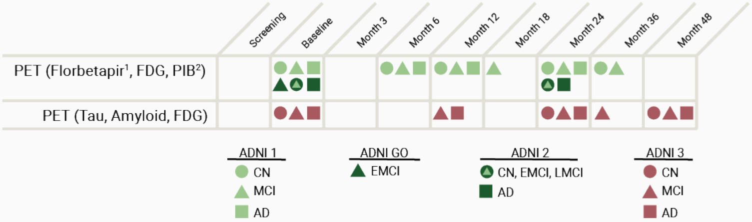
Causal structure learning and sampling using Markov Monte Carlo with momentum

Moritz Schauer
Chalmers University of Technology | University of Gothenburg
With Marcel Wienöbst, University of Lübeck.

Cramér Society 2023

Alzheimer's Disease Neuroimaging Initiative

Alzheimer's Disease Neuroimaging Initiative (ADNI)



[1]

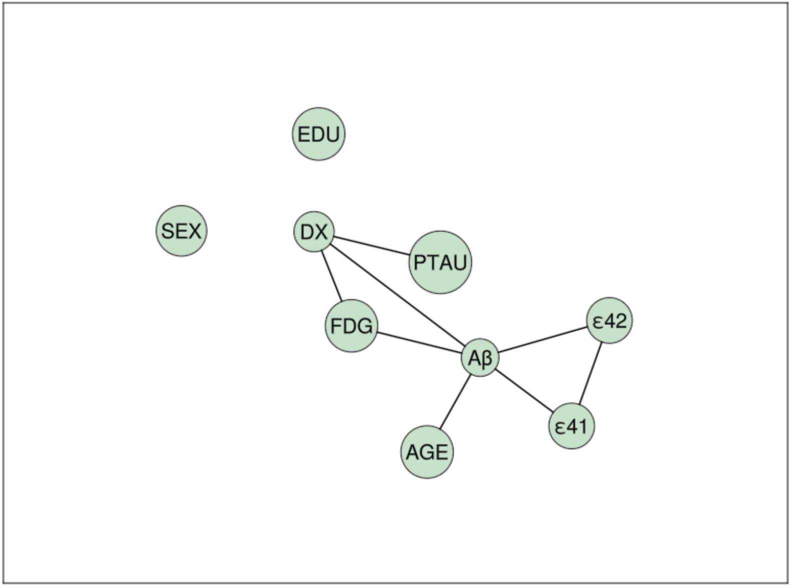
ADNI is a longitudinal study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of Alzheimer's disease (AD).

[1] Image: ADNI.

ADNI data

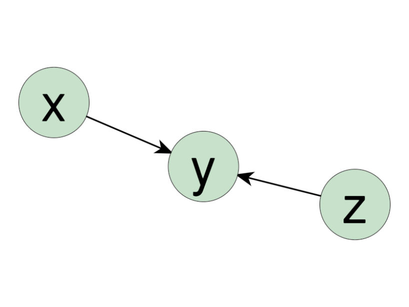
SEX	FDG	DX	Aβ	ε4	PTAU	EDU	AGE
Male	1.13615	CN	$\times$	0	0	16	78.3
Male	1.3086	Dementia	721.5	2	22.83	18	81.3
Male	$\times$	MCI	1501	0	13.29	10	67.5
Male	1.25956	CN	547.3	0	31.43	16	70.7
Female	$\times$	MCI	$\times$	0	$\times$	13	81.4
...							

Associations between variables



Graphical models

DAGs



A **DAG** is a directed graph such that following arrows it is impossible to return to any vertex (no cycles). A **PDAG** has additional undirected edges.

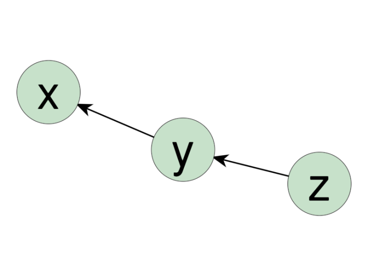
Vertices x, y, z correspond to stochastic variables.

