Journal of Machine Learning Research 1 (2000) 1-48 Submitted 4/00; Published 10/00

Belief net structure learning from uncertain interventions

Daniel Eaton DEATON@CS.UBC.CA
Kevin Murphy MURPHYK@CS.UBC.CA
University of British Columbia
Department of Computer Science
2366 Main Mall
Vancouver, BC V6T 1Z4
Canada

Editor: TheEditor

Abstract

We show how to learn causal structure from interventions with unknown effects and/or side effects by adding the intervention variables to the graph and using Bayesian inference to learn the resulting two-layered graph structure. We show that, on a datatset consisting of protein phosphorylation levels measured under various perturbations, learning the targets of intervention results in models that fit the data better than falsely assuming the interventions are perfect. Furthermore, learning the children of the intervention nodes is useful for such tasks as drug and disease target discovery, where we wish to distinguish direct effects from indirect effects. We illustrate the latter by correctly identifying known targets of genetic mutation in various forms of leukemia using microarray expression data.

Keywords: Bayesian Networks, Structure Learning, Causality, Interventions, Drug Target Discovery

1. Introduction

The use of belief networks (directed graphical models) to represent causal models has become increasingly popular (Pearl, 2000; Spirtes et al., 2000). In particular, there is much interest in learning the structure of these models from data, particularly in the area of systems biology (Friedman, 2004) and cognitive science (Gopnik and Schulz, 2007). However, given observational data, it is only possible to identify the structure up to Markov equivalence. For example, the three models $X \rightarrow Y \rightarrow Z$, $X \leftarrow Y \leftarrow Z$, and $X \leftarrow Y \rightarrow Z$ all encode the same conditional independency statement, $X \perp Z | Y$. To distinguish between such models, we need interventional (experimental) data (Cooper and Yoo, 1999; Pearl, 2000; Spirtes et al., 2000; Eberhardt et al., 2005; Korb and Nyberg, 2006).

Most previous work has focused on the case of "perfect" or "ideal" interventions, in which it is assumed that an intervention sets a single variable to a specific state (as in a randomized experiment). This is the basis of the "do-calculus" (as in the verb "to do") of Pearl (2000). A perfect intervention essentially "cuts off" the influence of the parents to the intervened node, and can be modeled as a structural change by performing "graph surgery" (removing incoming edges from the intervened node). Although some real-world interventions can be modeled in this way (such as gene knockouts), most interventions are not so precise in their effects.

One possible relaxation of this model is to assume that interventions are "soft", and merely increase the probability the target will enter a specific state (Markowetz et al., 2005). (An example of this would be requesting a patient to take a certain drug.) A further relaxation is to assume that the effect of an intervention does not render the node independent of its parents, but simply changes the parameters of the local distribution; this has been called a "mechanism change" (Tian and Pearl, 2001a,b) or "parametric change" (Eberhardt et al., 2006). For many situations, this is a more realistic model than perfect interventions, since it is often impossible to force variables into specific states.

In this paper we propose a different relaxation of the notion of perfect intervention, and consider the case where the targets of intervention are uncertain.[1] Note that this is orthogonal to the issue of whether the interventions act on their targets in a perfect or imperfect way. Our approach is straightforward: we add the intervention nodes to the graph as binary indicator variables, and then learn the structure of this two-layer graph (the intervention nodes are in the top layer and act as parents to the regular nodes in the bottom layer or "backbone"). In other words, we do not assume a known 1:1 mapping from interventions to targets, but instead learn their targets (as well as the rest of the graph structure).

Although the approach of adding intervention nodes to the graph is not novel (e.g., it is mentioned in (Pearl, 2000) and used in (Pe'er et al., 2001)), previous work has not examined the consequences of uncertain interventions in any detail. We show, with experiments on synthetic data, that the consequences are quite benign, in the sense that one can learn structure almost as well as if the targets of intervention were known. Furthermore, we show, on a real biological dataset, that learning the targets of intervention can result in a much better fitting model than assuming the targets are known, perhaps because the actual interventions were not as ideal as expected.

In addition to learning more accurate structure between the nodes in the backbone, learning the targets of intervention is often of interest in itself. For example, in the area of drug target discovery (sometimes called identifying the "mode of action" of a compound (Marton et al., 1998; Dejori et al., 2004; Gardner et al., 2003; Hallen et al., 2006)), the goal is to identify which genes change their mRNA expression level as a direct result of adding a drug. This is difficult because many genes may change as a result of a perturbation, but some of these are indirect consequences of the intervention (due to the genes being downstream of the targets). By jointly learning the structure of the network between the genes and the connections from the interventions to the genes, we can "explain away" such indirect effects, even when there is not enough data to fully learn the structure of the backbone.

The structure of the paper is as follows. In Section 2, we provide a summary of various models of intervention that have been proposed in the literature, and we explain our model in more detail. In Section 3, we provide a brief summary of various algorithms for structure learning that have been proposed in the literature, and sketch a novel algorithm designed specifically for two-layer graphs. In Section 4, we provide extensive experimental results on two synthetic and two real data datasets. We conclude in Section 5.

[1] A preliminary version of this paper appeared in (Eaton and Murphy, 2007a).

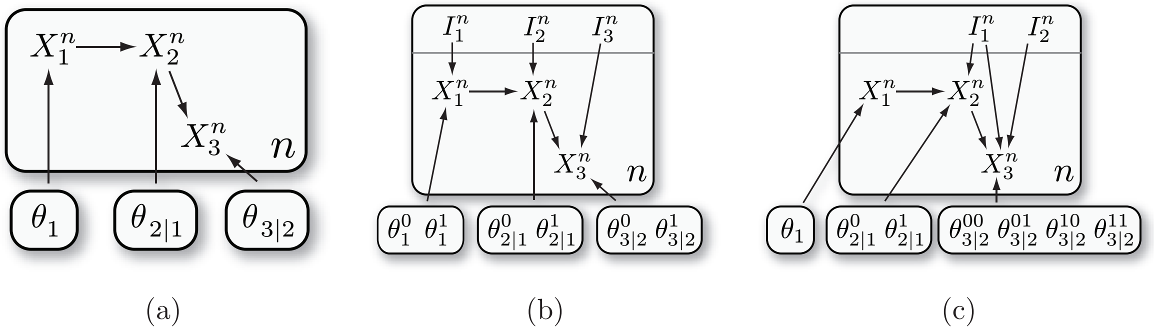

X_1^n
X_2^n
X_3^n
n
θ_1
$\theta_{2|1}$
$\theta_{3|2}$
(a)
I_1^n
I_2^n
I_3^n
X_1^n
X_2^n
X_3^n
n
θ_1^0 θ_1^1
$\theta_{2|1}^0$ $\theta_{2|1}^1$
$\theta_{3|2}^0$ $\theta_{3|2}^1$
(b)
I_1^n
I_2^n
X_1^n
X_2^n
X_3^n
n
θ_1
$\theta_{2|1}^0$ $\theta_{2|1}^1$
$\theta_{3|2}^{00}$ $\theta_{3|2}^{01}$ $\theta_{3|2}^{10}$ $\theta_{3|2}^{11}$
(c)

Figure 1: **An example network with different models of intervention.** X_i^n is backbone node i in data case n, and I_i^n is intervention node i in case n. θ_i are the parameters for backbone node i, outside the $n = 1 : N$ plate. (a) No interventions. (b) Known intervention targets (one I_i for every X_i). (c) Unknown intervention targets. Intervention 1 affects nodes 2 and 3 (and is thus said to have a "fat hand"); intervention 2 affects node 3. The parameters for node 3 are $\theta_{3|2}^{ij}(k, \ell)$, where $I_1 = i$, $I_2 = j$, $X_2 = k$ and $X_3 = \ell$.

2. Models of intervention

For the reader who is already familiar with standard Bayesian approaches[2] to learning belief net structure (see e.g., (Heckerman et al., 1995; Neapolitan, 2003)) we can explain our approach very easily by examining the example in Figure 1. On the left, we show a 3-node belief network, where X_i^n denotes the value of node i in case n, for $i = 1 : d$ and $n = 1 : N$. The parameters of X_i's conditional probability distribution (CPD) are denoted by θ_i: thus $p(X_i|X_{G_i}) = f_i(X_i, X_{G_i}, \theta_i)$, where G_i are i's parents and f_i is some parametric density function. We will call the X_i "backbone" nodes. In the middle, we show the same network, where each X_i has a unique intervention parent, I_i. This is a binary variable that acts like a "switching parent": if $I_i = 0$, the parameters for X_i's CPD are θ_i^0; if $I_1 = 1$, the parameters are θ_i^1. On the right, we show the same network, where we only have 2 intervention nodes. Some intervention nodes have multiple backbone children (these are called "fat hand" interventions), and some backbone nodes have multiple intervention parents. The goal of this paper is to learn graph structures of this type, and to compare their performance to models which assume known targets of intervention. In the sections below, we elaborate on this brief discussion, and explain variants of these basic model types that have been proposed by various authors, culminating in a more detailed description of our proposal.

[2] The Bayesian approach to structure learning avoids many of the conceptual problems that arise when trying to combine the results of potentially inconsistent conditional independency tests performed on data sampled from different interventional regimes (Eberhardt, 2006). Furthermore, it is particularly appropriate when the sample sizes are small, as in many systems biology and cognitive science experiments.

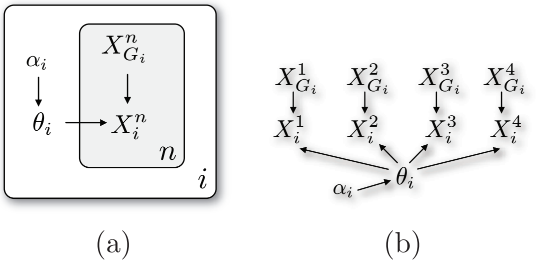


Figure 2: **No interventions.** (a) Plate notation, (b) the same model unrolled across 4 data cases.

2.1 No interventions

We will make the same widely-used assumptions as (Heckerman et al., 1995), which we summarize here for completeness. We assume parameters are a priori globally independent (see Figure 2), and that data is complete, which lets us write the marginal likelihood of a graph as

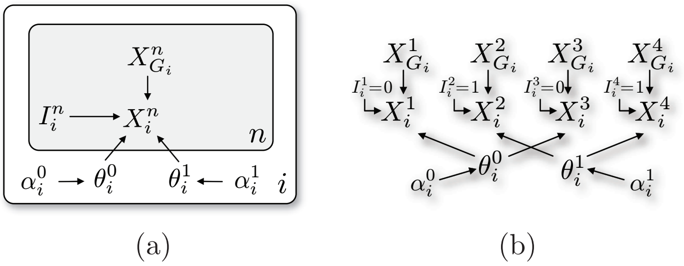
$$p(X^{1:N}|G) = \prod_{i=1}^{d} p(X_i^{1:N}|X_{G_i}^{1:N}) = \prod_{i=1}^{d} \int [\prod_{n=1}^{N} p(x_i^n|x_{G_i}^n, \theta_i)]p(\theta_i)d\theta_i \tag{1}$$

We assume the CPDs are represented as tables (conditional multinomials), $p(X_i = k|X_{G_i} = j, \theta_i) = \theta_{ijk}$, and that the parameter priors are conjugate and satisfy local independence. These assumptions imply that the prior is Dirichlet (Geiger and Heckerman, 1997):

$$p(\theta_i) = \prod_{j=1}^{q_i} Dir(\theta_{ij}|\alpha_{ij1}, \ldots, \alpha_{ijr_i}) \tag{2}$$

where r_i is the number of states of X_i, q_i is the number of states of X_{G_i} and α_{ijk} are the hyper-parameters. We will use the BDeu prior $\alpha_{ijk} = \alpha/q_i r_i$, where we set the prior strength to $\alpha = 1$. (See (Steck and Jaakkola, 2002) for a discussion on how to set α.) With these assumptions, the marginal likelihood of a family (a node and its parents) is given by the following equation:

$$\begin{aligned}
p(x_i^{1:N}|x_{G_i}^{1:N}) &= \int [\prod_{n=1}^{N} p(x_i^n|x_{G_i}^n, \theta_i)]p(\theta_i)d\theta_i \\
&= \prod_{j=1}^{q_i} \int [\prod_{n:x_{G_i}^n=j} p(x_i^n|\theta_{ij})]p(\theta_{ij})d\theta_{ij} \\
&= \prod_{j=1}^{q_i} \int \left[\prod_{k=1}^{r_i} \theta_{ijk}^{N_{ijk}}\right] Dir(\theta_{ij\cdot}|\alpha_{ij\cdot})d\theta_{ij} \\
&= \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij}+N_{ij})} \prod_{k=1}^{q_i} \frac{\Gamma(\alpha_{ijk}+N_{ijk})}{\Gamma(\alpha_{ijk})} \qquad (3)
\end{aligned}$$



(a) (b)

Figure 3: **Imperfect interventions.** (a) Plate notation, (b) the same model unrolled across 4 data cases. In this example, cases 1 and 3 are "normal", and cases 2 and 4 are interventional (for node i).

