Combining Experiments to Discover Linear Cyclic Models with Latent Variables

Frederick Eberhardt
Department of Philosophy
Washington University in St Louis

Patrik O. Hoyer
CSAIL, MIT
& HIIT, Univ. of Helsinki

Richard Scheines
Department of Philosophy
Carnegie Mellon University

Abstract

We present an algorithm to infer causal relations between a set of measured variables on the basis of experiments on these variables. The algorithm assumes that the causal relations are linear, but is otherwise completely general: It provides consistent estimates when the true causal structure contains feedback loops and latent variables, while the experiments can involve surgical or 'soft' interventions on one or multiple variables at a time. The algorithm is 'online' in the sense that it combines the results from any set of available experiments, can incorporate background knowledge and resolves conflicts that arise from combining results from different experiments. In addition we provide a necessary and sufficient condition that (i) determines when the algorithm can uniquely return the true graph, and (ii) can be used to select the next best experiment until this condition is satisfied. We demonstrate the method by applying it to simulated data and the flow cytometry data of Sachs et al (2005).

1 INTRODUCTION

Causal knowledge is key to supporting inferences about the effects of interventions on a system, hence much of applied science is devoted to identifying causal relationships between various measured quantities. For this purpose the randomized controlled experiment is generally the tool of choice whenever such experiments are feasible.

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

While the standard theory of experimental design provides procedures and inference techniques for a given set of potential causes (treatment variables) and effects (outcomes), it does not provide guidance on how to determine the full set of causal relationships among the measured variables (i.e. the causal 'interaction graph'). Randomized experiments directly provide the effects of the intervened variables *in the experiment*, but breaking these effects down into direct and indirect effects (with respect to the measured variables), or combining them to determine the total effect, is typically not straightforward. This problem arises in particular when correlations in the data may be partly attributable to unknown confounding variables, or when the system of interest exhibits feedback phenomena, as for example is common in bioinformatics and economics. Feedback phenomena in particular invalidate standard approaches based on directed acyclic graphs. Furthermore, experiments can involve different types of manipulations (e.g. randomizing a single vs. multiple variables simultaneously per experiment), and sometimes it is only possible to perform 'soft' interventions, in which the variables influenced by the experimenter still (partly) depend on their natural causes. Because of difficulties such as these, there is a definite need for algorithms and procedures that (i) combine the (possibly conflicting) results from different types of experiments, (ii) incorporate background knowledge where available, (iii) provide guidance on the selection of experiments, and (iv) given a set of stated assumptions, are able to identify consistently and efficiently the causal structure from experimental data, or do as well as possible (with regard to some measure) in cases of underdetermination or conflicting evidence.

As a simple and concrete example, consider the two alternative graphs of Figure 1 as explanations of the causal relationships between the variables x_1, x_2, and x_3. If in three separate experiments we randomize each of the variables one at a time while measuring the two others, we could deduce from the resulting dependencies that there exist directed paths from x_1 to x_2, from

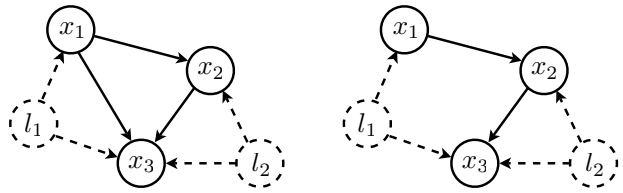

Figure 1: Two distinct causal graphs which cannot be distinguished based on (marginal and conditional) independencies alone, when using only a combination of observational data and experiments randomizing at most one variable per experiment.

x_1 to x_3, and from x_2 to x_3. However, based on dependencies and independencies alone it is impossible to determine whether there is a direct effect $x_1 \to x_3$ in addition to the indirect effect through x_2, because the latents l_1 and l_2 preclude us from using statistical conditioning to identify this feature. Note that we can identify all edges among the observed variables if we can in an experiment simultaneously randomize *both* x_1 and x_2 while measuring x_3. This holds generally: By simultaneously randomizing *all* variables except a target variable, it is possible to determine all direct causal effects with respect to the randomized variables. By performing such a 'leave-one-out' experiment for each variable, and combining the results, the complete graph can be determined. Unfortunately, it is seldom feasible or economical to perform such experiments.

There already exists a substantial body of work in this area. For instance, Cooper & Yoo (1999) use a Bayesian score to combine experimental and observational data in the acyclic no-hidden-variable case. Tong & Koller (2001), Murphy (2001), Eberhardt (2008) and He & Geng (2008) provide strategies for this setting to optimally select the experiments to learn as much as possible about the underlying structure. Furthermore, Eaton & Murphy (2007) describe how to handle 'uncertain' interventions where the targets of a given intervention is unknown, while Nyberg & Korb (2006) analyze the implications of soft interventions.

All of the above methods assume the restricted case of acyclic structures and no confounding latent variables. While Richardson (1996) discusses learning cyclic models based on observational data alone, there is little work that formally integrates experimental results into the search procedure. One exception to this is recent work by Schmidt and Murphy (2009) who adapt a formulation for undirected graphs to model cyclic directed graphs with no latents. In contrast, the standard literature on experimental design permits latent variables but only considers a very restricted set of possible causal structures.

