When causality matters for prediction: Investigating the practical tradeoffs

Robert E. Tillman   *Peter Spirtes*

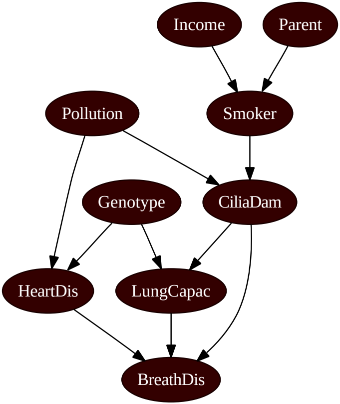
Carnegie Mellon

Department of Philosophy
College of Humanities and Social Sciences

Machine Learning Department
School of Computer Science

NIPS 2008 Workshop on Causality: Objectives and Assessment

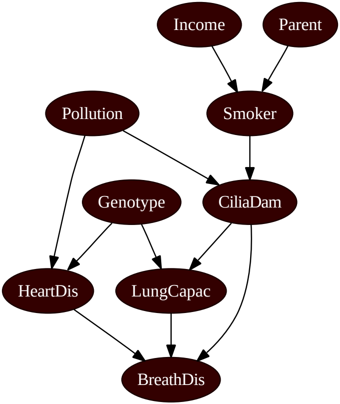
Causal Discovery



The Usual Setup:

- Unobserved data generating process
- i.i.d. sample

Causal Discovery



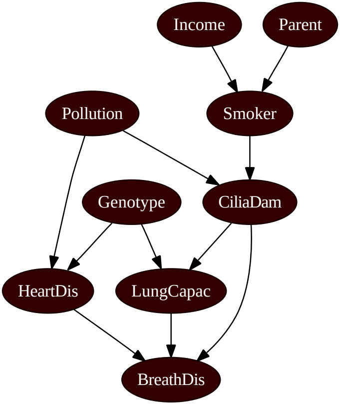
The Usual Setup:

- Unobserved data generating process
- i.i.d. sample

Objective:

- Learn structure, e.g. causal Bayesian network

Causal Discovery



The Usual Setup:

- Unobserved data generating process
- i.i.d. sample

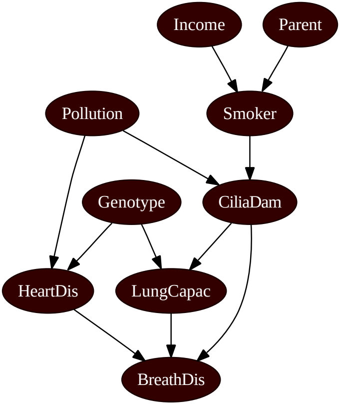
Objective:

- Learn structure, e.g. causal Bayesian network

Assessment:

- Compare to "ground truth", i.e. simulations, experimental studies, expert knowledge

Causal Discovery



The Usual Setup:

- Unobserved data generating process
- i.i.d. sample

Objective:

- Learn structure, e.g. causal Bayesian network

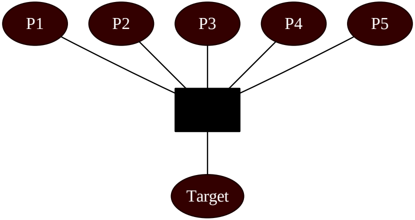
Assessment:

- Compare to "ground truth", i.e. simulations, experimental studies, expert knowledge

Focus:

- Learn network models that accurately depict the data generating mechanism

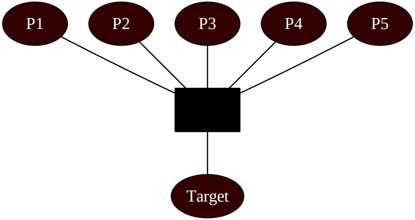
Prediction



The Standard Problem:

- “Target” variable associated with “predictor” variables
- i.i.d sample (training data)

