Causality 4 Climate

Team CoCaLa (Copenhagen Causality Lab), University of Copenhagen

December 14th, 2019, NeurIPS Competition Track Day 2

Gherardo Varando

Lasse Petersen

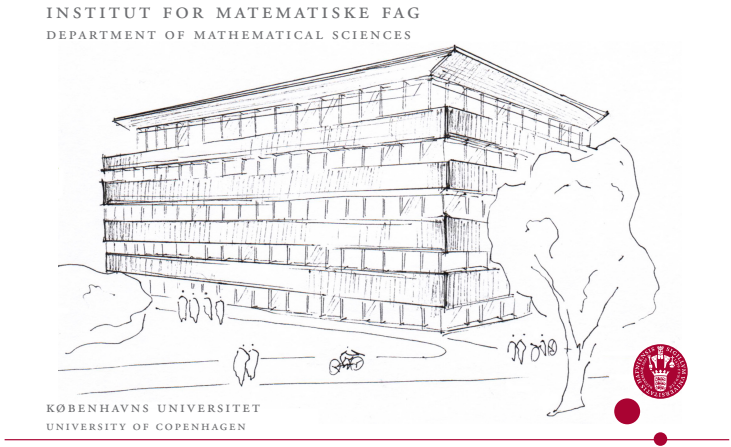
Martin E Jakobsen

Nikolaj Thams

Phillip B Mogensen

Sebastian Weichwald

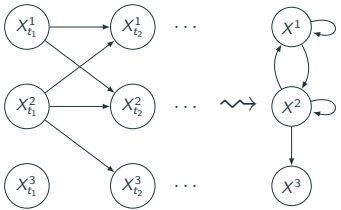
Team: 4 PhDs and 2 Postdocs from the Copenhagen Causality Lab (CoCaLa).



CoCaLa is a group of 12 people in Statistics at the University of Copenhagen doing research in causal inference.

Causal challenges in dynamical systems are different to those on iid data.

- With no unobserved process, it is well-known that structure is identifiable.
- Here: Different graphs cannot be Markov equivalent.
- This is because time helps to orient the edges.



(Meek 2014; Mogensen and Hansen 2018)

Overview of final methods

File	Description	Edgescore	Datasets
`ridge.py`	Ridge regression $\Delta X_t \approx f(X_{t-1})$. Bootstrap sampling, aggregation by quantiles.	Absolute values of regression coefficients	Climate data
`varvar.py`	OLS regression $X_t \approx f(X_{t-1,\ldots,t-\text{lags}})$ or iterative Lasso $X_t - f(X_{t-1}) \approx f(X_{t-2})$. Bootstrap sampling, random number of lags.	Absolute values of regression coefficients	Weather data, VAR models
`selvar.f`	Var model with variable and lag selection. OLS greedy search, p-values by likelihood-ratio testing.	p-value	non-linear data

Overview of final methods

File	Description	Edgescore	Datasets
`ridge.py`	Ridge regression $\Delta X_t \approx f(X_{t-1})$. Bootstrap sampling, aggregation by quantiles.	Absolute values of regression coefficients	Climate data
`varvar.py`	OLS regression $X_t \approx f(X_{t-1,\ldots,t-\text{lags}})$ or iterative Lasso $X_t - f(X_{t-1}) \approx f(X_{t-2})$. Bootstrap sampling, random number of lags.	Absolute values of regression coefficients	Weather data, VAR models
`selvar.f`	Var model with variable and lag selection. OLS greedy search, p-values by likelihood-ratio testing.	p-value	non-linear data

- We also tried: PC-type algorithms, neural-networks with input-perturbations, state space models, residual entropy estimation, ...

Ridge regression and bootstrap quantiling

Ridge and bootstrap quantiling

Target $Y_t = X_t - X_{t-1}$, **Regressor** X_{t-1}

for $i \in \{1, \dots, N\}$ **do**:

Draw bootstrap samples Y^i and X^i

Perform **Ridge** regression $Y^i \sim X^i$, obtain estimate $\hat{A}_i \in \mathbb{R}^{d \times d}$.