In this paper we provide a procedure for discovery of *linear* causal models that uses experimental data and is completely general with regard to structural assumptions. Given a set of datasets over the variables of interest where in each dataset a different combination of variables has been subject to intervention, our procedure identifies the linear constraints on the causal effects among the variables. For an underdetermined system it returns a model that provides a minimal representation of the measured constraints. From here there are two ways to proceed: Either additional routines can be used to characterize the underdetermination, or, alternatively, a condition we provide identifies the next experiment that would add the most additional required constraints. Once the system is sufficiently constrained (available background knowledge can be used if available), the procedure identifies the direct causal effects relative to the measured variables, and additionally infers the presence and locations of confounding latent variables. If the system is overconstrained it returns the model with the minimum sum of squared errors to the constraints. The technique works for both cyclic and acyclic systems, and easily incorporates soft interventions.

2 MODEL

The data generating model can be represented as a set of structural equations that assign to each of the observed variables a linear combination of the other variables and an additive 'disturbance' term. Grouping the observed variables $x_i,\ i = 1, \ldots, n$ into the vector $\mathbf{x}$, the corresponding disturbances e_i into the vector $\mathbf{e}$, and the linear coefficients b_{ji} representing the direct effects $x_i \to x_j$ into the matrix $\mathbf{B}$, we have:

$$\mathbf{x} \ := \ \mathbf{B}\mathbf{x} + \mathbf{e}. \tag{1}$$

Here, latent variables are represented as non-zero off-diagonal entries in the covariance matrix $\mathbf{\Sigma_e} = E\{\mathbf{e}\mathbf{e}^T\}$ of the disturbances. Without loss of generality, all the x_i and the e_i have zero mean.

A surgical (edge-breaking, i.e. fully randomizing) intervention on a variable x_i is represented by a manipulated model where the row in $\mathbf{B}$ corresponding to x_i has been set to zero, indicating that x_i is no longer influenced by any of the other observed variables. Similarly, we set the corresponding disturbance e_i to zero, and instead assign an independent, unit variance experimental variable c_i to x_i. If several of the observed variables are surgically intervened on simultaneously (in the same experiment), we perform these steps simultaneously for each of the intervened variables. Thus from the original model $(\mathbf{B}, \mathbf{\Sigma_e})$ we obtain a manipulated model $(\mathbf{B}^*, \mathbf{\Sigma_e^*})$ for which

$$\mathbf{x} \ := \ \mathbf{B}^*\mathbf{x} + \mathbf{c} + \mathbf{e}^*,$$

where the intervention vector $\mathbf{c}$ is zero except for rows corresponding to variables that were subject to an intervention, while the manipulated disturbances $\mathbf{e}^*$ equal $\mathbf{e}$ except that elements for intervened variables are set to zero. This follows the standard view that interventions break all incoming arrows to any intervened node in a directed graph representing the direct effects between the observed variables (Pearl 2000).

In some situations an intervention that breaks the influence of the set of causes on a given variable may not be possible, while a 'soft' intervention, in which the intervention influences the variable but does not fully determine it, can be performed. In our framework such an intervention on a variable x_i does not affect the corresponding row of $\mathbf{B}$, nor the corresponding disturbance e_i, but it does add an independent experimental variable c_i. Assuming that c_i is measured, its correlation with $\mathbf{x}$ can be used to obtain an estimate of its causal effect, as in the standard instrumental variable setting.[1] In what follows, however, all interventions are surgical unless otherwise explicitly stated.

We assume that our data sets are generated from a sequence of experiments $(\mathcal{E}_k)_{k=1,\ldots,m}$ where each experiment $\mathcal{E}_k = (\mathbf{J}_k, \mathbf{U}_k)$ consists of a set $\mathbf{J}_k$ of one or more variables that are subject to an intervention (simultaneously and independently) and a set $\mathbf{U}_k$ denoting the other variables, which are passively observed.

In each experiment $\mathcal{E}_k = (\mathbf{J}_k, \mathbf{U}_k)$ our measurements consist of the *experimental effects* of each $x_j \in \mathbf{J}_k$ on each $x_u \in \mathbf{U}_k$, denoted by $t(x_j \rightsquigarrow x_u || \mathbf{J}_k)$, which is equal to the covariance of x_j and x_u in this experimental setting (since x_j has unit variance).The *total effect* $t(x_j \rightsquigarrow x_u)$ of x_j on x_u is standardly defined as the experimental effect in a (hypothetical if not actual) experiment in which *only* x_j is subject to an intervention, i.e. we have $t(x_j \rightsquigarrow x_u) \equiv t(x_j \rightsquigarrow x_u || \{x_j\})$. In experiments in which interventions are performed on more than one variable (i.e. $\mathbf{J}_k$ is larger than the set $\{x_j\}$), the experimental effect may differ from the total effect, so $t(x_j \rightsquigarrow x_u)$ may in general not equal $t(x_j \rightsquigarrow x_u || \mathbf{J}_k)$ when $\mathbf{J}_k \neq \{x_j\}$.

Additionally, if available, a set of passive observational data (for which $\mathbf{J}_k = \emptyset$) can be used to estimate the (passive observational) covariance matrix of the observed variables: $\mathbf{\Sigma_x} = E\{\mathbf{x}\mathbf{x}^T\}$.

