Sufficient Condition for Pooling Data from different Distributions

Frederick Eberhardt FDE@CMU.EDU
Department of Philosophy, Carnegie Mellon University, Pittsburgh, PA 15213, USA

Abstract

We consider the problems arising from using sequences of experiments to discover the causal structure among a set of variables, none of whom are known ahead of time to be an "outcome". In particular, we present various approaches to resolve conflicts in the experimental results arising from sampling variability in the experiments. We provide a sufficient condition that allows for pooling of data from experiments with different joint distributions over the variables. Satisfaction of the condition allows for more powerful independence tests that may resolve some of the conflicts in the experimental results. The pooling condition has its own problems, but should – due to its generality – be informative to techniques for meta-analysis.

1. Introduction

Knowledge of causal structure enables predictions in cases where the system under consideration has been subject to interventions. Discovery of causal structure can proceed in two ways: Inference to causal structure – as far as is possible – from the passive observation of the variables, or active search of causal structure using interventions that specify a particular distribution for a subset of the variables.

There is a vast literature on causal discovery using passive observational data. Two main approaches can be distinguished: On the one hand there are a variety of Bayesian approaches which start with a prior over causal structures. The likelihood of the (passive observational) data given each causal structure is computed and multiplied with the prior over structures to form a posterior. The most likely graph is then taken to be the one with the highest posterior probability. For large numbers of variables there is an enormous number of possible graphs, so generally some computational short cuts are necessary. These can be in the form of (i) tricks that avoid computing the posterior for the entire set of possible graphs (Chickering, 2002), (ii) assuming structural constraints (maximum degree of nodes or other sparsity assumptions), or (iii) applying hierarchical Bayes methods (Mansinghka et al., 2006). On the other hand are constraint based methods that search for causal structure by sequentially testing for independence relations that hold in the data and using the results to constrain the search space (Spirtes et al., 2000). There are convergence results that guarantee the consistency of these algorithms, i.e. that guarantee that the algorithms recover as much information about the causal structure as is possible when the conditional independence relations true in the population are known. Similar asymptotic results are only known about the GES Algorithm for the Bayesian approach (Chickering, 2002).

Both of these approaches are limited to discovering an equivalence class of causal graphs, known as the Markov equivalence class, in which each graph implies the same (conditional) independence constraints for the data. In order to *uniquely* determine the causal structure, interventions are required. An intervention, in our parlance, is a randomization of one variable, which means that the values of the variable are determined by a distribution exogenous to any variable in the system. The advantage of interventions, first recognized by Fisher (1935), is that (i) interventions break any confounding due to (unmeasured) common causes, (ii) interventions can determine causal direction (if A causes B and the intervention is on A, then A and B will appear correlated, whereas if the intervention is on B, A and B will appear independent), and (iii) interventions provide a reference distribution over the intervened variable that allows for further statistical analysis (e.g. the estimation of strength parameters of the causal influence).

We assume general familiarity with the framework of causal Bayes nets (Pearl, 2000; Spirtes et al., 2000).

Accepted at *ERROR Conference*, Virginia Tech, Blacksburg, VA, June 2006.

Causal Bayes nets are directed acyclic graphs that represent the conditional independence relations implied by a causal structure. The causal Markov and faithfulness assumption link the graph with probability distributions over the variables. The causal Markov assumption states that every variable is conditionally independent of its non-descendents given its graphical parents, and faithfulness is the assumption that the graph represents all the conditional independence relations true in the population. In the following discussion we assume in addition to acyclicity, faithfulness and causal Markov, that the set of variables is causally sufficient, i.e. that there are no latent variables.

Interventions on variables in a Bayes net $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ are represented by policy variables. Although the notion of policy variables can be generalized, we will here assume that there is a policy variable I_X for each variable $X \in \mathbf{V}$. Each policy variable has a single arrow into the variable whose intervention it specifies ($I_X \rightarrow X$) and is exogenous to the variables under consideration, i.e. is – for all intents and purposes here – uncaused. A policy variable has two states, 0 and 1. If $I_X = 0$, the passive observational distribution over X obtains, and X is dependent on its normal causes, its graphical parents $pa(X)$, i.e. $P(X|pa(X), I_X = 0) = P(X|pa(X))$. If $I_X = 1$, then X is subject to an intervention. In that case the distribution over X is determined entirely by I_X. This is often referred to as a "hard" or "surgical" intervention, since the intervention breaks the dependency of X on its normal causes, i.e. $P(X|pa(X), I_X = 1) = P(X|I_X = 1)$.