Collapse $\hat{A}_1, \dots, \hat{A}_N$ into one $\hat{A} \in \mathbb{R}^{d \times d}$ by entrywise taking the q^{th} quantile.

return abs$(\hat{A})$ as scores for causal links

Ridge regression and bootstrap quantiling

Ridge and bootstrap quantiling

Target $Y_t = X_t - X_{t-1}$, **Regressor** X_{t-1}

for $i \in \{1, \dots, N\}$ **do**:

Draw bootstrap samples Y^i and X^i

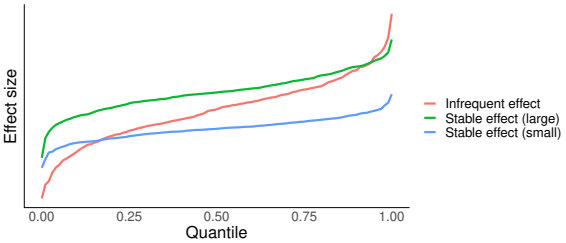
Perform **Ridge** regression $Y^i \sim X^i$, obtain estimate $\hat{A}_i \in \mathbb{R}^{d \times d}$.

Collapse $\hat{A}_1, \dots, \hat{A}_N$ into one $\hat{A} \in \mathbb{R}^{d \times d}$ by entrywise taking the q^{th} quantile.

return abs($\hat{A}$) as scores for causal links

Linear regression simulation study

- Choosing q large corresponds to searching for largest *local* effect.
- Choosing q small corresponds to large *minimal* effect.



Correlation does not imply causation.

Correlation does not imply causation.

What about covariance?

- Consider iid observations from the well-known linear acyclic model

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{bmatrix} = \begin{bmatrix} 0 & 0 & \dots & 0 \\ b_{21} & 0 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ b_{d1} & b_{d2} & \dots & 0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{bmatrix} + \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_d \end{bmatrix}$$

  where $E \sim \mathcal{N}(0, \sigma^2 I)$.

- Consider iid observations from the well-known linear acyclic model

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{bmatrix} = \begin{bmatrix} 0 & 0 & \dots & 0 \\ b_{21} & 0 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ b_{d1} & b_{d2} & \dots & 0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{bmatrix} + \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_d \end{bmatrix}$$

  where $E \sim \mathcal{N}(0, \sigma^2 I)$.

- Regress X_j onto X_i, for all pairs $j \neq i$, and obtain the OLS regression coefficient

$$\widehat{b}_{i \to j} = \frac{\widehat{\text{cov}}(X_i, X_j)}{\widehat{\text{cov}}(X_i, X_i)}.$$

- Consider iid observations from the well-known linear acyclic model

$$\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{bmatrix} = \begin{bmatrix} 0 & 0 & \dots & 0 \\ b_{21} & 0 & \dots & 0 \\ \vdots & & \ddots & \vdots \\ b_{d1} & b_{d2} & \dots & 0 \end{bmatrix} \begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_d \end{bmatrix} + \begin{bmatrix} E_1 \\ E_2 \\ \vdots \\ E_d \end{bmatrix}$$

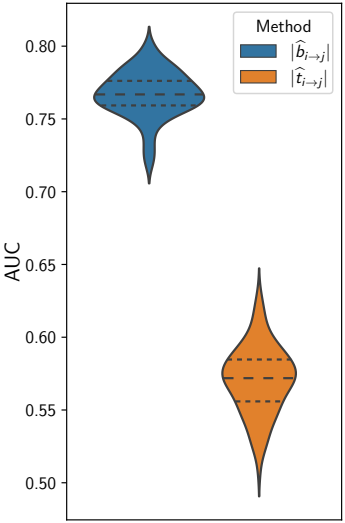
  where $E \sim \mathcal{N}(0, \sigma^2 I)$.

