Simplicity, Induction and the Causal Truth

Kevin T. Kelly
Konstantin Genin

Carnegie Mellon University
kk3n@andrew.cmu.edu
kgenin@andrew.cmu.edu

7th June 2014

Model Selection and Simplicity

Ockham's Razor

Ockham: "choose the *simplest* model compatible with the data".

Figure : Third and Twelfth Degree Fitted Polynomials

Model Selection: Finding True Structure

Question

- How could a fixed simplicity bias help one find the *true* model?
- *Truth* means getting the *counterfactual* predictions right.
- E.g. causal direction.

The Frequentist Story

The Over-fitting Argument

- Preferring the simpler theory minimizes out-of-sample *prediction* error at small sample sizes.

The Frequentist Story

The Over-fitting Argument

- Preferring the simpler theory minimizes out-of-sample *prediction* error at small sample sizes.
- But accuracy in the *sample* population does not imply accuracy in the *manipulated* population.

The Frequentist Story

The Over-fitting Argument

- Preferring the simpler theory minimizes out-of-sample *prediction* error at small sample sizes.
- But accuracy in the *sample* population does not imply accuracy in the *manipulated* population.
- Anyway, what *is* the over-fitting argument?

The Frequentist Story

The Over-fitting Argument

Suppose the true model is:

$$y = f(x) + \varepsilon,$$

where $\mathrm{Var}(\varepsilon) = \sigma^2$ and σ is known.

The Frequentist Story

The Over-fitting Argument

The *true risk* of an estimator is the expected distance from the true predictor:

$$E[(\hat{f}(x) - f(x))^2].$$

How can we estimate the true risk of our estimator?

The Frequentist Story

The Over-fitting Argument

Suppose the sample is: $(x_1, y_1), (x_2, y_2), ...(x_n, y_n)$.

The in-sample error is given by:

$$\sum_{i=1}^{n} (\hat{f}(x_i) - y_i)^2.$$

But that is bound to *underestimate* the true risk.

The Frequentist Story

The Over-fitting Arugment

$$\begin{aligned}
\text{True Risk} &= \mathsf{E}\,[\text{ in-sample error }] + \text{complexity} + \text{noise}. \\
\mathrm{E}[(\hat{f}(x) - f(x))^2] &= \mathsf{E}\left[\sum_{i=1}^{n} (\hat{f}(x_i) - y_i)^2\right] + 2\sigma^2 \mathrm{df}(\hat{f}) \quad + n\sigma^2. \\
\mathrm{df}(\hat{f}) &= \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{Cov}(\hat{f}(x_i), y_i).
\end{aligned}$$

The Frequentist Story

The Over-fitting Argument

- True Risk - E [in-sample error] = complexity + noise.
- So:

  in-sample error + complexity + noise

  is an *unbiased estimate* of the true risk!

The Frequentist Story

The Over-fitting Argument

- So why not do the following?
  1. Let $\hat{f}_M$ be the maximum likelihood estimator for model M;
  2. *select* the model M^* that minimizes the unbiased estimate of the risk of using $\hat{f}_M$;
  3. then output the estimate $\hat{f}_{M^*}$.
- Call the estimator just defined by that whole procedure as $\hat{f}^*$.
- e.g. AIC, ERM, cross-validation, etc.

The Frequentist Story

The Over-fitting Argument

- But the Ockham estimator $\hat{f}^*$ is *not* the estimator $\hat{f}_M$ of the M so chosen.
- It chooses *different* $\hat{f}_M$'s on different samples!
- What does the risk of the $\hat{f}^*$ estimator actually look like?

The Frequentist Story

The Over-fitting Argument

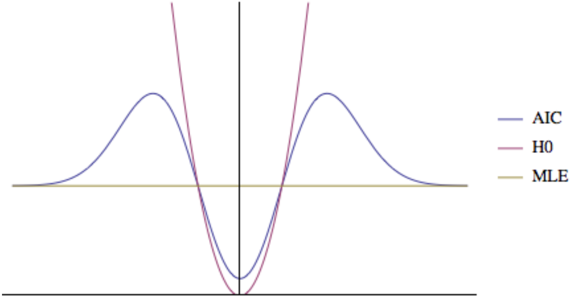
- Let $X_1, ...X_n \sim \mathcal{N}(\mu, \sigma^2)$ where σ is known. Suppose we are interested in estimating μ.
- Let:

$$\begin{aligned} \hat{\mu}_0 &= 0; \\ \hat{\mu}_{\mathrm{MLE}} &= \overline{X}. \end{aligned}$$

The Frequentist Story

The over-fitting argument as a decision

- In what sense do Ockham-like methods "minimize risk"?
- As a frequentist, it's *cheating* to appeal to area!