This form of "edge-breaking" intervention captures the notions of randomized trials as well as clamping (fixing a variable to a particular value), which is a degenerate form of a "surgical" intervention.[1] The policy variable represents the decision of whether or not a particular variable is subject to an intervention. Its non-zero state is associated with a particular distribution over the intervened variable. For the most part a policy variable behaves like any other variable, although we do not specify a distribution over the values of a policy variable. We refer to the subset of variables in $\mathbf{V}$ subject to an intervention in a given experiment (i.e. whose policy variable $I_X = 1$) as $\mathbf{I}$ and the corresponding set of policy variables as $\mathbf{Pol}$.

Even when using interventions of the type described above, there are causal graphs for which a single experiment involving a single intervention on one variable or multiple simultaneous interventions on a set of variables is insufficient to recover the complete causal structure among the variables. Sequences of experiments involving a combination of different interventions are needed. Eberhardt et al. (2006) show that $N - 1$ experiments are sufficient and in the worst case necessary to recover the causal structure among a set of N variables if only a single variable can be subject to an intervention in any one experiment. This bound can be reduced to $\lfloor \log_2(N) + 1 \rfloor$ if multiple simultaneous interventions are allowed (Eberhardt et al., 2005).

Only the combination of results from the $N - 1$ or $\lfloor \log_2(N) + 1 \rfloor$ different experiments allows the unique identification of the underlying causal structure. Consider an example with two variables X, Y: While independence of X and Y is identifiable by passive observational data, dependence underdetermines the causal structure, since it might be $X \rightarrow Y$ or $X \leftarrow Y$ (assuming causal sufficiency). However, an intervention on X would disambiguate the evidence, since we would find X and Y to be associated in case of the first structure, whereas we would fail to do so for the second, since the intervention destroys the correlation between X and Y. Similarly, for an intervention on Y. However, the combination of passive observational evidence with evidence from an experiment in which one of the pair of variables was subject to an intervention uniquely identifies the causal structure for that pair of variables. We refer to this as a combination of a structural adjacency test (since passive observation tests for adjacencies) and a structural direction test (since the intervention determines causal direction). Similarly, the combination of evidence from an experiment in which X was subject to an intervention and Y was passively observed, with evidence from a further experiment in which Y was subject to an intervention and X was passively observed, will uniquely determine the causal structure among X and Y. If X and Y are non-adjacent, both experiments will show independence, if $X \rightarrow Y$, we will only find independence in the experiment that intervenes on Y, and if $X \leftarrow Y$, then only the intervention on X will return independence. We refer to this set-up as a combination of two opposing structural direction tests.

More formally, let $< E >_n$ be a sequence of experiments on a set of variables $\mathbf{V}$. Each experiment E_i consists of a set of variables $\mathbf{I_i} \subseteq \mathbf{V}$ that are subject to an intervention, and a set of variables $\mathbf{U_i} \subseteq \mathbf{V}$ that are passively observed. $\mathbf{U_i} \cup \mathbf{I_i} = \mathbf{V}$ and $\mathbf{U_i} \cap \mathbf{I_i} = \emptyset$. Any experiment E_i is a *structural adjacency test* with respect to a pair of variables X, Y if both X and Y are passively observed, i.e. $X, Y \in \mathbf{U_i}$. E_i is a *structural X-direction* test, if $X \in \mathbf{I_i}$ and $Y \in \mathbf{U_i}$; E_i is a

[1] Note, that the notion of an intervention can be generalized to weaker interventions which do not break the influence of the other causes of X, or I_X might have many states which specify different distributions over X.

structural Y-direction test, if $Y \in \mathbf{I_i}$ and $X \in \mathbf{U_i}$.

As indicated above and discussed in more detail in Eberhardt et al. (2005), if we restrict our search algorithm to qualitative features of the data (independence relations), then we need either one structural adjacency test and one structural direction test, or two opposing structural direction tests to uniquely determine the causal structure among a pair of variables.

The bounds shown in Eberhardt et al. (2005) are implied by combinatorial constraints that result if one wants to subject each pair of variables to either one of these combination of structural tests. The bounds do not take into account the statistical issues that errors might occur due to statistical fluctuations.

2. The Problem

In any realistic experimental setting, conditional independence tests will be subject to errors. In search for the causal structure among a set of variables the absence of a particular causal arrow between X and Y is determined by the existence of *some* conditioning set that makes the two variables independent. However, the presence of an edge implies that the two variables in question remain dependent for *all* possible conditioning sets. In fact, not all possible conditioning sets need to be tested if the graph is sparse, but there does remain an asymmetry in the test requirements for the two possibilities. In large dense networks the search may require a very large number of conditioning sets to determine adjacency. Consequently, the likelihood of all independence tests returning the correct result descreases as the number of tests increases. This is exaggerated by the fact that the available number of data points for a particular independence test gets smaller as the conditioning sets increase. Thus, we can expect to obtain conflicting results from different experiments about a particular pair of variables, since some experiments might return the pair as adjacent, while others do not. Of course, if $X \rightarrow Y$ is in the true graph, we expect X and Y to be independent in an experiment where Y is subject to an intervention, whereas we expect dependence for all conditioning sets if the intervention is on X. The combination of these experimental results does not amount to a conflict, since the results are coherent with one true generating structure. We have a *conflict* if the results from different experiments are inconsistent with any causal structure that – appropriately manipulated given the interventions of the specific experiment – is assumed to generate the data in both cases.

