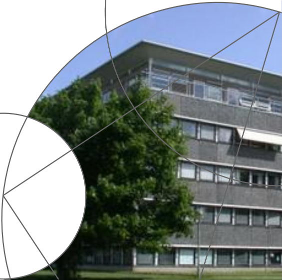
Faculty of Science

Greedy Learning of Causal Structures in Additive Noise Models

Master thesis defense

Phillip Bredahl Mogensen
Department of Mathematics

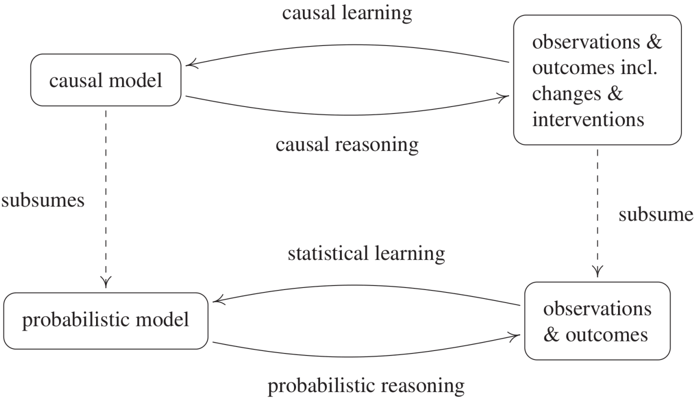


Agenda

1. Causal discovery and Additive Noise Models
2. Entropy scores and the Greedy entropy-search
3. A few simulations and real data
4. Conclusion

Causal discovery and Additive Noise Models

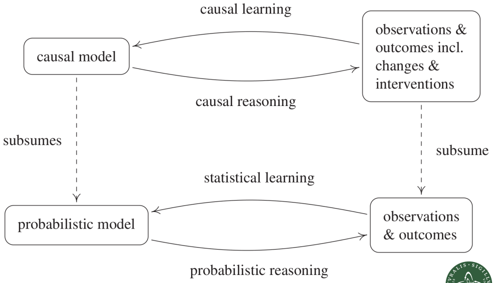
Aim of the thesis



Source: Peters et al. 2017, Page 6.

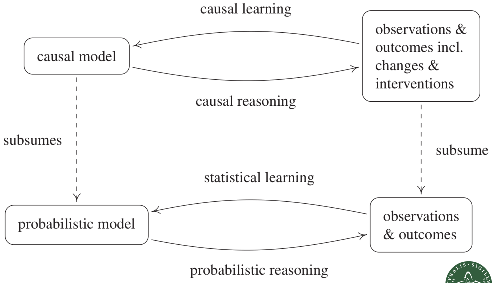
Aim of the thesis

- Population case, i.e. known distribution $\mathbb{P}$.



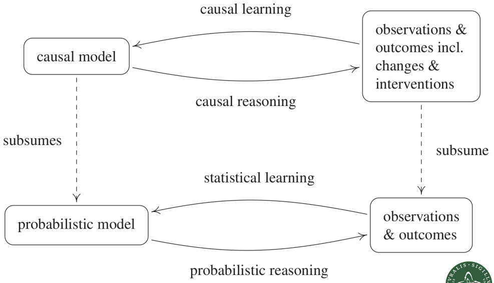
Aim of the thesis

- Population case, i.e. known distribution $\mathbb{P}$.
- We want to recover the graph of the causal model that generated $\mathbb{P}$.



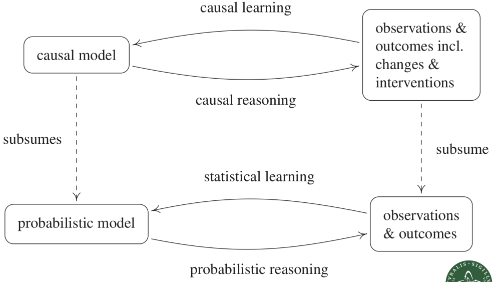
Aim of the thesis

- Population case, i.e. known distribution $\mathbb{P}$.
- We want to recover the graph of the causal model that generated $\mathbb{P}$. Two main problems:



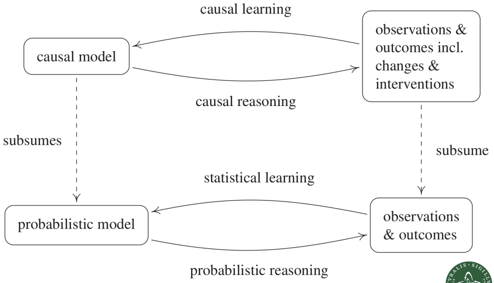
Aim of the thesis

- Population case, i.e. known distribution $\mathbb{P}$.
- We want to recover the graph of the causal model that generated $\mathbb{P}$. Two main problems:
  - Identifiability.



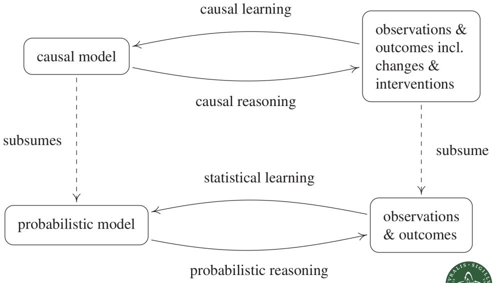
Aim of the thesis

- Population case, i.e. known distribution $\mathbb{P}$.
- We want to recover the graph of the causal model that generated $\mathbb{P}$. Two main problems:
  - Identifiability.
  - Computability.



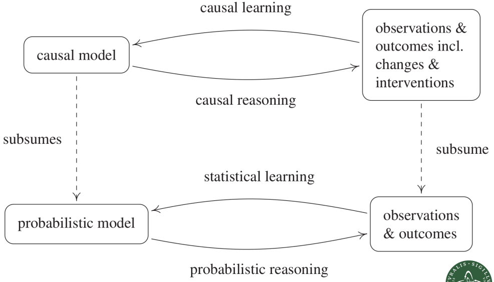
Aim of the thesis

- Population case, i.e. known distribution $\mathbb{P}$.
- We want to recover the graph of the causal model that generated $\mathbb{P}$. Two main problems:
  - Identifiability.
  - Computability.
- Additive Noise Models.



Aim of the thesis

- Population case, i.e. known distribution $\mathbb{P}$.
- We want to recover the graph of the causal model that generated $\mathbb{P}$. Two main problems:
  - Identifiability.
  - Computability.
- Additive Noise Models.
- A score-based Greedy search.



Additive Noise Models

Definition 1

An Additive Noise Model (ANM), $\mathscr{C}$, is a collection of assignments, $\mathbf{S}$, and mutually independent noise distributions. An ANM has an associated Directed Acyclic Graph (DAG).

Additive Noise Models

Definition 1

An Additive Noise Model (ANM), $\mathscr{C}$, is a collection of assignments, **S**, and mutually independent noise distributions. An ANM has an associated Directed Acyclic Graph (DAG).

$$S_\nu := \left\{ X_\nu := \sum_{\gamma \in \mathbf{PA}_{\mathcal{G}}(X_\nu)} f^0_{\nu,\gamma}(X_\gamma) + N_\nu \right\}$$

Additive Noise Models

Definition 1

An Additive Noise Model (ANM), $\mathscr{C}$, is a collection of assignments, $\mathbf{S}$, and mutually independent noise distributions. An ANM has an associated Directed Acyclic Graph (DAG).

$$\boxed{S_\nu := \left\{ X_\nu := \sum_{\gamma \in \mathbf{PA}_{\mathcal{G}}(X_\nu)} f^0_{\nu,\gamma}(X_\gamma) + N_\nu \right\}}$$

- The functions $f^0_{\cdot,\cdot}$ belong to a class called $\mathcal{F}$.

Additive Noise Models

Definition 1

An Additive Noise Model (ANM), $\mathscr{C}$, is a collection of assignments, $\mathbf{S}$, and mutually independent noise distributions. An ANM has an associated Directed Acyclic Graph (DAG).

$$S_\nu := \left\{ X_\nu := \sum_{\gamma \in \mathbf{PA}_{\mathcal{G}}(X_\nu)} f^0_{\nu,\gamma}(X_\gamma) + N_\nu \right\}$$

- The functions $f^0_{\cdot,\cdot}$ belong to a class called $\mathcal{F}$.
- The ANM $\mathscr{C}$ implies a unique, joint distribution, $\mathbb{P}$.

Additive Noise Models

Definition 1

An Additive Noise Model (ANM), $\mathscr{C}$, is a collection of assignments, $\mathbf{S}$, and mutually independent noise distributions. An ANM has an associated Directed Acyclic Graph (DAG).

$$S_\nu := \left\{ X_\nu := \sum_{\gamma \in \mathbf{PA}_{\mathcal{G}}(X_\nu)} f^0_{\nu,\gamma}(X_\gamma) + N_\nu \right\}$$

- The functions $f^0_{\cdot,\cdot}$ belong to a class called $\mathcal{F}$.
- The ANM $\mathscr{C}$ implies a unique, joint distribution, $\mathbb{P}$.
- We call $\mathscr{C}$ identifiable when $\mathbb{P}$ can *only* be implied by $\mathscr{C}$.

Additive Noise Models

Definition 1

An Additive Noise Model (ANM), $\mathscr{C}$, is a collection of assignments, **S**, and mutually independent noise distributions. An ANM has an associated Directed Acyclic Graph (DAG).

$$S_\nu := \left\{ X_\nu := \sum_{\gamma \in \mathbf{PA}_{\mathcal{G}}(X_\nu)} f^0_{\nu,\gamma}(X_\gamma) + N_\nu \right\}$$

- The functions $f^0_{\cdot,\cdot}$ belong to a class called $\mathcal{F}$.
- The ANM $\mathscr{C}$ implies a unique, joint distribution, $\mathbb{P}$.
- We call $\mathscr{C}$ identifiable when $\mathbb{P}$ can *only* be implied by $\mathscr{C}$.
- To get identifiability, we lean on existing results

Identifiability of ANMs – continued

We assume that:

Identifiability of ANMs – continued

We assume that:

(A1,2) $\mathcal{F}$ consists of non-linear C^3 functions + some regularity conditions.

(A3) The densities of the noise variables have only discretely many solutions to the differential equation $(\log f)'' = 0$.

(A4) The noise variables have full support and their densities are in C^3 and strictly positive.

(A5,6) All noise variables and all $\mathcal{F}$-transformations of them have second moment.

Identifiability of ANMs – continued

We assume that:

(A1,2) $\mathcal{F}$ consists of non-linear C^3 functions + some regularity conditions.

(A3) The densities of the noise variables have only discretely many solutions to the differential equation $(\log f)'' = 0$.

(A4) The noise variables have full support and their densities are in C^3 and strictly positive.

(A5,6) All noise variables and all $\mathcal{F}$-transformations of them have second moment.