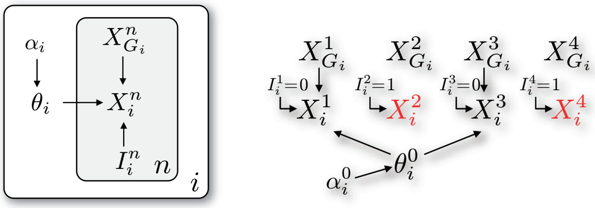
where $N_{ijk} = \sum_{n=1}^{N} I(x_i^n = k, x_{G_i}^n = j)$ and $N_{ij} = \sum_k N_{ijk}$ are the counts and α_{ijk} and $\alpha_{ij} = \sum_k \alpha_{ijk}$ are the pseudo counts. Note that the counts can be efficiently computed for all families from large datasets using ADtrees (Moore and Lee, 1998). An analogous formula can be derived for the normal-Gamma case (Geiger and Heckerman, 2002). However, none of what follows relies on being able to compute the marginal likelihood exactly (which is only possible for certain conjugate prior-CPD pairs). For example, we could use a BIC approximation instead.

2.2 Imperfect interventions

A simple way to model interventions is to introduce intervention nodes, that act like "switching parents": if $I_i^n = 1$, then we have performed an intervention on node i in case n and we use a different set of parameters than if $I_i^n = 0$, when we use the "normal" parameters. Specifically, we set $p(X_i|X_{G_i}, I_i = 0, \theta, G) = f_i(X_i|X_{G_i}, \theta_i^0)$ and $p(X_i|X_{G_i}, I_i = 1, \theta, G) = f_i(X_i|X_{G_i}, \theta_i^1)$. See Figure 3. (Note that the assumption that the functional form f_i does not change is made without loss of generality, since θ_i can encode within it the specific type of function.) Tian and Pearl (2001a,b) refer to this as a "mechanism change". To simplify notation, we assume every node has its own intervention node; if a node i is not intervenable, we simply clamp $I_i^n = 0$ for all n.

When we have interventional data, we modify the local marginal likelihood formula by partitioning the data into those cases in which X_i was passively observed, and those in which X_i was set by intervention:

$$\begin{aligned} p(x_i^{1:N}|x_{G_i}^{1:N}, I_i^{1:N}) &= \int \left[\prod_{n:I_i^n=0} p(x_i^n|x_{G_i}, \theta_i^0) \right] p(\theta_i^0) d\theta_i^0 \\ &\times \int \left[\prod_{n:I_i^n=1} p(x_i^n|x_{G_i}, \theta_i^1) \right] p(\theta_i^1) d\theta_i^1 \end{aligned} \tag{4}$$



(a) (b)

Figure 4: **Perfect interventions.** (a) Plate notation, (b) the same model unrolled across 4 data cases. The red nodes (cases 2 and 4) have been set by perfect intervention, so X_i is cut off from X_{G_i}.

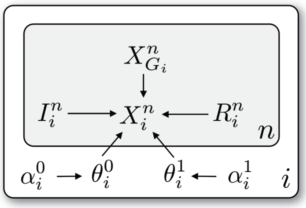


Figure 5: **Unreliable interventions.** We add another switch node R^n_i, so that an intervention is actually a mixture of θ^0_i and θ^1_i.

2.3 Perfect interventions

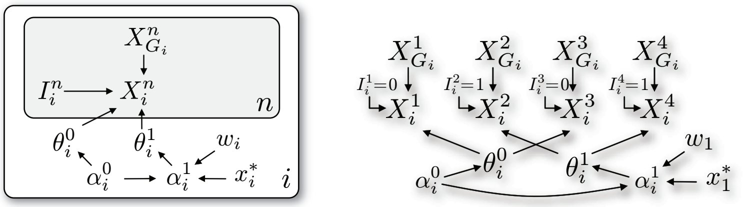
If we perform a perfect intervention on node i in data case n, then we set $X^n_i = x^*_i$, where x^*_i is the desired "target state" for node i. Hence we define the CPD as $p(X_i|X_{G_i}, \theta, I_i = 1) = \delta(X_i - x^*_i)$. Thus a perfect intervention can be seen as a special case of an imperfect intervention where θ^1_i (the parameters that are used when $I_i = 1$) encodes this delta function. (Formally $\theta^1_i = \vec{e}_t$, where $t = x^*_i$ is the target value for node i, and $\vec{e}_t = (0, \ldots, 0, 1, 0, \ldots, 0)$ with a 1 in the t'th position.) From Figure 4, we see that X_i is effectively "cut off" from its parents X_{G_i}. In this case, the second term of Equation 4 evaluates to 1 (assuming $x^n_i = x^*_i$ for each forced term), so the marginal likelihood simplifies as follows:

$$p(x^{1:N}_i | x^{1:N}_{G_i}, I^{1:N}_i) \quad = \quad \int [\prod_{n:I^n_i=0} p(x^n_i | x_{G_i}, \theta^0_i)] p(\theta^0_i) d\theta^0_i \tag{5}$$

In other words, we just drop cases in which node i was set by intervention (Cooper and Yoo, 1999).

2.4 Unreliable interventions

An orthogonal issue to whether the intervention is perfect or imperfect is the reliability of the intervention, i.e., how often does the intervention succeed? One way to model this is to assume that each attempted intervention succeeds with probability ϕ_i and fails with



(a) (b)

Figure 6: **Soft interventions.** (a) Plate notation, (b) the same model unrolled across 4 data cases. Proposed by Markowetz et al. (2005). x^*_i is the known state into which we wish to force node i when we perform an intervention on it; w_i is the strength of this intervention. α^1_i is a deterministic function of α^0_i, x^*_i and w_i.

probability $1 - \phi_i$; this is what Korb et al. (2004) call the degree of "effectiveness" of the intervention. We can associate a latent binary variable R^n_i to represent whether or not the intervention succeedeed or failed in case n, resulting in the mixture model

$$\begin{aligned} p(X_i|X_{G_i}, I_i = 1, \theta_i) &= \sum_r p(R_i = r) p(X_i|X_{G_i}, I_i = 1, R_i = r, \theta_i) \\ &= \phi_i f_i(X_i|X_{G_i}, \theta^1_i) + (1 - \phi_i) f_i(X_i|X_{G_i}, \theta^0_i). \end{aligned} \tag{6}$$

Figure 5 illustrates the idea of the unreliable intervention model. Although Figure 5 adds the indicator R^n_i to the imperfect model only, any of the other models of intervention under discussion could be augmented with the unreliable assumption also. For example, an unreliable, but otherwise perfect, intervention is modeled by setting

$$p(X_i|X_{G_i}, I_i = 1, R_i = 1, \theta, G) = \delta(X_i - x^*_i). \tag{7}$$

Unfortunately, computing the exact marginal likelihood of the data now becomes exponential in the number of R variables, because we have to sum over all latent assignments. Consequently we will not consider this model in this paper.

2.5 Soft interventions

Another way to model imperfect interventions is as "soft" interventions, in which an intervention just increases the likelihood that a node enters its target state x^*_i. Markowetz et al. (2005) suggest using the same model of $p(X_i|X_{G_i}, I_i, \theta_i)$ as before, but now the parameters θ^0_i and θ^1_i have *dependent* hyper-parameters. In particular, for the multinomial-Dirichlet case, $\theta^{0/1}_{ij\cdot} \sim Dir(\alpha^{0/1}_{ij\cdot})$, they assume the deterministic relation $\alpha^1_{ij\cdot} = \alpha^0_{ij\cdot} + w_i \vec{e}_t$, where j indexes states (conditioning cases) of x_{G_i}, $t = x^*_i$ is the target value for node i, $\vec{e}_t = (0, \ldots, 0, 1, 0, \ldots, 0)$ with a 1 in the t'th position, and w_i is the strength of the intervention. In other words, we "tilt" the Dirichlet distribution in the direction of state x^*_i, but don't force it to be a delta function. As $w_i \to \infty$, this becomes a perfect intervention, while if $w_i = 0$ it reduces to an imperfect intervention (where there are no "target states"). If the intervention strength w_i is unknown, Markowetz et al. (2005) suggest putting a mixture model on w_i, but it may be more appropriate to use the R_i mixture model mentioned in Section 2.4, where an intervention can succeed or fail on a case by case basis. Figure 6 shows the model graphically using plate notation.

2.6 Uncertain interventions

Finally we come to our proposed model for representing interventions with uncertain targets, as well as uncertain effects. We no longer assume a one to one correspondence between intervention nodes I_i and backbone nodes X_i. Instead, we assume that each intervention node I_i may have multiple backbone children. (Such interventions are sometimes said to be due to a "fat hand", which "touches" many variables at once.) If a backbone node has multiple intervention parents, we create a new parameter vector for each possible combination of intervention parents: see Figure 1(c) for an example.

We are interested in learning the connections from the intervention nodes to the backbone nodes, as well as between the backbone nodes. We do not allow connections between the intervention nodes, or from the backbone nodes back to the intervention nodes, since we assume the intervention nodes are exogenous and fixed. We enforce these constraints by using a two layered graph structure, $V = \mathcal{X} \cup \mathcal{I}$, where $\mathcal{X}$ are the backbone nodes and $\mathcal{I}$ are the intervention nodes. The addition of $\mathcal{I}$ motivates new notation, since the augmented adjacency matrix has a special block structure. The full adjacency matrix, denoted by H, is comprised of the *intervention* block F containing $\mathcal{I}$ nodes, and the *backbone* block G comprised of $\mathcal{X}$ nodes:

$$H = \begin{pmatrix} G & 0 \\ F & 0 \end{pmatrix}$$

where G is a $d \times d$ binary matrix and F is a $e \times d$ binary matrix, where d is the number of backbone nodes and e is the number of intervention nodes. We call the elements of F "target edges" since they correspond to edges $\mathcal{I} \rightarrow \mathcal{X}$ and the elements of G "backbone edges".

To explain how we modify the marginal likelihood function, we need some more notation. Let X_{G_i} be the backbone parents of node i, and I_{F_i} be the intervention parents. Let θ_i^ℓ be the parameters for node i given that its intervention parents have state ℓ. Then the marginal likelihood for a family becomes

$$p(x_i^{1:N}|x_{G_i}^{1:N}, I_{G_i}^{1:N}) \quad = \quad \prod_{\ell \in \mathrm{states}(I_{F_i})} \int \left[\prod_{n:I_{G_i}^n = \ell} p(x_i^n | x_{G_i}^n, \theta_i^\ell) \right] p(\theta_i^\ell) d\theta_i^\ell. \tag{8}$$

This is just the obvious extension of Equation 4 to multiple intervention parents. Note that we have a different parameter vector for each value of I_{F_i}, regardless of whether the CPD for X_i is tabular or not. Of course, we are free to choose a more parsimonious model of how interventions change their target distributions.

