Statistical Decidability in Linear, Non-Gaussian Causal Models

Konstantin Genin[*]
Cluster of Excellence – Machine Learning for Science
Universität Tübingen

konstantin.genin@uni-tuebingen.de

Conor Mayo-Wilson[†]
Department of Philosophy
University of Washington

conormw@uw.edu

Orthodox Statistics

	X	*Y*	*Z*
Sample point 1	0	1	1
Sample point 2	1	1	1
Sample point 3	0	0	1

Sample

Statistics

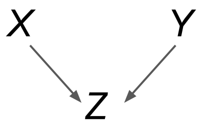
Probability Theory

$X \sim$ Bernoulli(.25)
$Y \sim$ Bernoulli(.8)
$X \perp Y$
Z = max(X,Y)

"True" Distribution

Causal Discovery

Model Assumptions
(e.g., Linear, Normal, Faithful, etc.)



$Z := \max(X,Y)$

$X \sim \text{Bernoulli}(.25)$
$Y \sim \text{Bernoulli}(.8)$
$X \perp Y$
$Z = \max(X,Y)$

Causal Structure

“True” Distribution

Causal Discovery

	X	*Y*	*Z*
Sample point 1	0	1	1
Sample point 2	1	1	1
Sample point 3	0	0	1

Sample

Counterfactual and Hypothetical Predictions

Orthodox statistics attempts to estimate only the distribution from which one samples.

Causal discovery attempts to estimate and predict outcomes of hypothetical (or counterfactual) interventions, which are features of distributions other than those from which one samples.

Identifiability vs. Success Criteria (e.g. Consistency)

Researchers often prove a set of causal models is ***identifiable,*** i.e., that different models give rise to different sampling distributions.

But identifiability doesn't guarantee the existence of ***consistent estimator***, i.e., a method that produces increasingly accurate estimates with increasing probability (Gabrielsen 1978).

- Trivial Example: Estimating Bernoulli parameter with the discrete metric.

Identifiability is a relation between models/parameters and distributions; it is not a success criteria for estimators (like consistency).

Genin, Arne (1978). "Consistency and Identifiability," *Journal of Econometrics.* 8. 261-263.

Success Criteria

Most research focuses on only two asymptotic, success/reliability criteria for estimators:

- (Pointwise) consistency/convergence - Estimates become more accurate with increasing probability, but there's no bound on how much data might be necessary.

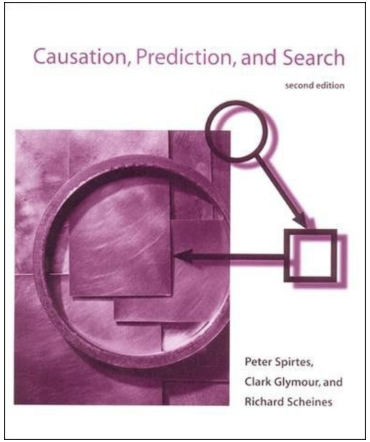
- Uniform consistency - One can name, *a priori*, how much data is necessary to achieve particular error bounds with particular probabilities.

Arithmetical Hierarchy vs. Statistical Hierarchy

Computability Theory	**Statistics**
Known bound on number of computational steps before terminating	Uniform consistency
Δ^0_2 i.e., "Trial and Error predicates" (Putnam 1965)	Consistency

Putnam, Hilary . "Trial and Error Predicates and a Solution to a Problem of Mostowski." *Journal of Symbolic Logic* 7, 30 (1) 1965.

The Linear Gaussian Model



Theorem (Spirtes et al., 2001)**.** When

1. noise terms are **independent** and **Gaussian**,
2. functional relationships are **linear** and **a-cyclic** and
3. there are no unobserved confounders,

it is possible to converge (pointwise) to the **Markov equivalence class** of the DAG generating the data.

The LiNGAM Model

Theorem (Shimizu et al., 2006)**.** When

1. noise terms are **independent** and **non-**Gaussian**,**
2. functional relationships are **linear** and **a-cyclic** and
3. there are no unobserved confounders,

it is possible to converge (pointwise) to the **DAG** generating the data.

Shimizu, Shohei, Patrik O. Hoyer, Aapo Hyvärinen, Aapo, and Antti Kerminen. “A Linear Non-Gaussian Acyclic Model for Causal Discovery.” *Journal of Machine Learning Research* 7, no. 72 (2006): 2003–30.

Carnap’s Critique of Consistency

“Reichenbach is **right** …

any procedure, which does not [converge in the limit] is **inferior** to his rule of induction.

However, … the same holds for an **infinite** number of **other** rules of induction ….”

Carnap’s Critique of Consistency

“... therefore we need a **more general** and **stronger** method for examining and comparing any two given rules of induction.”

On Inductive Logic, 1945.

Pitfalls of Pointwise

Further, **pointwise convergence** is compatible with all kinds of short run behavior.

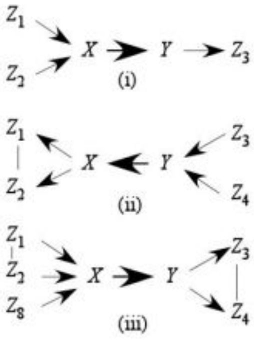


Figure 1: Causal Flips

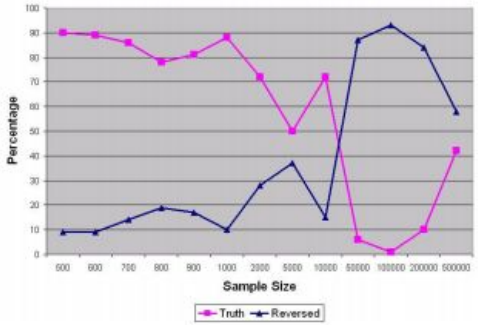


Figure 5: FCI Algorithm

Kelly, Kevin T, and Conor Mayo-Wilson (2010). “Causal Conclusions That Flip Repeatedly and Their Justification,” Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI 2010). https://arxiv.org/abs/1203.3488v1

Pitfalls of Pointwise

If noise is Gaussian, causal conclusion can **flip** arbitrarily often as data accumulates.