In short: 'Nice', nonlinear functions + 'nice' noise variables give identifiability.

Identifiability of ANMs – continued

We assume that:

(A1,2) $\mathcal{F}$ consists of non-linear C^3 functions + some regularity conditions.

(A3) The densities of the noise variables have only discretely many solutions to the differential equation $(\log f)'' = 0$.

(A4) The noise variables have full support and their densities are in C^3 and strictly positive.

(A5,6) All noise variables and all $\mathcal{F}$-transformations of them have second moment.

In short: 'Nice', nonlinear functions + 'nice' noise variables give identifiability.

Entropy scores and the Greedy entropy-search

Set-up

Suppose we know the distribution of an identifiable model, $\mathscr{C}$, with true graph $\mathcal{G}^0$.

Set-up

Suppose we know the distribution of an identifiable model, $\mathscr{C}$, with true graph $\mathcal{G}^0$.

Q: How do we rank candidate graphs?

Set-up

Suppose we know the distribution of an identifiable model, $\mathscr{C}$, with true graph $\mathcal{G}^0$.

Q: How do we rank candidate graphs?

A: Entropy scores:

Set-up

Suppose we know the distribution of an identifiable model, $\mathscr{C}$, with true graph $\mathcal{G}^0$.

Q: How do we rank candidate graphs?

A: Entropy scores:

Definition 2

The entropy score of a graph $\mathcal{G}$ under $\mathscr{C}$ is

$$\ell(\mathcal{G}) := -\sum_{\nu \in V(\mathcal{G})} \mathbb{H}\left(X_\nu - \sum_{\gamma \in \mathbf{PA}_{\mathcal{G}}(\nu)} \hat{f}_{\nu,\gamma}(X_\gamma)\right).$$

Set-up

Suppose we know the distribution of an identifiable model, $\mathscr{C}$, with true graph $\mathcal{G}^0$.

Q: How do we rank candidate graphs?

A: Entropy scores:

Definition 2

The entropy score of a graph $\mathcal{G}$ under $\mathscr{C}$ is

$$\ell(\mathcal{G}) := -\sum_{\nu \in V(\mathcal{G})} \mathbb{H}\left(X_\nu - \sum_{\gamma \in \mathbf{PA}_{\mathcal{G}}(\nu)} \hat{f}_{\nu,\gamma}(X_\gamma)\right).$$

$$= \underset{\mathcal{F} \cup \mathcal{C}}{\arg\min}\, \mathbb{E}\left(X_\nu - \sum_{\gamma \in \mathbf{PA}_{\mathcal{G}}(\nu)} f_{\nu,\gamma}(X_\gamma)\right)^2$$

Set-up

Suppose we know the distribution of an identifiable model, $\mathscr{C}$, with true graph $\mathcal{G}^0$.

Q: How do we rank candidate graphs?

A: Entropy scores:

Definition 2

The entropy score of a graph $\mathcal{G}$ under $\mathscr{C}$ is

$$\ell(\mathcal{G}) := -\sum_{\nu \in V(\mathcal{G})} \mathbb{H}\left(X_\nu - \sum_{\gamma \in \mathbf{PA}_{\mathcal{G}}(\nu)} \hat{f}_{\nu,\gamma}(X_\gamma) \right).$$

$$= \underset{\mathcal{F} \cup \mathcal{C}}{\arg\min}\, \mathrm{MSE}$$

Set-up

Suppose we know the distribution of an identifiable model, $\mathscr{C}$, with true graph $\mathcal{G}^0$.

Q: How do we rank candidate graphs?

A: Entropy scores:

Definition 2

The entropy score of a graph $\mathcal{G}$ under $\mathscr{C}$ is

$$\ell(\mathcal{G}) := -\sum_{\nu \in V(\mathcal{G})} \mathbb{H}\left(\text{residuals of } X_\nu\right).$$

Set-up

Suppose we know the distribution of an identifiable model, $\mathscr{C}$, with true graph $\mathcal{G}^0$.

Q: How do we rank candidate graphs?

A: Entropy scores:

Definition 2

The entropy score of a graph $\mathcal{G}$ under $\mathscr{C}$ is

$$\ell(\mathcal{G}) := -\sum_{\nu \in V(\mathcal{G})} \mathbb{H}\left(\text{residuals of } X_\nu\right).$$

Q: Why this score function?

Theorem 3

If $\mathcal{G}^0$ is the true graph of an ANM that satisfies (A1)–(A6), then

$$\mathcal{G}^0 = \arg\max_{\mathcal{G}} \ell(\mathcal{G}).$$

Theorem 3

If $\mathcal{G}^0$ is the true graph of an ANM that satisfies (A1)–(A6), then

$$\mathcal{G}^0 = \arg\max_{\mathcal{G}} \ell(\mathcal{G}).$$

- **Implication:** To find $\mathcal{G}^0$, just maximize ℓ.

Theorem 3

If $\mathcal{G}^0$ is the true graph of an ANM that satisfies (A1)–(A6), then

$$\mathcal{G}^0 = \arg\max_{\mathcal{G}} \ell(\mathcal{G}).$$

- **Implication:** To find $\mathcal{G}^0$, just maximize ℓ.
- We just check every possible graph!

The problem

How many graphs do we need to check if we have p variables?

[1] Calculations based on McKay et al. [2003]

The problem

How many graphs do we need to check if we have p variables?

Number of graphs[1]	p
1	1

[1] Calculations based on McKay et al. [2003]

The problem

How many graphs do we need to check if we have p variables?

Number of graphs[1]	p
1	1
3	2

[1] Calculations based on McKay et al. [2003]

The problem

How many graphs do we need to check if we have p variables?

Number of graphs[1]	p
1	1
3	2
25	3

[1] Calculations based on McKay et al. [2003]

The problem

How many graphs do we need to check if we have p variables?

Number of graphs[1]	p
1	1
3	2
25	3
543	4

[1] Calculations based on McKay et al. [2003]

The problem

How many graphs do we need to check if we have p variables?

Number of graphs[1]	p
1	1
3	2
25	3
543	4
$\vdots$	$\vdots$

[1] Calculations based on McKay et al. [2003]

The problem

How many graphs do we need to check if we have p variables?

Number of graphs[1]	p
1	1
3	2
25	3
543	4
$\vdots$	$\vdots$
2377252655534104384260462682682226888620 26	15

[1] Calculations based on McKay et al. [2003]

The problem

How many graphs do we need to check if we have p variables?

Number of graphs[1]	p
1	1
3	2
25	3
543	4
$\vdots$	$\vdots$
237725265553410438426046268268222688862026	15

Too many! Computationally an impossible task.

[1] Calculations based on McKay et al. [2003]

The problem

How many graphs do we need to check if we have p variables?

Number of graphs[1]	p
1	1
3	2
25	3
543	4
$\vdots$	$\vdots$
237725265553410438426046268268222688862026	15

Too many! Computationally an impossible task.
Instead, we do a Greedy search.

[1] Calculations based on McKay et al. [2003]

Getting to optimality – prerequisites

A few technicalities

- Introduce the Gaussian score:

$$\ell^{g}(\mathcal{G}) := -\sum_{\nu \in V(\mathcal{G})} \log \mathbb{V}\left(\text{residuals of } X_{\nu}\right)$$

Getting to optimality – prerequisites

A few technicalities

- Introduce the Gaussian score:

$$\ell^{g}(\mathcal{G}) := - \sum_{\nu \in V(\mathcal{G})} \log \mathbb{V}\left(\text{residuals of } X_\nu\right)$$

- Intuition: A worst-case-scenario entropy score.

Getting to optimality – prerequisites

A few technicalities

- Introduce the Gaussian score:

$$\ell^{g}(\mathcal{G}) := -\sum_{\nu \in V(\mathcal{G})} \log \mathbb{V}\left(\text{residuals of } X_\nu\right)$$

- Intuition: A worst-case-scenario entropy score.
- Restrict attention to unrelated graphs.

Getting to optimality – prerequisites

A few technicalities

- Introduce the Gaussian score:

$$\ell^{g}(\mathcal{G}) := -\sum_{\nu \in V(\mathcal{G})} \log \mathbb{V}\left(\text{residuals of } X_{\nu}\right)$$

- Intuition: A worst-case-scenario entropy score.
- Restrict attention to unrelated graphs.

All parents are mutually independent

Getting to optimality – prerequisites

A few technicalities

- Introduce the Gaussian score:

$$\ell^{g}(\mathcal{G}) := -\sum_{\nu \in V(\mathcal{G})} \log \mathbb{V}\,(\text{residuals of } X_\nu)$$

- Intuition: A worst-case-scenario entropy score.
- Restrict attention to unrelated graphs.

All cycles have at least three colliders

The Greedy entropy-search

Input: Graph $\mathcal{G}$, distribution $\mathbb{P}$

The Greedy entropy-search

Input: Graph $\mathcal{G}$, distribution $\mathbb{P}$
Select $(\alpha - \beta)$ that maximizes $\Delta\ell^{g}$

The Greedy entropy-search

Input: Graph $\mathcal{G}$, distribution $\mathbb{P}$
Select $(\alpha - \beta)$ that maximizes $\Delta\ell^{g}$
if $\Delta\ell^{g}$ *is negative* **then**
| **return** $\mathcal{G}$
end

The Greedy entropy-search

Input: Graph $\mathcal{G}$, distribution $\mathbb{P}$
Select $(\alpha - \beta)$ that maximizes $\Delta\ell^g$
if $\Delta\ell^g$ *is negative* **then**
  return $\mathcal{G}$
end
else
  Direct $(\alpha - \beta)$ by maximizing $\Delta\ell$.
end

The Greedy entropy-search

Input: Graph $\mathcal{G}$, distribution $\mathbb{P}$
Select $(\alpha - \beta)$ that maximizes $\Delta\ell^{g}$
if $\Delta\ell^{g}$ *is negative* **then**
  return $\mathcal{G}$
end
else
  Direct $(\alpha - \beta)$ by maximizing $\Delta\ell$.
end
Add directed edge to $\mathcal{G}$

The Greedy entropy-search

Input: Graph $\mathcal{G}$, distribution $\mathbb{P}$
Select $(\alpha - \beta)$ that maximizes $\Delta\ell^{g}$
if $\Delta\ell^{g}$ *is negative* **then**
| **return** $\mathcal{G}$
end
else
| Direct $(\alpha - \beta)$ by maximizing $\Delta\ell$.
end
Add directed edge to $\mathcal{G}$
Repeat until $\Delta\ell^{g}$ does not increase anymore

The Greedy entropy-search

Input: Graph $\mathcal{G}$, distribution $\mathbb{P}$
Select $(\alpha - \beta)$ that maximizes $\Delta\ell^g$
if $\Delta\ell^g$ *is negative* **then**
  return $\mathcal{G}$
end
else
  Direct $(\alpha - \beta)$ by maximizing $\Delta\ell$.
end
Add directed edge to $\mathcal{G}$
Repeat until $\Delta\ell^g$ does not increase anymore
return $\mathcal{G}$

The Greedy entropy-search

Input: Graph $\mathcal{G}$, distribution $\mathbb{P}$
Select $(\alpha - \beta)$ that maximizes $\Delta\ell^g$
if $\Delta\ell^g$ *is negative* **then**
| **return** $\mathcal{G}$
end
else
| Direct $(\alpha - \beta)$ by maximizing $\Delta\ell$.
end
Add directed edge to $\mathcal{G}$
Repeat until $\Delta\ell^g$ does not increase anymore
return $\mathcal{G}$

- Greedy search makes locally optimal choices.

