When causality matters for prediction: investigating the practical tradeoffs

Robert E. Tillman RTILLMAN@ANDREW.CMU.EDU
Department of Philosophy and Machine Learning Department, School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, United States

Peter Spirtes PS7Z@ANDREW.CMU.EDU
Department of Philosophy
Carnegie Mellon University
Pittsburgh, PA 15213, United States

Editor: Isabelle Guyon, Dominik Janzing and Bernhard Schölkopf

Abstract

Recent evaluations have indicated that in practice, general methods for prediction which do not account for changes in the conditional distribution of a target variable given feature values in some cases outperform causal discovery based methods for prediction which *can* account for such changes. We investigate some possibilities which may explain these findings. We give theoretical conditions, which are confirmed experimentally, for when particular manipulations of variables should not affect predictions for a target. We then consider the tradeoff between errors related to causality, i.e. not accounting for changes in a distribution after variables are manipulated, and errors resulting from sample bias, overfitting, and assuming specific parametric forms that do not fit the data, which most existing causal discovery based methods are particularly prone to making.

Keywords: causal discovery, prediction, interventions

1. Introduction

Most methods in machine learning are intended primarily for *prediction*. Given *training data* for a target variable T to be predicted and a set of associated *predictor* variables $\mathbf{X}$, the goal is to use the training data to learn a prediction function $T = f(\mathbf{X})$ that can be used to predict values for the target given values for the predictor variables, assuming the conditional distribution $P(T|\mathbf{X})$ does not change after the training data is collected. In general, we are not concerned with whether the prediction function actually depicts true causal relationships between variables in the underlying data generating mechanism; we care only whether it makes accurate predictions.

The advantage of *causal discovery* methods is that they can be be used to learn models that depict true data generating mechanisms. We can discover particular causal relationships between variables to determine which variables should be manipulated when setting policies to achieve a desired effect. We can use the resulting *causal models* to predict the effects of such manipulations or make predictions for a target variable when we have data for the predictor variables even if some variables have been manipulated since the training data was collected.

Evaluations of causal discovery methods have focused primarily on how closely the resulting causal models resemble true data generating mechanisms obtained either through simulations or data from controlled experiments. There has been little focus on how accurate predictions made using causal discovery methods after variables are manipulated are relative to known values for a predicted variable, i.e. using test data from the manipulated population, which is the primary means for evaluating most other methods in machine learning. Recent results from a causality challenge[1] have raised questions as to whether existing causal discovery methods are useful for making such predictions. In the challenge, some participants used prediction methods which ignored causality to predict a target variable after predictor variables were manipulated, i.e.. applying support vector machines trained using the unmanipulated data to make predictions for the manipulated data without any adjustments to account for the manipulations, and in some cases achieved results that were better than any of the participants who used causal discovery based methods for prediction.

One possible explanation for these results is that there is a noticeable tradeoff when using causal discovery methods to make such predictions: while causal discovery methods may make the proper adjustments to account for the change in a distribution after variables are manipulated, prediction with causal discovery based methods may result in significant errors due to overfitting and sampling bias as well as parametric assumptions, i.e. linearity, Gaussianity, which do not hold. There is nothing inherent in causal models or causal inference that requires parametric assumptions that are more restrictive than other machine learning methods; however, most[2] existing causal discovery algorithms do require such assumptions. Thus, most causal discovery methods for prediction may result in considerably more of this second type of error than many other nonparametric methods in machine learning, such as support vector machines. Furthermore, in many cases, manipulating a particular variable in a causal system will have no effect on the predicted value of a particular target, e.g. if the manipulated variable is conditionally independent of the target variable given the set of predictor variables. Thus, if certain parametric assumptions made by causal discovery algorithms do not hold for some data, then we should expect nonparametric methods for predictions and methods which make less strict parametric assumptions to outperform causal discovery based methods for predictions even for some cases where variables are manipulated after the training data is collected.

In this paper, we begin to investigate this tradeoff. We review the relevant terminology in section 2. In section 3, we present theoretical conditions which distinguish manipulations which do affect predictions for a target from those which do not and demonstrate how causal discovery methods used for prediction can account for the change in distribution. In section 4, we then experimentally test these conditions using synthetic data to confirm that they at least hold in the cases favorable for casual discovery methods. Conclusions are offerred in section 5.