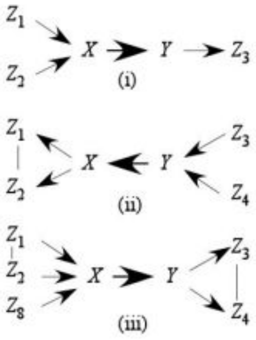


Figure 1: Causal Flips

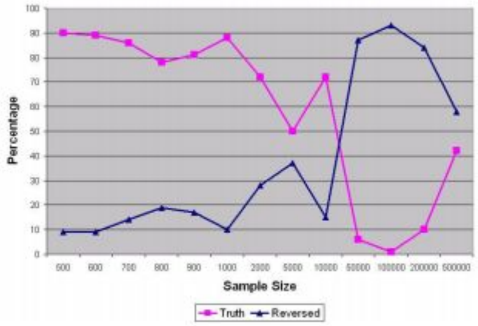


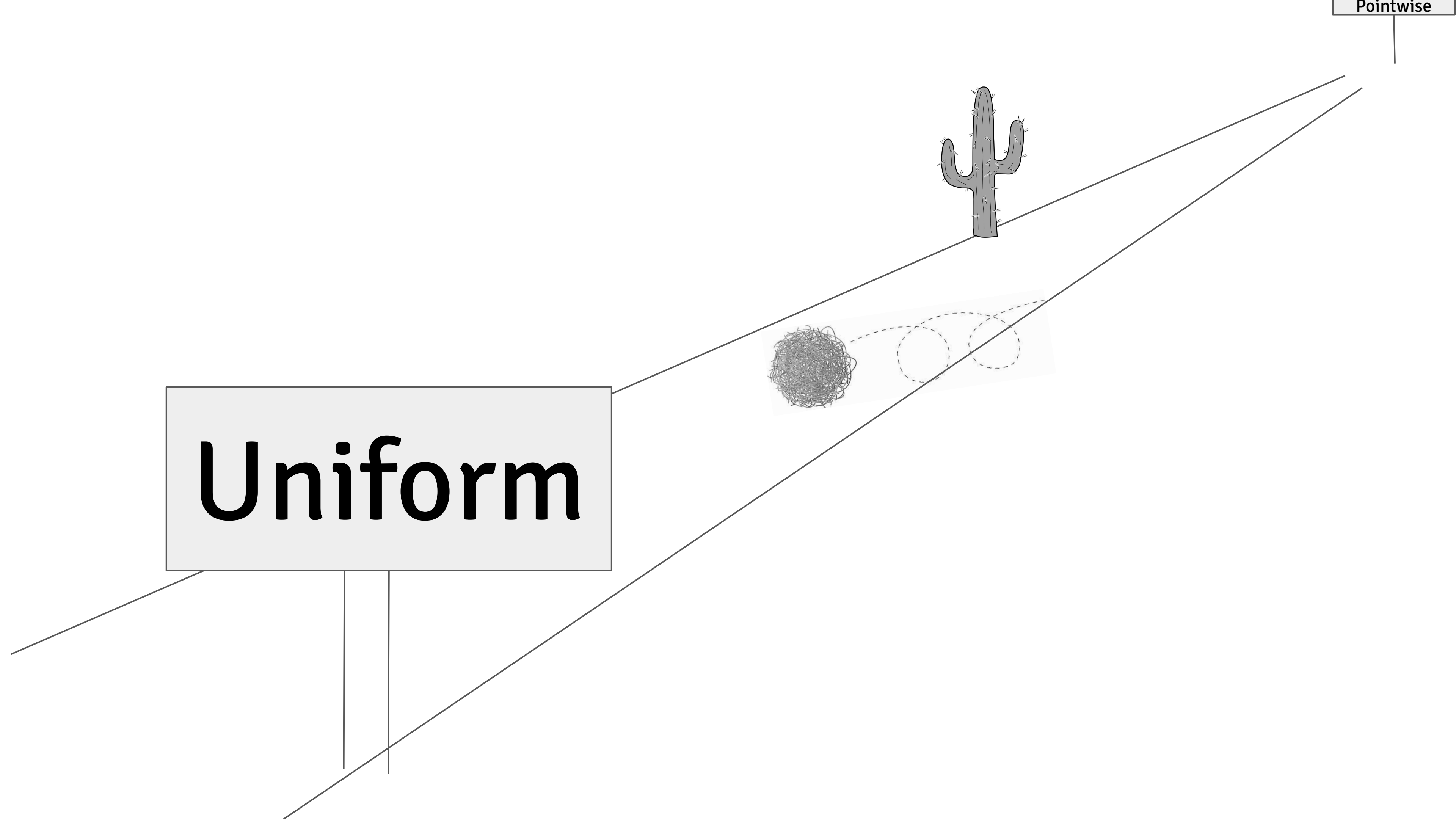
Figure 5: FCI Algorithm

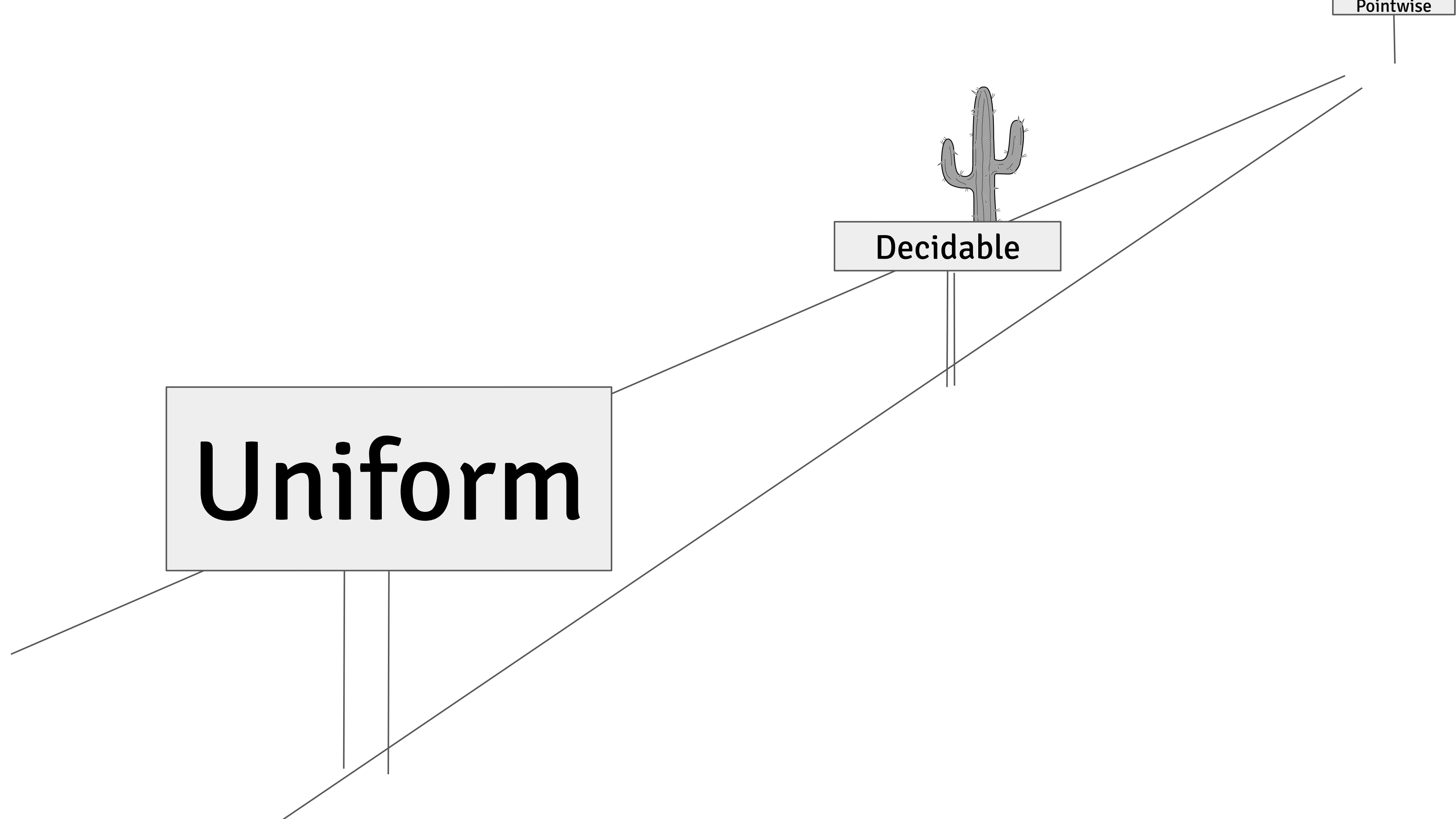
Kelly, Kevin T, and Conor Mayo-Wilson (2010). "Causal Conclusions That Flip Repeatedly and Their Justification," Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence (UAI 2010). https://arxiv.org/abs/1203.3488v1

Uniform Convergence is Impossible

Shimizu 2006 provide it is possible to converge (pointwise) to the **DAG** generating the data in the LiNGAM framework.

But **uniform** convergence to the true DAG is provably **impossible** in the LiNGAM framework.