The Greedy entropy-search

Input: Graph $\mathcal{G}$, distribution $\mathbb{P}$
Select $(\alpha - \beta)$ that maximizes $\Delta\ell^{g}$
if $\Delta\ell^{g}$ *is negative* **then**
| **return** $\mathcal{G}$
end
else
| Direct $(\alpha - \beta)$ by maximizing $\Delta\ell$.
end
Add directed edge to $\mathcal{G}$
Repeat until $\Delta\ell^{g}$ does not increase anymore
return $\mathcal{G}$

- Greedy search makes locally optimal choices.
- **Q:** Can we be sure it gives a global optimum?

The Greedy entropy-search

Input: Graph $\mathcal{G}$, distribution $\mathbb{P}$
Select $(\alpha - \beta)$ that maximizes $\Delta\ell^g$
if $\Delta\ell^g$ *is negative* **then**
  return $\mathcal{G}$
end
else
  Direct $(\alpha - \beta)$ by maximizing $\Delta\ell$.
end
Add directed edge to $\mathcal{G}$
Repeat until $\Delta\ell^g$ does not increase anymore
return $\mathcal{G}$

- Greedy search makes locally optimal choices.
- **Q:** Can we be sure it gives a global optimum?
- **A:** Yes!

The Greedy entropy-search

Input: Graph $\mathcal{G}$, distribution $\mathbb{P}$
Select $(\alpha - \beta)$ that maximizes $\Delta\ell^{g}$
if $\Delta\ell^{g}$ *is negative* **then**
| **return** $\mathcal{G}$
end
else
| Direct $(\alpha - \beta)$ by maximizing $\Delta\ell$.
end
Add directed edge to $\mathcal{G}$
Repeat until $\Delta\ell^{g}$ does not increase anymore
return $\mathcal{G}$

- Greedy search makes locally optimal choices.
- **Q:** Can we be sure it gives a global optimum?
- **A:** Yes! But it requires a bit of work.

Main theorem

Theorem 4

Let $\mathscr{C}$ be an ANM with unrelated graph, $\mathcal{G}^0$. Under regularity conditions, $\mathcal{G}^0$ can be recovered from the distribution of $\mathscr{C}$ by the Greedy entropy-search.

Main theorem

Theorem 4

Let $\mathscr{C}$ be an ANM with unrelated graph, $\mathcal{G}^0$. Under regularity conditions, $\mathcal{G}^0$ can be recovered from the distribution of $\mathscr{C}$ by the Greedy entropy-search.

- To prove: Show that we never

Main theorem

Theorem 4

Let $\mathscr{C}$ be an ANM with unrelated graph, $\mathcal{G}^0$. Under regularity conditions, $\mathcal{G}^0$ can be recovered from the distribution of $\mathscr{C}$ by the Greedy entropy-search.

- To prove: Show that we never
  1. add edges not in the skeleton of $\mathcal{G}^0$,

Main theorem

Theorem 4

Let $\mathscr{C}$ be an ANM with unrelated graph, $\mathcal{G}^0$. Under regularity conditions, $\mathcal{G}^0$ can be recovered from the distribution of $\mathscr{C}$ by the Greedy entropy-search.

- To prove: Show that we never
  1. add edges not in the skeleton of $\mathcal{G}^0$,
  2. misdirect edges.

Main theorem

Theorem 4

Let $\mathscr{C}$ be an ANM with unrelated graph, $\mathcal{G}^0$. Under regularity conditions, $\mathcal{G}^0$ can be recovered from the distribution of $\mathscr{C}$ by the Greedy entropy-search.

- To prove: Show that we never
  1. add edges not in the skeleton of $\mathcal{G}^0$,
  2. misdirect edges.

Includes (A1)–(A6). The rest we find during the proof

Proving optimality – part one

- Let $(\alpha - \beta)$ be the edge chosen by the GEnS at step s.

Proving optimality – part one

- Let $(\alpha - \beta)$ be the edge chosen by the GEnS at step s.
- Strategy:

Proving optimality – part one

- Let $(\alpha - \beta)$ be the edge chosen by the GEnS at step s.
- Strategy:
  - Assume for contradiction that $(\alpha - \beta) \notin \text{ske}(\mathcal{G}^0)$.

Proving optimality – part one

- Let $(\alpha - \beta)$ be the edge chosen by the GEnS at step s.
- Strategy:
    - Assume for contradiction that $(\alpha - \beta) \notin \text{ske}(\mathcal{G}^0)$.
    - Divide into cases.

Proving optimality – part one

- Let $(\alpha - \beta)$ be the edge chosen by the GEnS at step s.
- Strategy:
    - Assume for contradiction that $(\alpha - \beta) \notin \text{ske}(\mathcal{G}^0)$.
    - Divide into cases.
    - Find edges that have a higher score than $(\alpha - \beta)$.

Proving optimality – part one

- Let $(\alpha - \beta)$ be the edge chosen by the GEnS at step s.
- Strategy:
  - Assume for contradiction that $(\alpha - \beta) \notin \text{ske}(\mathcal{G}^0)$.
  - Divide into cases.
  - Find edges that have a higher score than $(\alpha - \beta)$.
  - Contradiction. ☺

Proving optimality – part one

- Let $(\alpha - \beta)$ be the edge chosen by the GEnS at step s.
- Strategy:
  - Assume for contradiction that $(\alpha - \beta) \notin \mathrm{ske}(\mathcal{G}^0)$.
  - Divide into cases.
  - Find edges that have a higher score than $(\alpha - \beta)$.
  - Contradiction. ☺
- To get there, we rely on three results.

Proving optimality – part one

Prerequisites

Proposition 5

Let $\mathscr{C}$ be an ANM with graph $\mathcal{G}^0$ and let $\mathcal{G}$ be a subgraph. If $\alpha \perp_d \beta \mid \mathbf{PA}_{\mathcal{G}}(\beta)$, then

$$\Delta\ell^g(\mathcal{G}, \alpha \to \beta) = 0.$$

Proving optimality – part one

Prerequisites

Proposition 5

Let $\mathscr{C}$ be an ANM with graph $\mathcal{G}^0$ and let $\mathcal{G}$ be a subgraph. If $\alpha \perp_d \beta \mid \mathbf{PA}_{\mathcal{G}}(\beta)$, then

$$\Delta\ell^g(\mathcal{G}, \alpha \to \beta) = 0.$$

Proposition 6

Same set-up as above. If $\alpha \to \beta$ is in $\mathcal{G}^0$ but not in $\mathcal{G}$, then

$$\Delta\ell^g(\mathcal{G}, \alpha \to \beta) > 0.$$

Proving optimality – part one

Prerequisites continued

Lemma 7

Let $(X, Y) \sim \mathbb{P}_{(X,Y)}$ *and* $N \sim \mathbb{P}_N$ *with* $N \perp\!\!\!\perp Y$*.*

Proving optimality – part one

Prerequisites continued

Lemma 7

Let $(X, Y) \sim \mathbb{P}_{(X,Y)}$ *and* $N \sim \mathbb{P}_N$ *with* $N \perp\!\!\!\perp Y$. *Assume that* $x \mapsto \mathbb{E}_N f(g(x) + N)$ *is in* $\mathcal{F}$ *whenever* $f, g \in \mathcal{F}$.

Proving optimality – part one

Prerequisites continued

Lemma 7

Let $(X, Y) \sim \mathbb{P}_{(X,Y)}$ and $N \sim \mathbb{P}_N$ with $N \perp\!\!\!\perp Y$. Assume that $x \mapsto \mathbb{E}_N f(g(x) + N)$ is in $\mathcal{F}$ whenever $f, g \in \mathcal{F}$. Then

$$\min_{f \in \mathcal{F}} \mathbb{E}_{(X,Y)} (Y - f(X))^2 < \min_{f \in \mathcal{F}} \mathbb{E}_{(X,Y,N)} (Y - f(g(X) + N))^2$$

Proving optimality – part one

Prerequisites continued

Lemma 7

Let $(X, Y) \sim \mathbb{P}_{(X,Y)}$ and $N \sim \mathbb{P}_N$ with $N \perp\!\!\!\perp Y$. Assume that $x \mapsto \mathbb{E}_N f(g(x) + N)$ is in $\mathcal{F}$ whenever $f, g \in \mathcal{F}$. Then

$$\min_{f \in \mathcal{F}} \mathbb{E}_{(X,Y)} (Y - f(X))^2 < \min_{f \in \mathcal{F}} \mathbb{E}_{(X,Y,N)} (Y - f(g(X) + N))^2$$

Intuition:

MSE of regressing Y onto X $<$
MSE of regressing Y onto noisy version of X.

Proving optimality – part one

- We are now ready to prove part one.

Proving optimality – part one

- We are now ready to prove part one.
- Divide into three cases.

Proving optimality – part one

- We are now ready to prove part one.
- Divide into three cases.
    1. α and β are not *d*-connected in $\mathcal{G}^0$.

Proving optimality – part one

- We are now ready to prove part one.
- Divide into three cases.
  1. α and β are not *d*-connected in $\mathcal{G}^0$.
  2. α and β are *d*-connected in $\mathcal{G}^0$ through $\mathbf{PA}_{\mathcal{G}^0}(\alpha)$.

Proving optimality – part one

- We are now ready to prove part one.
- Divide into three cases.
  1. α and β are not *d*-connected in $\mathcal{G}^0$.
  2. α and β are *d*-connected in $\mathcal{G}^0$ through $\mathbf{PA}_{\mathcal{G}^0}(\alpha)$.
  3. α and β are *d*-connected in $\mathcal{G}^0$ through $\mathbf{CH}_{\mathcal{G}^0}(\alpha)$.

Proving optimality – part one

- We are now ready to prove part one.
- Divide into three cases.
  1. α and β are not *d*-connected in $\mathcal{G}^0$.
  2. α and β are *d*-connected in $\mathcal{G}^0$ through $\mathbf{PA}_{\mathcal{G}^0}(\alpha)$.
  3. α and β are *d*-connected in $\mathcal{G}^0$ through $\mathbf{CH}_{\mathcal{G}^0}(\alpha)$.
- We briefly go through case 2.