- Regress X_j onto X_i, for all pairs $j \neq i$, and obtain the OLS regression coefficient

$$\widehat{b}_{i \to j} = \frac{\widehat{\mathrm{cov}}(X_i, X_j)}{\widehat{\mathrm{cov}}(X_i, X_i)}.$$

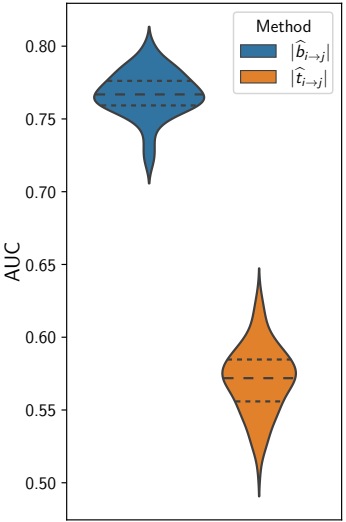
- Compare the AUC when scoring edges $X_i \to X_j$ either by the
  (a) absolute regression coefficients $|\widehat{b}_{i \to j}|$, or by the
  (b) corresponding absolute t-test statistics $|\widehat{t}_{i \to j}|$.

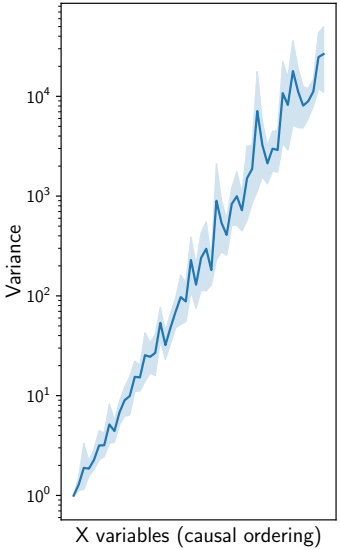
Equal error variance



$d = 50, n = 200, p_0 = 0.75, 100$ repetitions

Equal error variance





$d = 50, n = 200, p_0 = 0.75, 100$ repetitions

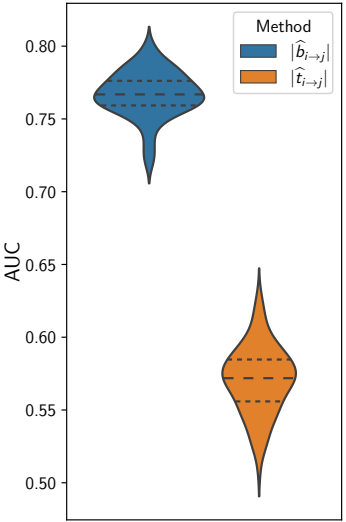
From absolute coefficients to t-test statistics

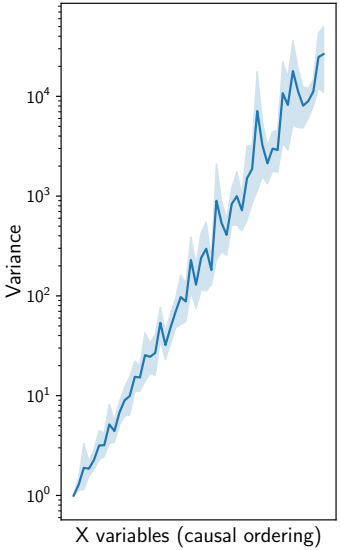
- $|\widehat{b}_{i\to j}|$
- $|\widehat{t}_{i\to j}| = |\widehat{b}_{i\to j}| \sqrt{\frac{\widehat{\text{var}}(X_i)}{\widehat{\text{var}}(X_j)}} \sqrt{\frac{(n-2)}{(1-\widehat{\rho}^2_{X_i,X_j})}}$ (pairwise regression)