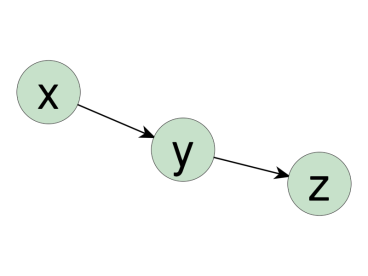
Classical (Bayesian) statistics

A single joint density on unknowns and observables. Different possible factorizations:

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

Classical (Bayesian) statistics





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

Typically some Markovian properties, for example here

$$x \perp\!\!\!\perp z \mid y$$

and corresponding factorizations of that density.

Faithfulness assumption

Assume **perfect correspondence** between law p and DAG $\mathcal{G}$

$$x \underset{p}{\perp\!\!\!\perp} z \mid y \quad \Leftrightarrow \quad x \underset{\mathcal{G}}{\perp\!\!\!\perp} z \mid y$$

Faithfulness assumption

Assume **perfect correspondence** between law p and DAG $\mathcal{G}$

$$x \underset{p}{\perp\!\!\!\perp} z \mid y \quad \Leftrightarrow \quad x \underset{\mathcal{G}}{\perp\!\!\!\perp} z \mid y$$

For example $\mathcal{G}$ having no edges is equivalent to complete independence under p.

Faithfulness assumption

Assume **perfect correspondence** between law p and DAG $\mathcal{G}$

$$x \underset{p}{\perp\!\!\!\perp} z \mid y \quad \Leftrightarrow \quad x \underset{\mathcal{G}}{\perp\!\!\!\perp} z \mid y$$

For example $\mathcal{G}$ having no edges is equivalent to complete independence under p.

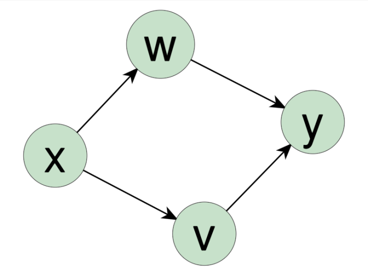
The implication "$\Rightarrow$" is **faithfulness**.

Faithfulness violation

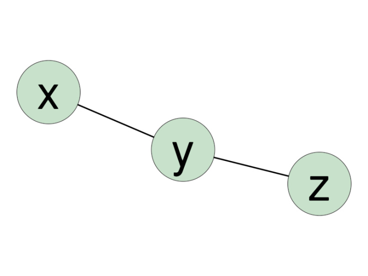
$z_1, \dots, z_4$ independent noise and

$$x = z_1$$
$$w = x + z_2$$
$$v = -x + z_3$$
$$y = v + w + z_4.$$

Independence $x \perp\!\!\!\perp y$ not implied by the DAG.

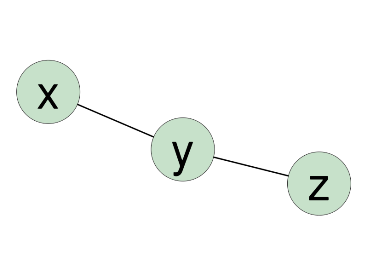


DAG discovery



What are the possible DAG models (edge orientations) under faithfulness?

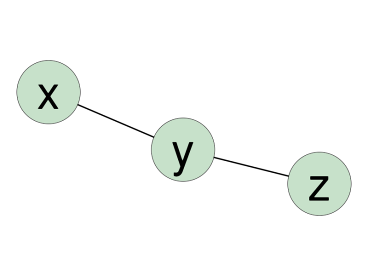
DAG discovery



What are the possible DAG models (edge orientations) under faithfulness?

- $x \perp\!\!\!\perp z$ implies $x \rightarrow y \leftarrow z$

DAG discovery

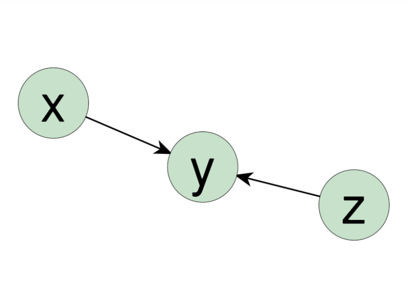


What are the possible DAG models (edge orientations) under faithfulness?