Proving optimality – part one

- We are now ready to prove part one.
- Divide into three cases.
  1. α and β are not *d*-connected in $\mathcal{G}^0$.
  2. α and β are *d*-connected in $\mathcal{G}^0$ through $\mathbf{PA}_{\mathcal{G}^0}(\alpha)$.
  3. α and β are *d*-connected in $\mathcal{G}^0$ through $\mathbf{CH}_{\mathcal{G}^0}(\alpha)$.
- We briefly go through case 2.
- Case 3 turns out to follow from case 2.

Proving optimality – part one

Case 2:

- Observe: There is only one *d*-connection, ϵ, between α and β.

Proving optimality – part one

Case 2:

- Observe: There is only one *d*-connection, ϵ, between α and β.
  - Otherwise, there would be a cycle with < 3 colliders.

Proving optimality – part one

Case 2:

- Observe: There is only one *d*-connection, ϵ, between α and β.
    - Otherwise, there would be a cycle with < 3 colliders.
- We let π be the parent of α on ϵ and ρ be the neighbor of β along ϵ.

Proving optimality – part one

Case 2 continued:

Suppose first that ρ is a parent of β.

Proving optimality – part one

Case 2 continued:

Suppose first that ρ is a parent of β.

$$
\begin{tikzcd}
\Omega_\beta \arrow[d] & & & \Omega_\alpha \setminus \{\pi\} \arrow[d] \\
\beta & \rho \arrow[l, dashed] \arrow[r, leftrightsquigarrow, "d"] & \pi \arrow[r, dashed] & \alpha \\
\Theta_\beta \setminus \{\rho\} \arrow[u, dotted] & & & \Theta_\alpha \arrow[u, dotted]
\end{tikzcd}
$$

Proving optimality – part one

Case 2 continued:

Suppose first that ρ is a parent of β.

$$
\begin{tikzcd}
\Omega_\beta \arrow[d] & & & \Omega_\alpha \backslash \{\pi\} \arrow[d] \\
\beta & \rho \arrow[l, dashed] \arrow[r, leftrightsquigarrow, "d"] & \pi \arrow[r, dashed] & \alpha \\
\Theta_\beta \backslash \{\rho\} \arrow[u, dotted] & & & \Theta_\alpha \arrow[u, dotted]
\end{tikzcd}
$$

- If π is in $\mathcal{G}^s$: $\Delta \ell^{\mathrm{g}} (\mathcal{G}^s, \beta \to \alpha) = 0$ by Proposition 5.

Proving optimality – part one

Case 2 continued:

Suppose first that ρ is a parent of β.

$$\begin{tikzcd}
\Omega_\beta \arrow[d] & & & \Omega_\alpha\setminus\{\pi\} \arrow[d] \\
\beta & \rho \arrow[l, dashed] \arrow[r, leftrightsquigarrow, "d"] & \pi \arrow[r, dashed] & \alpha \\
\Theta_\beta\setminus\{\rho\} \arrow[u, dotted] & & & \Theta_\alpha \arrow[u, dotted]
\end{tikzcd}$$

- If π is in $\mathcal{G}^s$: $\Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \beta \to \alpha\right) = 0$ by Proposition 5.
- If ρ is in $\mathcal{G}^s$: $\Delta\ell^{\mathrm{g}}\left(\mathcal{G}^s, \alpha \to \beta\right) = 0$ by Proposition 5.

Proving optimality – part one

Case 2 continued:

Suppose first that ρ is a parent of β.

$$
\begin{tikzcd}
\Omega_\beta \arrow[d] & & & \Omega_\alpha \backslash \{\pi\} \arrow[d] \\
\beta & \rho \arrow[l, dashed] \arrow[r, leftrightsquigarrow, "d"] & \pi \arrow[r, dashed] & \alpha \\
\Theta_\beta \backslash \{\rho\} \arrow[u, dotted] & & & \Theta_\alpha \arrow[u, dotted]
\end{tikzcd}
$$

- If π is <u>not</u> in $\mathcal{G}^s$: $\Delta\ell^{\mathrm{g}}(\mathcal{G}^s, \rho \to \alpha) > \Delta\ell^{\mathrm{g}}(\mathcal{G}^s, \beta \to \alpha)$ by Lemma 7.
- If ρ is in $\mathcal{G}^s$: $\Delta\ell^{\mathrm{g}}(\mathcal{G}^s, \alpha \to \beta) = 0$ by Proposition 5.

Proving optimality – part one

Case 2 continued:

Suppose first that ρ is a parent of β.

$$
\begin{tikzcd}
\Omega_\beta \arrow[d] & & & \Omega_\alpha \backslash \{\pi\} \arrow[d] \\
\beta & \rho \arrow[l, dashed] \arrow[r, leftrightsquigarrow, "d"] & \pi \arrow[r, dashed] & \alpha \\
\Theta_\beta \backslash \{\rho\} \arrow[u, dotted] & & & \Theta_\alpha \arrow[u, dotted]
\end{tikzcd}
$$

- If π is <u>not</u> in $\mathcal{G}^s$: $\Delta\ell^{\text{g}}(\mathcal{G}^s, \rho \to \alpha) > \Delta\ell^{\text{g}}(\mathcal{G}^s, \beta \to \alpha)$ by Lemma 7.
- If ρ is <u>not</u> in $\mathcal{G}^s$: $\Delta\ell^{\text{g}}(\mathcal{G}^s, \pi \to \beta) > \Delta\ell^{\text{g}}(\mathcal{G}^s, \alpha \to \beta)$ by Lemma 7.

Proving optimality – part one

Case 2 continued:

Suppose now ρ is a child of β.

Proving optimality – part one

Case 2 continued:

Suppose now ρ is a <u>child</u> of β.

$$
\begin{tikzcd}
\Omega_\beta \arrow[d] & & & \Omega_\alpha \backslash \{\pi\} \arrow[d] \\
\beta \arrow[r, dashed] & \rho \arrow[r, dashed, maps to] & \pi \arrow[r, dashed] & \alpha \\
\Theta_\beta \arrow[u, dotted] & & & \Theta_\alpha \arrow[u, dotted]
\end{tikzcd}
$$

Proving optimality – part one

Case 2 continued:

Suppose now ρ is a child of β.

$$
\begin{tikzcd}
\Omega_\beta \arrow[d] & & & \Omega_\alpha\backslash\{\pi\} \arrow[d] \\
\beta \arrow[r, dashed] & \rho \arrow[r, dashed, mapsto] & \pi \arrow[r, dashed] & \alpha \\
\Theta_\beta \arrow[u, dotted] & & & \Theta_\alpha \arrow[u, dotted]
\end{tikzcd}
$$

- As before, $(\pi \to \beta)$ is better than $(\alpha \to \beta)$.

Proving optimality – part one

Case 2 continued:

Suppose now ρ is a child of β.

$$\begin{tikzcd}
\Omega_\beta \arrow[d] & & & \Omega_\alpha\backslash\{\pi\} \arrow[d] \\
\beta \arrow[r, dashed] & \rho \arrow[r, dashed, mapsto] & \pi \arrow[r, dashed] & \alpha \\
\Theta_\beta \arrow[u, dotted] & & & \Theta_\alpha \arrow[u, dotted]
\end{tikzcd}$$

- As before, $(\pi \to \beta)$ is better than $(\alpha \to \beta)$.
- New argument for not including $(\beta \to \alpha)$.

Proving optimality – part one

Case 2 continued:

Suppose now ρ is a <u>child</u> of β.

$$\begin{tikzcd}
\Omega_\beta \arrow[d] & & & \Omega_\alpha \backslash \{\pi\} \arrow[d] \\
\beta \arrow[r, dashed] & \rho \arrow[r, dashed, mapsto] & \pi \arrow[r, dashed] & \alpha \\
\Theta_\beta \arrow[u, dotted] & & & \Theta_\alpha \arrow[u, dotted]
\end{tikzcd}$$

- As before, $(\pi \to \beta)$ is better than $(\alpha \to \beta)$.
- New argument for not including $(\beta \to \alpha)$.
    - Boils down to exploiting the independence structure of $\mathscr{C}$. But we won't show this.

Proving optimality – part one

- In summary:

Proving optimality – part one

- In summary:
  - If $(\alpha - \beta)$ is not in ske$(\mathcal{G}^0)$, we can always find a better alternative.

Proving optimality – part one

- In summary:
  - If $(\alpha - \beta)$ is not in ske$(\mathcal{G}^0)$, we can always find a better alternative.
  - This contradicts that we selected $(\alpha - \beta)$.

Proving optimality – part one

- In summary:
  - If $(\alpha - \beta)$ is not in ske$(\mathcal{G}^0)$, we can always find a better alternative.
  - This contradicts that we selected $(\alpha - \beta)$.
  - $\Rightarrow (\alpha - \beta)$ must be in ske$(\mathcal{G}^0)$.

Proving optimality – part one

- In summary:
  - If $(\alpha - \beta)$ is not in ske$(\mathcal{G}^0)$, we can always find a better alternative.
  - This contradicts that we selected $(\alpha - \beta)$.
  - $\Rightarrow (\alpha - \beta)$ must be in ske$(\mathcal{G}^0)$.
- We then move on to part 2.

Proving optimality – part two

- Assume w.l.o.g. that $\Delta\ell$ is highest in $(\alpha \to \beta)$.

Proving optimality – part two

- Assume w.l.o.g. that $\Delta\ell$ is highest in $(\alpha \to \beta)$.
- Proof strategy:

Proving optimality – part two

- Assume w.l.o.g. that $\Delta\ell$ is highest in $(\alpha \to \beta)$.
- Proof strategy:
  - Assume for contradiction that $(\alpha \to \beta)$ is <u>not</u> in $\mathcal{G}^0$.

Proving optimality – part two

- Assume w.l.o.g. that $\Delta\ell$ is highest in $(\alpha \to \beta)$.
- Proof strategy:
  - Assume for contradiction that $(\alpha \to \beta)$ is not in $\mathcal{G}^0$.
  - This implies that $(\beta \to \alpha)$ is.

Proving optimality – part two

- Assume w.l.o.g. that $\Delta\ell$ is highest in $(\alpha \to \beta)$.
- Proof strategy:
  - Assume for contradiction that $(\alpha \to \beta)$ is <u>not</u> in $\mathcal{G}^0$.
  - This implies that $(\beta \to \alpha)$ is.
  - We look a subgraph of $\mathcal{G}^s$, $\tilde{\mathcal{G}}$.

Proving optimality – part two

- Assume w.l.o.g. that $\Delta\ell$ is highest in $(\alpha \to \beta)$.
- Proof strategy:
  - Assume for contradiction that $(\alpha \to \beta)$ is not in $\mathcal{G}^0$.
  - This implies that $(\beta \to \alpha)$ is.
  - We look a subgraph of $\mathcal{G}^s$, $\tilde{\mathcal{G}}$.