From absolute coefficients to t-test statistics

- $|\widehat{b}_{i\to j}|$
- $|\widehat{t}_{i\to j}| = |\widehat{b}_{i\to j}| \sqrt{\frac{\widehat{\text{var}}(X_i)}{\widehat{\text{var}}(X_j)}} \sqrt{\frac{(n-2)}{(1-\widehat{\rho}^2_{X_i,X_j})}}$ (pairwise regression)
- $|\widehat{t}_{i\to j}| = |\widehat{b}_{i\to j}| \frac{\widehat{\text{sre}}(X_i|X_{\setminus i})}{\widehat{\text{sre}}(X_j|X_{\setminus j})} \sqrt{\frac{(n-d)}{(1-\widehat{\rho}^2_{X_i,X_j|X_{\setminus\{i,j\}}})}}$ (multivariate regression)

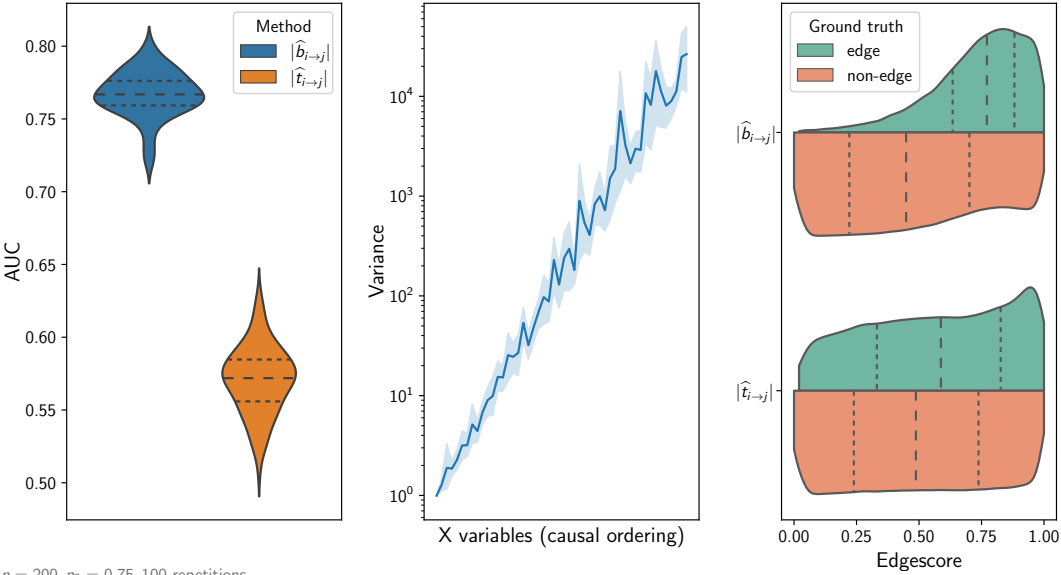
Equal error variance





$d = 50, n = 200, p_0 = 0.75, 100$ repetitions

Equal error variance



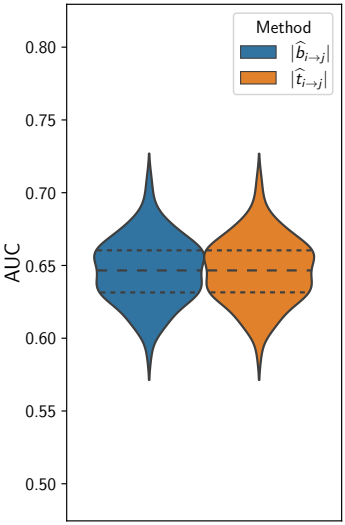
$d = 50, n = 200, p_0 = 0.75, 100$ repetitions