2. Formal preliminaries

We first introduce some terminology. A *directed graph* $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ is a set of nodes $\mathcal{V}$, which represent variables, and a set of directed edges $\mathcal{E}$ connecting distinct nodes. For a node $V \in \mathcal{V}$, $\mathbf{Pa}_V^{\mathcal{G}}$ refers to the set of nodes that are parents of V (nodes with an edge directed into V), $\mathbf{Ch}_V^{\mathcal{G}}$ refers to the children of V (nodes with an edge directed out of V), and $\mathbf{Co}_V^{\mathcal{G}}$ refers to the coparents (spouses) of V (parents of children of V other than V). A *trail* in $\mathcal{G}$ is a sequence of nodes such

[1] See http://www.causality.inf.ethz.ch/challenge.php for details.

[2] There have been several recent proposals which require less restrictive parametric assumptions, i.e. Shimizu et al. (2006), Hoyer et al. (2008), Hoyer et al. (2009).

that each adjacent pair in the sequence is connected by an edge (ignoring directions), and no node appears more than once in the sequence. A trail is a *directed path* if every edge points in the same direction. $\mathcal{G}$ is a *directed acyclic graph* (DAG) if for every pair $\{X,Y\} \subseteq \mathcal{V}$, there are not directed paths from both X to Y and Y to X (no directed cycles). X is an *ancestor* (*descendant*) of Y if there is a directed path from X to Y (Y to X). A *v-structure* (*collider*) is a triple of nodes $\langle X,Y,Z \rangle$ such that X and Z are parents of Y.[3] A trail is *active* given a conditioning set $\mathbf{C} \subseteq \mathcal{V}$ if (i) for every v-structure $\langle X,Y,Z \rangle$ in the trail either $Y \in \mathbf{C}$ or some descendant of Y is in $\mathbf{C}$ and (ii) no other node in the trail is in $\mathbf{C}$. For disjoint sets of nodes, $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$, $\mathbf{X}$ is d-separated from $\mathbf{Y}$ given $\mathbf{Z}$ if and only if there are no active trails between any $X \in \mathbf{X}$ and any $Y \in \mathbf{Y}$ given $\mathbf{Z}$.

A *Bayesian network* $\mathcal{B}$ is a pair $\langle \mathcal{G},\mathcal{P} \rangle$, where $\mathcal{G} = \langle \mathcal{V},\mathcal{E} \rangle$ is a DAG and $\mathcal{P}$ is a joint probability distribution over the variables represented by the nodes in $\mathcal{V}$ such that $\mathcal{P}$ can be factored as follows:

$$\mathcal{P}(\mathcal{V}) = \prod_{V \in \mathcal{V}} P(V|\mathbf{Pa}_V^{\mathcal{G}})$$

If $\mathbf{X}$ is d-separated from $\mathbf{Y}$ given $\mathbf{Z}$ in $\mathcal{G}$, then $\mathbf{X}$ is conditionally independent of $\mathbf{Y}$ given $\mathbf{Z}$ in $\mathcal{P}$ (Pearl, 1988). For disjoint sets of nodes, $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{Z}$ in $\mathcal{V}$, $\mathcal{P}$ is *faithful* to $\mathcal{G}$ if $\mathbf{X}$ is d-separated from $\mathbf{Y}$ given $\mathbf{Z}$ in $\mathcal{G}$ whenever $\mathbf{X}$ is conditionally independent of $\mathbf{Y}$ given $\mathbf{Z}$ in $\mathcal{P}$ (Spirtes et al., 2000). $\mathcal{B}$ is said to be a *causal* Bayesian network if an edge from X to Y indicates that X is a direct cause of Y relative to $\mathcal{V}$. When performing causal inference, it is generally assumed that the distribution over the observed variables $\mathcal{P}$ factors according to a DAG $\mathcal{G}$ in a causal Bayesian network $\mathcal{B} = \langle \mathcal{G},\mathcal{P} \rangle$ and $\mathcal{P}$ is faithful to $\mathcal{G}$. In this paper, we assume that there are no unmeasured common causes of variables in $\mathcal{V}$.