Proving optimality – part two

- Assume w.l.o.g. that $\Delta\ell$ is highest in $(\alpha \to \beta)$.
- Proof strategy:
    - Assume for contradiction that $(\alpha \to \beta)$ is <u>not</u> in $\mathcal{G}^0$.
    - This implies that $(\beta \to \alpha)$ is.
    - We look a subgraph of $\mathcal{G}^s$, $\tilde{\mathcal{G}}$.
        - Consists of α, β and their $\mathcal{G}^s$ parents.

Proving optimality – part two

- Assume w.l.o.g. that $\Delta\ell$ is highest in $(\alpha \to \beta)$.
- Proof strategy:
  - Assume for contradiction that $(\alpha \to \beta)$ is <u>not</u> in $\mathcal{G}^0$.
  - This implies that $(\beta \to \alpha)$ is.
  - We look a subgraph of $\mathcal{G}^s$, $\tilde{\mathcal{G}}$.
    - Consists of α, β and their $\mathcal{G}^s$ parents.

$$
\begin{tikzcd}
\mathbf{PA}_{\mathcal{G}^s}(\beta) \arrow[d] & \mathbf{PA}_{\mathcal{G}^s}(\alpha) \arrow[d] \\
\beta & \alpha
\end{tikzcd}
$$

Proving optimality – part two

- Assume w.l.o.g. that $\Delta\ell$ is highest in $(\alpha \to \beta)$.
- Proof strategy:
  - Assume for contradiction that $(\alpha \to \beta)$ is <u>not</u> in $\mathcal{G}^0$.
  - This implies that $(\beta \to \alpha)$ is.
  - We look a subgraph of $\mathcal{G}^s$, $\tilde{\mathcal{G}}$.
    - Consists of α, β and their $\mathcal{G}^s$ parents.
    - Add $(\beta \to \alpha)$ to $\tilde{\mathcal{G}}$

$$\begin{array}{ccc} \mathbf{PA}_{\mathcal{G}^s}(\beta) & \qquad & \mathbf{PA}_{\mathcal{G}^s}(\alpha) \\ \downarrow & & \downarrow \\ \beta & & \alpha \end{array}$$

Proving optimality – part two

- Assume w.l.o.g. that $\Delta\ell$ is highest in $(\alpha \to \beta)$.
- Proof strategy:
  - Assume for contradiction that $(\alpha \to \beta)$ is <u>not</u> in $\mathcal{G}^0$.
  - This implies that $(\beta \to \alpha)$ is.
  - We look a subgraph of $\mathcal{G}^s$, $\tilde{\mathcal{G}}$.
    - Consists of α, β and their $\mathcal{G}^s$ parents.
    - Add $(\beta \to \alpha)$ to $\tilde{\mathcal{G}}$

$$\begin{tikzcd}
\mathbf{PA}_{\mathcal{G}^s}(\beta) \arrow[d] & \mathbf{PA}_{\mathcal{G}^s}(\alpha) \arrow[d] \\
\beta \arrow[r] & \alpha
\end{tikzcd}$$

Proving optimality – part two

- Assume w.l.o.g. that $\Delta\ell$ is highest in $(\alpha \to \beta)$.
- Proof strategy:
  - Assume for contradiction that $(\alpha \to \beta)$ is <u>not</u> in $\mathcal{G}^0$.
  - This implies that $(\beta \to \alpha)$ is.
  - We look a subgraph of $\mathcal{G}^s$, $\tilde{\mathcal{G}}$.
    - Consists of α, β and their $\mathcal{G}^s$ parents.
    - Add $(\beta \to \alpha)$ to $\tilde{\mathcal{G}}$
  - We then apply Theorem 3 to reach a contradiction.

$$\begin{tikzcd} \mathbf{PA}_{\mathcal{G}^s}(\beta) \arrow[d] & \mathbf{PA}_{\mathcal{G}^s}(\alpha) \arrow[d] \\ \beta \arrow[r] & \alpha \end{tikzcd}$$

Theorem 3

If $\mathcal{G}^0$ is the true graph of an ANM that satisfies (A1)–(A6), then

$$\mathcal{G}^0 = \arg\max_{\mathcal{G}} \ell(\mathcal{G}).$$

Proving optimality – part two

Difficulties in proving part two

- To apply Theorem 3, we need $\tilde{\mathcal{G}}$ to be the graph of an identifiable ANM, $\tilde{\mathscr{C}}$.

Proving optimality – part two

Difficulties in proving part two

- To apply Theorem 3, we need $\tilde{\mathcal{G}}$ to be the graph of an identifiable ANM, $\tilde{\mathscr{C}}$.
- $\Leftrightarrow \tilde{\mathscr{C}}$ needs to satisfy assumptions (A1)–(A6)

Proving optimality – part two

Difficulties in proving part two

- To apply Theorem 3, we need $\tilde{\mathcal{G}}$ to be the graph of an identifiable ANM, $\tilde{\mathscr{C}}$.
- $\Leftrightarrow$ $\tilde{\mathscr{C}}$ needs to satisfy assumptions (A1)–(A6)
- What does $\tilde{\mathscr{C}}$ look like?

Proving optimality – part two

Difficulties in proving part two

- To apply Theorem 3, we need $\tilde{\mathcal{G}}$ to be the graph of an identifiable ANM, $\tilde{\mathscr{C}}$.
- $\Leftrightarrow$ $\tilde{\mathscr{C}}$ needs to satisfy assumptions (A1)–(A6)
- What does $\tilde{\mathscr{C}}$ look like?

Proving optimality – part two

Difficulties in proving part two

- To apply Theorem 3, we need $\tilde{\mathcal{G}}$ to be the graph of an identifiable ANM, $\tilde{\mathscr{C}}$.
- $\Leftrightarrow$ $\tilde{\mathscr{C}}$ needs to satisfy assumptions (A1)–(A6)
- What does $\tilde{\mathscr{C}}$ look like? We can write the assignments as

$$X_\nu := \sum_{\gamma \in \mathbf{PA}_{\tilde{\mathcal{G}}}(\nu)} f^0_{\nu,\gamma}(X_\gamma) + \tilde{N}_\nu$$

Proving optimality – part two

Difficulties in proving part two

- To apply Theorem 3, we need $\tilde{\mathcal{G}}$ to be the graph of an identifiable ANM, $\tilde{\mathscr{C}}$.
- $\Leftrightarrow$ $\tilde{\mathscr{C}}$ needs to satisfy assumptions (A1)–(A6)
- What does $\tilde{\mathscr{C}}$ look like? We can write the assignments as

$$X_\nu := \sum_{\gamma \in \mathbf{PA}_{\tilde{\mathcal{G}}}(\nu)} f^0_{\nu,\gamma}(X_\gamma) + \tilde{N}_\nu$$

where

$$\tilde{N}_\nu = \sum_{\gamma \in \mathbf{PA}_{\mathcal{G}^0}(\nu) \setminus \mathbf{PA}_{\tilde{\mathcal{G}}}(X_\nu)} f^0_{\nu,\gamma}(X_\gamma) + N_\nu.$$

Proving optimality – part two

Difficulties in proving part two

- To apply Theorem 3, we need $\tilde{\mathcal{G}}$ to be the graph of an identifiable ANM, $\tilde{\mathscr{C}}$.
- $\Leftrightarrow$ $\tilde{\mathscr{C}}$ needs to satisfy assumptions (A1)–(A6)
- What does $\tilde{\mathscr{C}}$ look like? We can write the assignments as

$$X_\nu := \sum_{\gamma \in \mathbf{PA}_{\tilde{\mathcal{G}}}(\nu)} f^0_{\nu,\gamma}(X_\gamma) + \tilde{N}_\nu$$

where

$$\tilde{N}_\nu = \sum_{\gamma \in \mathbf{PA}_{\mathcal{G}^0}(\nu) \setminus \mathbf{PA}_{\tilde{\mathcal{G}}}(X_\nu)} f^0_{\nu,\gamma}(X_\gamma) + N_\nu.$$

- Notice that the $\tilde{N}$ variables are mutually independent.

Proving optimality – part two

(A1,2) $\mathcal{F}$ consists of non-linear C^3 functions + some regularity conditions.

(A3) The densities of the noise variables have only discretely many solutions to the differential equation $(\log f)'' = 0$.

(A4) The noise variables have full support and their densities are in C^3 and strictly positive.

(A5,6) All noise variables and all $\mathcal{F}$-transformations of them have second moment.

Proving optimality – part two

(A1,2) $\mathcal{F}$ consists of non-linear C^3 functions + some regularity conditions.

(A3) The densities of the noise variables have only discretely many solutions to the differential equation $(\log f)'' = 0$.

(A4) The noise variables have full support and their densities are in C^3 and strictly positive.

(A5,6) All noise variables and all $\mathcal{F}$-transformations of them have second moment.

Proving optimality – part two

Difficulties in proving part two continued

- Then, for $\tilde{\mathscr{C}}$ to satisfy (A3):

Proving optimality – part two

Difficulties in proving part two continued

- Then, for $\tilde{\mathscr{C}}$ to satisfy (A3):
  - Convolutions of N_ν and ν-parents must have at most discretely many solutions to $(\log f)'' = 0$.

Proving optimality – part two

Difficulties in proving part two continued

- Then, for $\tilde{\mathscr{C}}$ to satisfy (A3):
  - Convolutions of N_ν and ν-parents must have at most discretely many solutions to $(\log f)'' = 0$.
- How do we ensure this holds?

Proving optimality – part two

Difficulties in proving part two continued

- Then, for $\tilde{\mathscr{C}}$ to satisfy (A3):
  - Convolutions of N_ν and ν-parents must have at most discretely many solutions to $(\log f)'' = 0$.
- How do we ensure this holds?
- Start by solving the equation.

Proving optimality – part two

Difficulties in proving part two continued

- Then, for $\tilde{\mathscr{C}}$ to satisfy (A3):
  - Convolutions of N_ν and ν-parents must have at most discretely many solutions to $(\log f)'' = 0$.
- How do we ensure this holds?
- Start by solving the equation.
- $(\log f)'' = 0 \Leftrightarrow f(x) = exp(c_1 \cdot x + c_2)$.

Proving optimality – part two

Difficulties in proving part two continued

- Then, for $\tilde{\mathscr{C}}$ to satisfy (A3):
  - Convolutions of N_ν and ν-parents must have at most discretely many solutions to $(\log f)'' = 0$.
- How do we ensure this holds?
- Start by solving the equation.
- $(\log f)'' = 0 \Leftrightarrow f(x) = exp(c_1 \cdot x + c_2)$.
- We call these functions **log-linear**.

Proving optimality – part two

Solving the convolution problem

Lemma 8

Let f be a real analytic function.

Proving optimality – part two

Solving the convolution problem

Lemma 8

Let f be a real analytic function.

C^∞ and has a convergent power series representation in a neighborhood of every point. Symbol: C^ω.