The simplest such conflict occurs if we have two experiments E_1 and E_2 on the same set of variables $\mathbf{V}$. We assume that the same (pre-manipulation) causal structure underlies both experiments and that we would – if the experiments involved no interventions – observe the same joint distributions over the set of variables. Let $X, Y \in \mathbf{V}$ both be passively observed in both experiments, i.e. both experiments are structural adjacency tests with respect to this pair of variables. Suppose X and Y are independent for some conditioning set C in E_1, indicating that neither variable is a direct cause of the other. Further assume that X and Y are dependent for all conditioning sets in E_2, suggesting that either $X \rightarrow Y$ or $X \leftarrow Y$. We must conclude that at least one of the independence tests must have returned an erroneous result, because their implications are not consistent. E_1 says there is no edge, while E_2 says there is one. If we assume we assume that the same underlying causal structure generated the data in both cases, then the results are inconsistent..

Since the bounds described in Eberhardt et al. (2005) optimize the combinatorics of structural direction and adjacency tests, conflicts can arise from a variety of situations: First, it is possible that a pair of variables is subject to the same structural test repeatedly in a sequence of experiments. With statistical errors, the results for the same test might not be consistent across experiments. Second, conflicts may arise because results from different structural tests cannot be combined coherently.

Let X and Y refer to some pair of variables in $\mathbf{V}$ and let E_i refer to some experiment in the sequence of experiments, different indices indicate different experiments. Conflicts are given in the following situations:

1. X and Y are passively observed in E_k and E_l, but E_k indicates they are non-adjacent, whereas E_l indicates they are adjacent.
2. X is randomized in E_k and found to be adjacent to Y (in fact it would be a direct cause of Y), but in E_l both variables are passively observed and found to be non-adjacent.
3. X is randomized in E_k and found to be a direct cause of Y and Y is randomized in E_l and found to be a direct cause of X (a conflict, since we assume acyclicity).
4. X and Y are passively observed in E_k and found to be adjacent, but X is randomized in E_l and not found to be adjacent to Y and Y is randomized in E_m and not found to be adjacent to X.

Given these types of conflicts, under what circumstances is it possible to resolve them? Of course one could re-run one of the experiments, possibly with a larger sample size, pool the data from the original and the repeated experiment and perform the crucial independence tests again, now with a greater power. But additional experiments may be expensive to perform.

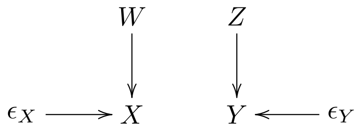
In order to recover the causal structure uniquely, we have to perform a sequence of experiments anyway. Is it possible to utilize information gained in the other experiments to resolve our conflicts?

The concern is that we cannot simply pool the data relevant to a particular independence test from two experiments, because different experiments in a sequence have different joint distributions over the variables due to the different interventions. If different variables are subject to interventions, this implies different manipulated graphs over the variables, representing the different joint distributions. Pooling data from different distributions may lead to spurious changes in correlations.

Consider the following examples.

2.0.1. Example 1: Independence to Dependence

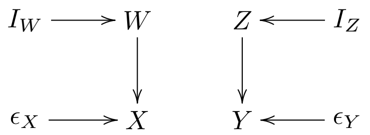
Suppose we have the following linear structural equation model with gaussian error terms.



$$\begin{aligned}
W &\sim N(0, 2.25)\\
Z &\sim N(0, 5.29)\\
\epsilon_X &\sim N(0, 3.24)\\
\epsilon_Y &\sim N(0, 6.25)\\
X &= 2W + \epsilon_X\\
Y &= 3Z + \epsilon_Y
\end{aligned}$$

Under passive observation of X and Y we obtain the data shown in Figure 1. X and Y are found to be independent with high significance on any standard independence test – as we would expect from the causal structure.

Now consider the same causal structure, but where simultaneous and independent interventions impose distributions on W and Z that are different to their passive observational distribution, i.e. the following causal structure:



$$\begin{aligned}
W_i | I_W = 1 &\sim N(3, 1)\\
Z_i | I_Z = 1 &\sim N(-2, 1)\\
\epsilon_X &\sim N(0, 3.24)\\
\epsilon_Y &\sim N(0, 6.25)\\
X &= 2W + \epsilon_X\\
Y &= 3Z + \epsilon_Y
\end{aligned}$$

In this case we obtain the data over X and Y shown in Figure 2. Again, as expected from the causal structure, X and Y are independent and the result is significant.