In the special case when the variables can be ordered in such a way that the matrix $\mathbf{B}$ is strictly lower-triangular (corresponding to a directed *acyclic* graph) the interpretation of the model is straightforward. This case is known as a 'recursive SEM' in the literature and corresponds to a causal Bayes network with linear relationships over continuous variables.

The interpretation of models which contain directed cycles (i.e. are 'non-recursive') requires more care. Such cases arise from feedback phenomena and represent systems that equilibriate.[2] To ensure that an equilibrium exists the absolute values of the eigenvalues of $\mathbf{B}$ must all be less than 1. This condition must also be satisfied for the manipulated matrix $\mathbf{B}^*$ in all possible experiments (both hypothetical and actual).[3]

While self-loops ($[\mathbf{B}]_{ii} \neq 0$ for any i) are permitted in the true model, the estimated model will not represent them explicitly; instead the effect of the self-loop will be included in the incoming edges to the variable in question.[4] The only underdetermination is the path and the speed to convergence to the equilibrium, not the equilibrium point itself. Since all of our measurments are presumed to come from the equilibrium, this underdetermination is not only natural, but inevitable. Hence, we assume in the following that $\mathbf{B}$ has been standardized to have a diagonal of zeros.

Finally, we note that our representation of latent variables is only implicit (in correlations among the components of $\mathbf{e}$) rather than explicit. Thus, while the true model may contain some latent variables that simultaneously confound more than two variables, these will be rendered as a set of non-zero entries in the estimated covariance matrix $\mathbf{\Sigma_e}$, connecting any pair of variables that are confounded by the hidden variables.

3 PROCEDURE

The experiments only directly supply measures of the experimental effects, but our main interest lies in the direct effects represented by the matrix $\mathbf{B}$ in the model of Section 2. Our procedure for deriving the direct effects essentially consists of two steps: The experimental effects are combined to infer total effects, and the total effects are subsequently used to derive the direct effects. We first describe the latter step.

[1] We here assume that the parametrization $(\mathbf{B}, \mathbf{\Sigma_e})$ of the original model does not change as a result of the soft intervention. One could imagine other soft interventions where this is not the case.

[2] Our procedure can handle both stochastic and deterministic equilibria for cyclic models, but the details are beyond the scope of this paper.

[3] In order to avoid special cases, in which identifiability depends on the experiments performed *and* the particular parameterization, we use this slightly stronger assumption than is needed, including all possible experiments.

[4] If variable x_i has an incoming edge with edge-coefficient a and a self-loop with edge-coefficient b in the true model $\mathbf{B}$, then the estimated model will not return the self loop but instead return the edge coefficient on the incoming edge as $a/(1-b)$.

3.1 TOTAL TO DIRECT EFFECTS

An experiment $\mathcal{E}_k$ consisting of a single surgical intervention (i.e. $\mathbf{J}_k = \{x_j\}$) directly supplies the total effect of that variable (x_j) on all the other measured variables. (Note that the presence of latent variables does not affect the total effect of a variable that is subject to a surgical intervention.[5]) If we have such an experiment for each measured variable, then all total effects can be estimated and represented by a total effects matrix $\mathbf{T}$, where $[\mathbf{T}]_{ji} = t(x_i \rightsquigarrow x_j)$ and we define $t(x_i \rightsquigarrow x_i) = 1$ for all i.

By solving equation (1) for $\mathbf{x}$ as a function of $\mathbf{e}$, we obtain $\mathbf{x} = \mathbf{A}\mathbf{e}$ with $\mathbf{A} = (\mathbf{I} - \mathbf{B})^{-1}$, where $\mathbf{I}$ refers to the $n \times n$ identity matrix.[6] It turns out that $\mathbf{T} = \mathbf{A}\mathbf{D}$ where $\mathbf{D}$ is a diagonal matrix that rescales $\mathbf{A}$ to have a diagonal of ones. Thus, given an estimated total effects matrix $\mathbf{T}$ we can infer the direct effects as

$$\mathbf{B} = \mathbf{I} - \mathbf{A}^{-1} = \mathbf{I} - \mathbf{D}\mathbf{T}^{-1}.$$

Because $\mathbf{B}$ is standardized to have a diagonal of zeros, it follows that $\mathbf{D}$ rescales $\mathbf{T}^{-1}$ to have a diagonal of ones. Furthermore, given $\mathbf{A}$ and the covariance matrix $\mathbf{\Sigma_e}$ over the disturbances, $\mathbf{\Sigma_x}$ is determined by

$$\mathbf{\Sigma_x} = E\{\mathbf{x}\mathbf{x}^T\} = E\{\mathbf{A}\mathbf{e}\mathbf{e}^T\mathbf{A}^T\} = \mathbf{A}\mathbf{\Sigma_e}\mathbf{A}^T.$$

Given passive observational data from which we can estimate $\mathbf{\Sigma_x}$, we can now infer the sought covariance as $\mathbf{\Sigma_e} = \mathbf{A}^{-1}\mathbf{\Sigma_x}\mathbf{A}^{-T}$ to determine the latent variables.

