A New Method for Estimating Causal Model Learning Accuracy

Erich Kummerfeld, PhD[1], Gregory F. Cooper, MD, PhD[2]

[1]University of Minnesota; [2]University of Pittsburgh

Nov 4, 2017

The Problem

Scenario:

The Problem

Scenario:
You learned causal model *M* from real world data *D* generated from unknown true model *T*.

The Problem

Scenario:
You learned causal model M from real world data D generated from unknown true model T.

Question:

The Problem

Scenario:
You learned causal model M from real world data D generated from unknown true model T.

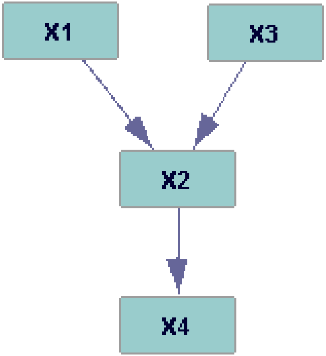
Question:
How close is M to T?

Solution Strategies

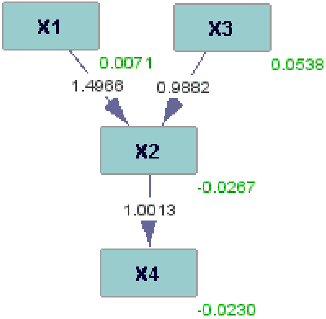
- **Statistical measures of fit**: individual score is not informative; best fit could still be inaccurate
- **Benchmark simulations**: incomplete; may not apply to this type of data; may not even be able to know if they apply
- **Resimulation**: benchmark against data that is similar to D

Resimulation 0: Data $D1$ and Learned Graph $G1$

X1	X2	X3	X4
0.3596	-1.2491	-2.9277	-3.5328
1.2639	4.0011	1.2282	4.1915
0.8749	-1.7419	-1.6859	-2.2926
-2.1222	-0.3536	2.465	-0.2342
-0.9151	-2.7165	-2.3928	-4.5982
-0.5706	-3.802	0.0331	-4.7854
1.2468	0.5542	-0.7107	1.0888
1.1232	5.1059	0.7407	6.571
-1.4056	1.5811	-0.527	0.5514
-0.2384	0.7289	-0.133	0.8222
0.0751	-2.5419	-1.708	-1.8866
0.8523	2.0218	0.6163	2.7466



Resimulation 1: Fit $G1$ to $D1$, making model $M1$

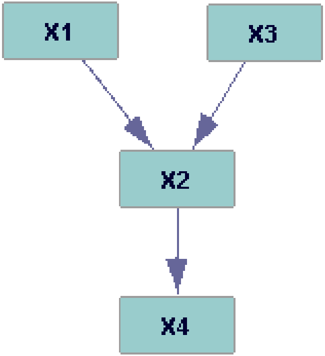


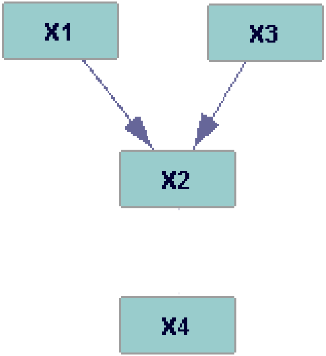
Resimulation 2: Sample *D2* from *M1*

X1	X2	X3	X4
0.8708	-4.2755	-3.8422	-1.4316
-0.4746	0.2453	2.3116	3.6107
-0.8326	1.6374	1.3244	-1.3061
0.8904	1.3817	0.5551	1.6176
-1.5868	-0.2379	-0.8964	-0.4021
0.9449	-0.4699	-1.6115	-0.4532
-1.4363	-1.9608	-0.0541	-2.2113
-0.1365	-1.5573	0.0807	0.2054
2.7841	6.6639	2.0372	7.3468
0.2111	-0.9978	-0.3473	1.5266
-0.6065	3.208	1.9332	4.4616
0.9039	-0.7902	-0.1446	-0.7825
[illegible]	[illegible]	[illegible]	[illegible]

Each row sampled from $P_{M1}(X1, X2, X3, X4)$

Resimulation 3: Learn *G2* from *D2*, compare to *G1*

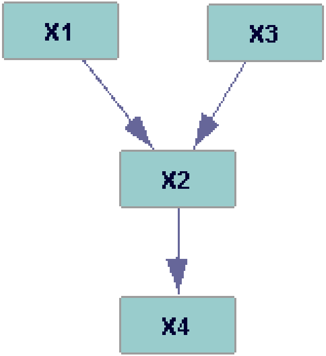




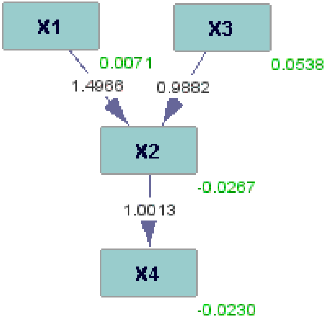
G2 (right) contains 2 of the 3 edges in G1 (left), and no additional edges.