The question of whether or not the targets of intervention are known is orthogonal to the question of what effect the interventions have on their targets. Ignoring soft and unreliable interventions, we can identify four combinations, as shown in Table 1. The standard assumption is that there is a known one-to-one correspondence between intervention nodes and targets. There could be fewer intervention nodes than backbone nodes, but without

	perfect	imperfect
known	$F = I$,$\theta^1 = \delta$	$F = I$
unknown	$\theta^1 = \delta$	-

Table 1: **Four types of intervention model**, corresponding to different assumptions about intervention edge topology F (rows) and interventional parameters θ^1 (columns).

loss of generality, we can assume there is an equal number, and hence $F = I_d$, the $d \times d$ identity matrix. The assumption of perfect interventions is that $p(X_i|X_{G_i}, I_i = 1)$ is a delta function; we denote this event by $\theta^1 = \delta$. If we assume interventions are perfect, then we can have at most one intervention parent for each backbone node (unless we modify the definition of perfect intervention to allow resolution of conflicting target states). We can combine the perfect intervention assumption with the unknown target assumption, but doing so seems a little unnatural. Hence we will mostly focus on just three combinations: known targets + perfect interventions, known targets + imperfect interventions, and unknown targets + imperfect interventions; we will call these "perfect", "imperfect" and "uncertain" for brevity.

2.7 Interventional vs conditional density models

One might ask what the difference is between a conditional density model and an interventional model. The crucial difference is that in the latter, we assume that the interventions have local (albeit unknown) effects. If they did not, we would no be able to pool the data sampled from the different conditional distributions (Eberhardt, 2006). To see this, suppose, for simplicity, we have a single intervention node I_i. If all the backbone nodes depend on I_i, then we have a mixture of belief networks (sometimes called a "Bayesian multinet" (Thiesson et al., 1998; Bilmes, 2000)); but if only some nodes depend on I_i, we get a "factored mixture", in which only some components of the density depend on the mixing indicator, the rest being invariant across conditions.

We can see this difference in more detail by examining the marginal likelihood for the two proposed models (Tian and Pearl, 2001a). Suppose we observe N_0 cases in which $I_i = 0$ and N_1 cases in which $I_i = 1$. Let N^0_{ijk} be the counts in the first batch, N^1_{ijk} be the counts in the second batch, and $N_{ijk} = N^0_{ijk} + N^1_{ijk}$. If the post interventional distribution is unconstrained (i.e., the parameters that generated the second batch of data are unrelated to the first set of parameters), then we get

$$\begin{aligned}
&p(X_{1:N_1}, X_{N_1+1:N_2} | I_i^{1:N_1} = 0, I_i^{N_1+1:N_2} = 1, G) \\
&= \prod_{i=1}^{d} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N^0_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N^0_{ijk})}{\Gamma(\alpha_{ijk})} \\
&\times \prod_{i=1}^{d} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N^1_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N^1_{ijk})}{\Gamma(\alpha_{ijk})},
\end{aligned}$$

which is a product of two regular BDe (Bayesian Dirichlet likelihood equivalent) scores. Thus the data from the two regimes is not combined. But if we constrain the intervention to only affect node ℓ, we get

$$\begin{aligned}
&p(X_{1:N_1}, X_{N_1+1:N_2} | I_i^{1:N_1} = 0, I_i^{N_1+1:N_2} = 1, G, ch(I_i) = \ell) \\
&= \prod_{i \neq \ell} \prod_{j=1}^{q_i} \frac{\Gamma(\alpha_{ij})}{\Gamma(\alpha_{ij} + N_{ij})} \prod_{k=1}^{r_i} \frac{\Gamma(\alpha_{ijk} + N_{ijk})}{\Gamma(\alpha_{ijk})} \\
&\quad \times \prod_{j=1}^{q_\ell} \frac{\Gamma(\alpha_{\ell,j})}{\Gamma(\alpha_{\ell,j} + N^0_{\ell,j})} \prod_{k=1}^{r_\ell} \frac{\Gamma(\alpha_{\ell,j,k} + N^0_{\ell,j,k})}{\Gamma(\alpha_{\ell,j,k})} \\
&\quad \times \prod_{j=1}^{q_\ell} \frac{\Gamma(\alpha_{\ell,j})}{\Gamma(\alpha_{\ell,j} + N^1_{\ell,j})} \prod_{k=1}^{r_\ell} \frac{\Gamma(\alpha_{\ell,j,k} + N^1_{\ell,j,k})}{\Gamma(\alpha_{\ell,j,k})}.
\end{aligned}$$

where the first line pools the data for all $i \neq \ell$, and the second and third lines partition the data for node ℓ. It is this difference in likelihood that lets us distinguish local targets of intervention from global targets.

3. Algorithms for structure learning

In the previous section, we defined a variety of possible models for interventional data. In this section, we briefly discuss relevant computational issues. The most important question is: what are we trying to infer? One possible goal is to compute the most probable backbone graph, assuming the targets of intervention are known and are perfect:

$$G_{CMAP} = \arg\max_G p(X^{1:N} | I^{1:N}, G, F = I, \theta^1 = \delta) p(G) \tag{9}$$

where CMAP stands for conditional MAP (since we condition on F and θ^1). Another option is to learn the most probable backbone graph and the most probable set of intervention targets:

$$H_{MAP} = \arg\max_{G,F} p(X^{1:N} | I^{1:N}, G, F) p(G, F) \tag{10}$$

In some cases, such as drug target discovery, only the targets of intervention are of interest, and G is a nuisance variable, so we can compute

$$F_{MMAP} = \arg\max_F \sum_G p(X^{1:N} | I^{1:N}, G, F) p(G, F) \tag{11}$$

where MMAP stands for marginal MAP. However, since max-sum-product problems are typically harder than pure sum-product or max-product (Park and Darwiche, 2004), we follow the common practice of approximate this by maximizing over G instead of summing over G.

In cases where the sample size N is low, there may be considerable uncertainty about the MAP model. A more robust approach is to use Bayes model averaging (BMA) to compute the posterior over various features of interest, such as the existence of edges (Friedman and Koller, 2003). For example, we might compute edge marginals assuming the targets of intervention are known and perfect

$$p(G_{ij} = 1|D, F = I, \theta^1 = \delta) = \sum_G I(G(i,j) = 1)p(G|D) \tag{12}$$

where $D = (X^{1:N}, I^{1:N})$ is the data. Or we might compute edge marginals in the expanded graph, where we don't assume the targets of intervention are known:

$$p(H_{ij} = 1|D) = \sum_H I(H(i,j) = 1)p(H|D) \tag{13}$$

We can compute G_{CMAP}[3] and $p(G_{ij} = 1|D, F = I, \theta^1 = \delta)$ exactly[4] using dynamic programming (DP) in $O(d2^d)$ time and space (Koivisto and Sood, 2004; Koivisto, 2006; Silander and Myllmaki, 2006; Singh and Moore, 2005). If the interventions are unknown, we can compute H_{MAP} and $p(H_{ij} = 1|D)$ in $O((d+e)2^{d+e})$ time, where e is the number of intervention nodes. If we restrict each regular node to have at most one intervention parent, the cost becomes $O((d+e)2^d)$, since we don't have to search over subsets of intervention parents. Thus, assuming $e = O(d)$, the total cost is $O(2d2^d)$, so we see that structure learning with uncertain interventions is only a factor of two slower than structure learning assuming known intervention targets.

The dynamic programming algorithms are only practical for $d \leq 22$ nodes (unless we have additional prior knowledge, such as layering constraints (Koivisto and Sood, 2004)). To scale up, we need to resort to heuristic local search methods (for the MAP problem) or MCMC methods (for the BMA problem). See e.g., (Heckerman et al., 1995; Friedman and Koller, 2003) for details. Since these algorithms are not guaranteed to find the exact answer, we will represent their output by $\hat{G}_{CMAP}$, $\hat{p}(G_{ij} = 1|D, F = I, \theta^1 = \delta)$, etc.

The two-layered nature of our structure learning problem suggests the following iterative algorithm for estimating $\hat{H}_{MAP}$: first estimate F (the targets of intervention), and use this to partition the data; then estimate G (the backbone) given this partitioning; and iterate until convergence. This is appealing because we can estimate G_{CMAP} using standard software, even if it cannot handle interventions. Unfortunately, preliminary experiments suggest this method is more prone to getting stuck in local maxima than searching through H-space directly. Hence when we cannot afford to use DP, we resort to standard local search techniques on the two-layer graph. In Section 4.2, we will show empirically that local search does a good at finding posterior modes.

4. Experimental results

We first present some results on synthetic data generated from two belief networks of known structure, and then present results on two real biological datasets. The advantage of the

[3] Note that there may be more than one MAP-optimal DAG, but they are all Markov equivalent, sothe algorithm just returns one of these.

[4] Although the BMA computations are exact, they require a rather unnatural (and highly non-uniform) "modular" prior on graph structures $p(G)$. See (Eaton and Murphy, 2007b) for a way to use the Metropolis-Hastings algorithm to correct for this bias. The MAP computations can use a uniform graph prior, $p(G) \propto 1$, and do not suffer from this bias. We have not found it necessary to include an explicit complexity-penalizing term in $p(G)$, although this is of course possible.

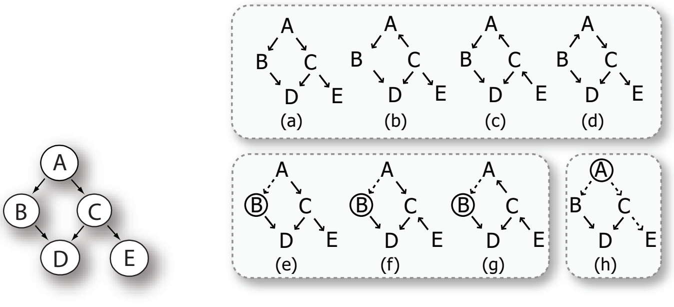


Figure 7: **Cancer network.** Left: The cancer network, from (Friedman et al., 1998). Right: (a-d) are Markov equivalent. (c-g) are equivalent under an intervention on B. (h) is the unique member under an intervention on A. Based on (Tian and Pearl, 2001a).

synthetic networks is that we can easily assess the ability of different algorithms/ models to recover the true (generating) structure. In Sections 4.1 and 4.2, we restrict ourselves to small models (up to 12 nodes), so that we can compare to the results of using exact inference algorithms. We sample data from the known model assuming perfect interventions, and then try to recover the graph, either using the knowledge that the interventions were perfect, or not using such knowledge. We show that the latter does almost as well as the former. (If we generate data from imperfect interventions, the perfect model cannot fit the data, since it assumes intervened upon nodes act deterministically.)

In Section 4.3, we compute G_{CMAP} and H_{MAP} using some data that measures the phosphorylation levels of 11 proteins in part of the T-cell signalling pathway. We show that H_{MAP} has much better predictive abilities than G_{CMAP}, indicating that the perfect intervention assumption is not warranted in this case. Finally, in Section 4.4, we compute $\hat{H}_{MAP}$ using some microarray data that measures expression levels of 271 genes, from cells with 7 different types of cancer. Each cancer type causes a different primary gene to mutate, which in turn causes downstream changes in other genes. We show that we are able to identify the primary affected genes.

4.1 Synthetic "cancer" network

The ability to recover the true causal structure (assuming no latent variables) using perfect and imperfect interventions has already been demonstrated both theoretically (Eberhardt et al., 2005, 2006; Tian and Pearl, 2001b,a) and empirically (Cooper and Yoo, 1999; Markowetz and Spang, 2003; Tian and Pearl, 2001b,a; Werhli et al., 2006). Specifically, each intervention determines the direction of the edges between the intervened nodes and its neighbors; this in turn may result in the direction of other edges being "compelled" (Chickering, 1995).

