Causal Discovery from Changes: a Bayesian Approach*

Jin Tian and Judea Pearl
Cognitive Systems Laboratory
Computer Science Department
University of California, Los Angeles, CA 90024
{*jtian, judea* }*@cs.ucla.edu*

Abstract

We propose a new method of discovering causal structures, based on the detection of local, spontaneous changes in the underlying data-generating model. We derive expressions for the Bayesian score that a causal structure should obtain from streams of data produced by locally changing distributions. Simulation experiments indicate that dynamic information may improve the power of discovery up to the theoretical limits set by statistical indistinguishability.

1 Introduction

In recent years, several graph-based algorithms have been developed for the purpose of inferring causal structures from empirical data. Some are based on detecting patterns of conditional independence relationships [Pearl and Verma, 1991, Spirtes *et al.*, 1993], and some are based on Bayesian approaches [Heckerman *et al.*, 1997, Cooper, 1999]. These discovery methods assume static environment, that is, a time-invariant distribution and a time-invariant data-generating model, while attempting to infer structures that encode dynamic aspects of the environment, for example, how probabilities would change as a result of interventions. This transition, from static to dynamic information, constitutes a major inferential leap, and is severely limited by the inherent indistinguishability (or equivalence) relation that governs Bayesian networks [Verma and Pearl, 1990].

One way of overcoming this basic limitation is to augment the data with partial causal knowledge, if such is available. [Spirtes *et al.*, 1993], for example, discussed the use of experimental data to identify causal relationships. [Cooper and Yoo, 1999] discussed a Bayesian method of causal discovery from a mixture of observational and experimental data.

We propose a new method of discovering causal relations in data, based on the detection and interpretation of local spontaneous changes in the environment. While previous methods are based on static statistical features of the data, our proposal aims at exploiting dynamic changes in that data. Such changes are always present in any realistic domain that is embedded in a larger background of dynamically changing conditions. For example, natural disasters, armed conflicts, epidemics, labor disputes, and even mundane decisions by other agents, are unexpected eventualities that are not naturally captured in distribution functions. The occurrence of such eventualities tend to *alter* the distribution under study and yield changes that are markedly different from ordinary statistical fluctuations. Whereas static analysis views these changes as nuisance, and attempts to adjust and compensate for them, we will view them as a source of information about the data-generating process. A controlled experimental study may be thought of as a special case of these environmental changes, where the external influence involves fixing a designated variable to some predetermined value. In general, however, the external influence may be milder, merely changing the conditional probability of a variable, given its causes. Moreover, in marked contrast to controlled experiments, we may not know in advance the nature of the change, its location, or even whether it took place; these may need to be inferred from the data itself.

The basic idea has its roots in the economic literature. The economist Kevin Hoover (1990) attempted to infer the direction of causal influences among economic variables (e.g., employment and money supply) by observing the changes that sudden modifications in the economy (e.g., tax reform, labor dispute) induced in the statistics of these variables. Hoover assumed that the conditional probability of an effect given its causes remains invariant to changes in the mechanism that generates the cause, while the conditional probability

*This is part-II of a two-part paper submitted to UAI-01. Part-I is entitled: Causal Discovery from Changes.

of a cause given the effect would not remain invariant under such changes. This asymmetry may be useful in distinguishing cause and effect.

We will assume that we have data generated from a dynamically changing environment and our task is to recover the actual causal structures. In a companion paper [Tian and Pearl, 2001] we have analyzed the patterns of distributional changes that such datasets may induce, and we proposed recovery methods that infer causal directionality information from those changes. In this paper, we investigate the Bayesian approach. The Bayesian approach [Heckerman *et al.*, 1997] gives us a consistent way of combining dynamic datasets to get an overall estimation of causal structures. We show how to derive a Bayesian scoring metric from various types of dynamic data by assigning appropriate priors over probability parameters. The Bayesian scores obtained are extensions of previously derived Bayesian scores [Cooper and Herskovits, 1992, Heckerman *et al.*, 1995]. For mixed observational and experimental data we obtained the same score as given in [Cooper and Yoo, 1999]. We show that dynamic data increase our power of causal discovery beyond the limits set by independence equivalence.

