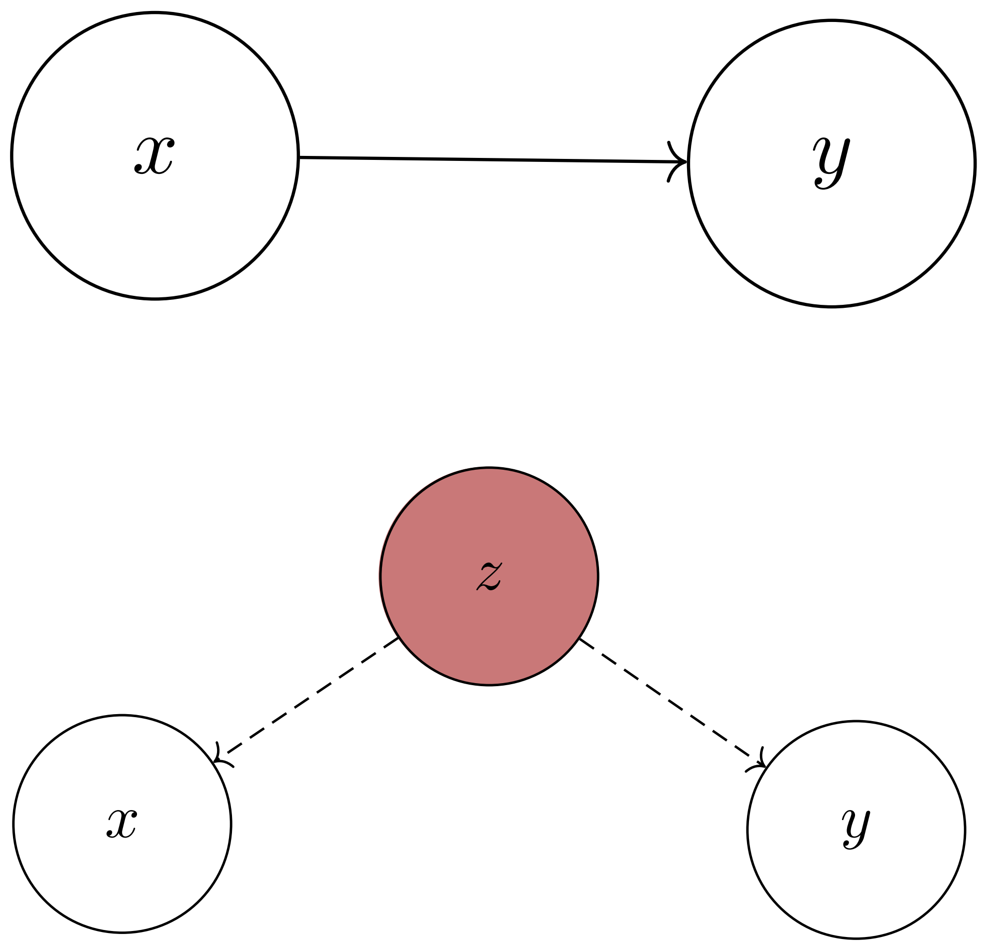


Definition: if there exists a variable Z causing at least two other variables, it is known as a confounder



Confounders complicate causal inference

Confounders Are Everywhere

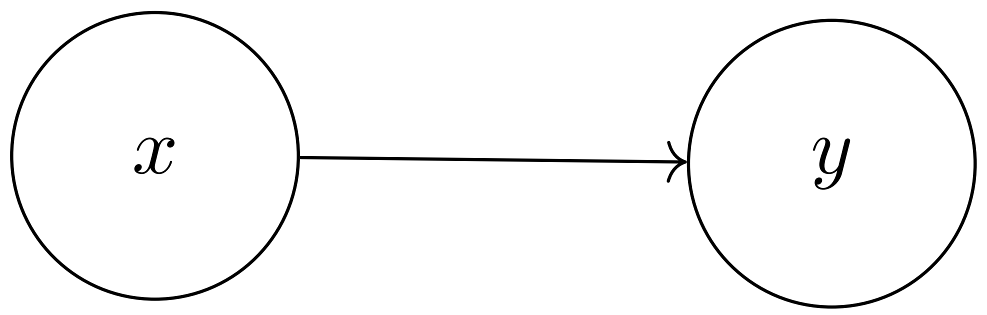


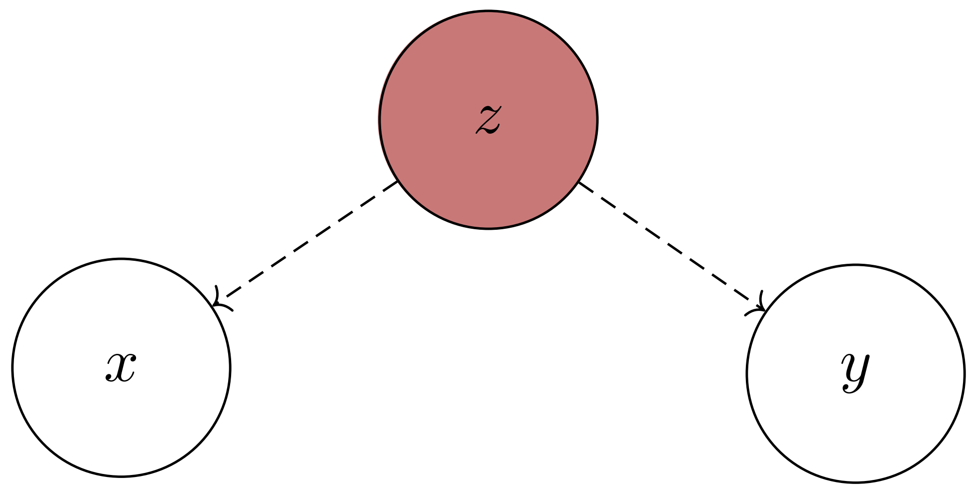
Definition: if there exists a variable Z causing at least two other variables, it is known as a confounder

Confounders complicate causal inference

This complication gets much worse if Z is unobserved

Confounders Are Everywhere





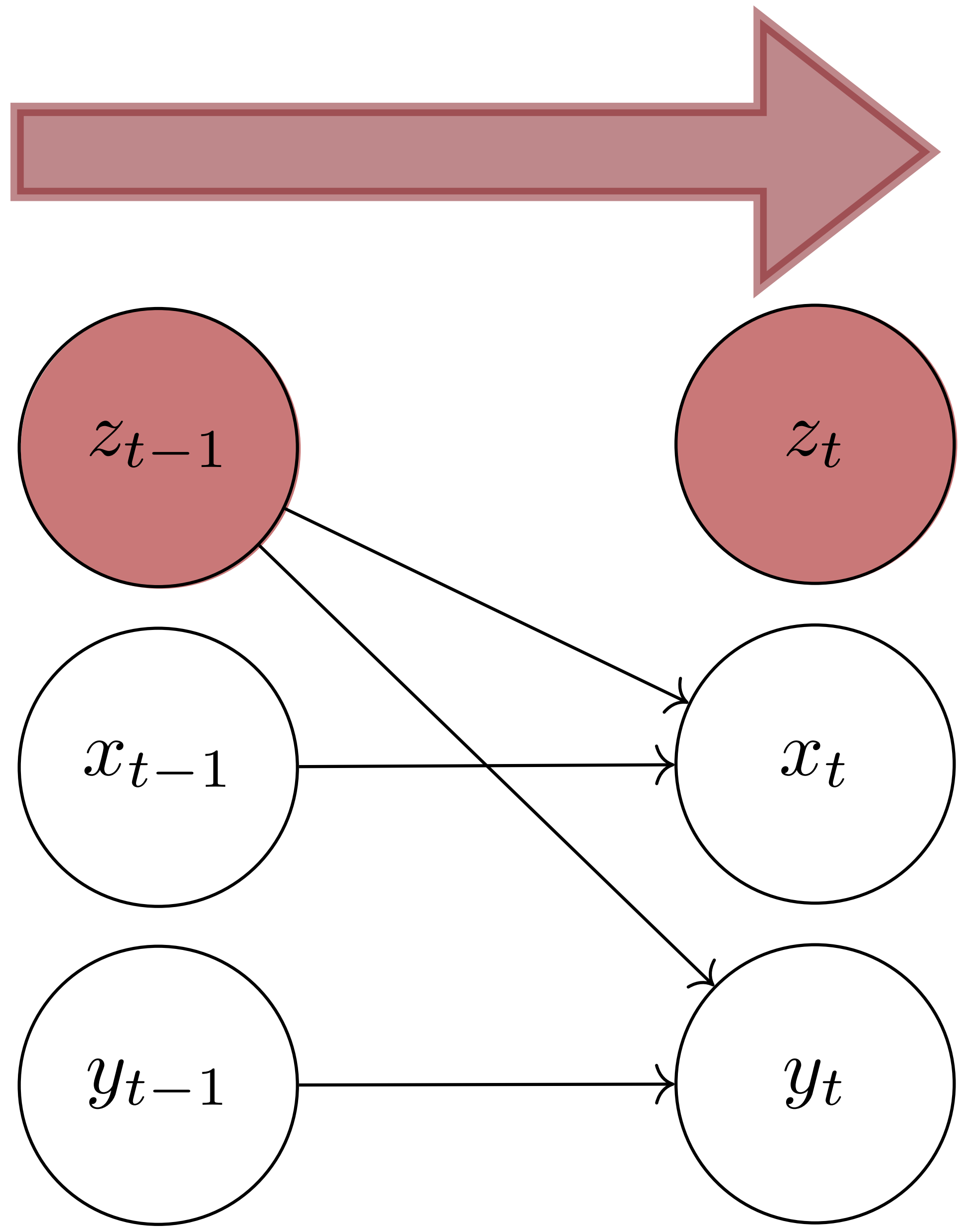
Definition: if there exists a variable Z causing at least two other variables, it is known as a confounder

Confounders complicate causal inference

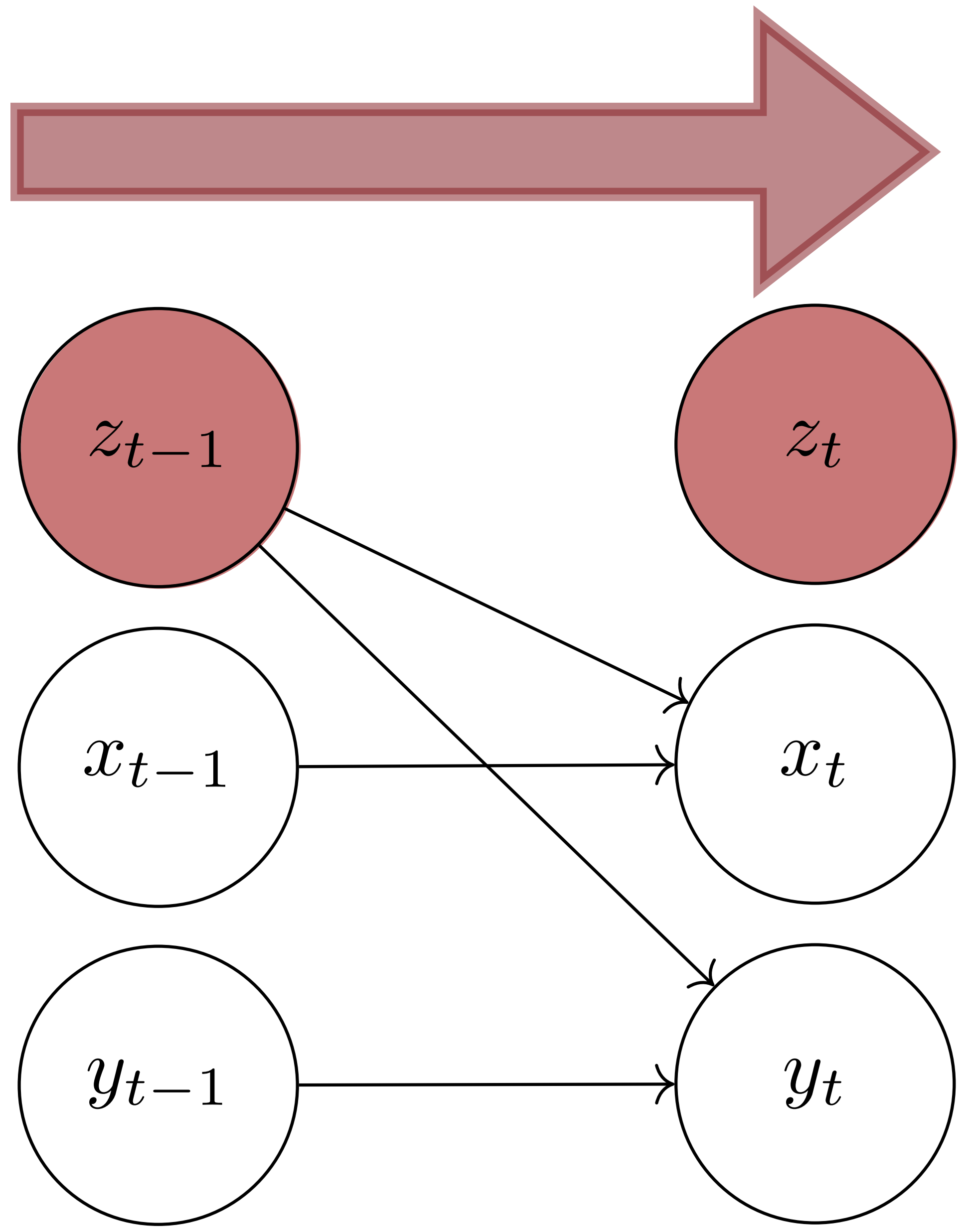
This complication gets much worse if Z is unobserved

One of the most common assumptions is causal sufficiency, i.e. there are no latent confounders

Time Simplifies Causal Structure

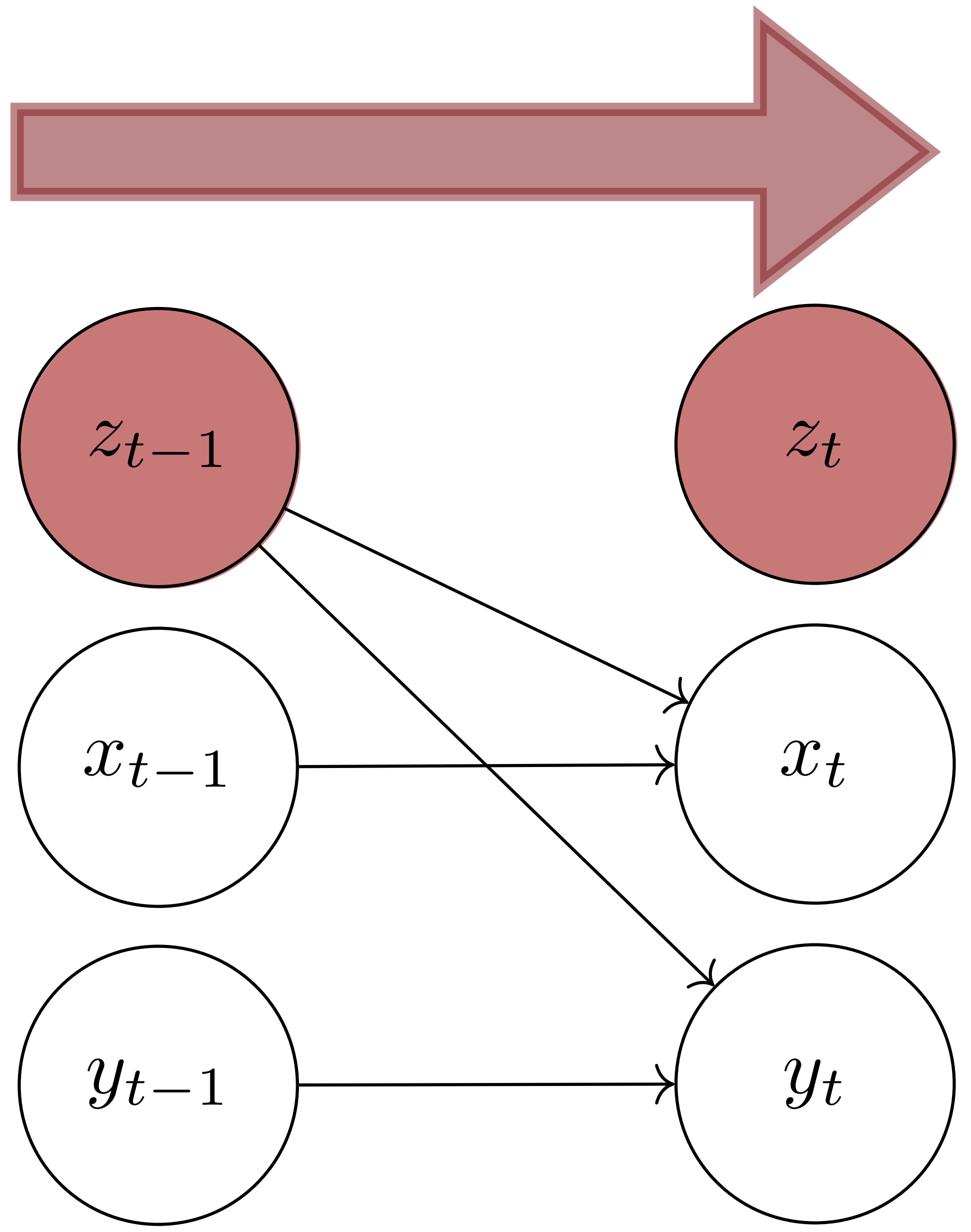


Time Simplifies Causal Structure



In time series, we have the "arrow of time"

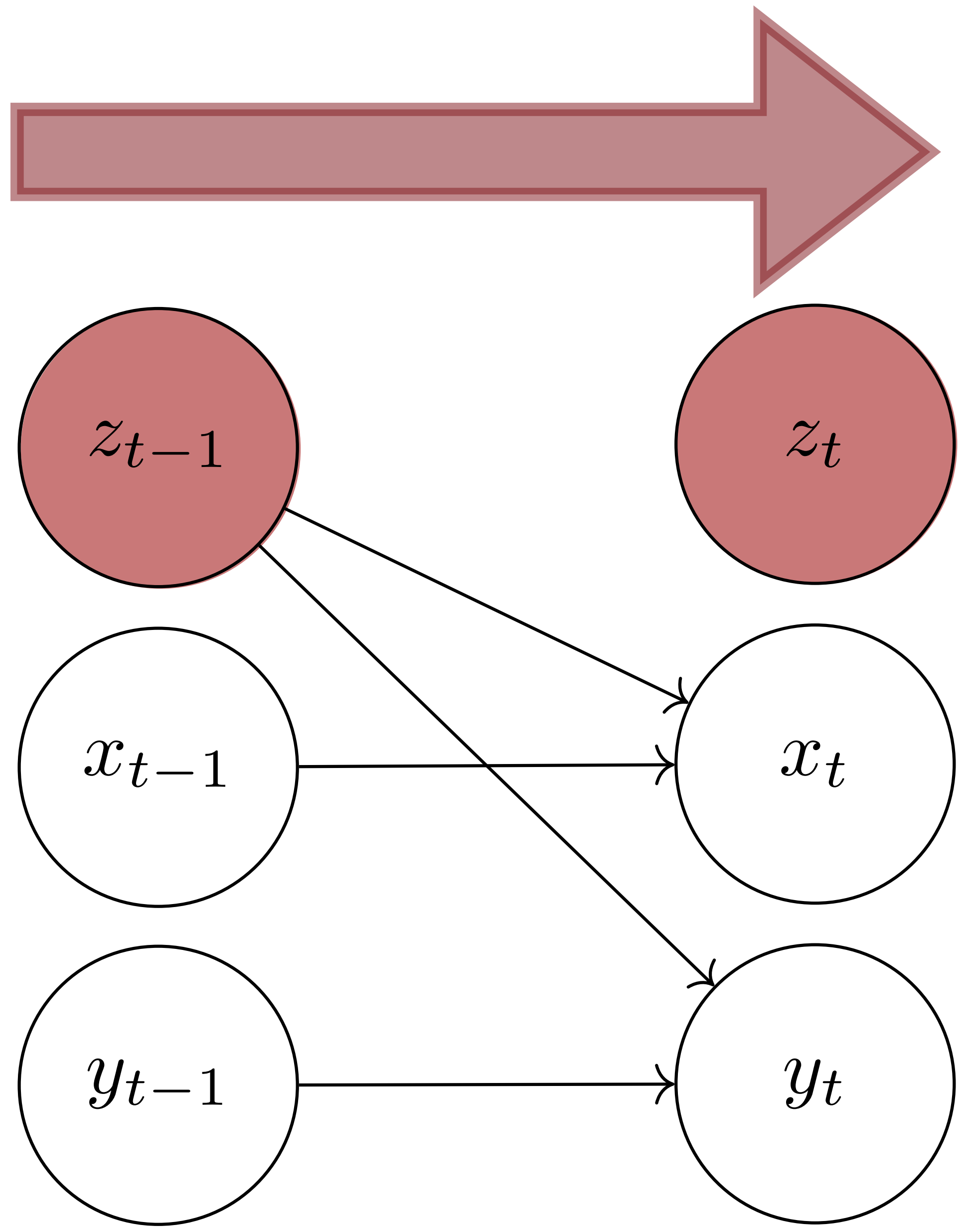
Time Simplifies Causal Structure



In time series, we have the "arrow of time"

This disallows causes from the future affecting the present

Time Simplifies Causal Structure

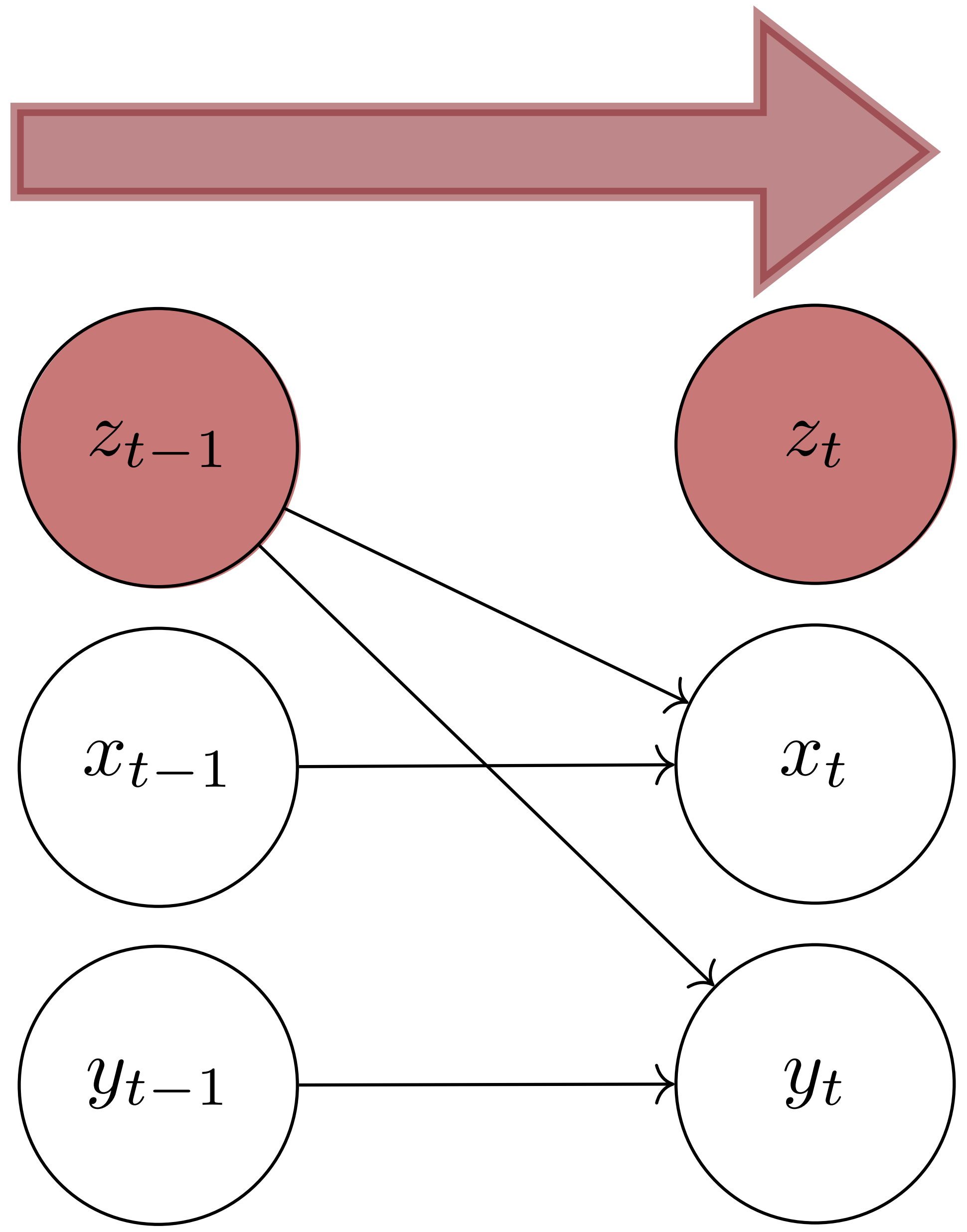


In time series, we have the “arrow of time”

This disallows causes from the future affecting the present

If sampled with sufficiently high rate, we can also disallow “instantaneous causes”

Time Simplifies Causal Structure



In time series, we have the “arrow of time”

This disallows causes from the future affecting the present

If sampled with sufficiently high rate, we can also disallow “instantaneous causes”

The assumption of no instantaneous causes turns discovery into ordinary regression

Some Existing Work on Confounders

Some Existing Work on Confounders

Structure Learning

Confounder Detection

Some Existing Work on Confounders

Structure Learning

- Finds a modified *causal graph* with possible confoundedness indicated

Confounder Detection

Some Existing Work on Confounders

Structure Learning

- Finds a modified *causal graph* with possible confoundedness indicated
    - Constraint-based methods (e.g., FCI [Spirtes et al., MIT Press 2000])

Confounder Detection

Some Existing Work on Confounders

Structure Learning

- Finds a modified *causal graph* with possible confoundedness indicated
  - Constraint-based methods (e.g., FCI [Spirtes et al., MIT Press 2000])
  - Score-based methods (e.g., ICF [Drton & Richardson, UAI 2004])

Confounder Detection

Some Existing Work on Confounders

Structure Learning

- Finds a modified *causal graph* with possible confoundedness indicated
  - Constraint-based methods (e.g., FCI [Spirtes et al., MIT Press 2000])
  - Score-based methods (e.g., ICF [Drton & Richardson, UAI 2004])
  - Asymmetry methods (e.g., LiNGAM [Shimizu et al., JMLR 2006])

Confounder Detection

Some Existing Work on Confounders

Structure Learning

- Finds a modified *causal graph* with possible confoundedness indicated
  - Constraint-based methods (e.g., FCI [Spirtes et al., MIT Press 2000])
  - Score-based methods (e.g., ICF [Drton & Richardson, UAI 2004])
  - Asymmetry methods (e.g., LiNGAM [Shimizu et al., JMLR 2006])
- Some extensions to time series available

Confounder Detection

Some Existing Work on Confounders

Structure Learning

- Finds a modified *causal graph* with possible confoundedness indicated
    - Constraint-based methods (e.g., FCI [Spirtes et al., MIT Press 2000])
    - Score-based methods (e.g., ICF [Drton & Richardson, UAI 2004])
    - Asymmetry methods (e.g., LiNGAM [Shimizu et al., JMLR 2006])
- Some extensions to time series available
    - e.g., LPCMCI [Gerhardus & Runge, NeurIPS 2020] and VAR-LiNGAM [Hyvärinen et al., ICML 2008]

Confounder Detection

Some Existing Work on Confounders

Structure Learning

- Finds a modified *causal graph* with possible confoundedness indicated
    - Constraint-based methods (e.g., FCI [Spirtes et al., MIT Press 2000])
    - Score-based methods (e.g., ICF [Drton & Richardson, UAI 2004])
    - Asymmetry methods (e.g., LiNGAM [Shimizu et al., JMLR 2006])
- Some extensions to time series available
    - e.g., LPCMCI [Gerhardus & Runge, NeurIPS 2020] and VAR-LiNGAM [Hyvärinen et al., ICML 2008]

Confounder Detection

- Aims to *detect* the presence of a latent confounder

Some Existing Work on Confounders

Structure Learning

- Finds a modified *causal graph* with possible confoundedness indicated
  - Constraint-based methods (e.g., FCI [Spirtes et al., MIT Press 2000])
  - Score-based methods (e.g., ICF [Drton & Richardson, UAI 2004])
  - Asymmetry methods (e.g., LiNGAM [Shimizu et al., JMLR 2006])
- Some extensions to time series available
  - e.g., LPCMCI [Gerhardus & Runge, NeurIPS 2020] and VAR-LiNGAM [Hyvärinen et al., ICML 2008]

Confounder Detection

- Aims to *detect* the presence of a latent confounder
  - Spectral methods (e.g., [Janzing & Scholkopf, JCI 2017])

Some Existing Work on Confounders

Structure Learning

- Finds a modified *causal graph* with possible confoundedness indicated
  - Constraint-based methods (e.g., FCI [Spirtes et al., MIT Press 2000])
  - Score-based methods (e.g., ICF [Drton & Richardson, UAI 2004])
  - Asymmetry methods (e.g., LiNGAM [Shimizu et al., JMLR 2006])
- Some extensions to time series available
  - e.g., LPCMCI [Gerhardus & Runge, NeurIPS 2020] and VAR-LiNGAM [Hyvärinen et al., ICML 2008]

Confounder Detection

- Aims to *detect* the presence of a latent confounder
  - Spectral methods (e.g., [Janzing & Scholkopf, JCI 2017])
  - ICA-based methods (e.g., [Janzing & Scholkopf, ICML 2018])

Some Existing Work on Confounders

Structure Learning

- Finds a modified *causal graph* with possible confoundedness indicated
  - Constraint-based methods (e.g., FCI [Spirtes et al., MIT Press 2000])
  - Score-based methods (e.g., ICF [Drton & Richardson, UAI 2004])
  - Asymmetry methods (e.g., LiNGAM [Shimizu et al., JMLR 2006])
- Some extensions to time series available
  - e.g., LPCMCI [Gerhardus & Runge, NeurIPS 2020] and VAR-LiNGAM [Hyvärinen et al., ICML 2008]

Confounder Detection

- Aims to *detect* the presence of a latent confounder
  - Spectral methods (e.g., [Janzing & Scholkopf, JCI 2017])
  - ICA-based methods (e.g., [Janzing & Scholkopf, ICML 2018])
  - Information-theoretic techniques (e.g., [Kaltenpoth & Vreeken, SDM 2019])

Some Existing Work on Confounders

Structure Learning

- Finds a modified *causal graph* with possible confoundedness indicated
  - Constraint-based methods (e.g., FCI [Spirtes et al., MIT Press 2000])
  - Score-based methods (e.g., ICF [Drton & Richardson, UAI 2004])
  - Asymmetry methods (e.g., LiNGAM [Shimizu et al., JMLR 2006])
- Some extensions to time series available
  - e.g., LPCMCI [Gerhardus & Runge, NeurIPS 2020] and VAR-LiNGAM [Hyvärinen et al., ICML 2008]

Confounder Detection

- Aims to *detect* the presence of a latent confounder
  - Spectral methods (e.g., [Janzing & Scholkopf, JCI 2017])
  - ICA-based methods (e.g., [Janzing & Scholkopf, ICML 2018])
  - Information-theoretic techniques (e.g., [Kaltenpoth & Vreeken, SDM 2019])
- Our work: extension to time series using *latent variable models (LVMs) and differential causal effect (DCE)*

Detecting Confounders Through Causal Strength

Detecting Confounders Through Causal Strength

It is impossible to infer what we can’t see

Detecting Confounders Through Causal Strength

It is impossible to infer what we can’t see

It is possible to infer what we can’t see (up to diffeomorphism)

Detecting Confounders Through Causal Strength

It is impossible to infer what we can't see

It is possible to infer what we can't see (up to diffeomorphism)

Idea:

Detecting Confounders Through Causal Strength

Learn an LVM

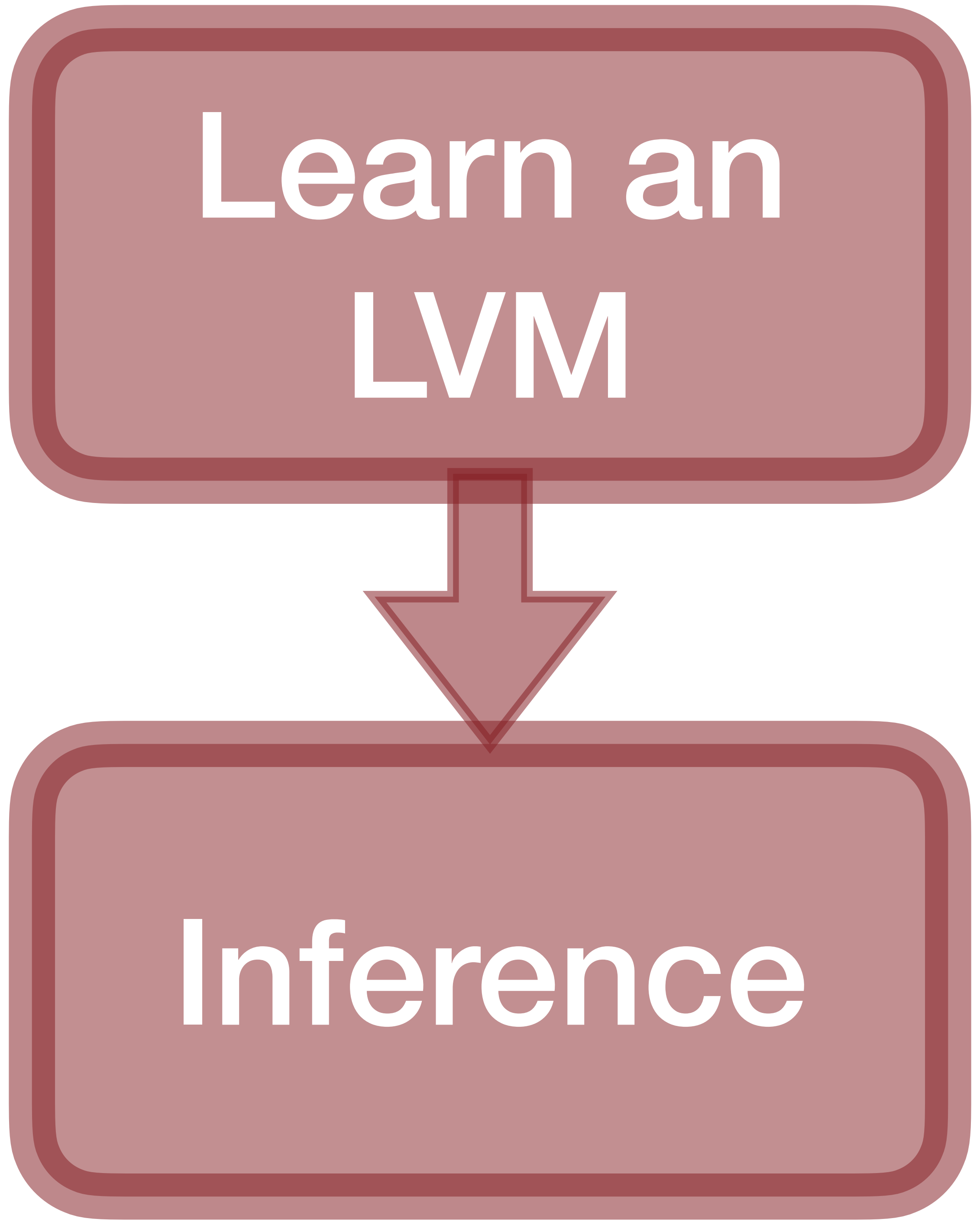
It is impossible to infer what we can't see

It is possible to infer what we can't see (up to diffeomorphism)

Idea:

1. Learn a latent variable model (LV)

Detecting Confounders Through Causal Strength



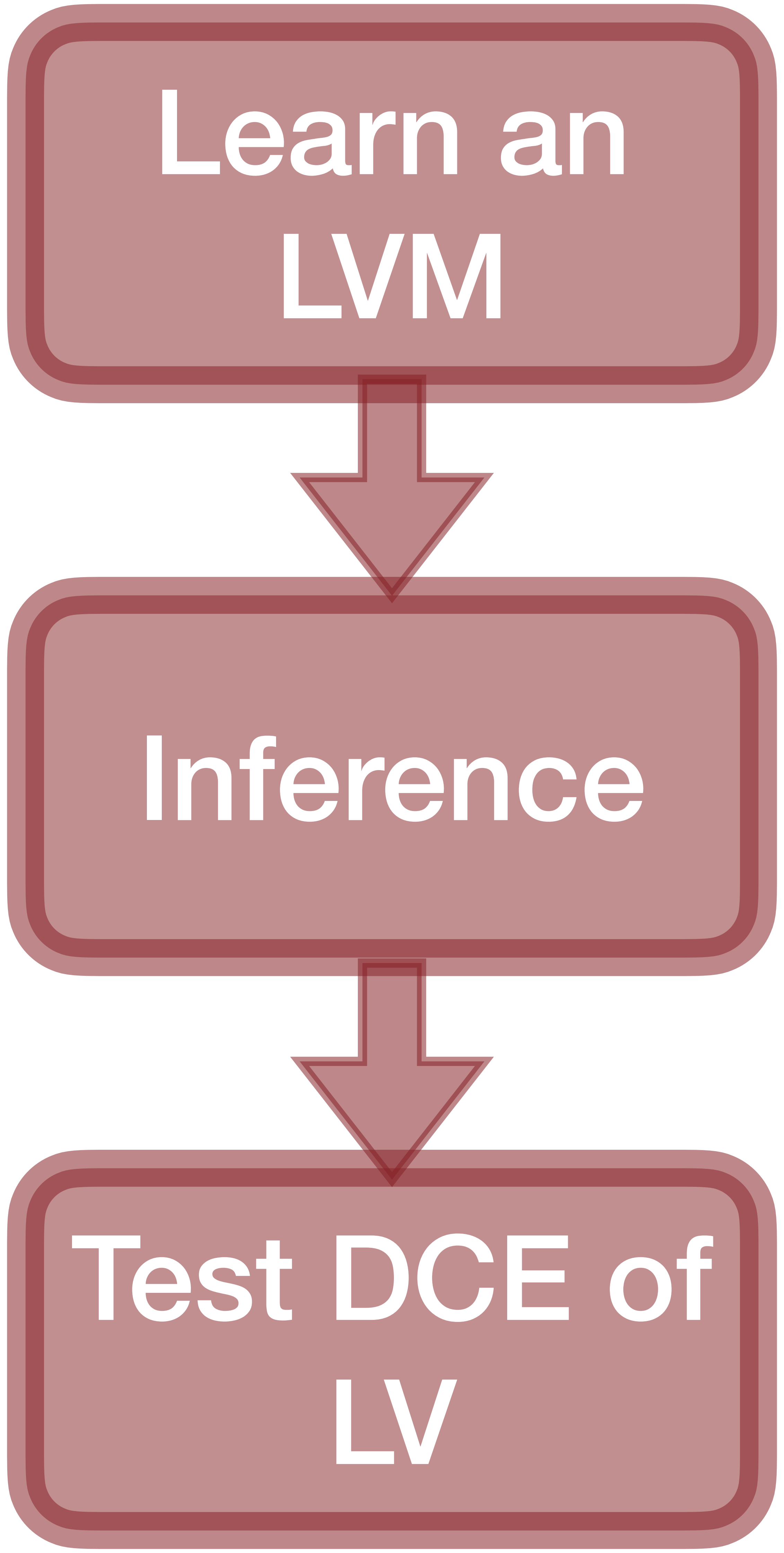
It is impossible to infer what we can't see

It is possible to infer what we can't see (up to diffeomorphism)

Idea:

1. Learn a latent variable model (LV)
2. Perform inference of latent time series z_t

Detecting Confounders Through Causal Strength



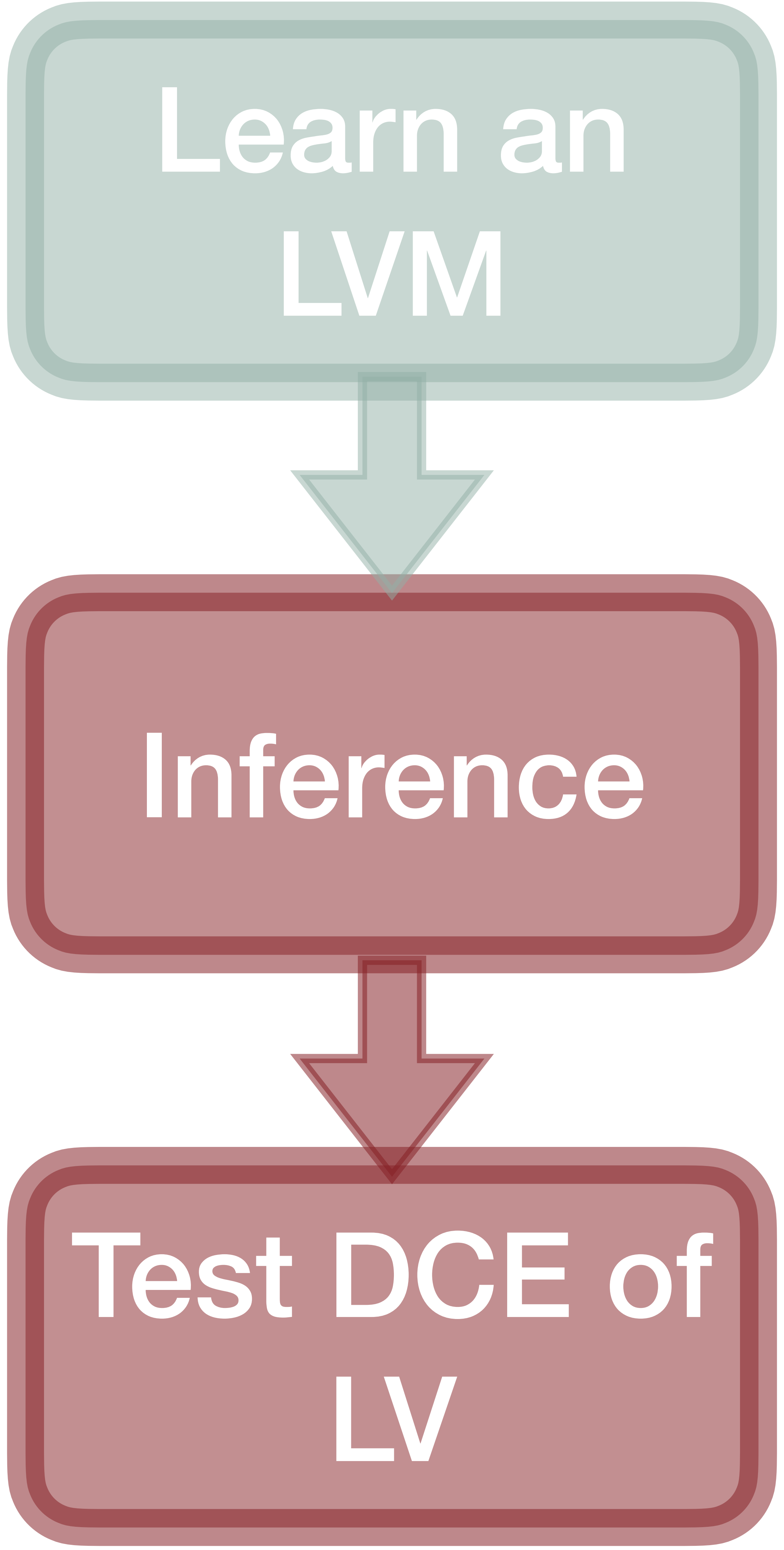
It is impossible to infer what we can't see

It is possible to infer what we can't see (up to diffeomorphism)

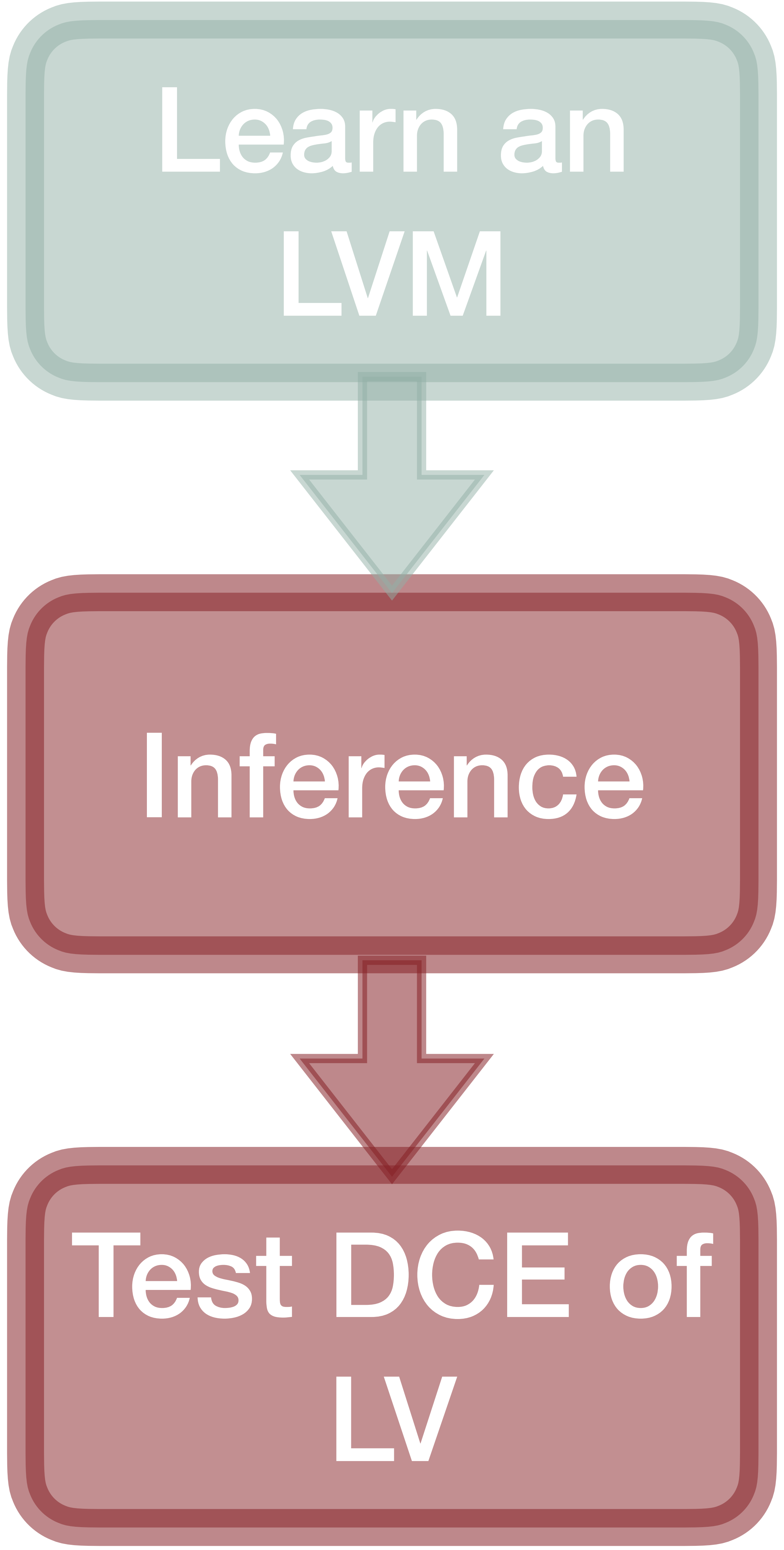
Idea:

1. Learn a latent variable model (LV)
2. Perform inference of latent time series z_t
3. Test the DCE of z_{t-i} to y_t

Gaussian Processes for State Space Models

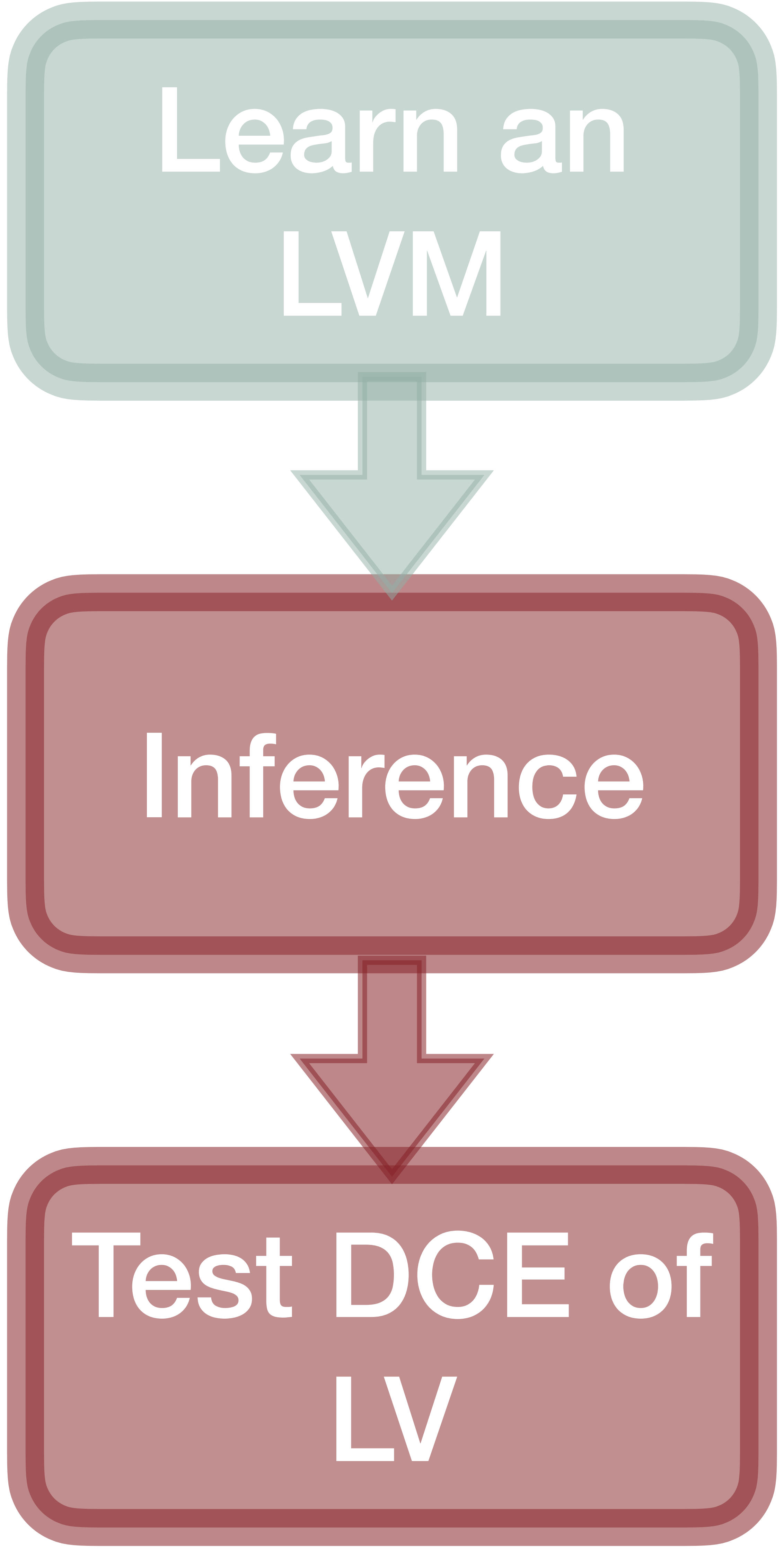


Gaussian Processes for State Space Models



The exact LVM is not so important, but we desire online learning in multivariate time series

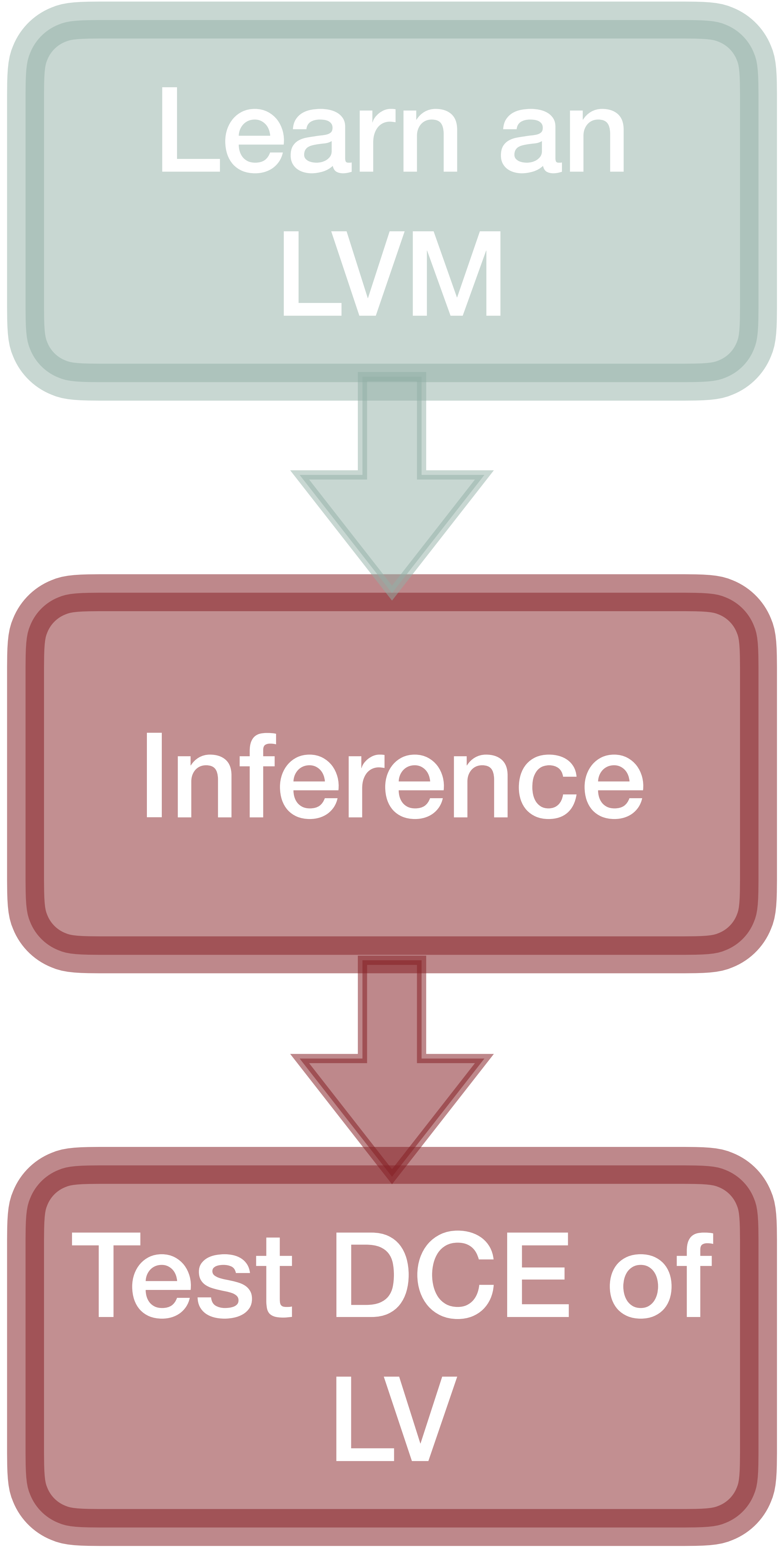
Gaussian Processes for State Space Models



The exact LVM is not so important, but we desire online learning in multivariate time series

We use deep Gaussian process state-space models [Liu et al., TSP 2023]

Gaussian Processes for State Space Models

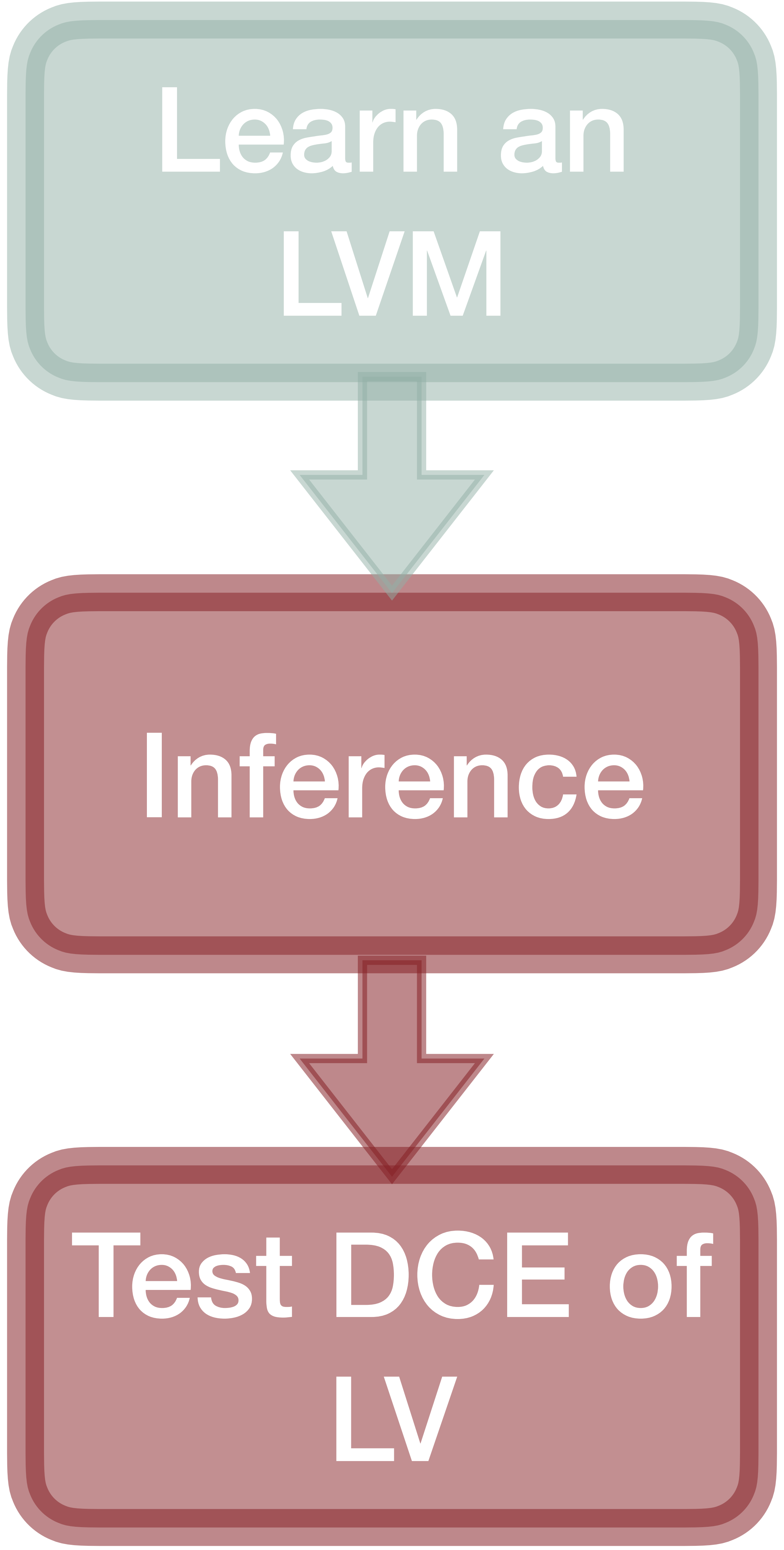


The exact LVM is not so important, but we desire online learning in multivariate time series

We use deep Gaussian process state-space models [Liu et al., TSP 2023]

These write time-series auto regressively with Gaussian processes (GPs)

Gaussian Processes for State Space Models



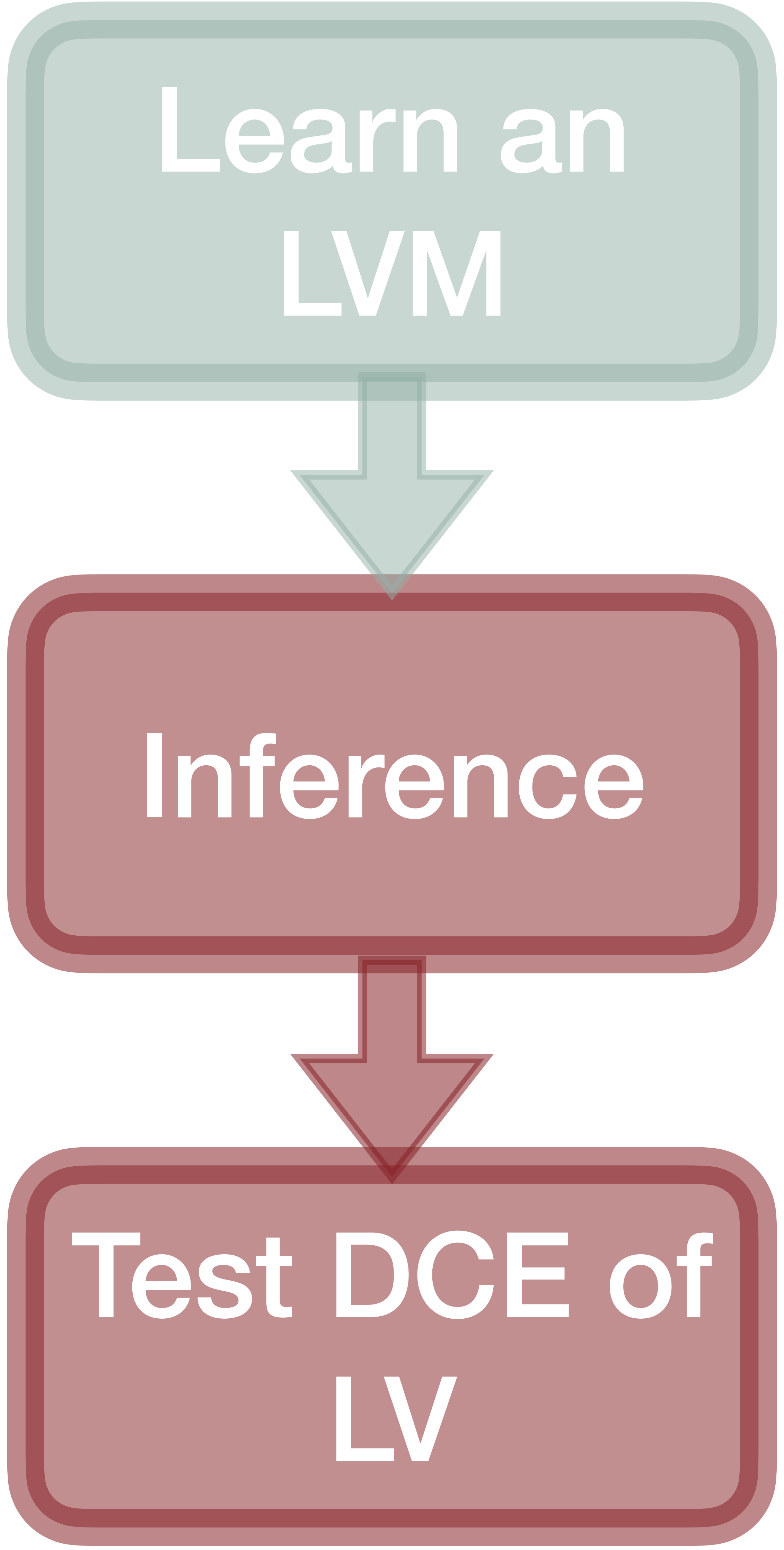
The exact LVM is not so important, but we desire online learning in multivariate time series

We use deep Gaussian process state-space models [Liu et al., TSP 2023]

