Evaluating Causal Models by Comparing Interventional Distributions

Dan Garant and David Jensen
Knowledge Discovery Laboratory
College of Information and Computer Sciences
University of Massachusetts Amherst

Findings

- Existing approaches to evaluation are strictly structural, and do not characterize the full causal inference pipeline
- Statistical distances can be used to evaluate interventional distribution quality
- Evaluation with statistical distance can lead to different conclusions about algorithmic performance

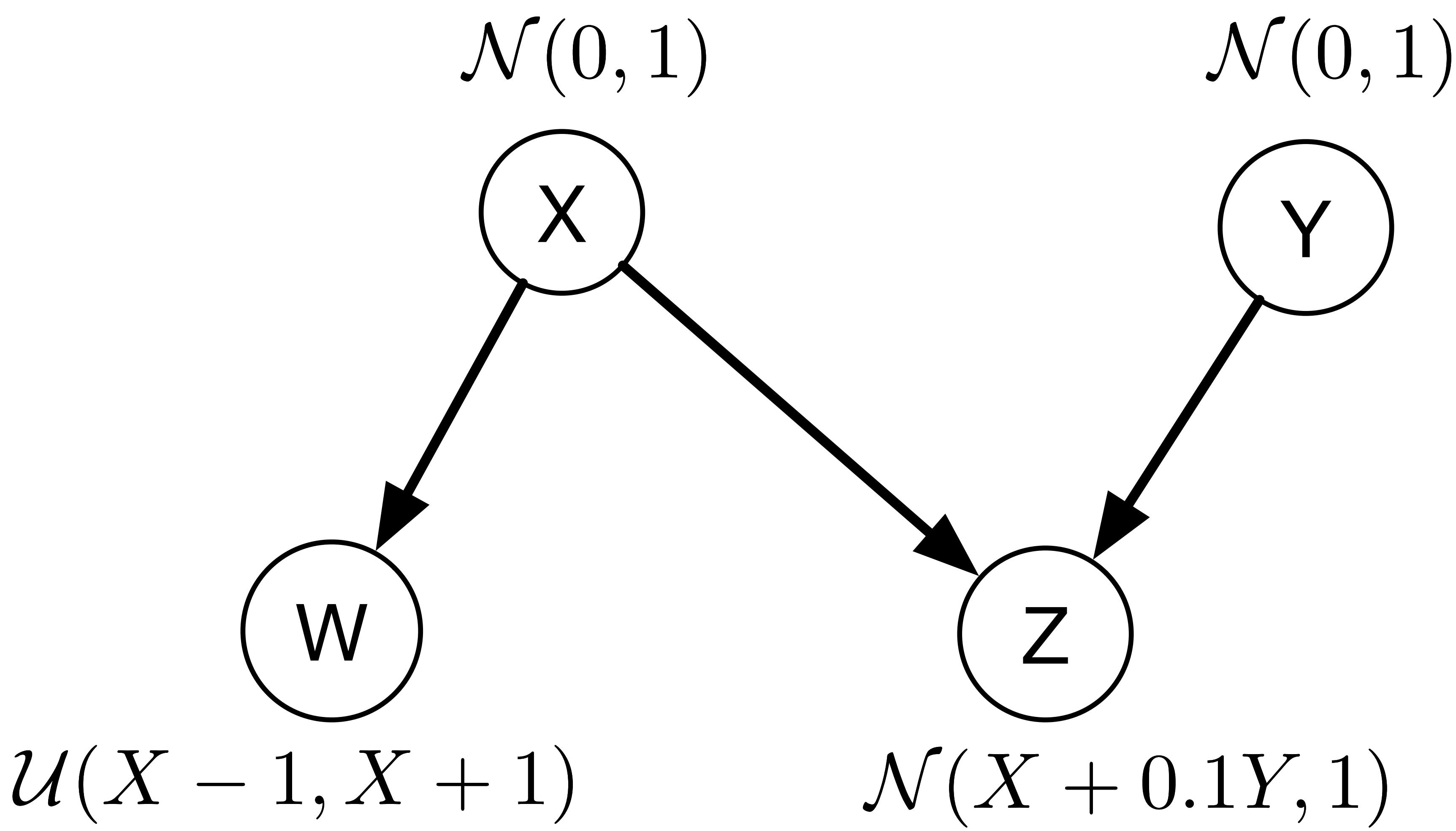
Overview

- Causal Graphical Models
- Current Approaches to Evaluation
- Evaluation with Statistical Distance
- Comparative Results