2 Causal Models and Mechanism Change

Let our problem domain be a set of discrete random variables $V = \{V_1, \ldots, V_n\}$. We assume that a *causal model* over V is a pair $M = \langle G, \Theta_G \rangle$, where G is a DAG over V, called a *causal diagram*, and Θ_G is a set of probability parameters. We assume that each variable V_i can take values from a finite domain, $Dm(V_i) = \{v_{i1}, \ldots, v_{ir_i}\}$, where r_i is the number of states of V_i. We use Pa_i to represent the set of parents of V_i in a causal diagram G and $Dm(Pa_i)$ to represent the set of states of Pa_i. Let $\theta_{v_i;pa_i}, v_i \in Dm(V_i), pa_i \in Dm(Pa_i)$ denote the multinomial parameter corresponding to the conditional probability $P(v_i|pa_i)$. We will use the following notations: $\vec{\theta}_{pa_i} = \{\theta_{v_i;pa_i} | v_i \in Dm(V_i)\}$, $\Psi_i = \cup_{pa_i \in Dm(Pa_i)} \vec{\theta}_{pa_i}$, $\Theta_G = \cup_{i=1}^n \Psi_i$. Assuming the Causal Markov condition [Spirtes *et al.*, 1993], a causal model $M = \langle G, \Theta_G \rangle$ generates a probability distribution

$$P(v) = \prod_i \theta_{v_i;pa_i}. \tag{1}$$

A probability distribution $P(V)$ is said to be *compatible* with a causal diagram G if $P(V)$ can be generated by some causal model $M = \langle G, \Theta_G \rangle$.

The factorization in Eq. (1) obtains causal character through the assumption of *modularity*; each family in the causal diagram represents an autonomous physical mechanism and is subjected to change without influencing other mechanisms. We formally define mechanism change as follows.

Definition 1 (Mechanism Change) *A mechanism change to a causal model $M = \langle G, \Theta_G \rangle$ at a variable V_i is a transformation of M that produces a new model, $M_{V_i} = \langle G, \Theta'_G \rangle$, where $\Theta'_G = \Psi'_i \cup (\Theta_G \setminus \Psi_i)$ and Ψ'_i is a set of parameters having different values with the parameters in Ψ_i.*

We assume in this paper that the parent set Pa_i does not change in a mechanism change. An intervention that fixes V_i to a particular value is a special case of a mechanism change. Let $P(V)$ be the distribution generated by M, as in Eq. (1). Then the distribution generated by M_{V_i} is given by

$$P_{V_i}(v) = \theta'_{v_i;pa_i} \prod_{j \neq i} \theta_{v_j;pa_j}. \tag{2}$$

We will call (P, P_{V_i}) a *transition pair (TP)* and V_i the *focal variable* of the transition. Assume that a series of mechanism changes occurred successively to a causal model $M = \langle G, \Theta_G^0 \rangle$, and let $F = (V_{i_1}, \ldots, V_{i_k})$ denote the corresponding sequence of focal variables. We use $P_{TS} = (P^0, P^1, \ldots, P^k)$ to denote the sequence of distributions generated by such a series, and call the pair (P_{TS}, F) a *transition sequence (TS)*, where each pair (P^{j-1}, P^j) is a TP with V_{i_j} as the focal variable. Assume that a series of mechanism changes occurred to a same causal model $M = \langle G, \Theta_G^0 \rangle$, and let $F = (V_{i_1}, \ldots, V_{i_k})$ denote the sequence of focal variables, and $P_{ES} = (P^0, P^1, \ldots, P^k)$ the corresponding sequence of distributions, where each pair (P^0, P^j) is a TP with V_{i_j} as the focal variable. We will call the pair (P_{ES}, F) an *experimental sequence (ES)*. An example of an ES is a series of experimental studies performed on a model.

As oracles for cause-and-effect relations, causal models can predict the effects that any external or spontaneous changes have on the distributions. Conversely, from probability distributions resulted from various mechanism changes, we obtain information on the structure of the model generating those distributions. In this paper, we assume that we are given a TS (P_{TS}, F) or an ES (P_{ES}, F) corresponding to some causal diagram G, and our task is to recover G. We will then assume that we have a sequence of datasets $\mathbb{D} = \{D^0, \ldots, D^k\}$, where each D^i is a set of random samples from a distribution P^i, and we will derive a Bayesian scoring metric for learning causal structures from this dynamic data. First, we introduce the Bayesian approach for causal discovery.