For a Bayesian network $\mathcal{B} = \langle \mathcal{G},\mathcal{P} \rangle$, where $\mathcal{G} = \langle \mathcal{V},\mathcal{E} \rangle$, a *Markov blanket* for some node $V \in \mathcal{V}$ in $\mathcal{G}$, $\mathbf{MB}_V^{\mathcal{G}}$, is a minimal set of variables in $\mathcal{V}/\{V\}$ such that V is conditionally independent of $\mathcal{V}/\{\mathbf{MB}_V^{\mathcal{G}} \cup \{V\}\}$ given $\mathbf{MB}_V^{\mathcal{G}}$. If $\mathcal{P}$ is faithful to $\mathcal{G}$, then $\mathbf{MB}_V^{\mathcal{G}} = \mathbf{Pa}_V^{\mathcal{G}} \cup \mathbf{Ch}_V^{\mathcal{G}} \cup \mathbf{Co}_V^{\mathcal{G}}$, for any $V \in \mathcal{V}$ (Pearl, 1988).

We represent manipulations of variables $\mathbf{Z} \subseteq \mathcal{V}$ in a causal Bayesian network $\mathcal{B} = \langle \mathcal{G},\mathcal{P} \rangle$ where $\mathcal{G} = \langle \mathcal{V},\mathcal{E} \rangle$, by forming the new DAG $\mathcal{G}(Policy(\mathbf{Z}))$, where we introduce a new exogenous node (node without parents) $Policy(Z)$ to $\mathcal{G}$ that is a parent of only Z, for each $Z \in \mathbf{Z}$. For disjoint sets of non-policy nodes $\mathbf{X}$ and $\mathbf{Y}$, a conditional distribution $P(\mathbf{Y}|\mathbf{X})$ is *invariant* under the manipulation of $\mathbf{Z}$ if $P(\mathbf{Y}|\mathbf{X})$ is the same when the variables in $\mathbf{Z}$ are manipulated and the variables in $\mathbf{Z}$ are unmanipulated. If **Policy(Z)**, the set of all policy nodes, is d-separated from $\mathbf{Y}$ given $\mathbf{X}$ in $\mathcal{G}(Policy(\mathbf{Z}))$, then $P(\mathbf{Y}|\mathbf{X})$ is invariant under manipulation of $\mathbf{Z}$ (Spirtes et al., 2000). $\mathcal{P}_M$, the distribution resulting from manipulating the variables in $\mathbf{Z}$, factors according to the DAG $\mathcal{G}_M$, which is $\mathcal{G}$ changed by removing every edge that is directed into some $Z \in \mathbf{Z}$ (Spirtes et al., 2000).

3. Invariance of predictions under manipulations

In some cases, it is obvious that manipulations of certain variables will not affect predictions for other variables. Consider the simple case of two variables X and Y, where X causes Y and there are no common causes of X and Y. If X is manipulated, this does not change the distribution of $P(Y|X)$, which produces the Bayes optimal prediction for Y. Thus, a classifier or regression method trained using data from a population where X is not manipulated should correctly make predictions for Y using test data from a population where X is manipulated without any adjustments to account for the manipulation. The following theorem distinguishes

[3] We are using the definition given in Koller and Friedman (2008). Other sources use v-structure to refer to only such triples where X and Z are not adjacent (an *immorality* or *unshielded collider*).

the more complicated cases where manipulations do not affect predictions for a target variable from those where manipulations do affect predictions.

Theorem 1 *Let $\mathcal{B} = \langle \mathcal{G}, \mathcal{P} \rangle$ be a Bayesian network over variables $\mathcal{V}$, $T \in \mathcal{V}$ a target predicted variable, $\mathbf{X} \subseteq \mathcal{V}$ a set of predictor variables, and $\mathbf{Z} \subseteq \mathcal{V}$ a set of variables that are manipulated. If $\forall Y \in \mathbf{Y}$, $Y \neq T$ and $Y \notin \mathbf{Ch}_T^{\mathcal{G}}$, then $P(T|\mathbf{X})$ is invariant under the manipulation.*