Prediction



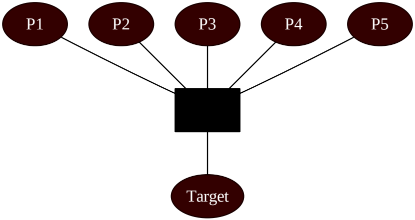
The Standard Problem:

- “Target” variable associated with “predictor” variables
- i.i.d sample (training data)

Objective:

- Predict target from values of predictor variables

Prediction



The Standard Problem:

- "Target" variable associated with "predictor" variables
- i.i.d sample (training data)

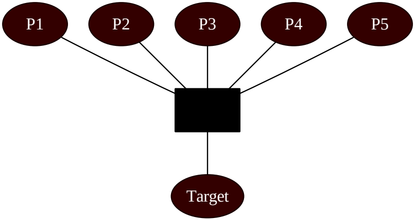
Objective:

- Predict target from values of predictor variables

Assessment:

- Compare predictions to known target values, i.e. testing data, cross validation

Prediction



The Standard Problem:

- “Target” variable associated with “predictor” variables
- i.i.d sample (training data)

Objective:

- Predict target from values of predictor variables

Assessment:

- Compare predictions to known target values, i.e. testing data, cross validation

Focus:

- Train classifier/regression model that minimizes loss function, e.g. makes accurate predictions
- Model need not resemble the true data generating mechanism, i.e. Naive Bayes

Causal Discovery and Prediction

Previous focus: predicting the effects of possible interventions:

- Specify the distribution for a manipulated population
- Counterfactuals

Causal Discovery and Prediction

Previous focus: predicting the effects of possible interventions:

- Specify the distribution for a manipulated population
- Counterfactuals
- Assume intervention has not been performed, e.g. no data from manipulated population

Causal Discovery and Prediction

Previous focus: predicting the effects of possible interventions:

- Specify the distribution for a manipulated population
- Counterfactuals
- Assume intervention has not been performed, e.g. no data from manipulated population

Causation and Prediction Challenge:

- Training data from unmanipulated population

Causal Discovery and Prediction

Previous focus: predicting the effects of possible interventions:

- Specify the distribution for a manipulated population
- Counterfactuals
- Assume intervention has not been performed, e.g. no data from manipulated population

Causation and Prediction Challenge:

- Training data from unmanipulated population
- (Structural) intervention is performed
- System stabilizes

Causal Discovery and Prediction

Previous focus: predicting the effects of possible interventions:

- Specify the distribution for a manipulated population
- Counterfactuals
- Assume intervention has not been performed, e.g. no data from manipulated population

Causation and Prediction Challenge:

- Training data from unmanipulated population
- (Structural) intervention is performed
- System stabilizes
- Draw i.i.d sample for predictors from manipulated population
- Predict target using predictor values from stabilized manipulated distribution

Causation and Prediction Challenge

Results:

- In some instances, noncausal methods outperformed causal methods

Causation and Prediction Challenge

Results:

- In some instances, noncausal methods outperformed causal methods

Questions:

- Is causality useful for standard prediction tasks?

Causation and Prediction Challenge

Results:

- In some instances, noncausal methods outperformed causal methods

Questions:

- Is causality useful for standard prediction tasks?
- Is it useful in practice?

Causation and Prediction Challenge

Results:

- In some instances, noncausal methods outperformed causal methods

Questions:

- Is causality useful for standard prediction tasks?
- Is it useful in practice?
- Is this a realistic scenario?

Causation and Prediction Challenge

Results:

- In some instances, noncausal methods outperformed causal methods

Questions:

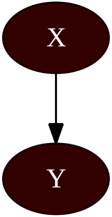
- Is causality useful for standard prediction tasks?
- Is it useful in practice?
- Is this a realistic scenario?

Possible Explanations:

- Sampling error, overfitting
- Parametric assumptions do not hold, i.e. linearity, Gaussianity
- Prediction for target is invariant under the manipulation.