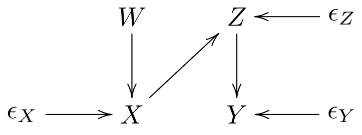
However, when we pool the two distributions, X and Y are no longer independent, as can be seen in Figure 3. The gradient of the regression of X on Y is -0.5 with a p-value smaller than 0.001.

This is a very simple case where two samples in which X and Y are independent in each sample (as they should be in accordance with the causal structure), become dependent when the samples are pooled. The interventions were not even on the variables themselves. Of course, this particular case is not problematic, since in the case of causal discovery we know the separate samples before and could normalize the data from each sample before pooling and then we would find X and Y to be independent. But normalizing is only possible when the distribution over the variables in the intervened case is of the same type as in the passive observational case. If the intervention distributions on W and Z had been, say, χ^2 distributions, then such normalizing and pooling would not be possible, similarly, if the interventions had been a restriction in the range of values that W and Z could take.

2.0.2. Example 2: Dependence to Independence

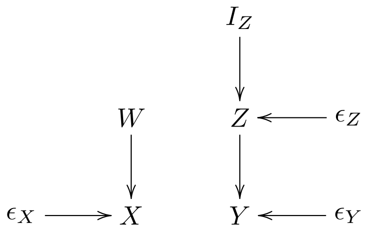
In the second example we consider a case where dependence in one experiment might be washed out to render the two variables independent, when pooled with a separate sample.

Suppose the true causal structure corresponds to the following linear structural equation model with gaussian error terms:



$$
\begin{aligned}
W &\sim N(0, 1.9)\\
\epsilon_X &\sim N(0, 0.49)\\
\epsilon_Z &\sim N(0, 4.84)\\
\epsilon_Y &\sim N(0, 1.21)\\
X &= 2W + \epsilon_X\\
Z &= 0.1X + \epsilon_Z\\
Y &= 3Z + \epsilon_Y
\end{aligned}
$$

A plot of 1,000 samples of X and Y are shown in Figure 4. The main point to note is that X and Y are found to be dependent with high significance. If we now perform an experiment in which we intervene to set the distribution on Z, then we break the arrow from X to Z and have the following causal structure:



$$
\begin{aligned}
W &\sim N(0, 1.9)\\
\epsilon_X &\sim N(0, 0.49)\\
\epsilon_Z &\sim N(0, 4.84)\\
\epsilon_Y &\sim N(0, 1.21)\\
X &= 2W + \epsilon_X\\
Z_i|I_Z = 1 &= \epsilon_Z\\
Y &= 3Z_i + \epsilon_Y
\end{aligned}
$$

A plot of 1,000 samples of X and Y from the manipulated distribution are shown in Figure 5. As expected, X and Y are independent.

However, if we pool the data from the two experiments (Figure 6), then we find that X and Y are independent. It is not clear what the normative solution is in this case since we are mixing distributions, one where X and Y are dependent and one where they are independent. But what we find is that pooling the samples removed the dependence in the first sample. This example can be made more extreme with a graph where there are to causal connections between X and Y which are both strong, but almost cancel each other out[2]. If there are two experiments, each of which blocks one of the paths, then the pooled result might return independence (since the total effect is too weak to register in the total sample size), even though there is a strong correlation in each subsample.

The two examples illustrate that pooling samples from different distributions can lead to spurious changes in the correlation between two variables, and the change can be in either direction. In light of these examples, how can we resolve any conflicts?

3. Solutions

3.1. Voting

The first solution that comes to mind to resolve conflicts without re-doing one of the experiments is some type of voting procedure. Given a sequence of experiments one could select those experiments that are informative about a particular pair of variables X and Y, i.e. those experiments that do not simultaneously intervene on both X and Y. Among these experiments a simple vote decides whether $X \rightarrow Y$, $X \leftarrow Y$ or X and Y are non-adjacent. However, it is not that simple: Although there are three possibilities of what might be going on between a pair of variables, the structural tests are only binary. A structural direction test can decide whether there is an edge from the intervened variable to the other variable, but cannot distinguish between an edge incedent on the intervened variable and no edge at all. Similarly, a structural adjacency test can tell wether there is an edge at all, but is unable to distinguish directions. It seems therefore, that the vote of a particular test should be evenly split between the options it cannot distinguish. That is, if after an intervention on X we find X and Y to be independent, then both non-adjacency and $X \leftarrow Y$ should receive half a vote each.