- $x \perp\!\!\!\perp z$ implies $x \rightarrow y \leftarrow z$
- $x \not\!\perp\!\!\!\perp z$ is compatible with all others

$$\underbrace{x \leftarrow y \leftarrow z \qquad x \rightarrow y \rightarrow z \qquad x \leftarrow y \rightarrow z}_{\text{Markov equivalence class (MEC)}}$$

v-structures



x, y, and z in a DAG form an **v-structure** if $x \to y \leftarrow z$ and x and z are not adjacent.

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

Collider bias

Among babies of low birth weight (y) maternal smoking (x) was associated with lower infant mortality.

Collider bias

Among babies of low birth weight (y) maternal smoking (x) was associated with lower infant mortality.

Think: if z are other independent causes of low birth weight, then

$$x \to y \leftarrow z \quad \text{and} \quad x \not\perp\!\!\!\perp z \mid y.$$

Markov equivalence classes

All DAGs on a vertex set V with n vertices with the same set of v-structures and the same set of adjacencies are observationally equivalent and form the **Markov equivalence class (MEC)** denoted $\mathcal{M}_n$ (Verma and Pearl, 1990)

CPDAG

A way to represent a MEC is the CPDAG (completed PDAG):

Arrows $x \rightarrow y$ only if all members of the equivalence class agree on the direction; undirected edges $x - y$ otherwise.

CPDAG

A way to represent a MEC is the CPDAG (completed PDAG):

Arrows $x \to y$ only if all members of the equivalence class agree on the direction; undirected edges $x - y$ otherwise.

$\mathcal{M}_n$ is the space of CPDAGs or MECs with elments denoted $\gamma, \eta, \dots \in \mathcal{M}_n$.

Two variables

What are the MECs on two variables x and y?

Two variables

What are the MECs on two variables x and y?

No v-structures, so

$$\gamma_0 = \text{“}x \quad y\text{''} = \{\text{“}x \quad y\text{”}\}$$
$$\gamma_1 = \text{“}x - y\text{''} = \{\text{“}x \rightarrow y\text{”}, \text{“}x \leftarrow y\text{”}\}$$

Two variables

What are the MECs on two variables x and y?

No v-structures, so

$$\gamma_0 = \text{``}x \quad y\text{''} = \{\text{``}x \quad y\text{''}\}$$
$$\gamma_1 = \text{``}x - y\text{''} = \{\text{``}x \rightarrow y\text{''}, \text{``}x \leftarrow y\text{''}\}$$

Not so nice when trying to infer causation from association.

Two variables

What are the MECs on two variables x and y?

No v-structures, so

$$\gamma_0 = \text{“}x \quad y\text{''} = \{\text{“}x \quad y\text{”}\}$$
$$\gamma_1 = \text{“}x - y\text{''} = \{\text{“}x \to y\text{”}, \text{“}x \leftarrow y\text{”}\}$$

Not so nice when trying to infer causation from association.

But: Re-factorizing $p(\text{data}|\theta)p(\theta)$ as $p(\theta|\text{data})p(\text{data})$ has some applications.

Characterisation of CPDAGs

THEOREM 4.1 (Characterization of D^*). *A graph $G \equiv (V, E)$ is equal to D^* for some ADG D if and only if G satisfies the following four conditions.*

(i) *G is a chain graph.*
(ii) *For every chain component τ of G, G_τ is chordal.*
(iii) *The configuration $a \to b — c$ does not occur as an induced subgraph of G.*
(iv) *Every arrow $a \to b \in G$ is strongly protected in G.*

S. A. Andersson, D. Madigan and M. D. Perlman, "A characterization of Markov equivalence classes for acyclic digraphs", Annals of Statistics 25 (1997) 505-541.