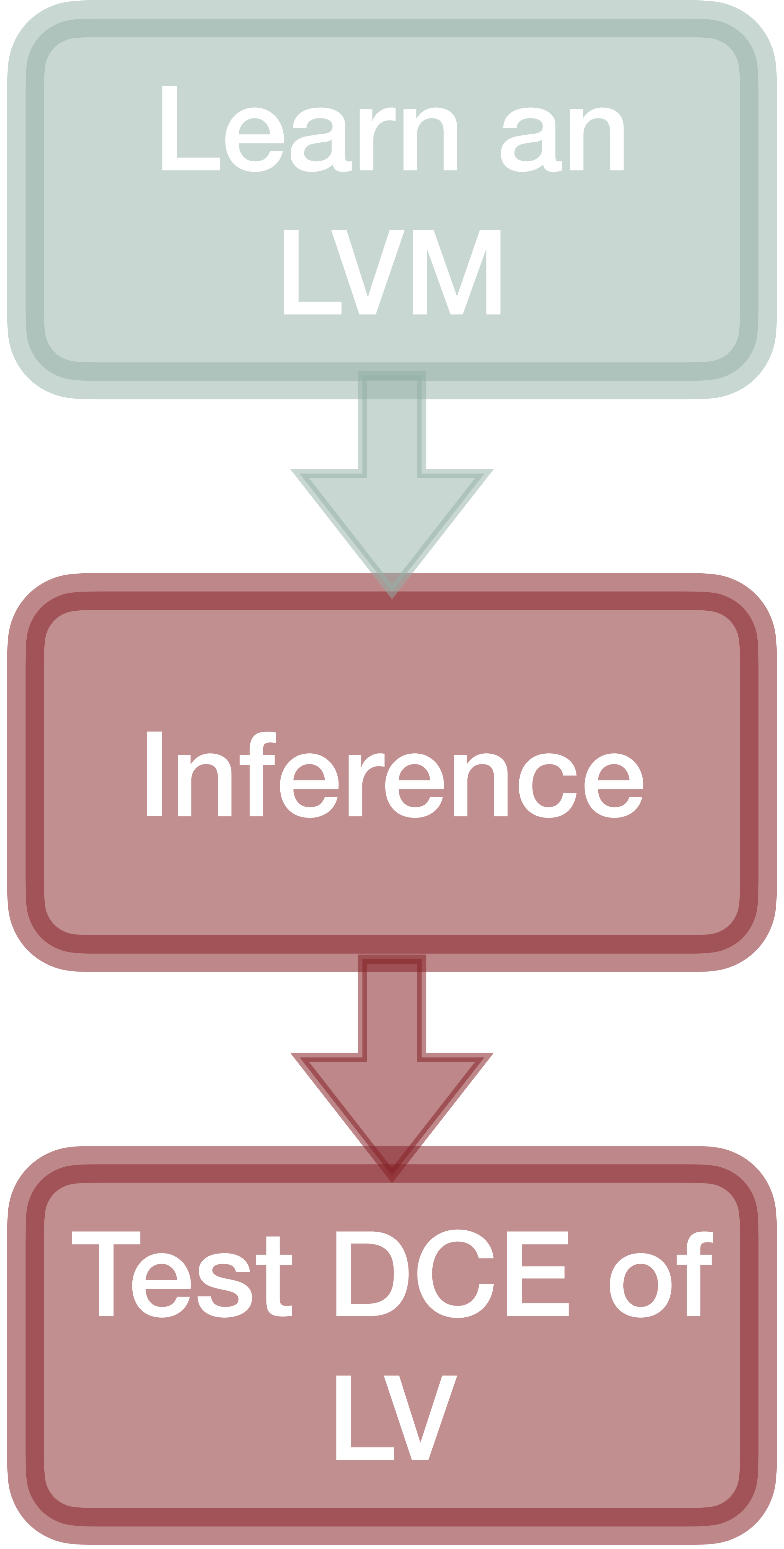
These write time-series auto regressively with Gaussian processes (GPs)

Using a specific GP approximation, they filter on the latent state and the GP parameters

Gaussian Processes for State Space Models

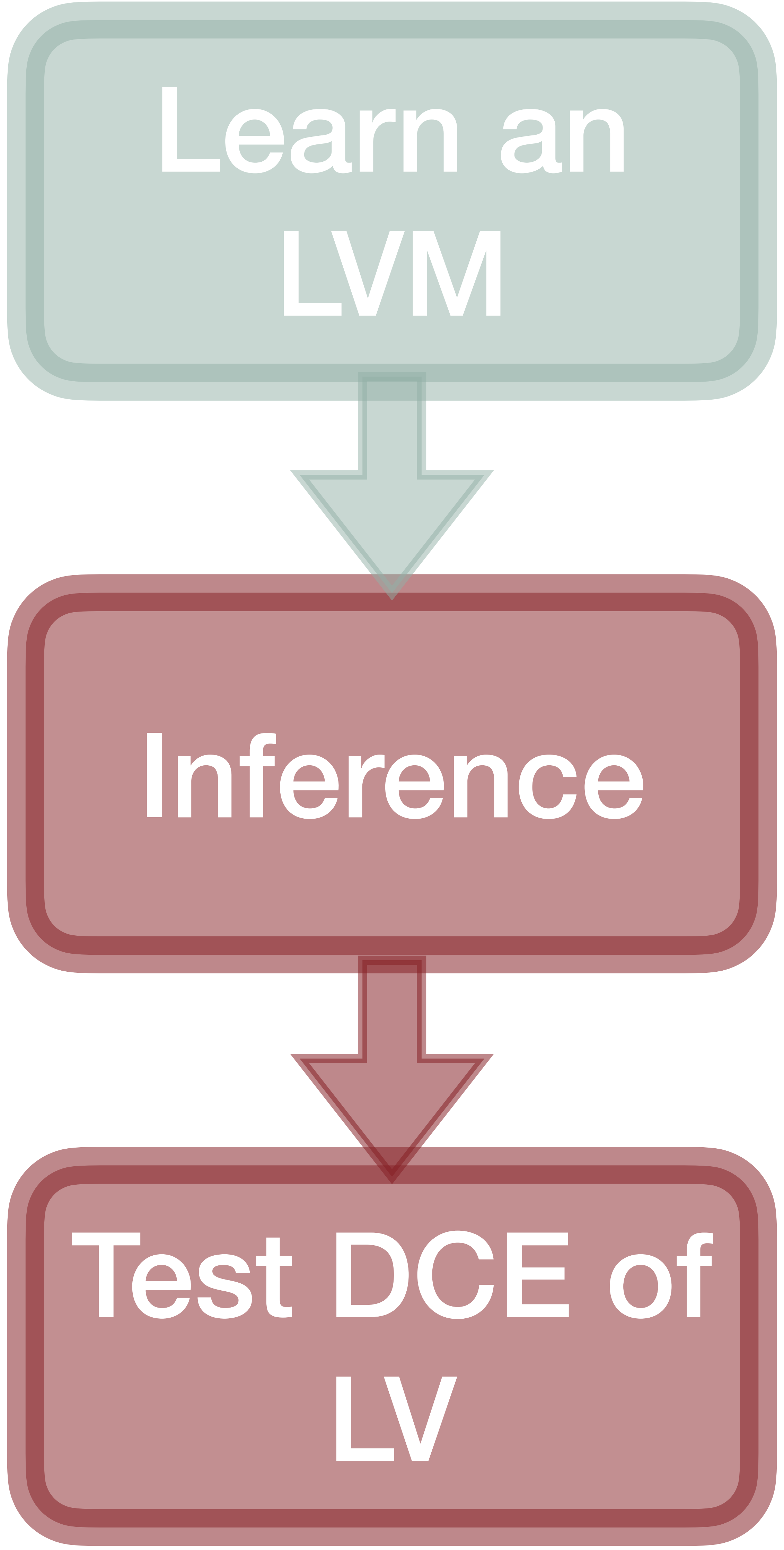


Gaussian Processes for State Space Models



Let $\mathbf{z}_t, \mathbf{x}_t, y_t$ be the time series of interest

Gaussian Processes for State Space Models

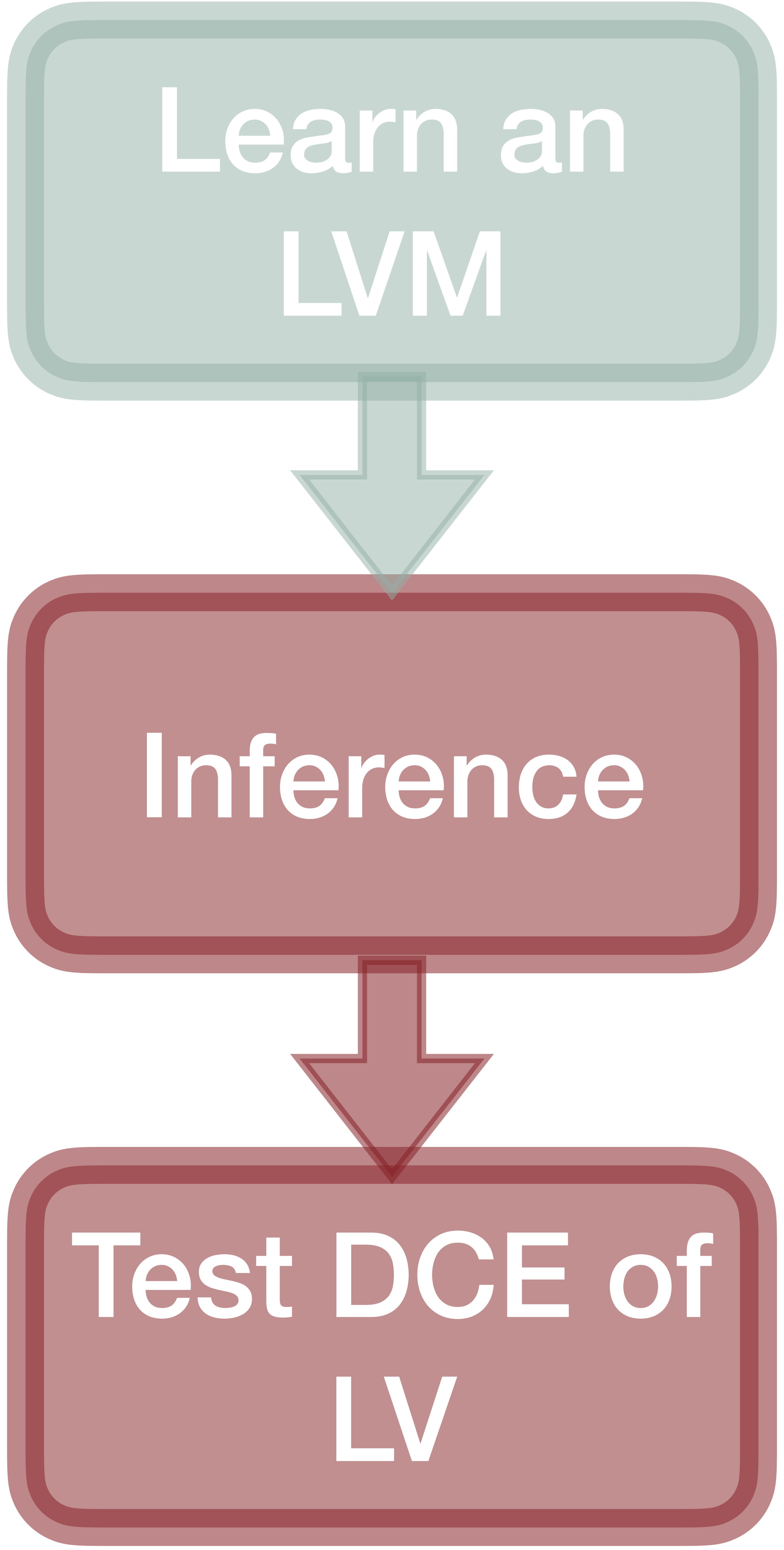


Let $\mathbf{z}_t, \mathbf{x}_t, y_t$ be the time series of interest

Then with unknown functions f, g, h and white Gaussian noise $\mathbf{u}_t, \mathbf{v}_t, e_t$:

$$\begin{aligned}
\mathbf{z}_t &= f(\mathbf{z}_{t-l_{zz}:t-1}, \mathbf{x}_{t-l_{zx}:t-1}, y_{t-l_{zy}:t-1}) + \mathbf{u}_t, \\
\mathbf{x}_t &= h(\mathbf{z}_{t-l_{xz}:t-1}, \mathbf{x}_{t-l_{xx}:t-1}, y_{t-l_{xy}:t-1}) + \mathbf{v}_t, \\
y_t &= g(\mathbf{z}_{t-l_{yz}:t-1}, \mathbf{x}_{t-l_{yx}:t-1}, y_{t-l_{yy}:t-1}) + e_t\,.
\end{aligned}$$

Gaussian Processes for State Space Models



Let $\mathbf{z}_t, \mathbf{x}_t, y_t$ be the time series of interest

Then with unknown functions f, g, h and white Gaussian noise $\mathbf{u}_t, \mathbf{v}_t, e_t$:

$$\begin{aligned}
\mathbf{z}_t &= f(\mathbf{z}_{t-l_{zz}:t-1}, \mathbf{x}_{t-l_{zx}:t-1}, y_{t-l_{zy}:t-1}) + \mathbf{u}_t, \\
\mathbf{x}_t &= h(\mathbf{z}_{t-l_{xz}:t-1}, \mathbf{x}_{t-l_{xx}:t-1}, y_{t-l_{xy}:t-1}) + \mathbf{v}_t, \\
y_t &= g(\mathbf{z}_{t-l_{yz}:t-1}, \mathbf{x}_{t-l_{yx}:t-1}, y_{t-l_{yy}:t-1}) + e_t\,.
\end{aligned}$$

Any LVM that learns f, g, h and $\mathbf{z}_t$ is good for us

Gaussian Processes for State Space Models

Learn an LVM

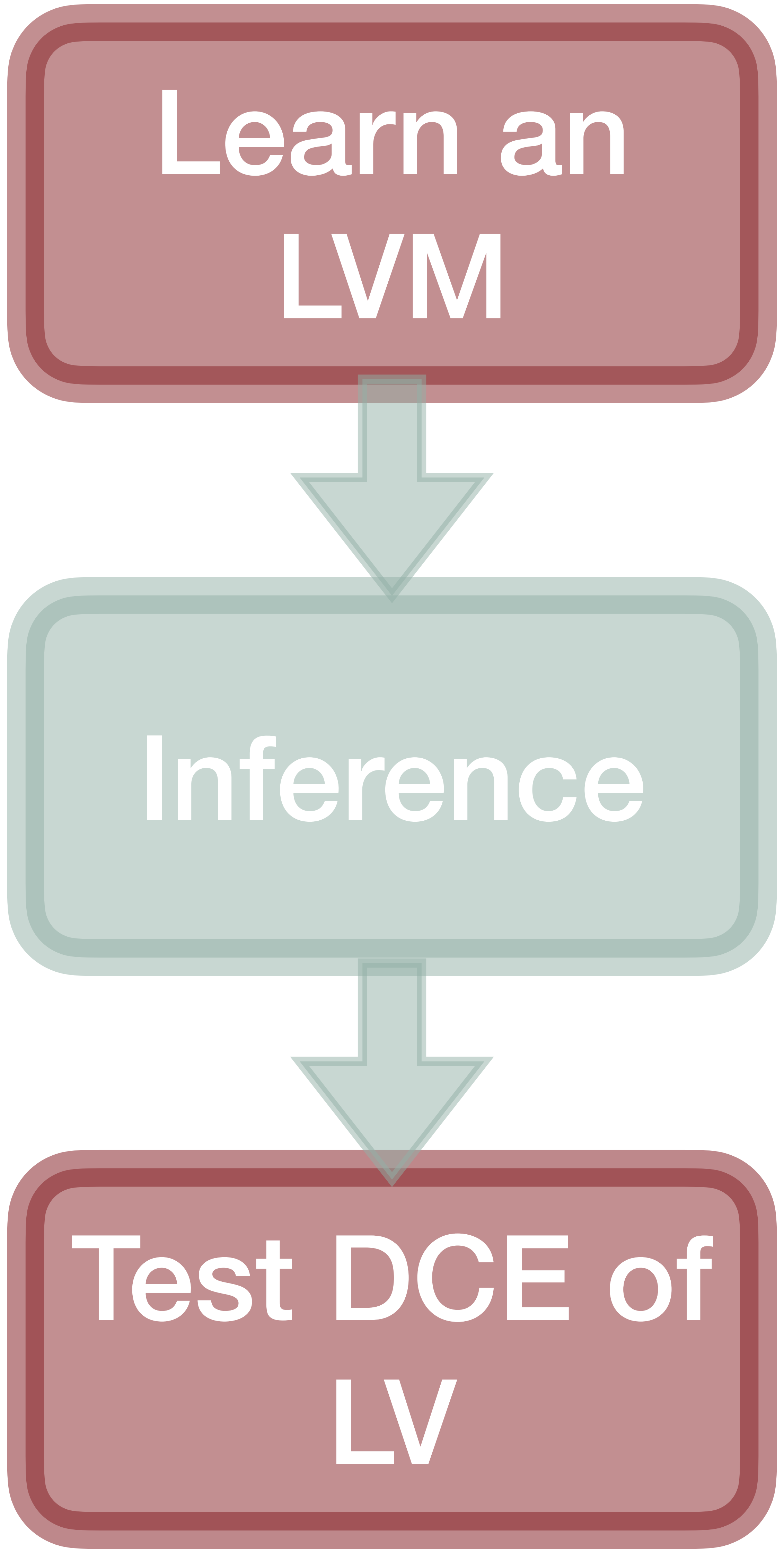
↓

Inference

↓

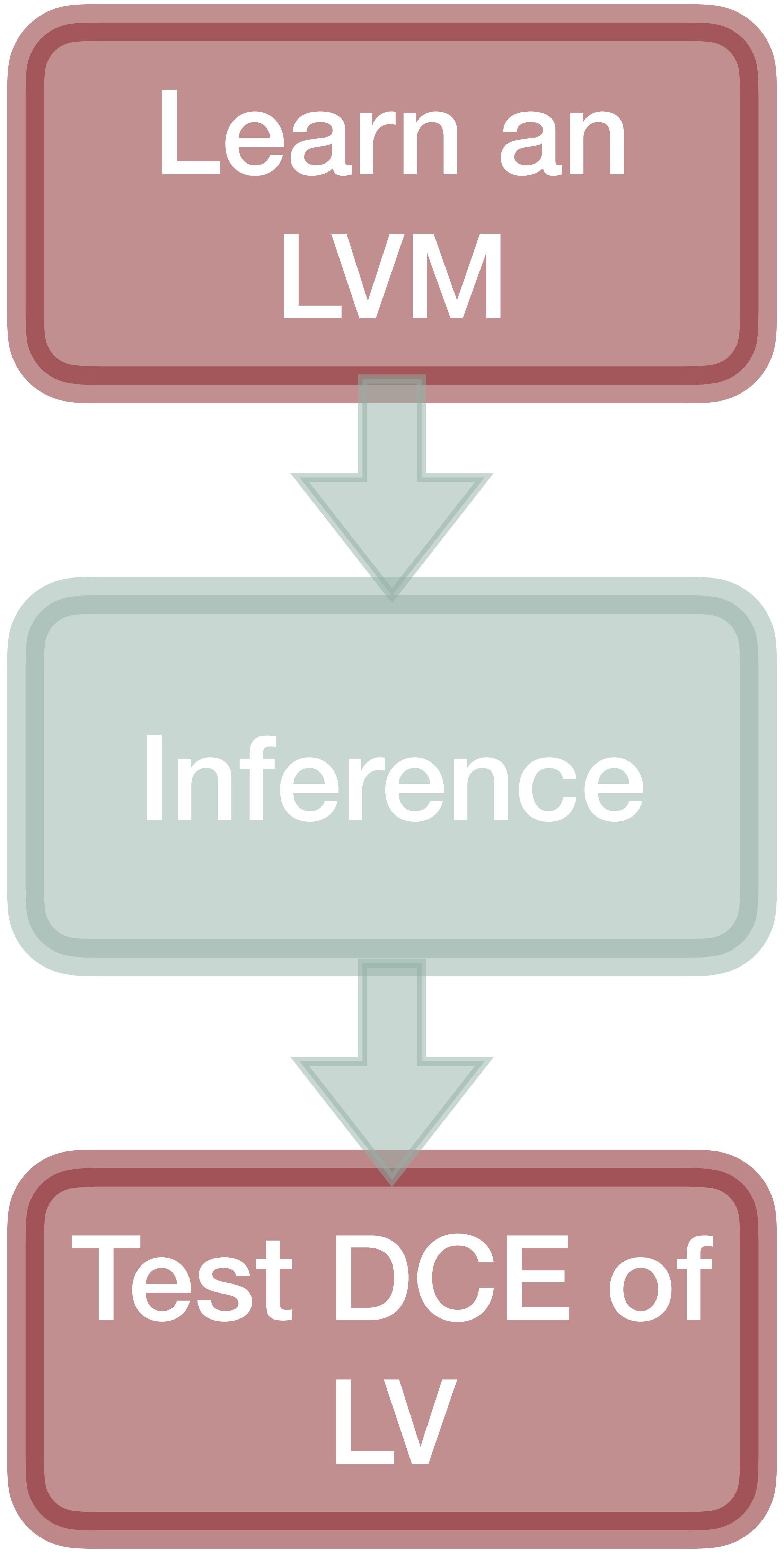
Test DCE of LV

Gaussian Processes for State Space Models



In the inference step, two goals:

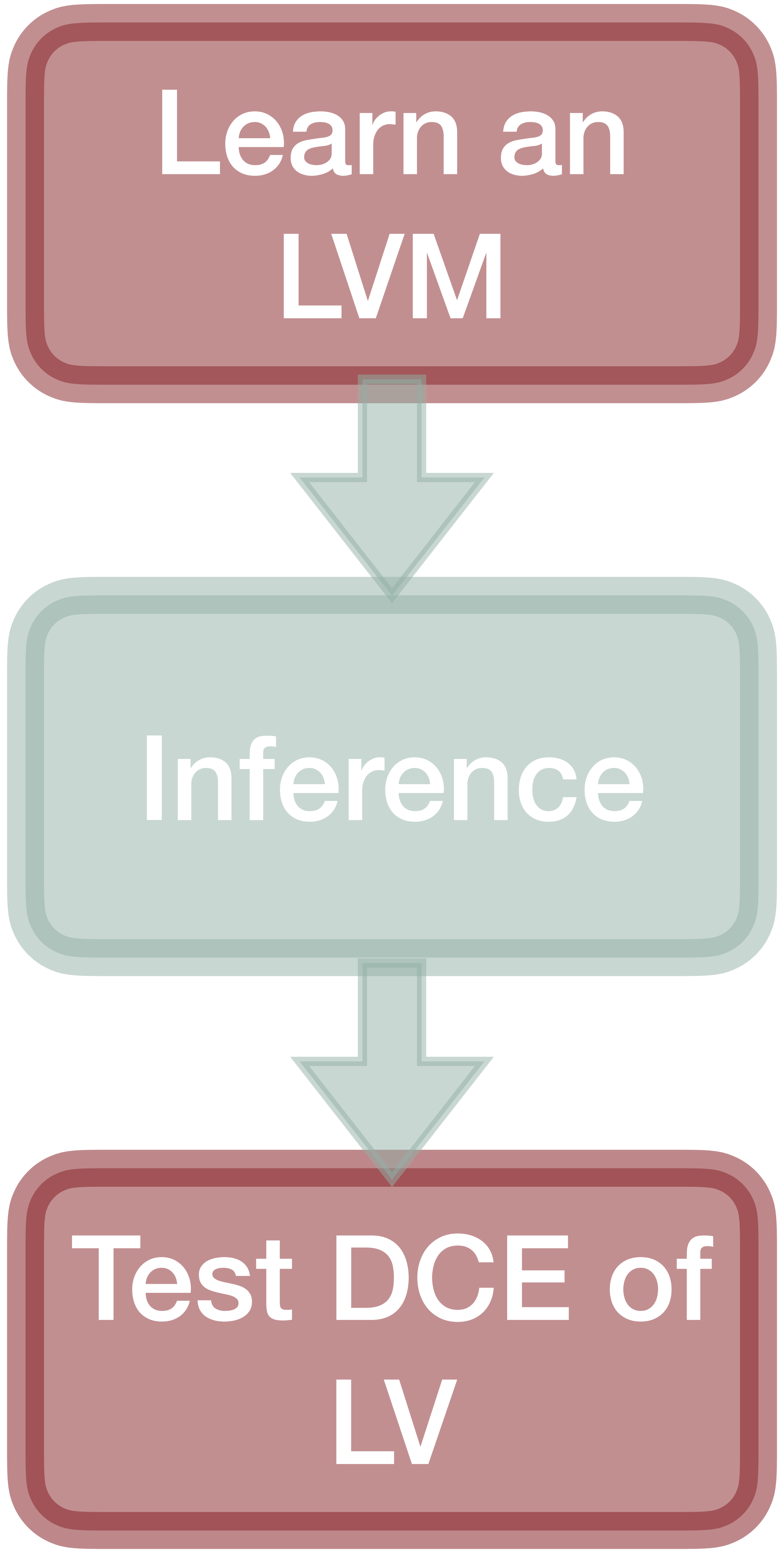
Gaussian Processes for State Space Models



In the inference step, two goals:

1. Infer $\mathbf{z}_t$ for $t = 1,\ldots,T$

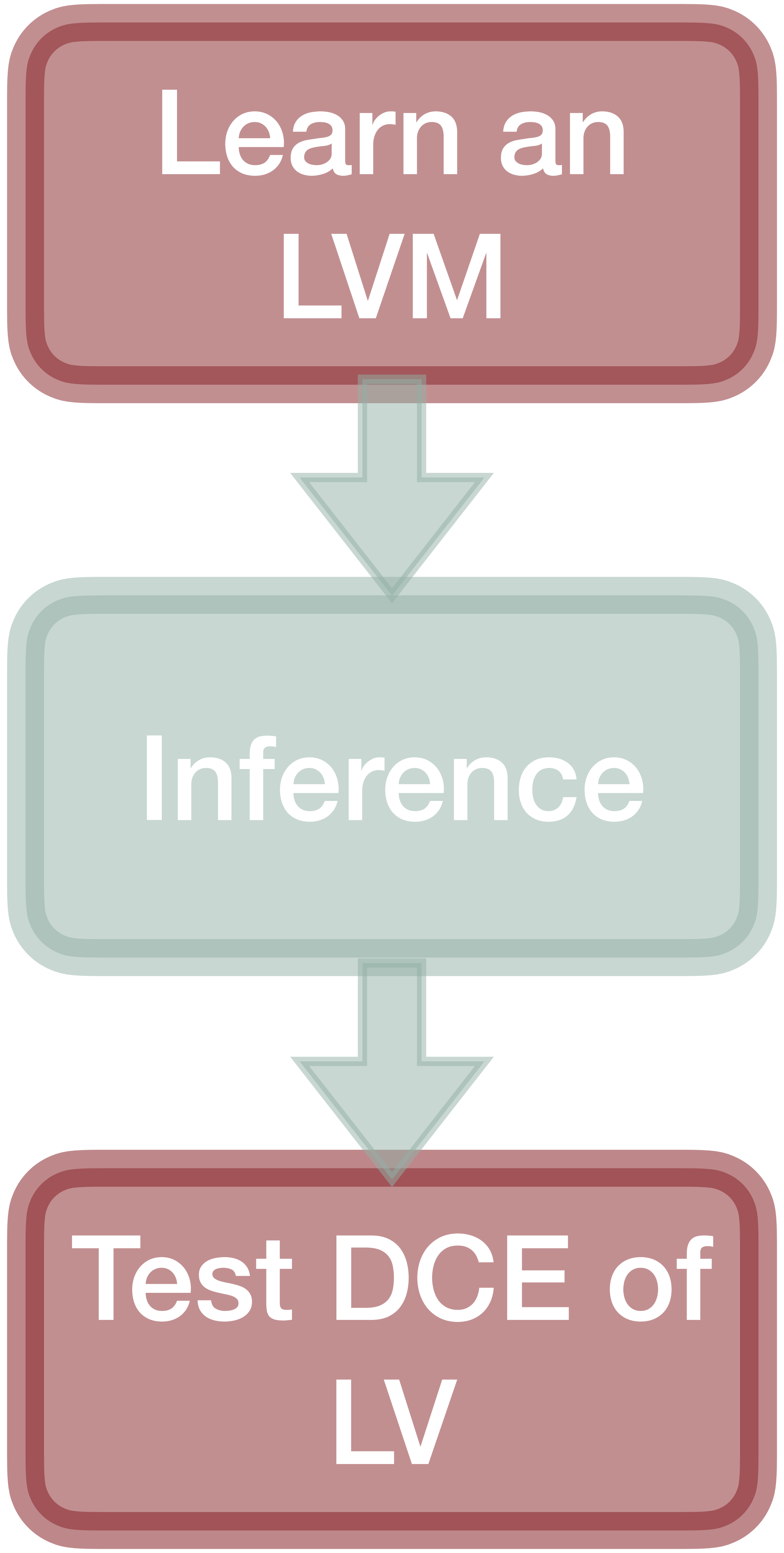
Gaussian Processes for State Space Models



In the inference step, two goals:

1. Infer $\mathbf{z}_t$ for $t = 1, \ldots, T$
2. Infer $\mathrm{DCE}_{z_{k,t-i} \to y_t}$ for $k = 1, \ldots, K$ and $i = 1, \ldots, l_{yz} - 1$

Gaussian Processes for State Space Models

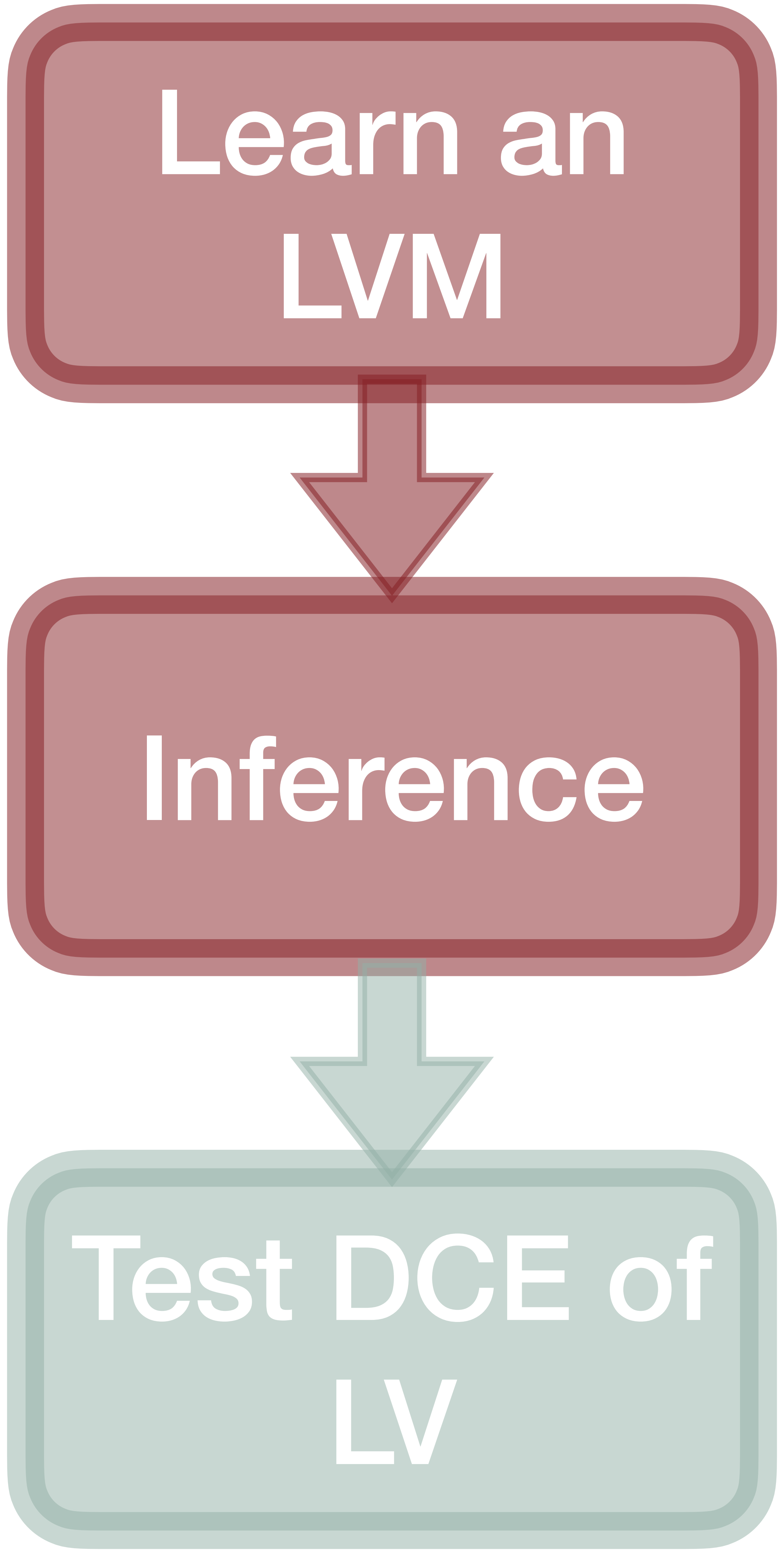


In the inference step, two goals:

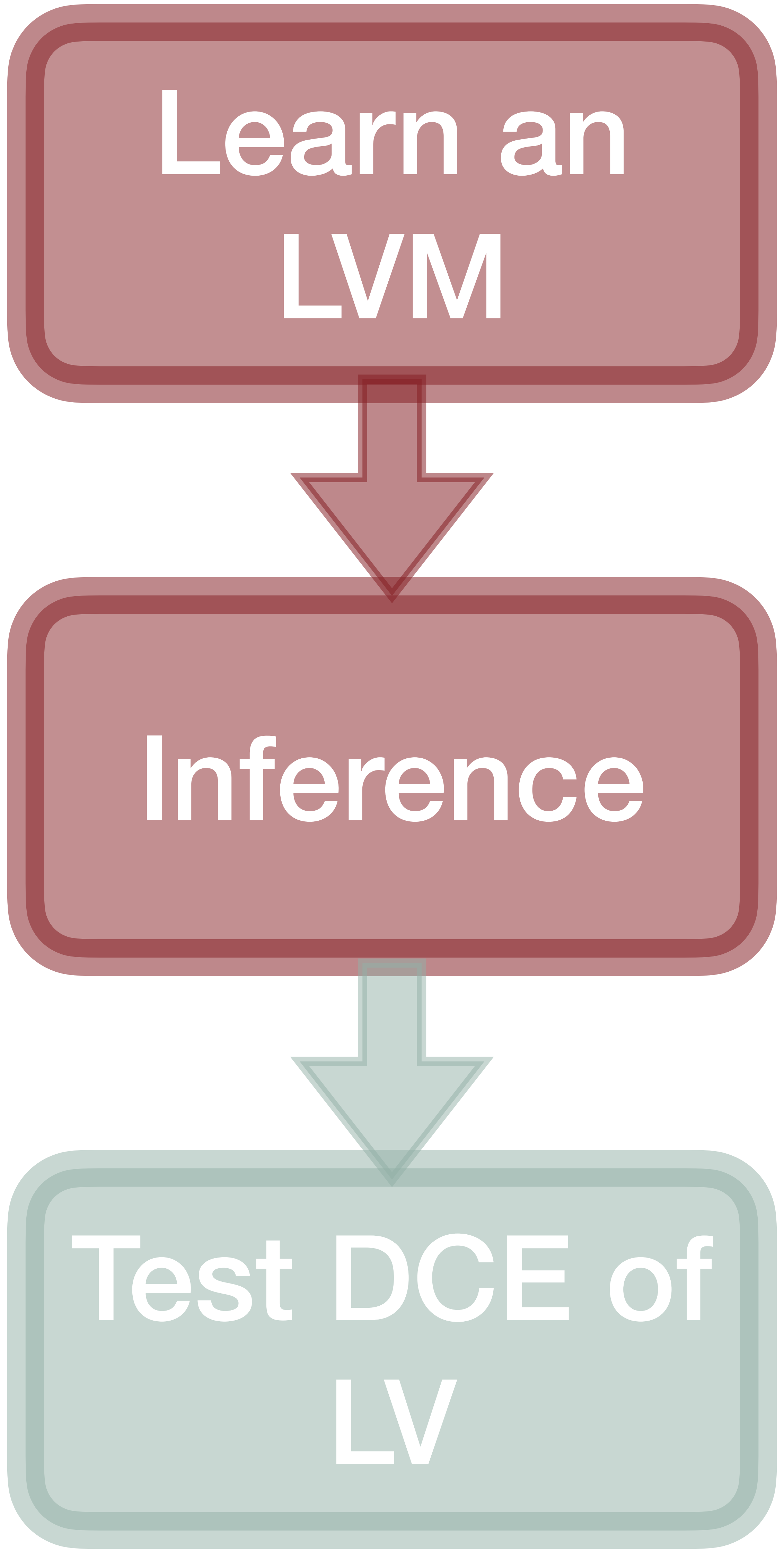
1. Infer $\mathbf{z}_t$ for $t = 1,\ldots,T$
2. Infer $\mathrm{DCE}_{z_{k,t-i}\rightarrow y_t}$ for $k = 1,\ldots,K$ and $i = 1,\ldots,l_{yz} - 1$

In our case, $\mathrm{DCE}_{z_{k,t-i}\rightarrow y_t}$ is available in a convenient closed form

Gaussian Processes for State Space Models

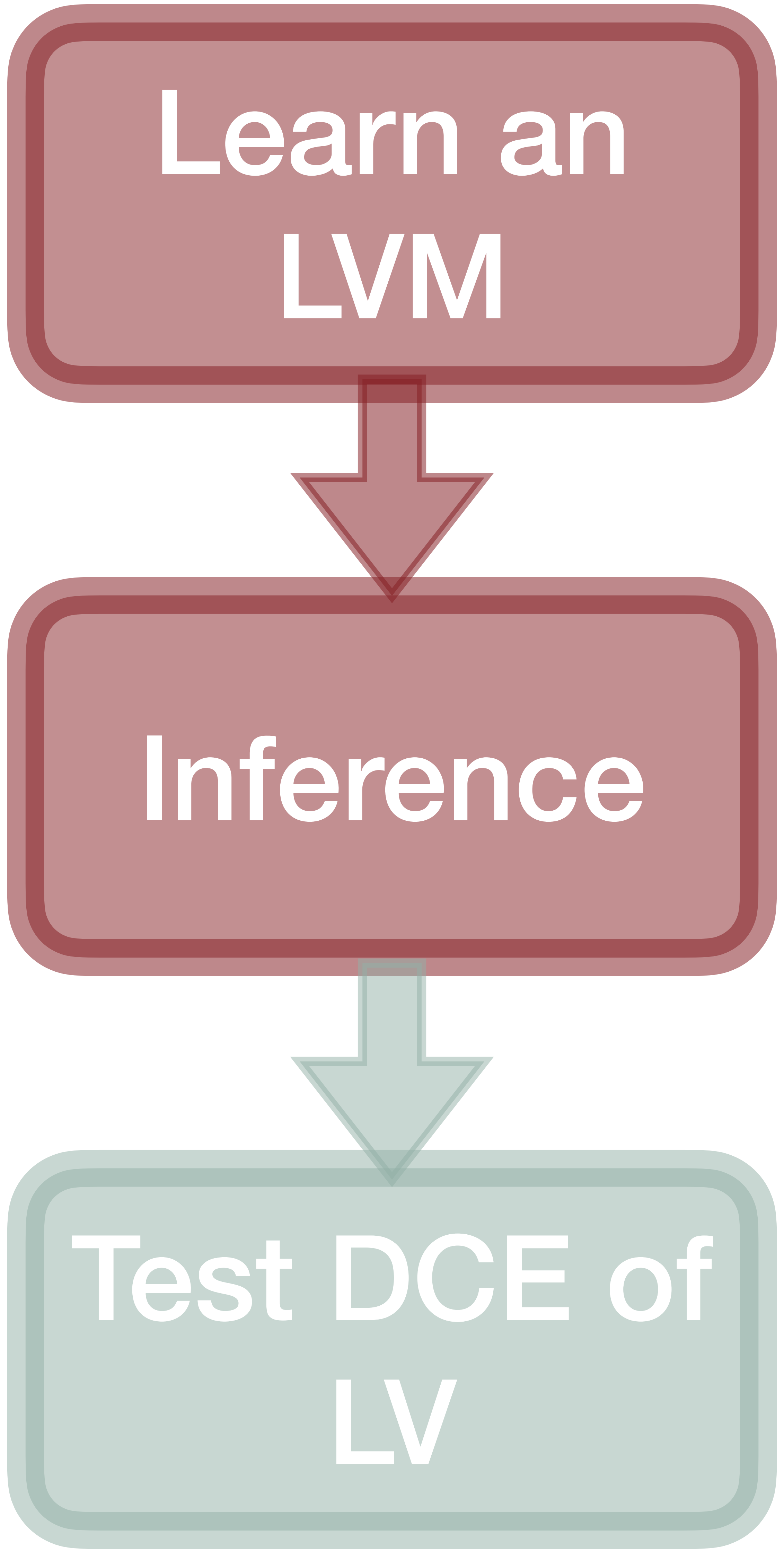


Gaussian Processes for State Space Models



For testing, the goal is to decide if $\mathrm{DCE}_{z_{k,t-i} \to y_t} = 0$ (no influence) or $\mathrm{DCE}_{z_{k,t-i} \to y_t} \neq 0$ (influence)

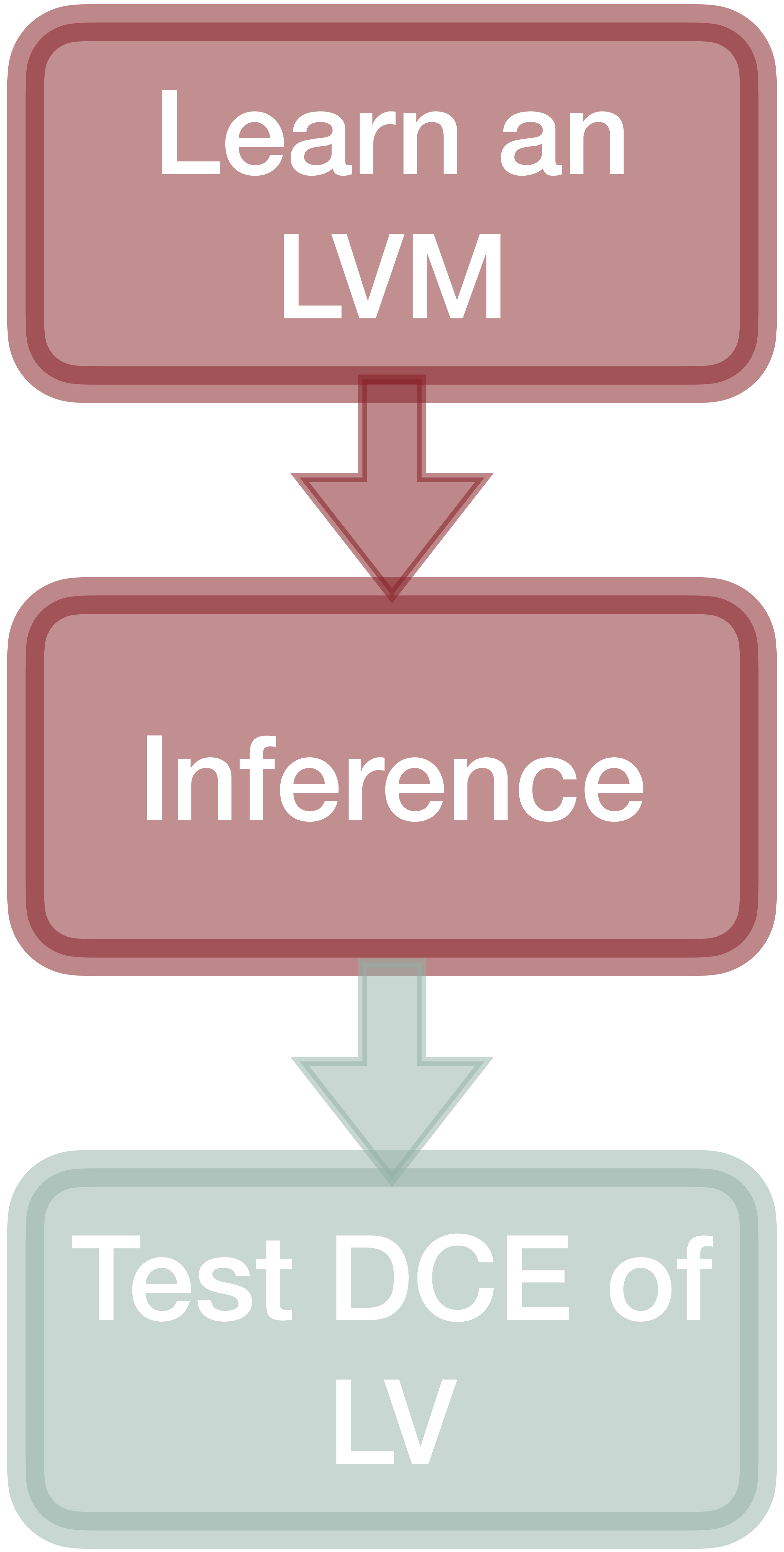
Gaussian Processes for State Space Models



For testing, the goal is to decide if $\mathrm{DCE}_{z_{k,t-i} \to y_t} = 0$ (no influence) or $\mathrm{DCE}_{z_{k,t-i} \to y_t} \neq 0$ (influence)

The best way to do this is unresolved

Gaussian Processes for State Space Models

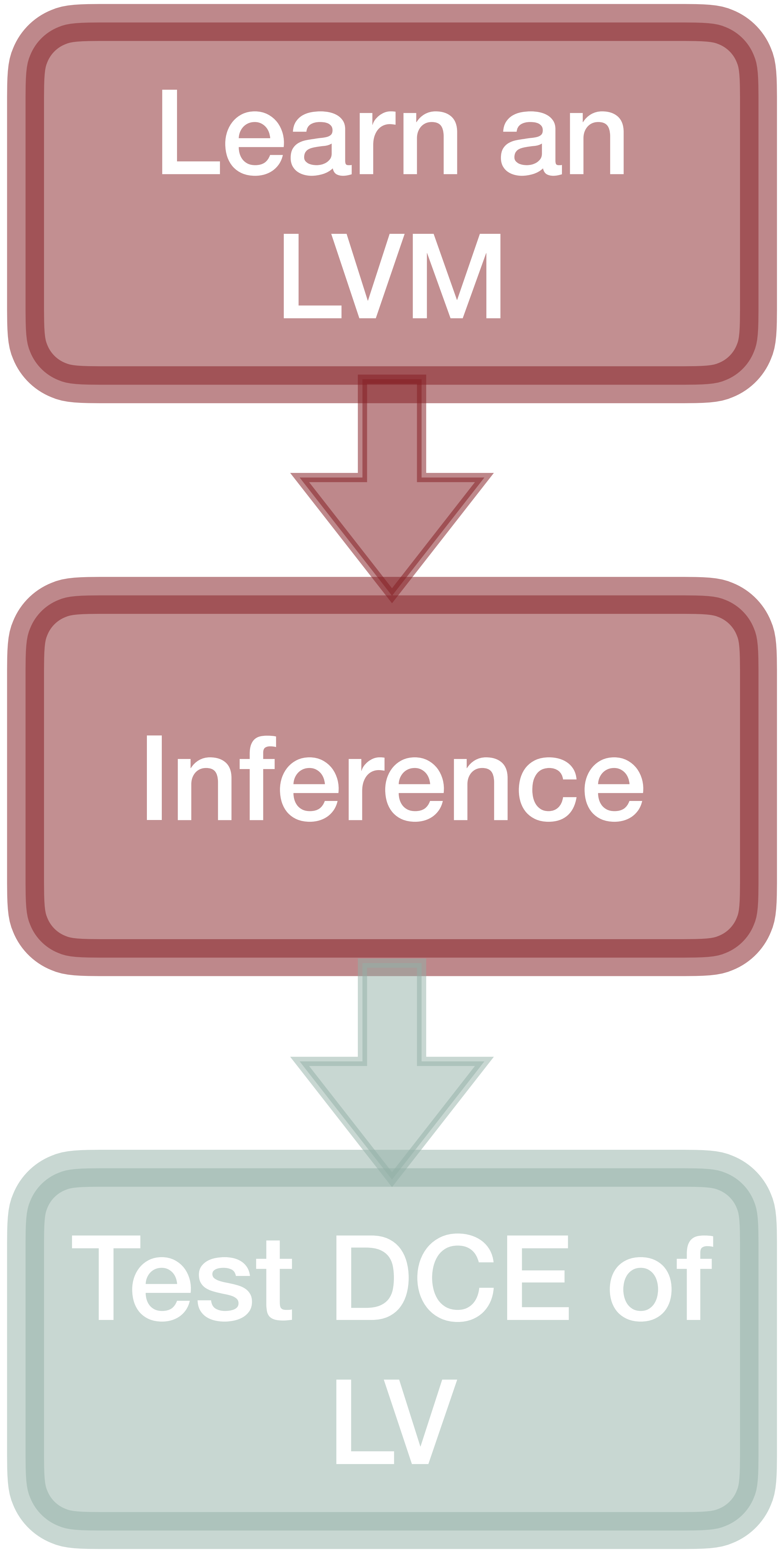


For testing, the goal is to decide if $\mathrm{DCE}_{z_{k,t-i} \to y_t} = 0$ (no influence) or $\mathrm{DCE}_{z_{k,t-i} \to y_t} \neq 0$ (influence)

The best way to do this is unresolved

Two ideas:

Gaussian Processes for State Space Models



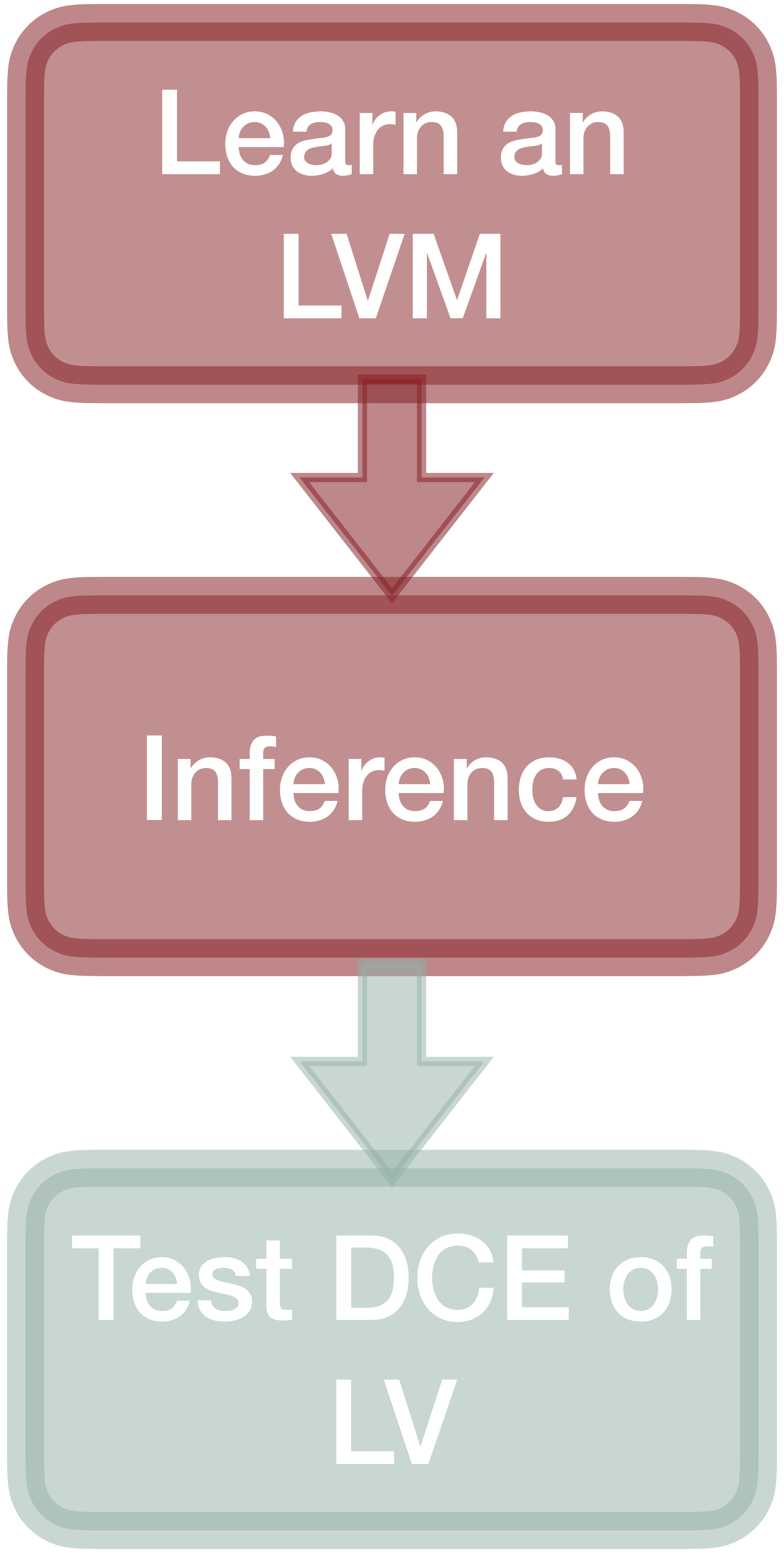
For testing, the goal is to decide if $\mathrm{DCE}_{z_{k,t-i} \to y_t} = 0$ (no influence) or $\mathrm{DCE}_{z_{k,t-i} \to y_t} \neq 0$ (influence)

The best way to do this is unresolved

Two ideas:

1. Test if the $p\,\%$ credible interval of $\mathrm{DCE}_{z_{k,t-i} \to y_t}$ contains 0

Gaussian Processes for State Space Models



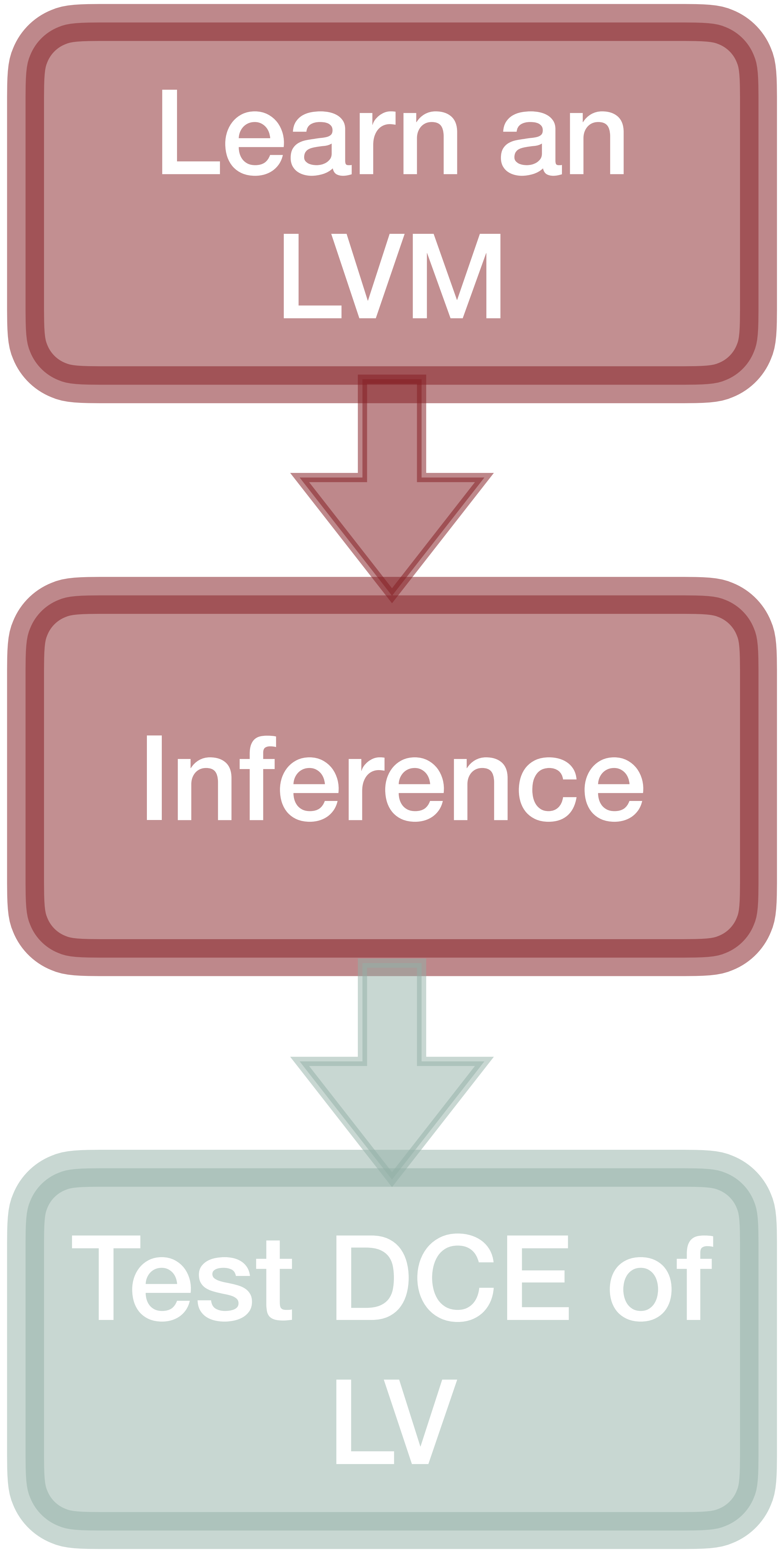
For testing, the goal is to decide if $\mathrm{DCE}_{z_{k,t-i} \to y_t} = 0$ (no influence) or $\mathrm{DCE}_{z_{k,t-i} \to y_t} \neq 0$ (influence)

The best way to do this is unresolved

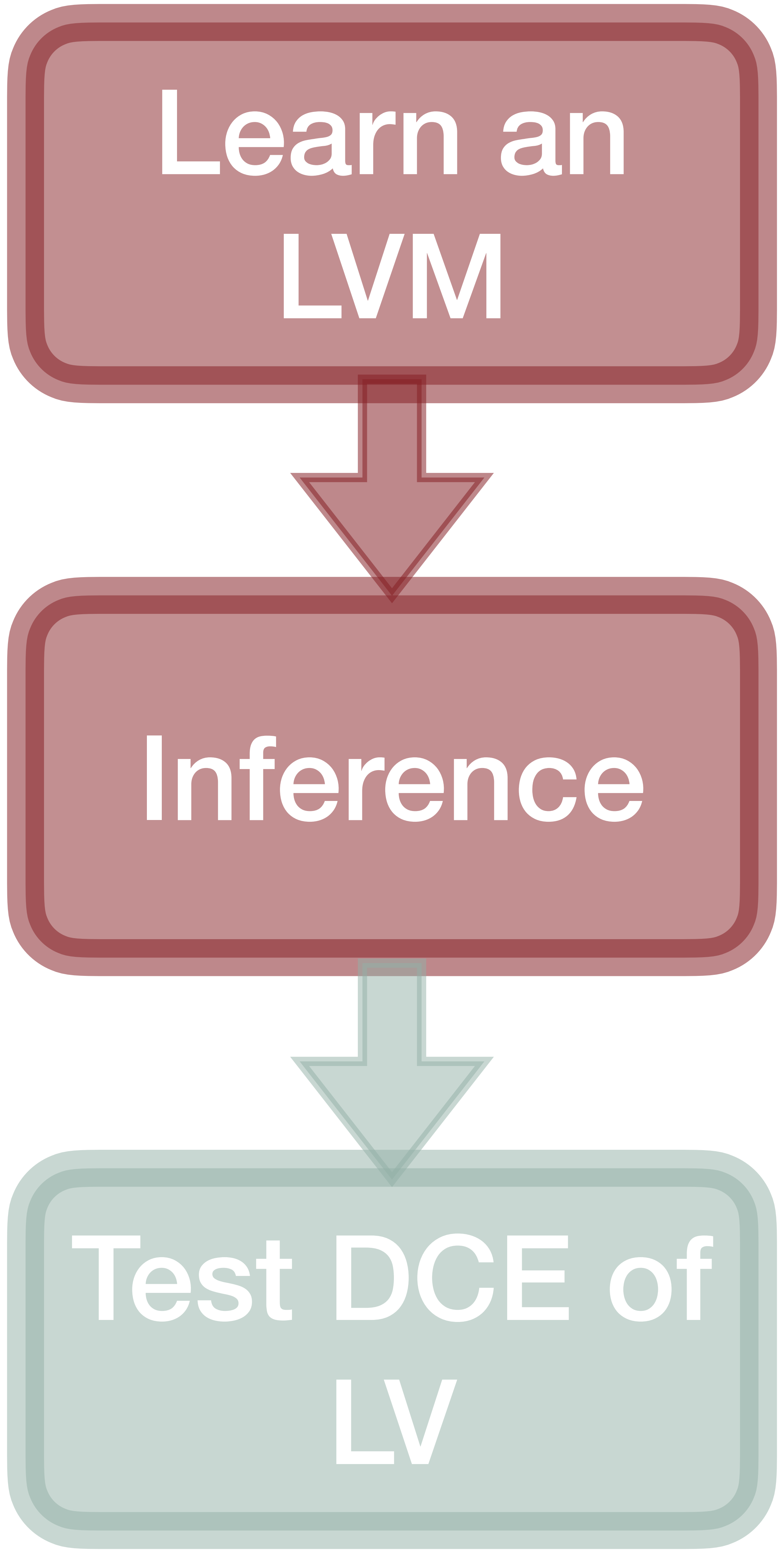
Two ideas:

1. Test if the $p\,\%$ credible interval of $\mathrm{DCE}_{z_{k,t-i} \to y_t}$ contains 0
2. Test if $\Pr\left(\mathrm{DCE}_{z_{k,t-i} \to y_t} \in (-\epsilon, \epsilon)\right)$ exceeds some threshold

Using the DCE is Justified

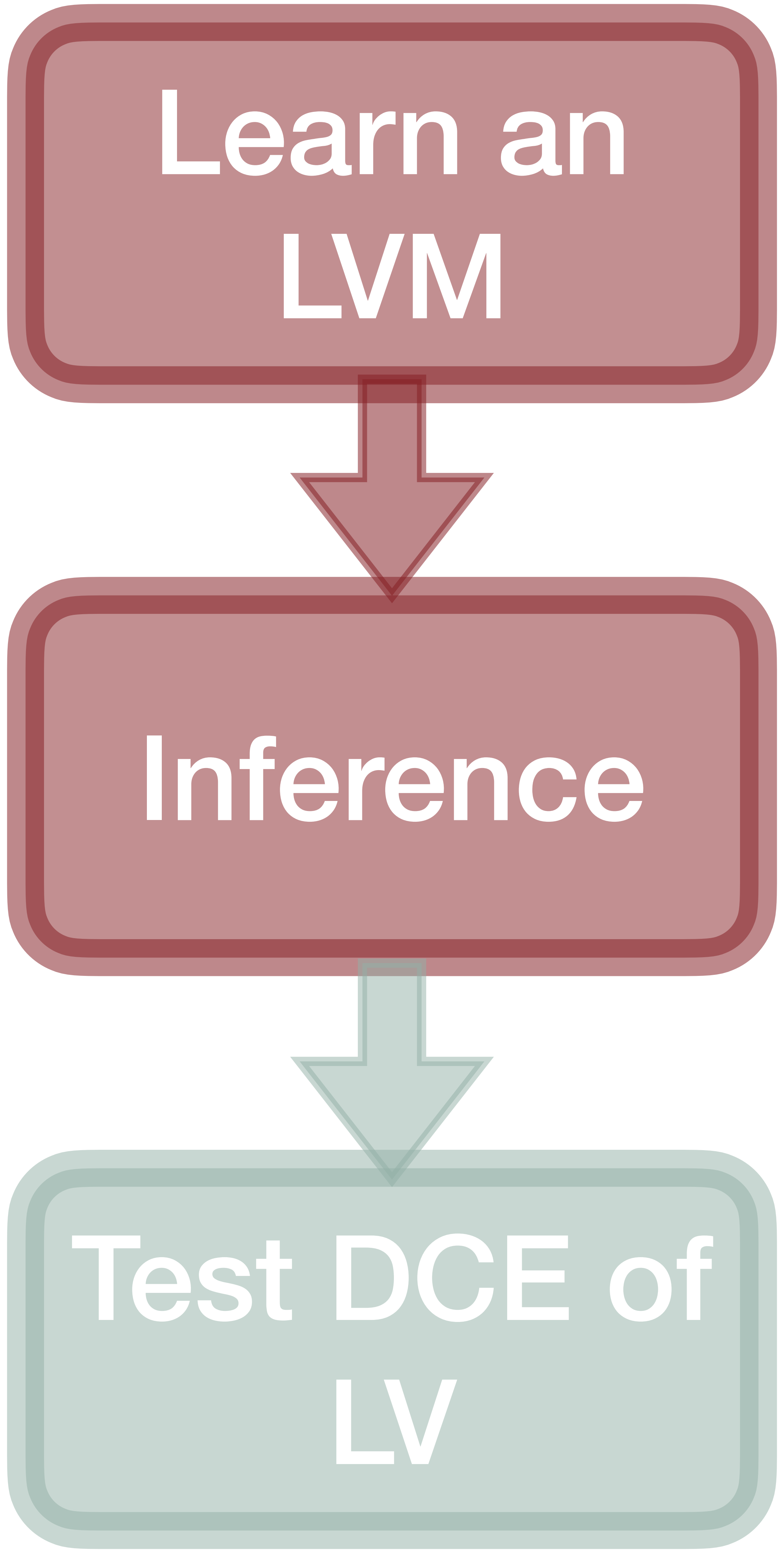


Using the DCE is Justified



Is the causal strength of a latent variable well-defined?