But even now, the approach does not take into account that votes from different experiments are votes from different joint distributions, which may make the discovery of a particular (in)dependence harder or easier. In order to reflect the importance of the result from any particular experiment, the votes could be weighted by the p-values of the independence tests they represent. But now we run into trouble with the asymmetry of the search procedure: In order to discover *non-adjacency* we have to find *one* conditioning set that makes the two variables independent. The PC-

[2] As long as the two causal connections do not exactly cancel each other out, this would not amount to a violation of faithfulness.

algorithm iterates through the independence tests in order of complexity (size of the conditioning set), so that there always is a well-defined independence test that determines non-adjacency. The p-value from this test could be used to determine the weight of the vote from this particular experiment. However, *adjacency* is established if there is *no* conditioning set that makes the variables independent, i.e. *all* the independence tests fail. Consequently there is no unique p-value that could be used to weight the vote. There is no guarantee that there is a corresponding independence test in each experiment so that we could reduce the conflict to a set of independence tests. And even if there were, then we are aggregating votes weighted by p-values from different distributions. It is not clear what the justification for such a procedure may be, since different distributions may yield p-values that distort the vote.

Quite apart from the above matters, issues of judgment aggregation arise. Since the combination of independence relations imply other (in particular, higher order) independence relations the outcome of a voting procedure might depend on how votes are aggregated and it is not clear at all, how an aggregation procedure here would have to be designed to be in some sense “truth tracking”, i.e. that we could have any hope that using some voting method will get us closer to the true graph.

The bottom line is that voting may well work as a useful heuristic to resolve conflicts, but the worry is that the ad hoc decisions made in order to have a well-defined voting procedure destroy the consistency guarantees of the overall search algorithm.

3.2. Bayesian Approach

For a Bayesian, conflicts of the type described above do not arise. A strictly Bayesian approach would place a prior over all possible structures and all possible parameterizations of those structures. Given the set of variables subject to an intervention in the first experiment, one would compute all the post manipulation graphs for all the structures and compute the likelihood of the data obtained from the experiment given each manipulated graph. The likelihood is then multiplied with the prior, where the prior probability for each manipulated structure for this experiment is specified by the prior of the corresponding unmanipulated structure. The variables specifying the parameterizations are integrated out to yield a posterior distribution over all possible (manipulated) structures. This posterior can then be used as a prior over all possible unmanipulated structures for the next experiment. Conflicts do not arise explicitly but are taken care of implicitly in the likelihood and updating procedure.

This Bayesian approach – if fully implemented – would preserve the consistency results of the search algorithm. However, the computation cost is enormous: Even for six variables, there are between 2^{15} and 3^{15} possible structures (three possibilities for each edge, but excluding all cyclic structures). So, even if the integrals were simple to compute, there would be a huge number of them for large graphs. It is insufficient to keep track of only the most likely graph, since there are cases where only particular combinations of interventions render the true graph the most likely. Until these have been performed, it not clear why the true graph would be the most likely graph. In addition, in general the prior over the structures will not be simple, since it will be the posterior of the previous experiment(s), for which there is no reason to think that it will be simple.

Computational limitations will restrict the feasibility of this approach to toy problems. Alternatively, a variety of assumptions have to be added to simplify the calculations (e.g. particle filtering, hierarchical methods etc.). But at least it is a solution.

4. A sufficient Condition for a Solution

The key difficulty is to account for the differences of the joint distributions due to the different interventions, and to figure out how and when these differences affect the independence tests relevant to a conflict. Failure to take these differenes into account can lead to spurious correlations or independencies when data is pooled. However, if we can ensure that the distribution *relevant* to the conditional independence test in question is the same in both experiments, then the data can be pooled to obtain an independence test with more power.

For example, suppose the variables, say X and Y, whose independence is in question, are graphically disconnected, i.e. causally separate, from the other variables $W_1, \ldots, W_n$ in the causal structure. If there are two experiments, one which is an intervention on W_i and another with an intervention on W_j, then clearly the changes in the interventions will have no effect on the marginal distribution over X and Y and the data from the experiment can be pooled for the independence tests on X and Y. This is a very strong condition to ensure the validity of pooling, but we show that it can be weakened.

In the following we provide a sufficient condition whose satisfaction allows for pooling of data from different joint distributions.

Let $< E >_n = E_1, E_2 \dots E_n$ be a sequence of experiments on the set of variables $\mathbf{V}$. Each experiment is represented by a triple of sets $E_i = (\mathbf{I_i}, \mathbf{U_i}, \mathbf{Pol_i})$, where $\mathbf{I_i}$ represents the subset of $\mathbf{V}$ that is subject to an intervention in E_i, $\mathbf{Pol_i}$ is the corresponding set of policy variables, and $\mathbf{U_i}$ contains the remaining passively observed variables.