Learn the Markov equivalence class

Causal Structure Learning With Momentum: Sampling Distributions Over Markov Equivalence Classes of DAGs

Moritz Schauer
Chalmers University of Technology
and University of Gothenburg

Marcel Wienöbst
Institute for Theoretical Computer Science,
University of Lübeck

Abstract

In the context of inferring a Bayesian network structure (directed acyclic graph, DAG for short), we devise a non-reversible continu-

1 Introduction

A Bayesian network is a probabilistic graphical model that represents a set of random variables and their conditional (in)dependencies using a directed acyclic graph (DAG). Graph and random variables are linked

Learn the Markov equivalence class

Causal Structure Learning With Momentum: Sampling Distributions Over Markov Equivalence Classes of DAGs

Moritz Schauer
Chalmers University of Technology
and University of Gothenburg

Marcel Wienöbst
Institute for Theoretical Computer Science,
University of Lübeck

Abstract

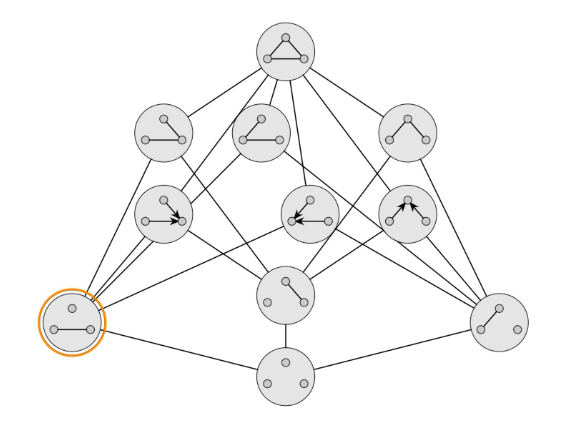
In the context of inferring a Bayesian network structure (directed acyclic graph, DAG for short), we devise a non-reversible continu-

1 Introduction

A Bayesian network is a probabilistic graphical model that represents a set of random variables and their conditional (in)dependencies using a directed acyclic graph (DAG). Graph and random variables are linked

Random walk on CPDAGs

Random walks for causal models



Uniform random walk on a graph

Random walk visits vertices with many edges more often, so needs to spends less time there:

$$\mathbb{E}[\tau_v] = \frac{1}{\deg(v)}, \quad \tau_v \text{ residence time in } v$$

After τ_v time units, the process jumps to a neighbour (picked at random.)

Uniform random walk on a graph

Random walk visits vertices with many edges more often, so needs to spends less time there:

$$\mathbb{E}[\tau_v] = \frac{1}{\deg(v)}, \quad \tau_v \text{ residence time in } v$$

After τ_v time units, the process jumps to a neighbour (picked at random.)

With Markovianity

$$\tau_v \sim \mathrm{Exp}(\deg(v))$$

Neighbours??

Declare adjacency between CPDAGs:

$$\gamma = \{\text{“}x \to y \quad z\text{”}, \text{“}x \leftarrow y \quad z\text{”}\}(= \text{“}x - y \quad z\text{”})$$

and

$$\eta = \{x \to y \leftarrow z\}$$

are **neighbours**, because I can insert an edge into “$x \to y \quad z$” to obtain “$x \to y \leftarrow z$”.

Neighbours??

Declare adjacency between CPDAGs:

$$\gamma = \{\text{``}x \rightarrow y \quad z\text{''}, \text{``}x \leftarrow y \quad z\text{''}\} (= \text{``}x - y \quad z\text{''})$$

and

$$\eta = \{x \rightarrow y \leftarrow z\}$$

are **neighbours**, because I can insert an edge into "$x \rightarrow y \quad z$" to obtain "$x \rightarrow y \leftarrow z$".

Notation: $\eta \in \mathrm{Insert}(\gamma)$, $\gamma \in \mathrm{Delete}(\eta)$.

Chickering's operators

The operator Insert(γ, x, y, T) inserts the edge $x \rightarrow y$ to the CPDAG γ and directs previously undirected edges $t - y$ to $t \rightarrow y$ for $t \in T$, such that vertices $t \in T$ become "tails" of a v-structure $t \rightarrow y \leftarrow x$.

Chickering's operators

The operator Insert(γ, x, y, T) inserts the edge $x \to y$ to the CPDAG γ and directs previously undirected edges $t - y$ to $t \to y$ for $t \in T$, such that vertices $t \in T$ become "tails" of a v-structure $t \to y \leftarrow x$.