3 The Bayesian Approach

Assume that we have a set of random samples D generated from a causal model $M = \langle G, \Theta_G \rangle$. In the Bayesian approach, we compute the posterior probability of a causal diagram G given the dataset D as:

$$P(G|D,\xi) = \frac{P(D|G,\xi)P(G|\xi)}{P(D|\xi)}, \tag{3}$$

where ξ represents our background knowledge. The *marginal likelihood* of the data given G is computed as

$$P(D|G,\xi) = \int P(D|\Theta_G, G, \xi)P(\Theta_G|G,\xi)d\Theta_G. \tag{4}$$

The term $P(D|\Theta_G, G, \xi)$ is the probability of the data given a Bayesian network and is computable. We need to provide prior distributions for the probability parameters, $P(\Theta_G|G,\xi)$, and causal diagrams, $P(G|\xi)$. The term $P(D|\xi)$ is just a proportional constant.

We can then compute the posterior probability of any hypothesis of interest by averaging over all possible causal models. For example, the posterior probability that X causes Y is computed as

$$P(X \longrightarrow Y|D,\xi) = \sum_{X \to Y \in G} P(G|D,\xi), \tag{5}$$

where the summation is over all causal diagrams which contain the edge $X \longrightarrow Y$. Since the number of possible diagrams is exponential in the number of variables n, it is impractical to sum over all diagrams unless for very small n. One way to deal with this problem is to use the relative posterior probability $P(D, G|\xi)$ as a *scoring metric* and search for diagrams with high scores.

4 Derivation of Bayesian Score

For the case that the dataset D is from a static distribution, closed form expressions for $P(D|G,\xi)$ have been derived [Cooper and Herskovits, 1992, Heckerman *et al.*, 1995]. We will extend previous derivations to incorporate dynamic data.

Assume that we have two data sets, D and D', generated from a causal diagram G but with different parameters, Θ_G and Θ'_G respectively. The marginal likelihood is computed as:

$$P(D, D'|G,\xi) = \int P(D, D'|\Theta_G, \Theta'_G, G, \xi)P(\Theta_G, \Theta'_G|G,\xi)d\Theta_G d\Theta'_G. \tag{6}$$

Assuming that data cases are random samples, and that the data are *complete*, that is, every variable is assigned a value in all data cases, we have

$$\begin{aligned} &P(D, D'|\Theta_G, \Theta'_G, G, \xi) \\ &= P(D|\Theta_G, G, \xi)P(D'|\Theta'_G, G, \xi) \\ &= \prod_{l=1}^{N} P(C_l|\Theta_G, G, \xi) \prod_{l=1}^{N'} P(C'_l|\Theta'_G, G, \xi) \\ &= \prod_{i=1}^{n} \prod_{v_i} \prod_{pa_i} \theta_{v_i;pa_i}^{N_{v_i,pa_i}} \theta'^{N'_{v_i,pa_i}}_{v_i;pa_i}, \end{aligned} \tag{7}$$

where N is the number of cases in D, C_l represents a specific case in D, and N_{v_i,pa_i} is the number of cases in D for which V_i takes the value v_i and its parents Pa_i takes the value pa_i. We use $\prod_{v_i}$ as a shorthand for $\prod_{v_i \in Dm(V_i)}$ and $\prod_{pa_i}$ for $\prod_{pa_i \in Dm(Pa_i)}$.

Consider the prior distribution $P(\Theta_G, \Theta'_G|G,\xi)$. Assume that, as a background knowledge, the two datasets D and D' are from a TP (P, P') with known focal variable V_l. Therefore, the two sets of parameters Θ_G and Θ'_G differ only by those parameters in Ψ_l. With this knowledge, we assume the following prior:

$$\begin{aligned} &P(\Theta_G, \Theta'_G|G, V_l, \xi) \\ &= P(\Theta_G|G,\xi)P(\Psi'_l|G,\xi) \prod_{i \neq l} \delta(\Psi_i - \Psi'_i), \end{aligned} \tag{8}$$

where $\delta(x)$ is the Dirac delta function. Eq. (8) says that for $i \neq l$, $\Psi'_i = \Psi_i$, and the reader can verify that $P(\Theta_G, \Theta'_G|G, V_l, \xi)$ integrates to 1 and is a valid density function. We have put V_l as a condition to reflect the fact that V_l is known as the focal variable of the TP.