The Frequentist Story

Not about the truth

- Even if the over-fitting argument were an argument, it wouldn't pertain to estimates of policy outcomes.
- The estimate of risk is unbiased only in the *training* distribution.
- Banning ash trays doesn't prevent lung cancer.

The Bayesian Story

Beg the Question

- Simpler theories are more "probably true".
- But that just *is* a personal bias toward simplicity!
- Why *should* we have one?

The Bayesian Story

Or Subtly Beg the Question

- Suppose:
- $M_0 = \{\theta_0\}$;
- $M_1 = \{\theta_1, ..., \theta_n\}$;
- $\mathrm{P}(D \mid \theta_0) \approx 1$;
- $\mathrm{P}(D \mid \theta_1) \approx 1$;
- $\mathrm{P}(D \mid \theta_2) \approx 0$;

  $\vdots$
- $\mathrm{P}(D \mid \theta_n) \approx 0$.
- $\mathrm{P}(M_0) \approx \mathrm{P}(M_1)$.

The Bayesian Story

Bayes Theorem

$$\frac{\mathrm{P}(M_0 \mid D)}{\mathrm{P}(M_1 \mid D)} \approx \frac{\mathrm{P}(D \mid M_0)}{\mathrm{P}(D \mid M_1)}.$$

The Bayesian Story

Total Probability

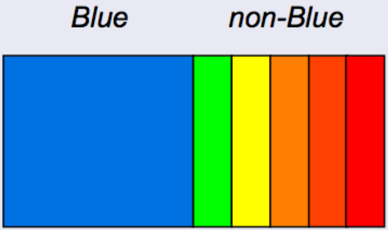
$$\frac{\mathrm{P}(M_0 \mid D)}{\mathrm{P}(M_1 \mid D)} \approx \frac{\sum_{i=0}^{0} \mathrm{P}(D \mid \theta_0)\mathrm{P}(\theta_0)}{\sum_{i=1}^{n} \mathrm{P}(D \mid \theta_i)\mathrm{P}(\theta_i)} \approx \frac{\mathrm{P}(\theta_0)}{\mathrm{P}(\theta_1)} = n.$$

- The outcome is just the *prior ratio* $\mathrm{P}(\theta_0)/\mathrm{P}(\theta_1)$.

The Bayesian Story

The *Bad* Paradox of Indifference

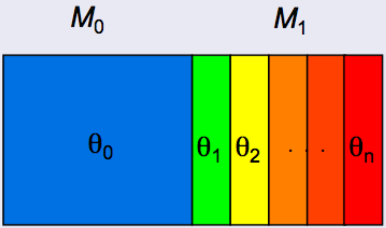
- *Ignorance* whether blue generates *"knowledge"* against green.



The Bayesian Story

The *Good* Bayes Factor Argument for Simplicity

- *Ignorance* whether M_0 generates *"knowledge"* against θ_1.



The Bayesian Story

Circularity

- The Bayesian arguments for Ockham's razor *pass along* a *prior bias* toward simplicity.
- The prior bias is not *reliable* unless one *assumes* that the simple model is true.

The Simulation Story

Doggone it, Ockham *Smells* the Truth on Simulated Samples!

- Did you generate the data from a model with parameters set by an uninformative prior density?
- If so, you are just doing a Monte-Carlo simulation of the Bayesian simplicity bias just described.

The CMU Transcendental Deduction

Just assume whatever's nececessary

- At least it's honest.
- Also, it is hopeless to expect a method to work when an illusion is *perfect* forever.
- But strong faithfulness is another matter (Uhler 2014)!

Pursuit of Truth

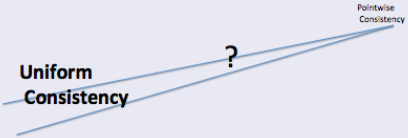
The Status Quo

- All linear Gaussian search strategies rely heavily on Ockham's razor.
- Wouldn't it be nice to have a *non-circular* argument that Ockham's razor is the *best* strategy for finding the *true* model?