Pointwise
Uniform


Pointwise
Decidable
Uniform

Arithmetical Hierarchy vs. Statistical Hierarchy

Computability Theory	**Statistics**
Known bound on number of computational steps before terminating	Uniform consistency
Decidable/Recursive - Δ^0_1	*Statistically Decidable - Consistency with bounded probability of error at every sample size*
Δ^0_2 i.e., “Trial and Error predicates” (Putnam 1965)	Consistent estimator exists

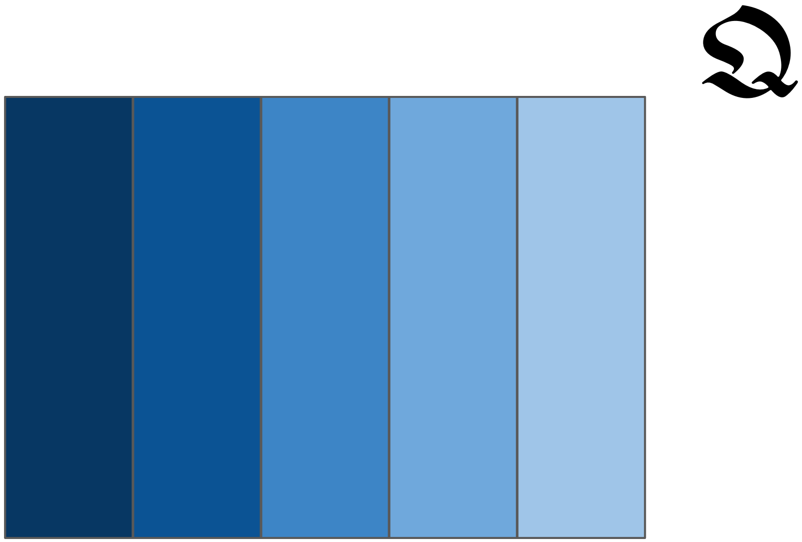
Putnam, Hilary . “Trial and Error Predicates and a Solution to a Problem of Mostowski.” *Journal of Symbolic Logic* 7, 30 (1) 1965.

Statistical Questions

Let $\mathcal{M}$ be a set of statistical models.