3.2 EXPERIMENTAL TO TOTAL EFFECTS

Experiments intervening on more than one variable enable search strategies that are more efficient (in the number of experiments *and* in total sample size) than single intervention experiments (see Section 5, Fig. 2). In such multi-intervention experiments the experimental effects do not necessarily correspond to the total effects: Any path of the form $x_j \rightarrow \ldots \rightarrow x_i \rightarrow \ldots \rightarrow x_u$ is broken if x_i is surgically manipulated. Nevertheless, the experimental effects provide partial information on the total effects in the form of linear constraints: In an experiment $\mathcal{E}_k = (\mathbf{J}_k, \mathbf{U}_k)$ with $x_j \in \mathbf{J}_k$ and $x_u \in \mathbf{U}_k$ the *total* effect $[\mathbf{T}]_{uj} = t(x_j \rightsquigarrow x_u)$ from x_j to x_u can be represented as

$$[\mathbf{T}]_{uj} = t(x_j \rightsquigarrow x_u) \;=\; \sum_{x_i \in \mathbf{J}_k} t(x_j \rightsquigarrow x_i)\, t(x_i \rightsquigarrow x_u || \mathbf{J}_k) \quad (2)$$

where we define $t(x_j \rightsquigarrow x_j) = 1$. That is, the total effect of x_j on x_u can be redescribed as a decomposition of the total effects of x_j on each variable x_i in the intervention set and the experimental effects of the x_i on the passively observed variable x_u. In the experiment only the experimental effects $t(x_i \rightsquigarrow x_u || \mathbf{J}_k)$ can be measured, but given the intervention set $\mathbf{J}_k$ we can specify for each experiment $q_k = |\mathbf{J}_k| \times |\mathbf{U}_k|$ linear constraints of the form of equation (2) on the total effects. Since there are $n^2 - n$ total effects for n measured variables, several experiments will be required to obtain a sufficient number of constraints.

We can now combine the set of $q = \sum_i q_i$ linear constraints (from all available experiments) into

$$\mathbf{H}\mathbf{t} \;=\; \mathbf{h} \quad (3)$$

where the $(q \times (n^2 - n))$-matrix $\mathbf{H}$ and the vector $\mathbf{h}$ of q scalars contain the measured experimental effects, and $\mathbf{t}$ is the $(n^2 - n)$-vector of desired total effects that can be re-arranged into the matrix $\mathbf{T}$.

In principle, any number of constraints can be represented in this equation: constraints that are linearly dependent (by coincidence or because two sets of measurements, possibly from different experiments, constrain the same total effects and agree), constraints that conflict (because two sets of measurements constrain the same total effects, but do not agree), or added constraints that represent background knowledge (e.g. x_j is not an ancestor of x_i: $t(x_j \rightsquigarrow x_i) = 0$, or more generally that the total effect of x_j on x_i is equal to some scalar c: $t(x_j \rightsquigarrow x_i) = c$).

In most cases $\mathbf{H}$ is not square nor full-rank, and hence equation (3) is not straightforwardly solvable for $\mathbf{t}$. When there are no conflicting constraints and the rank of $\mathbf{H}$ is less than $n^2 - n$ then the system is underdetermined, so there are several directed graphs that satisfy the experimental constraints. If on the other hand the rank of $\mathbf{H}$ is $n^2 - n$ then there is either a unique causal structure that satisfies all the constraints, or if there are conflicting constraints, there is no such model.

In principle, there are several ways to proceed. One could select non-conflicting constraints until full rank is achieved and proceed to the unique solution ignoring the other constraints, thereby avoiding (denying) conflict and remaining agnostic in the underdetermined case. We proceed differently: We use the Moore-Penrose pseudoinverse $\mathbf{H}^\dagger$ to invert $\mathbf{H}$ using whatever constraints are available. In the underdetermined case the pseudoinverse returns the solution that minimizes the l^2-norm of the total effects; in the square invertible case we obtain the unique solution, and in the overdetermined case we obtain the solution that minimizes the sum of the squared errors of the constraints. We

[5] In the case of soft interventions techniques used for instrumental variables can be applied with the same result.

[6] Recall that for stability of the linear system we have to require that the absolute values of all eigenvalues of $\mathbf{B}$ are smaller than 1. Thus for each $\lambda \in \text{eigenvalues}(\mathbf{B})$ we have $1 - \lambda \in \text{eigenvalues}(\mathbf{I} - \mathbf{B})$, so the stability condition ensures that no eigenvalue of $\mathbf{I} - \mathbf{B}$ is zero and the matrix is always invertible.

Program 1 Algorithm pseudocode. A full implementation of this procedure is available at:
`http://www.cs.helsinki.fi/u/phoyer/code/LLC/`

EXPERIMENTS, LLC: (linear, latents, cyclic)

Given m datasets $\mathcal{D}_k$ from a sequence of experiments $(\mathcal{E}_k)_{k=1,\ldots,m}$ and, if available, also a dataset $\mathcal{D}_o$ of passive observational data over a set of measured variables:

For each experiment $\mathcal{E}_k = (\mathbf{J}_k, \mathbf{U}_k)$ with dataset $\mathcal{D}_k$

For each ordered pair of variables (x_j, x_u) with $x_j \in \mathbf{J}_k$ and $x_u \in \mathbf{U}_k$

Determine from $\mathcal{D}_k$ the constraint $t(x_j \rightsquigarrow x_u) = \sum_{x_i \in \mathbf{J}_k} t(x_j \rightsquigarrow x_i) t(x_i \rightsquigarrow x_u || \mathbf{J}_k)$.