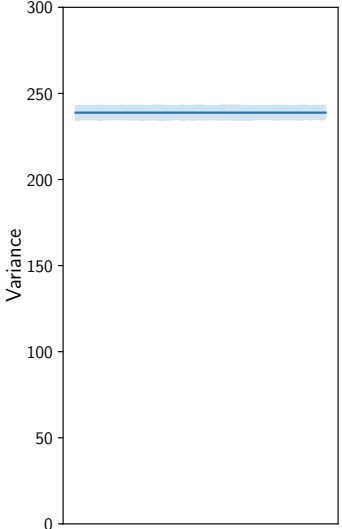
- Consider, as before, $X = BX + E$ with $E \sim \mathcal{N}(0, \text{diag}(\sigma_1^2, \ldots, \sigma_d^2))$.
- This time *(somewhat artificially)* ensure equal marginal variances

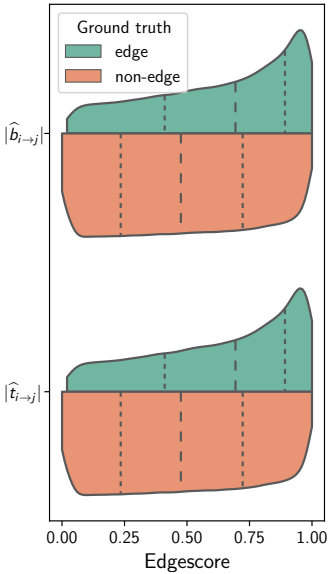
$$\text{var}(X_i) = c, \ \ i = 1, \ldots, d,$$

by rescaling rows of B and the σ_i^2 appropriately.

Equal marginal variance







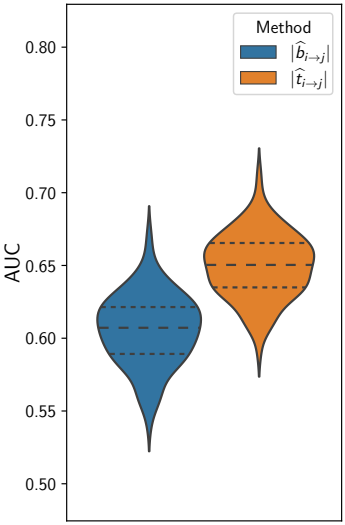
$d = 50, n = 200, p_0 = 0.75, 100$ repetitions

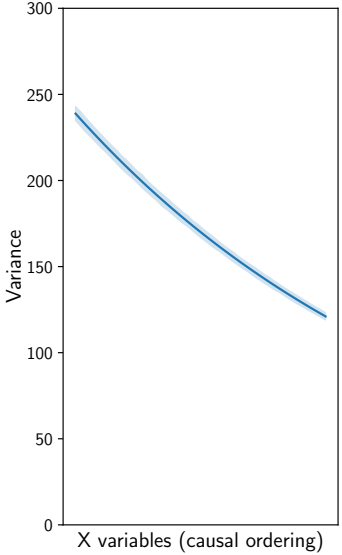
- This time *(somewhat artificially)* ensure decreasing marginal variances,

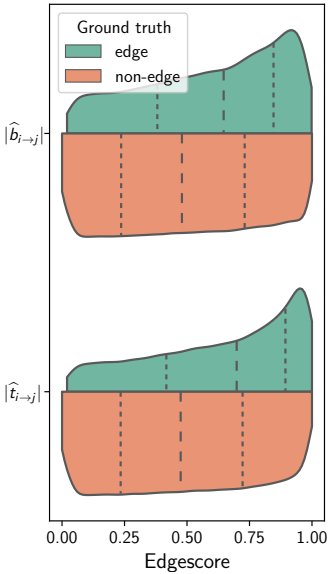
$$\forall i < j : \quad \text{var}(X_i) \geq \text{var}(X_j),$$

by rescaling rows of B and the σ_i^2 appropriately.

Decreasing marginal variance







$d = 50, n = 200, p_0 = 0.75, 100$ repetitions

Correlation does not imply causation.

What about covariance?

Correlation does not imply causation.

What about covariance?

Sometimes covariance or absolute regression coefficients may rank causal links well.

Correlation does not imply causation.

What about covariance?

Sometimes covariance or absolute regression coefficients may rank causal links well.