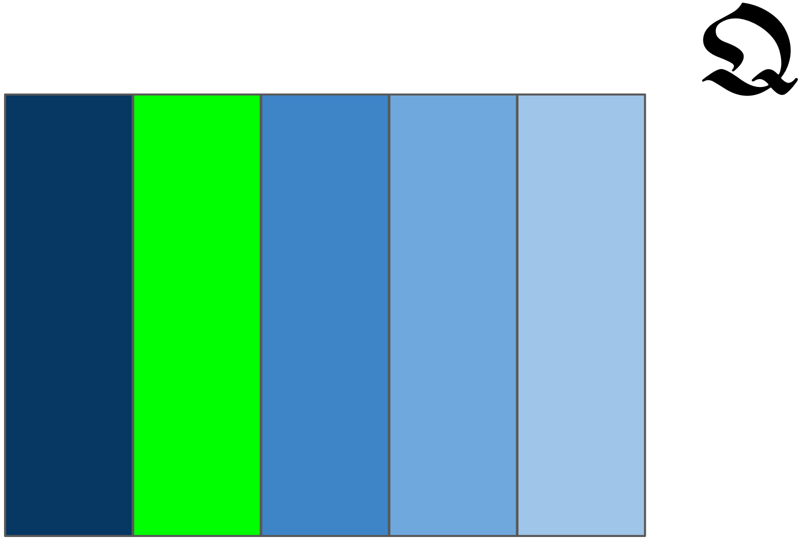
Statistical Questions

A **question** $\mathfrak{Q}$, partitioning $\mathscr{M}$ into a countable set of **answers.**



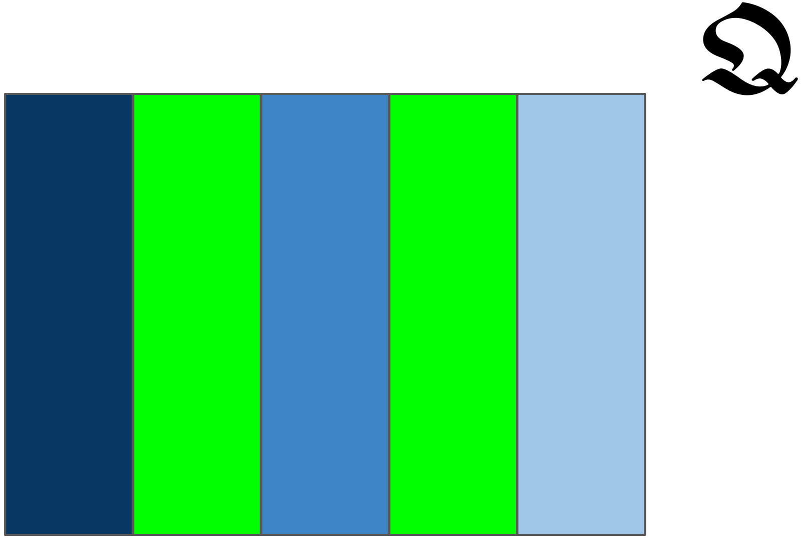
Statistical Questions

A **relevant response** is a **union** of answers.



Statistical Questions

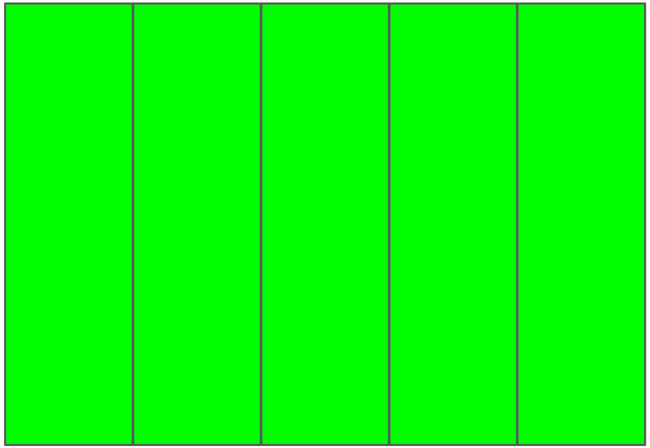
A **relevant response** is a **union** of answers.



Statistical Questions

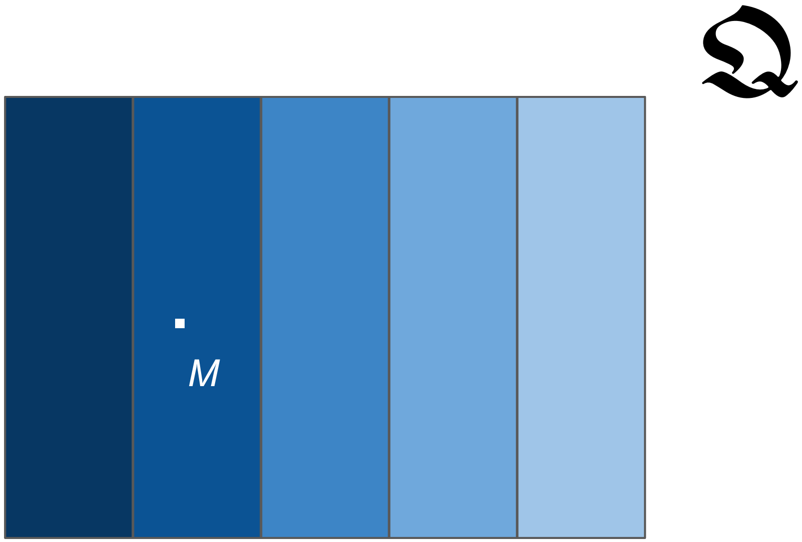
A **relevant response** is a **union** of answers.