Proof Let P be a trail between T and $Policy(Y)$ for some $Y \in \mathbf{Y}$. If P is into T, e.g. some node Z in P is a parent of T, then Z and T do not form a v-structure with some parent or child of Z in P and $Z \in \mathbf{X}$ since $Z \in \mathbf{Pa}_T^{\mathcal{G}}$ so P is not an active trail for T given $\mathbf{X}$. If P is out of T, e.g. some node Z in P is a child of T, and some child U of Z is in P, then $\langle U, Z, T \rangle$ is not a v-structure and $Z \in \mathbf{X}$ since $Z \in \mathbf{Ch}_T^{\mathcal{G}}$ so P is not an active trail for T given $\mathbf{X}$. If P is out of T and some parent U of Z other than T is in P, then U and Z do not form a v-structure with some parent or child of U and $Z \in \mathbf{X}$ and $U \in \mathbf{X}$ since $Z \in \mathbf{Pa}_T^{\mathcal{G}}$ and $U \in \mathbf{Co}_T^{\mathcal{G}}$ so P is not an active trail for T given $\mathbf{X}$. This exhausts all cases so $Policy(Y)$ and T are d-separated given $\mathbf{X}$. ■

Thus, as long as a set of predictors includes the Markov blanket for a target node, prediction will be unaffected by any manipulation which does not change the value of a child of the target, even if the manipulation changes the value of another variable in the Markov blanket. In such cases, causal knowledge will not improve the accuracy of predicted values in any way (though we will not know whether prediction is affected by a manipulation unless we know the causal relationships in the underlying data generating mechanism). While this is a straightforward result, it is important in practice. Most methods for prediction which do an explicit or implicit feature selection will likely assign high weight to (at least most of) the features in the Markov blanket. Since in most cases, the children of a particular target will consist of only a small percentage of the nodes in a graph, it is unlikely that the target will have many children that are manipulated to unlikely values, which would lead to high error in prediction. In many cases, errors resulting from manipulated children of a target that are used as features may be negligible and canceled out by other gains made by a prediction method that combats overfitting well or does not make restrictive parametric assumptions. In other cases where we may expect many children to be manipulated to unlikely values, we can use causal knowledge to select the correct set of predictors for the manipulated distribution and avoid errors resulting from manipulated children of a target that are used as features.

Theorem 2 *Let $\mathcal{B} = \langle \mathcal{G}, \mathcal{P} \rangle$ be a Bayesian network over variables $\mathcal{V}$, $T \in \mathcal{V}$ a target predicted variable, $\mathbf{X} \subseteq \mathcal{V}$ a set of predictor variables, and $\mathbf{Z} \subseteq \mathcal{V}$ a set of variables that are manipulated. If $\mathbf{X} = \mathbf{MB}_T^{\mathcal{G}_M}$, then $P(T|\mathbf{MB}_T^{\mathcal{G}_M})$ is invariant under the manipulation of $\mathbf{Z}$ if $\forall Z \in \{\mathbf{Z} \cap \mathbf{Ch}_T^{\mathcal{G}}\}$, Z is not an ancestor of some $X \in \mathbf{Ch}_T^{\mathcal{G}}$ such that $X \notin \mathbf{Z}$.*