Re-order the constraint and concatenate it to the matrix equation $\mathbf{Ht} = \mathbf{h}$.

Check $\mathbf{H}$ and $\mathbf{h}$ and report rank and the existence of conflicts, if any, to determine whether the system is underconstrained, conflicted or exactly determined.

Compute $\mathbf{t} = \mathbf{H}^{\dagger}\mathbf{h}$, and determine the total effects matrix $\mathbf{T}$ by reordering $\mathbf{t}$ into matrix form (note that the diagonal of $\mathbf{T}$ consists of 1's.).

Determine the direct effects matrix $\mathbf{B} = \mathbf{I} - \mathbf{D}\mathbf{T}^{-1}$, where $\mathbf{I}$ is the $n \times n$ identity matrix.

If passive observational dataset $\mathcal{D}_o$ is available then

Estimate the passive observational covariance matrix $\mathbf{\Sigma_x} = E\{\mathbf{x}\mathbf{x}^T\}$ from dataset $\mathcal{D}_o$.

Infer the error covariance matrix $\mathbf{\Sigma_e} = \mathbf{A}^{-1}\mathbf{\Sigma_x}\mathbf{A}^{-T}$.

Return the estimated $\mathbf{B}$ (and $\mathbf{\Sigma_e}$ if available).

are thus able to provide an 'online' algorithm (Program 1), which can integrate background knowledge and, given the constraints, does as well as possible with regard to a well-defined measure.

4 IDENTIFIABILITY

Ideally, the constraints obtained from the experiments would determine the total effects matrix $\mathbf{T}$ uniquely. However, this will almost never be the case, since most combinations of experiments provide either too few or too many constraints, and in the latter case conflicts will be common. Hence the crucial question for the scientist is how the identifiability of the causal structure is related to the experiments. Given a set of measured variables $\mathbf{V}$, the following condition enables exactly this connection between identifiability and experimental set-up.

Definition 1 *Given experiments $(\mathcal{E}_k)_{k=1,\ldots,m}$ and an ordered pair of variables $(x_i, x_j) \in \mathbf{V} \times \mathbf{V}$ with $i \neq j$ we say that the* pair condition *is satisfied for this variable pair w.r.t. the experiments whenever there is a k such that $x_i \in \mathbf{J}_k$ and $x_j \in \mathbf{U}_k$ for some $\mathcal{E}_k = (\mathbf{J}_k, \mathbf{U}_k)$.*

Theorem 1 *Given the* experimental *effects from the experiments $(\mathcal{E}_k)_{k=1,\ldots,m}$ over the variables in $\mathbf{V}$, all* total *effects $t(x_i \rightsquigarrow x_j)$ are identified if and only if the pair condition is satisfied for all ordered pairs of variables w.r.t. these experiments.*[7]

Sketch of proof of necessity: Suppose the pair condition is not satisfied for an ordered pair of variables (x_i, x_j). Then for any experiment $\mathcal{E}_k = (\mathbf{J}_k, \mathbf{U}_k)$, either (i) $x_i, x_j \in \mathbf{J}_k$ or (ii) $x_i, x_j \in \mathbf{U}_k$ or (iii) $x_i \in \mathbf{U}_k$ and $x_j \in \mathbf{J}_k$. Consider $t(x_i \rightsquigarrow x_j)$. Experiments of type (iii) place no constraint on $t(x_i \rightsquigarrow x_j)$. Experiments of type (i) provide for each $w \in \mathbf{U}_k$ a constraint of the form $t(x_i \rightsquigarrow w) = t(x_i \rightsquigarrow x_j) t(x_j \rightsquigarrow w || \mathbf{J}_k) + \text{const}$, but since $t(x_i \rightsquigarrow w)$ is unknown, these constraints are insufficient to determine $t(x_i \rightsquigarrow x_j)$. The only relevant constraints from experiments of type (ii) marginalize over x_i making it impossible to separate the effect of some $w \in \mathbf{J}_k$ via x_i from the more direct effects:

$$\begin{aligned} t(w \rightsquigarrow x_j || \mathbf{J}_k) &= t(w \rightsquigarrow x_i || \mathbf{J}_k) t(x_i \rightsquigarrow x_j || \mathbf{J}_k \cup \{x_i\}) \\ &\quad + t(w \rightsquigarrow x_j || \mathbf{J}_k \cup \{x_i\}) \end{aligned}$$

Since there will be a constraint of this type for each w, $t(x_i \rightsquigarrow x_j || \mathbf{J}_k \cup \{x_i\})$ is undetermined. But it is needed to determine the total effect of x_i on x_j in terms of what is known, since:

$$\begin{aligned} t(x_i \rightsquigarrow x_j) &= t(x_i \rightsquigarrow x_j || \mathbf{J}_k \cup \{x_i\}) \\ &\quad + \sum_{x_k \in \mathbf{J}_k} t(x_i \rightsquigarrow x_k) t(x_k \rightsquigarrow x_j || \mathbf{J}_k \cup \{x_i\}) \end{aligned}$$

Hence the total effect $t(x_i \rightsquigarrow x_j)$ is underdetermined, so the total effects matrix $\mathbf{T}$ is not fully determined. □

Sketch of proof of sufficiency: Since the pair condition is satisfied for each ordered pair of variables (x_i, x_j), we can select for each total effect $t(x_i \rightsquigarrow x_j)$ one such constraint for our matrix equation $\mathbf{Ht} = \mathbf{h}$.