$\mathfrak{Q}$



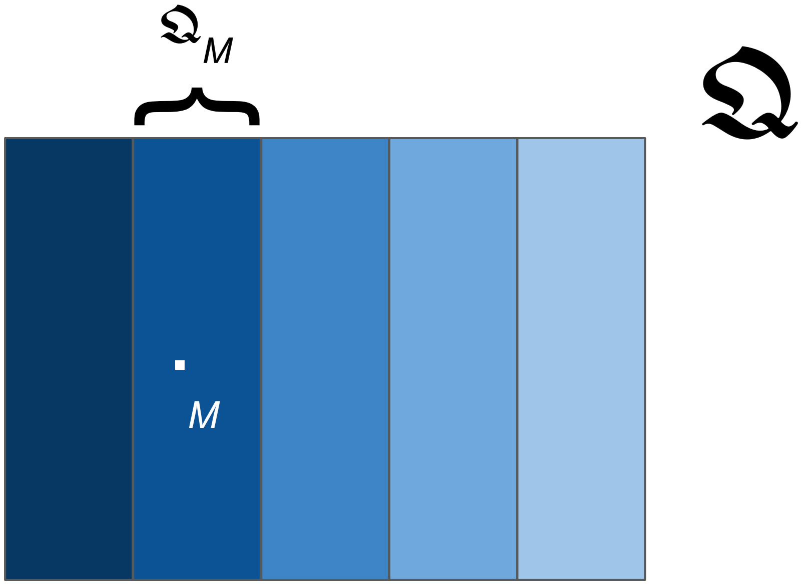
Statistical Questions

If $M \in \mathcal{M}$,



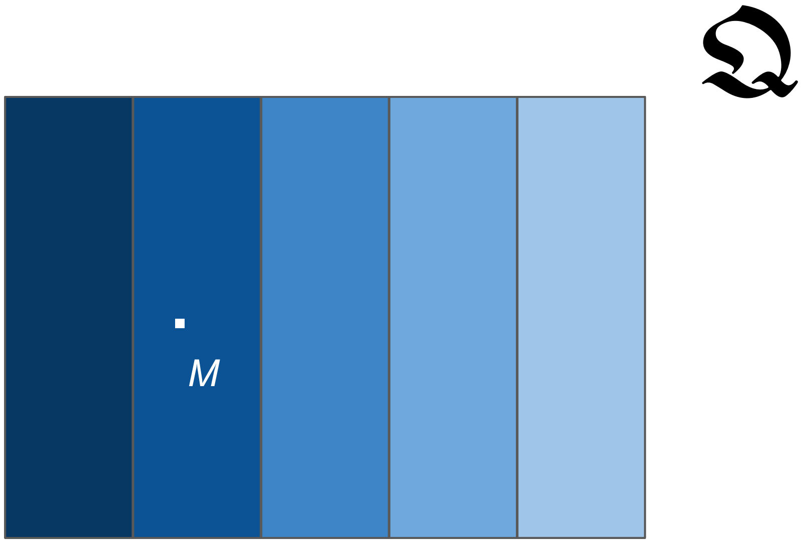
Statistical Questions

If $M \in \mathcal{M}$, let $\mathfrak{Q}_M$ be the answer true in M.



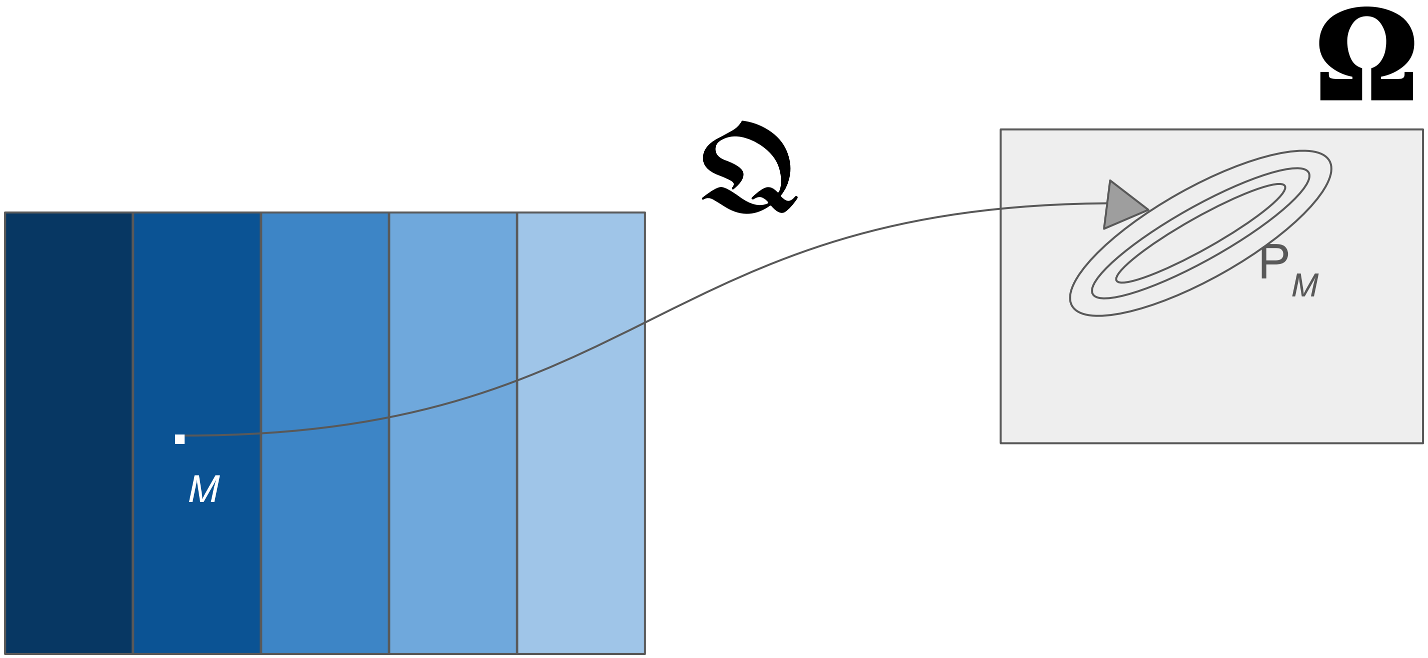
Statistical Questions

If $M \in \mathcal{M}$,



Statistical Questions

If $M \in \mathcal{M}$, let P_M be the distribution induced by M over observables.