For the parameter priors $P(\Theta_G|G,\xi)$ and $P(\Psi'_l|G,\xi)$, we use the following assumptions given in [Heckerman *et al.*, 1995]:

- *Global Parameter Independence*:

$$P(\Theta_G|G,\xi) = \prod_{i=1}^{n} P(\Psi_i|G,\xi) \tag{9}$$

- *Local Parameter Independence*:

$$P(\Psi_i|G,\xi) = \prod_{pa_i} P(\vec{\theta}_{pa_i}|G,\xi), i = 1, \ldots, n. \tag{10}$$

- *Parameter Modularity*: if V_i has the same parents in two causal diagrams G_1 and G_2, then

$$P(\vec{\theta}_{pa_i}|G_1,\xi) = P(\vec{\theta}_{pa_i}|G_2,\xi), \; pa_i \in Dm(Pa_i). \tag{11}$$

While these assumptions were originally made for learning Bayesian networks, [Heckerman, 1995] discussed their implications for causal Bayesian networks.

Using Eq.s (7)–(11), and integrating out $\Theta'_G \setminus \Psi'_l$, Eq. (6) is transformed to

$$
\begin{aligned}
&P(\mathbb{D}_{TP}|G, V_l, \xi) \\
&= \prod_{i \neq l} \prod_{pa_i} \int (\prod_{v_i} \theta_{v_i;pa_i}^{M_{v_i,pa_i}}) P(\vec{\theta}_{pa_i}|\xi) d\vec{\theta}_{pa_i} \\
&\times \prod_{pa_l} \int (\prod_{v_l} \theta_{v_l;pa_l}^{N_{v_l,pa_l}}) P(\vec{\theta}_{pa_l}|\xi) d\vec{\theta}_{pa_l} \\
&\times \prod_{pa_l} \int (\prod_{v_l} \theta'^{N'_{v_l,pa_l}}_{v_l;pa_l}) P(\vec{\theta'}_{pa_l}|\xi) d\vec{\theta'}_{pa_l},
\end{aligned}
\tag{12}
$$

where

$$
M_{v_i,pa_i} = N_{v_i,pa_i} + N'_{v_i,pa_i}. \tag{13}
$$

We use the notation $\mathbb{D}_{TP} = \{D, D'\}$ and put V_l as a condition to emphasize that Eq. (12) is obtained under the assumption that the datasets D and D' are from a TP with known focal variable V_l. The standard assumption for $P(\vec{\theta}_{pa_i}|\xi)$ is a *Dirichlet distribution*:

$$
P(\vec{\theta}_{pa_i}|\xi) = Dir(\vec{\theta}_{pa_i}|\vec{\alpha}_{pa_i}), \tag{14}
$$

where $\vec{\alpha}_{pa_i} = \{\alpha_{v_i;pa_i} | v_i \in Dm(V_i)\}$ denotes the set of parameters for the Dirichlet distribution. Assuming that the set of parameters $\vec{\theta'}_{pa_l}$ have the same prior distribution as $\vec{\theta}_{pa_l}$ given by Eq. (14), we obtain

$$
\begin{aligned}
&P(\mathbb{D}_{TP}|G, V_l, \xi) \\
&= \prod_{i \neq l} \prod_{pa_i} \frac{\Gamma(\alpha_{pa_i})}{\Gamma(\alpha_{pa_i} + M_{pa_i})} \prod_{v_i} \frac{\Gamma(\alpha_{v_i;pa_i} + M_{v_i,pa_i})}{\Gamma(\alpha_{v_i;pa_i})} \\
&\times \prod_{pa_l} \frac{\Gamma(\alpha_{pa_l})}{\Gamma(\alpha_{pa_l} + N_{pa_l})} \prod_{v_l} \frac{\Gamma(\alpha_{v_l;pa_l} + N_{v_l,pa_l})}{\Gamma(\alpha_{v_l;pa_l})} \\
&\times \prod_{pa_l} \frac{\Gamma(\alpha_{pa_l})}{\Gamma(\alpha_{pa_l} + N'_{pa_l})} \prod_{v_l} \frac{\Gamma(\alpha_{v_l;pa_l} + N'_{v_l,pa_l})}{\Gamma(\alpha_{v_l;pa_l})},
\end{aligned}
\tag{15}
$$

where $\Gamma(\cdot)$ is the Gamma function, and

$$
\alpha_{pa_i} = \sum_{v_i} \alpha_{v_i;pa_i}, \; N_{pa_i} = \sum_{v_i} N_{v_i,pa_i}, \; M_{pa_i} = \sum_{v_i} M_{v_i,pa_i}.
$$