Proving optimality – part two

Solving the convolution problem

Lemma 8

Let f be a real analytic function. If f is log-linear on $[a, b]$, then f is log-linear on all of $\mathbb{R}$.

Proving optimality – part two

Solving the convolution problem

Lemma 8

Let f be a real analytic function. If f is log-linear on $[a, b]$, then f is log-linear on all of $\mathbb{R}$.

Theorem 9

Let $f \in C^\omega \cap \mathcal{L}^\infty$ and $g \in \mathcal{L}^1$.

Proving optimality – part two

Solving the convolution problem

Lemma 8

Let f be a real analytic function. If f is log-linear on $[a, b]$, then f is log-linear on all of $\mathbb{R}$.

Theorem 9

Let $f \in C^\omega \cap \mathcal{L}^\infty$ and $g \in \mathcal{L}^1$. Then $f * g \in C^\omega$.

Proving optimality – part two

Solving the convolution problem continued

Corollary 10

Let $f \in C_{+}^{\omega}$ and $g \in \mathcal{L}^{\infty}$ be densities.

Proving optimality – part two

Solving the convolution problem continued

Corollary 10

Let $f \in C_+^\omega$ and $g \in \mathcal{L}^\infty$ be densities. The convolution $f * g$ is not log-linear on any interval.

Proving optimality – part two

Solving the convolution problem continued

Corollary 10

Let $f \in C_+^\omega$ and $g \in \mathcal{L}^\infty$ be densities. The convolution $f * g$ is not log-linear on any interval.

Proof.

1. Observe that $f * g$ is integrable and real analytic.

□

Proving optimality – part two

Solving the convolution problem continued

Corollary 10

Let $f \in C_+^\omega$ and $g \in \mathcal{L}^\infty$ be densities. The convolution $f * g$ is not log-linear on any interval.

Proof.

1. Observe that $f * g$ is integrable and real analytic.
2. All integrable functions that are log-linear on all of $\mathbb{R}$ are on the form $h(x) = \exp(c_1 \cdot |x| + c_2)$, $c_1 < 0$.

$\square$

Proving optimality – part two

Solving the convolution problem continued

Corollary 10

Let $f \in C_+^\omega$ and $g \in \mathcal{L}^\infty$ be densities. The convolution $f * g$ is not log-linear on any interval.

Proof.

1. Observe that $f * g$ is integrable and real analytic.
2. All integrable functions that are log-linear on all of $\mathbb{R}$ are on the form $h(x) = \exp(c_1 \cdot |x| + c_2)$, $c_1 < 0$.
3. $h(x)$ is <u>not</u> real analytic $\Rightarrow f * g$ is not log-linear on all of $\mathbb{R}$.

$\square$

Proving optimality – part two

Solving the convolution problem continued

Corollary 10

Let $f \in C_+^\omega$ and $g \in \mathcal{L}^\infty$ be densities. The convolution $f * g$ is not log-linear on any interval.

Proof.

1. Observe that $f * g$ is integrable and real analytic.
2. All integrable functions that are log-linear on all of $\mathbb{R}$ are on the form $h(x) = \exp(c_1 \cdot |x| + c_2)$, $c_1 < 0$.
3. $h(x)$ is not real analytic $\Rightarrow f * g$ is not log-linear on all of $\mathbb{R}$.
4. $\Rightarrow f * g$ is not log-linear on any interval.

$\square$

Proving optimality – part two

Solving the convolution problem continued

Corollary 10

Let $f \in C_+^\omega$ and $g \in \mathcal{L}^\infty$ be densities. The convolution $f * g$ is not log-linear on any interval.

- **Implication:** All N's have real analytic densities $\Rightarrow$ all $\tilde{N}$'s satisfy (A3).

Proving optimality – part two

Solving the convolution problem continued

Corollary 10

Let $f \in C^{\omega}_{+}$ and $g \in \mathcal{L}^{\infty}$ be densities. The convolution $f * g$ is not log-linear on any interval.

- **Implication:** All N's have real analytic densities $\Rightarrow$ all $\tilde{N}$'s satisfy (A3).
- Includes Gaussian variables, among others.

Proving optimality – part two

Solving the convolution problem continued

Corollary 10

Let $f \in C_+^\omega$ and $g \in \mathcal{L}^\infty$ be densities. The convolution $f * g$ is not log-linear on any interval.

- **Implication:** All N's have real analytic densities $\Rightarrow$ all $\tilde{N}$'s satisfy (A3).
- Includes Gaussian variables, among others.
- Assume all noises have real analytic density

Proving optimality – part two

Solving the convolution problem continued

Corollary 10

Let $f \in C_+^\omega$ and $g \in \mathcal{L}^\infty$ be densities. The convolution $f * g$ is not log-linear on any interval.

- **Implication:** All N's have real analytic densities $\Rightarrow$ all $\tilde{N}$'s satisfy (A3).
- Includes Gaussian variables, among others.
- Assume all noises have real analytic density
- $\tilde{\mathscr{C}}$ is now identifiable!

Proving optimality – part two

- From here, it's easy:

Proving optimality – part two

- From here, it's easy:
- Let $\tilde{\mathcal{G}}_{\alpha\to\beta}$ be a version of $\tilde{\mathcal{G}}$ where we flip $(\beta \to \alpha)$.

Proving optimality – part two

- From here, it's easy:
- Let $\tilde{\mathcal{G}}_{\alpha \to \beta}$ be a version of $\tilde{\mathcal{G}}$ where we flip $(\beta \to \alpha)$.

$$
\begin{tikzcd}
\mathbf{PA}_{\mathcal{G}^s}(\beta) \arrow[d] & \mathbf{PA}_{\mathcal{G}^s}(\alpha) \arrow[d] \\
\beta \arrow[r] & \alpha
\end{tikzcd}
\qquad
\begin{tikzcd}
\mathbf{PA}_{\mathcal{G}^s}(\beta) \arrow[d] & \mathbf{PA}_{\mathcal{G}^s}(\alpha) \arrow[d] \\
\beta & \alpha \arrow[l]
\end{tikzcd}
$$

Proving optimality – part two

- From here, it's easy:
- Let $\tilde{\mathcal{G}}_{\alpha\to\beta}$ be a version of $\tilde{\mathcal{G}}$ where we flip $(\beta \to \alpha)$.
- By Theorem 3

$$\ell(\tilde{\mathcal{G}}) > \ell(\tilde{\mathcal{G}}_{\alpha\to\beta})$$

$$\begin{tikzcd}
\mathbf{PA}_{\mathcal{G}^s}(\beta) \arrow[d] & \mathbf{PA}_{\mathcal{G}^s}(\alpha) \arrow[d] \\
\beta \arrow[r] & \alpha
\end{tikzcd}
\qquad\qquad
\begin{tikzcd}
\mathbf{PA}_{\mathcal{G}^s}(\beta) \arrow[d] & \mathbf{PA}_{\mathcal{G}^s}(\alpha) \arrow[d] \\
\beta & \alpha \arrow[l]
\end{tikzcd}$$

Proving optimality – part two

- From here, it's easy:
- Let $\tilde{\mathcal{G}}_{\alpha\to\beta}$ be a version of $\tilde{\mathcal{G}}$ where we flip $(\beta \to \alpha)$.
- By Theorem 3

$$\ell(\tilde{\mathcal{G}}) > \ell(\tilde{\mathcal{G}}_{\alpha\to\beta})$$

- Write out and rearrange...

$$\begin{tikzcd}
\mathbf{PA}_{\mathcal{G}^s}(\beta) \arrow[d] & \mathbf{PA}_{\mathcal{G}^s}(\alpha) \arrow[d] \\
\beta \arrow[r] & \alpha
\end{tikzcd}
\qquad\qquad
\begin{tikzcd}
\mathbf{PA}_{\mathcal{G}^s}(\beta) \arrow[d] & \mathbf{PA}_{\mathcal{G}^s}(\alpha) \arrow[d] \\
\beta & \alpha \arrow[l]
\end{tikzcd}$$

Proving optimality – part two

- From here, it's easy:
- Let $\tilde{\mathcal{G}}_{\alpha \to \beta}$ be a version of $\tilde{\mathcal{G}}$ where we flip $(\beta \to \alpha)$.
- By Theorem 3

$$\ell(\tilde{\mathcal{G}}) > \ell(\tilde{\mathcal{G}}_{\alpha \to \beta})$$

- Write out and rearrange...
- We get

$$\Delta \ell \left(\mathcal{G}^s, \beta \to \alpha\right) > \Delta \ell \left(\mathcal{G}^s, \alpha \to \beta\right)$$

which is a contradiction.

$$\begin{tikzcd}
\mathbf{PA}_{\mathcal{G}^s}(\beta) \arrow[d] & \mathbf{PA}_{\mathcal{G}^s}(\alpha) \arrow[d] \\
\beta \arrow[r] & \alpha
\end{tikzcd}
\qquad
\begin{tikzcd}
\mathbf{PA}_{\mathcal{G}^s}(\beta) \arrow[d] & \mathbf{PA}_{\mathcal{G}^s}(\alpha) \arrow[d] \\
\beta & \alpha \arrow[l]
\end{tikzcd}$$

Proving optimality – part two

- From here, it's easy:
- Let $\tilde{\mathcal{G}}_{\alpha\to\beta}$ be a version of $\tilde{\mathcal{G}}$ where we flip $(\beta \to \alpha)$.
- By Theorem 3

$$\ell(\tilde{\mathcal{G}}) > \ell(\tilde{\mathcal{G}}_{\alpha\to\beta})$$

- Write out and rearrange...
- We get

$$\Delta\ell\left(\mathcal{G}^s, \beta \to \alpha\right) > \Delta\ell\left(\mathcal{G}^s, \alpha \to \beta\right)$$

  which is a contradiction.
- Which completes the proof!

Proving optimality – part two

- From here, it's easy:
- Let $\tilde{\mathcal{G}}_{\alpha\to\beta}$ be a version of $\tilde{\mathcal{G}}$ where we flip $(\beta \to \alpha)$.
- By Theorem 3

$$\ell(\tilde{\mathcal{G}}) > \ell(\tilde{\mathcal{G}}_{\alpha\to\beta})$$

- Write out and rearrange...
- We get

$$\Delta\ell\left(\mathcal{G}^s, \beta \to \alpha\right) > \Delta\ell\left(\mathcal{G}^s, \alpha \to \beta\right)$$

  which is a contradiction.
- Which completes the proof!
- **Remark:** Proof can be made to work with linear assignments.

A few simulations and real data

Set-up

- Non-parametric regression with `gam` from `mgcv` package.