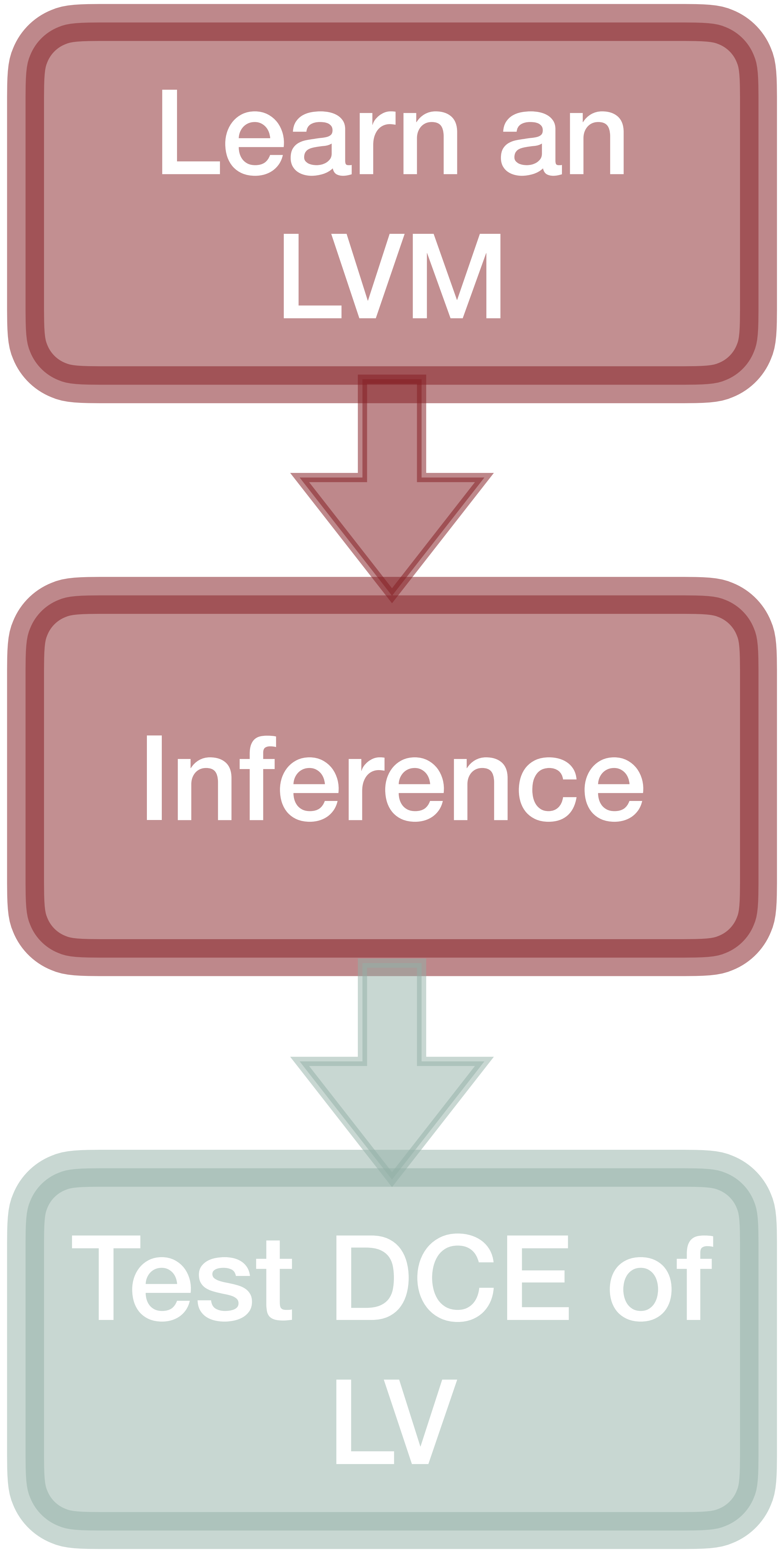
Using the DCE is Justified



Is the causal strength of a latent variable well-defined?

No…

Using the DCE is Justified

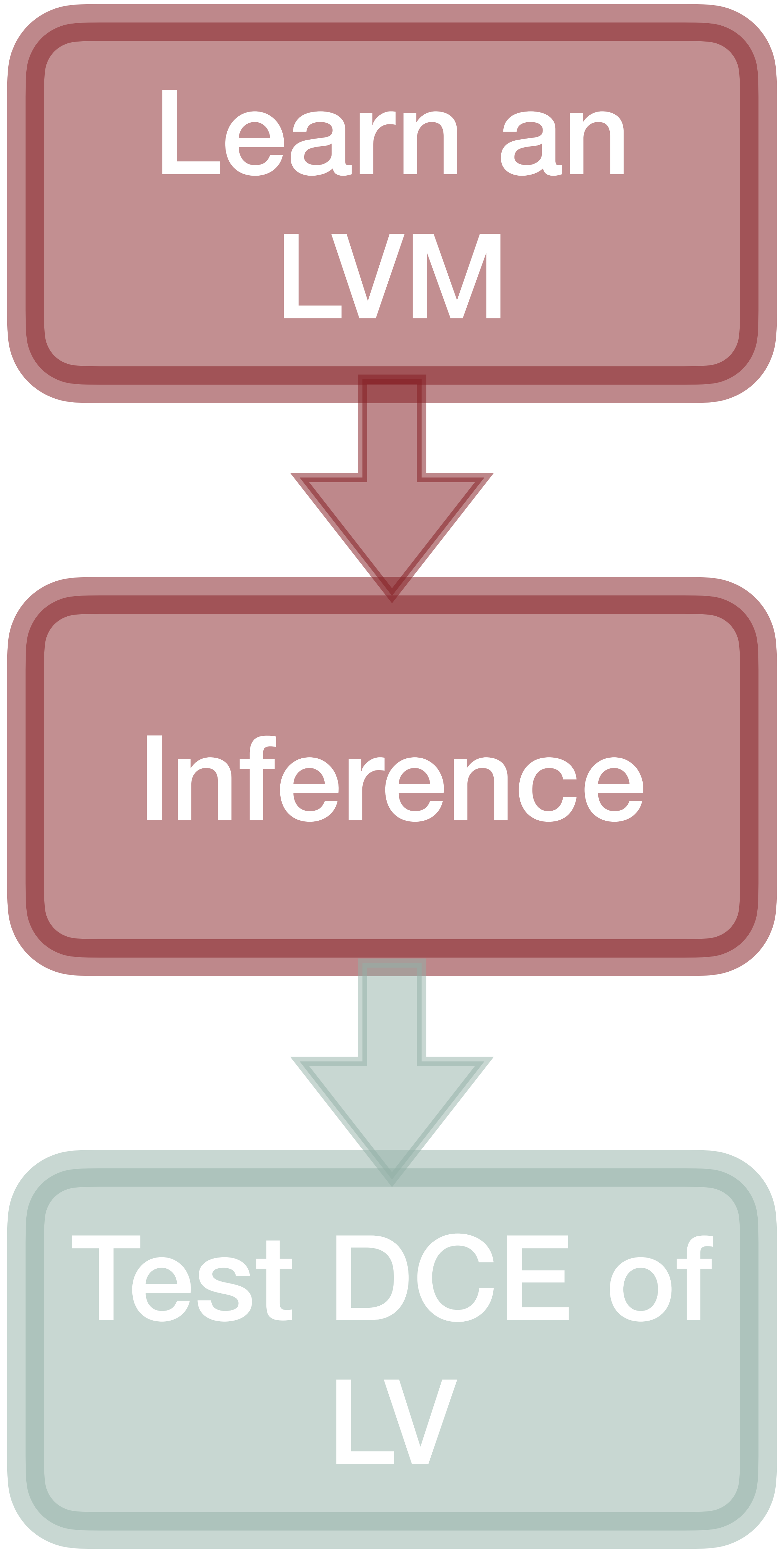


Is the causal strength of a latent variable well-defined?

No…

But for scalars, zeroness of the DCE is invariant to a change in coordinates:

Using the DCE is Justified



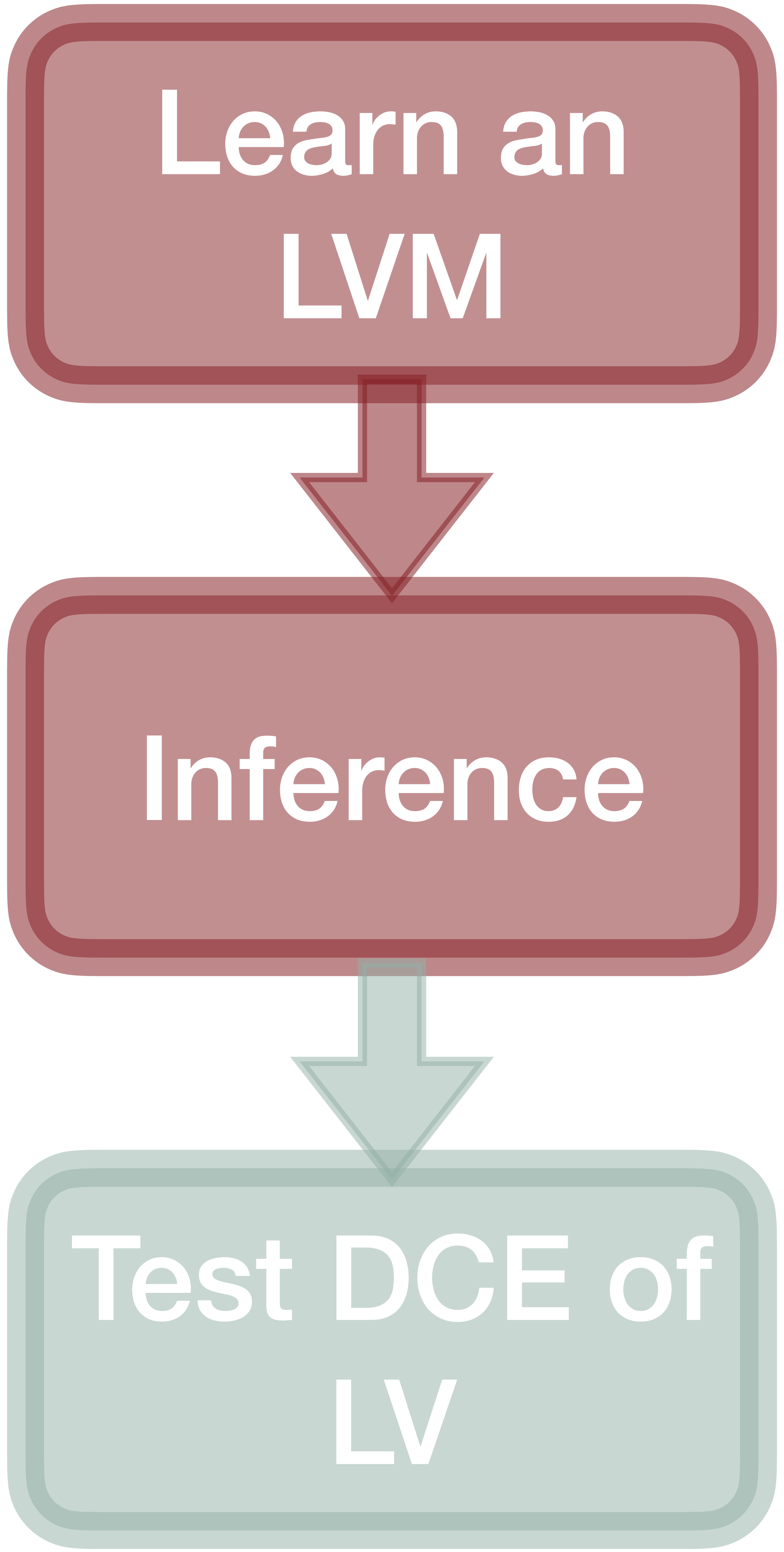
Is the causal strength of a latent variable well-defined?

No…

But for scalars, zeroness of the DCE is invariant to a change in coordinates:

$$\frac{\partial y_t}{\partial z_{t-i}} = \frac{\partial y_t}{\partial z'_{t-i}} \frac{\partial z'_{t-i}}{\partial z_{t-i}} = 0 \times \frac{\partial z'_{t-i}}{\partial z_{t-i}} = 0.$$

Using the DCE is Justified



Is the causal strength of a latent variable well-defined?

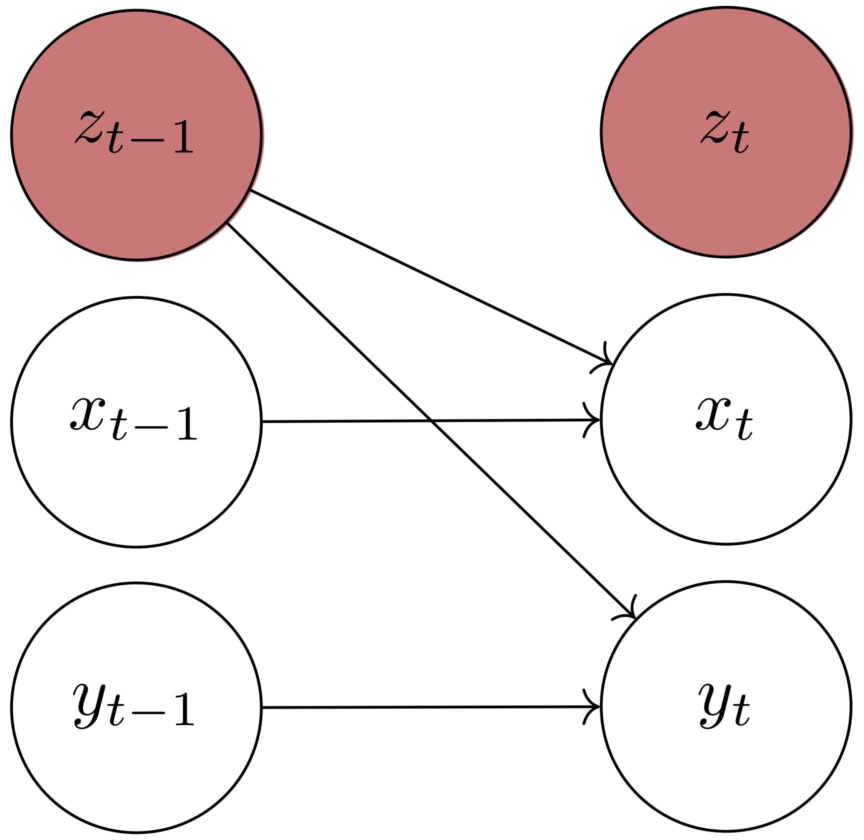
No…

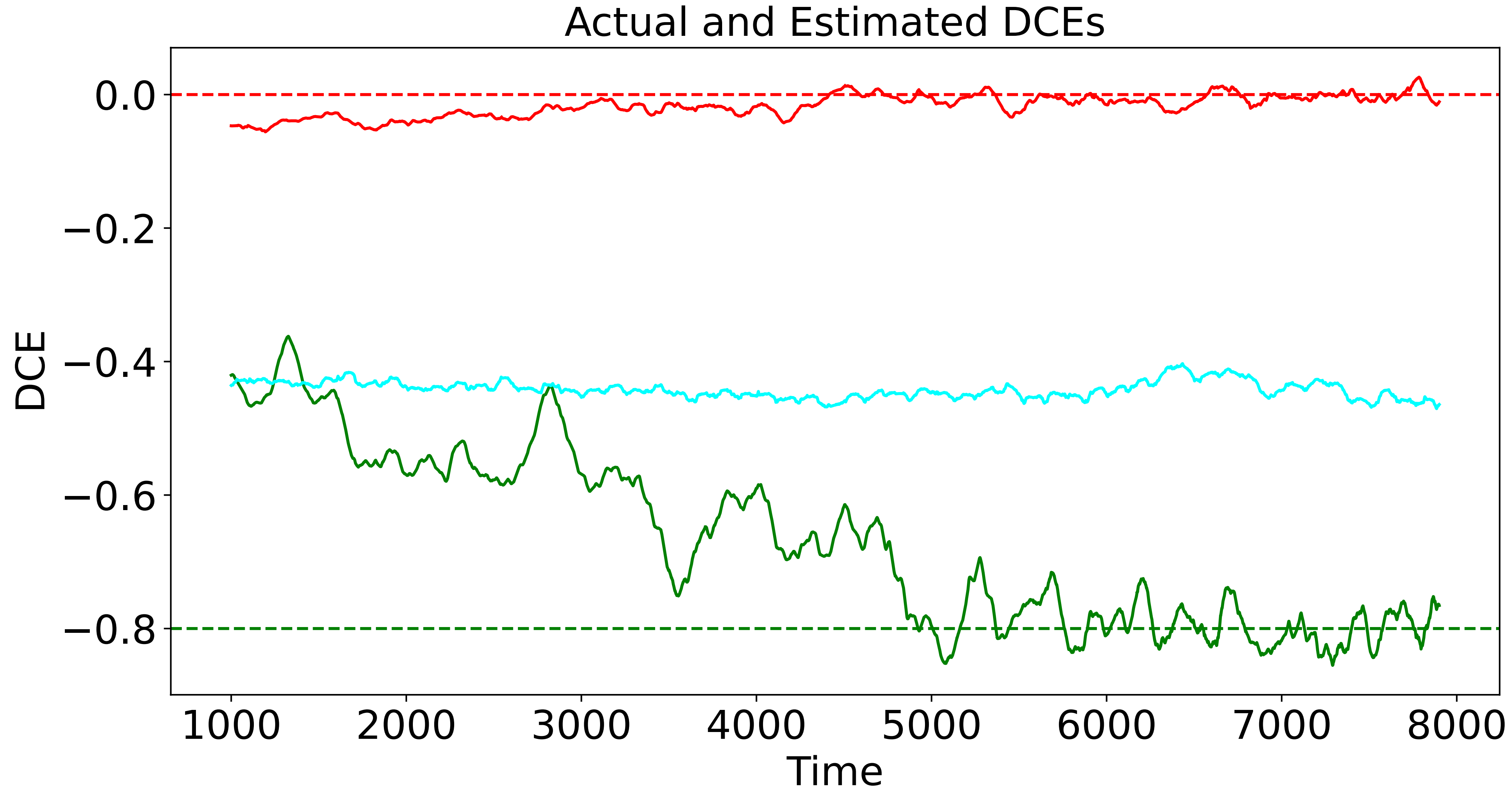
But for scalars, zeroness of the DCE is invariant to a change in coordinates:

$$\frac{\partial y_t}{\partial z_{t-i}} = \frac{\partial y_t}{\partial z'_{t-i}} \frac{\partial z'_{t-i}}{\partial z_{t-i}} = 0 \times \frac{\partial z'_{t-i}}{\partial z_{t-i}} = 0.$$

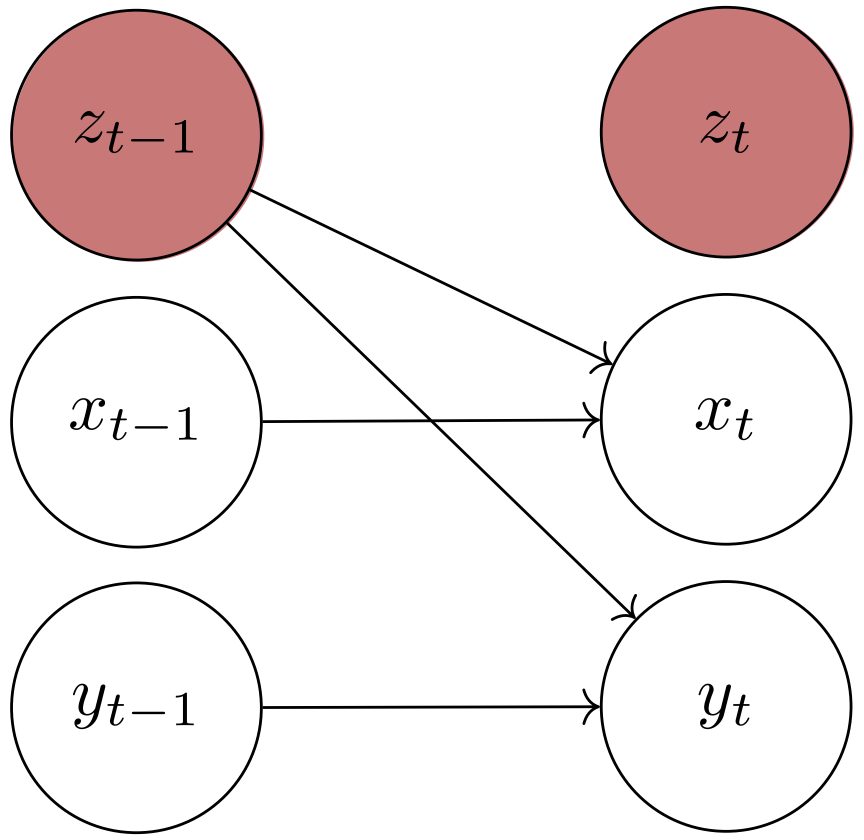
Therefore, testing for zero-ness is well-defined

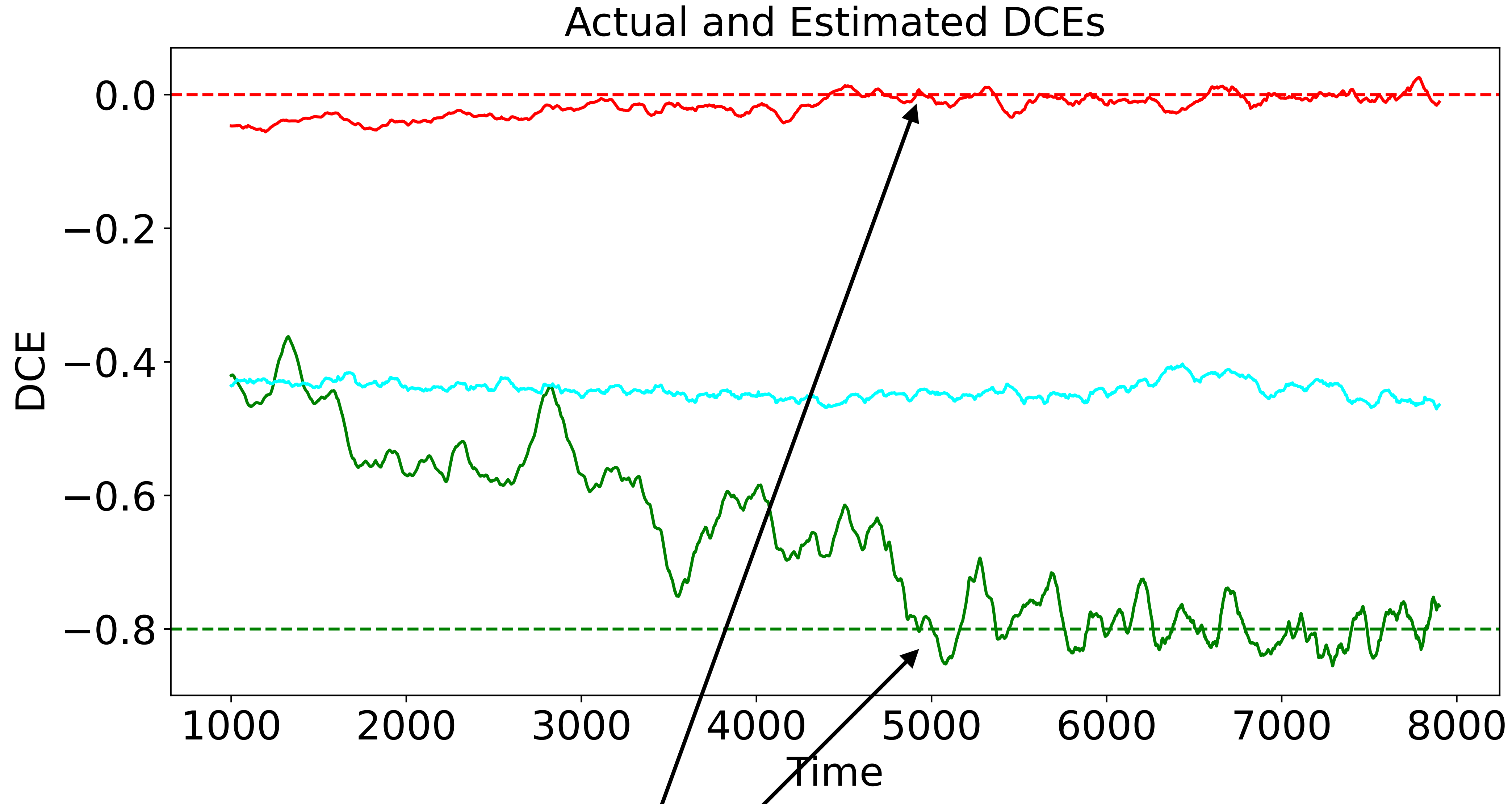
Our Method Can Detect Confounders In the Static Case





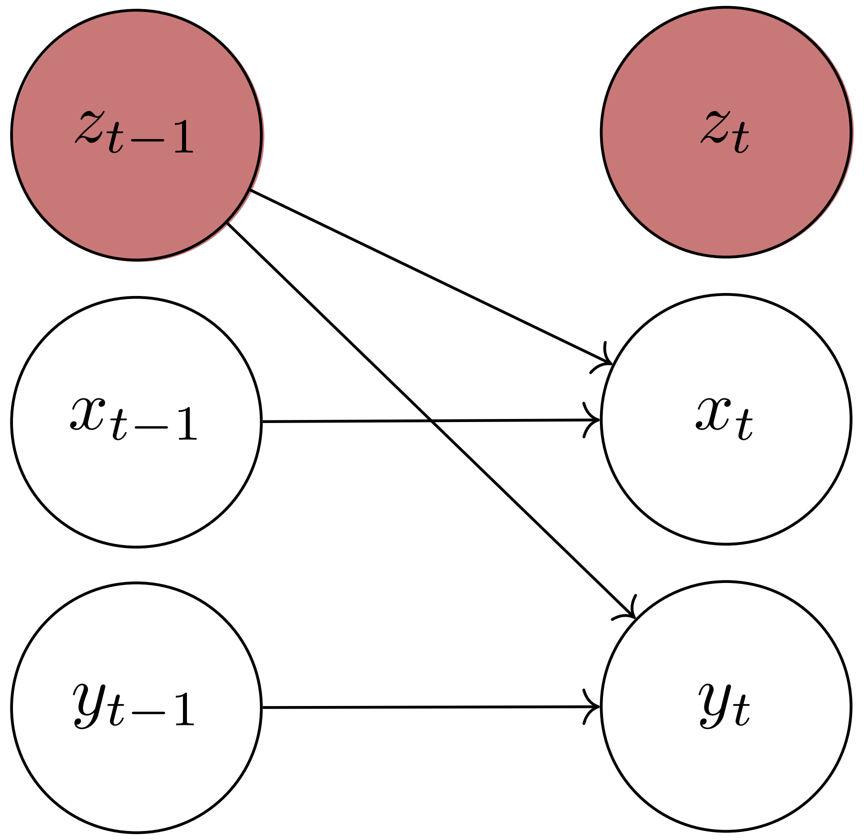
Our Method Can Detect Confounders In the Static Case