5 Likelihood Equivalence

For two independence-equivalent causal diagrams G_1 and G_2, any distribution compatible with G_1 is also compatible with G_2. Hence, it is reasonable to assume that a dataset D from a static distribution cannot distinguish between independence-equivalent causal diagrams, or, $P(D|G_1, \xi) = P(D|G_2, \xi)$. [Heckerman *et al.*, 1995] call this assumption *likelihood equivalence*. They show that it constrains the space of prior parameters $\alpha_{v_i;pa_i}$ and call the resulting likelihood-equivalent Bayesian scoring metric the BDe metric. We will use prior parameters that satisfy the likelihood equivalence property, and call the associated metric $P(\mathbb{D}_{TP}, G|V_l, \xi) = P(\mathbb{D}_{TP}|G, V_l, \xi)P(G|\xi)$ the BDe_TP metric.

The BDe_TP metric is *not* likelihood equivalent, and for a good reason. A TP can indeed distinguish independence-equivalent diagrams: among those independence-equivalent diagrams compatible with both P and P_{V_l}, a TP (P, P_{V_l}) can distinguish those that can generate P_{V_l} from P with a *single* mechanism change from those that can not. A causal diagram G is said to be *compatible with a transition pair* (P, P_{V_i}) if P can be generated by a causal model $M = <G, \Theta_G>$ and P_{V_i} can be generated by a causal model $M_{V_i} = <G, \Theta'_G>$ resulted from a mechanism change to M at V_i. Two causal diagrams G_1 and G_2 are called *transition pair equivalent* with respect to a TP with focal variable V_i, or V_i*-transition equivalent*, if every TP (P, P_{V_i}) compatible with G_1 is also compatible with G_2. The graphical conditions for TP equivalence are given by the following theorem [Tian and Pearl, 2001].

Theorem 1 (Transition Pair Equivalence) *Two causal diagrams G_1 and G_2 are V_i-transition equivalent if and only if they have the same skeletons, the same sets of v-structures, that is, two converging arrows whose tails are not connected, and the same sets of parents for V_i.*

See Figure 1 for an example of TP equivalence.

It is natural to extend the likelihood equivalence requirement and define a new property: a marginal likelihood $P(\mathbb{D}|G, \xi)$ is said to be V_l*-transition likelihood equivalent* if for any dataset $\mathbb{D}$ and two V_l-transition equivalent causal diagrams G_1 and G_2, $P(\mathbb{D}|G_1, \xi) = P(\mathbb{D}|G_2, \xi)$.

Theorem 2 *The marginal likelihood $P(\mathbb{D}_{TP}|G, V_l, \xi)$ given by Eq. (15) is V_l-transition likelihood equivalent.*

Proof: Eq. (15) can be rewritten as

$$
\begin{aligned}
&P(\mathbb{D}_{TP}|G, V_l, \xi) \\
&= \prod_{i} \prod_{pa_i} \frac{\Gamma(\alpha_{pa_i})}{\Gamma(\alpha_{pa_i} + M_{pa_i})} \prod_{v_i} \frac{\Gamma(\alpha_{v_i;pa_i} + M_{v_i,pa_i})}{\Gamma(\alpha_{v_i;pa_i})} \\
&\times \Big(\prod_{pa_l} \frac{\Gamma(\alpha_{pa_l})\Gamma(\alpha_{pa_l} + M_{pa_l})}{\Gamma(\alpha_{pa_l} + N_{pa_l})\Gamma(\alpha_{pa_l} + N'_{pa_l})} \\
&\prod_{v_l} \frac{\Gamma(\alpha_{v_l;pa_l} + N_{v_l,pa_l})\Gamma(\alpha_{v_l;pa_l} + N'_{v_l,pa_l})}{\Gamma(\alpha_{v_l;pa_l})\Gamma(\alpha_{v_l;pa_l} + M_{v_l,pa_l})} \Big).
\end{aligned}
\tag{16}
$$