Suppose that the pair of variables $X, Y \in \mathbf{V}$ is subject to a conflict: There is some experiment E_i which renders X and Y adjacent, whereas some experiment E_j renders X and Y non-adjacenct. Hence, there is some conditioning set $\mathbf{C}$ such that X and Y are independent conditional on $\mathbf{C}$ in E_j. Let this independence test be $T_{X,Y|\mathbf{C}}$. Let $\mathbf{M_{i,j}} \subseteq \mathbf{V}$ be the set of variables that is subject to an intervention in E_i, but not in E_j, i.e. $\mathbf{M_{i,j}} = \mathbf{I_i} \setminus \mathbf{I_j}$, and let $\mathbf{Pol_{M_{i,j}}}$ be the set of policy variables corresponding to $\mathbf{M_{i,j}}$. Similarly for $\mathbf{M_{j,i}}$ and $\mathbf{Pol_{M_{j,i}}}$.

In addition, there may be some variables $Z \in \mathbf{V}$ that are subject to an intervention in both experiments, but the distribution imposed on Z by the interventions is different in each case, i.e policy variable $I_{Z_i} \neq I_{Z_j}$. Let $\mathbf{Pol_{N_i,N_j}}$ contain all the policy variables from E_i corresponding to such variables Z. Similarly for $\mathbf{Pol_{N_j,N_i}}$.

Let $\mathbf{S_i} = \mathbf{Pol_{M_{i,j}}} \cup \mathbf{Pol_{N_i,N_j}}$. Similarly for $\mathbf{S_j}$. $\mathbf{S_i}$ and $\mathbf{S_j}$ contain all the policy variables that *change* between the two experiments E_i and E_j. The basic idea is that we only have to worry about those interventions that change across experiments, since they are the only changes that differentiate the joint distributions. These changing interventons are contained in $\mathbf{S_i}$ and $\mathbf{S_j}$.

If we could ensure that the distribution relevant to a particular independence test is invariant to the differences in the joint distributions in the two experiments, then we could pool the data and obtain more powerful independence tests. These would be more powerful tests of the adjacency relations between variables and could be used to resolve conflicts. The following theorem specifies a sufficient condition for this invariance.

Theorem 4.1 *If the set of variables $\{X, Y\}$ is d-separated from the set of changing policy variables $\mathbf{S_i}$ given the conditioning set $\mathbf{C}$ in experiment E_i and if the set of variables $\{X, Y\}$ is d-separated from the set of changing policy variables $\mathbf{S_j}$ given the conditioning set $\mathbf{C}$ in experiment E_j, then the distributions relevant for independence test $T_{X,Y|C}$ are invariant across experiments E_i and E_j and the data relevant to the test can be pooled.*

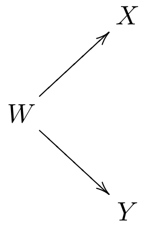
Proof: Let $P_i(\mathbf{V})$ be the distribution over variables $\mathbf{V}$ in experiment E_i, while $P(\mathbf{V})$ is the passive observational distribution. If $\{X, Y\}$ is d-separated from the changing policy variables $\mathbf{S_i}$ given $\mathbf{C}$, the joint distribution simply factorizes:

$$\begin{aligned} P_i(X, Y|\mathbf{C}) &= \sum_{\mathbf{S_i}} P_i(X, Y, \mathbf{S_i}|\mathbf{C}) \\ &= \sum_{\mathbf{S_i}} P(X, Y|\mathbf{C}) P_i(\mathbf{S_i}|\mathbf{C}) \\ &= P(X, Y|\mathbf{C}) \sum_{\mathbf{S_i}} P_i(\mathbf{S_i}|\mathbf{C}) \\ &= P(X, Y|\mathbf{C}) \end{aligned}$$

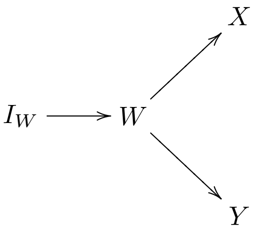
Similarly for $P_j(X, Y|\mathbf{C})$. The joint (conditional) distribution over $X, Y|\mathbf{C}$ is invariant to the changing interventions in E_i and E_j. It follows as a trivial consequence that the marginals $P(X|\mathbf{C})$ and $P(Y|\mathbf{C})$ are also invariant. Consequently, the independence test $T_{X,Y|\mathbf{C}}$ is invariant to the distributions of both experiments and the data relevant to this test can be pooled. □

We have specified a sufficient condition that allows for pooling of data. A simple example will illustrate the main claim.

Suppose we have the following true graph:



We have two experiments. One is passive observational (above), one is an intervention on W, i.e.



The theorem says that we cannot pool for the (unconditional) test $T_{X,Y}$ (whether X is independent of Y), because I_W is not d-separated from $\{X, Y\}$ in the second experiment. However, we can pool for the test $T_{X,Y|W}$ (whether X and Y are independent conditional on W).

The theorem does not specify a *necessary* condition, since the intervention distributions can be tweaked in such ways as to preserve the invariance properties of the distributions relevant to the independence test even if the d-separation condition is not satisfied. Trivially, this can be done if the intervention distribution of a variable is essentially the same as the passive observational distribution for that variable.