Proof If $P(T|\mathbf{MB}_T^{\mathcal{G}_M})$ is not invariant under the manipulation of $\mathbf{Z}$, then there is an active trail R between T and $Policy(Z)$ for some $Z \in \mathbf{Z}$ given $\mathbf{MB}_T^{\mathcal{G}_M}$ in $\mathcal{G}(Policy(\mathbf{Z}))$. Let X be the node in R connected to T, and U the node connected to X in R other than T. If X is a parent of T, then $X \in \mathbf{MB}_T^{\mathcal{G}_M}$ and $\langle U, X, T \rangle$ is not a v-structure so R is not active. Thus, X is a child of T. We have the following 2 cases. Case 1: $\langle U, X, T \rangle$ is not a v-structure. R is active given $\mathbf{MB}_T^{\mathcal{G}_M}$ so $X \notin \mathbf{MB}_T^{\mathcal{G}_M}$. $X \in \mathbf{Ch}_T^{\mathcal{G}}$ and $X \notin \mathbf{MB}_T^{\mathcal{G}_M}$ so $X \in \mathbf{Z}$. $Policy(Z)$ and T are both parents in R, so X is an ancestor of the middle node of a v-structure in R. R is active so the middle node of this v-structure either is contained in or has a descendant in $\mathbf{MB}_T^{\mathcal{G}_M}$. Thus, X is an ancestor of some $W \in \mathbf{MB}_T^{\mathcal{G}_M}$. Case 2: $\langle U, X, T \rangle$ is a v-structure in R. R is active given $\mathbf{MB}_T^{\mathcal{G}_M}$ so $U \notin \mathbf{MB}_T^{\mathcal{G}_M}$. $U \in \mathbf{Co}_T^{\mathcal{G}}$ and $U \notin \mathbf{MB}_T^{\mathcal{G}_M}$ so $X \in \mathbf{Z}$. R is active given $\mathbf{MB}_T^{\mathcal{G}_M}$ so X is either contained in or has a descendant in $\mathbf{MB}_T^{\mathcal{G}_M}$. Thus, for cases 1 and 2, $X \in \mathbf{Z}$ and either X is an ancestor of some

$W \in \mathbf{MB}_T^{\mathcal{G}_M}$ or $X \in \mathbf{MB}_T^{\mathcal{G}_M}$. If $W \in \mathbf{Pa}_T^{\mathcal{G}}$ ($X \in \mathbf{Pa}_T^{\mathcal{G}}$), then there is a directed path from T to W (X) and a directed path from W (X) to T. $\mathcal{G}$ is acyclic so W (X) $\in \mathbf{Ch}_T^{\mathcal{G}} \cup \mathbf{Co}_T^{\mathcal{G}}$ and either W (X) $\notin \mathbf{Z}$ or W (X) is a parent of some $Q \in \mathbf{Ch}_T^{\mathcal{G}}$ such that $Q \notin \mathbf{Z}$. But since $X \in \mathbf{Z}$, there are only three cases: (i) $X \in \mathbf{Ch}_T^{\mathcal{G}} \cup \mathbf{Co}_T^{\mathcal{G}}$ and X is a parent of some $Q \in \mathbf{Ch}_T^{\mathcal{G}}$ such that $Q \notin \mathbf{Z}$, (ii) $W \in \mathbf{Ch}_T^{\mathcal{G}} \cup \mathbf{Co}_T^{\mathcal{G}}$ and $W \notin \mathbf{Z}$ and (iii) $W \in \mathbf{Ch}_T^{\mathcal{G}} \cup \mathbf{Co}_T^{\mathcal{G}}$ and W is a parent of some $Q \in \mathbf{Ch}_T^{\mathcal{G}}$ such that $Q \notin \mathbf{Z}$. In all three cases X is an ancestor of some $S \in \mathbf{Ch}_T^{\mathcal{G}}$ such that $S \notin \mathbf{Z}$. ■

As long as we are using the the Markov blanket for the manipulated structure, e.g. after policy nodes are added, as our set of predictors, which requires causal knowledge, predictions will not be affected by manipulated children unless there is some manipulated child of the target that is an ancestor of an unmanipulated child of the target. When it is the case that a manipulated child of the target is an ancestor of some unmanipulated child, we can still make predictions using the Markov blanket for the manipulated structure, but a correction[4] needs to be made to subtract out the influence from the manipulated variable. In practice, however, failure to make this correction usually has little effect on predictions. The importance of this theorem is that it allows us to select the correct set of predictors to account for changes in a distribution resulting from manipulations, regardless of what variables are manipulated. Thus, we should expect causal discovery based methods for prediction to perform increasingly better than methods for prediction which ignore causality as we increase the number of children of a target variable that are manipulated if the parametric assumptions made by the causal discovery methods are reasonable for some given data and error due to overfitting and sampling bias are reasonably low, even if such instances are not representative of the majority of cases. This hypothesis is evaluated in the next section.