Pursuit of Truth

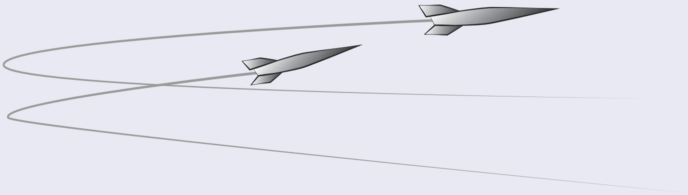
Reliability as Most Direct Approach to the Truth

- In some cases, it is *impossible* to find the causal truth reliably in the *short* run.
- *Other* methods besides Ockham's razor find the truth in the *long* run.
- Perhaps Ockham's razor is best in some *intermediate* sense.



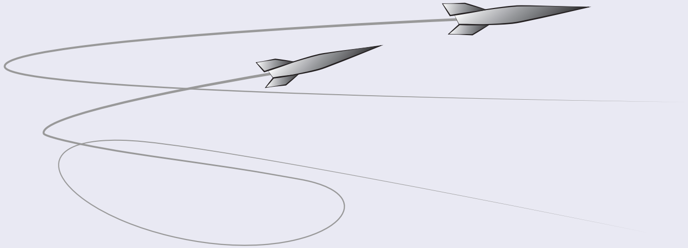
Pursuit of Truth

Optimally Direct Pursuit



Pursuit of Truth

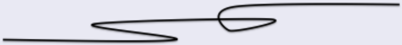
Needlessly Indirect Pursuit



Optimal Pursuit of Truth

Inductive Justification

- Deductive inference is *direct*.
- Inductive inference is *indirect*.
- But it should still be as *direct as possible*!
- Measures of indirectness are *course reversals* and *loop length*.



Optimal Pursuit of Truth

Reversals

- Saying A and then saying B inconsistent with A.

Loops

- Saying A, then saying B inconsistent with A, and then saying C that entails A.

Optimal Pursuit of Truth

Reversals in Chance

- The chance of saying A goes down and the chance of an answer inconsistent with A goes up.

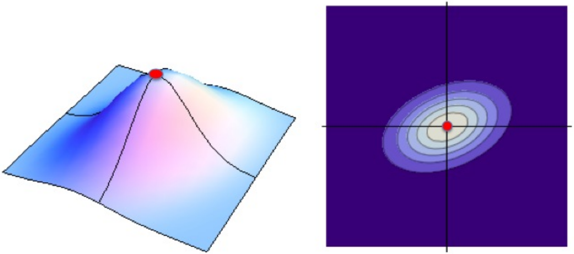
Loops in Chance

- The chance of saying A goes down, the chance of saying B inconsistent with A goes up, and then the chance of saying C that entails A goes up.

Statistical Reversals

Bivariate Normal Mean Problem

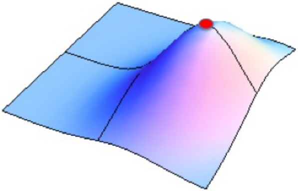
How many mean components are non-zero?

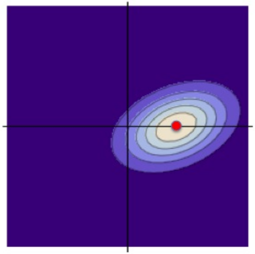


Statistical Reversals

Bivariate Normal Mean Problem

How many mean components are non-zero?

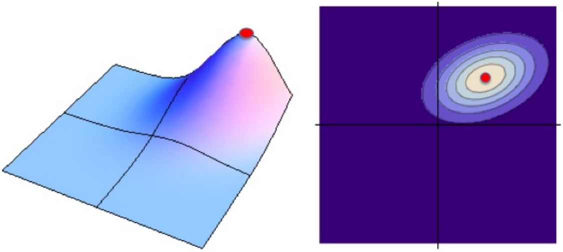




Statistical Reversals

Bivariate Normal Mean Problem

How many mean components are non-zero?



BIC, no correlation.

(Click to play movie)

BIC, with correlation.

(Click to play movie)

Bayes, no correlation.

(Click to play movie)

Bayes, with correlation.

(Click to play movie)

Agnostic Bayes, no correlation.

(Click to play movie)

Agnostic Bayes, with correlation.

(Click to play movie)

Improved BIC, no correlation.

(Click to play movie)

Improved BIC, with correlation.

(Click to play movie)

To see the simulations:

www.andrew.cmu.edu/user/kk3n/ockham/probsims/statsims.html

Simplicity

Simplicity as Topology

- Let A, B be sets of sampling distributions.
- Topologize faithful distributions by total variation metric.
- Define the pre-order:

$$A \preceq B \quad \Leftrightarrow \quad A \subseteq \mathrm{bdry}(B).$$

- The statistical problem of induction.
- No possible statistical technique could reliably rule out B if A is true.