For example, in Figure 7, we see that there are 4 graphs that are Markov equivalent to the true structure; given observational data alone, this is all we can infer. However, given enough interventions (perfect or imperfect) on B, we can eliminate the fourth graph (d), since it has the wrong parents for B. Given enough interventions on A, we can uniquely identify the graph, since we can identify the arcs out of A by intervention, the arcs into D since it is a v-structure, and the $C \rightarrow E$ arc since it is compelled. (We cannot orient the arc as $E \rightarrow C$ without creating a new v-structure.) In general, given a set of interventions and observational data, we can identify a graph up to intervention equivalence (see (Tian and Pearl, 2001b) for a precise definition).

In this section, we experimentally study the question of whether one can still learn the true structure, even when the targets of intervention are a priori unknown, and if so, how much more data one needs compared to the case where the intervention targets are known. We assessed this using the following experimental protocol. We take the 5 node "cancer" network shown in Figure 7, and generated random multinomial CPDs by sampling from a Dirichlet distribution with hyper-parameters chosen by the method described in (Chickering and Meek, 2002), which ensures strong dependencies between the nodes.[5] (Stronger dependencies increase the likelihood that the distribution will be numerically faithful to the conditional independency assumptions encoded by the structure.) For simplicity, we used binary nodes. We then generated data using forwards sampling; the first 1000 cases D_0 were from the original model, the second 1000 cases D_1 from a "mutated" model, in which we performed a perfect intervention either on A or B, forcing it to the "off" state in each case.

Next we tried to learn back the structure using varying sample sizes of $N \in \{10, 25, 250, 1000\}$. Specifically we used N observational samples and N interventional samples, $D = (D_0^{1:N}, D_1^{1:N})$. We used exact BMA to compute edge marginals under increasingly weak assumptions: (1) using the perfect interventions model; (2) using the soft interventions model[6]; (3) using the imperfect model; and (4) using the uncertain interventions model. In the latter case, we also learned the children of the intervention node. As a control, we also tried just using observational data, $D = D_0^{1:2N}$ (equivalent to clamping $I_i^n = 0$ for all i and n).

Our results for the perfect and uncertain models are shown in Figure 8. (On this network, the imperfect and soft intervention models perform very similarly to the perfect case, though they require more data to achieve the same accuracy; hence, for brevity, we only show the results of perfect interventions.) We see that with observational data alone, we are only able to recover the v-structure $B \rightarrow D \leftarrow C$, with the directions of the other arcs being uncertain (e.g., $P(C \rightarrow E) \approx 0.75$, since 3 out of 4 Markov equivalence classes have the edge oriented in this way). With perfect interventions on B, we can additionally recover the $A \rightarrow B$ arc, and with perfect interventions on A, we can recover the graph uniquely, consistent with the theoretical results mentioned above.

With uncertain interventions, we see that there is more uncertainty about the graph structure, but this uncertainty reduces with sample size, and eventually the posterior converges to a delta function on the intervention equivalence class. We also see that we need less data to recover the target edges than the backbone edges (the I^* rows are less entropic

[5] The method of (Chickering and Meek, 2002) works as follows. Consider a node i with 3 states and 4 parent states. We pick a "basis vector" $(1, 1/2, 1/3)$, and then, for the j'th parent state, we sample $\theta_{ij\cdot} \sim Dir(s\alpha_{ij\cdot})$, where $\alpha_{i1\cdot} \propto (1, 1/2, 1/3)$, $\alpha_{i2\cdot} \propto (1/2, 1/3, 1)$, $\alpha_{i3\cdot} \propto (1/3, 1, 1/2)$, $\alpha_{i4\cdot} \propto (1, 1/2, 1/3)$, and $s = 10$ is an effective sample size. Obviously other methods of generating CPDs with strong dependencies are possible.

[6] Markowetz et al. (2005) do not discuss how to set the pushing strength w_i. We set it equal to $0.5N$, so that the data does not overwhelm the hyper-parameter α_{ijk}^1.

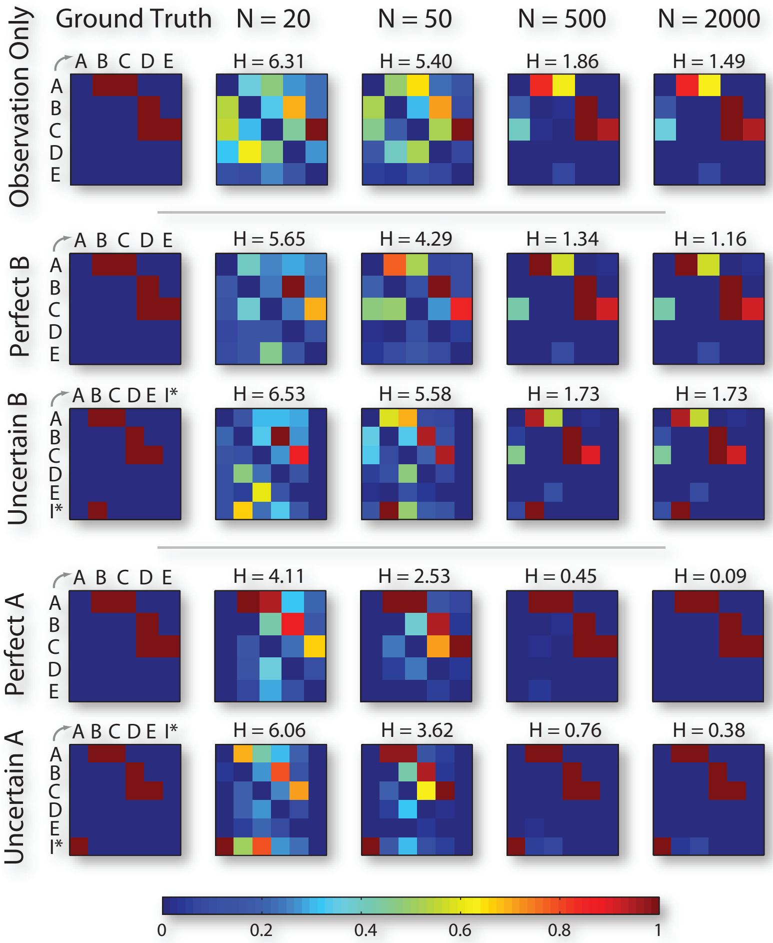


Figure 8: **Results of structure learning on the cancer network**. Left column: ground truth. Subsequent columns: posterior edge probabilities for increasing sample sizes N, where dark red denotes 1.0 and dark blue denotes 0.0. I^* denotes the intervention node. For the rows labeled "perfect", we show $p(G_{ij} = 1|D, F)$, whereas for the rows labeled "uncertain", we show $p(H_{ij} = 1|D)$. The titles of the form $H = x$ means the entropy of the factored posterior $\prod_{ij} p(G_{ij}|D, F)$ is x. (In the uncertain case, we ignore the uncertainty in the interventional edge when computing the entropy, to make comparisons fair.) This figure is best viewed in colour.

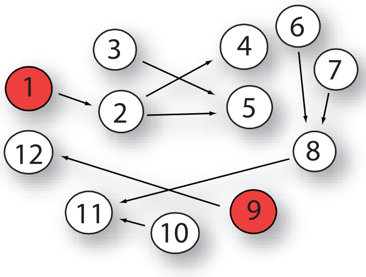


Figure 9: **Car diagnosis network**, introduced in (Heckerman et al., 1994). By selecting the appropriate two intervention nodes, marked here in red, it is possible to uniquely recover the structure.

than the other rows at small sample sizes), although this might be an artefact of using perfect interventions in the generating model. We obtain similar results with other experiments on random graphs. This suggests that we do not lose much in terms of statistical efficiency by learning the targets of intervention, rather than assuming they are known and perfect. In Section 4.3, we will show that the weaker assumption actually works better on a particular biological dataset.

4.2 Synthetic "cars" network

In this section, we perform a more quantitative analysis of the ability to learn structure using a slightly larger (12 node) network. In addition, we compare the performance of exact methods for computing H_{MAP} with local search methods for approximating $\hat{H}_{MAP}$.

We consider the synthetic "Car Diagnosis" network, shown in Figure 9. Again, we assign multinomial CPDs according to the method of (Chickering and Meek, 2002). Without interventional data, it would be impossible to learn the orientation of some edges; however, if we intervene on nodes 1 and 9, it is possible to uniquely recover the original structure. We sampled 10 data sets, each of size $N \in \{20, 200, 2000\}$. For a given N, half of the data is observational and half interventional, with interventions generated according to the perfect model. Next, we attempted to learn the structure back from each data set, using known or unknown targets of intervention.

Following Koivisto (2006); Husmeier (2003), we summarize performance in terms of ROC curves applied to the exact BMA marginals $p(G_{ij} = 1|D, F = I, \theta^1 = \delta)$ assuming known perfect interventions and $p(H_{ij} = 1|D)$ using learned targets. See Figure 10. (The results of known imperfect interventions are similar to known perfect interventions, so are omitted.) We see that we recover all the edges as the sample size increases, and that the performance using uncertain interventions is only slightly worse than assuming perfect interventions. We also see that we need less data to recover the target edges than the backbone edges (the mean AUC for the target edges is always higher than for the backbone edges).

Next we wish to compare exact and approximate methods for finding H_{MAP}. The exact method uses the dynamic programming algorithm of (Silander and Myllmaki, 2006); the

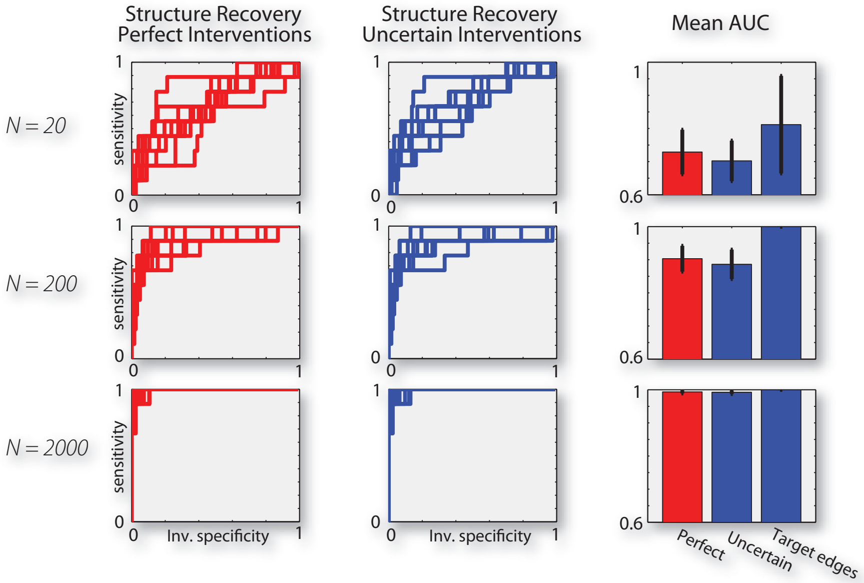


Figure 10: **Performance on cars network of BMA**. Each row denotes a different sample size ($N = 200$ means there were 100 observational and 100 interventional data cases). The first two columns contain ROC curves; there are 10 curves per plot, corresponding to the 10 datasets generated per sample size. In the right column, the ROC curves have been summarized using area under the curve. The column labeled "perfect" is measuring the performance of estimating $p(G_{ij} = 1|D, F = I, \theta^1 = \delta)$; the column labeled "uncertain" is measuring the performance of estimating $p(G_{ij} = 1|D)$; and the column labeled "target edges" is measuring the performance of estimating $p(F_{ij} = 1|D)$;