4. Experimental results

We first constructed the graph of the causal environment around the target node T shown in figure 1a to use in the following experiments. The graph was constructed to be similar to the causal environment we learned for the target variable in one of the challenge datasets. We chose random linear Gaussian parameters for the variables in this structure and used forward sampling to generate a synthetic training dataset of size $N = 1000$. We then generated test datasets of sizes $N = 1000$ after various manipulations were made to variables in the structure. We first manipulated 0, 5, and 10 random non-children of T and then for each we also manipulated from 0 to 9 children of T. In our simulations, we manipulated variables by setting each manipulated variable's dependency with all parents to 0, its mean to an unlikely value, and its variance to a small value. Each simulation described below was repeated 100 times and the results averaged.

Before considering a realistic prediction scenario, we first show the isolated causal component of prediction errors to confirm that the distributions are changing as variables are manipulated in our simulations according to the theorems from section 3. Using the chosen parameters for the models, we calculated ground truth regression equations for the target variable in the unmanipulated model and each manipulated model. We then calculated predicted values for T using the values of all predictor variables in each test dataset with these ground truth equations. Figure 1b shows the average squared difference between the predicted values using the equations for the unmanipulated model and each manipulated model. When the number of manipulated variables in $\mathbf{Ch}_T^{\mathcal{G}}$ is 0, there is no difference between the predictions, even when 5 or 10 non-children of T are manipulated. However, as we increase the number of variables in $\mathbf{Ch}_T^{\mathcal{G}}$

[4] To compute this correction, we (i) replace the original equations for predicting the manipulated variables with the new manipulated equations (e.g. if Z is manipulated to 3, then the new equation is Z = 3), (ii) calculate the implied covariance matrix for the manipulated set of equations, and (iii) use the implied covariance matrix for the manipulated set of equations to calculate the equation for predicting the target variable in the usual way.

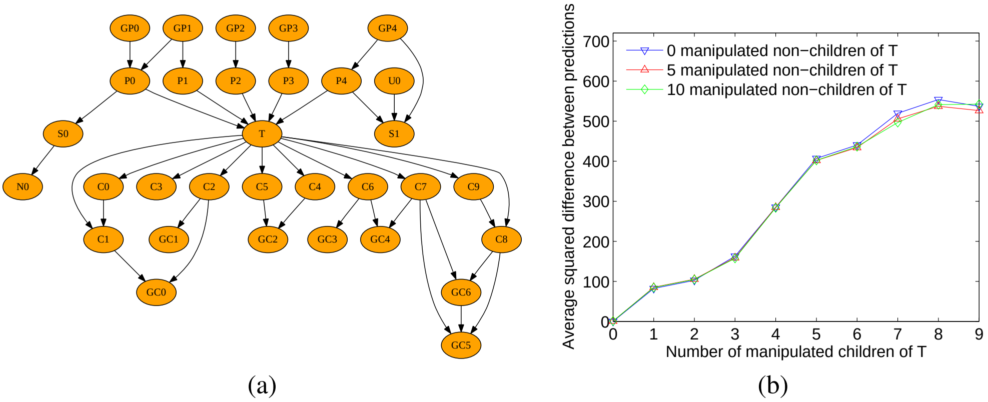

Figure 1: (a) causal structure used in the simulations, (b) averaged squared differences in predictions using the ground truth regression equations for the manipulated and unmanipulated datasets

that are manipulated, the difference between the predicted values increases at approximately the same rate regardless of the number of non-children of T that are manipulated. This confirms theorem 1. To confirm theorem 2, we repeated this procedure using only the variables in $\mathbf{MB}_T^{\mathcal{G}_M}$ as predictors for T and made corrections for manipulated children of T that are ancestors of unmanipulated children of T in a given simulation. In each of these cases, there was no difference in the predictions for T when using the ground truth regression equations for either the unmanipulated or manipulated models, indicating that the change in the distribution was correctly accounted for using the causal information.