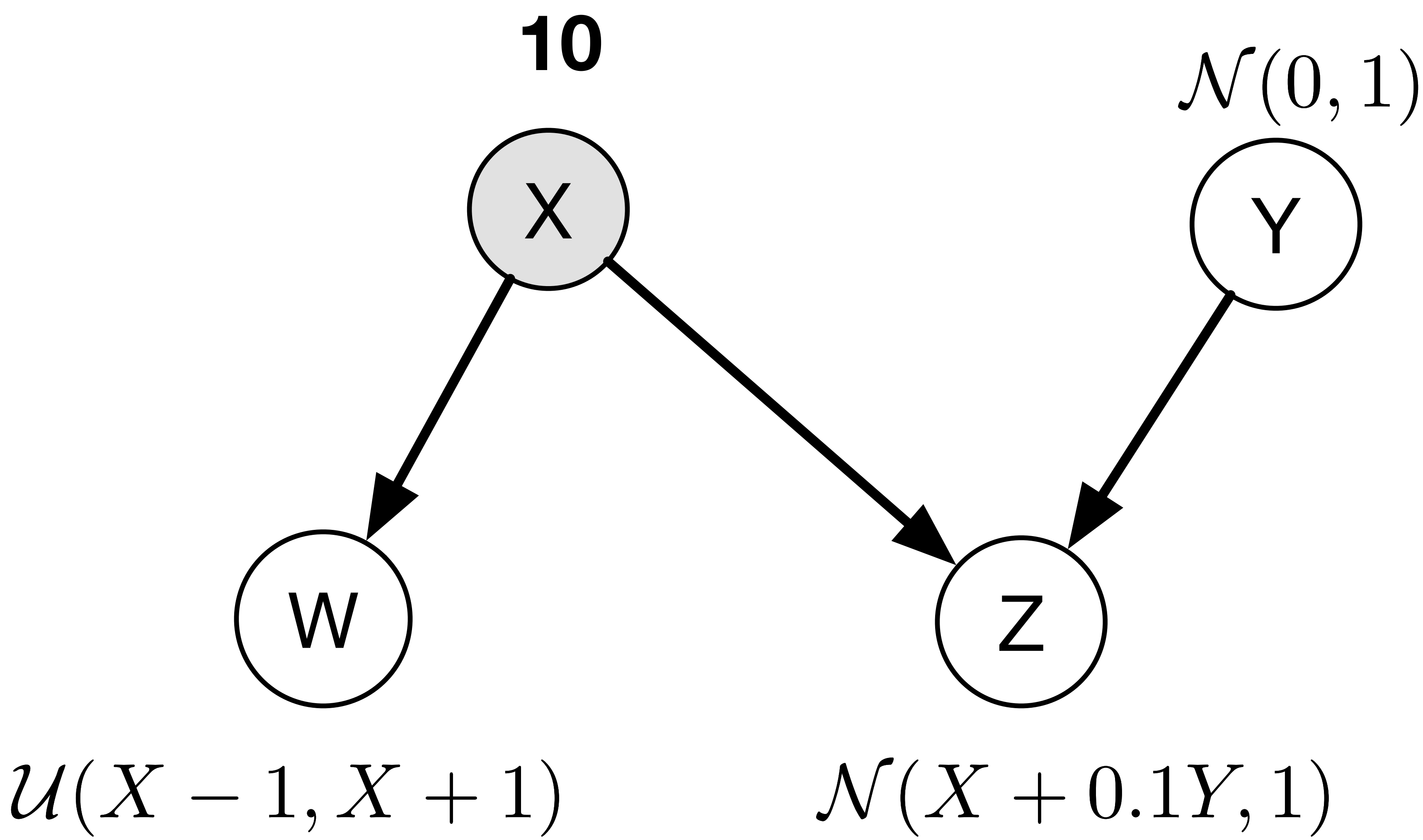
Overview

- **Causal Graphical Models**
- Current Approaches to Evaluation
- Evaluation with Statistical Distance
- Comparative Results

Causal Graphical Models



Causal Graphical Models



Use Cases

- Qualitative assessment of causal structure (does intervening on X influence Z?)

- Estimation of interventional distributions

$$P(Z|\mathrm{do}(X = 10))$$

Use Cases

- Qualitative assessment of causal structure (does intervening on X influence Z?)

- **Estimation of interventional distributions**

$$P(Z|\mathrm{do}(X = 10))$$

Structure Learning

- PC (Spirtes et al. 2000): Use conditional independence tests to derive constraints on possible structure
- GES (Chickering 2002): Perform local updates in order to maximize a global score on structures, maximizing structure likelihood
- MMHC (Tsamardinos et al. 2006): Combines constraint-based and score-based approaches

Chickering, D. M. (2002). Optimal structure identification with greedy search. Journal of machine learning research, 3(Nov), 507-554.

Spirtes, P., Glymour, C. N., & Scheines, R. (2000). Causation, prediction, and search. MIT press.

Tsamardinos, I., Brown, L. E., & Aliferis, C. F. (2006). The max-min hill-climbing Bayesian network structure learning algorithm. Machine learning, 65(1), 31-78.