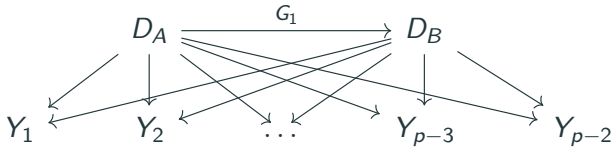
- What does it mean for two causal systems to be ‘close‘ to one another?

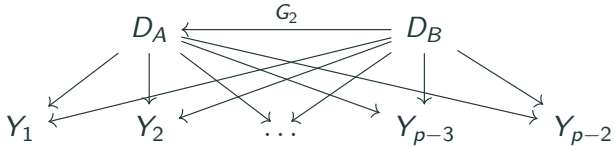
- What does it mean for two causal systems to be ‘close‘ to one another?
- In the competition, AUC was used to measure closeness of graphs.

- What does it mean for two causal systems to be ‘close‘ to one another?
- In the competition, AUC was used to measure closeness of graphs.
- For inferring causal structures, we argue that it may not be sufficient to consider the AUC.

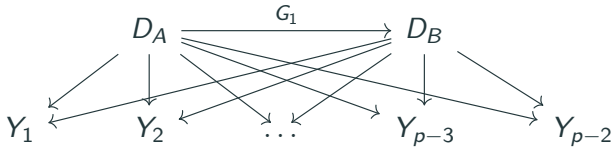
- We consider two graphs with binary adjacency matrices:

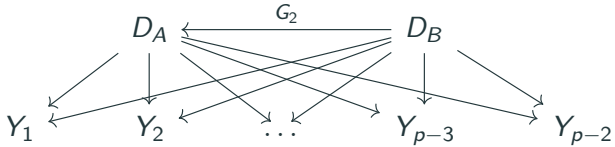
- We consider two graphs with binary adjacency matrices:





- We consider two graphs with binary adjacency matrices:



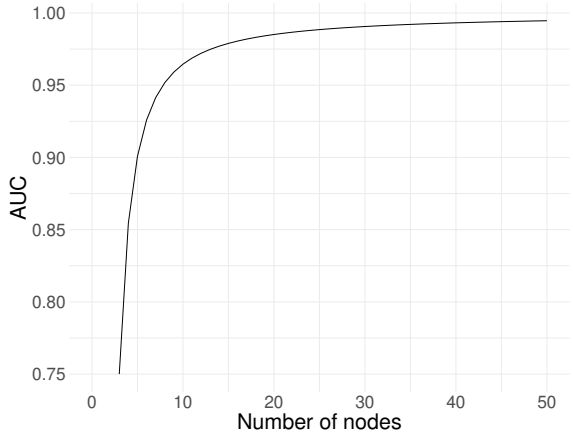


- Only difference: Flipped edge between the two drivers.

- The Structural Hamming Distance (SHD) between the two is constant at 1 or 2.

- The Structural Hamming Distance (SHD) between the two is constant at 1 or 2.
- The AUC (as calculated in the competition) increases quickly as p grows.

- The Structural Hamming Distance (SHD) between the two is constant at 1 or 2.
- The AUC (as calculated in the competition) increases quickly as p grows.



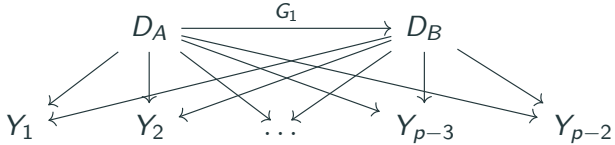
- The Structural Hamming Distance (SHD) between the two is constant at 1 or 2.
- The AUC (as calculated in the competition) increases quickly as p grows.
- The number of incorrectly inferred interventional distributions (SID, see Peters and Bühlmann 2015) is $2(p-1)$.