5. Discussion

While the above theorem specifies a sufficient condition for pooling which might resolve some conflicts in sequences of experiments, it requires substantial knowledge about the causal structure. Whether or not we can pool for a particular independence test depends on whether the changing interventions are d-separated from the changing interventions and – more importantly – whether we *know* that this d-separation condition is satisfied. But we are trying to *discover* the causal structure in the first place. Only in very sparse graphs, or if only very few conflicts occur in our sequence of experiments will we be able to know whether the condition is satisfied. Furthermore, it requires the search algorithm to store information about which independence test determined a non-adjacency in each experiment, so that the problematic test can be identified for possible conflict resolution afterwards (or one has to find it again). The result therefore, does not provide a simple solution for sequences of experiments.

However, the result applies generally and is not specific to particular families of distributions. It is therefore directly relevant to techniques in meta-analysis. In particular, if the d-separation condition specified in the theorem is known to fail, then there is particular reason for concern if unexpected correlations or independencies occur when data is pooled. If the d-separation relation is known to hold, then there is a possibility of obtaining results with a higher significance even for cases where joint distributions are known to be different or where there are correlated interventions.

The description of possible conflicts is not exhaustive. We list all the conflicts that occur if the search algorithm only determines adjacencies in each experiment. However, there are well known cases where the direction can be determined as well: First, if an adjacency is found between an intervened variable and any other variable in the system, then we know that the direction of the arrow is out of the intervened variable. Second, if there are three passively observed variables X, Y, Z and (i) X and Y are dependent, (ii) Y and Z are dependent, (iii) X and Z are independent, and (iv) X and Z are dependent conditional on Y, then we know that Y is a common effect of X and Z, i.e. $X \rightarrow Y$ and $Z \rightarrow Y$. This structure is often referred to as an "unshielded collider" and is identifiable in passive observational data. If these techniques to determine direction were included in the structure search algorithm, a variety of further conflicts may arise pertaining to directional information. It is not known whether this addition would result in faster structure search or more conflicted results when sampling errors are taken into account.

6. Conclusion

We have provided a sufficient condition for pooling data from different joint distributions when variables have been subject to interventions. The result is almost trivial once the semantics of the policy variables has been made explicit. In that sense, this work is more a contribution to developing a clear understanding of policy variables, how they are represented in causal Bayes nets and how they affect the joint distribution over the variables in question. The result vidicates the representation of policy variables as distinct variables with particular structural constraints in the graph, since many (though not all) of the properties for ordinary Bayes net variables carry over. However, the semantics presented here rejects the notion that policy variables are just placeholders for fixing a distribution over the intervened variable. They are better understood as decision points for input exogenous to the system of variables under consideration. The question we do not address is whether or not there should be a distribution over the values of the policy variable. Whether this would be sensible is really a framing question, since the answer depends on the reference frame we use for the discovery problem and on how the decisions for or against interventions are made.

The key problem with regard to the application of the central pooling theorem to structure search is that in many cases we will not know whether the d-separation condition is satisfied or not. Nevertheless, it specifies a clear and fairly general condition for meta-analysis, since it is intervention and distribution independent. Perhaps more importantly to meta-analysis, however, is that the use of this theorem is not restricted to independence tests. Since it guarantees the invariance of the marginal/conditiona distributions, these can also be used for parameter estimation. This is useful in meta-analyses where the causal structure is (largely) known, but where more data is needed to perform accurate parameter estimation. Here satisfaction of the d-separation condition can be checked easily and data pooled accordingly.

Acknowledgements

I am extremely grateful for the many discussions I had with my advisors Richard Scheines and Clark Glymour. I would also like to thank Joseph Ramsey for making me worry about this problem, and Peter Spirtes and Jiji Zhang for helpful comments.

References

Chickering, D. M. (2002). Optimal structure identification with greedy search. *Journal of Machine Learning Research*, *3*, 507–554.

Eberhardt, F., Glymour, C., & Scheines, R. (2005). On the number of experiments sufficient and in the worst case necessary to identify all causal relations among n variables. *Proceedings of the 21 st Conference on Uncertainty and Artificial Intelligence* (pp. 178–184). AUAI Press, Corvallis, Oregon.

Eberhardt, F., Glymour, C., & Scheines, R. (2006). N-1 experiments suffice to determine the causal relations among n variables. In D. E. Holmes and L. C. Jain (Eds.), *Innovations in machine learning*, vol. 194 of *Theory and Applications Series: Studies in Fuzziness and Soft Computing*. Springer-Verlag.

Fisher, R. (1935). *The design of experiments*. Hafner.

Mansinghka, V. K., Kemp, C., Tenenbaum, J. B., & Griffiths, T. L. (2006). Structured priors for structure learning.

Murphy, K. P. (2001). *Active learning of causal bayes net structure* (Technical Report). Department of Computer Science, U.C. Berkeley.