Set-up

- Non-parametric regression with `gam` from `mgcv` package.
- Entropy estimation with a resubstitution estimator:

$$\hat{\mathbb{H}}_n(\hat{\boldsymbol{N}}_\nu) := -\frac{1}{n}\sum_{i=1}^{n} \log \hat{p}_{\hat{N}_\nu}(\hat{N}_\nu^i).$$

Set-up

- Non-parametric regression with `gam` from `mgcv` package.
- Entropy estimation with a resubstitution estimator:

$$\hat{\mathbb{H}}_n(\hat{\boldsymbol{N}}_\nu) := -\frac{1}{n}\sum_{i=1}^{n} \log \hat{p}_{\hat{N}_\nu}(\hat{N}_\nu^i).$$

Estimated residuals

Set-up

- Non-parametric regression with `gam` from `mgcv` package.
- Entropy estimation with a resubstitution estimator:

$$\hat{\mathbb{H}}_n(\hat{\boldsymbol{N}}_\nu) := -\frac{1}{n}\sum_{i=1}^{n} \log \hat{p}_{\hat{N}_\nu}(\hat{N}_\nu^i).$$

Estimated residuals

Density estimate using `logspline`

Set-up

- Non-parametric regression with `gam` from `mgcv` package.
- Entropy estimation with a resubstitution estimator:

$$\underbrace{\hat{\mathbb{H}}_n(\hat{\boldsymbol{N}}_\nu)}_{\text{Estimated residuals}} := -\frac{1}{n}\sum_{i=1}^{n} \log \underbrace{\hat{p}_{\hat{N}_\nu}}_{\text{Density estimate using logspline}}(\hat{N}_\nu^i).$$

- Modified exit condition:

Set-up

- Non-parametric regression with `gam` from `mgcv` package.
- Entropy estimation with a resubstitution estimator:

$$\hat{\mathbb{H}}_n(\hat{\boldsymbol{N}}_\nu) := -\frac{1}{n}\sum_{i=1}^{n} \log \hat{p}_{\hat{N}_\nu}(\hat{N}^i_\nu).$$

Estimated residuals

Density estimate using `logspline`

- Modified exit condition:
  - Test for significance of marginal entropy.

Set-up

- Non-parametric regression with `gam` from `mgcv` package.
- Entropy estimation with a resubstitution estimator:

$$\hat{\mathbb{H}}_n(\hat{\boldsymbol{N}}_\nu) := -\frac{1}{n}\sum_{i=1}^{n} \log \hat{p}_{\hat{N}_\nu}(\hat{N}_\nu^i).$$

Estimated residuals

Density estimate using `logspline`

- Modified exit condition:
  - Test for significance of marginal entropy.
  - Exit after a set number of attempts at adding non-significant edges.

Set-up

- Non-parametric regression with `gam` from `mgcv` package.
- Entropy estimation with a resubstitution estimator:

$$\hat{\mathbb{H}}_n(\hat{\boldsymbol{N}}_\nu) := -\frac{1}{n}\sum_{i=1}^{n} \log \hat{p}_{\hat{N}_\nu}(\hat{N}_\nu^i).$$

Estimated residuals

Density estimate using `logspline`

- Modified exit condition:
  - Test for significance of marginal entropy.
  - Exit after a set number of attempts at adding non-significant edges.
- We compare methods using SHD and SID.

Set-up

- Non-parametric regression with `gam` from `mgcv` package.
- Entropy estimation with a resubstitution estimator:

$$\hat{\mathbb{H}}_n(\hat{\boldsymbol{N}}_\nu) := -\frac{1}{n}\sum_{i=1}^{n} \log \hat{p}_{\hat{N}_\nu}(\hat{N}_\nu^i).$$

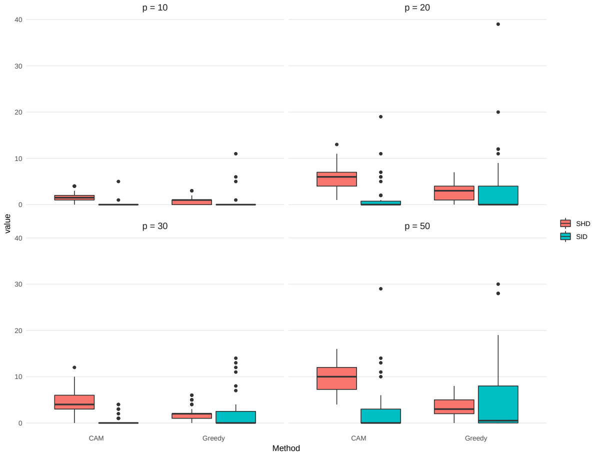
Estimated residuals

Density estimate using `logspline`

- Modified exit condition:
  - Test for significance of marginal entropy.
  - Exit after a set number of attempts at adding non-significant edges.
- We compare methods using SHD and SID.
- We try it on random DAGs with random edge functions.

Non-linear, Gaussian case

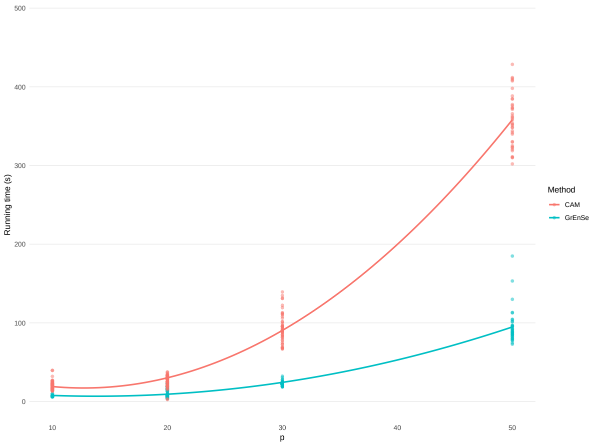
Comparison to CAM



50 repetitions, $N = 300$. p is number of nodes.

Non-linear, Gaussian case

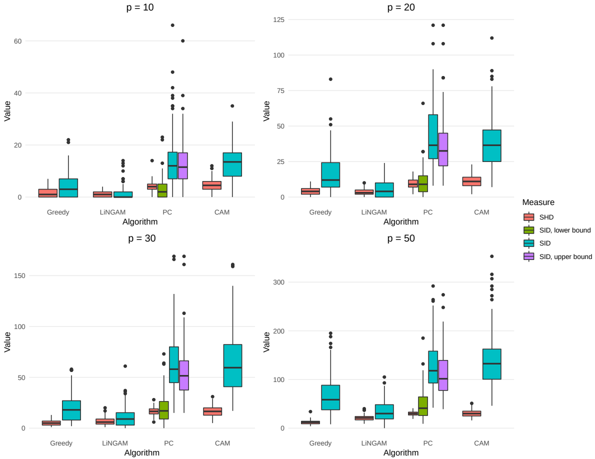
Comparison to CAM – computation time



50 repetitions, $N = 300$. p is number of nodes.

Linear, non-Gaussian

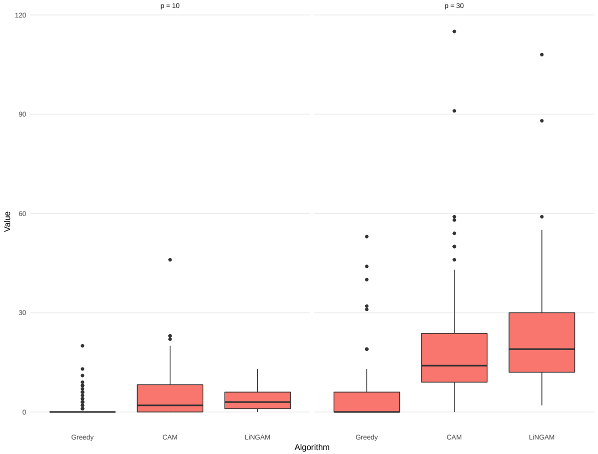
Comparison to other other methods



100 repetitions, $N = 1000$. p is number of nodes. Noise is Hyperbolic Secant

Non-Gaussian, no assumption on linearity

Comparison to other other methods – only SID



100 repetitions, $N = 1000$. p is number of nodes. Noise is Hyperbolic Secant

Real data

- 96 cause-effect pairs[2] with known ground truth.

[2] http://webdav.tuebingen.mpg.de/cause-effect/

Real data

- 96 cause-effect pairs[2] with known ground truth.
- Correctly identified 58 (60.4%) of cases.

[2] http://webdav.tuebingen.mpg.de/cause-effect/

Real data

- 96 cause-effect pairs[2] with known ground truth.
- Correctly identified 58 (60.4%) of cases.
- Weighted accuracy: 65.2%.

[2] http://webdav.tuebingen.mpg.de/cause-effect/

Conclusion

Conclusion

- Proven optimality in Population case.

Conclusion

- Proven optimality in Population case.
- Simulations:

Conclusion

- Proven optimality in Population case.
- Simulations:
  - Non-linear, Gaussian: Similar to CAM; runs faster than, but achieves slightly worse SID.

Conclusion

- Proven optimality in Population case.
- Simulations:
  - Non-linear, Gaussian: Similar to CAM; runs faster than, but achieves slightly worse SID.
  - Linear, non-Gaussian: Slightly outperformed by LiNGAM on SID, similar in terms of SHD. Outperforms both CAM and PC.

Conclusion

- Proven optimality in Population case.
- Simulations:
  - Non-linear, Gaussian: Similar to CAM; runs faster than, but achieves slightly worse SID.
  - Linear, non-Gaussian: Slightly outperformed by LiNGAM on SID, similar in terms of SHD. Outperforms both CAM and PC.
  - Greedy entropy-search perhaps more stable to linear functions/non-Gaussian noise?

Conclusion

- Proven optimality in Population case.
- Simulations:
  - Non-linear, Gaussian: Similar to CAM; runs faster than, but achieves slightly worse SID.
  - Linear, non-Gaussian: Slightly outperformed by LiNGAM on SID, similar in terms of SHD. Outperforms both CAM and PC.
  - Greedy entropy-search perhaps more stable to linear functions/non-Gaussian noise?
- Method appears comparable to others on real data.

Conclusion

- Proven optimality in Population case.
- Simulations:
  - Non-linear, Gaussian: Similar to CAM; runs faster than, but achieves slightly worse SID.
  - Linear, non-Gaussian: Slightly outperformed by LiNGAM on SID, similar in terms of SHD. Outperforms both CAM and PC.
  - Greedy entropy-search perhaps more stable to linear functions/non-Gaussian noise?
- Method appears comparable to others on real data.
- Simulations were small-scale.

Conclusion

- Proven optimality in Population case.
- Simulations:
  - Non-linear, Gaussian: Similar to CAM; runs faster than, but achieves slightly worse SID.
  - Linear, non-Gaussian: Slightly outperformed by LiNGAM on SID, similar in terms of SHD. Outperforms both CAM and PC.
  - Greedy entropy-search perhaps more stable to linear functions/non-Gaussian noise?
- Method appears comparable to others on real data.
- Simulations were small-scale.
- We did not attempt to optimize runtimes with CAM – could possibly be sped up.

Future work

- Find consistent estimators.

Future work

- Find consistent estimators.
- Test in simulations on larger scale.