Fineprint: Here x and y are not adjacent and T are (undirected) neighbours of y that are not adjacent to x. The resulting PDAG is then completed.

Valid moves

Denote by $\mathrm{NA}_x(y)$ the (undirected) neighbours of y that are adjacent to x.

Valid moves

Denote by $\mathrm{NA}_x(y)$ the (undirected) neighbours of y that are adjacent to x.

$\mathrm{Insert}(\gamma, x, y, T)$ is a valid move, if and only if

- $\mathrm{NA}_x(y)$ and the elements of T form a clique and
- any path from y to x without a directed edge pointing towards y (such a path is called semi-directed) contains a vertex in $\mathrm{NA}_x(y) \cup T$.

Valid moves

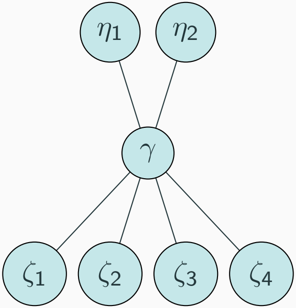
Denote by $\mathrm{NA}_x(y)$ the (undirected) neighbours of y that are adjacent to x.

Insert(γ, x, y, T) is a valid move, if and only if

- $\mathrm{NA}_x(y)$ and the elements of T form a clique and
- any path from y to x without a directed edge pointing towards y (such a path is called semi-directed) contains a vertex in $\mathrm{NA}_x(y) \cup T$.

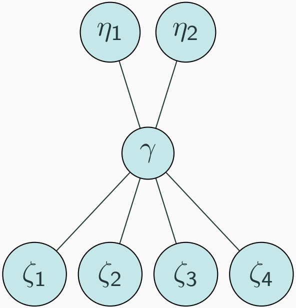
Story for the delete operator is a bit simpler

Random walk on CPDAGs



A MEC γ with two neighbours η_1, η_2 in Insert(γ) and four neighbours $\zeta_1, \ldots, \zeta_4$ in Delete(γ). This is a lattice!

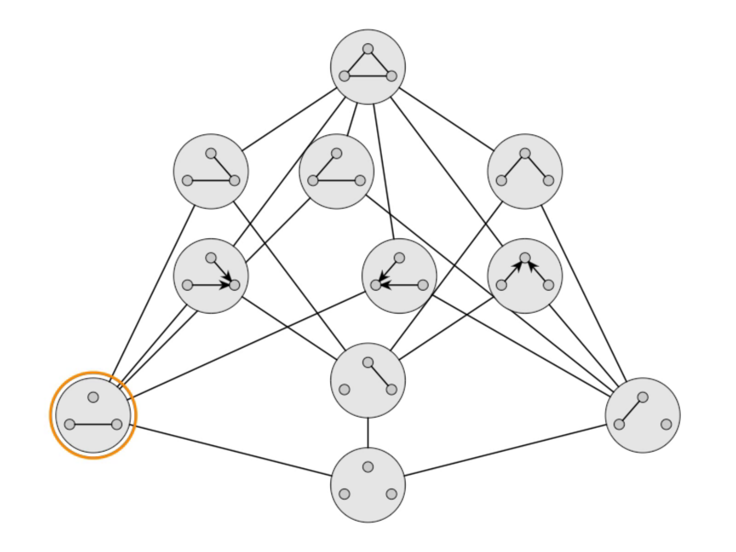
Random walk on CPDAGs



Random walk will leave γ after an exponentially distributed time with total rate $\Lambda(\gamma) = 6$ towards one of the six neighbours drawn from $\kappa_\gamma = \mathrm{U}(\{\eta_1, \eta_2, \zeta_1, \zeta_2, \zeta_3, \zeta_4\})$. (Not so easy to count for large graphs...)