Hsim 0: Data $D1$ and Learned Graph $G1$

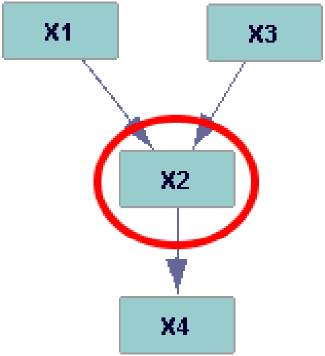
X1	X2	X3	X4
0.3596	-1.2491	-2.9277	-3.5328
1.2639	4.0011	1.2282	4.1915
0.8749	-1.7419	-1.6859	-2.2926
-2.1222	-0.3536	2.465	-0.2342
-0.9151	-2.7165	-2.3928	-4.5982
-0.5706	-3.802	0.0331	-4.7854
1.2468	0.5542	-0.7107	1.0888
1.1232	5.1059	0.7407	6.571
-1.4056	1.5811	-0.527	0.5514
-0.2384	0.7289	-0.133	0.8222
0.0751	-2.5419	-1.708	-1.8866
0.8523	2.0218	0.6163	2.7466



Hsim 1: Fit $G1$ to $D1$, making model $M1$



Hsim 2: Pick variables to resimulate



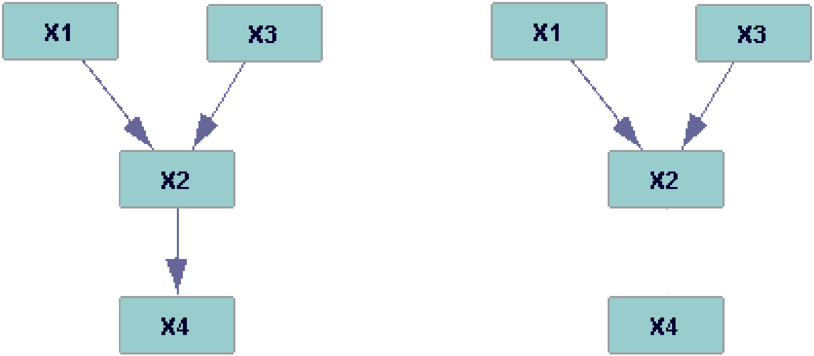
Variables can be selected or chosen at random.

Hsim 3: Sample *D2* from *M1*

X1	X2	X3	X4
0.3596	-0.3461	-2.9277	-3.5328
1.2639	-4.2367	1.2282	4.1915
0.8749	-2.4683	-1.6859	-2.2926
-2.1222	3.1447	2.465	-0.2342
-0.9151	-1.2284	-2.3928	-4.5982
-0.5706	-2.2541	0.0331	-4.7854
1.2468	-5.6872	-0.7107	1.0888
1.1232	-0.2238	0.7407	6.571
-1.4056	-4.6249	-0.527	0.5514
-0.2384	5.6388	-0.133	0.8222
0.0751	-1.8405	-1.708	-1.8866
0.8523	5.0506	0.6163	2.7466
0.2449	1.0469	0.6777	0.0993

Each row sampled from $P_{M1}(X2 | X1 = x1, X3 = x3, X4 = x4)$

Hsim 4: Learn *G2* from *D2*, compare to *G1*



G2 (right) contains all edges oriented towards X2 in G1 (left). G2 contains no additional edges connected to X2.

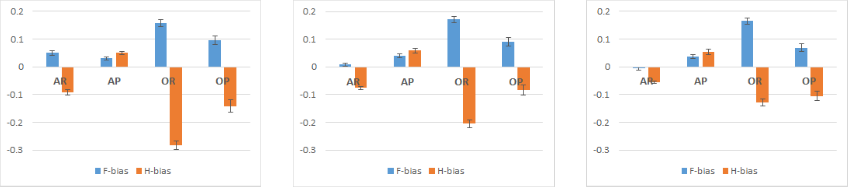
Simulation Parameters

- Simulated 500 “true” graphs and sampled data.
- Run FGES and calculate actual accuracy measures.
- Estimate accuracy with full and hybrid resimulation.

Model parameters:

- Gaussian noise
- Functional relationships:
  - Linear
  - Nonlinear

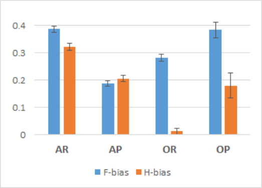
Linear



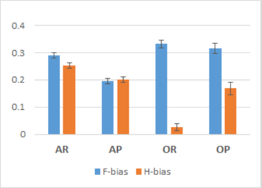
(100) (300) (900)

Figure: Simulation study results for linear models, showing mean estimation errors for AR, AP, OR, and OP at sample sizes 100, 300, and 900. Error bars represent 95% confidence intervals of the mean estimates shown.

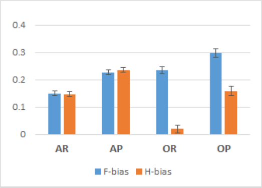
Nonlinear



(100)



(300)



(900)

Figure: Simulation study results for nonlinear models, showing estimation errors for AR, AP, OR, and OP at sample sizes 100, 300, and 900. Error bars represent 95% confidence intervals of the mean estimates shown.

Acknowledgements

The authors would like to thank Bryan Andrews for assistance with generating the nonlinear data, and the UCI Machine Learning Repository for hosting the Breast Cancer Wisconsin (Diagnostic) Data Set. Research reported in this publication was supported by grant U54HG008540 awarded by the National Human Genome Research Institute through funds provided by the trans-NIH Big Data to Knowledge (BD2K) initiative. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. This project was also funded, in part, by grant #4100070287 from the Pennsylvania Department of Health. The Department specifically disclaims responsibility for any analyses, interpretations, or conclusions.