Future work

- Find consistent estimators.
- Test in simulations on larger scale.
    - Could be interesting to try with no assumptions on linearity and random distributions.

Future work

- Find consistent estimators.
- Test in simulations on larger scale.
    - Could be interesting to try with no assumptions on linearity and random distributions.
- Relax assumptions of main proof.

Thanks for listening ☺

Proof of Lemma 7

Proof.

Choose any $f, g \in \mathcal{F}$. By Jensen's inequality

$$(Y - \mathbb{E}_N f(g(X) + N))^2 < \mathbb{E}_N (Y - f(g(X) + N))^2.$$

Take $\mathbb{E}_{(X,Y)}$ on both sides and use Tonelli:

$$\mathbb{E}_{(X,Y)}(Y - \mathbb{E}_N f(g(X) + N))^2 < \mathbb{E}_{(X,Y,N)}(Y - f(g(X) + N))^2.$$

By assumption, this implies:

$$\min_{f \in \mathcal{F}} \mathbb{E}_{(X,Y)}(Y - f(X))^2 < \min_{f \in \mathcal{F}} \mathbb{E}_{(X,Y,N)}(Y - f(g(X) + N))^2.$$

$\square$

Proof of Lemma 8

Proof.

Assume for contradiction that f does not solve $(\log f)'' = 0$ on $\mathbb{R} \backslash [a, b]$. By assumption

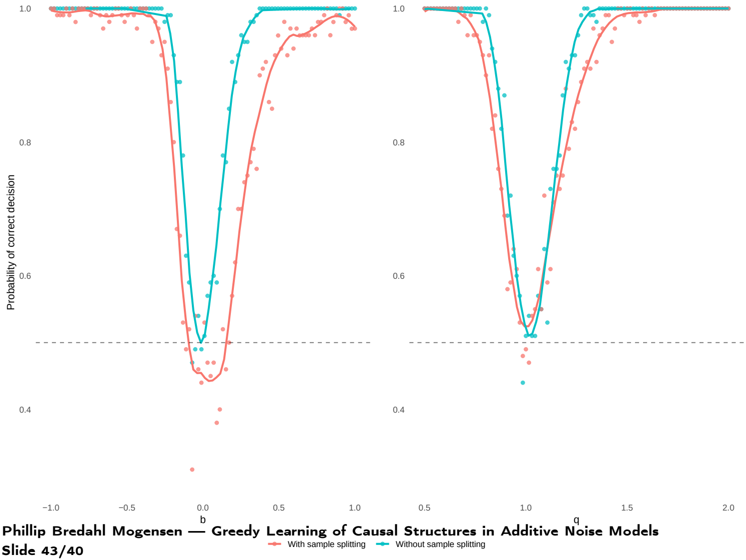
$$\forall n \in \mathbb{N}_0 : \quad f^{(n)}(a) = \exp(c_2) \cdot c_1^n \cdot \exp(c_1 \cdot a) =: \tilde{k} \cdot c_1^n.$$

Taylor-expand f around a:

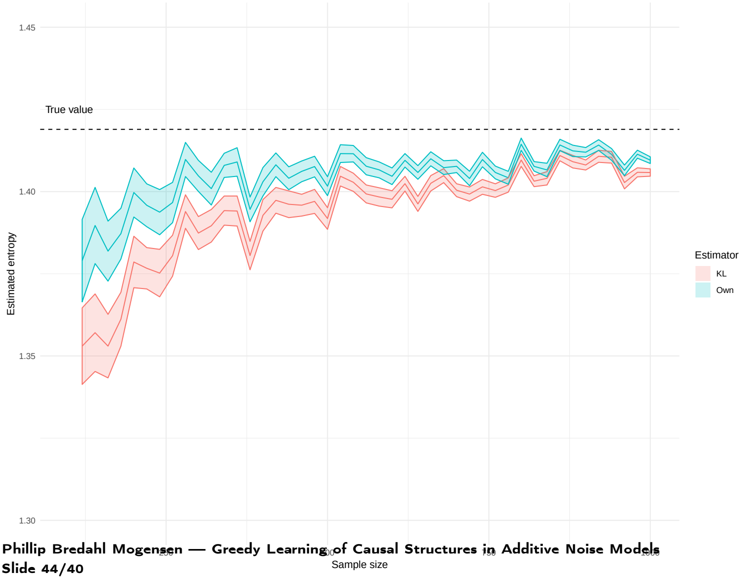
$$f(x) = \cdots = \tilde{k} \sum_{i=0}^{\infty} \frac{c_1^i}{i!}(x-a)^i = \tilde{k} \exp(c_1 \cdot (x-a)).$$

This holds in an open neighborhood of a, which gives us a contradiction. □

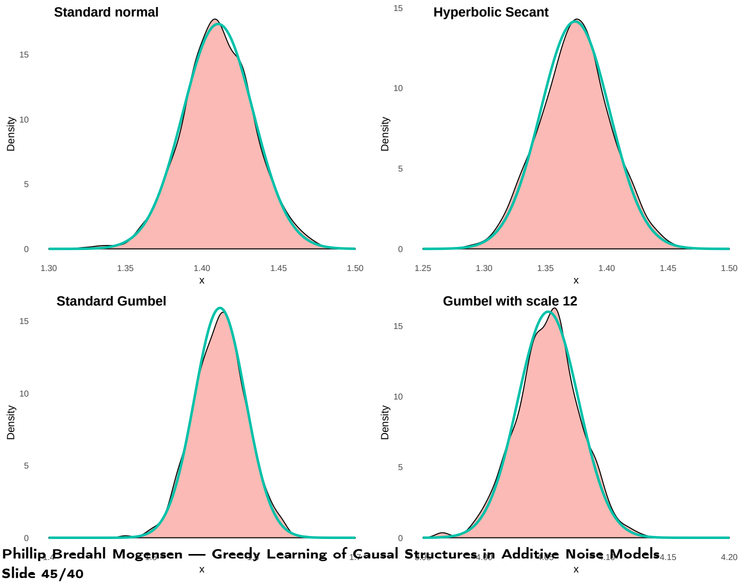
Bivariate case



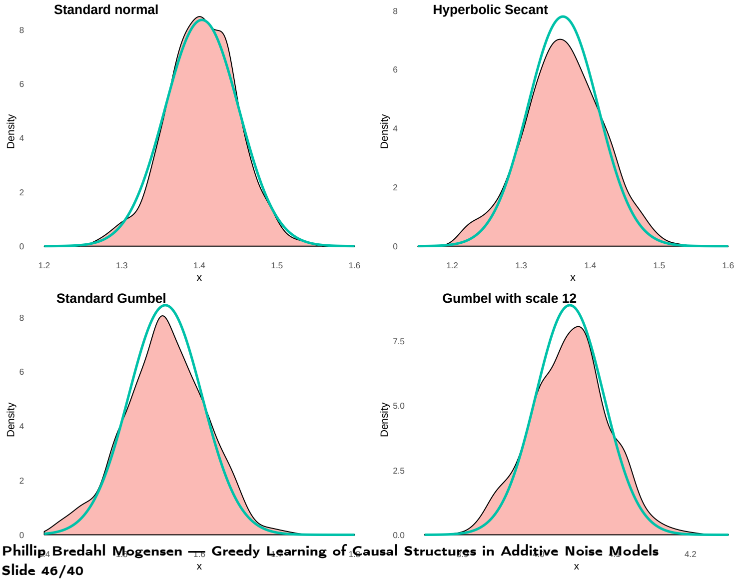
KL vs. logspline estimation



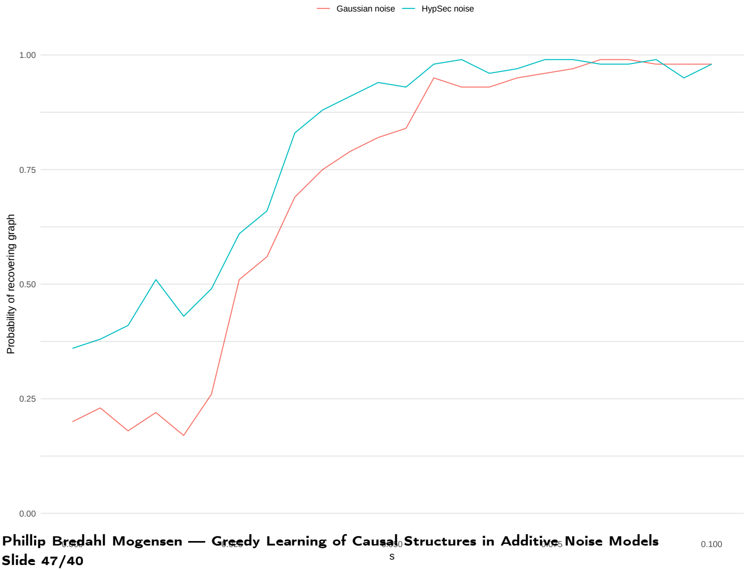
Asymptotics of entropy estimator



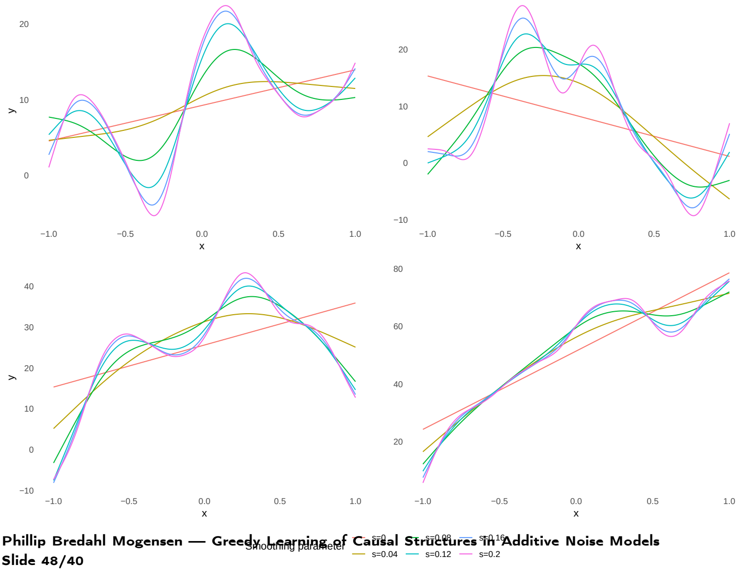
Asymptotics of entropy estimator, $n = 250$



Linear assignments



Examples of random functions



References I

Brendan D. McKay, Frederique E. Oggier, Gordon F. Royle, N. J. A. Sloane, Ian M. Wanless, and Herbert S. Wilf. Acyclic digraphs and eigenvalues of (0,1)-matrices. 2003.

Jonas Martin Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of causal inference, foundations and learning algorithms*. Adaptive computation and machine learning series. MIT Press, Cambridge, MA, 2017. ISBN 9780262037310.