Invariance of prediction under manipulations

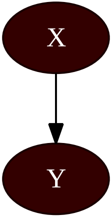
Simple example:



Bayes optimal prediction for Y is $P(Y|X)$

Invariance of prediction under manipulations

Simple example:

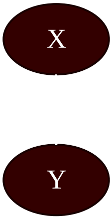


Bayes optimal prediction for Y is $P(Y|X)$

- Manipulating X does not change distribution of $P(Y|X)$, still Bayes optimal
- Prediction (once system stabilizes) is invariant under manipulation

Invariance of prediction under manipulations

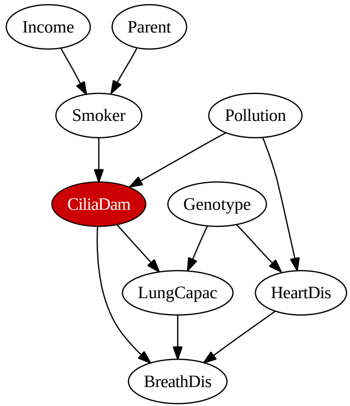
Simple example:



Bayes optimal prediction for Y is $P(Y|X)$

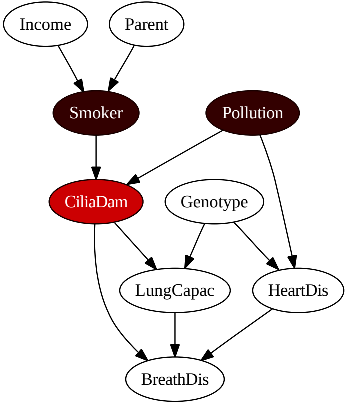
- Manipulating Y does change distribution of $P(Y|X)$, Y depends on manipulation
- Incorrect predictions in stabilized manipulated population

Terminology



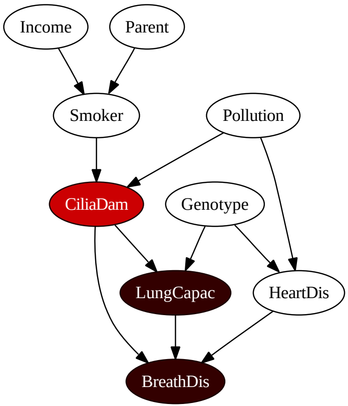
Predict CiliaDam

Terminology



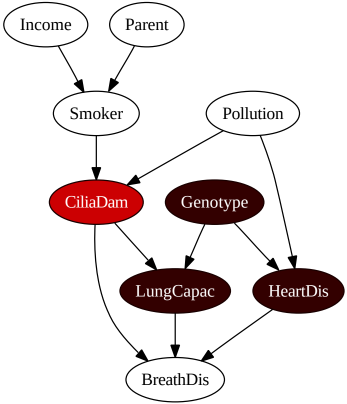
Parents of CiliaDam

Terminology



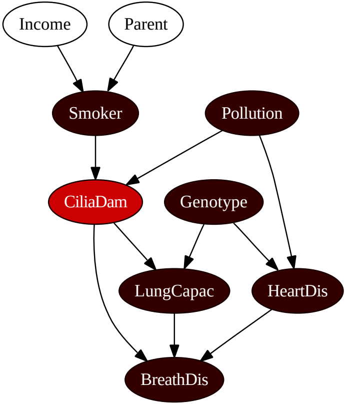
Children of CiliaDam

Terminology



Coparents (spouses) of CiliaDam

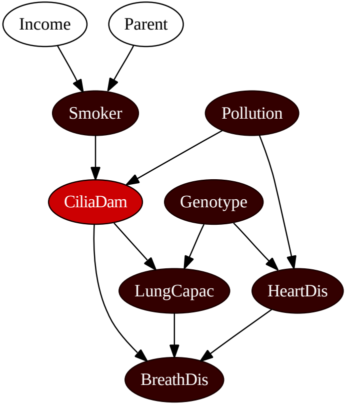
Terminology



Definition

In a causal Bayesian network $\mathcal{B} = \langle \mathcal{G}, P \rangle$ over variables $\mathbf{V}$, the Markov Blanket for $X \in \mathbf{V}$ is the minimal set of variables $\mathbf{MB}_X^{\mathcal{G}} \subseteq \mathbf{V}/\{X\}$ such that $X \perp\!\!\!\perp \mathbf{V}/\mathbf{MB}_X^{\mathcal{G}} \mid \mathbf{MB}_X^{\mathcal{G}}$.