Need for Quantitative Evaluation

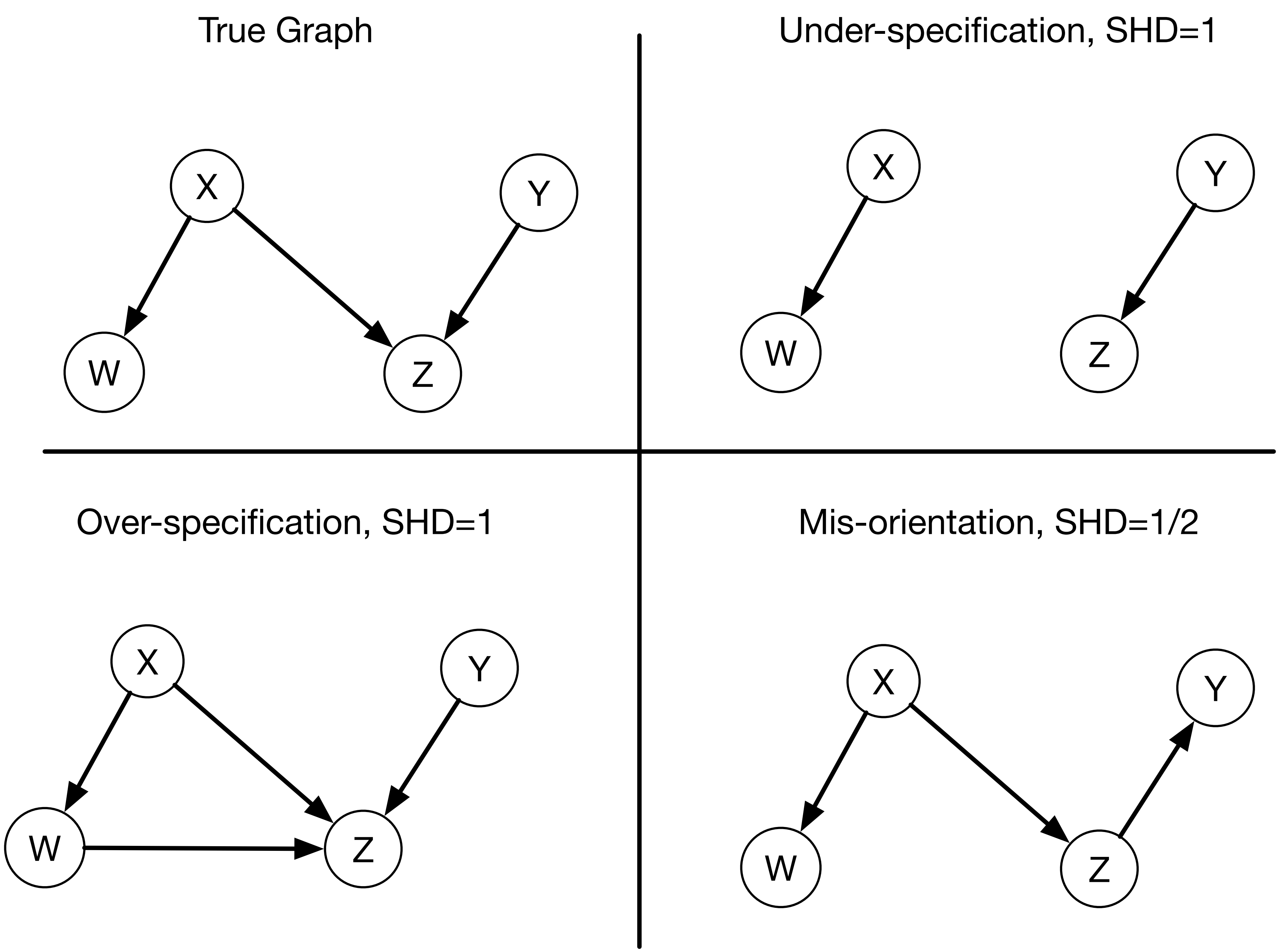
- How well do these algorithms work in practice? Under what circumstances do they perform better or worse?

- Which algorithm should I use? Does performance depend on domain characteristics?

Overview

- Causal Graphical Models
- **Current Approaches to Evaluation**
- Evaluation with Statistical Distance
- Comparative Results

Structural Hamming Distance (SHD)



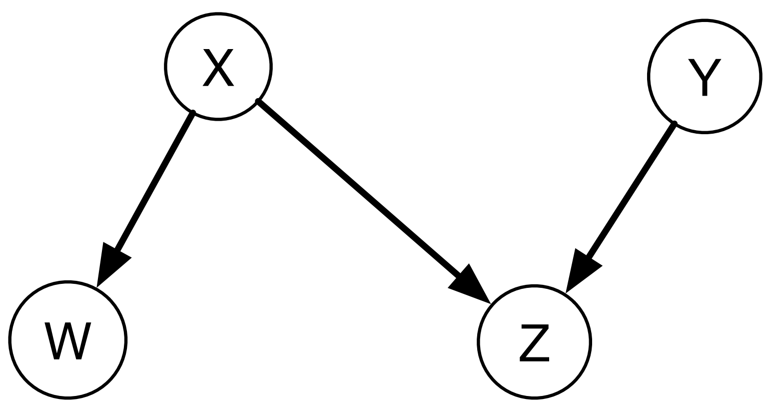
Structural Intervention Distance (SID)

- Graph mis-specification is not fundamentally related to quality of a causal model (Peters & Bühlmann 2015)
  - Including superfluous edges does not necessarily bias a causal model
  - Reversing or omitting edges can potentially induce bias in *many* interventional distributions
- Structural intervention distance: Count number of mis-specified pairwise interventional distributions

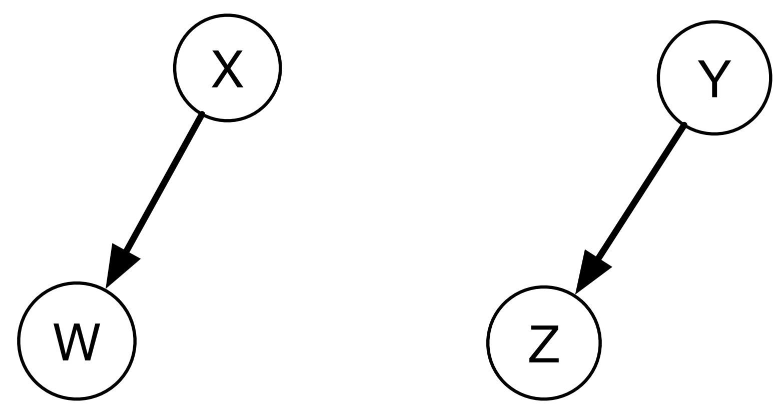
Peters, J., & Bühlmann, P. (2015). Structural intervention distance for evaluating causal graphs. Neural computation.