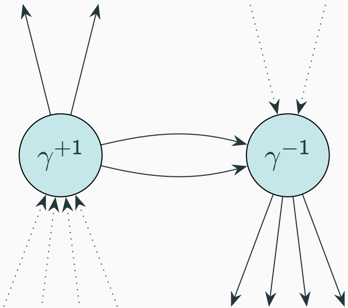
Adding momentum

Lifted random walk for causal models



Lifted random walk

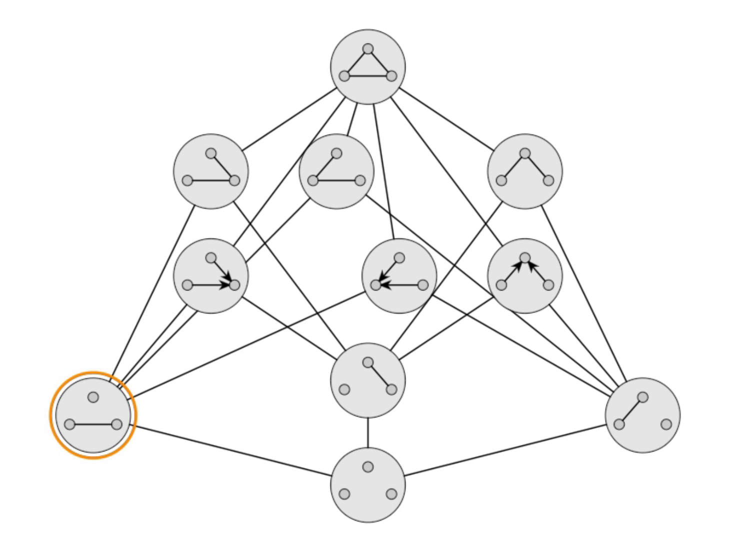
$\gamma^{+1} := (\gamma, +1) \qquad \gamma^{-1} := (\gamma, -1)$



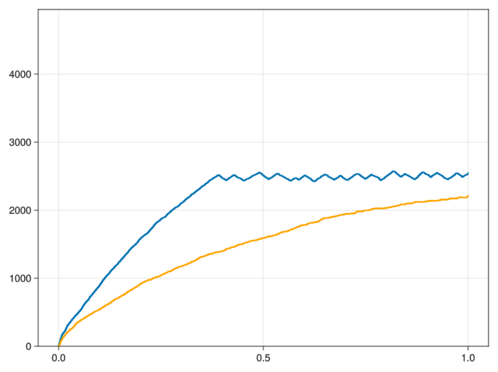
If $\gamma \in \mathcal{M}_n$ has 2 direct neighbours in Insert(γ) and 4 direct neighbours in Delete(γ):

Move up from γ^{+1} with total rate 2, move from γ^{+1} to γ^{-1} with rate $2 = 4 - 2$ and down from γ^{-1} with total rate 4.

Lifted random walk



Mixing

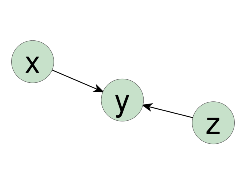


Continuous-time trace of the number of edges of the sampled graphs when targeting a uniform distribution on CPDAGs with 100 vertices. Blue: Lifted, orange: Normal.

The total time of 1 unit corresponds to 5 000 jumps.

Causal discovery

Causality



One DAG and a corresponding factorization

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

can describe a family different of joint densities corresponding to different interventions:

$$p_{do(z=z_0)}(x, y) = p(x)p(y \mid x, z_0) \neq p(x, y)$$

$$p_{do(y=y_0)}(x, z) = p(x)p(z) = p(x, z)$$

Causal discovery

Difficult problem: Learn a causal model from observational data.

Causal discovery

Difficult problem: Learn a causal model from observational data.

Assuming that all relevant variables are observed, the causal model is in the observational MEC.

Causal discovery

Difficult problem: Learn a causal model from observational data.

Assuming that all relevant variables are observed, the causal model is in the observational MEC.

If you know the MEC, you can think of experiments to pin down the causal relationships further, e.g. by gene knockouts.

Score based causal discovery

Markov equivalent score

A scoring function for DAGs is a **Markov equivalent score** if it assigns the same score to any DAG in the same MEC.

Markov equivalent score

A scoring function for DAGs is a **Markov equivalent score** if it assigns the same score to any DAG in the same MEC.