Ockham's Razor

Simplicity as Topology

- The $\preceq$ order is only a pre-order, so simplicity cycles are possible.
- But indistinguishability classes (i.c.'s) of linear Gaussian distributions are *locally closed*.
- Then $\preceq$ is a partial order.

Ockham's Razor

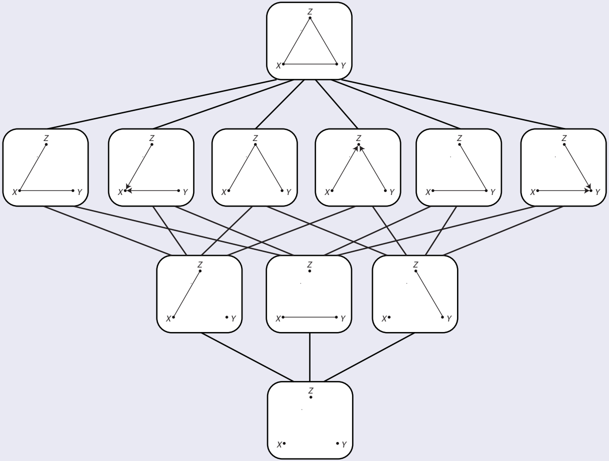
Simplicity as Topology

- Let A, B be faithful sets of conditional *dependencies*.
- Let A^*, B^* be the corresponding sets of faithful, linear Gaussian and discrete Bayes distributions.
- Then:

$$A \subseteq B \quad \Leftrightarrow \quad A^* \preceq B^*.$$

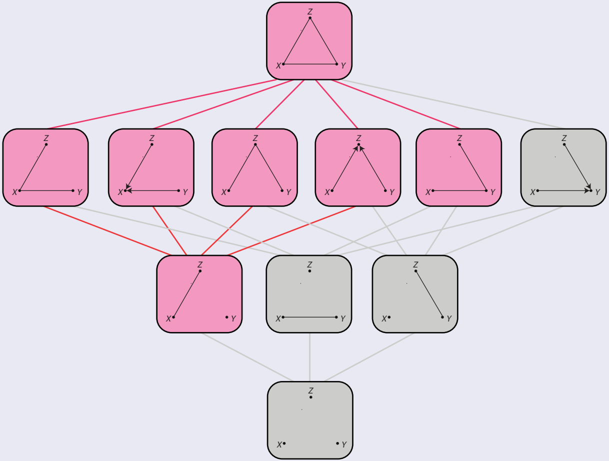
Causal Simplicity

Linear Gaussian Simplicity, Three Variables



Ockham's Razor

Ambiguous Data: $X \not\perp\!\!\!\perp Y$



Ockham's Razor

Skips

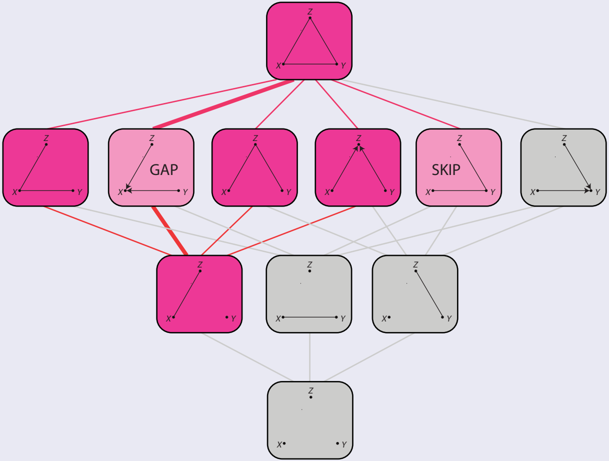
- Output rules out some *minimal* i.c. compatible with current information.
- Results in *extra reversals*.

Gaps

- Output is *not closed downward* among i.c.'s compatible with current information.
- Results in *extra loops*.

Ockham's Razor

Ambiguous Data: $X \not\perp\!\!\!\perp Y$



Ockham's Razor

Horizontal

- Avoid *skips*!
- Minimizes worst-case *reversals* in each i.c..

Vertical

- Avoid *gaps*!
- Minimizes worst-case *loops* in each i.c..

Ockham's Razor in Chance

Skips in Chance

- Some chance in P of producing an answer false in P that is not true in some simpler P'.
- Results in extra *reversals in chance*.