SHD vs SID

True Graph

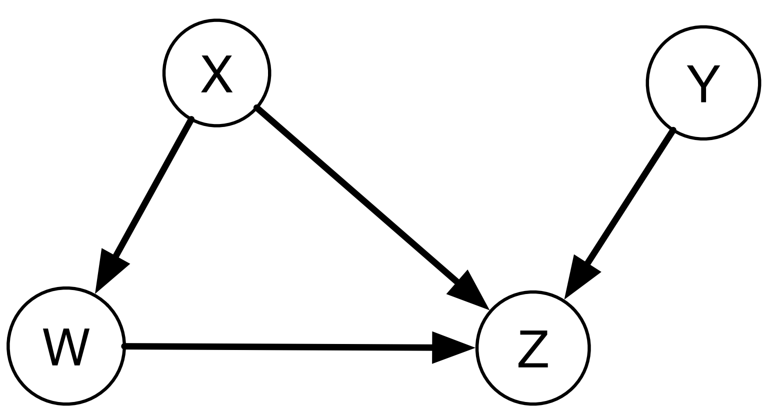


Under-specification, SHD=1, SID=1

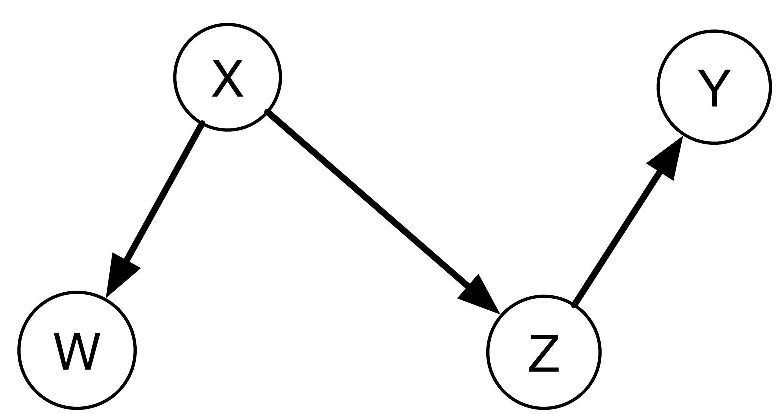


$P(Z|do(X))$

Over-specification, SHD=1, SID=0



Mis-orientation, SID=1/2, SID=3



$P(Y|do(X))$

$P(Z|do(Y))$

$P(Y|do(Z))$

Problems with Structural Distances

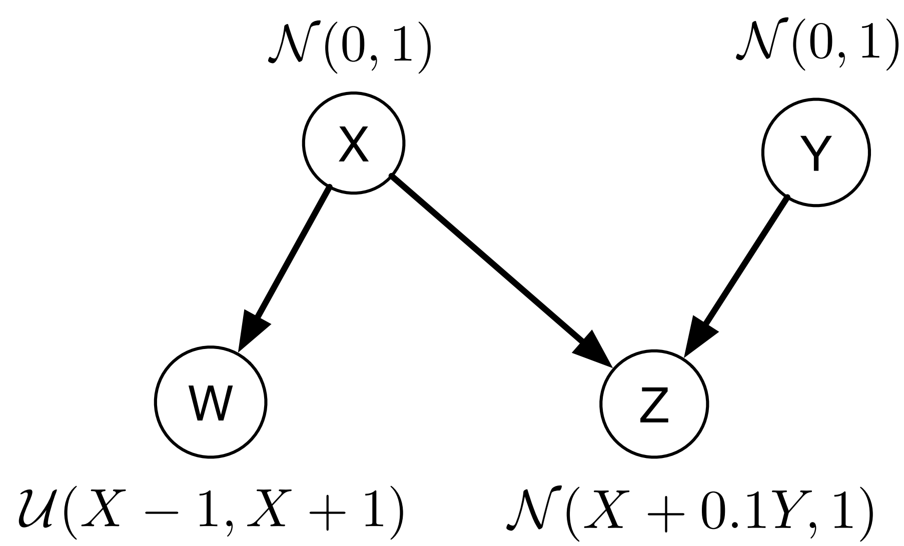
- Structural measures fail to characterize the full causal inference pipeline. To reach an interventional distribution, we also need to learn parameters and perform inference

- Some interventional distributions may be more biased than others

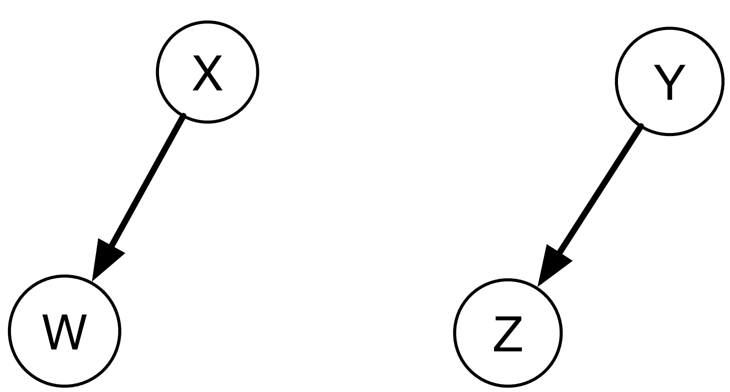
- In finite sample settings, variance matters too. A biased model with low variance may be better than an unbiased model with high variance