Pearl, J. (2000). *Causality*. Oxford University Press.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction and search*. MIT Press. 2 edition.

Tong, S., & Koller, D. (2001). Active learning for structure in bayesian networks. *Proceedings of the International Joint Conference on Artificial Intelligence*.

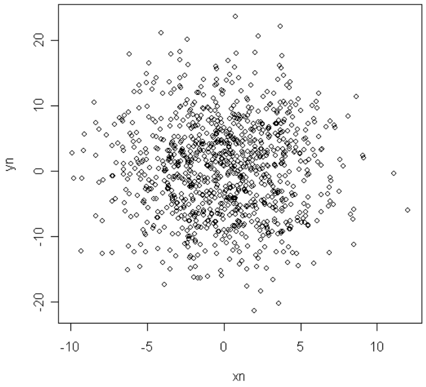


Figure 1. A plot of 1,000 samples of X and Y from the linear structural equation model with gaussian error terms under passive observation in Example 1. As expected given the causal structure, X and Y appear independent.

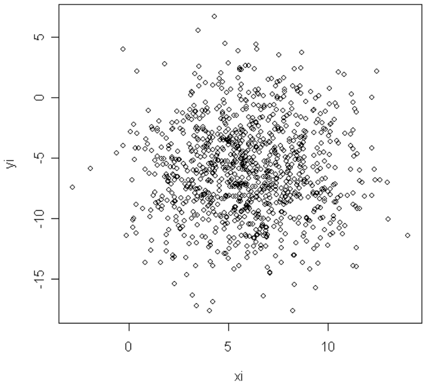


Figure 2. A plot of 1,000 samples of X and Y from the linear structural equation model with gaussian error terms with interventions on W and Z in Example 1. Again, as expected given the manipulated causal structure, X and Y appear independent.

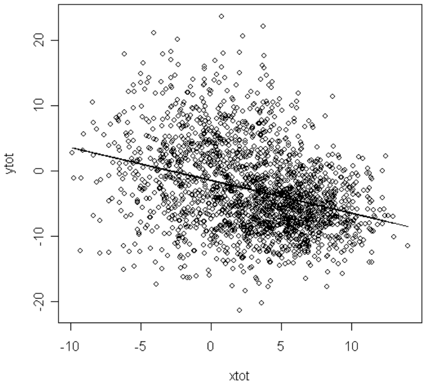


Figure 3. Samples of X and Y from the passive observational distribution mixed with samples from the distribution, where W and Z were subject to an intervention shown with a fitted regression line. X and Y no longer appear independent.

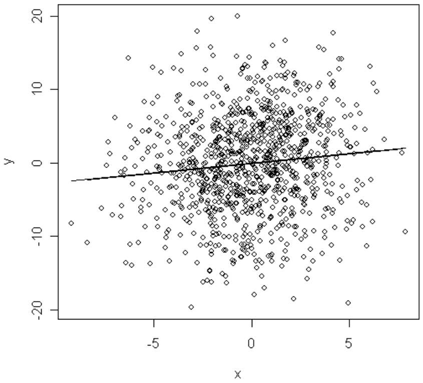


Figure 4. Plot of 1,000 samples of X and Y drawn from the passively observed causal structure in Example 2. As expected, X and Y are found to be dependent with a high significance ($p < 0.001$). The regression line is shown, the slope is slight, since the $X \rightarrow Z$ connection is weak, but the significance is what matters here.

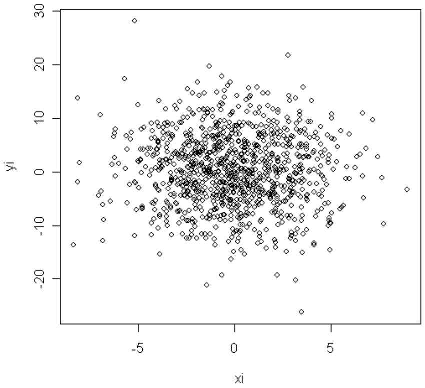


Figure 5. Plot of 1,000 samples of X and Y drawn from the manipulated distribution where Z was subject to an intervention in Example 2. As expected, X and Y are independent.

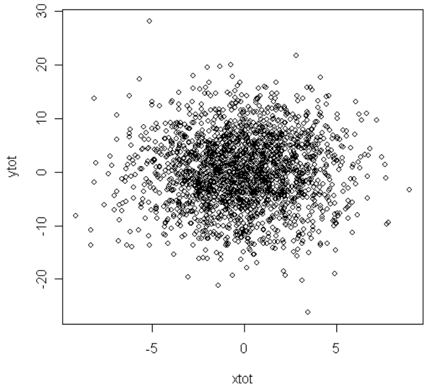


Figure 6. Samples from both the passively observed and manipulated distributions are pooled. X and Y are independent.