We now consider the scenario from the causality challenge where we have only training data from the unmanipulated population and test data from some manipulated population and we know the variables that were manipulated. We used two simple causal discovery based methods for prediction: LR-MB/C and LR-MB/C*. For both of these methods we used the training data to calculate parameters for the model, then made the appropriate changes to the model and parameters to account for the manipulations, and finally used the parameters to calculate a regression equation for T using only the variables in $\mathbf{MB}_T^{\mathcal{G}_M}$, which was used to obtain a predicted value for T. For LR-MB/C* we added the additional step of correcting for manipulated nodes that are ancestors of unmanipulated nodes, as described in section 3. We used six other methods for prediction where causality was ignored: LR-ALL, LR-MB, LASSO, SVR-RBF, and RVR-RBF. In each case, a prediction function for T was learned using the training data and then applied to the manipulated test data without accounting for the manipulated variables in any way. LR-ALL and LR-MB are simply linear regression using all of the variables other than T as predictors and only the variables in $\mathbf{MB}_T^{\mathcal{G}}$ as predictors, respectively. LASSO is the “least absolute shrinkage and selection operator”, which uses the L_1 penalty to obtain a sparse linear regression model (Tibshirani, 1996). SVR-RBF is support vector regression with a Gaussian RBF kernel (Smola and Schölkopf, 1998), RVR-RBF is relevance vector regression with a Gaussian RBF kernel (Tipping, 2001). Figures 2, 3, and 4 show the mean squared errors for the predicted values for T for each method as the number of manipulated children increases from 0 to 9, when 0, 5, and 10 non-children of T are manipulated, respectively.

As expected, the methods which take advantage of the causal structure perform no better than the methods that ignore causality when we manipulate 0, 5, or 10 non-children of T as

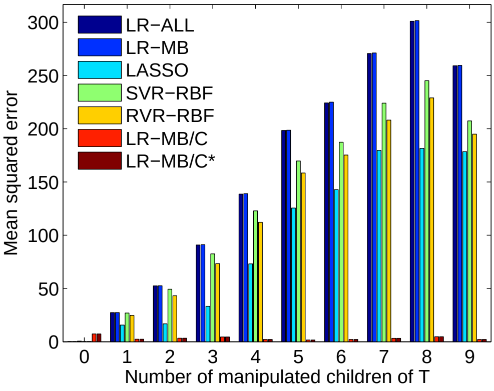

Figure 2: Mean squared error when 0 non-children of T are manipulated

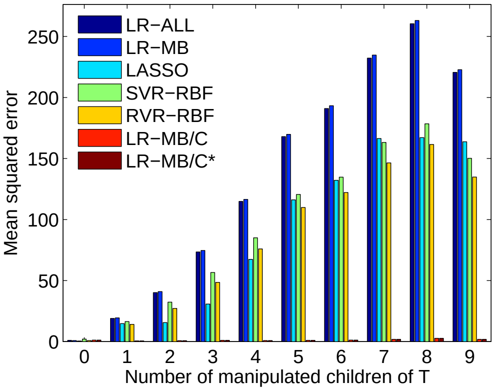

Figure 3: Mean squared error when 5 non-children of T are manipulated

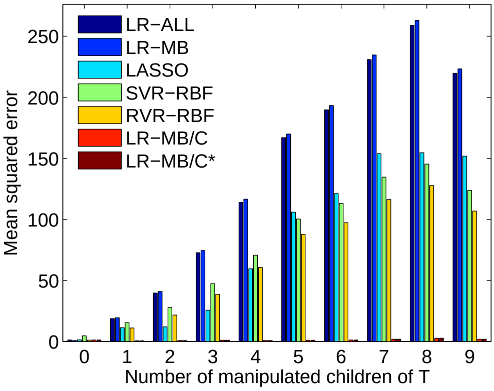

Figure 4: Mean squared error when 10 non-children of T are manipulated

long as no children of T are manipulated. If fact, when no variables are manipulated, the causal methods show the highest error. However, as we manipulate children of T, the accuracy of the causal methods does not change, but the non-causal methods begin to perform progressively worse. The trend as the number of manipulated children increases appears relatively constant for the 0, 5, and 10 manipulated non-children cases. We also note that the difference between LR-MB/C and LR-MB/C* are not noticeable in any case, indicating that the correction applied with the LR-MB/C* method does not make a considerable difference in practice. We attempted the same simulations after adding nonlinear dependencies between the variables to test a case where the parametric assumptions made by the causal methods do not hold, but the results were not very informative. We simply note that the nonparametric SVR-RBF and RVR-RBF methods performed best, as we might expect, but made slightly more errors when children of T were manipulated.