Statistical Effects of Model Errors

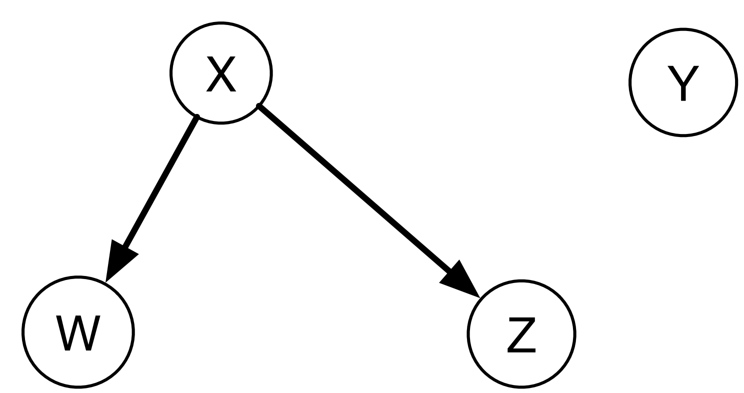
True Graph



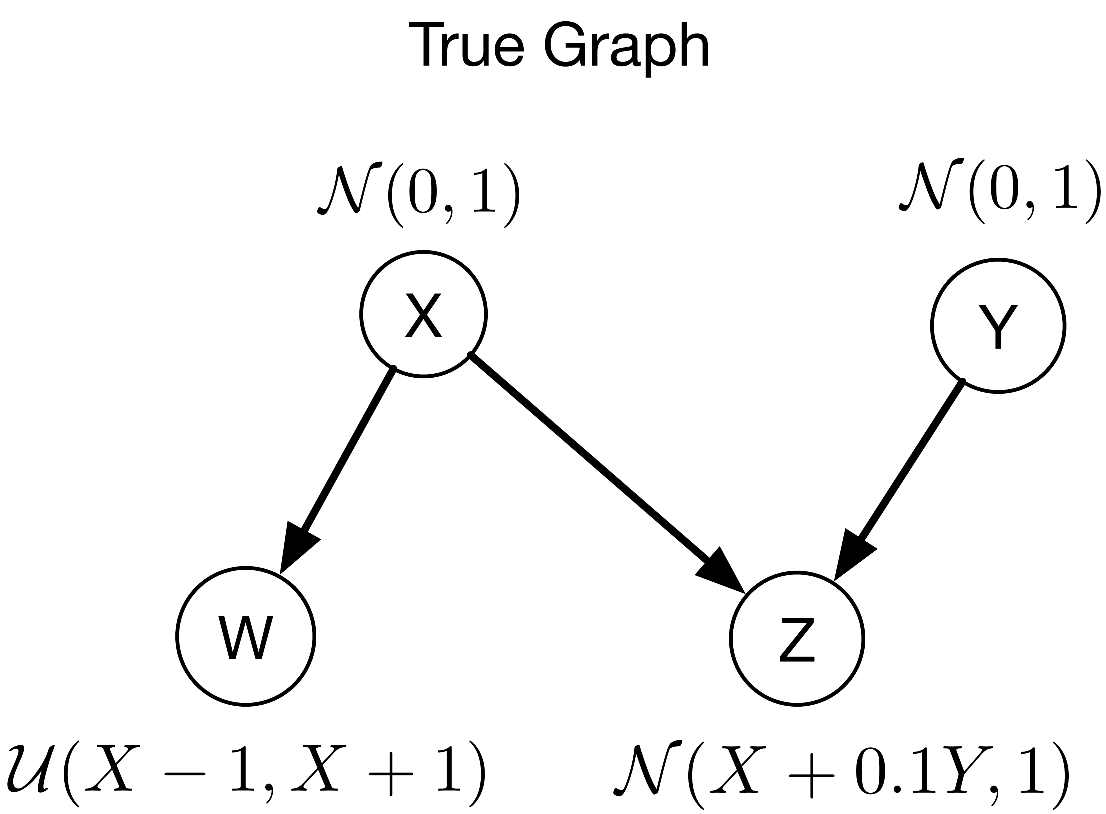
Under-specification, SHD=1, SID=2

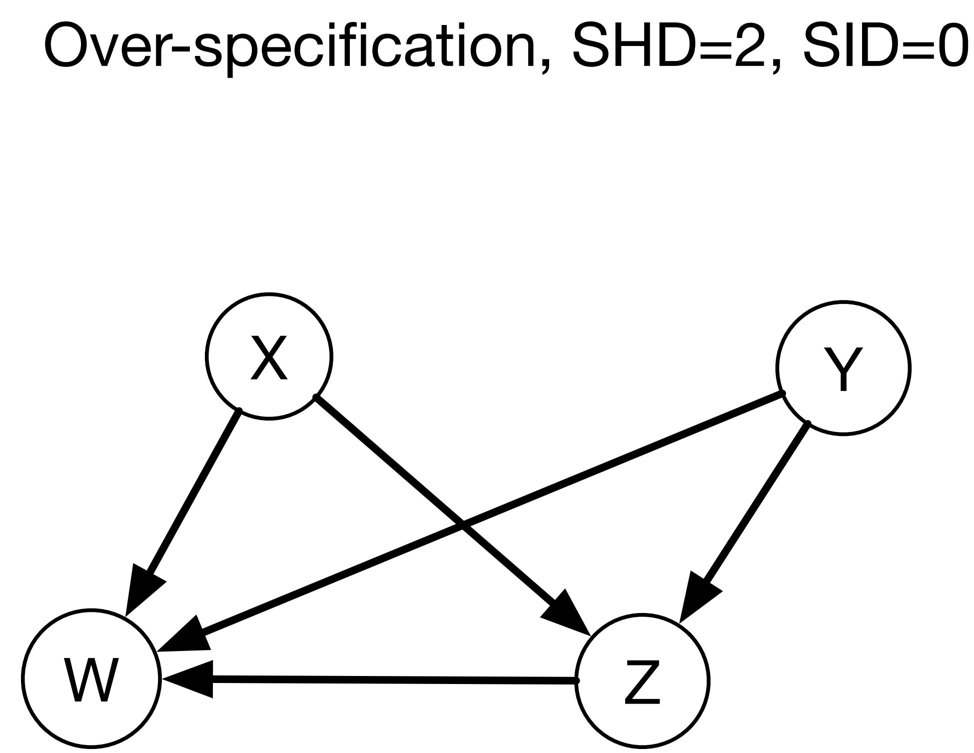


Under-specification, SHD=1, SID=2



Statistical Effects of Model Errors





Overview

- Causal Graphical Models
- Current Approaches to Evaluation
- **Evaluation with Statistical Distance**
- Comparative Results