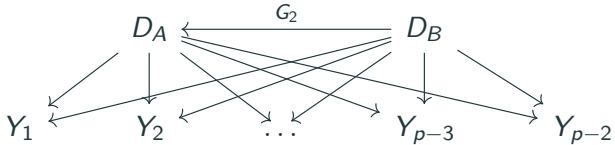
- The Structural Hamming Distance (SHD) between the two is constant at 1 or 2.
- The AUC (as calculated in the competition) increases quickly as p grows.
- The number of incorrectly inferred interventional distributions (SID, see Peters and Bühlmann 2015) is $2(p-1)$.
- The following distributions are incorrectly inferred (for fixed $c \in \mathbb{R}$):

$$\forall k \in \{1, \dots, p-2\}: \quad \mathbb{P}_{D_A}^{\mathrm{do}(D_B=c)}, \quad \mathbb{P}_{D_B}^{\mathrm{do}(D_A=c)}, \quad \mathbb{P}_{Y_k}^{\mathrm{do}(D_A=c)}, \quad \mathbb{P}_{Y_k}^{\mathrm{do}(D_B=c)}$$

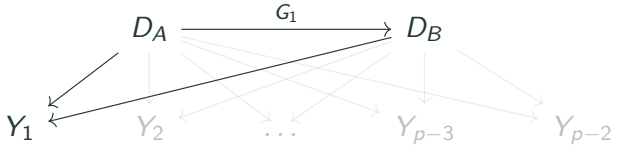
- Consider the case with $p = 3$, linear assignments and gaussian noise. The 'true' model is G_1.

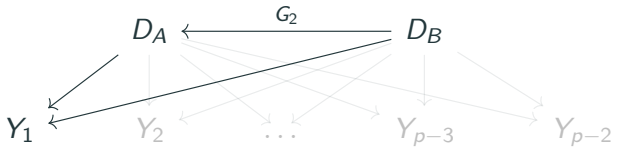
- Consider the case with $p = 3$, linear assignments and gaussian noise. The 'true' model is G_1.





- Consider the case with $p = 3$, linear assignments and gaussian noise. The 'true' model is G_1.



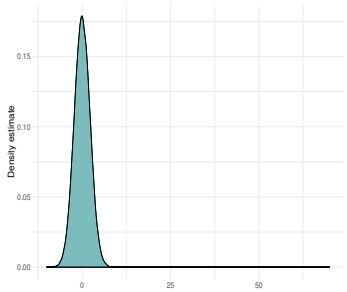


- Consider the case with $p = 3$, linear assignments and gaussian noise. The 'true' model is G_1.

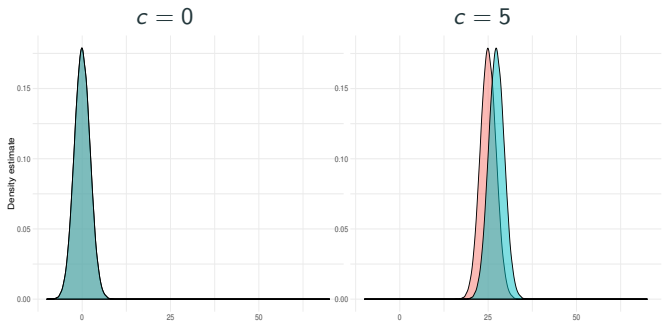
- Consider the case with $p = 3$, linear assignments and gaussian noise. The 'true' model is G_1.
- The distribution of Y under $\text{do}(D_B = c)$ depends on the graph it is calculated under.