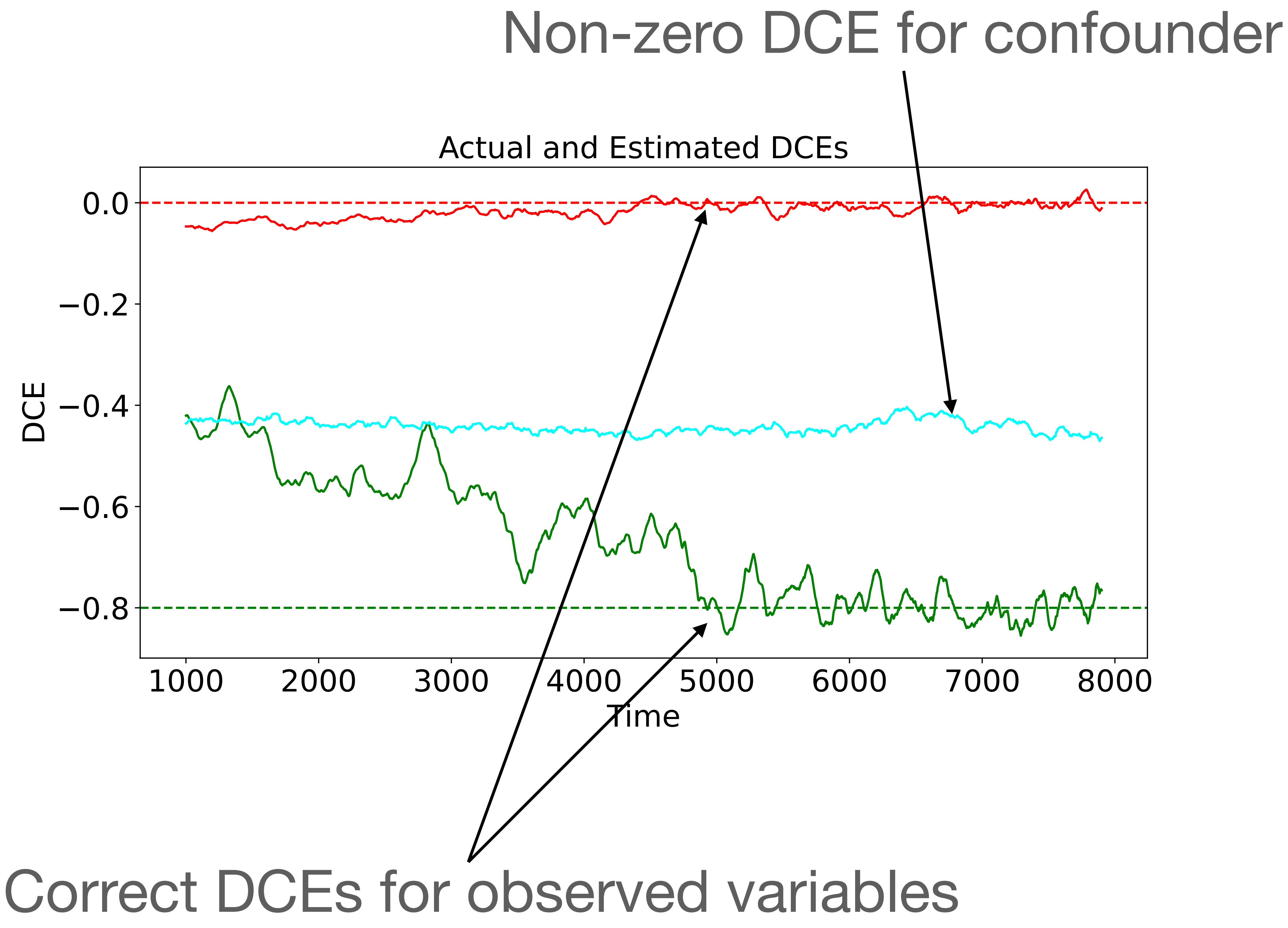




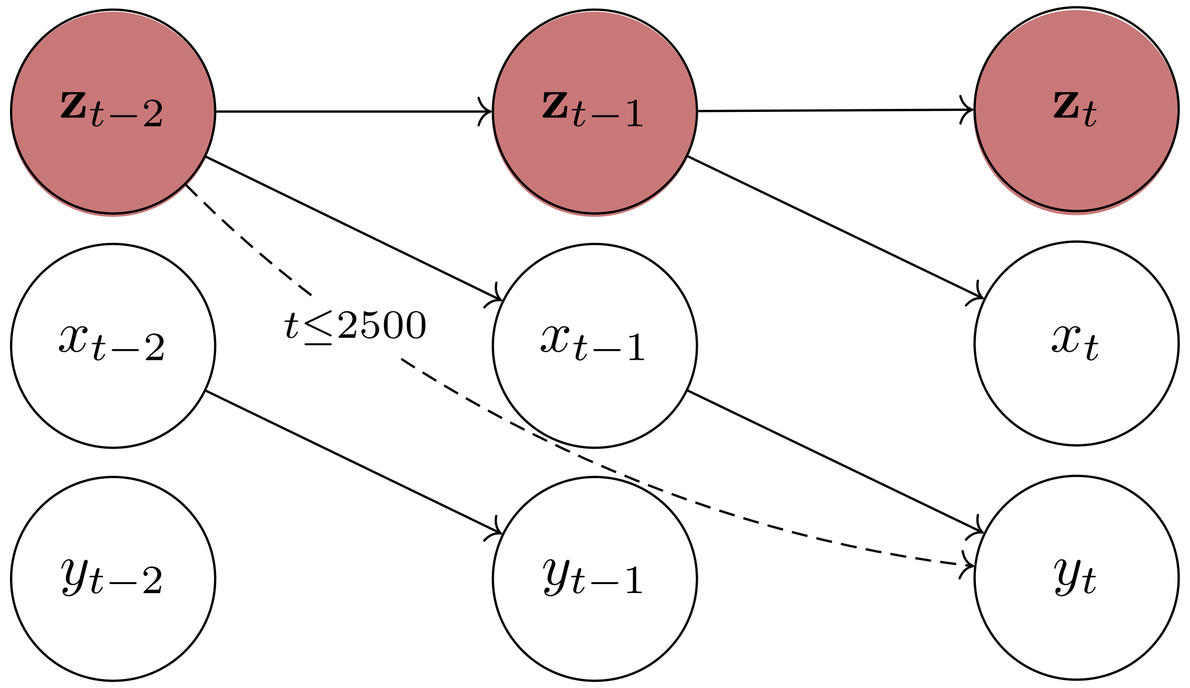
Correct DCEs for observed variables

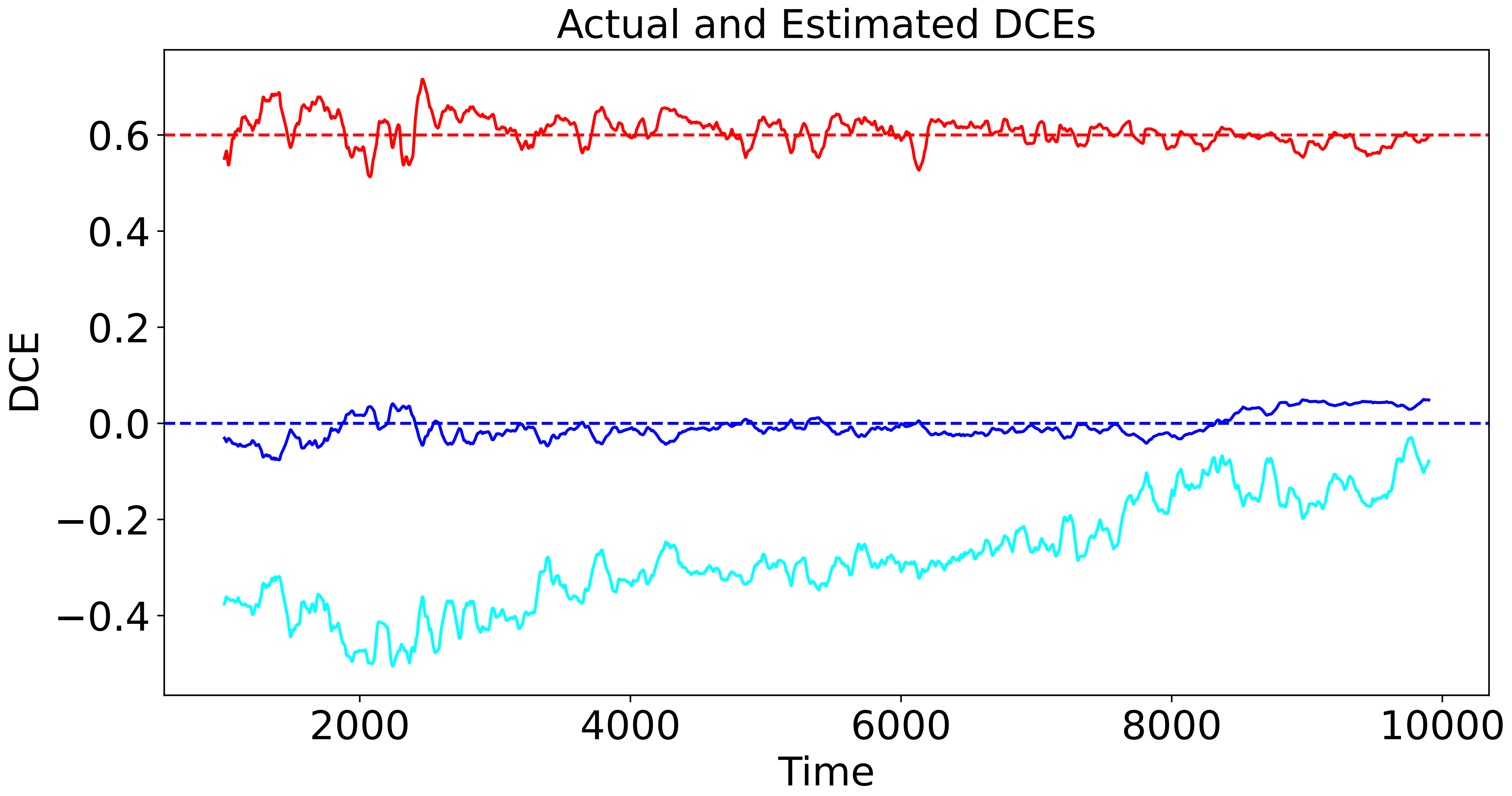
Our Method Can Detect Confounders
In the Static Case



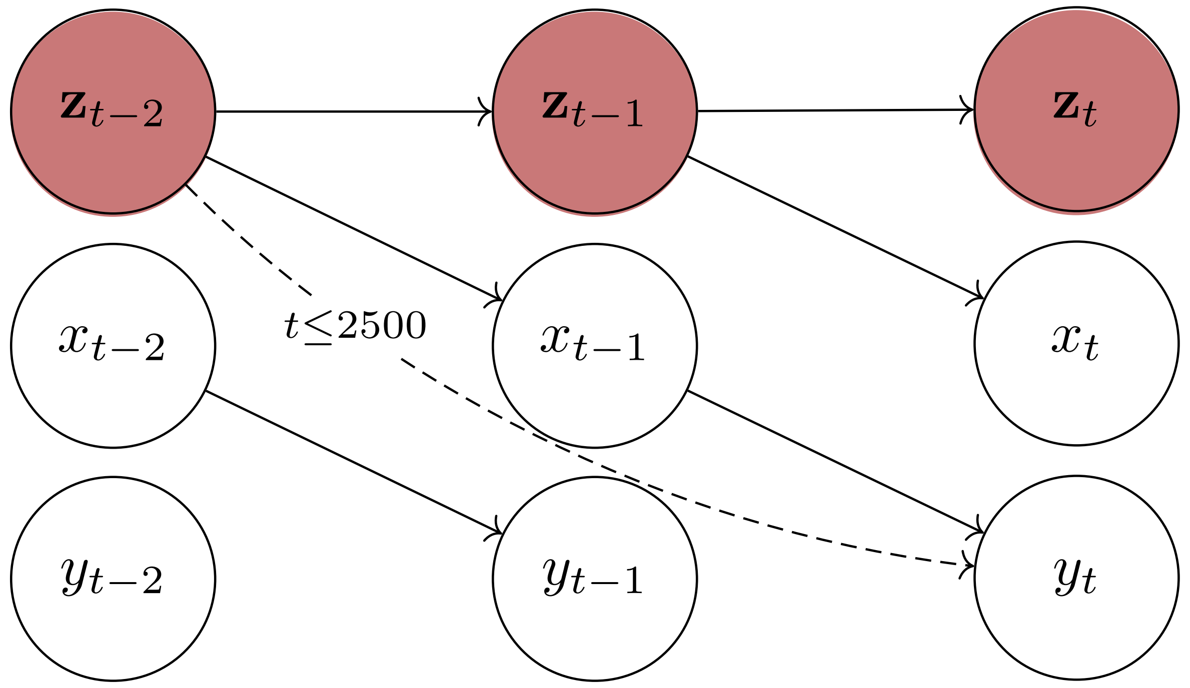


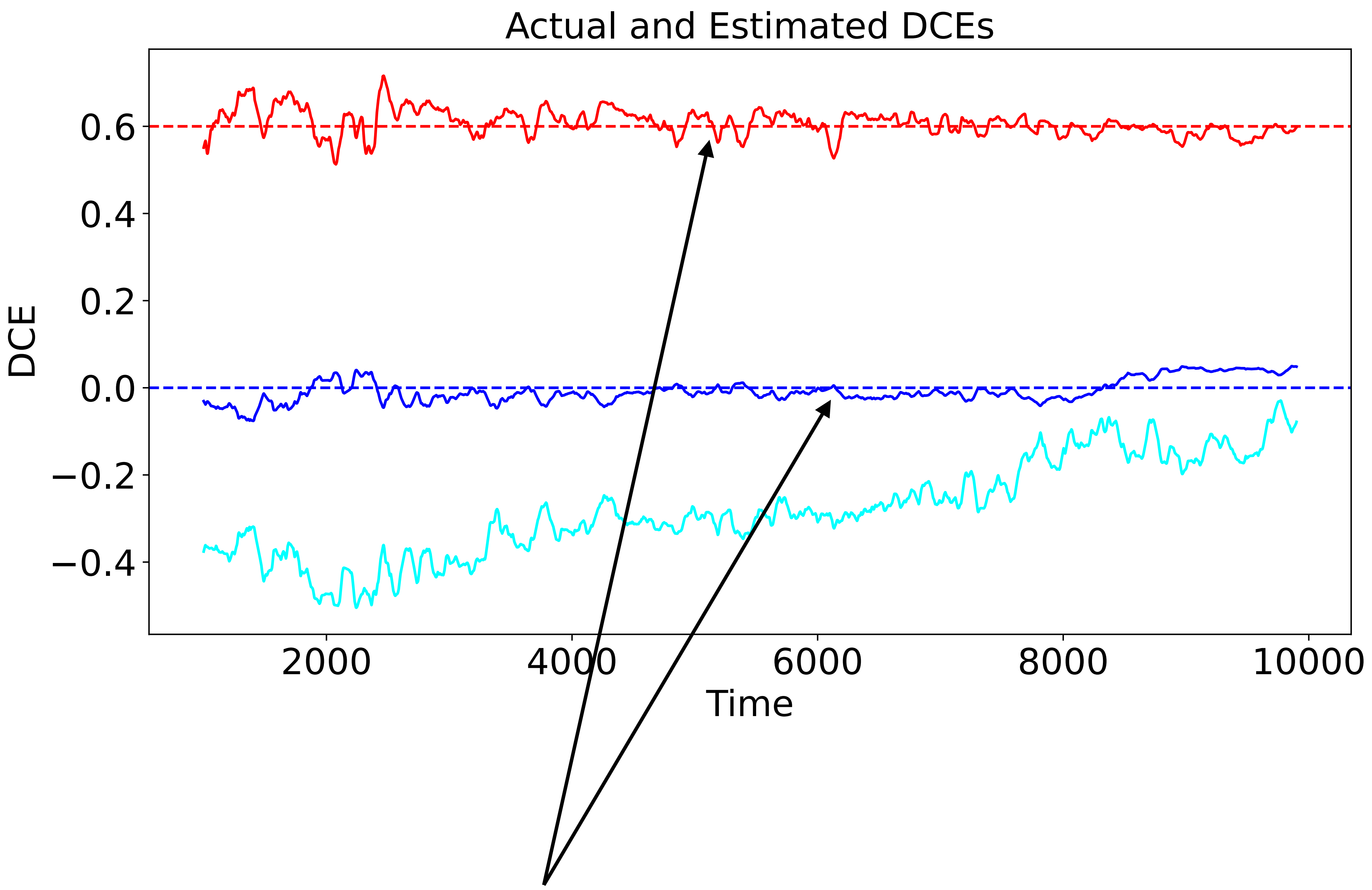
Our Method Can Detect Confounders In the Dynamic Case





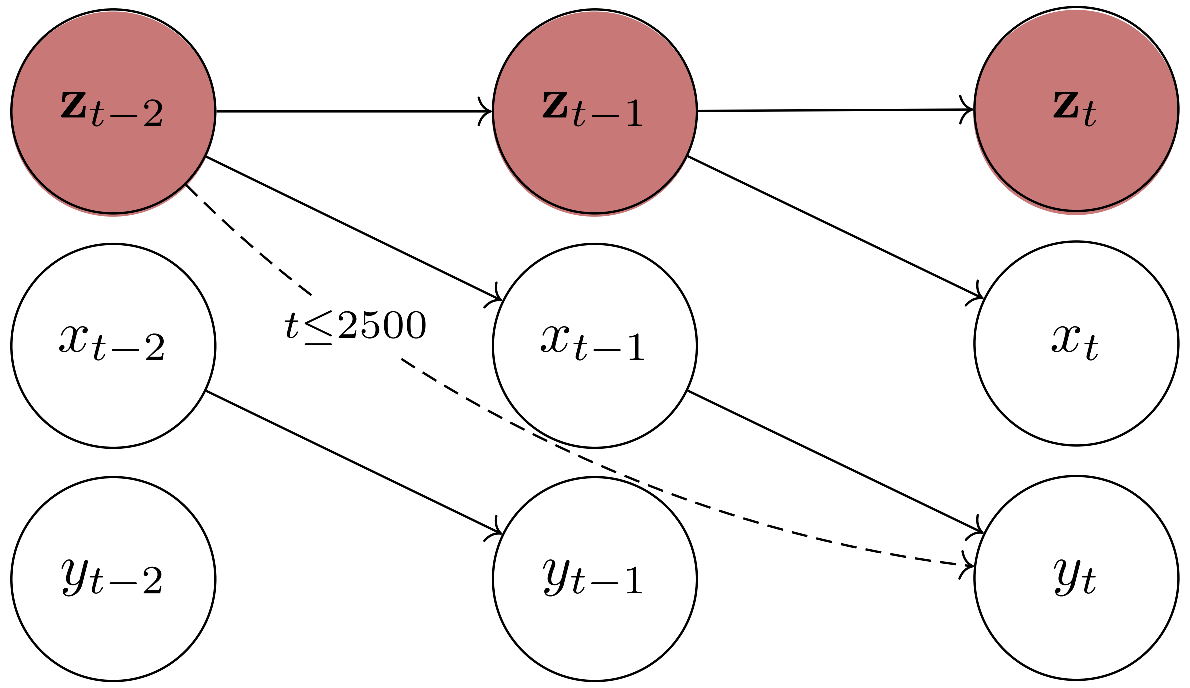
Our Method Can Detect Confounders In the Dynamic Case



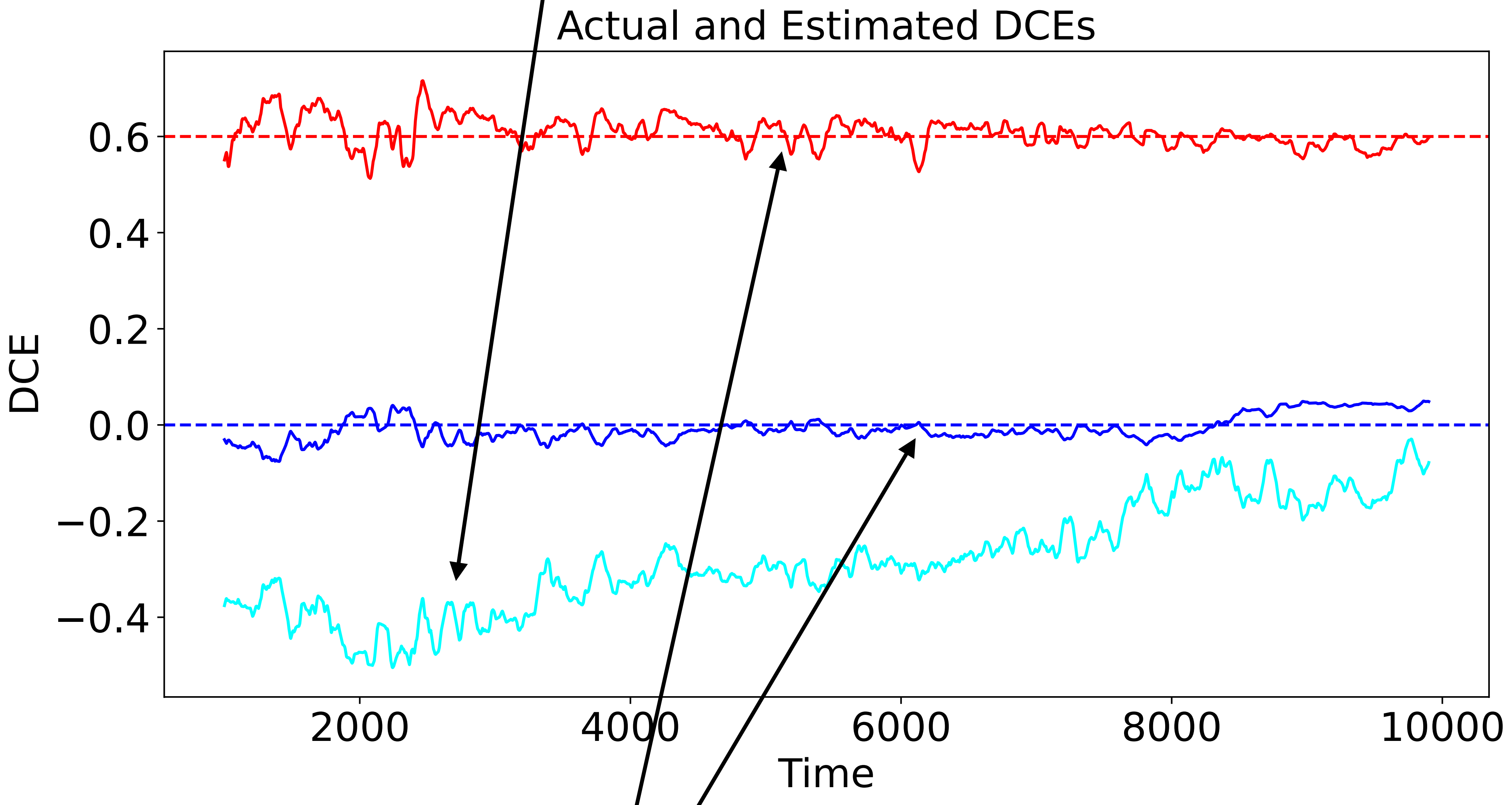


Correct DCEs for observed variables

Our Method Can Detect Confounders In the Dynamic Case



Initially nonzero DCE, decaying to zero for confounder



Correct DCEs for observed variables

Causal Discovery

D. Waxman, K. Butler, and P. M. Djurić
“DAGMA-DCE: Interpretable, Non-Parametric Differentiable Causal Discovery”
Submitted.

Discovering Causal Relationships Requires Assumptions

Discovering Causal Relationships Requires Assumptions

Causal discovery, or causal structure learning, entails learning the causal graph

Discovering Causal Relationships Requires Assumptions

Causal discovery, or causal structure learning, entails learning the causal graph

As before: assumptions, assumptions, and more assumptions

Discovering Causal Relationships Requires Assumptions

Causal discovery, or causal structure learning, entails learning the causal graph

As before: assumptions, assumptions, and more assumptions

Largely, two categories:

Discovering Causal Relationships Requires Assumptions

Causal discovery, or causal structure learning, entails learning the causal graph

As before: assumptions, assumptions, and more assumptions

Largely, two categories:

1. Constraint-based methods

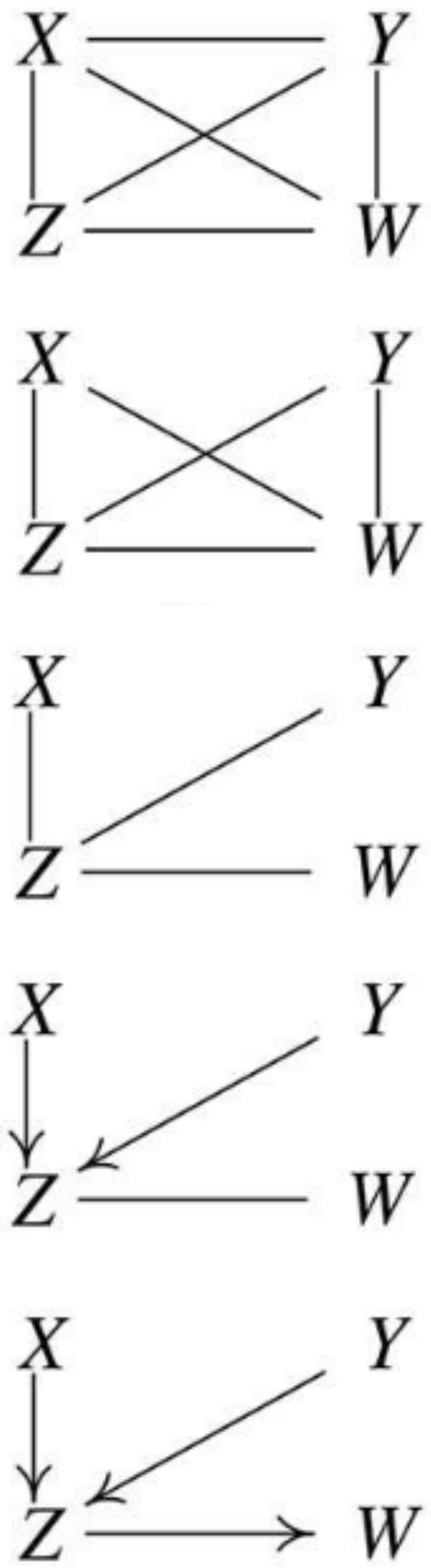
Discovering Causal Relationships Requires Assumptions

Causal discovery, or causal structure learning, entails learning the causal graph

As before: assumptions, assumptions, and more assumptions

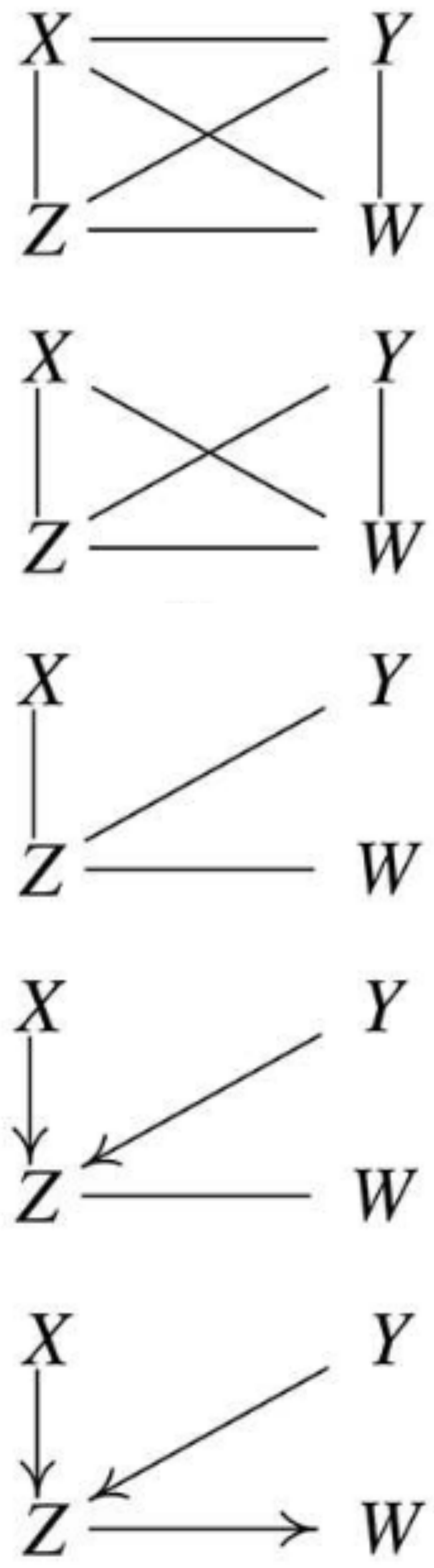
Largely, two categories:
1. Constraint-based methods
2. Score-based methods

Constraint-Based Methods Exploit Independencies



C. Glymour, K. Zhang, & P. Spirtes. (2019). Review of causal discovery methods based on graphical models. *Frontiers in genetics*, *10*, 524.

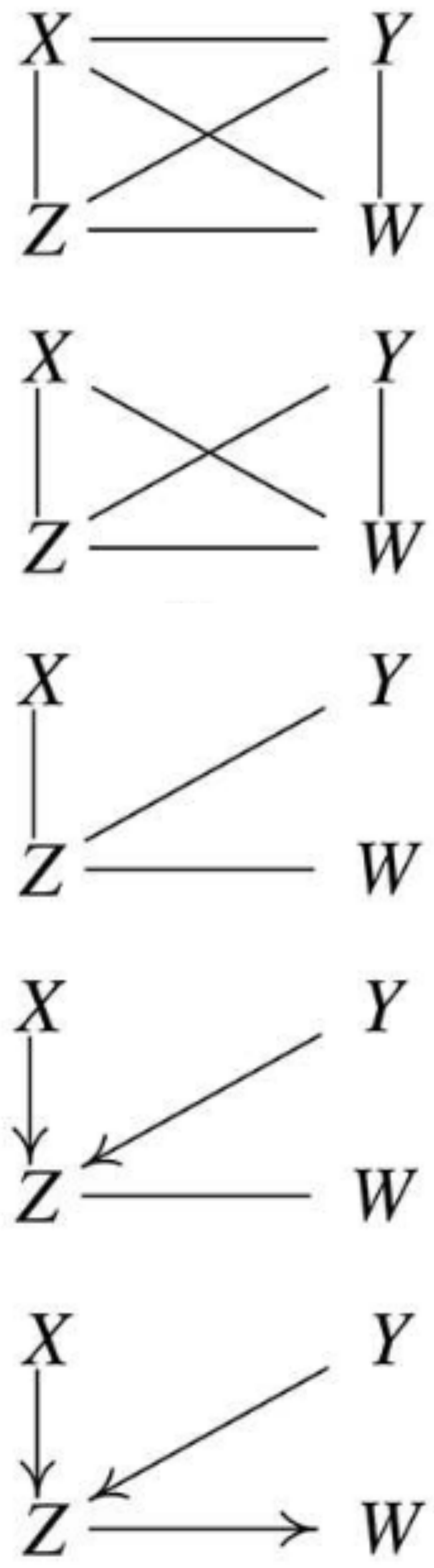
Constraint-Based Methods Exploit Independencies



C. Glymour, K. Zhang, & P. Spirtes. (2019). Review of causal discovery methods based on graphical models. *Frontiers in genetics*, *10*, 524.

The most historically popular methods are constraint-based

Constraint-Based Methods Exploit Independencies



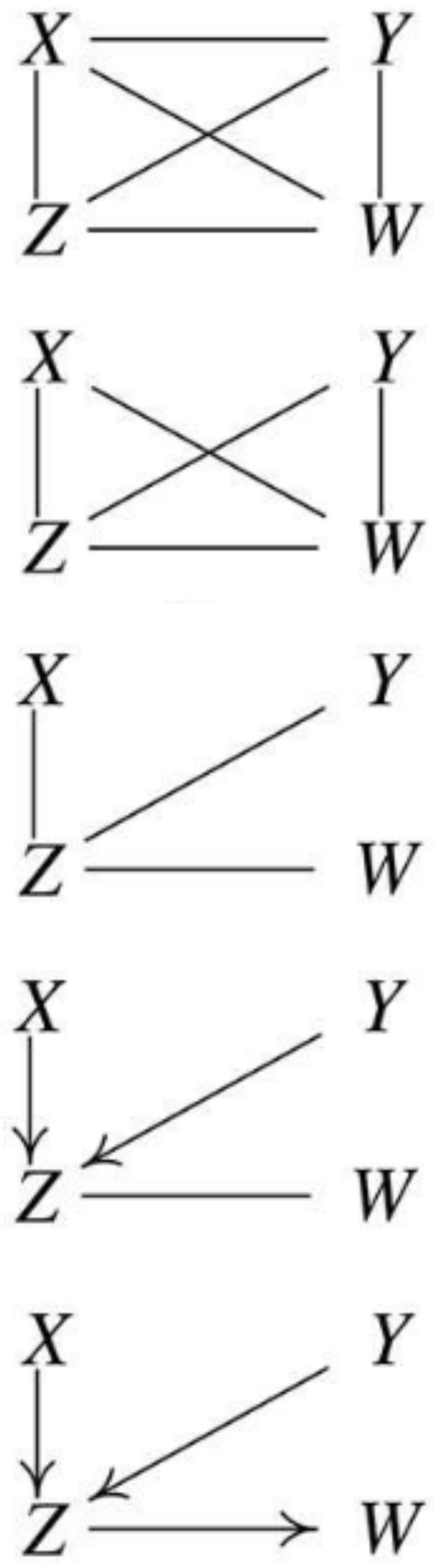
C. Glymour, K. Zhang, & P. Spirtes. (2019). Review of causal discovery methods based on graphical models. *Frontiers in genetics*, *10*, 524.

The most historically popular methods are constraint-based

Recall the “causal Markov” and “causal faithfulness assumptions”

Together, statistical conditional independence if and only if encoded by a causal relationship

Constraint-Based Methods Exploit Independencies



C. Glymour, K. Zhang, & P. Spirtes. (2019). Review of causal discovery methods based on graphical models. *Frontiers in genetics*, *10*, 524.

The most historically popular methods are constraint-based

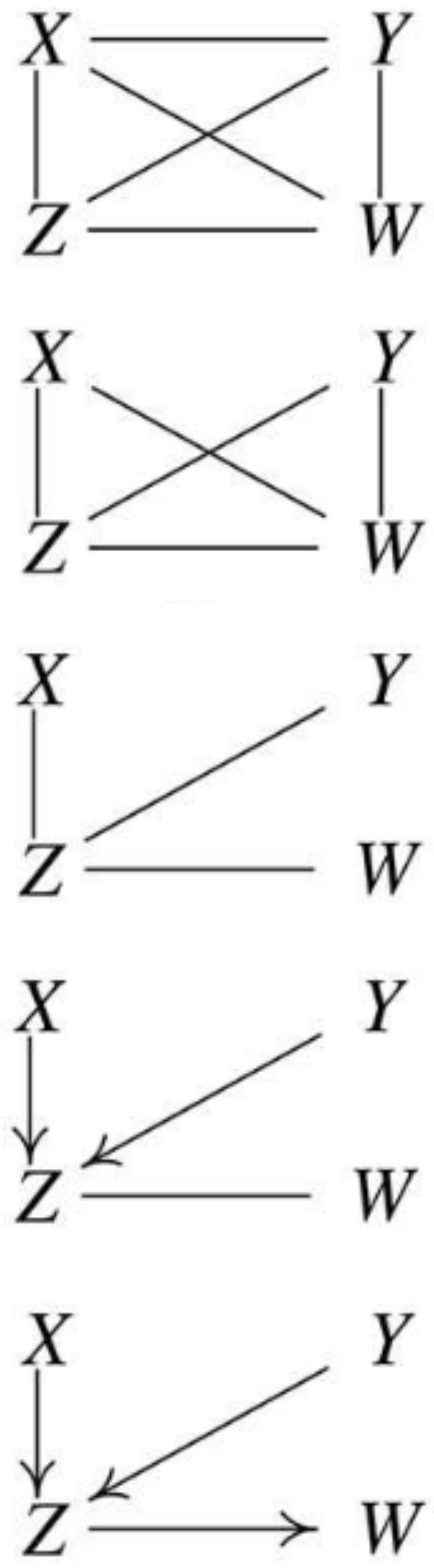
Recall the “causal Markov” and “causal faithfulness assumptions”

Together, statistical conditional independence if and only if encoded by a causal relationship

Constraint-based methods test conditional independencies

This can falsify many causal graphs

Constraint-Based Methods Exploit Independencies



C. Glymour, K. Zhang, & P. Spirtes. (2019). Review of causal discovery methods based on graphical models. *Frontiers in genetics*, *10*, 524.