Statistical Solution

A set of measurable functions (T_n) is a **method** if each one is a function from samples of size n to **relevant responses** (unions of answers).

Statistical Solution

A method (T_n) is a **solution** to $\mathfrak{Q}$ iff for all M $\in \mathcal{M}$,

- $P_M(T_n = \mathfrak{Q}_M) \longrightarrow 1$ as $n \longrightarrow \infty$.

Statistical Decision

A method (T_n) is an **α-decision procedure** for $\mathfrak{Q}$ iff it is a solution to $\mathfrak{Q}$ and

- for all sample sizes *n*, $P_M(\mathfrak{Q}_M \not\subseteq T_n) < \alpha$.

A question $\mathfrak{Q}$ is statistically **decidable** iff it has an α-decision procedure for all $\alpha > 0$.

Progressive Solutions

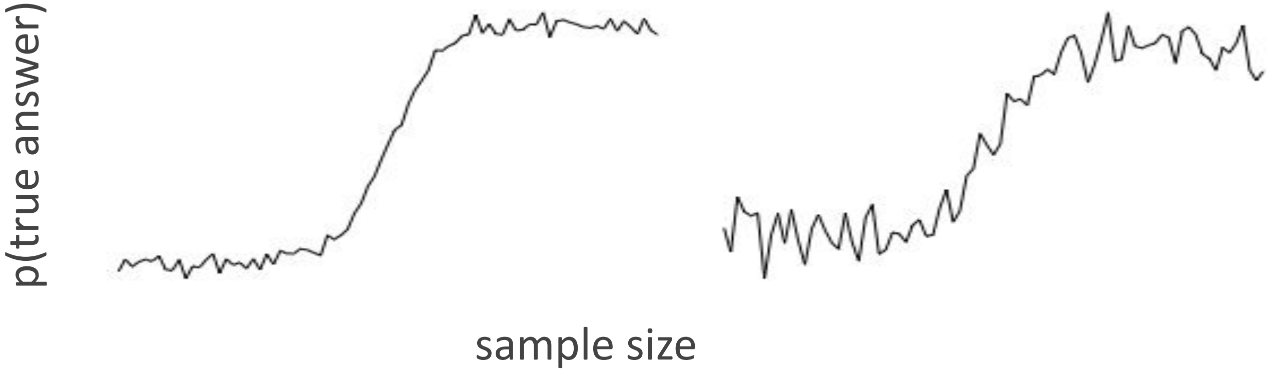
A method (T_n) is a **progressive** solution for $\mathfrak{Q}$ iff it is a solution to $\mathfrak{Q}$ and

- for all sample sizes $n_1 < n_2$, $P_M(T_{n1} = \mathfrak{Q}_M) < P_M(T_{n2} = \mathfrak{Q}_M)$.