Let G_1 and G_2 be two V_l-transition-equivalent causal diagrams. Then G_1 and G_2 are independence equivalent and have the same parent set Pa_l by

Theorem 1. The first term in Eq. (16) has exactly the same form as the BDe score and takes the same values for two independence-equivalent diagrams [Heckerman *et al.*, 1995]. The second term obtains the same values for G_1 and G_2 since they have the same Pa_l set. □

We see that given data from a TP, previously indistinguishable independence-equivalent causal diagrams may now be distinguished, and in this sense, two datasets generated from a same causal structure but with different parameters give us more power to learn the structure. This power comes from our assumption (or knowledge) that only a *single* causal mechanism has changed in generating the two datasets. Indeed, if we have no knowledge on how the two sets of parameters Θ_G and Θ'_G differ, we may only assume that they are independent and have the same distributions:

$$P(\Theta_G, \Theta'_G|G,\xi) = P(\Theta_G|G,\xi)P(\Theta'_G|G,\xi), \qquad (17)$$

which leads to a marginal likelihood given by

$$\begin{aligned} &P(D, D'|G,\xi) \\ &= P(D|G,\xi)P(D'|G,\xi) \\ &= \prod_i \prod_{pa_i} \frac{\Gamma(\alpha_{pa_i})}{\Gamma(\alpha_{pa_i}+N_{pa_i})} \prod_{v_i} \frac{\Gamma(\alpha_{v_i;pa_i}+N_{v_i,pa_i})}{\Gamma(\alpha_{v_i;pa_i})} \\ &\times \prod_i \prod_{pa_i} \frac{\Gamma(\alpha_{pa_i})}{\Gamma(\alpha_{pa_i}+N'_{pa_i})} \prod_{v_i} \frac{\Gamma(\alpha_{v_i;pa_i}+N'_{v_i,pa_i})}{\Gamma(\alpha_{v_i;pa_i})}. \end{aligned} \qquad (18)$$

Eq. (18) is a product of two BDe likelihood applied on datasets D and D' respectively, and is still likelihood equivalent. Hence, without knowledge on how they came about, two datasets do not increase our power of discrimination, save for providing more samples.

6 Incorporating Experimental Data

Now assume that our knowledge is that the cases in D' are from an experimental study in which the variable V_l is fixed to a value $v_{lj} \in Dm(V_l)$, denoted by $do(V_l = v_{lj})$ or $do(v_{lj})$. Then instead of the Dirichlet distribution, we assign the following prior distribution to the parameter set $\vec{\theta'}_{pa_l}$:

$$P(\vec{\theta'}_{pa_l}|do(v_{lj}),\xi) = \delta(\theta'_{v_{lj};pa_l} - 1) \prod_{v_l \neq v_{lj}} \delta(\theta'_{v_l;pa_l}), \qquad (19)$$

which asserts that

$$\theta'_{v_l;pa_l} = \begin{cases} 1 & \text{if } v_l = v_{lj} \\ 0 & \text{otherwise} \end{cases}$$

Plugging Eq. (19) into Eq. (12), we obtain

$$\begin{aligned} &P(\mathbb{D}_{TP}|G, do(v_{lj}),\xi) \\ &= \prod_{i\neq l} \prod_{pa_i} \frac{\Gamma(\alpha_{pa_i})}{\Gamma(\alpha_{pa_i}+M_{pa_i})} \prod_{v_i} \frac{\Gamma(\alpha_{v_i;pa_i}+M_{v_i,pa_i})}{\Gamma(\alpha_{v_i;pa_i})} \\ &\times \prod_{pa_l} \frac{\Gamma(\alpha_{pa_l})}{\Gamma(\alpha_{pa_l}+N_{pa_l})} \prod_{v_l} \frac{\Gamma(\alpha_{v_l;pa_l}+N_{v_l,pa_l})}{\Gamma(\alpha_{v_l;pa_l})}. \end{aligned} \qquad (20)$$

Eq. (20) has been given in [Cooper and Yoo, 1999]. Here we show that it can be derived by providing an informative parameter prior as given by Eqs. (8) and (19). In the derivation of Eq. (20), we have used the following equation

$$\int (\prod_{v_l} {\theta'}_{v_l;pa_l}^{N'_{v_l,pa_l}}) \delta(\theta'_{v_{lj};pa_l} - 1) \prod_{v_l \neq v_{lj}} \delta(\theta'_{v_l;pa_l}) d\vec{\theta'}_{pa_l} = 1, \qquad (21)$$

which follows from that for $v_l \neq v_{lj}$, $N'_{v_l,pa_l} = 0$.