The most historically popular methods are constraint-based

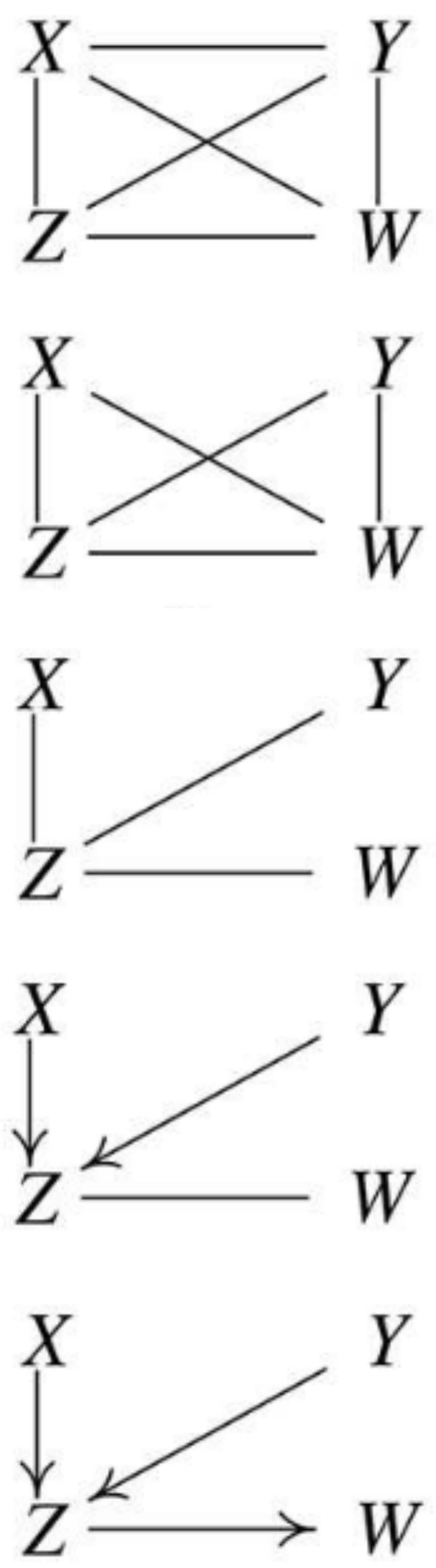
Recall the “causal Markov” and “causal faithfulness assumptions”

Together, statistical conditional independence if and only if encoded by a causal relationship

Constraint-based methods test conditional independencies

This can falsify many causal graphs

Constraint-Based Methods Exploit Independencies



C. Glymour, K. Zhang, & P. Spirtes. (2019). Review of causal discovery methods based on graphical models. *Frontiers in genetics*, *10*, 524.

The most historically popular methods are constraint-based

Recall the "causal Markov" and "causal faithfulness assumptions"

Together, statistical conditional independence if and only if encoded by a causal relationship

Constraint-based methods test conditional independencies

This can falsify many causal graphs

Famous examples include the PC Algorithm and Fast Causal Inference (FCI)

Constraint-Based Learning is Imperfect

Constraint-Based Learning is Imperfect

There are a few issues with constraint-based learning:

Constraint-Based Learning is Imperfect

There are a few issues with constraint-based learning:

- Require lots of data (due to CI tests)

Constraint-Based Learning is Imperfect

There are a few issues with constraint-based learning:

- Require lots of data (due to CI tests)
- Worst-case exponential run time

Constraint-Based Learning is Imperfect

There are a few issues with constraint-based learning:

- Require lots of data (due to CI tests)
- Worst-case exponential run time
- Only an equivalence class of graphs is recovered

Constraint-Based Learning is Imperfect

There are a few issues with constraint-based learning:

- Require lots of data (due to CI tests)
- Worst-case exponential run time
- Only an equivalence class of graphs is recovered

Still, they can be a great tool

Constraint-Based Learning is Imperfect

There are a few issues with constraint-based learning:

- Require lots of data (due to CI tests)
- Worst-case exponential run time
- Only an equivalence class of graphs is recovered

Still, they can be a great tool

But we'll work on score-based methods instead

Score-Based Methods Exploit Identifiability

Score-Based Methods Exploit Identifiability

MLE-type (or AIC/BIC-type) procedures are very nice

Score-Based Methods Exploit Identifiability

MLE-type (or AIC/BIC-type) procedures are very nice

Surprisingly mild assumptions ensure identifiability

Score-Based Methods Exploit Identifiability

MLE-type (or AIC/BIC-type) procedures are very nice

Surprisingly mild assumptions ensure identifiability

For example:

Score-Based Methods Exploit Identifiability

MLE-type (or AIC/BIC-type) procedures are very nice

Surprisingly mild assumptions ensure identifiability

For example:

- Linear model with additive non-Gaussian noise (LiNGAM)

Score-Based Methods Exploit Identifiability

MLE-type (or AIC/BIC-type) procedures are very nice

Surprisingly mild assumptions ensure identifiability

For example:

- Linear model with additive non-Gaussian noise (LiNGAM)
- Three-times differentiable, strictly nonlinear with additive Gaussian noise

Score-Based Methods Exploit Identifiability

MLE-type (or AIC/BIC-type) procedures are very nice

Surprisingly mild assumptions ensure identifiability

For example:

- Linear model with additive non-Gaussian noise (LiNGAM)
- Three-times differentiable, strictly nonlinear with additive Gaussian noise

Score-based methods search over DAGs to minimize the MLE/ AIC/BIC

There are too Many DAGs to do This Quickly

```
1: 1

2: 3

3: 25

4: 543

5: 29281

...
```

14: 1.4×10^{36}

There are too Many DAGs to do This Quickly

The space of DAGs grows super-exponentially

1: 1

2: 3

3: 25

4: 543

5: 29281

...

14: 1.4×10^{36}

There are too Many DAGs to do This Quickly

The space of DAGs grows super-exponentially

This poses issues for discrete optimization

```
1: 1

2: 3

3: 25

4: 543

5: 29281

...
```

14: 1.4×10^{36}

There are too Many DAGs to do This Quickly

```
1: 1

2: 3

3: 25

4: 543

5: 29281

...
```

14: 1.4×10^{36}

The space of DAGs grows super-exponentially

This poses issues for discrete optimization

Clever approaches (e.g. GES, GEIS) help, but are still slow

There are too Many DAGs to do This Quickly

1: 1

2: 3

3: 25

4: 543

5: 29281

...

14: 1.4×10^{36}

The space of DAGs grows super-exponentially

This poses issues for discrete optimization

Clever approaches (e.g. GES, GEIS) help, but are still slow

We need to avoid searching all DAGs

Using Continuous Optimization To Search

Using Continuous Optimization To Search

Let $\mathbf{A}$ = binary adjacency matrix of $\mathscr{G}$

Using Continuous Optimization To Search

Let $\mathbf{A}$ = binary adjacency matrix of $\mathscr{G}$

$\left[\mathbf{A}^k\right]_{ij}$ = # of paths of length k from x_i to x_j

Using Continuous Optimization To Search

Let $\mathbf{A}$ = binary adjacency matrix of $\mathcal{G}$

$\left[\mathbf{A}^k\right]_{ij}$ = # of paths of length k from x_i to x_j

Let's look at $k = 2$:

$$\left[A^2\right]_{ij} = \sum a_{ik} a_{kj}$$

Using Continuous Optimization To Search

Let $\mathbf{A}$ = binary adjacency matrix of $\mathscr{G}$

$\left[\mathbf{A}^k\right]_{ij}$ = # of paths of length k from x_i to x_j

Let's look at $k = 2$:

$$\left[A^2\right]_{ij} = \sum a_{ik}a_{kj}$$

NOTEARS [Zheng et al., NeurIPS 2018] characterizes

$\mathscr{G}$ is a DAG $\iff$ $\mathrm{trace}(\exp A) - d = 0$

Using Continuous Optimization To Search

Let $\mathbf{A}$ = binary adjacency matrix of $\mathcal{G}$

$\left[\mathbf{A}^k\right]_{ij}$ = # of paths of length k from x_i to x_j

Let's look at $k = 2$:

$$\left[A^2\right]_{ij} = \sum a_{ik} a_{kj}$$

NOTEARS [Zheng et al., NeurIPS 2018] characterizes

$\mathcal{G}$ is a DAG $\iff$ $\mathrm{trace}(\exp A) - d = 0$

Searches over DAGs can then be constrained, continuous optimization

Linear NOTEARS

Linear NOTEARS

Linear case [Zheng et al., NeurIPS 2018] is the easiest

Linear NOTEARS

Linear case [Zheng et al., NeurIPS 2018] is the easiest

Use the linear coefficients as $\mathbf{A}$, then solve

$$\min_{\mathbf{A}} \|\mathbf{X} - \mathbf{X}\mathbf{A}\|_F^2 + \lambda \|\mathbf{A}\|_1$$

$$\text{s.t. } \operatorname{tr}\left(\exp\left(\mathbf{A} \odot \mathbf{A}\right)\right) - d = 0$$

Linear NOTEARS

Linear case [Zheng et al., NeurIPS 2018] is the easiest

Use the linear coefficients as $\mathbf{A}$, then solve

$$\min_{\mathbf{A}} \|\mathbf{X} - \mathbf{X}\mathbf{A}\|_F^2 + \lambda \|\mathbf{A}\|_1$$

$$\text{s.t. } \operatorname{tr}\left(\exp\left(\mathbf{A} \odot \mathbf{A}\right)\right) - d = 0$$

This is well-posed and gets very nice results

Nonlinear NOTEARS

Nonlinear NOTEARS

Nonlinear case [Zheng et al., AISTATS 2020] is similar

Nonlinear NOTEARS

Nonlinear case [Zheng et al., AISTATS 2020] is similar

Idea: define $\left[\mathbf{A}\right]_{ij} = \|\partial_i f_j\|_2$

Nonlinear NOTEARS

Nonlinear case [Zheng et al., AISTATS 2020] is similar

Idea: define $\left[\mathbf{A}\right]_{ij} = \|\partial_i f_j\|_2$

Second idea: use a proxy that maintains zeroness

For example, in an MLP, take the L_2 norm of the first layer

Nonlinear NOTEARS

Nonlinear case [Zheng et al., AISTATS 2020] is similar

Idea: define $\left[\mathbf{A}\right]_{ij} = \|\partial_i f_j\|_2$

Second idea: use a proxy that maintains zeroness

For example, in an MLP, take the L_2 norm of the first layer

For parameterized model $\mathscr{M}_\theta$

$$\min_\theta \|\mathbf{X} - f_\theta(\mathbf{X})\|_2^2 + \lambda\|\mathbf{A}_\theta\|_1$$

$$\text{s.t. } \operatorname{tr}\left(\exp\left(\mathbf{A}_\theta \odot \mathbf{A}_\theta\right)\right) - d = 0$$

Fixing NOTEARS' Gradients

Fixing NOTEARS’ Gradients

NOTEARS has poorly behaved gradients

Fixing NOTEARS' Gradients

NOTEARS has poorly behaved gradients

It also uses the augmented Laplacian

Fixing NOTEARS' Gradients

NOTEARS has poorly behaved gradients

It also uses the augmented Laplacian

Idea: define a class of matrices so that barrier methods work

In this case, M-matrices from econometrics

Fixing NOTEARS' Gradients

NOTEARS has poorly behaved gradients

It also uses the augmented Laplacian

Idea: define a class of matrices so that barrier methods work

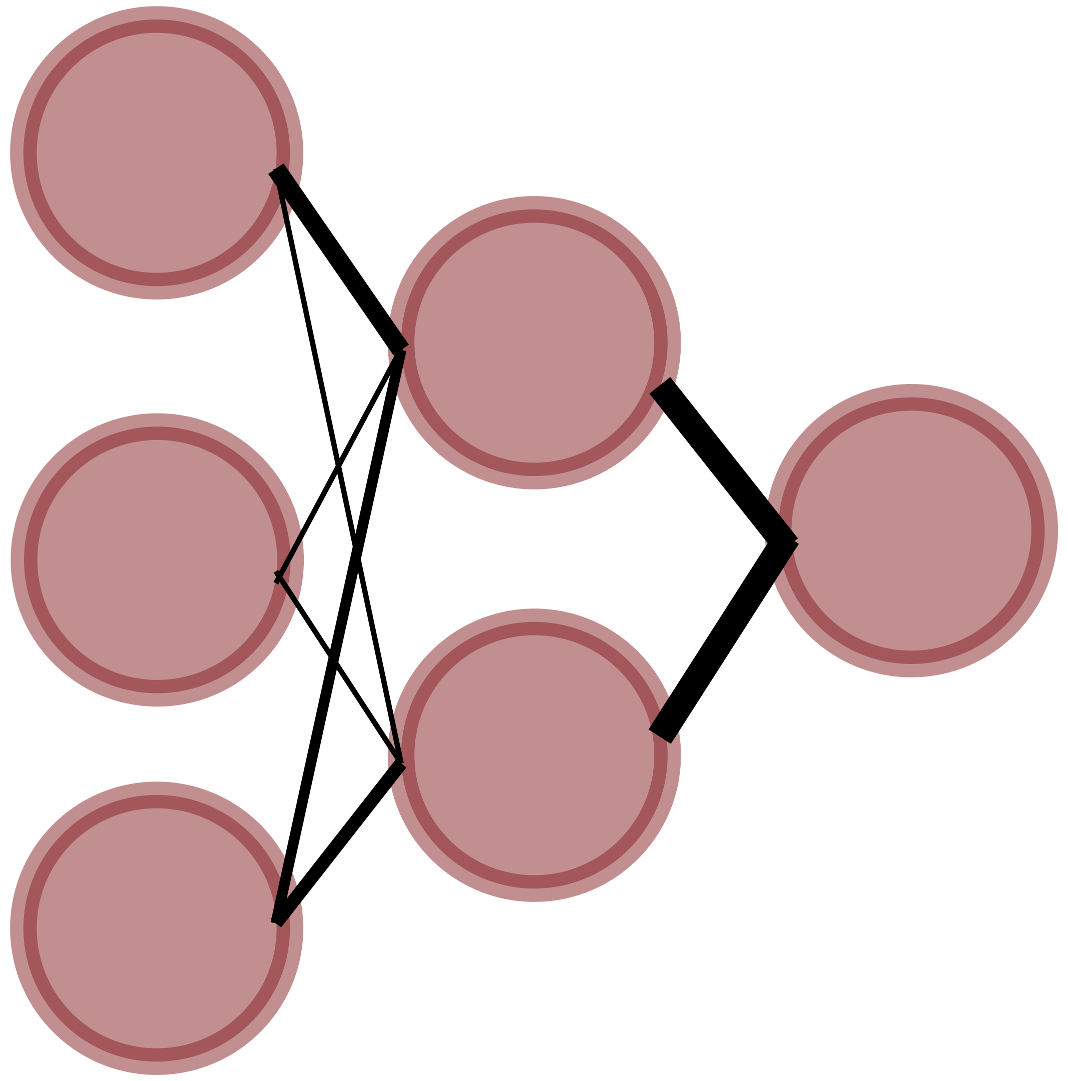
In this case, M-matrices from econometrics

DAGMA [Bello et al., NeurIPS 2022] then gives a different constraint:

$$\min_{\theta} ||\mathbf{X} - f_\theta(\mathbf{X})||_F^2 + \lambda ||\mathbf{A}_\theta||_1$$

$$\text{s.t.} \; -\log\left(\det\left(s\mathbb{I} - \mathbf{A}_\theta \odot \mathbf{A}_\theta\right)\right) + d\log s = 0$$

$\mathbf{A}_\theta$ is Arbitrarily Misspecified

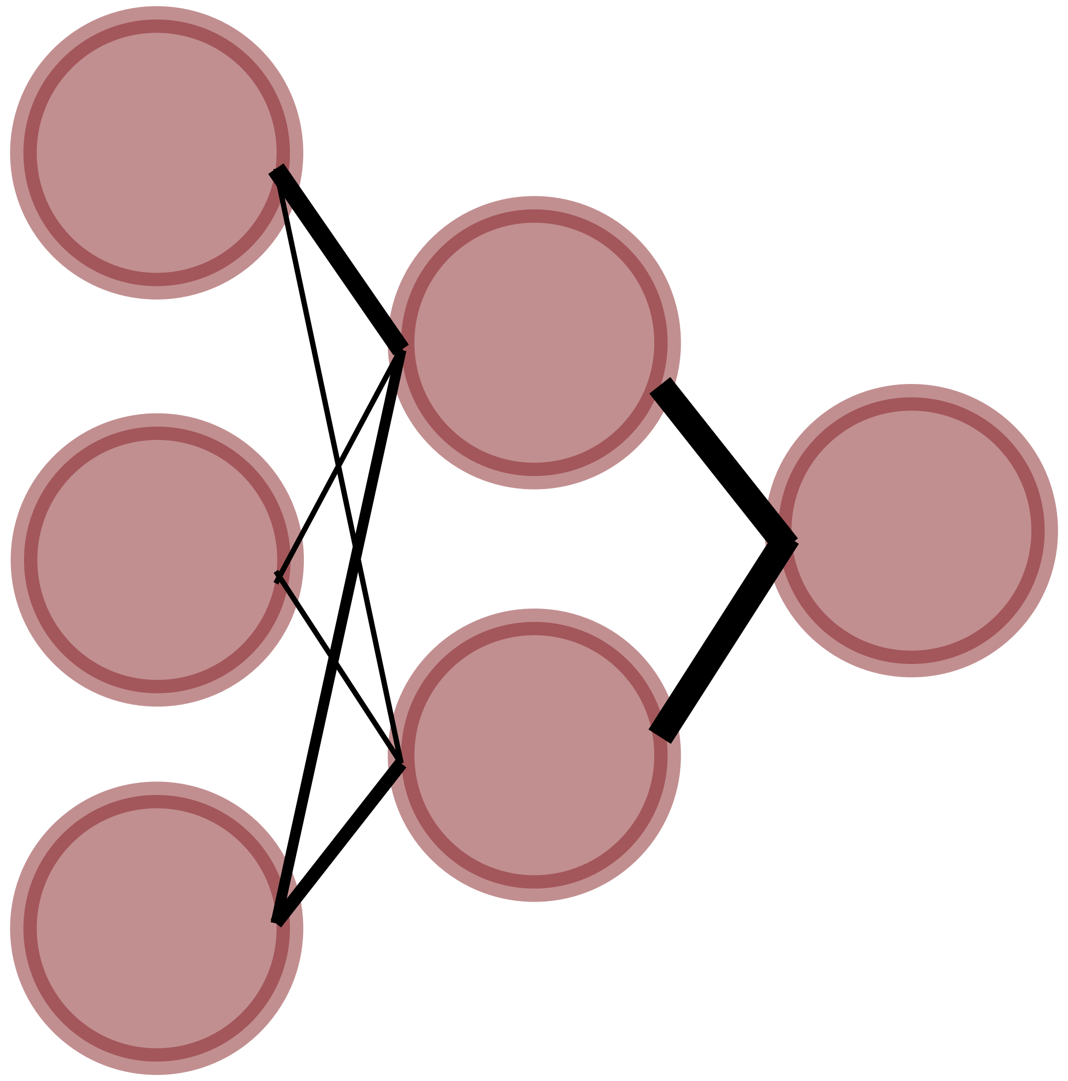


$$|\boldsymbol{+}| = |$$

$$\|\partial_i f_j\| = \mathbf{|}$$

$\mathbf{A}_\theta$ is Arbitrarily Misspecified

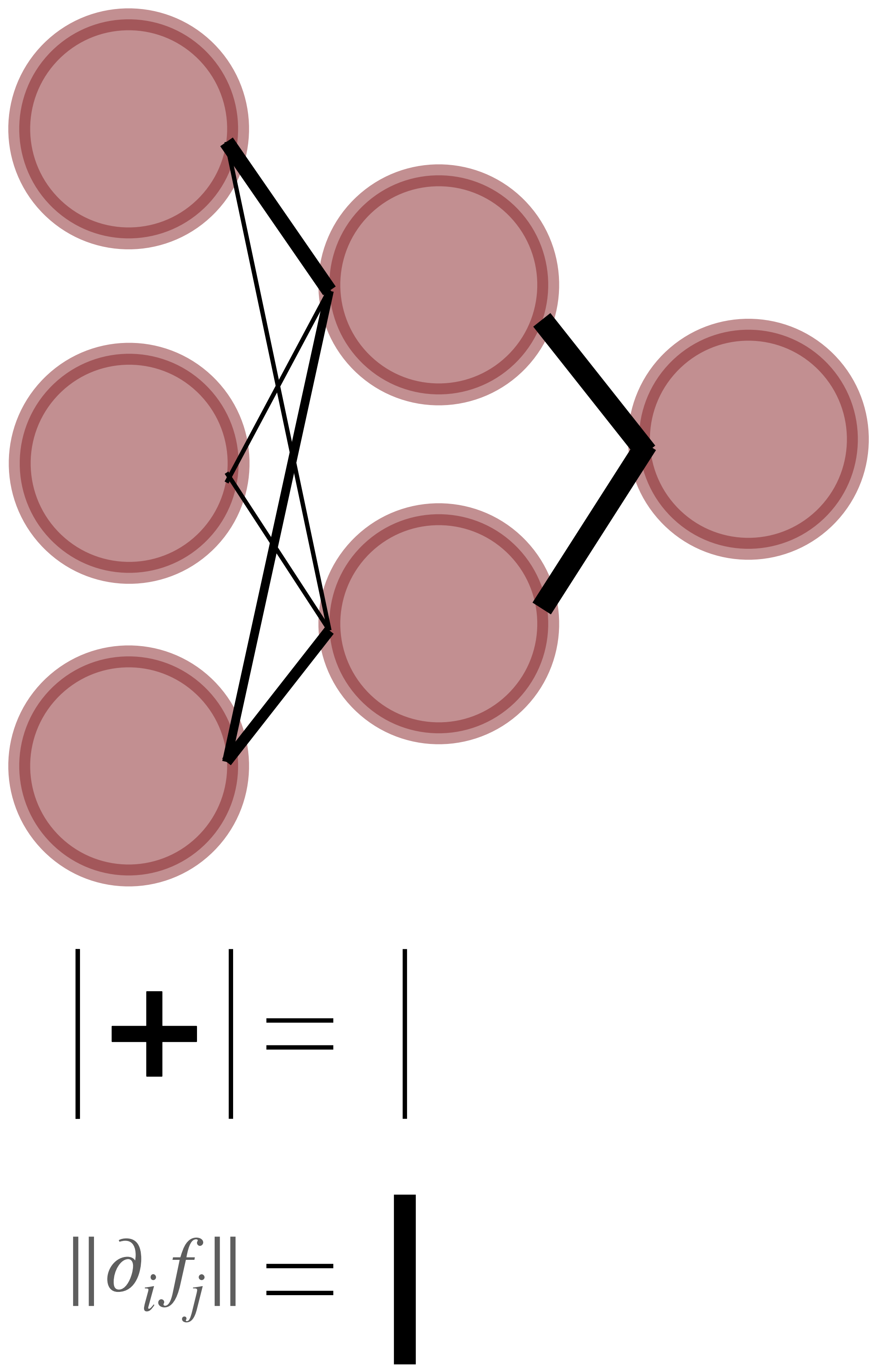
DAGMA-MLP defines $\mathbf{A}_\theta$ using the L^2 norm



$|\mathbf{+}| =$

$\|\partial_i f_j\| =$

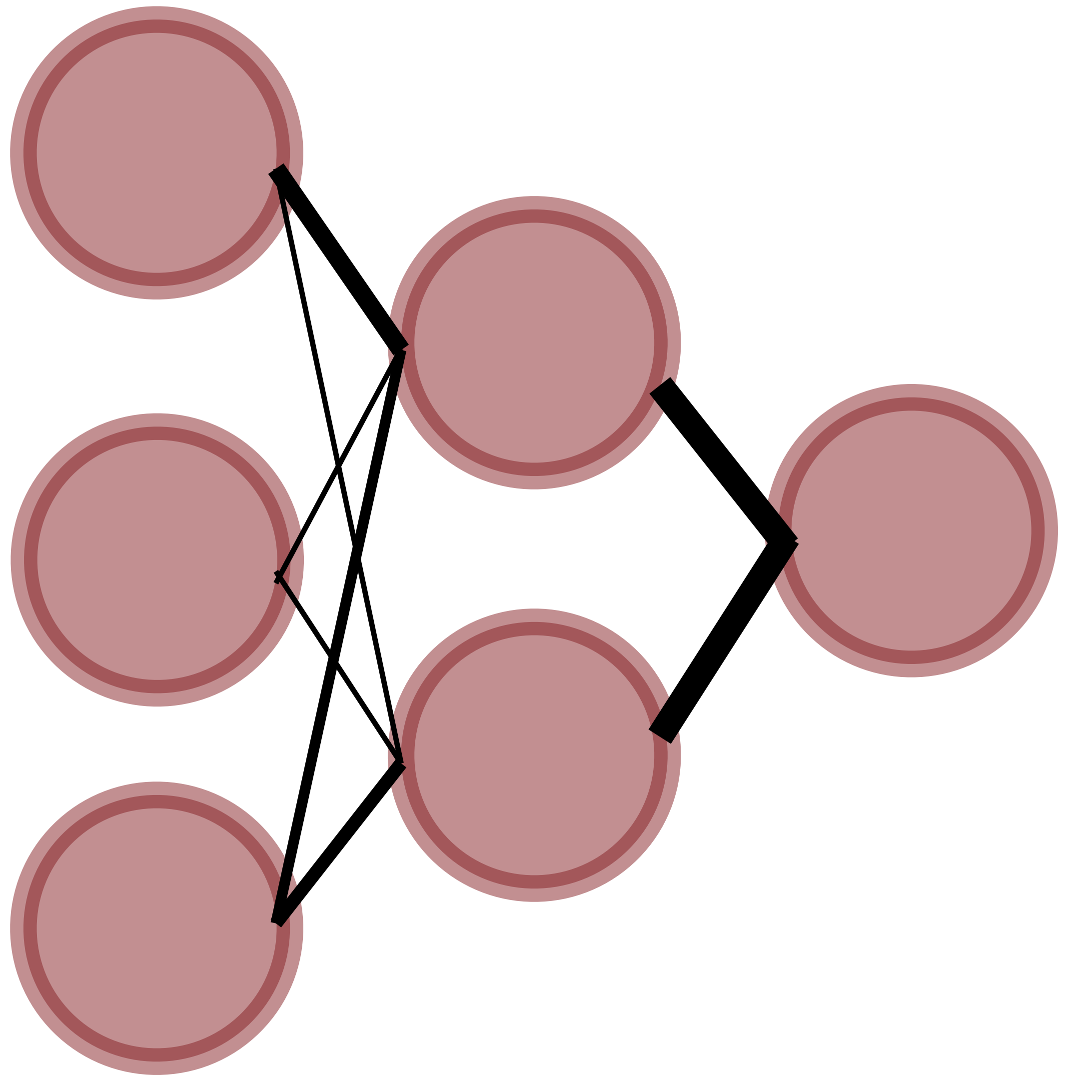
$\mathbf{A}_\theta$ is Arbitrarily Misspecified



DAGMA-MLP defines $\mathbf{A}_\theta$ using the L^2 norm

This is arbitrarily detached from $||\partial_i f_j||_2$

$\mathbf{A}_\theta$ is Arbitrarily Misspecified



$\left| \mathbf{+} \right| = \ |$

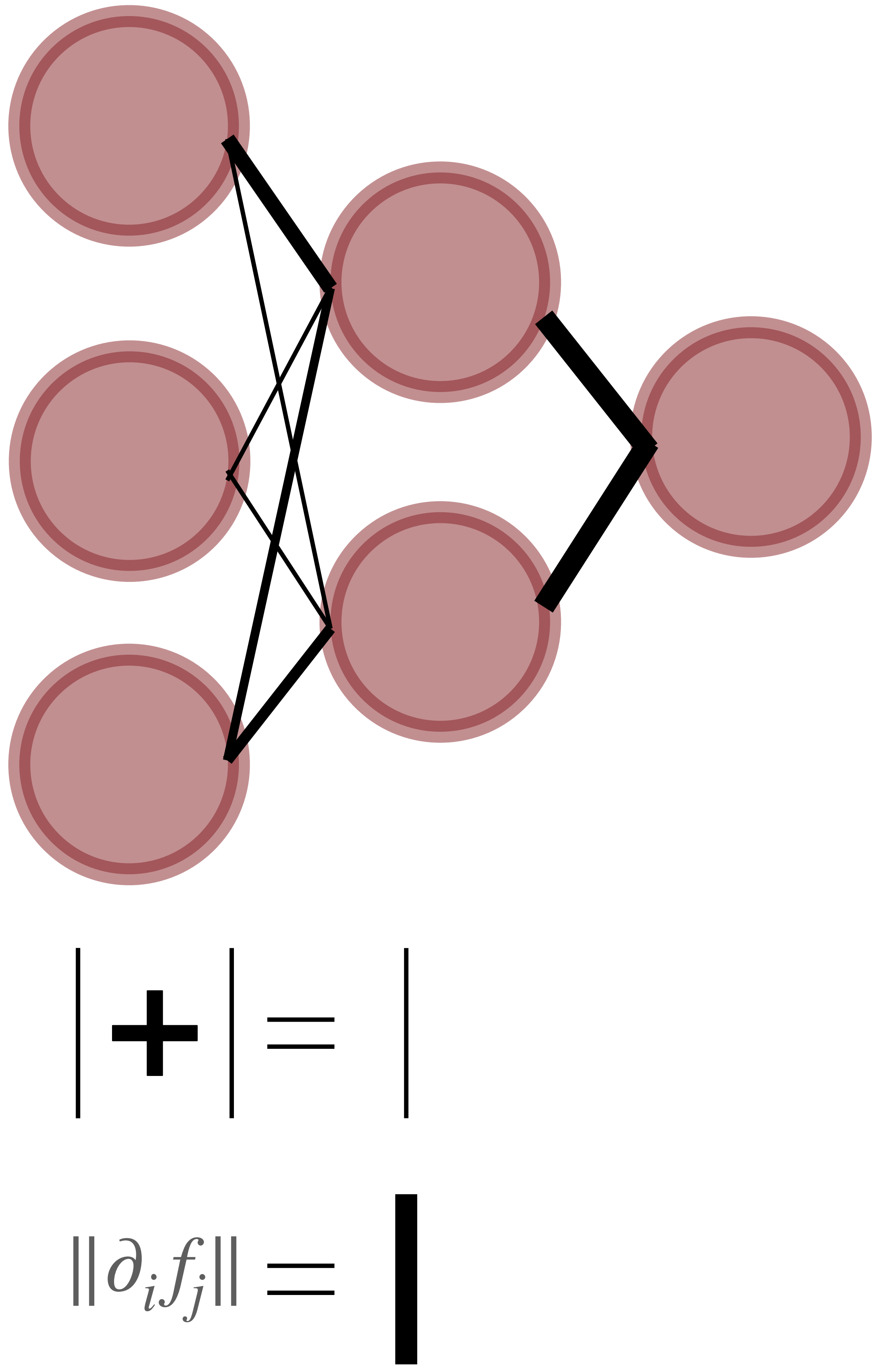
$\|\partial_i f_j\| = \ \mathbf{|}$

DAGMA-MLP defines $\mathbf{A}_\theta$ using the L^2 norm

This is arbitrarily detached from $||\partial_i f_j||_2$

Lemma: There exists an MLP with weight matrices $B^{(1)}, \ldots, B^{(M)}$ and sigmoidal activation such that $||B_1^{(1)}||_2 < \epsilon$ but $||\partial_i f_j||_2 > \delta$.