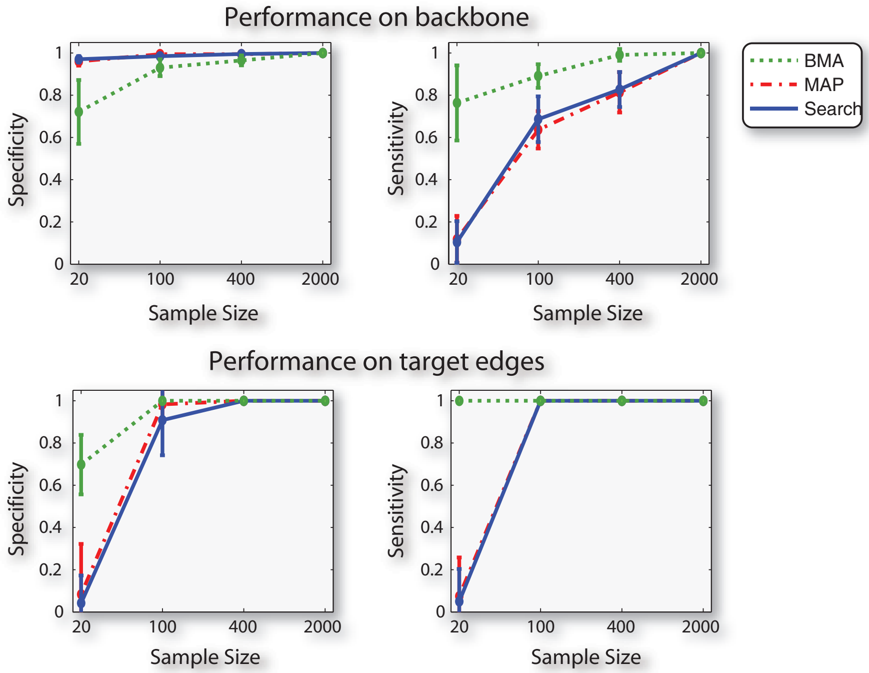
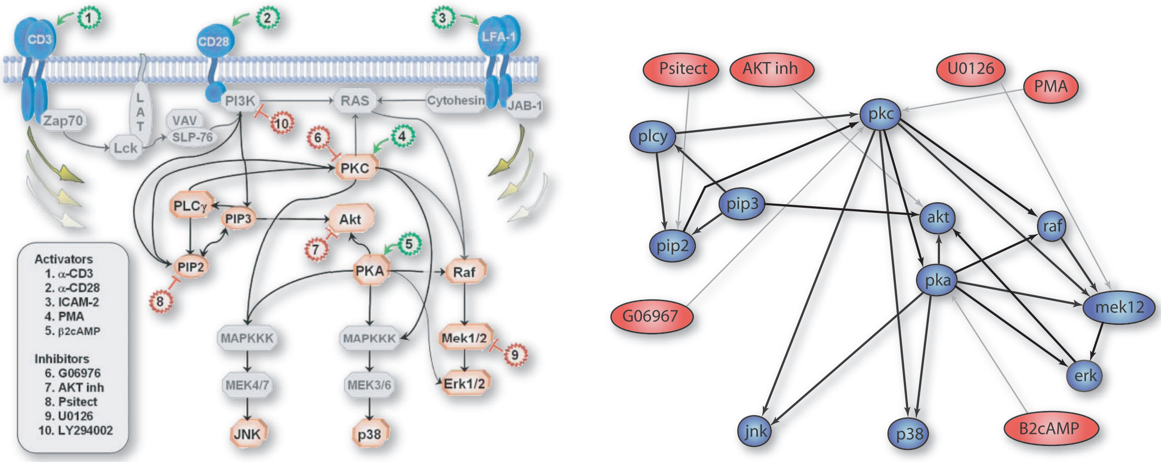


Figure 11: **Performance on cars network of point estimates**. "BMA" refers to $p(H_{ij} = 1|D) > \theta$, where θ is the (approximate) equal error rate threshold; "MAP" refers to H_{MAP} computed using DP; "search" refers to $\hat{H}_{MAP}$ computed using multiple-restart hill-climbing.

approximate method uses multiple-restart hill-climbing (Heckerman et al., 1995). Local search was allowed to run for 20 seconds (3 times as long as it took to compute the exact H_{MAP}).

Since we cannot compute a ROC curve from a point estimate H_{MAP}, we instead compute a single value for the sensitivity and specifity of the resulting estimated structure. We also compare to using exact BMA, and evaluate the point estimate $\hat{H} = I(p(H_{ij} = 1|D) > \theta)$, where θ is the equal error rate threshold.[7] The results are shown in Figure 11. We see that for low sample sizes, BMA works better, but eventually all techniques converge to the true structure, including local search.



(a) (b)

Figure 12: **Ground truth model of T-cell signaling pathway**. (a) biologically realistic model with latent variables. The small round circles with numbers represent various interventions (green = activators, red = inhibitors). From (Sachs et al., 2005). Reprinted with permission from AAAS. (b) the projection of this onto the 11 backbone variables (circles) and 6 intervention variables (ovals) suggested in (Sachs et al., 2005). Intervention edges are in light gray. This figure is best viewed in colour.

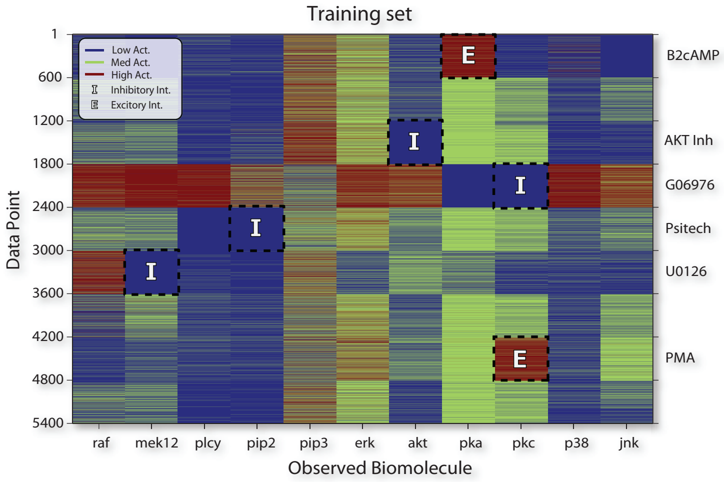
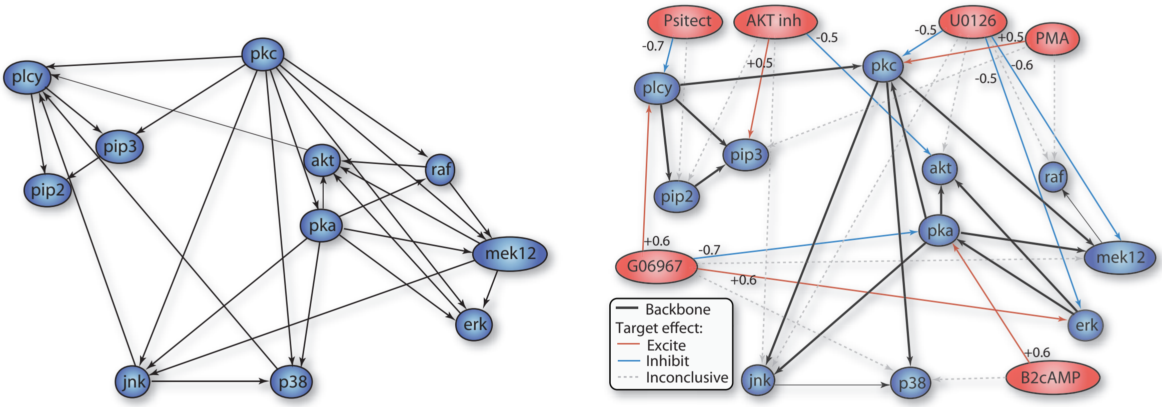


Figure 13: **T-cell data.** 3-state training data from (Sachs et al., 2005). Columns are the 11 measured proteins, rows are the 9 experimental conditions, 3 of which are "general stimulation" rather than specific interventions. The name of the chemical that was added in each case is shown on the right. The intended primary target is indicated by an E (for excitation) or I (for inhibition). There are 600 measurements per condition. This figure is best viewed in colour.

4.3 T-cell data

We now analyze an interesting dataset from (Sachs et al., 2005). This consists of 11 protein concentration levels measured under 6 different interventions, plus 3 unperturbed conditions. The proteins in question constitute part of a biochemical signaling network of human T-cells, and therefore play a vital role in the immune system. See Figure 12(a) for the "ground truth" network (derived after many years of experimental research). Since the ground truth network contains unmeasured variables we cannot expect any belief net learning algorithm to recover this structure exactly unless we devise a way to uncover hidden variables (which is beyond the scope of this paper). The best we can hope for is to recover the closest "projection" of the true structure onto the space of fully observed belief nets. The model proposed in (Sachs et al., 2005) as "ground truth DAG" is shown in Figure 12(b). It is clear that this does not capture all of the relevant biology. For example, in the true model, there is a bidirectional edge between PIP2 and PIP3, whereas this is disallowed in a DAG. Also, in the DAG, there is an edge from ERK to AKT, which does not seem to be present in the

[7] The equal error rate is the threshold at which sensitivity(θ) = specifity(θ) (or equivalently, false negative rate equals false positive rate), which can be found graphically by intersecting the ROC curve with a diagonal line from the top left to the bottom right. In practice, the ROC curve may not intersect this curve exactly, so instead we choose the threshold such that sensitivity(θ) + specifity(θ) is maximized. Werhli et al. (2006) proposed an alternative way to select the threshold, namely choosing θ such that the number of false positives is fixed (they used 5). Note that the full BMA ROC curve (using all thresholds) is shown in Figure 10.



(a) (b)

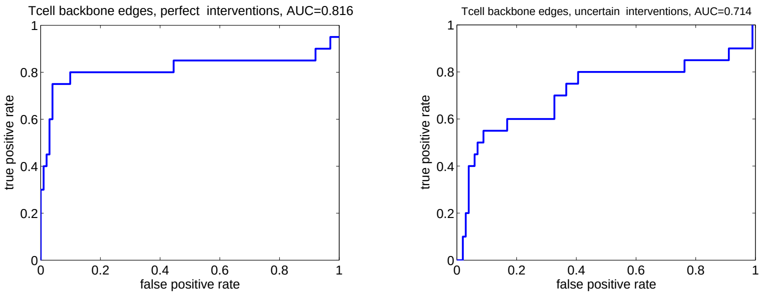
Figure 14: **MAP structures learned on T-cell data**. (a) G_{MAP} assuming known perfect interventions. (b) H_{MAP} assuming unknown imperfect interventions. Red nodes are intervention nodes. Red edges are excitory, blue edges are inhibitory, dotted gray edges have ambiguous polarity. Numbers represent the probability the target is excited/ inhibited. Backbones edges are solid black.

standard biological model (they cite (Fukuda et al., 2003) as evidence for the existence of this arc). In view of this, in addition to comparing to this "ground truth DAG", we will look at alternative measures of performance.

The data in question were gathered using a technique called flow cytometry, which can record phosphorylation levels of proteins in individual cells. This has two advantages compared to other measurement techniques, such as microarrays: first, it avoids the information loss commonly incurred by averaging over ensembles of cells; second, it creates relatively large sample sizes (600 cells under 9 conditions yielding $N = 5400$ data points). The raw data, which is available online[8] was discretized into 3 states, representing low, medium and high activity, using a technique described in Hartemink (2001). We obtained this discretized data directly from Karen Sachs; see Figure 13.

Various methods have been used to learn belief networks from this data, including multiple restart simulated annealing in the space of DAGs (Sachs et al., 2005), Metropolis-Hastings in the space of node orderings (Werhli et al., 2006), and equi-energy sampling in the space of node orderings (Ellis and Wong, 2006). All of these techniques assumed perfect interventions.

[8] See `http://www.sciencemag.org/cgi/content/full/sci;308/5721/523/DC1`.



(a) (b)

Figure 15: **ROC curves for the T-cell dataset**. (a) $P(G_{ij} = 1|D, F = I, \theta^1 = \delta)$ assuming known perfect interventions, AUC = 0.816. (b) $P(G_{ij} = 1|D)$ assuming uncertain, imperfect interventions, AUC = 0.714. (We exclude the intervention edges from the ROC plot and AUC computation, in order to make the results comparable.)