5. Conclusions

The conditions from section 3, confirmed experimentally in section 4, may help to explain the surprising results from the recent causality challenge. There is a tradeoff between gains resulting from using the correct causal set of predictors and losses resulting from overfitting and sample bias as well as when parametric assumptions made by causal discovery algorithms do not hold. While the results given in section 4 indicate that there are cases where we should expect causal discovery based methods for prediction to strongly outperform prediction methods which do not account for causality, these cases, where many variables of a target node are manipulated, may not arise frequently in practice, and the exact parametric forms used when generating the data in the experiments may not be reflective of real world data. Thus, in practice, we may see greater performance when nonparametric methods and methods which make less restrictive assumptions about parametric forms and combat overfitting well that ignore causality are used, since even though these methods may make errors related to causality, i.e. not accounting for changes in a distribution after variables are manipulated, these errors may be small when compared to errors resulting from overfitting and sampling bias and when parametric assumptions do not reflect the data when causal discovery algorithms are used.

One possibility for achieving accurate results while still accounting for causality is to use methods which perform well in the prediction scenario with only the causally correct set of variables for each particular setting where certain variables are manipulated, e.g. retrain support vector machines using different sets of “causal” features for each prediction setting where manipulations vary. While using more sophisticated methods for prediction with such causal features may certainly produce models that are less likely to make errors related to sampling bias and overfitting, this still may not overcome problems resulting from assuming a parametric form which does not fit the data. In the experiments in section 4, we assumed that we were able to learn the correct Bayesian network for the data using a causal discovery algorithm, since the variables were linear Gaussians. However, if the data are very nonlinear then the DAG learned may be far from the truth. Thus, it would make more sense to use the features selected by a nonparametric method which does not account for causality, since the causally relevant set of variables for a particular setting where variables are manipulated that is indicated by the DAG may remove important variables and include problematic variables due to errors made by the causal discovery algorithm when a parametric form which does not fit the data is assumed. Fortunately, there has been much recent work in developing causal discovery algorithms which make less restrictive assumptions about the parametric forms, i.e. Shimizu et al. (2006), Hoyer et al. (2008), Hoyer et al. (2009). This may indeed become a possibility for obtaining accurate predictions that are sensitive to changes in a distribution when variables are manipulated in the future.

There are also many other factors which can affect prediction in these contexts. We merely highlighted a few factors relevant for the causality challenge. In particular, we considered only structural or perfect manipulations. In practice, manipulations may not completely break edges into a manipulated node and instead only change the conditional distribution of the node, and may affect other variables as well. We also have not considered the effects of unobserved variables which are causes of more than one of the observed variables on predictions or how well the children of a target variable predict the target compared to other variables in the Markov blanket. A more thorough investigation which considers some of these factors and uses more realistic data for testing may provide a more complete understanding of when causality is useful for making predictions.

Acknowledgments

We thank several anonymous reviewers for helpful comments and suggestions. R.E.T. was supported by the James S. McDonnell Foundation Causal Learning Collaborative Initiative.

References

P. O. Hoyer, A. Hyvärinen, R. Scheines, P. Spirtes, J. Ramsey, G. Lacerda, and S. Shimizu. Causal discovery of linear acyclic models with arbitrary distributions. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*. AUAI Press, 2008.

P. O. Hoyer, D. Janzing, J. Mooij, J. Peters, and B. Schölkopf. Nonlinear causal discovery with additive noise models. In *Advances in Neural Information Processing Systems*, volume 21. MIT Press, 2009.

D. Koller and N. Friedman. *Structured Probabilistic Models: Principles and Techniques*. Draft Textbook, 2008.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kauffmann Publishers, 1988.

S. Shimizu, P. O. Hoyer, A. Hyvärinen, and A. Kerminen. A linear non-gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, 7:2003–2030, 2006.

A. Smola and B. Schölkopf. A tutorial on support vector regression. Technical Report NC2-TR-1998-030, NeuroCOLT, 1998.

P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.

R. Tibshirani. Regression shrinkage and selction via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.

M. E. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.