Theorem 3 *The likelihood* $P(\mathbb{D}_{TP}|G, do(v_{lj}),\xi)$ *given by Eq. (20) is* V_l*-transition likelihood equivalent.*

Proof: The same proof for Theorem 2. □

7 Combining Various Types of Dynamic Data

So far we have only considered the situations with two datasets. The discussions can be easily extended to the situations with a sequence of datasets, generated from a TS or an ES. Let $\mathbb{D} = \{D^0, D^1, \ldots, D^k\}$ be a sequence of datasets generated from some causal diagram G with parameters $\Theta_G^0, \ldots, \Theta_G^k$ respectively, and let $\Xi_G = \cup_{i=0}^k \Theta_G^i$. The marginal likelihood is computed as

$$P(\mathbb{D}|G,\xi) = \int P(\mathbb{D}|\Xi_G, G, \xi) P(\Xi_G|G,\xi) d\Xi_G. \qquad (22)$$

The term $P(\mathbb{D}|\Xi_G, G,\xi)$ can be computed as in Eq. (7). To give an appropriate parameter prior $P(\Xi_G|G,\xi)$, we need to know how these datasets in $\mathbb{D}$ came about. Assume that we have the knowledge that the sequence of datasets, which will now be denoted by $\mathbb{D}_{TS}$, are from a TS with a sequence of focal variables $F = (V_{i_1}, \ldots, V_{i_k})$. Then, we assume the following prior:

$$\begin{aligned} &P(\Xi_G|G, F,\xi) \\ &= P(\Theta_G^0|G,\xi) \Big(P(\Psi_{i_1}^1|G,\xi) \prod_{i\neq i_1} \delta(\Psi_i^1 - \Psi_i^0) \Big) \\ &\quad \Big(P(\Psi_{i_2}^2|G,\xi) \prod_{i\neq i_2} \delta(\Psi_i^2 - \Psi_i^1) \Big) \\ &\quad \cdots \Big(P(\Psi_{i_k}^k|G,\xi) \prod_{i\neq i_k} \delta(\Psi_i^k - \Psi_i^{k-1}) \Big), \end{aligned} \qquad (23)$$

where we have used the notation $\Theta_G^j = \cup_{i=1}^n \Psi_i^j, j = 0, \ldots, k$ as before. Eq. (23) is an extension of Eq. (8), and says that the set of parameters Θ_G^j differs with Θ_G^{j-1} only by the parameters in $\Psi_{i_j}^j$. Let $I = \{i_1, \ldots, i_k\}$ be the set of indexes for focal variables. Using the Dirichlet priors, we obtain the following expression for the marginal likelihood (22):

$$\begin{aligned} &P(\mathbb{D}_{TS}|G,F,\xi) \\ &= \prod_{i \notin I} \prod_{pa_i} \frac{\Gamma(\alpha_{pa_i})}{\Gamma(\alpha_{pa_i} + M_{pa_i})} \prod_{v_i} \frac{\Gamma(\alpha_{v_i;pa_i} + M_{v_i,pa_i})}{\Gamma(\alpha_{v_i;pa_i})} \\ &\times \prod_{l=1}^{k} \prod_{pa_{i_l}} \frac{\Gamma(\alpha_{pa_{i_l}})}{\Gamma(\alpha_{pa_{i_l}} + M_{pa_{i_l}}^l)} \prod_{v_{i_l}} \frac{\Gamma(\alpha_{v_{i_l};pa_{i_l}} + M_{v_{i_l},pa_{i_l}}^l)}{\Gamma(\alpha_{v_{i_l};pa_{i_l}})} \\ &\times \prod_{l=1}^{k} \prod_{pa_{i_l}} \frac{\Gamma(\alpha_{pa_{i_l}})}{\Gamma(\alpha_{pa_{i_l}} + L_{pa_{i_l}}^l)} \prod_{v_{i_l}} \frac{\Gamma(\alpha_{v_{i_l};pa_{i_l}} + L_{v_{i_l},pa_{i_l}}^l)}{\Gamma(\alpha_{v_{i_l};pa_{i_l}})}, \end{aligned} \tag{24}$$