Terminology



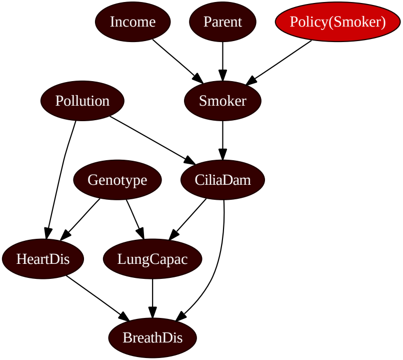
Definition

In a causal Bayesian network $\mathcal{B} = \langle \mathcal{G}, P \rangle$ over variables $\mathbf{V}$, the Markov Blanket for $X \in \mathbf{V}$ is the minimal set of variables $\mathbf{MB}_X^{\mathcal{G}} \subseteq \mathbf{V}/\{X\}$ such that $X \perp\!\!\!\perp \mathbf{V}/\mathbf{MB}_X^{\mathcal{G}} \mid \mathbf{MB}_X^{\mathcal{G}}$.

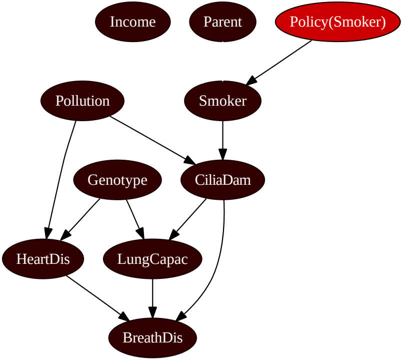
Theorem (Pearl, 1988)

The Markov blanket for X consists of the parents, children and coparents of X in $\mathcal{G}$.

Interventions

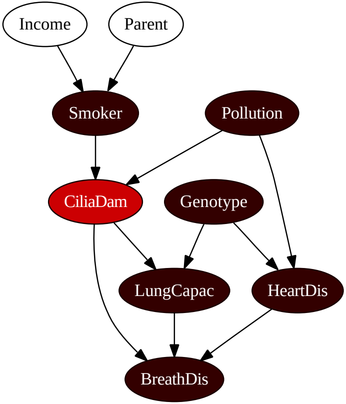


Policy(Smoker)=0



Policy(Smoker)=1

Conditions for invariance of prediction under manipulations



Theorem (Prediction invariance)

In a causal Bayesian network $\mathcal{B} = \langle \mathcal{G}, P \rangle$ over variables $\mathbf{V}$, let $T \in \mathbf{V}$ be a target, $\mathbf{X} \subseteq \mathbf{V}$ a set of predictor variables, and $\mathbf{Y} \subseteq \mathbf{V}$ the set of manipulated variables. If $\mathbf{X} \supseteq \mathbf{MB}_T^{\mathcal{G}}$ and $\forall Y \in \mathbf{Y}, Y \neq T$ and $Y \notin \mathbf{Children}(T)$, then prediction of T using $\mathbf{X}$ is invariant under the manipulation.