Interventional Distribution Quality

- Ultimately, we care about the quality of interventional distributions rather than only the quality of the graph structure
- To evaluate distributions, we need:
  - Parameterized models
  - Inference algorithms
  - A measure of distributional accuracy

Total Variation Distance

$$TV_{P,\hat{P},T=t}(O) = \frac{1}{2} \sum_{o \in \Omega(O)} \left| P\left(O = o \middle| do(T = t)\right) - \hat{P}\left(O = o \middle| do(T = t)\right) \right|$$

Enumerating Distributions

- To evaluate an entire DAG, we need to enumerate pairs of treatments and outcomes

$$TV_{DAG}(G, \hat{G}) = \sum_{V \in \mathbf{V}(G), V' \in \mathbf{V}(G) \setminus \{V\}} TV_{P_G, P_{\hat{G}}, v'=v'_*}(V)$$

- Performing these inferences is expensive, but these are precisely the inferences that must be performed to use the model

Overview

- Causal Graphical Models
- Current Approaches to Evaluation
- Evaluation with Statistical Distance
- **Comparative Experiments**

Synthetic Domains

- **Logistic:** Binary data, each node is a logistic function of its parents

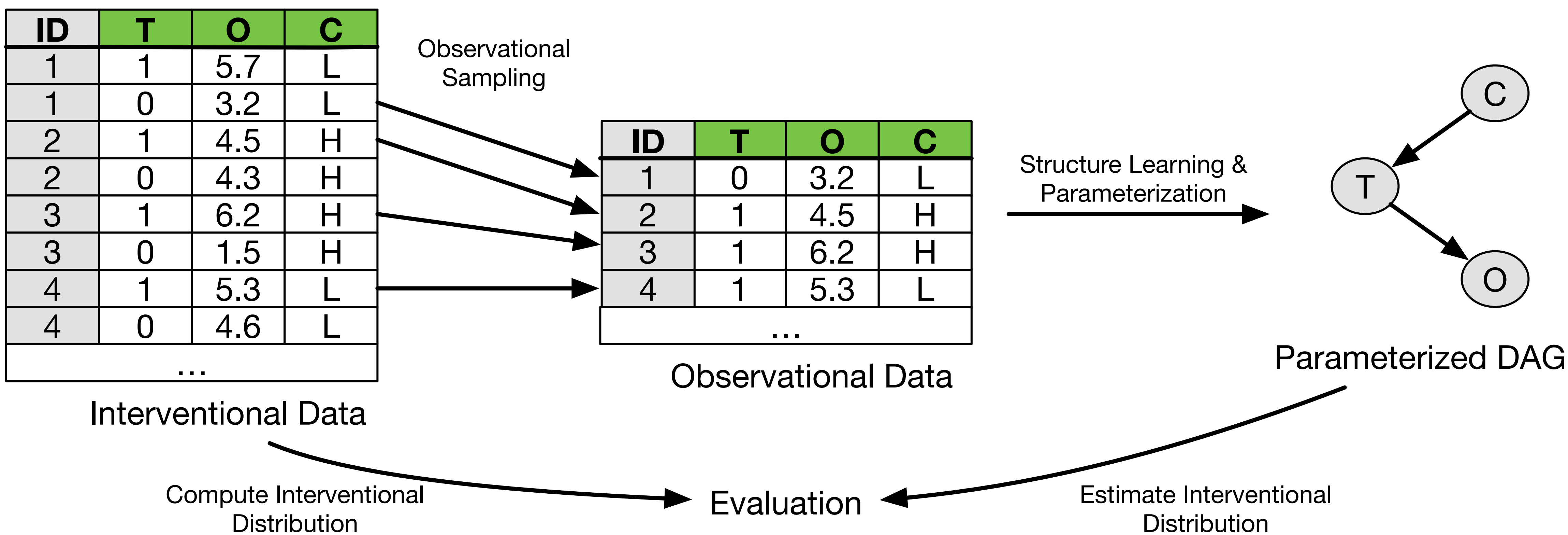
- **Linear-Gaussian:** Real-valued data, values for each node are normally distributed around a linear combination of parent values

- **Dirichlet:** Discrete data, CPD for each node is sampled from a Dirichlet distribution determined by parent values

Software Domains

- We instrumented and performed factorial experiments on three software domains:
    - Postgres
    - Java Development Kit
    - Web platforms
- Then, a biased sampling biased sampling routine is used to transform experimental data into observational data
- Ground-truth interventional distributions are computed on experimental data and compared to the distributions estimated from a learned model structure