Estimated densities of $\mathbb{P}_Y^{\mathrm{do}(D_B=c)}$

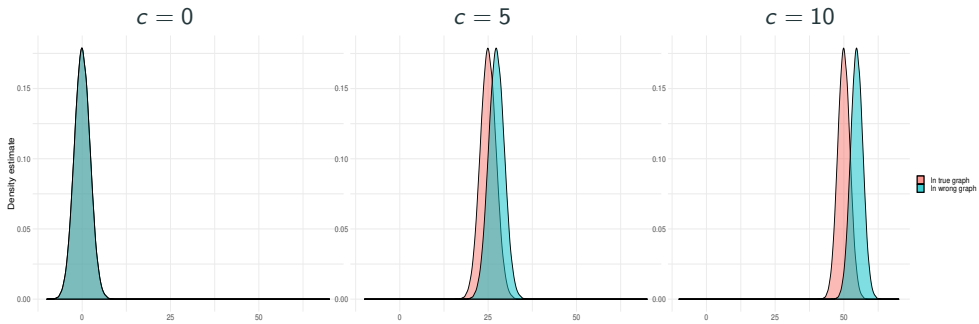
$c = 0$



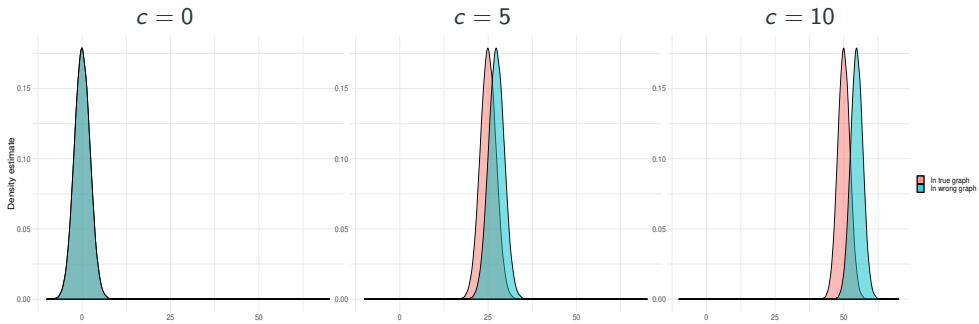
Estimated densities of $\mathbb{P}_Y^{\mathrm{do}(D_B=c)}$



Estimated densities of $\mathbb{P}_Y^{\mathrm{do}(D_B=c)}$



Estimated densities of $\mathbb{P}_Y^{\mathrm{do}(D_B=c)}$



- The estimated distribution can be arbitrarily wrong with just a single edge flipped.

- Open question: What does it mean for two causal systems to be 'close' to one another?
- In the competition, AUC was used to measure 'closeness' of graphs.
- For inferring causal structures, we argue that it may not be sufficient to consider the AUC.

- Open question: What does it mean for two causal systems to be 'close' to one another?
- In the competition, AUC was used to measure 'closeness' of graphs.
- For inferring causal structures, we argue that it may not be sufficient to consider the AUC.
- Take-home message: Inferring causal structures is difficult, but:

- Open question: What does it mean for two causal systems to be 'close' to one another?
- In the competition, AUC was used to measure 'closeness' of graphs.
- For inferring causal structures, we argue that it may not be sufficient to consider the AUC.
- Take-home message: Inferring causal structures is difficult, but:
  - If you are interested in an overall structure, the AUC is a good measure.

- Open question: What does it mean for two causal systems to be 'close' to one another?
- In the competition, AUC was used to measure 'closeness' of graphs.
- For inferring causal structures, we argue that it may not be sufficient to consider the AUC.
- Take-home message: Inferring causal structures is difficult, but:
  - If you are interested in an overall structure, the AUC is a good measure.
  - If you care about interventional distributions, it is not sufficient to consider only the AUC.

Final remarks

- Linear regression can beat ‘causality tailored’ and non-linear approaches.
- When is covariance a good indicator of causality?
- The task is only as causal as the score function!

Final remarks

- Linear regression can beat 'causality tailored' and non-linear approaches.
- When is covariance a good indicator of causality?
- The task is only as causal as the score function!

Thanks to the organizers!

CoCaLa: `https://www.math.ku.dk/english/research/spt/cocala/`

Code: `https://github.com/sweichwald/CoCaLa-CauseMe-NeurIPS-competition`

✈ Thanks also for accomodating our telepresence which saved $\sim$28500 kg in CO_2 emissions.