Gaps in Chance

- Some chance in P of producing an answer false in some more complex P and true in some even more complex P.
- Results in extra *loops in chance*.

Ockham's Razor in Chance

Horizontal

- Avoid *skips in chance*!
- Forces simple acceptance zones to take precedence over complex acceptance zones.
- Neyman-Pearson acceptance zone is a trivial case.
- Also forces the method to return disjunctions when sampling distributions of equally simple worlds with different answers overlap.
- **Conjecture:** Minimizes worst-case *reversals in chance* in each i.c..

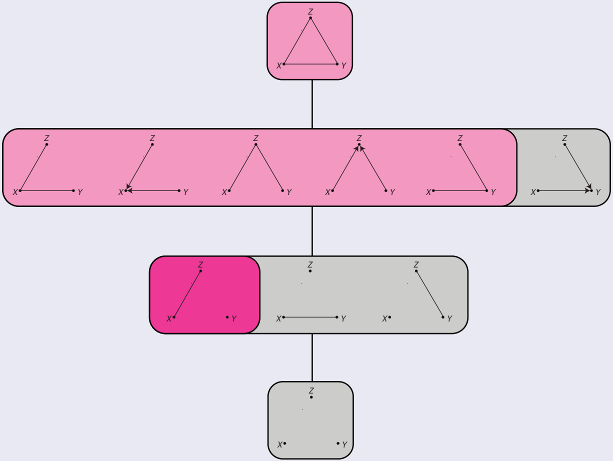
Ockham's Razor in Chance

Vertical

- Avoid *gaps in chance*!
- Forces simple acceptance zones to *overlap* complex acceptance zones.
- Allows for greedy favoritism over equally simple models.
- **Conjecture:** Minimizes worst-case *loops in chance* in each i.c..

Simplicity as Edge Count

Ambiguous Data: $X \not\perp Y$



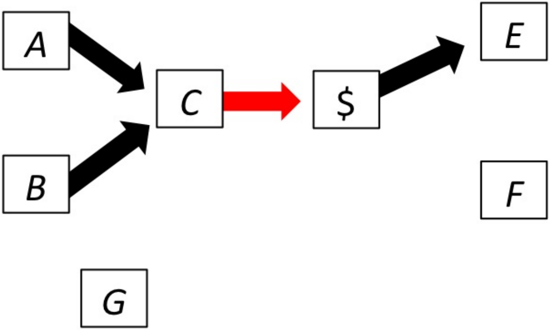
Simplicity as Edge Count

Epistemic Trade-off

- Simplicity *ranking* allows for *stronger* conclusions and *less computation*.
- Simplicity ranking *excuses* more reversals under its coarser worst-case bounds.
- No difference for loops.