Software Domains



ID	T	O	C
1	1	5.7	L
1	0	3.2	L
2	1	4.5	H
2	0	4.3	H
3	1	6.2	H
3	0	1.5	H
4	1	5.3	L
4	0	4.6	L
…			

ID	T	O	C
1	0	3.2	L
2	1	4.5	H
3	1	6.2	H
4	1	5.3	L
…			

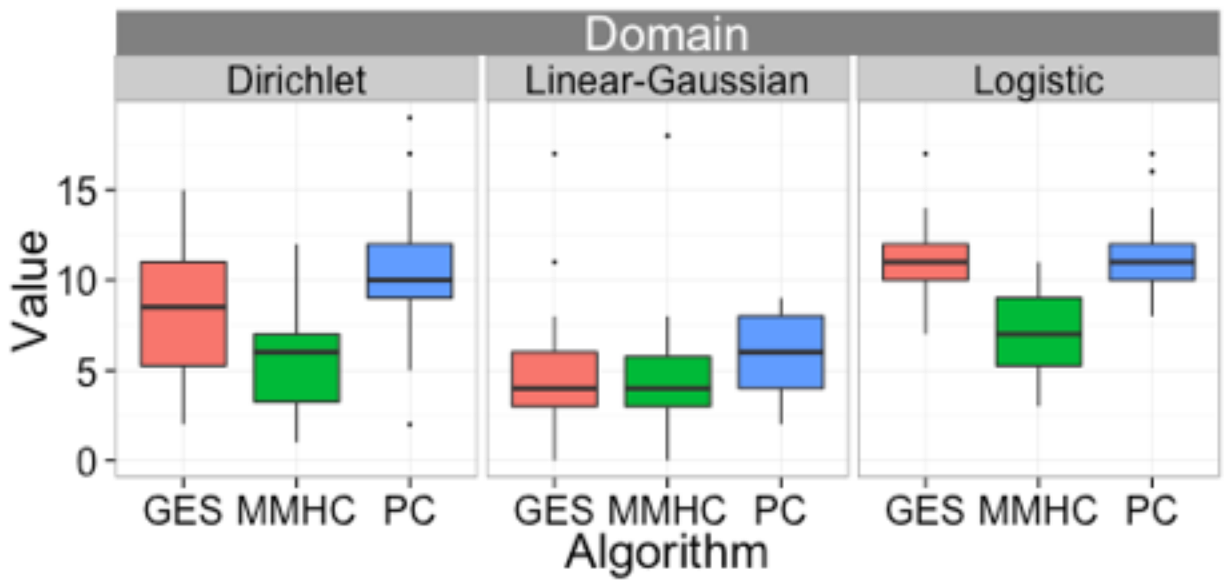
Over-specification and Under-specification

- We created DAG models derived from the true structure of our real software domains:
  - **Over-specified:** The parent set of each outcome is a strict superset of the true parent set
  - **Under-specified:** The parent set of each outcome is a strict subset of the true parent set
- Then, we evaluated these models against the ground truth structure and interventional distribution

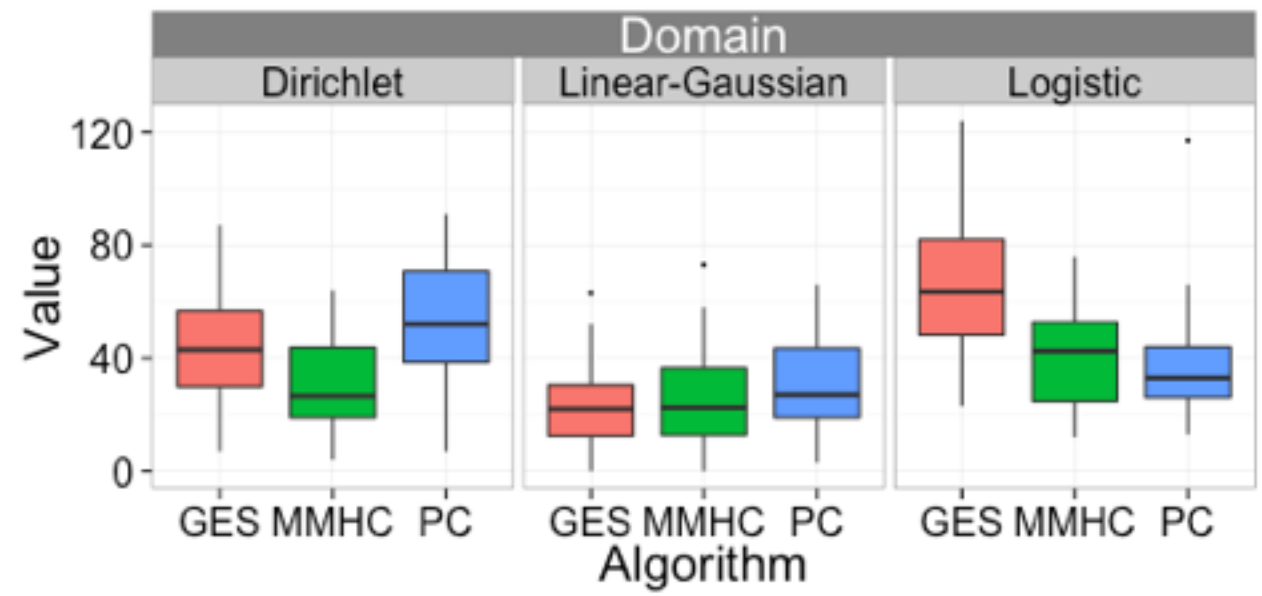
Domain	Subjects	Model Type	SID: Min, Median, Max			SHD: Min, Median, Max			TV: Min, Median, Max		
JDK	473	Over-specify	0	0	0	1	3	3	0.04	0.17	0.21
		Under-specify	4	5	9	2	2	4	0.22	0.41	0.58
Postgres	5,000	Over-specify	0	0	0	0	1	2	0.00	0.06	0.09
		Under-specify	4	6	8	3	4	5	0.17	0.35	0.61
HTTP	2,599	Over-specify	0	0	0	1	2	4	0.06	0.06	0.09
		Under-specify	2	6	10	1	3	4	0.22	0.25	0.30