Let $\mathbf{H}^*$ be the matrix of $n^2 - n$ constraints on the total effects obtained when for each variable there is an experiment intervening on all but that variable. Each constraint will have at most $n - 1$ non-zero entries in $\mathbf{H}^*$. These entries will correspond to direct effects of an intervened variable on the target variable. Appropriately re-ordering the rows, $\mathbf{H}^*$ can be organized into a block-diagonal matrix, where each block on the diagonal corresponds to a submatrix of $\mathbf{I} - \mathbf{B}$. One such block is shown in Table 1. Given the stability

[7] Full proofs supplied upon request. Due to space constraints we here give a proof sketch only.

Table 1: One of the blocks in $\mathbf{H}^*$. The $b(x_i \to x_j)$ are entries of the direct effects matrix $\mathbf{B}$.

$\mathbf{J}$	$t(x_1 \rightsquigarrow x_2)$	$t(x_1 \rightsquigarrow x_3)$	$t(x_1 \rightsquigarrow x_4)$
$\mathbf{V} \setminus \{x_2\}$	1	$-b(x_3 \to x_2)$	$-b(x_4 \to x_2)$
$\mathbf{V} \setminus \{x_3\}$	$-b(x_2 \to x_3)$	1	$-b(x_4 \to x_3)$
$\mathbf{V} \setminus \{x_4\}$	$-b(x_2 \to x_4)$	$-b(x_3 \to x_4)$	1

assumption on $\mathbf{B}$ (see Section 2), any such block, and consequently $\mathbf{H}^*$, is full rank. In cases where the pair condition is satisfied, but the experiments did not intervene on all but one variable, at least one of the off-diagonal elements in such a block of $\mathbf{H}$ must be zero. Leaving a variable out of the intervention set corresponds in the measurement of the constraint to a marginalization over that variable. It can be shown that marginalization correponds to elementary row operations on the corresponding block of $\mathbf{H}^*$. Further, since satisfaction of the pair condition ensures that each row of $\mathbf{H}$ consists of a constraint corresponding to a different marginalization, it can be shown that any such $\mathbf{H}$ can be obtained by row operations from the $\mathbf{H}^*$ matrix. Since $\mathbf{H}^*$ is full rank, and since row operations preserve rank, any such $\mathbf{H}$ is full rank, and hence invertible. □

For an underdetermined system, the theorem implies that there are ordered pairs of variables that do not satisfy the pair condition. One way to proceed is to select the next intervention set to maximally reduce the number of such pairs. Using brute force, this would require checking $\sum_{i=1}^{k} \binom{n}{i}$ possible experiments for n variables and maximum intervention set size $k \leq n/2$. However, faster heuristics are available or can be adapted (Eberhardt 2008). If feasible, an experiment that intervenes on $n/2$ variables, none of which have yet been subject to intervention, supplies the maximum number of additional pair constraints. Selecting the next best experiment is obviously greedy, and if this procedure is used to select a whole sequence of experiments starting with no prior knowledge, it will select a sequence of $2\log_2(n)$ experiments intervening on a different set of $n/2$ variables in each experiment. This sequence of experiments is sufficient for identifiability, but is known to be suboptimal for most n.

Instead of attempts at further resolution one might be interested in a characterization of the underdetermination. Given that the system is linear, the solution space for the total effects is easily specified, so further numerical or analytical routines can be used to specify the implications of the underdetermination for the direct effects matrix $\mathbf{B}$. Although analytical procedures could be devised, a naive numerical heuristic would sample total effects from the solution space, and invert $\mathbf{T}$ repeatedly to explore the variation in $\mathbf{B}$.

5 SIMULATIONS

Several factors will affect the accuracy of our estimation procedure. The number of samples used in each experimental condition will naturally be an important consideration. Another is the number of constraints obtained from the set of experiments, which depends not only on the number of experiments but also on how many variables are intervened per experiment. Finally, the relative efficiency of soft vs. surgical interventions is unclear. It is thus not obvious what the best experimental strategy is in any given practical situation.

To investigate the issue empirically, we generated 100 random cyclic models over 8 variables, each with 4 latents and possible self-loops, and sampled experimental data from these models for a variety of strategies. First, we considered surgical interventions only. We compared the strategy of intervening on each variable separately (8 experiments to satisfy the pair condition) to intervening on pairs of variables (also requiring 8 experiments) to intervening on four variables at a time (requiring only $2\log_2 8 = 6$ experiments), with a variety of sample sizes. Figure 2 (top row) shows the accuracy measured as the mean correlation between true and estimated values across the 100 graphs as a function of the *total number* of samples (divided evenly over the number of experiments) used for each strategy. The main result is that experiments involving interventions on a number of variables are more effective than experiments of single interventions, reflecting the larger number of constraints obtained per experiment.

Next, we investigated the efficiency of soft interventions, performing identical simulations to those given above but replacing surgical interventions with soft interventions (bottom row in Figure 2). Here, we additionally test the strategy of performing a *single experiment* intervening on *all* variables simultaneously. (Surgically intervening on all variables simultaneously does not provide any information, but softly intervening does.) Results are comparable to the surgical interventions case, except that softly intervening on all variables is a superior strategy in terms of the accuracy as a function of the total number of samples used.