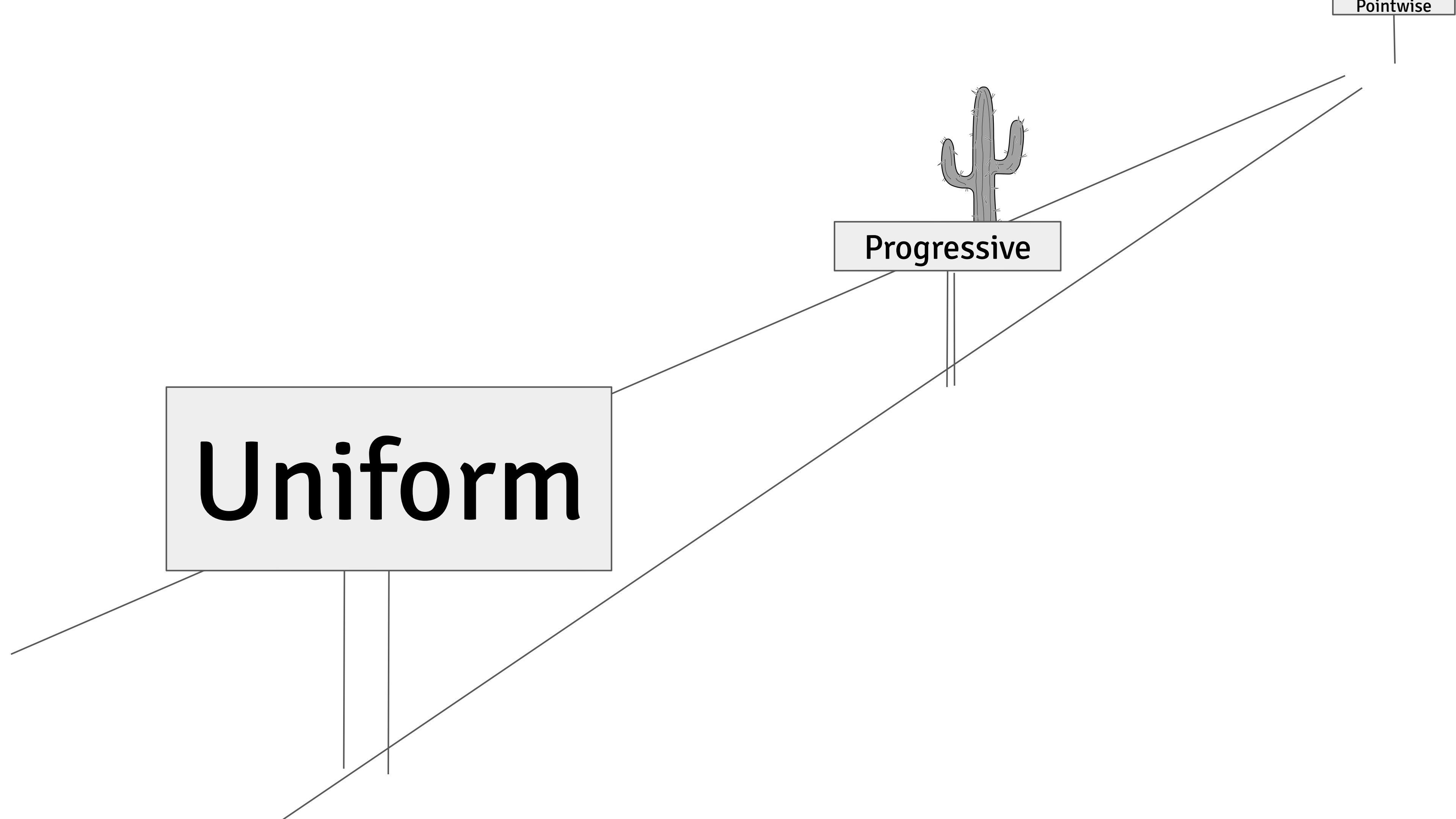
α-Progressive Solutions

A solution to $\mathcal{Q}$ (L_n) is α-**progressive** iff for all M in W and $n_1 < n_2$,

- $P_M(L_{n1} = \mathcal{Q}_M) < P_M(L_{n2} = \mathcal{Q}_M) + \alpha$.



Problem $\mathcal{Q}$ is **progressively solvable** iff it has an α-progressive solution for all $\alpha > 0$.


Pointwise
Progressive
Uniform

LiNGAM Questions

- Let $\textsc{Lng}_d$ be the set of all LiNGAM models on *d* observable variables.
- Let $\textsc{Lng}_d^c \subseteq \textsc{Lng}_d$ be the set of all models with causal coefficients bounded in absolute value by *c.*

(Suffices to let *c* be the number of particles in the universe.)

LiNGAM Questions

- $\mathcal{M}_{i\to j} \subseteq \mathrm{LNG}_d^c$ be the set of all models with $X_i \to X_j$,
- $\mathcal{M}_{i \circ j} \subseteq \mathrm{LNG}_d^c$ be the set of all models with X_i, X_j have no edge between them.

Main Results

Thm. The orientation question $\{\mathcal{M}_{i\to j}, \mathcal{M}_{i\to j}\}$ is **statistically decidable.**

Thm. The orientation question $\{\mathcal{M}_{ioj}, \mathcal{M}_{i\to j}, \mathcal{M}_{i\to j}\}$ is **progressively solvable.**

Thm. The DAG identification question $\{\mathcal{M}_G : G \in \mathrm{DAG}_d\}$ is **progressively solvable.**

Main Results

Thm. The orientation question $\{\mathcal{M}_{i\to j}, \mathcal{M}_{i\to j}\}$ is **statistically decidable.**

Thm. The orientation question $\{\mathcal{M}_{ioj}, \mathcal{M}_{i\to j}, \mathcal{M}_{i\to j}\}$ is **progressively solvable.**

Thm. The DAG identification question $\{\mathcal{M}_G : G \in \mathrm{DAG}_d\}$ is **progressively solvable.**

Flipping is Avoidable!

Existing algorithms make unnecessary mistakes/flips

- E.g., We applied Direct-LiNGAM to data from the model $X_1 \rightarrow X_2 \rightarrow X_3$ where
  - X_1 is uniform on {1,2…, 20},
  - $X_2=X_1+e_1$ & $X_3=X_2+e_2$, where e_1 and e_2 are independent Bernoulli variables with parameter ½.

Sample Size	$X_2 \rightarrow X_3$	$X_3 \rightarrow X_2$
50	55%	45%
5000	92%	8%

- **Moral:** Direct-LiNGAM orients edges when evidence is weak. Our main results show that algorithms could **wait** until evidence is strong enough to avoid reversing conclusions reached at earlier sample sizes.

LiNGAM + Confounding - Unfaithfulness

Theorem (Salehkaleybar et al., 2020)**.** When

1. noise terms are **independent** and **non-**Gaussian**,**
2. functional relationships are **linear** and **a-cyclic**,
3. there **may be** unobserved confounders, but
4. there are no cancelling paths (**faithfulness**),

then causal relationships between **observed** variables are identified.

Salehkaleybar, Saber, et al. (2020) "Learning Linear Non-Gaussian Causal Models in the Presence of Latent Variables." *Journal of Machine Learning Research* 21.39: 1-24.

LiNGAM + Confounding - Unfaithfulness

Theorem (Salehkaleybar et al., 2020)**.** When

1. noise terms are **independent** and **non-**Gaussian**,**
2. functional relationships are **linear** and **a-cyclic**,
3. there **may be** unobserved confounders, but
4. there are no cancelling paths (**faithfulness**),

then causal ancestry relationships between **observed** variables are **identified**.

But how *identified* are they, really?

Salehkaleybar, Saber, et al. (2020) "Learning Linear Non-Gaussian Causal Models in the Presence of Latent Variables." *Journal of Machine Learning Research* 21.39: 1-24.

Good News

Theorem (Genin, 2021)**.** When

1. noise terms are **independent** and **non-**Gaussian**,**
2. functional relationships are **linear** and **a-cyclic**,
3. there **may be** unobserved confounders, but
4. there are no cancelling paths (**faithfulness**),

then causal ancestry relationships between **observed** variables can be **consistently** identified.

Genin, Konstantin (2021). “Statistical Undecidability in Linear Non-Gaussian Models in the Presence of Latent Confounders,” *Neglected Assumptions in Causal Discovery Workshop,* ICML 2021*.*

Bad News

Theorem (Genin, 2021)**.** When

1. noise terms are **independent** and **non-**Gaussian**,**
2. functional relationships are **linear** and **a-cyclic**,
3. there **may be** unobserved confounders, but
4. there are no cancelling paths (**faithfulness**),

then causal ancestry relationships between **observed** variables are **not decidable**.