Conditions for invariance of prediction under manipulations

$$P(T \mid \mathbf{X}) = P(T \mid \mathbf{MB}_T^{\mathcal{G}})$$

Conditions for invariance of prediction under manipulations

$$P(T \mid \mathbf{X}) = P(T \mid \mathbf{MB}_T^{\mathcal{G}}) = \frac{P(T, \mathbf{MB}_T^{\mathcal{G}})}{\sum_T P(T, \mathbf{MB}_T^{\mathcal{G}})}$$

Conditions for invariance of prediction under manipulations

$$
\begin{aligned}
P(T \mid \mathbf{X}) &= P(T \mid \mathbf{MB}_T^{\mathcal{G}}) \\
&= \frac{P(T, \mathbf{MB}_T^{\mathcal{G}})}{\sum_T P(T, \mathbf{MB}_T^{\mathcal{G}})} \\
&= \frac{\prod_{X \in T \cup \mathbf{Children}(T) \cup \mathbf{Parents}(T) \cup \mathbf{Coparents}(T)} P(X \mid \mathbf{Parents}(T))}{\sum_T \prod_{X \in T \cup \mathbf{Children}(T) \cup \mathbf{Parents}(T) \cup \mathbf{Coparents}(T)} P(X \mid \mathbf{Parents}(T))}
\end{aligned}
$$

in the Markov blanket subgraph

Conditions for invariance of prediction under manipulations

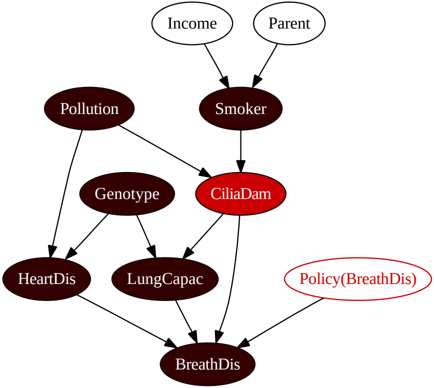
$$
\begin{aligned}
P(T \mid \mathbf{X}) &= P(T \mid \mathbf{MB}_T^{\mathcal{G}}) \\
&= \frac{P(T, \mathbf{MB}_T^{\mathcal{G}})}{\sum_T P(T, \mathbf{MB}_T^{\mathcal{G}})} \\
&= \frac{\prod_{X \in T \cup \mathbf{Children}(T) \cup \mathbf{Parents}(T) \cup \mathbf{Coparents}(T)} P(X \mid \mathbf{Parents}(T))}{\sum_T \prod_{X \in T \cup \mathbf{Children}(T) \cup \mathbf{Parents}(T) \cup \mathbf{Coparents}(T)} P(X \mid \mathbf{Parents}(T))}
\end{aligned}
$$

in the Markov blanket subgraph

$\ldots$

$$
= \frac{\prod_{X \in T \cup \mathbf{Children}(T)} P(X \mid \mathbf{Parents}(T))}{\sum_T \prod_{X \in T \cup \mathbf{Children}(T)} P(X \mid \mathbf{Parents}(T))}
$$

Correcting for manipulations

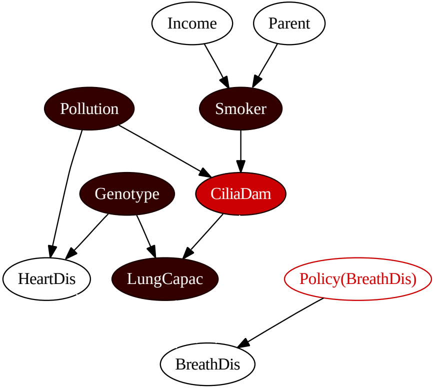


$$Policy(BreathDis) = 0$$

Theorem (Causal correction)

In a causal Bayesian network $\mathcal{B} = \langle \mathcal{G}, P \rangle$ over variables $\mathbf{V}$, let T be a target and $\mathbf{Y} \subseteq \mathbf{V}$ the set of manipulated variables.
$P\left(T \mid \mathbf{MB}_T^{\mathcal{G}(\text{Policy}(\mathbf{Y}))}\right)$, is invariant under the manipulation of $\mathbf{Y}$ if $\nexists Y \in \mathbf{Y}$, such that $Y \in \mathbf{Children}(T)$ and Y is an ancestor of some $C \in \mathbf{Children}(T) \cap \mathbf{V}/\mathbf{Y}$.