6 FLOW CYTOMETRY DATA

To demonstrate how our algorithm might be used in practice we show its application to the flow cytometry data of Sachs et al. (2005). This data consists of single cell recordings of the abundance of 11 different phosphoproteins and phospholipids under a number of experimental conditions in human T-cells. In each condition, measurements were obtained from some 700–900 single cells. These data have been previously an-

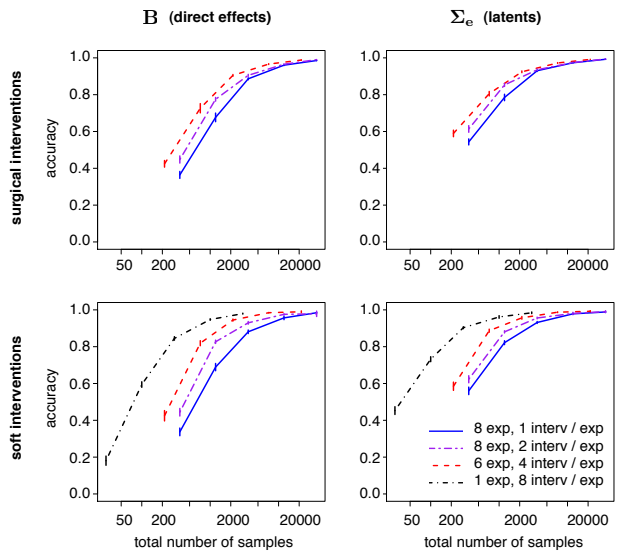

Figure 2: Results on simulated data with 8 observed variables. Each graph shows the accuracy (correlation between estimates and true values) as a function of the total number of samples used, with error bars indicated. The top plots show accuracy of the estimated $\mathbf{B}$ and $\mathbf{\Sigma_e}$ for *surgical* interventions, while the bottom row gives corresponding results for *soft* interventions.

alyzed using several techniques. Sachs et al. (2005) show that many known causal relationships between the measured signaling molecules can be found using standard Bayesian network learning methods on this data. Eaton & Murphy (2007) provide evidence that many of the 'specific perturbations' may in fact not have been localized to single targets. A recent analysis provided by Schmidt & Murphy (2009) was the first to apply a technique for learning cyclic models.

To give the reader a sense of the data we follow Ellis & Wong (2008) and plot in Figure 3 histograms of two out of the eleven measured variables, in five different conditions. Panel (a) shows the response of Raf while (b) displays the activity of Akt. The top rows show the responses to the 'background condition' of α-CD3/28 stimulation. The four bottom rows show corresponding histograms in four separate experiments where, in addition to α-CD3/28, inhibitors were selectively applied to Akt, PKC, PIP2, and Mek (top to bottom respectively). In panel (a) we see that the inhibitors applied to PKC and Mek have a large effect on Raf, while the effect (if any) of inhibition of Akt and PIP2 is minimal. In (b) we see that all four experimental conditions affect Akt, though the effect is again most pronounced with the inhibitor on PKC.

A number of issues are worth pointing out. First, note that the Akt inhibitor has a minimal (if any) effect on the measured quantity of Akt. This is because the

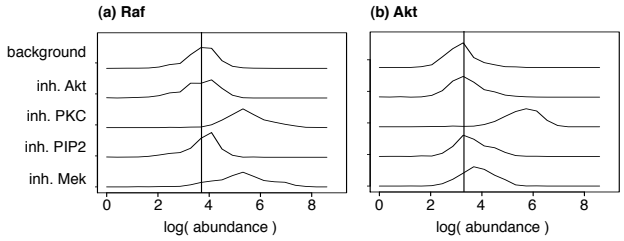

Figure 3: Histograms of Raf and Akt in different experimental conditions. Top row: histograms of Raf and Akt in the background condition (α-CD3/28). Second-to-bottom rows, histograms of Raf and Akt in the four stimulus conditions where selectively Akt, PKC, PIP2, and Mek (respectively) were manipulated.

inhibitors may just inactivate a given molecule rather than decrease its abundance (Sachs et al. 2005). Thus, measured quantities of the intervened-upon molecules are not reliable indicators of how active they are, and so we must impute a 'low' value for the inhibited variables. Second, note that there are strong effects in the data that cannot be captured using acyclic models with targeted interventions. For example, the strong effect of Mek on Raf as seen in panel (a) goes against the acyclic model given by Sachs et al. (2005). Although these effects may be accounted for using the framework of 'uncertain interventions' (Eaton and Murphy 2007), our approach is to account for them using cyclic models. Finally, we emphasize that some of the experimental settings with 'specific' perturbations in the original dataset were not actually targeted interventions because they changed the background conditions. In particular, all excitatory perturbations were performed without the α-CD3/28 background stimulus, and hence any differences to the passive observational case observed in these conditions may be due to either the excitatory stimulus or the change in the background condition. For this reason we exclude them from our analysis.