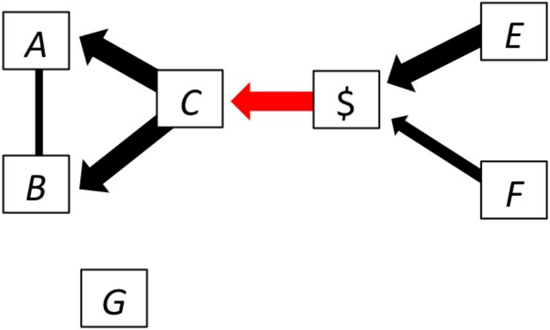
Example

Buy Gimme Pharmaceuticals.: $N = 2000$



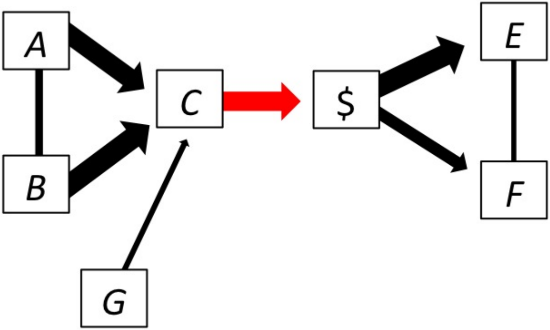
Example

Gimme Pharmaceuticals faces Chapter 11: $N = 50,000$



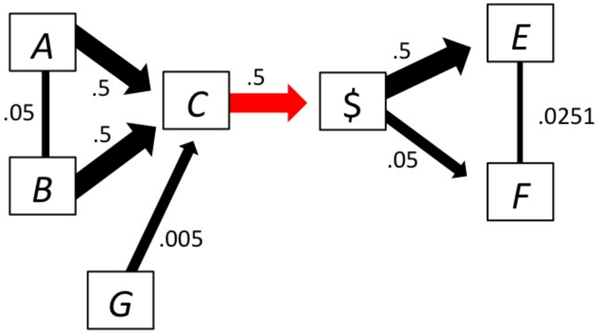
Example

Gimme Pharmaceuticals shares soar! $N = 1,000,000$



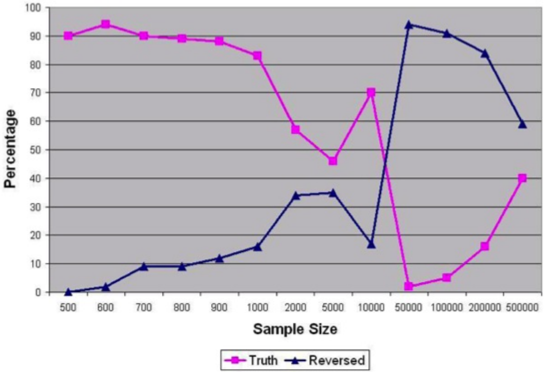
Example

The Underlying Truth: Impossible?



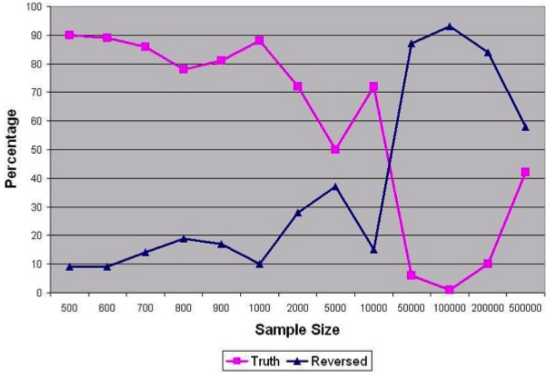
Simulation Studies

The PC Algorithm (c. 2012)



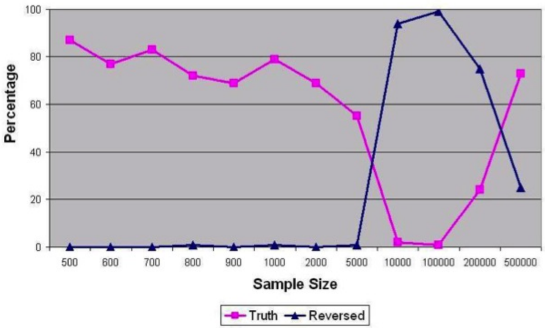
Simulation Studies

The FCI Algorithm (c. 2012)



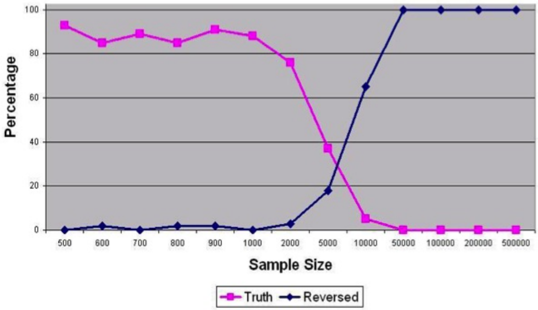
Simulation Studies

The CPC Algorithm (c. 2012)



Simulation Studies

The GES Algorithm (c. 2012)



Causal Discovery Nouveau

Non-Gaussian and Non-Linear

- Ironically, the standard case is the hardest case.
- *Assuming* that the model is non-Gaussian or non-linear, the problem of induction disappears, under reasonable assumptions.
- But if it doesn't disappear, linear Gaussian is in the *boundary* of the other two possibilities.
- So Ockham's Razor says to *favor the linear Gaussian case* until it is refuted!

Ockham and Expanded Faithfulness

Ockham Favors Linear Gaussian

- *Assuming* that the model is non-Gaussian or non-linear, the problem of induction disappears.
- But if it doesn't disappear, linear Gaussian models are in the *boundary* of the other two possibilities, so it is favored by Ockham.

Faithfulness

Is it Ockham's Razor?

- Typically, faithfulness *rules out* a boundary set of possibilities.
- Ockham's razor *favors* a boundary set of possibilities.
- Faithfulness is tied to the semantics of "cause" and takes precedence over Ockham's razor, which is a defeasible inferential principle.
- Given that the causal mechanisms are causally sufficient, a "dependence" among mechanisms requires a causal meta-connection by a natural extension of the causal Markov condition.