$\mathbf{A}_\theta$ is Arbitrarily Misspecified



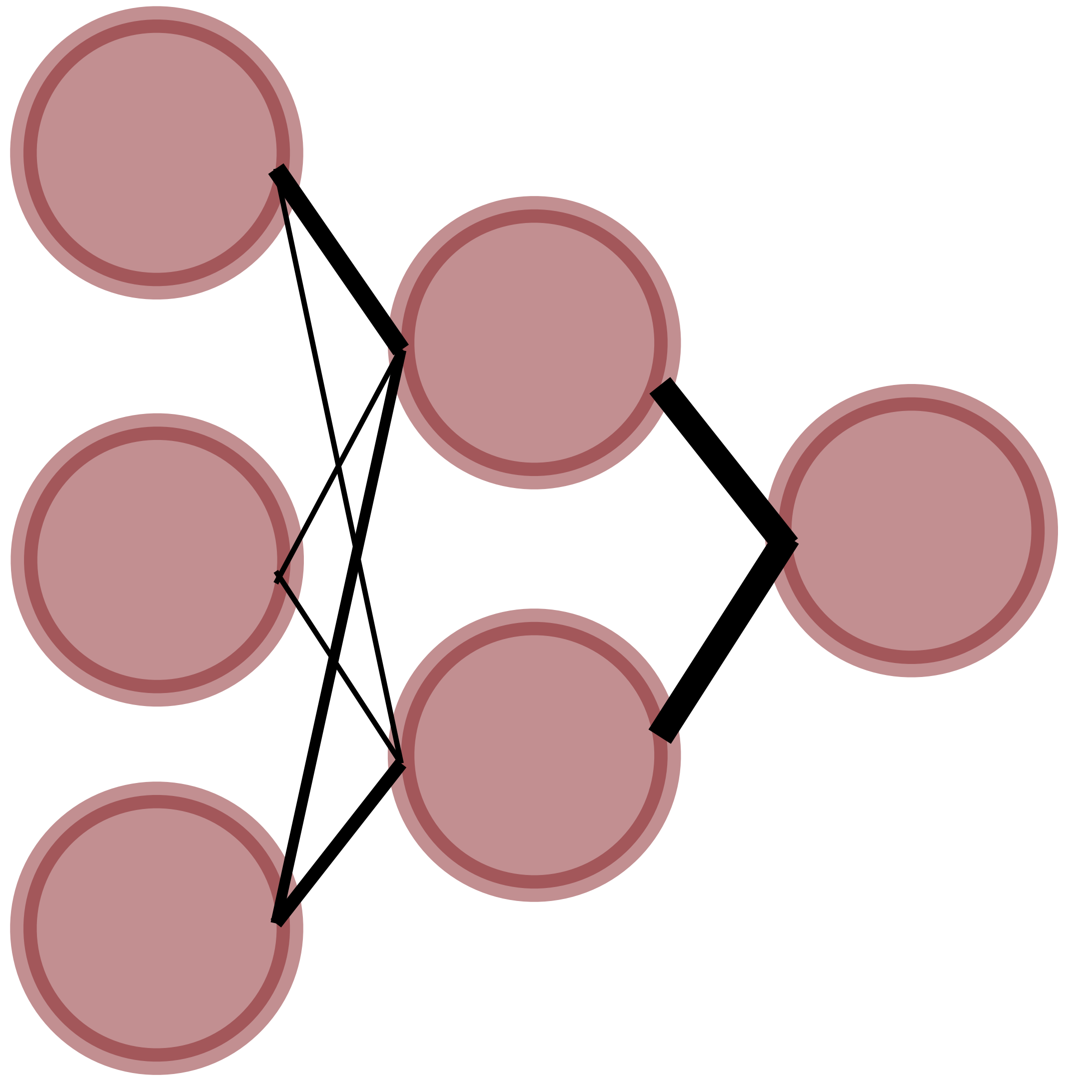
DAGMA-MLP defines $\mathbf{A}_\theta$ using the L^2 norm

This is arbitrarily detached from $\|\partial_i f_j\|_2$

Lemma: There exists an MLP with weight matrices $B^{(1)}, \ldots, B^{(M)}$ and sigmoidal activation such that $\|B_1^{(1)}\|_2 < \epsilon$ but $\|\partial_i f_j\|_2 > \delta$.

Proof Idea: for each outgoing edge of $B^{(1)}$ which is small, compensate with very large edges in $B^{(2)}$

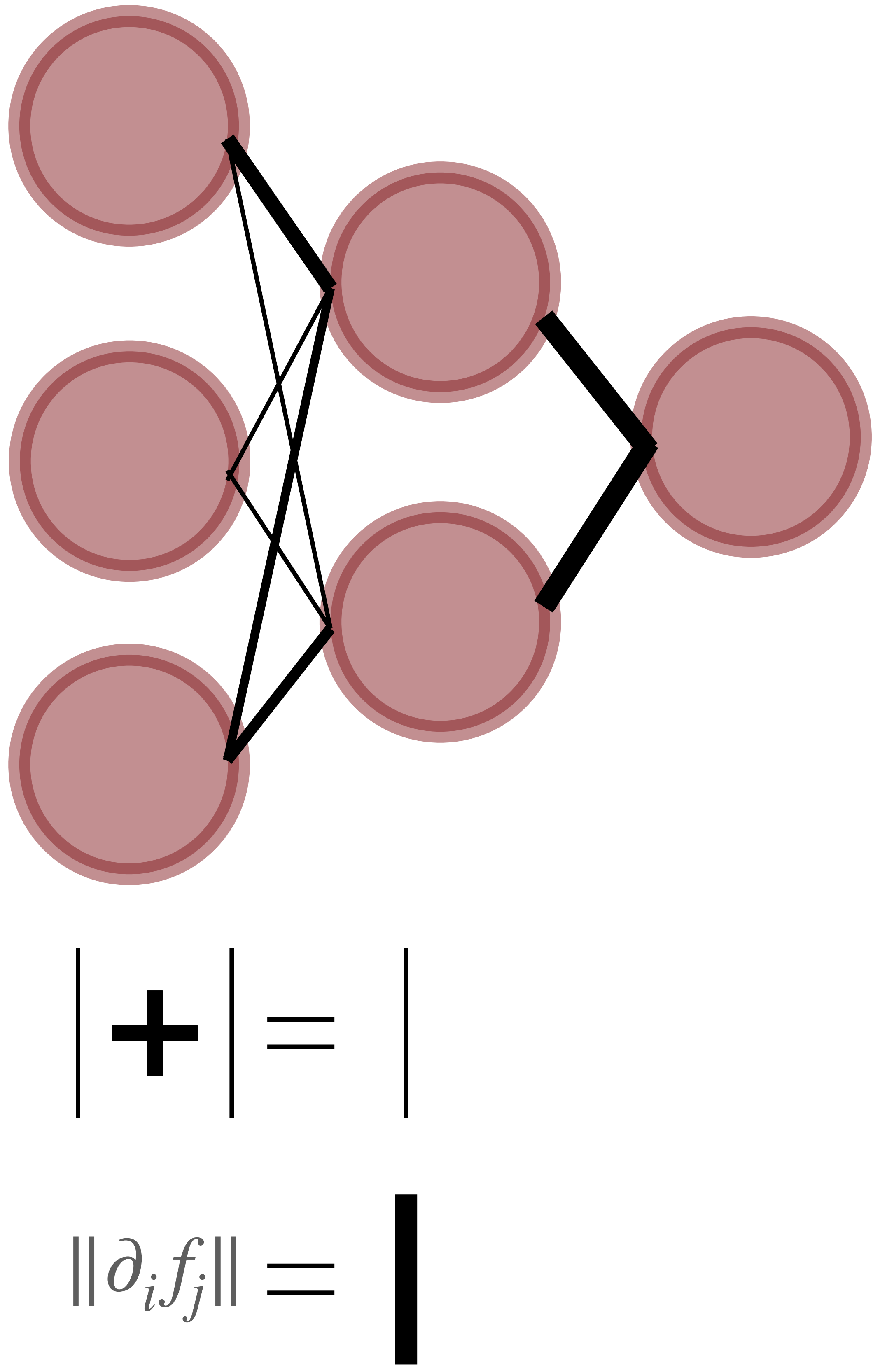
We Redefine $\mathbf{A}_\theta$ using DCE



$\left| + \right| =$

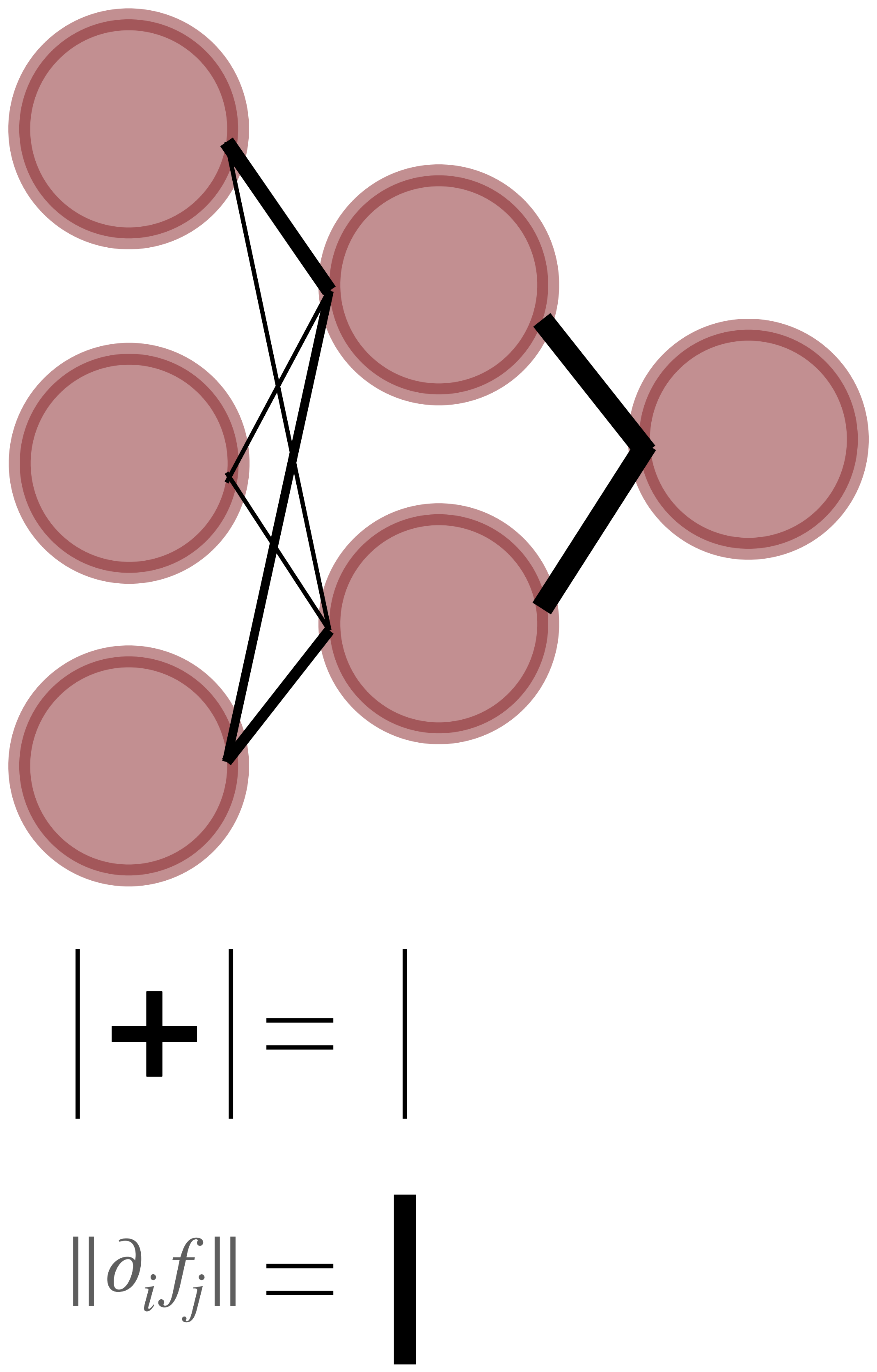
$\|\partial_i f_j\| =$

We Redefine $\mathbf{A}_\theta$ using DCE



Define instead $\mathbf{A}_\theta \triangleq \|\partial_i f_j\|_{L_2(\mathbb{P}^X)}$

We Redefine $\mathbf{A}_\theta$ using DCE

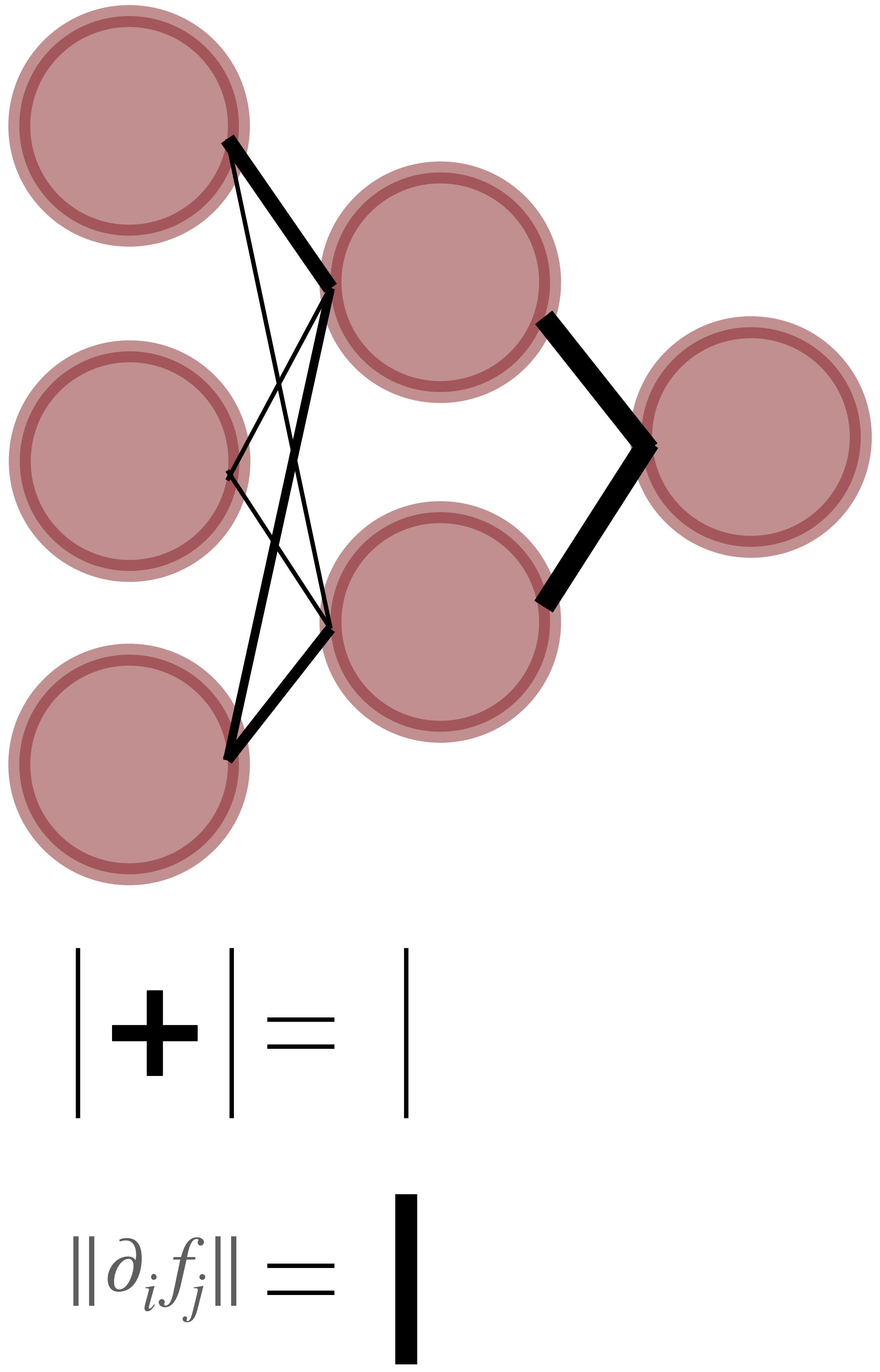


Define instead $\mathbf{A}_\theta \triangleq \|\partial_i f_j\|_{L_2(\mathbb{P}^X)}$

We can take a Monte Carlo approximation

$$\left[\mathbf{A}_\theta\right]_{ij} \approx \sqrt{\frac{1}{N}\sum_{n=1}^{N}\left(\partial_i f_j(x_n)\right)^2}$$

We Redefine $\mathbf{A}_\theta$ using DCE



Define instead $\mathbf{A}_\theta \triangleq \|\partial_i f_j\|_{L_2(\mathbb{P}^X)}$

We can take a Monte Carlo approximation

$$\left[\mathbf{A}_\theta\right]_{ij} \approx \sqrt{\frac{1}{N}\sum_{n=1}^{N}\Big(\partial_i f_j(x_n)\Big)^2}$$

This is the root-mean-square DCE

DAGMA-DCE

DAGMA-DCE

Our optimization problem stays the same

$$\min_{\theta} \|\mathbf{X} - f_\theta(\mathbf{X})\|_2^2 + \lambda \|\mathbf{A}_\theta\|_1$$

$$\text{s.t.} \; -\log\left(\det\left(s\mathbb{I} - \mathbf{A}_\theta \odot \mathbf{A}_\theta\right)\right) + d\log s = 0$$

DAGMA-DCE

Our optimization problem stays the same

$$\min_{\theta} \|\mathbf{X} - f_\theta(\mathbf{X})\|_2^2 + \lambda \|\mathbf{A}_\theta\|_1$$

$$\text{s.t.} \ -\log\left(\det\left(s\mathbb{I} - \mathbf{A}_\theta \odot \mathbf{A}_\theta\right)\right) + d\log s = 0$$

Notably, $\|\mathbf{A}_\theta\|_1$ is different!

DAGMA-DCE

Our optimization problem stays the same

$$\min_{\theta} \|\mathbf{X} - f_\theta(\mathbf{X})\|_2^2 + \lambda \|\mathbf{A}_\theta\|_1$$

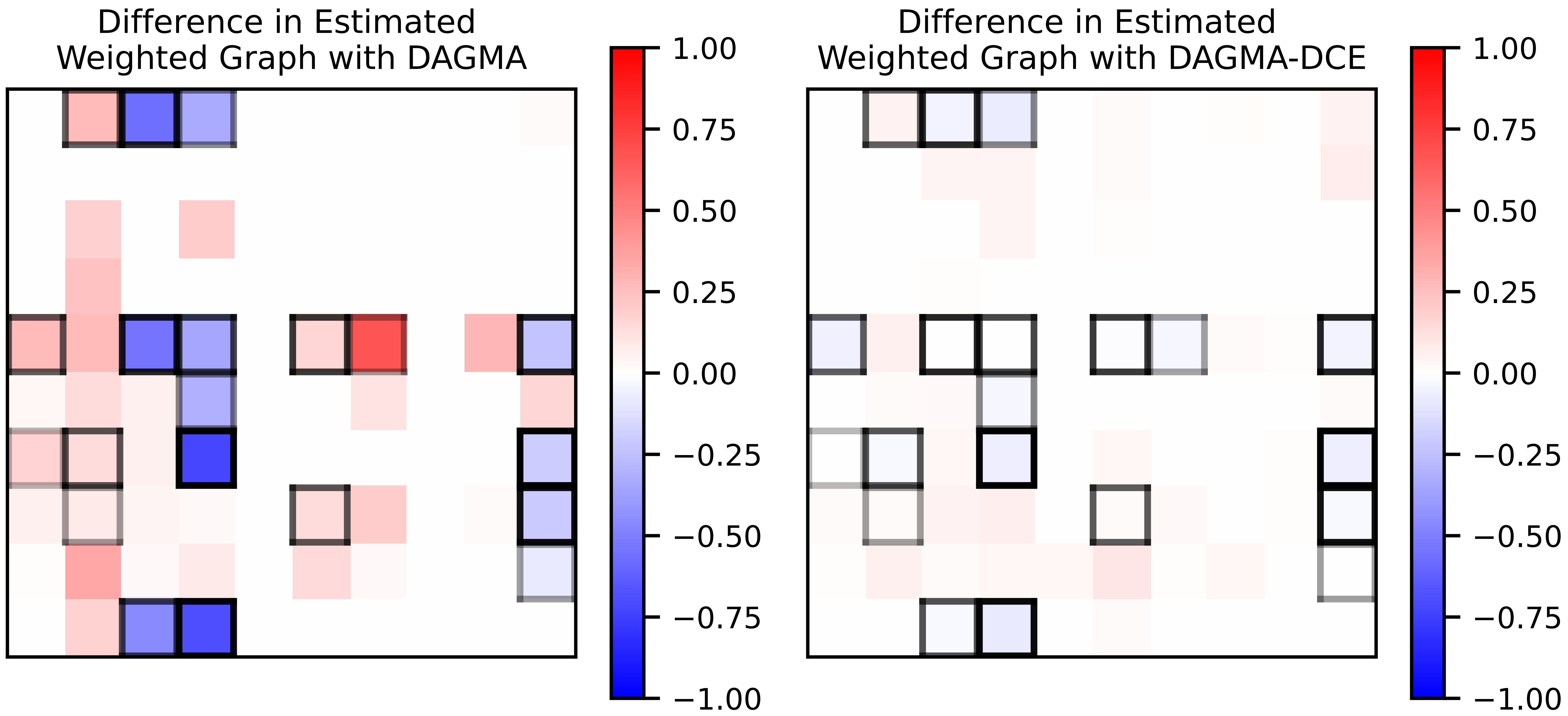
$$\text{s.t.} \; -\log\left(\det\left(s\mathbb{I} - \mathbf{A}_\theta \odot \mathbf{A}_\theta\right)\right) + d\log s = 0$$

Notably, $\|\mathbf{A}_\theta\|_1$ is different!

Whenever the DCE is well-defined and easy to compute, terms in this problem are too

DAGMA-DCE Recovers Linear Strength

Data was generated with a linear SEM

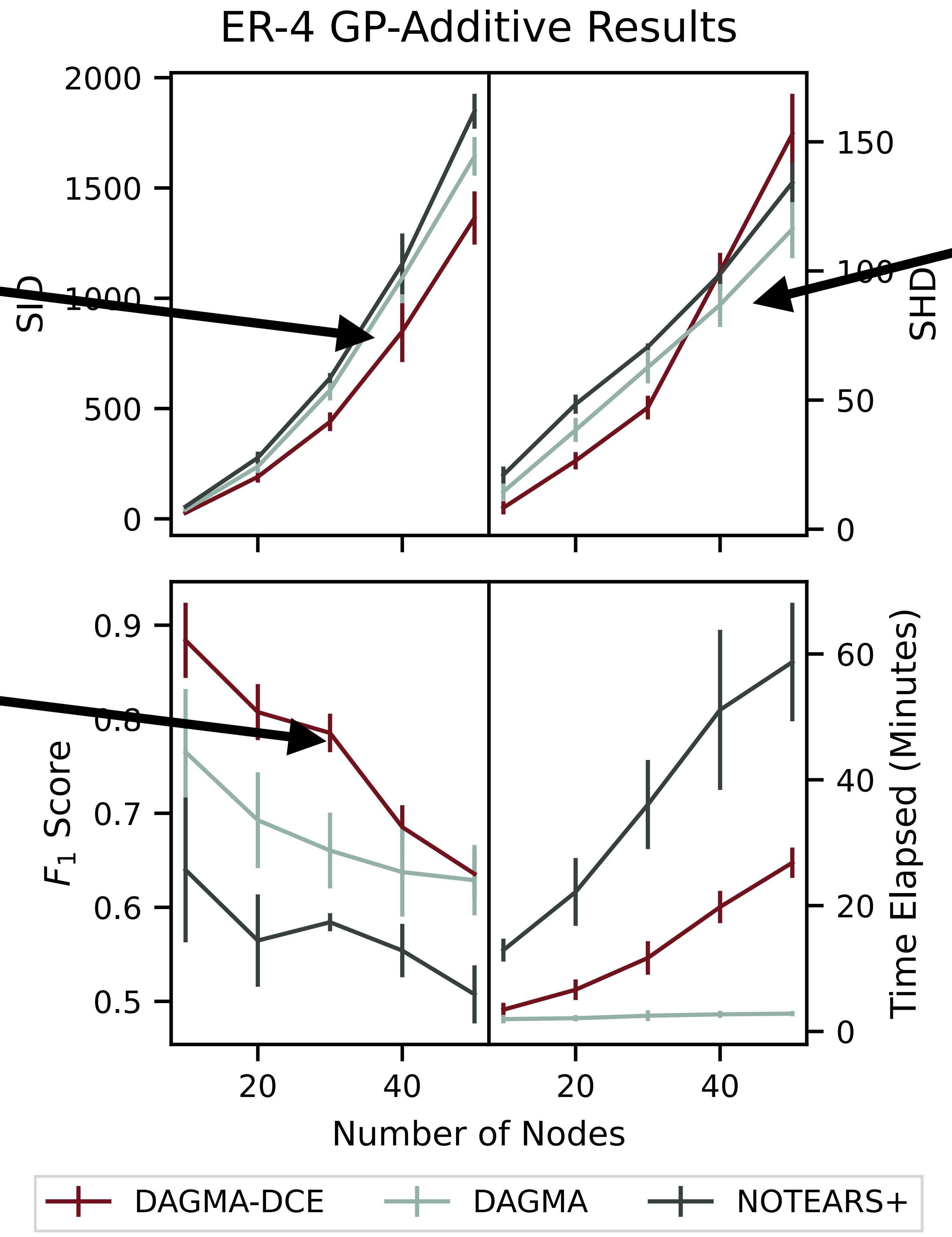


DAGMA-DCE Maintains Performance

Data was generated with additive Gaussian processes

Lower SID

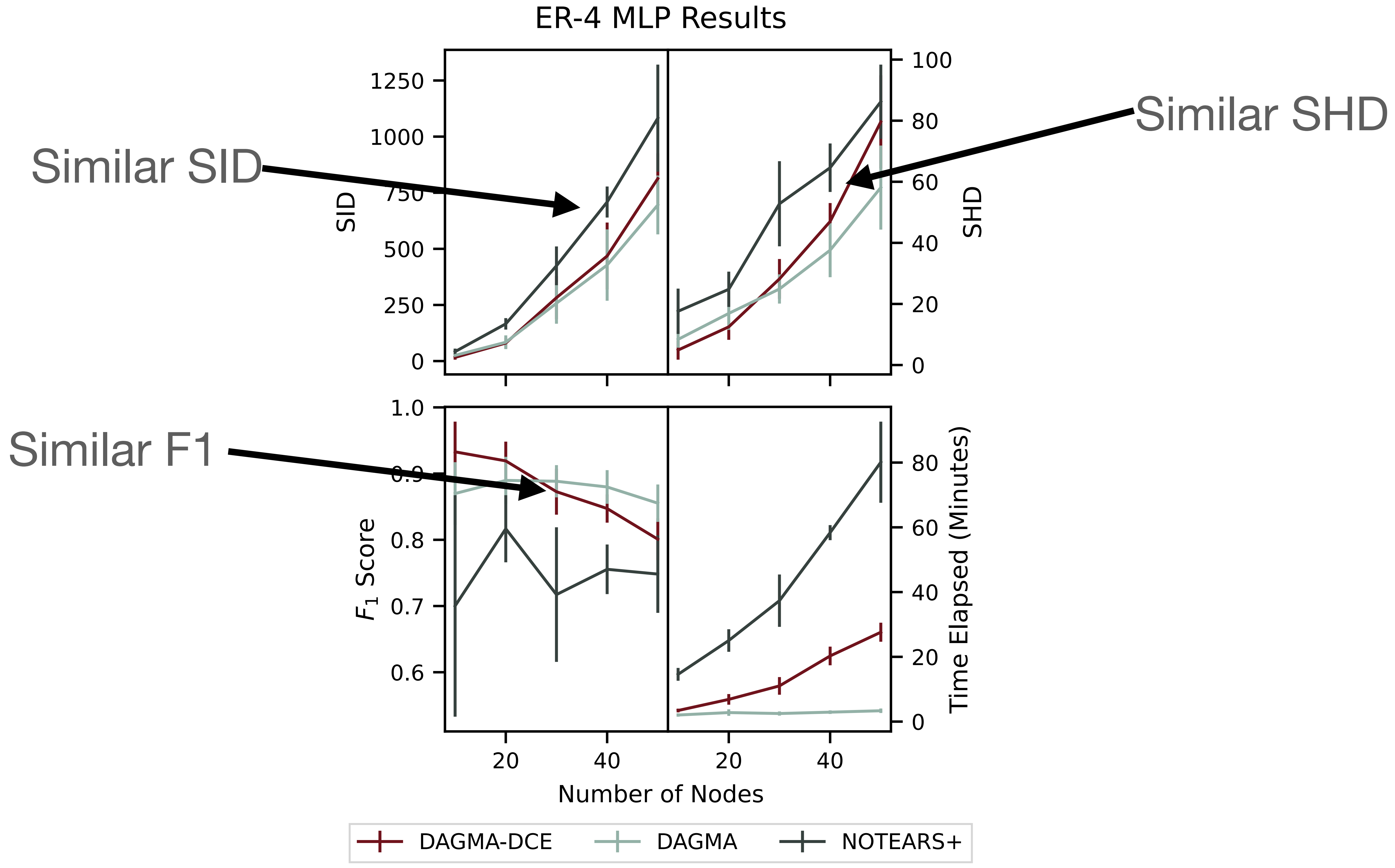
Higher F1



Similar SHD

… Even in Unfavorable Comparisons

Data was generated with MLPs, made to ensure DAGMA identifiability



DAGMA-DCE Orders Variables Differently

Dataset	Kendall's Tau	Spearman's Rho
Additive GPs	0.40 ± 0.09	0.53 ± 0.11
MLPs	0.55 ± 0.06	0.74 ± 0.07

DAGMA-DCE Orders Variables Differently

Can measure orderings with rank correlation

Dataset	Kendall's Tau	Spearman's Rho
Additive GPs	0.40 ± 0.09	0.53 ± 0.11
MLPs	0.55 ± 0.06	0.74 ± 0.07

DAGMA-DCE Orders Variables Differently

Can measure orderings with rank correlation

Both Kendall's τ and Spearman's ρ indicate different orderings

Dataset	Kendall's Tau	Spearman's Rho
Additive GPs	0.40 ± 0.09	0.53 ± 0.11
MLPs	0.55 ± 0.06	0.74 ± 0.07

DAGMA-DCE Does Principled Thresholding

DAGMA-DCE Does Principled Thresholding

One of the ad-hoc components of NOTEARS+/DAGMA was thresholding

DAGMA-DCE Does Principled Thresholding

One of the ad-hoc components of NOTEARS+/DAGMA was thresholding

DAGMA-DCE still thresholds, but the threshold is interpretable

DAGMA-DCE Does Principled Thresholding

One of the ad-hoc components of NOTEARS+/DAGMA was thresholding

DAGMA-DCE still thresholds, but the threshold is interpretable

This allows the expert to decide what’s a relevant effect

Concluding Remarks

Recap

Recap

Causal inference is important to our understanding of signals, systems, and their scientific context

Recap

Causal inference is important to our understanding of signals, systems, and their scientific context

Causal relationships have not only a direction, but a strength

Recap

Causal inference is important to our understanding of signals, systems, and their scientific context

Causal relationships have not only a direction, but a strength

Strength is often thrown away

Recap

Causal inference is important to our understanding of signals, systems, and their scientific context

Causal relationships have not only a direction, but a strength

Strength is often thrown away

By incorporating strength, we could

Recap

Causal inference is important to our understanding of signals, systems, and their scientific context

Causal relationships have not only a direction, but a strength

Strength is often thrown away

By incorporating strength, we could

- Detect confounders in multivariate time series

Recap

Causal inference is important to our understanding of signals, systems, and their scientific context

Causal relationships have not only a direction, but a strength

Strength is often thrown away

By incorporating strength, we could
- Detect confounders in multivariate time series
- Increase interpretability in differentiable causal discovery

Future

Future

Lots of other places to use ML and causal strength in causality

Future

Lots of other places to use ML and causal strength in causality

Future

Lots of other places to use ML and causal strength in causality

Interesting avenues with “hybrid” causal discovery methods

Future

Lots of other places to use ML and causal strength in causality

Interesting avenues with “hybrid” causal discovery methods

Future

Lots of other places to use ML and causal strength in causality

Interesting avenues with “hybrid” causal discovery methods

Interpretability brings opportunities for “workflows”

Future

Lots of other places to use ML and causal strength in causality

Interesting avenues with “hybrid” causal discovery methods

Interpretability brings opportunities for “workflows”

Future

Lots of other places to use ML and causal strength in causality

Interesting avenues with “hybrid” causal discovery methods

Interpretability brings opportunities for “workflows”

Together, these empower decision-makers

Thank You!

Collaborators

Petar Djurić

Kurt Butler

Chen Cui

Yuhao Liu