Example: Bayesian information criterion (BIC).

Markov equivalent score

A scoring function for DAGs is a **Markov equivalent score** if it assigns the same score to any DAG in the same MEC.

Example: Bayesian information criterion (BIC).

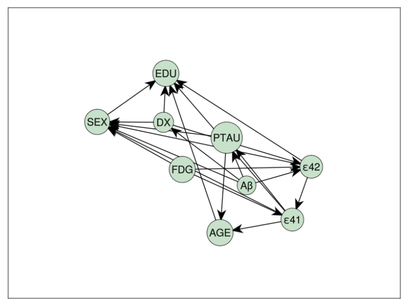
Exponentiated BIC score factorises over the DAGs

$$w(G, \mathrm{Data}) = \prod_{x \in V} w(\mathrm{Pa}_G(x), x, \mathrm{Data}),$$

Changes in w can be computed efficiently by comparing local scores.

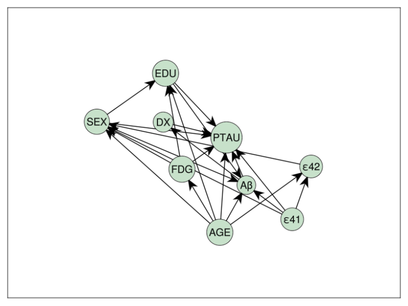
ADNI

There are 1213442454842881 (1.2 quadrillion) directed acyclic graphs on 9 vertices. These are some of them.



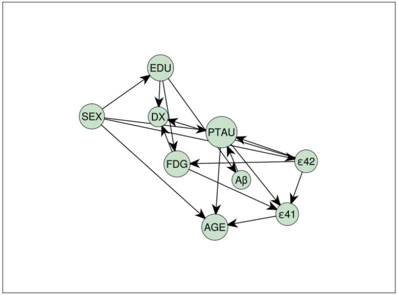
ADNI

There are 1213442454842881 (1.2 quadrillion) directed acyclic graphs on 9 vertices. These are some of them.



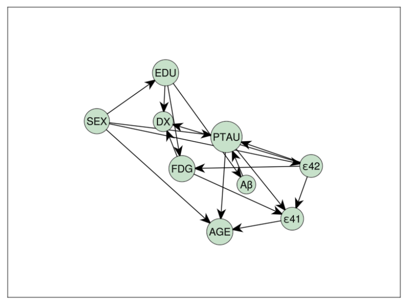
ADNI

There are 1213442454842881 (1.2 quadrillion) directed acyclic graphs on 9 vertices. These are some of them.



ADNI

There are 1213442454842881 (1.2 quadrillion) directed acyclic graphs on 9 vertices. These are some of them.



Given the graph structure, perform regression on the parents of a variable to obtain a model, e.g.

$$\mathrm{DX} = \beta_1 \cdot \mathrm{EDU} + \beta_2 \cdot \mathrm{FDG} + \beta_3 \cdot \mathrm{PTAU} + \text{error term.}$$

Zanella process

Continuous time random walk to sample from a distribution π defined on $\mathcal{M}_n$.

Zanella process

Continuous time random walk to sample from a distribution π defined on $\mathcal{M}_n$.

Like in Metropolis-Hastings we need a balancing function g such as $\sqrt{t}$ or $\min(1, t)$ with the property $g(t) = tg(1/t)$.

Zanella process

Continuous time random walk to sample from a distribution π defined on $\mathcal{M}_n$.

Like in Metropolis-Hastings we need a balancing function g such as $\sqrt{t}$ or $\min(1, t)$ with the property $g(t) = tg(1/t)$.

The Zanella process is defined by the jump intensity

$$\lambda(\gamma \curvearrowright \eta) = \begin{cases} g\left(\dfrac{\pi\{\eta\}}{\pi\{\gamma\}}\right) & \text{if } \eta \in \mathrm{Insert}(\gamma) \sqcup \mathrm{Delete}(\gamma) \\ 0 & \text{otherwise} \end{cases},$$

where $\gamma \in \mathcal{M}_n$.