Correcting for manipulations

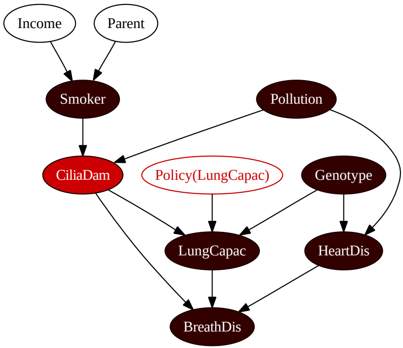


$$Policy(BreathDis) = 1$$

Theorem (Causal correction)

In a causal Bayesian network $\mathcal{B} = \langle \mathcal{G}, P \rangle$ over variables $\mathbf{V}$, let T be a target and $\mathbf{Y} \subseteq \mathbf{V}$ the set of manipulated variables. $P\left(T \mid \mathbf{MB}_T^{\mathcal{G}(\text{Policy}(\mathbf{Y}))}\right)$, is invariant under the manipulation of $\mathbf{Y}$ if $\nexists Y \in \mathbf{Y}$, such that $Y \in \mathbf{Children}(T)$ and Y is an ancestor of some $C \in \mathbf{Children}(T) \cap \mathbf{V}/\mathbf{Y}$.

Correcting for manipulations

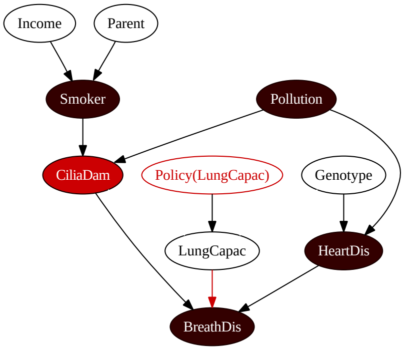


$$Policy(BreathDis) = 0$$

Theorem (Causal correction)

In a causal Bayesian network $\mathcal{B} = \langle \mathcal{G}, P \rangle$ over variables $\mathbf{V}$, let T be a target and $\mathbf{Y} \subseteq \mathbf{V}$ the set of manipulated variables. $P\left(T \mid \boldsymbol{MB}_T^{\mathcal{G}(\text{Policy}(\mathbf{Y}))}\right)$, is invariant under the manipulation of $\mathbf{Y}$ if $\nexists Y \in \mathbf{Y}$, such that $Y \in \boldsymbol{Children}(T)$ and Y is an ancestor of some $C \in \boldsymbol{Children}(T) \cap \mathbf{V}/\mathbf{Y}$.

Correcting for manipulations



$Policy(BreathDis) = 1$
Make Correction!

Theorem (Causal correction)

In a causal Bayesian network $\mathcal{B} = \langle \mathcal{G}, P \rangle$ over variables $\mathbf{V}$, let T be a target and $\mathbf{Y} \subseteq \mathbf{V}$ the set of manipulated variables. $P\left(T \mid \boldsymbol{MB}_T^{\mathcal{G}(Policy(\mathbf{Y}))}\right)$, is invariant under the manipulation of $\mathbf{Y}$ if $\nexists Y \in \mathbf{Y}$, such that $Y \in \boldsymbol{Children}(T)$ and Y is an ancestor of some $C \in \boldsymbol{Children}(T) \cap \mathbf{V}/\mathbf{Y}$.

Experiments

Hypotheses:

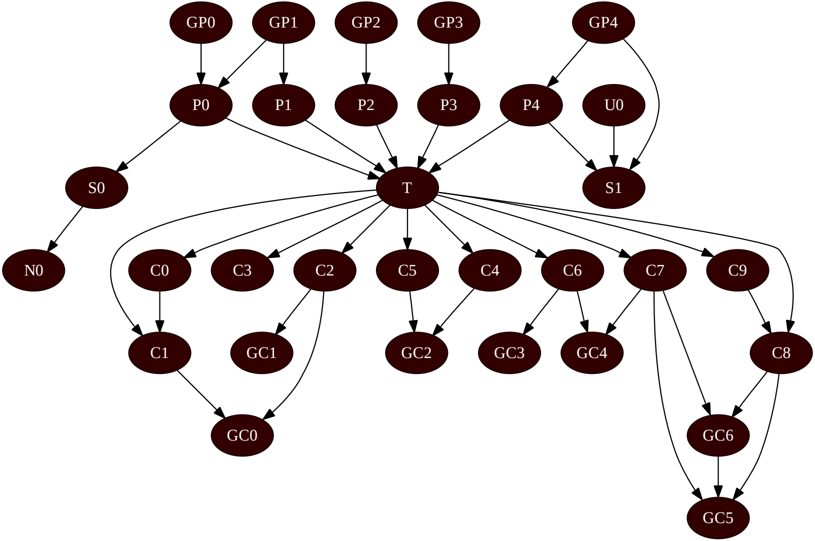
- Noncausal methods will be equivalent or better when no children are manipulated

Experiments

Hypotheses:

- Noncausal methods will be equivalent or better when no children are manipulated
- Causal methods will do increasingly better than noncausal methods as more children are manipulated

Model for experiments



Experiments

Method:

- Train causal and noncausal prediction methods on unmanipulated population (linear Gaussians)

Experiments

Method:

- Train causal and noncausal prediction methods on unmanipulated population (linear Gaussians)
- Manipulate 0, 5, 10 random nonchildren of T (including Markov blanket)

Experiments

Method:

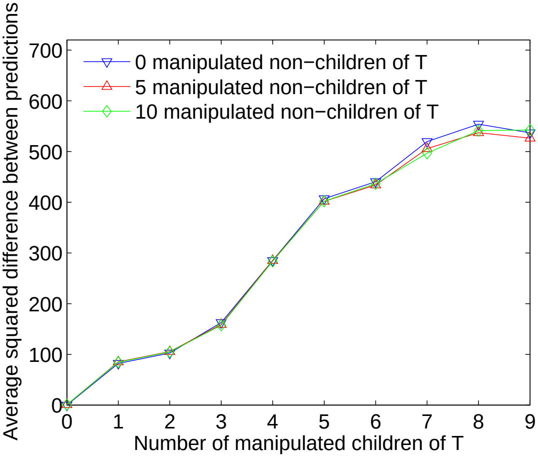
- Train causal and noncausal prediction methods on unmanipulated population (linear Gaussians)
- Manipulate 0, 5, 10 random nonchildren of T (including Markov blanket)
- Manipulate $0, \ldots, 9$ children of T in addition

Experiments

Method:

- Train causal and noncausal prediction methods on unmanipulated population (linear Gaussians)
- Manipulate 0, 5, 10 random nonchildren of T (including Markov blanket)
- Manipulate $0, \ldots, 9$ children of T in addition
- Predict T from manipulated distribution

Differences between distributions



Squared difference between ground truth predictions for T using unmanipulated and manipulated model

Prediction methods

Noncausal Methods:

LR-ALL	linear regression using all predictors
LR-MB	linear regression using only the Markov blanket
LASSO	"least absolute shrinkage and selection operator"
SVR-RBF	support vector regression using radial kernel
RVR-RBF	relevance vector regression using radial kernel

Prediction methods

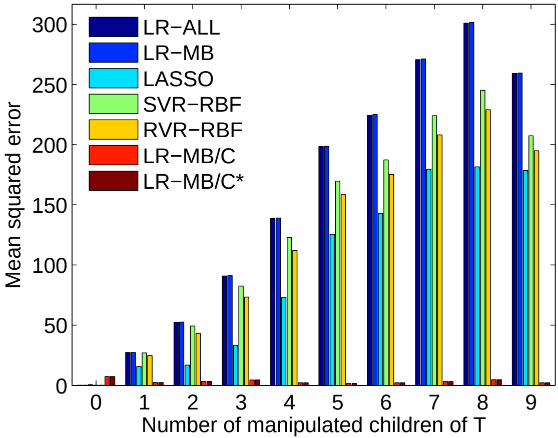
Noncausal Methods:

LR-ALL	linear regression using all predictors
LR-MB	linear regression using only the Markov blanket
LASSO	"least absolute shrinkage and selection operator"
SVR-RBF	support vector regression using radial kernel
RVR-RBF	relevance vector regression using radial kernel

Causal Methods:

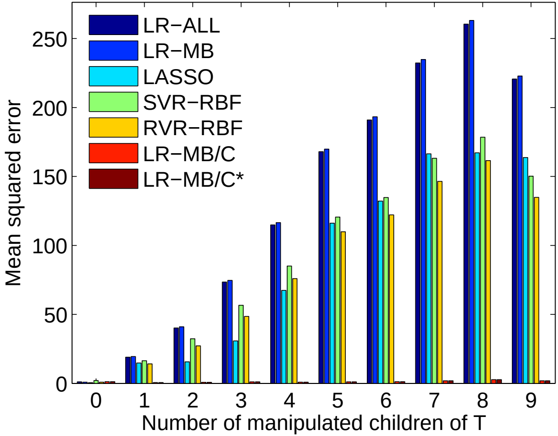
LR-MB/C	linear regression with Markov blanket correcting for manipulated children
LR-MB/C*	linear regression with Markov blanket correcting for manipulated children and active paths to unmanipulated children

Total prediction error



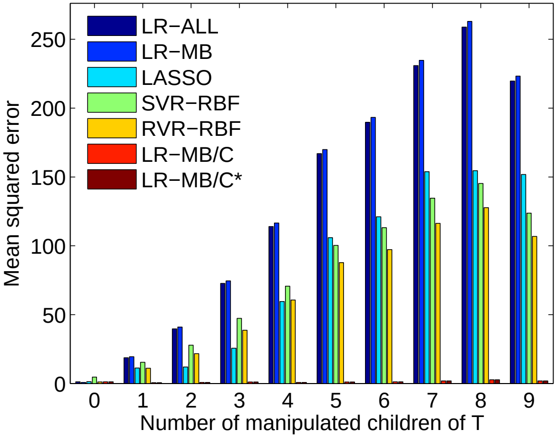
0 Manipulated Nonchildren of T

Total prediction error



5 Manipulated Nonchildren of T

Total prediction error



10 Manipulated Nonchildren of T

Nonlinear data

- Repeated previous simulations adding nonlinear dependencies

Nonlinear data

- Repeated previous simulations adding nonlinear dependencies
- Results so far inconclusive
- In general, nonparametric methods do best, though poor performance in all cases

Conclusions

Is causality relevant for prediction?

Conclusions

Is causality relevant for prediction?

- Unless noncausal method is invariant under the manipulation

Conclusions

Is causality relevant for prediction?

- Unless noncausal method is invariant under the manipulation
- But causality is needed to know noncausal methods are invariant!

Conclusions

Is causality relevant for prediction?

- Unless noncausal method is invariant under the manipulation
- But causality is needed to know noncausal methods are invariant!

In practice?

- Tradeoff between errors related to causality and errors related to parametric assumptions, overfitting, etc.

Conclusions

Is causality relevant for prediction?

- Unless noncausal method is invariant under the manipulation
- But causality is needed to know noncausal methods are invariant!

In practice?

- Tradeoff between errors related to causality and errors related to parametric assumptions, overfitting, etc.
- Noncausal prediction may be frequently invariant (or *almost* invariant)
- Advantages of nonparametric methods and methods which deal with overfitting well may cancel out errors related to causality

Conclusions

Is causality relevant for prediction?

- Unless noncausal method is invariant under the manipulation
- But causality is needed to know noncausal methods are invariant!

In practice?

- Tradeoff between errors related to causality and errors related to parametric assumptions, overfitting, etc.
- Noncausal prediction may be frequently invariant (or *almost* invariant)
- Advantages of nonparametric methods and methods which deal with overfitting well may cancel out errors related to causality
- Many other variables involved, analysis incomplete