Relative Performance of Algorithms

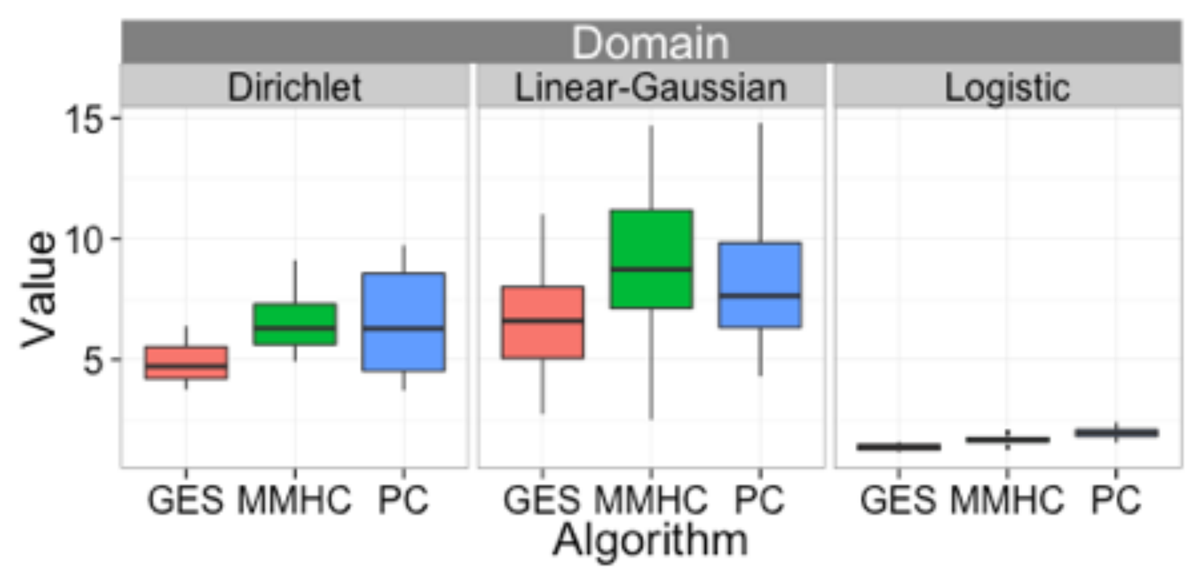
SID



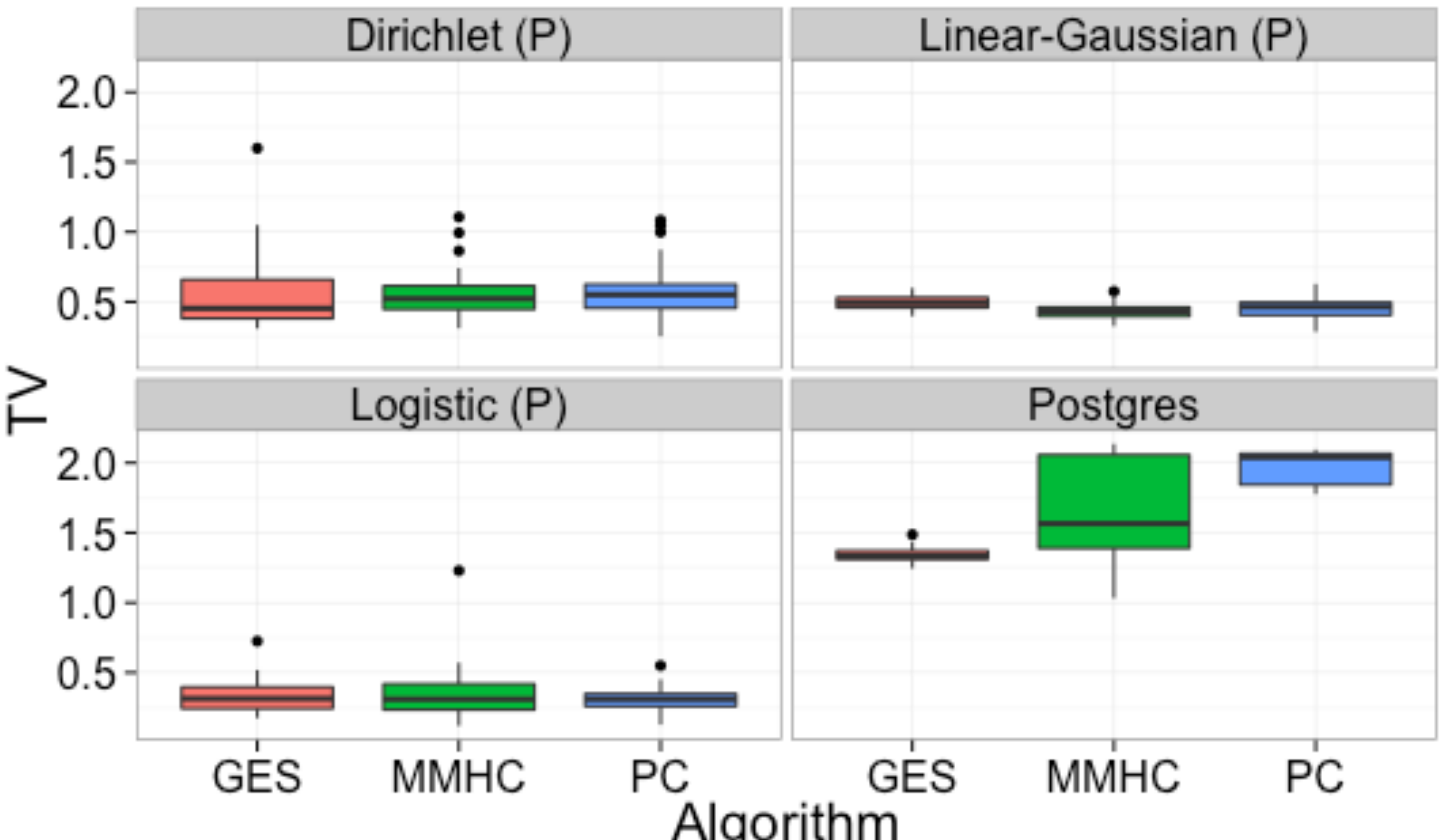
SHD



TV



Revisiting Synthetic Data Generation



Conclusions

- Existing approaches to evaluation are strictly structural, and do not characterize the full causal inference pipeline
- Statistical distances can be used to evaluate interventional distribution quality
- Evaluation with statistical distance can lead to different conclusions about algorithmic performance