Lifted Zanella process

A random walk on $\mathcal{M}_n \times \{+1, -1\}$ with correct marginal:

Lifted Zanella process

A random walk on $\mathcal{M}_n \times \{+1, -1\}$ with correct marginal:

For $\gamma \in \mathcal{M}_n$,

$$\lambda(\gamma^{+1} \curvearrowright \eta^{+1}) = \begin{cases} g\left(\frac{\pi\{\eta\}}{\pi\{\gamma\}}\right) & \text{if } \eta \in \mathrm{Insert}(\gamma) \\ 0 & \text{otherwise.} \end{cases}$$

Lifted Zanella process

A random walk on $\mathcal{M}_n \times \{+1, -1\}$ with correct marginal:

For $\gamma \in \mathcal{M}_n$,

$$\lambda(\gamma^{+1} \curvearrowright \eta^{+1}) = \begin{cases} g\left(\dfrac{\pi\{\eta\}}{\pi\{\gamma\}}\right) & \text{if } \eta \in \mathsf{Insert}(\gamma) \\ 0 & \text{otherwise.} \end{cases}$$

$$\lambda(\eta^{-1} \curvearrowright \gamma^{-1}) = \begin{cases} g\left(\dfrac{\pi\{\gamma\}}{\pi\{\eta\}}\right) & \text{if } \gamma \in \mathsf{Delete}(\eta) \\ 0 & \text{otherwise.} \end{cases}$$

Lifted Zanella process

A random walk on $\mathcal{M}_n \times \{+1, -1\}$ with correct marginal:

For $\gamma \in \mathcal{M}_n$,

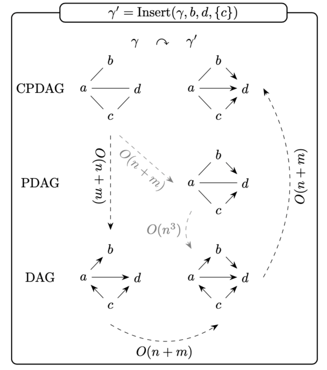
$$\lambda(\gamma^{+1} \curvearrowright \eta^{+1}) = \begin{cases} g\left(\frac{\pi\{\eta\}}{\pi\{\gamma\}}\right) & \text{if } \eta \in \mathrm{Insert}(\gamma) \\ 0 & \text{otherwise.} \end{cases}$$

$$\lambda(\eta^{-1} \curvearrowright \gamma^{-1}) = \begin{cases} g\left(\frac{\pi\{\gamma\}}{\pi\{\eta\}}\right) & \text{if } \gamma \in \mathrm{Delete}(\eta) \\ 0 & \text{otherwise.} \end{cases}$$

and for $\gamma \in \mathcal{M}_n$ and $d \in \{-1, +1\}$,

$$\lambda(\gamma^{d} \curvearrowright \gamma^{-d}) = \left(-\sum_{\eta} \lambda(\gamma^{d} \curvearrowright \eta^{d}) + \sum_{\eta} \lambda(\gamma^{-d} \curvearrowright \eta^{-d})\right)^{+}.$$

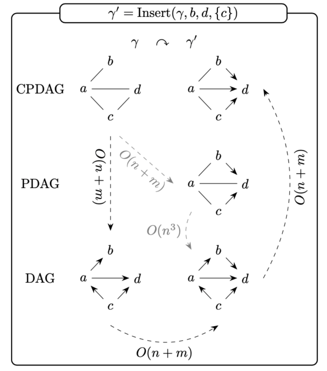
Moving efficiently



Linear-time approach for applying a GES operator.

Previous approaches: add the inserted edge to the initial CPDAG, obtaining a PDAG associated with the new MEC γ'.

Moving efficiently



Linear-time approach for applying a GES operator.

Our approach: find a consistent DAG extension of the initial CPDAG in time $O(n+m)$, which has the property that applying the operator directly yields a DAG from γ'.

What else is there?

- Plug and play: `CausalInference.jl`
- Intriguing connection to the GES algorithms (greedy search for the MEC which maximises score).
- Some ideas how to handle unobserved confounders.

What causal models does the ADNI data suggest?