Genin, Konstantin (2021). “Statistical Undecidability in Linear Non-Gaussian Models in the Presence of Latent Confounders,” *Neglected Assumptions in Causal Discovery Workshop,* ICML 2021*.*

Bad News

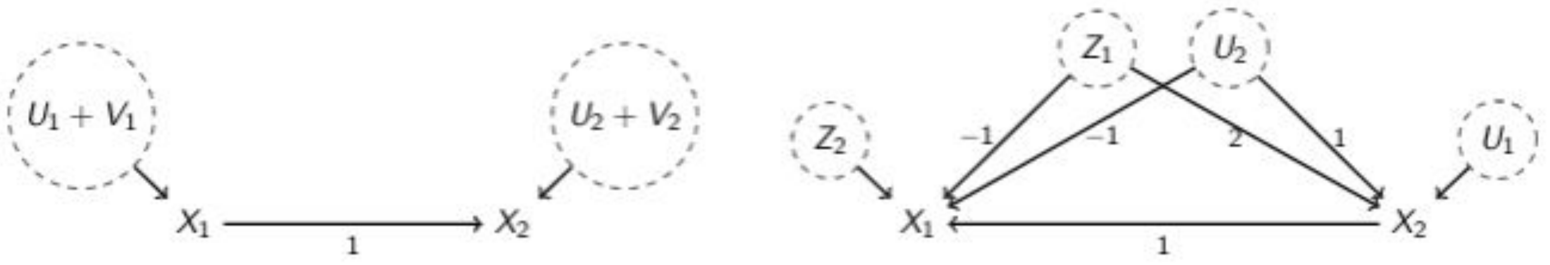
Flipping returns when we allow for unobserved confounders.

Although causal orientation is a solvable problem (assuming faithfulness), it is no longer decidable.

Genin, Konstantin (2021). “Statistical Undecidability in Linear Non-Gaussian Models in the Presence of Latent Confounders,” submitted.

Bad News

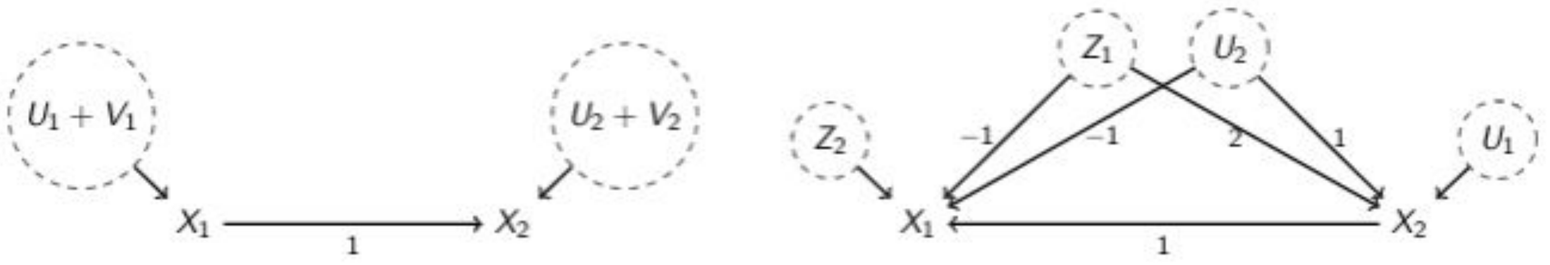
Let U_1, U_2 be independent, non-Gaussian. Let Z_1,Z_2 be independent Gaussians of equal variance. Then, V_1=Z_1+Z_2 and V_2=Z_1-Z_2 are independent and Gaussian. The following give rise to the same distribution over (X_1, X_2):



Genin, Konstantin (2021). “Statistical Undecidability in Linear Non-Gaussian Models in the Presence of Latent Confounders,” submitted.

Bad News

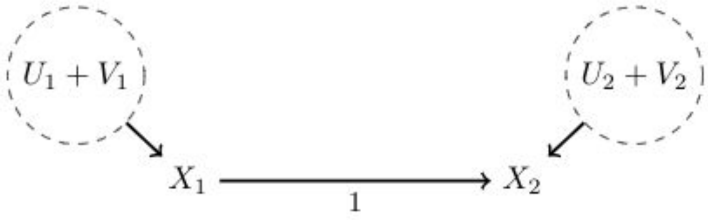
The lhs is a LiNGAM with no confounders, but the rhs model is a mess: it is unfaithful (because of U_2) and has a Gaussian noise term (Z_2) and Gaussian confounder (Z_1).

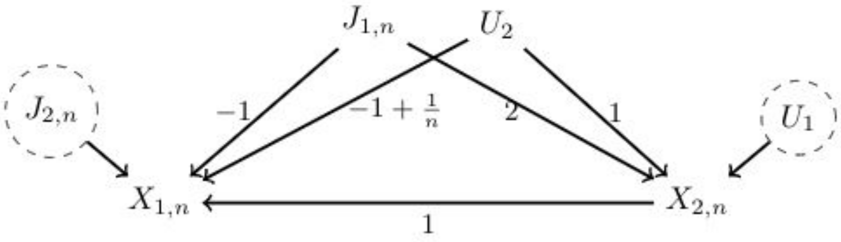


Genin, Konstantin (2021). “Statistical Undecidability in Linear Non-Gaussian Models in the Presence of Latent Confounders,” submitted.

Bad News

But we can approximate the rhs by a sequence of LiNGAMS. Let A_1,A_2 be independent, non-Gaussian. Let $J_{1,n}$ = Z_1+ (1/n)*A_2 and $J_{2,n}$ = Z_2 + (1/n)*A_2. Then ($X_{1,n}$, $X_{2,n}$) ⇒ (X_1,X_2).





Genin, Konstantin (2021). “Statistical Undecidability in Linear Non-Gaussian Models in the Presence of Latent Confounders,” submitted.

A Way Out?

Say that a r.v. X has **no Gaussian component** if it cannot be written as $X = Y + Z$, where Y,Z are independent and Z is Gaussian.

Conjecture: We can banish flipping in the potentially confounded case by slightly strengthening the LiNGAM assumptions to rule out noise terms with Gaussian components.