In Figure 14(a) we show G_{MAP} computed exactly using DP assuming perfect interventions.[9] The corresponding ROC plot for the BMA estimates are in Figure 15(a). We see that the method does a good job at recovering most of the ground truth DAG.[10]

In Figure 14(b) we show H_{MAP} computed exactly using DP assuming uncertain interventions. (Again, thresholding the BMA edge marginals results in an identical structure.) The corresponding ROC plot for the BMA estimates are in Figure 15(b); we see a slight drop in performance, consistent with the synthetic results in previous sections. However, by relaxing the perfect intervention assumption, we learn something interesting. While we see that we are able to learn the known targets of all but one of the 6 interventions (we missed the G06967 → pkc edge), we also found that the interventions have multiple children, even though they were designed to target specific proteins, i.e., they have "side effects". Upon further investigation, we found that each intervention typically affected a node and some of its immediate neighbors. For example, from the ground truth network in Figure 12(a), we see that Psitect (designated 8 in that figure) is known to inhibit pip2; in our learned

[9] Note that thresholding the edge marginals at almost any value in the range $0 < \theta < 1$ results in an identical graph structure to G_{MAP}, suggesting that the posterior is very "peaky". Indeed, if we plot the edge marginals as a "heat map", we see that they are strongly concentrated around 0 or 1 (data not shown). We verified that this result was not an artefact of the modular prior required by the DP algorithm by using a Metropolis-Hastings method (Eaton and Murphy, 2007b) with a DP proposal to compute BMA under a uniform graph prior $p(G) \propto 1$; this gave essentially identical results. This suggests that $N = 5400$ is sufficiently large to ensure the posterior is dominated by a single peak (or multiple intervention equivalent peaks), and that vaguer edge marginals (such as those obtained in (Sachs et al., 2005)) may be the result of poor MCMC convergence.

[10] In Werhli et al. (2006), they obtained an AUC of 0.7. However, their experimental setting differs in several respects: they only use 100 samples randomly chosen from the full 5400; they process the original continuous data by quantile normalization, ensuring all the marginals are Gaussian, rather than discretization; they use the BGe score (Geiger and Heckerman, 1994) (corresponding to a normal-normal-Gamma model) instead of BDe (corresponding to a multinomial-Dirichlet model); and they used MCMC sampling in order space rather than exact inference based on DP.

network (Figure 14(b)), we see that Psitect connects to pip2, but also to plcy, which is a neighbor of pip2. This is biologically plausible, since although these compounds often have specific targets (and are chosen for this reason), there may be latent pathways which cause indirect changes in several visible variables. Also, although we missed the G06967 → pkc edge, the other children of G06967 (plcy, pka, mek12, erk and p38) seem to be strongly affected by G06967 when looking at the data in Figure 13.

In addition to determining the targets of intervention, it is interesting to determine the type of interaction between an intervention and its targets. Given H_{MAP}, we compute the posterior mean parameters of $p(X_i|X_{G_i}, I_i = 1)$, $\overline{\theta}^1_{ijk} = (\alpha^1_{ijk} + N^1_{ijk})/(\alpha^1_{ij} + N^1_{ij})$, for each child X_i and intervention parent I_i. We can then marginalize out X_{G_i} to get the marginal effect of the intervention on its target:

$$p(X_i|I_i = 1) = \frac{\sum_{X_{G_i}} p(X_i|X_{G_i}, I_i = 1) p_u(X_{G_i})}{\sum_{X_i, X_{G_i}} p(X_i|X_{G_i}, I_i = 1) p_u(X_{G_i})}. \tag{14}$$

where $p_u(X_{G_i})$ is a uniform prior over the parent states. If most of the mass resides in the "overexpressed" state, $p(X_i = +1|I_i = 1) \approx 1$, we say the intervention is excitory, while if $p(X_i = -1|I_i = 1) \approx 1$, we say it is inhibitory. If the distribution is uniform, the polarity of the edge is inconclusive. In Figure 14(b), we color code the excitatory edges in red, and the inhibitory edges in blue; we also show the marginal probability $p(X_i = \pm 1|I_i = 1)$. We see that, as expected, the edges from B2cAMP and PMA are excitatory and the edges from Psitech and U0126 are inhibitory; however, while the edges from AKTInh and G06976 are inibitory on some of their targets (as desired), they also seem to have excitatory side effects.

Since the ground truth DAG (Figure 12(b)) does not reflect biological reality (Figure 12(a)) particularly well (due to the absence of cycles, latent variables, etc.), merely quoting an AUC score is potentially misleading. An alternative way to assess performance is to sample from the learned model to see if the resulting data resembles the training data. (See (Gelman et al., 2004) for more sophisticated forms of posterior predictive model checking.) For Dirichlet-multinomial models, we can compute the posterior predictive density by plugging in the posterior mean parameters

$$\begin{aligned} p(X|I, D, H_{MAP}) &= \prod_{i=1}^{d} \int p(X_i|X_{G_i}, I_{F_i}, \theta_i) p(\theta_i|D) d\theta_i & (15) \\ &= \prod_{i=1}^{d} \int \prod_j \prod_k \theta_{ijk}^{I(X_i=k, Y_{H_i}=j)} Dir(\theta_{ijk}|\alpha_{ijk} + N_{ijk}) d\theta_{ijk} & (16) \\ &= \prod_{i=1}^{d} \prod_j \prod_k \overline{\theta}_{ijk}^{I(X_i=k, Y_{H_i}=j)} & (17) \end{aligned}$$

where $Y = (X_{1:d}, I_{1:e})$ are all the nodes (so $Y_{H_i} = (X_{G_i}, I_{F_i})$). We can then sample from this using forwards sampling, setting the intervention nodes in the same way as done for the training data. The result of sampling from G_{MAP} (which assumes perfect interventions) is shown in Figure 16(a); sampling from the ground truth DAG topology (but learning the parameters) gives very similar results (not shown). The result of sampling from H_{MAP}

(which learns the interventions) is shown in Figure 16(b). It is clear that the latter method learns a model that fits the data much better, e.g., compare the rows labeled “G06976” with Figure 13. In fact, we can see from the original data in Figure 13 that the interventions were not perfect e.g., see the row labeled “PMA”, which is not a constant block of +1’s (this is easier to see in colour) despite being the target of an excitatory intervention.[11]

One might worry that H_{MAP} is overfitting.[12] We can assess this by comparing the predictive log-likelihood in a cross-validation framework. Specifically, we evaluate the average log-likelihood across folds using a plug-in estimate of the structure. For the uncertain interventions we have

$$CV(H_{MAP}) = \sum_f \frac{1}{N_f} \sum_{n=1}^{N_f} \log p(X^n | I^n, D^{-f}, H_{MAP}) \tag{18}$$

where D^{-f} is the data excluding fold f. We can compute $CV(G_{MAP})$ similarly.[13] The results of this CV comparison, for the perfect, imperfect, soft and uncertain intervention models are shown in Figure 17. We see that the uncertain interventions model is a significantly better predictor of the response of the system to interventions. In the following section, we will show the results of extrapolating beyond the training set (predicting the response to interventions that have not been seen before).

4.4 Leukemia data

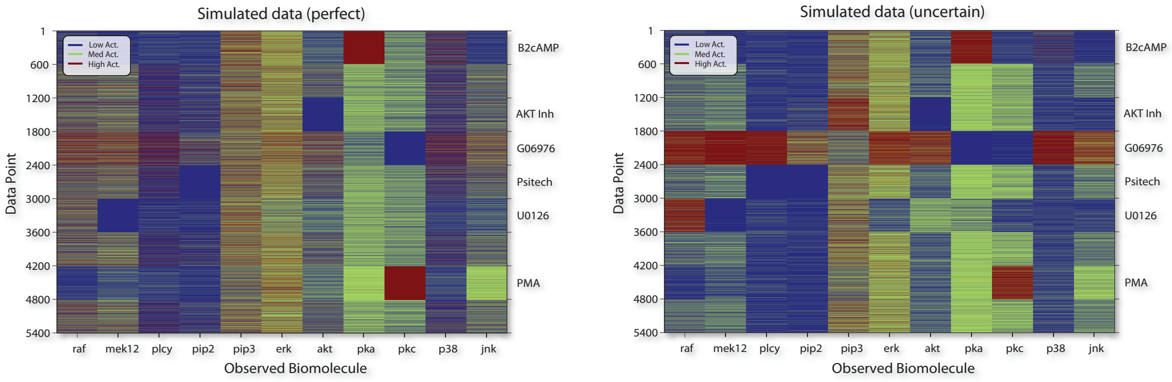
In this section we explore a promising application of the uncertain intervention model, namely that of drug/disease target discovery. In this setting, the primary goal is to infer the intervention edges; the backbone structure is merely used to “explain away” indirect changes. Although the backbone structure may be of interest, in many applications there will not be enough data to estimate it very reliably. However, as we saw in the synthetic data experiments, the amount of data needed to learn the intervention edges is usually much less.

Note that similar problems arise in various other settings, such as determining which genes are regulated by transcription factors. In general, the problem is to identify the “downward” edges in a directed, two-layer graph. The key to success is the assumption that the dependence between an intervention node and its true backbone target is stronger than the dependence between an intervention node and any other backbone target. One way to measure dependence is to use (conditional) mutual information, as in e.g., (Margolin

[11] The perfect intervention model excludes cases that were set by intervention, and hence does not detect this discrepancy. In Figure 17, to be described below, we show that the imperfect intervention model, which uses all the data, also does not fit the model as well as the uncertain intervention model, indicating that learning the topology of the intervention nodes, and not just allowing for stochastic effects, is important.

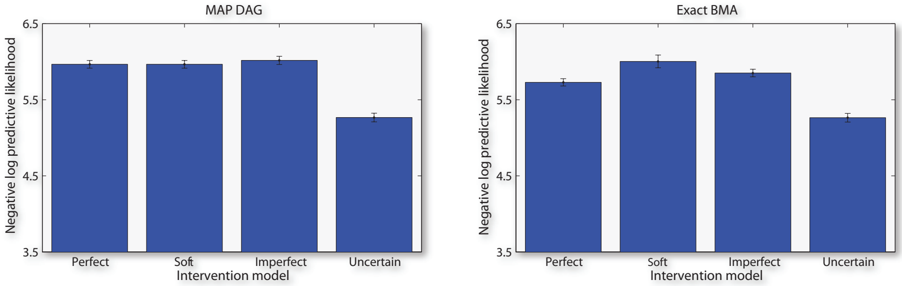
[12] In this domain, underfitting is also a potential problem, since if a model cannot capture important aspects of the data, we should be cautious drawing conclusions about its topology.

[13] We can also use the marginal likelihood $p(X^n|D) = p(X^n, D)/p(D)$ in the innermost sum above (using DP to marginalize out the graph structures) rather than using a plug-in estimate, but this is more computationally expensive. See (Eaton and Murphy, 2007b) for a cheap approximation that works better than a plug-in estimate but is faster than re-running DP for each X^n.



(a) (b)

Figure 16: **Data sampled from learned T-cell models.** (a) Data sample from G_{MAP}, learned using the perfect intervention assumption. (c) Data sample from H_{MAP}, using the learned targets of intervention. This figure is best viewed in colour.



(a) (b)

Figure 17: **Cross-validated negative log likelihoods on T-cell dataset.** Lower is better. (a) MAP plug-in. (b) Exact BMA. Result obtained across 10-fold validation.

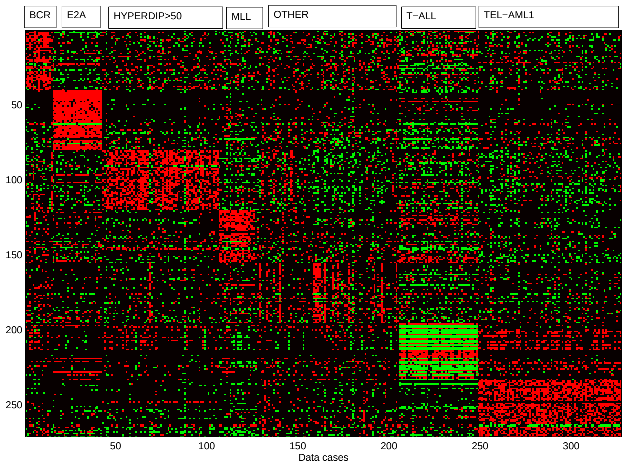


Figure 18: **ALL dataset.** This figure is best viewed in colour.

et al., 2006), but here we adopt a Bayesian approach. (See (Minka, 2003) for an interesting relationship between these two approaches.)