Since each of the four remaining inhibitory experiments only targeted a single molecule, the experimental effects here directly correspond to the total effects $t(x_i \rightsquigarrow x_j)$. Our measure of total effects essentially determines the deviation of the means of the variables in the experimental conditional from the passive observational condition, normalized by the standard deviation in the passive observational case. For instance, the total effects of PKC, PIP2, and Mek (respectively) onto Akt were -0.967 ± 0.021 , -0.153 ± 0.031, and -0.256 ± 0.048; the effects are negative because an *inhibition* of the cause *increased* the amounts of the effect molecule. (Estimates of errors were obtained using a bootstrap approach.) These numbers correspond to the difference between the top row and the bottom three rows (respectively) of Figure 3b.

Table 2: Estimated direct effects from the Sachs et al. data, as estimated from (a) the experiments with a background condition of α-CD3/28; and (b) with a background condition of α-CD3/28 + ICAM-2. The tables are read from column to row so, for instance, the direct effect of PKC on PIP2 is -0.90 and -0.61 in the two settings. Errors given by a bootstrap analysis were small compared to the size of the coefficients.

(a)

	Akt	PKC	PIP2	Mek
Akt		-0.91	-0.22	0.22
PKC	-0.15		-0.09	0.48
PIP2	-0.43	-0.90		0.40
Mek	-0.27	-1.69	-0.22	

(b)

	Akt	PKC	PIP2	Mek
Akt		-1.34	-0.24	-0.26
PKC	0.13		-0.10	-0.08
PIP2	0.09	-0.61		-0.01
Mek	0.10	-0.93	-0.11	

Given the total effects among the four intervened-upon molecules, we derived the corresponding direct effects among them. Note that these direct effects are direct only relative to these four molecules; the effects are presumably mediated by the other (both measured and unmeasured) molecules in the signaling system. The result is shown in Table 2a. The most pronounced result is that PKC has strong negative connections to the other three variables. Similarly, PIP2 and Akt have weak negative effects, and Mek weak positive direct effects on the other variables.

While it is obvious that linearity is a very strong (and indeed quite unreasonable) assumption for this data, we nevertheless hope that we can use it as a first and very rough approximation. Fortunately, the dataset contains some additional data that can be used to at least partly corroborate the results from our analysis: In addition to the α-CD3/28 background condition, there are similar experimental data in which a reagent called ICAM-2 was added to the background, both with or without the inhibitory targeted interventions. Using this additional data, we were able to confirm the strong inhibitory direct effect (with respect to these four variables) of PKC on the other three molecules, as shown in Table 2b. Similarly, PIP2 still has a weak negative effect on the other variables. On the other hand, in this additional data the effect of Akt and of Mek on the other variables is weak and of opposite sign to the results in (a), indicating that these connections should be treated as unreliable at best. Thus, in this respect, the model fits the 'ground truth' graph of Sachs et al. (2005) relatively well, as there were no directed paths from these variables to the other of the four variables in their graph.

7 CONCLUSIONS

We have described (and provided a software package for) a general algorithm for discovery of linear causal models using experimental data. The algorithm can handle cyclic structures and discover the presence and location of latent variables. It is 'online' in that it works with any set of experimental data, and indicates whether the system is underdetermined, overdetermined or exactly determined. The method of integrating data from different experiments resolves conflicts in a way that is optimal with regard to a specified measure. In an underdetermined system the algorithm provides a minimally satisficing result, and we have suggested how to proceed by selecting more experiments, or by characterizing the underdetermination.

It should be noted that in addition to the generality of structures we consider, we do not assume faithfulness. This results in the (admittedly demanding) identifiability condition we have shown to be necessary and sufficient. Similarly, if constraints other than those on the experimental effects are considered, model search can be made more efficient.

The current version of the algorithm relies on the assumption of linearity. While no corresponding solution can be supplied for general discrete networks, we are hopeful that a similar procedure for particular types of discrete parameterizations (e.g. noisy-or) is possible.

References

Cooper, G. and C. Yoo (1999). Causal discovery from a mixture of experimental and observational data. In *UAI '99*, pp. 116–125.

Eaton, D. and K. Murphy (2007). Exact Bayesian structure learning from uncertain interventions. In *AISTATS '07*.

Eberhardt, F. (2008). Almost optimal intervention sets for causal discovery. In *UAI '08*.

Ellis, B. and W. Wong (2008). Learning causal Bayesian network structures from experimental data. *J. Amer. Stat. Assoc. 103*, 778–789.

He, Y. and Z. Geng (2008). Active learning of causal networks with intervention experiments and optimal designs. *J. Mach. Learn. Res. 9*, 2523–2547.

Murphy, K. P. (2001). Active learning of causal Bayes net structure. Technical report, U.C. Berkeley.

Nyberg, E. and K. Korb (2006). Informative interventions. In *Causality and Probability in the Sciences*. College Publications, London.

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

Richardson, T. (1996). *Feedback Models: Interpretation and Discovery*. Ph. D. thesis, Carnegie Mellon.

Sachs, K., O. Perez, D. Pe'er, D. Lauffenburger, and G. Nolan (2005). Causal protein-signaling networks derived from multiparameter single-cell data. *Science 308*(5721), 523–529.

Schmidt, M. and K. Murphy (2009). Modeling discrete interventional data using directed cyclic graphical models. In *UAI '09*.

Tong, S. and D. Koller (2001). Active learning for structure in Bayesian networks. In *UAI '01*, pp. 863–869.