We analyze a dataset from (Yeoh et al., 2002) consisting of measurements of 12,000 genes from 327 humans suffering from different forms of acute lymphoblastic leukemia (ALL). ALL is a heterogeneous cancer, meaning that it is manifested by several subtypes that vary in their genetic cause and consequently in their response to treatment. The dataset of (Yeoh et al., 2002) contains 7 classes, called HYPERDIP $>$ 50, E2A-PBX1, BCR-ABL, TEL-AML1, MLL, T-ALL, and OTHER. It is believed that each of these subtypes is caused by a small number of gene mutations; in the case of E2A-PBX1 and BCR-ABL, the relevant genetic causes are known. We can think of each disease subtype as targeting a subset of genes, which in turn cause other genes to change their mRNA expression level. The goal is to identify the primary affected genes for each subtype, and to distinguish them from genes which only change indirectly.

Yeoh et al. (2002) used a chi-square-based filtering method to identify the 40 most differentially expressed genes in each subtype, yielding a total of 271 unique genes (9 were shared across subtypes). Their data is available online.[14] Dejori and Stetter (2004) discretized this data into 3 levels representing "underexpressed" (-1), "unchanged" (0) and "overexpressed" ($+1$) using the following technique: values less than $\mu_i - \sigma_i$ were mapped to -1, values greater than $\mu_i + \sigma_i$ were mapped to $+1$, and the remainder to 0 (where (μ_i, σ_i) are the mean and standard deviation of gene i). The resulting dataset is shown in Figure 18. We will use the same discretized data.

After discretization, Dejori and Stetter (2004) learned a belief network using simulated annealing search, ignoring the fact that the data samples come from different conditions (distributions). Using exact inference (state estimation), they then computed the posterior predictive distribution assuming that gene i is *observed* to be in its overexpressed or underexpressed state, $p(X|X_i = \pm 1, \hat{G}_{MAP}, \overline{\theta})$. (Note that this is not the same as setting gene i

[14] See `http://www.stjuderesearch.org/data/ALL1/all_datafiles.html`.

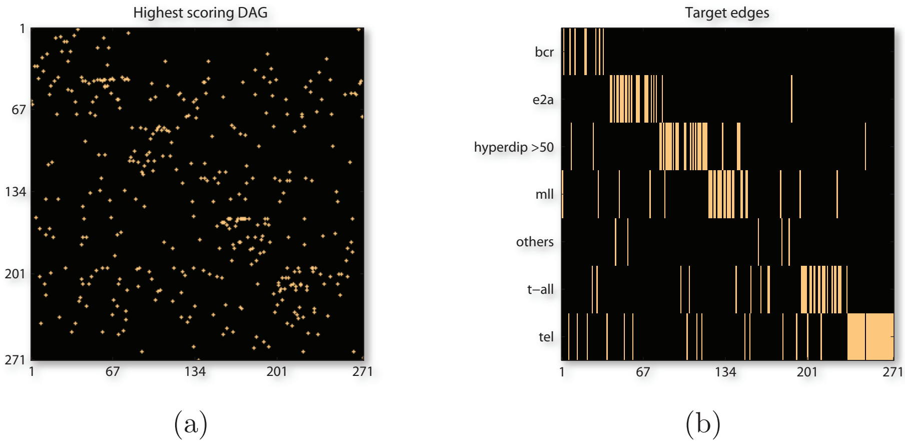


Figure 19: **Highest scoring DAG found on ALL dataset.** (a) DAG backbone, **G**, (b) corresponding target edges, **F**.

by intervention. Indeed, they make no attempt to interpret their learned structure causally, and simply use belief nets as a convenient density estimator for discrete data.) They then generated a sample of size 327 for each gene i, $D^*_i = \{x \sim p(X|X_i = \pm 1, \hat{G}_{MAP}, \overline{\theta})\}$, and compared the Euclidean distance between D^*_i and D_k for each $k = 1 : 7$, where D_k denote all the training data cases from subtype k:

$$\text{sim}(k, i) = \sum_{n=1}^{327} \exp(-||D_{k,n} - D^*_{i,n}||^2) \tag{19}$$

This gives them a "profile" over genes i for each subtype k; they then return the top 5 peaks as the most likely targets (causes) for that cancer subtype. (An alternative way to compute the profile would be to compute the likelihood of D_k under the conditional model $p(X|X_i = \pm 1, \hat{G}_{MAP}, \overline{\theta})$ using exact inference, thus bypassing the sampling and Euclidean similarity steps.)

There are several drawbacks to the above approach. Firstly, the real goal is to learn a causal model of the consequences of forcing a gene into a mutated state, rather than a conditional density model of the consequences of *observing* a gene in a mutated state. Second, in addition to estimating the graph structure, the above technique requires inference (state estimation) in the resulting graph to infer targets, which might be slow, particular for searching through subsets of multi-gene targets.

In our approach, we augment the 271 backbone variables with 7 binary intervention nodes encoding the presence or absence of the subtypes using a 1-of-7 encoding. We then learn $\hat{H}_{MAP}$ using multiple restart hill climbing. Figure 19 shows the best local maximum; the next-best structure had a score approximately e^{64} times lower, although it is structurally quite similar.

One way to assess the quality of our learned model is to sample from it, as described earlier. In Figure 20(a), we show samples drawn from $p(X|I, \hat{H}_{MAP}, \overline{\theta})$ where I is set in the same way as the training data. We see that we fit the data very well. Figure 20(b)

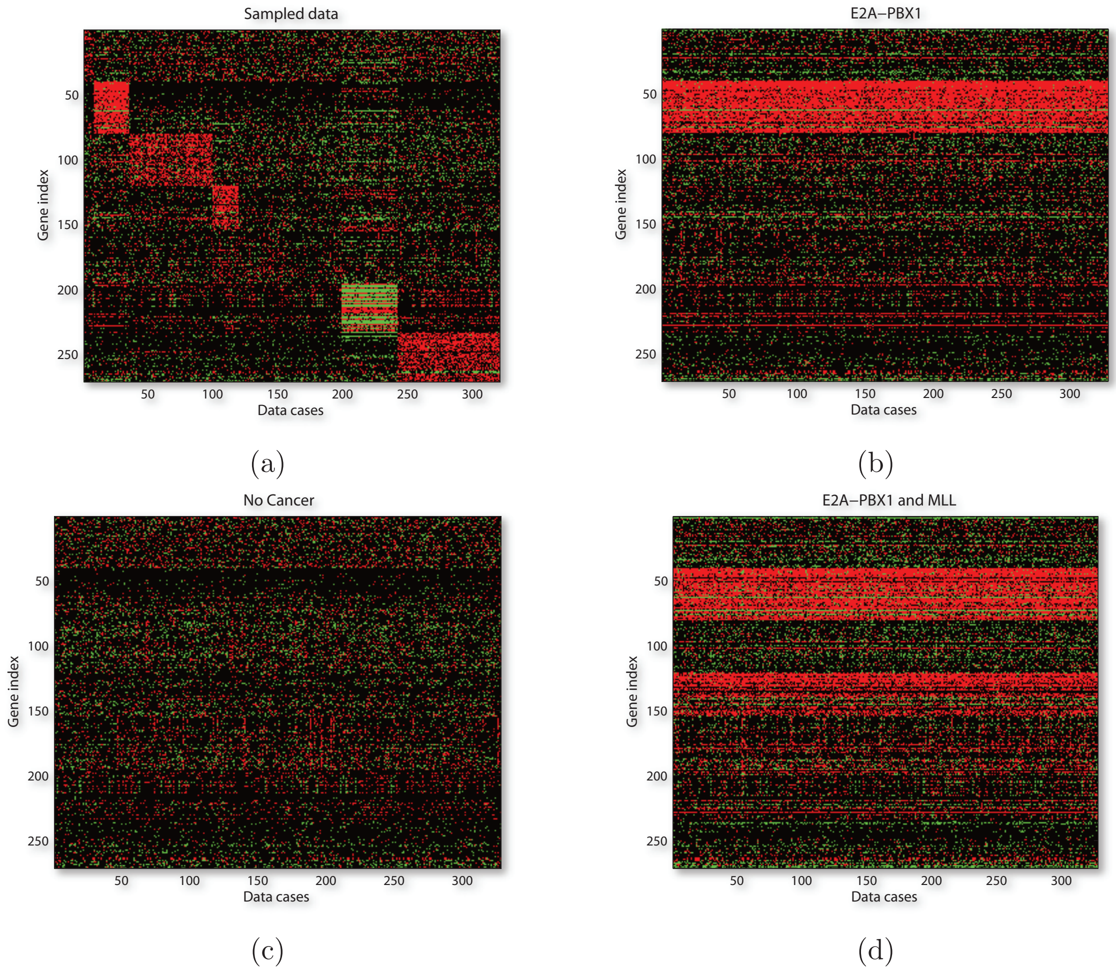


Figure 20: **Data sampled from the learned ALL model.** (a) Sampled under the same interventional conditions as the training data. (b) Sampled assuming E2A-BX1 is on. (c) Sampled assuming no interventions. (d) Sampled assuming E2A-BX1 and MLL are on.

shows samples where $I_{E2A} = 1$ and the other I nodes are off. However, we can also sample conditions that were not present in the training data. Figure 20(c) shows samples where all $I_i = 0$ (the "wildtype" condition), and Figure 20(d) shows samples where $I_{E2A} = 1$ and $I_{MLL} = 1$, but the other I nodes are off. Both of these extrapolations look plausible, suggesting the model can be used to predict the consequences of novel interventions. However, we would likely need a much more complex model (e.g., comparable in complexity to the one in Figure 12(a)) to make plausible large extrapolations.

Next we examine the targets of each intervention node (i.e., the putative cause of each cancer subtype). Figure 19(b) shows $\hat{F}_{MAP}$, but what we would like is a "profile" across all the potential gene targets. We therefore compute the following approximate Bayes factor for the edge $I_k \to X_i$

$$BF(k,i) = \log \frac{p(D|k \to i)}{p(D|k \not\to i)} \approx \log \frac{p(D|\hat{G}_{MAP}, k \to i)}{p(D|\hat{G}_{MAP}, k \not\to i)} \tag{20}$$

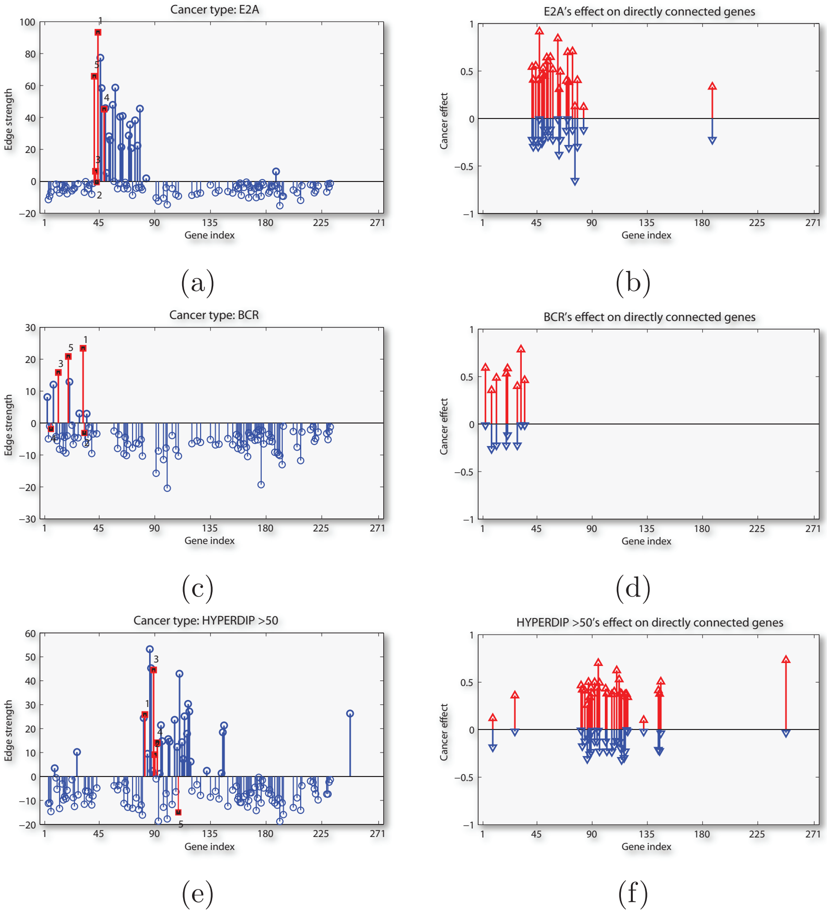
where $k \not\to i$ means there is no edge from k to i. (The approximation arises because we use the plug-in $\hat{G}_{MAP}$ rather than marginalizing over it, in addition to the approximation of using local search.) We can compute this by computing the marginal likelihood of the family for node X_i with and without I_k as a parent. This score is qualitatively similar to $\text{sim}(k,i)$ defined above. It will be positive for all the children of I_k in $\hat{H}_{MAP}$ (else $\hat{H}_{MAP}$ would not be a local maximum), and will be negative for the rest. In particular, it will be $-\infty$ for nodes i which already have another intervention parent, since we use the hard constraint of at most one intervention edge per backbone node, based on the assumption that different genes are targeted in each condition. (Of course, we could relax this assumption, but it seemed reasonable in light of (Dejori and Stetter, 2004).)

The results for three of the subtypes are shown Figures 21(a,c,e). The top 5 targets predicted by (Dejori and Stetter, 2004) are plotted in red, with their ranking shown as an integer. For subtypes E2A and BCR, the true oncogene targets (PBX1 and ABL1, respectively) is ranked first by our method and theirs. In other words, we both recover the true cause of these cancer subtypes. For the other subtypes, our results are also similar to those of (Dejori and Stetter, 2004), but since ground truth is unknown, it is hard to conclude too much from this. In Figures 21(b,d,f) we plot the polarity of the intervention edges in $\hat{H}_{MAP}$. In agreement with (Dejori and Stetter, 2004), we find that these cancer subtypes "work" by generally forcing an overexpression of their targets.

5. Conclusions and future work

In conclusion, we have shown that interventions are useful for learning causal structure, even if they have unknown (side) effects. By adopting a Bayesian approach and using standard structure learning algorithms, we can learn the targets of intervention and the graph structure at the same time. This results in better-fitting models and might also be useful for drug target discovery.

A natural next step is active learning, i.e., deciding which intervention to perform so as to identify the structure as quickly as possible. This has previously been studied in (Tong and Koller, 2001; Murphy, 2001), but the high variance of MCMC estimation limited the



(a) (b)

(c) (d)

(e) (f)

Figure 21: **Learned targets for various subtypes of cancer.** (a-b) E2A-PBX1, (c-d) BCR-ABL, (e-f) HYPERDIP>50. *Left:* Target edge strength $BF(k, i)$ for chosen edges (positive) and absent edges (negative). The top 5 genes chosen by Dejori and Stetter (2004) are shown in red, with their ranking shown as an integer. Gaps along the horizontal axis correspond to target edges which are impossible due to the fan-in constraint (i.e., genes which already have an intervention parent). *Right:* Probability that the cancer sets the target to +1 or -1 for the target edges in $\hat{H}_{MAP}$.

usefulness of the technique. It is possible that the recent introduction of DP algorithms for exact Bayesian inference will help.

Another important issue is how best to introduce latent variables into the model. This has been studied in the non-interventional setting by (Elidan et al., 2000), but interventions add a new twist: if an intervention node targets many children, it may be more parsimonious to say it targets a latent variable (e.g., representing a pathway) which is a hidden common cause of all the children. We leave such extensions to future work.

Acknowledgments

We would like to acknowledge support from NSERC.

References

J. Bilmes. Dynamic Bayesian multinets. In *UAI*, 2000.

D. Chickering. A transformational characterization of equivalent Bayesian network structures. In *UAI*, 1995.

D. Chickering and C. Meek. Finding Optimal Bayesian Networks. In *UAI*, 2002.

G. Cooper and C. Yoo. Causal discovery from a mixture of experimental and observational data. In *UAI*, 1999.

M. Dejori and M. Stetter. Identifying interventional and pathogenic mechanisms by generative inverse modeling of gene expression profiles. *Journal of Computational Biology*, 11 (6):1135–1148, 2004.

M. Dejori, B. Schuermann, and M. Stetter. Hunting drug targets by systems-level modeling of gene expression profiles. *IEEE Trans. on Nanobioscience*, 3(3), 2004.

D. Eaton and K. Murphy. Exact Bayesian structure learning from uncertain interventions. In *AI/Statistics*, 2007a.

D. Eaton and K. Murphy. Bayesian structure learning using dynamic programming and MCMC. In *UAI*, 2007b.

F. Eberhardt. Sufficient condition for pooling data from different distributions. In *First Symposium on Philosophy, History, and Methodology of Error*, 2006.

F. Eberhardt, C. Glymour, and R. Scheines. On the number of experiments sufficient and in the worst case necessary to identify all causal relations among N variables. In *UAI*, 2005.

F. Eberhardt, C. Glymour, and R. Scheines. Interventions and causal inference. In *20th Mtg. Philos. of Sci. Assoc.*, 2006.

G. Elidan, N. Lotner, N. Friedman, and D. Koller. Discovering hidden variables: A structure-based approach. In *NIPS*, 2000.

B. Ellis and W. Wong. Sampling Bayesian Networks quickly. In *Interface*, 2006.

N. Friedman. Inferring Cellular Networks Using Probabilistic Graphical Models. *Science*, 303(5659):799–805, February 2004.

N. Friedman and D. Koller. Being Bayesian about Network Structure: A Bayesian Approach to Structure Discovery in Bayesian Networks. *Machine Learning*, 50:95–126, 2003.

N. Friedman, K. Murphy, and S. Russell. Learning the structure of dynamic probabilistic networks. In *UAI*, 1998.

R. Fukuda, B. Kelly, and G. Semenza. Vascular endothelial growth factor gene expression in colon cancer cells exposed to prostaglandin e2 is mediated by hypoxia-inducible factor 1. *Cancer Research*, 63:2330–2334, 2003.

T. Gardner, D. di Bernardo, D. Lorenz, and J. Collins. Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*, 301:102–105, 2003.

D. Geiger and D. Heckerman. Parameter priors for directed acyclic graphical models and the characterization of several probability distributions. *The Annals of Statistics*, 30(5): 1412–1440, 2002.

D. Geiger and D. Heckerman. Learning Gaussian networks. In *UAI*, volume 10, pages 235–243, 1994.

D. Geiger and D. Heckerman. A characterization of Dirchlet distributions through local and global independence. *Annals of Statistics*, 25:1344–1368, 1997.

A. Gelman, J. Carlin, H. Stern, and D. Rubin. *Bayesian data analysis*. Chapman and Hall, 2004. 2nd edition.

A. Gopnik and L. Schulz. *Causal learning: Psychology, philosophy, and computation*. Oxford University Press, 2007.

K. Hallen, J. Bjorkegren, and J. Tegner. Detection of compound mode of action by computational integration of whole-genome measurements and genetic perturbations. *BMC Bioinformatics*, 7(51), 2006.

A. Hartemink. *Principled computational methods for the validation and discovery of genetic regulatory networks*. PhD thesis, MIT EECS, 2001.

D. Heckerman, J. Breese, and K. Rommelse. Troubleshooting under uncertainty. Technical Report MSR-TR-94-07, Microsoft Research, 1994.

D. Heckerman, D. Geiger, and M. Chickering. Learning Bayesian networks: the combination of knowledge and statistical data. *Machine Learning*, 20(3):197–243, 1995.

D. Husmeier. Sensitivity and specificity of inferring genetic regulatory interactions from microarray experiments with dynamic Bayesian networks. *Bioinformatics*, 19:2271–2282, 2003.

M. Koivisto. Advances in exact Bayesian structure discovery in Bayesian networks. In *UAI*, 2006.

M. Koivisto and K. Sood. Exact Bayesian structure discovery in Bayesian networks. *J. of Machine Learning Research*, 5:549–573, 2004.

K. Korb and E. Nyberg. The power of intervention. *Minds and Machines*, 16:289–302, 2006.

K. Korb, L. Hope, A. Nicholson, and K. Axnick. Varieties of causal intervention. In *Pacific Rim Conference on AI*, 2004.

A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, and R. Favera abd A. Califano. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bionformatics*, 7, 2006.

F. Markowetz and R. Spang. Evaluating the effect of perturbations in reconstructing network topologies. In *Proc. 3rd Intl. Wk. on Distrib. Stat. Computing*, 2003.

F. Markowetz, S. Grossmann, and R.Spang. Probabilistic soft interventions in conditional gaussian networks. In *10th AI/Stats*, 2005.

M. Marton, J. De Risi, H. Bennett, V. Iyer, M. Meyer, C. Roberts, R. Stoughton, J. Burchard, D. Slade, H. Dai, D. Bassett, L. Hartwell, P. Brown, and S. Friend. Drug tagret validation and identification of secondary drug target effects using DNA microarrays. *Nature Medicine*, 4(11), 1998.

Tom Minka. Bayesian inference, entropy and the multinomial distribution. Technical report, CMU, 2003.

A. Moore and M. Lee. Cached sufficient statistics for efficient machine learning with large datasets. *J. of AI Research*, 8:67–91, 1998.

K. Murphy. Active learning of causal Bayes net structure. Technical report, Comp. Sci. Div., UC Berkeley, 2001.

R. Neapolitan. *Learning Bayesian Networks*. Prentice Hall, 2003.

J. Park and A. Darwiche. Complexity Results and Approximation Strategies for MAP Explanations. *J. of AI Research*, 21:101–133, 2004.

J. Pearl. *Causality: Models, Reasoning and Inference*. Cambridge Univ. Press, 2000.

D. Pe'er, A. Regev, G. Elidan, and N. Friedman. Inferring subnetworks from peturbed expression profiles. *Bioinformatics*, 17, 2001. Supplement 1.

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

T. Silander and P. Myllmaki. A simple approach for finding the globally optimal Bayesian network structure. In *UAI*, 2006.

A. Singh and A. Moore. Finding optimal bayesian networks by dynamic programming. Technical report, Carnegie Mellon University, June 2005.

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

H. Steck and T. Jaakkola. On the dirichlet prior and bayesian regularization. In *NIPS*, 2002.

B. Thiesson, C. Meek, D. Chickering, and D. Heckerman. Learning mixtures of DAG models. In *UAI*, 1998.

J. Tian and J. Pearl. Causal discovery from changes: a Bayesian approach. Technical report, UCLA, 2001a.

J. Tian and J. Pearl. Causal discovery from changes. In *UAI*, 2001b.

S. Tong and D. Koller. Active learning for structure in Bayesian networks. In *Intl. Joint Conf. on AI*, 2001.

A. Werhli, M. Grzegorczyk, and D. Husmeier. Comparative evaluation of reverse engineering gene regulatory networks with relevance networks, graphical Gaussian models and Bayesian networks. *Bioinformatics*, 22(20):2523–2531, 2006.

EJ Yeoh, ME Rossa, SA Shurtleff, and WK Williams et al. Classification, subtype discovery, and prediction of outcome in pediatric acute lymphoblastic leukemia by gene expression profiling. *Cancer Cell*, 1:133–143, 2002.