where

$$M_{v_i,pa_i}^l = \sum_{j=0}^{l-1} N_{v_i,pa_i}^j, M_{v_i,pa_i} = M_{v_i,pa_i}^{k+1}, L_{v_i,pa_i}^l = \sum_{j=l}^{k} N_{v_i,pa_i}^j,$$
$$M_{pa_i}^l = \sum_{v_i} M_{v_i,pa_i}^l, \; M_{pa_i} = \sum_{v_i} M_{v_i,pa_i}, \; L_{pa_i}^l = \sum_{v_i} L_{v_i,pa_i}^l,$$

and N_{v_i,pa_i}^j is the number of cases in the dataset D^j for which V_i takes the value v_i and its parents Pa_i takes the value pa_i. Note that $M_{v_i,pa_i} = L_{v_i,pa_i}^l + M_{v_i,pa_i}^l$ is the number of cases in the whole dataset $\mathbb{D}_{TS}$ for which V_i takes the value v_i and its parents Pa_i takes the value pa_i. We will call the Bayesian scoring metric $P(\mathbb{D}_{TS}, G|F,\xi) = P(\mathbb{D}_{TS}|G,F,\xi)P(G|\xi)$ (with parameters $\alpha_{v_i;pa_i}$ satisfying the likelihood equivalence property) the BDe_TS metric.

A TS is simply a series of TP's. Accordingly, we say that a causal diagram is *compatible with a transition sequence* (P_{TS}, F) if it is compatible with each TP in the sequence. Likewise, two causal diagrams G_1 and G_2 are called *transition sequence equivalent* with respect to a TS (P_{TS}, F), or *F-transition equivalent*, if every TS (P_{TS}, F) compatible with G_1 is also compatible with G_2. The graphical conditions for TS equivalence are given by the following theorem.

Theorem 4 (Transition Sequence Equivalence) *Two causal diagrams are F-transition equivalent if and only if they have the same skeletons, the same sets of v-structures, and the same sets of parents for variables in F.*

A marginal likelihood $P(\mathbb{D}|G,\xi)$ is said to satisfy the property of *F-transition likelihood equivalence* if for two F-transition equivalent causal diagrams G_1 and G_2, $P(\mathbb{D}|G_1,\xi) = P(\mathbb{D}|G_2,\xi)$.

Theorem 5 *The marginal likelihood $P(\mathbb{D}_{TS}|G,F,\xi)$ given by Eq. (24) is F-transition likelihood equivalent.*

Proof: Similar to the proof of Theorem 2. □

Now assume that we have the knowledge that the sequence of datasets, which will now be denoted by $\mathbb{D}_{ES}$, are from an ES with the focal variables $F = (V_{i_1}, \ldots, V_{i_k})$. We then assume the following prior:

$$\begin{aligned} &P(\Xi_G|G,F,\xi) \\ &= P(\Theta_G^0|G,\xi)\Big(P(\Psi_{i_1}^1|G,\xi) \prod_{i \neq i_1} \delta(\Psi_i^1 - \Psi_i^0)\Big) \\ &\quad \Big(P(\Psi_{i_2}^2|G,\xi) \prod_{i \neq i_2} \delta(\Psi_i^2 - \Psi_i^0)\Big) \\ &\quad \cdots \Big(P(\Psi_{i_k}^k|G,\xi) \prod_{i \neq i_k} \delta(\Psi_i^k - \Psi_i^0)\Big). \end{aligned} \tag{25}$$

Eq. (25) is also an extension of Eq. (8), and says that the set of parameters Θ_G^j differs with Θ_G^0 only by the parameters in $\Psi_{i_j}^j$. Using the Dirichlet distribution, the marginal likelihood is given by

$$\begin{aligned} &P(\mathbb{D}_{ES}|G,F,\xi) \\ &= \prod_{i \notin I} \prod_{pa_i} \frac{\Gamma(\alpha_{pa_i})}{\Gamma(\alpha_{pa_i} + M_{pa_i})} \prod_{v_i} \frac{\Gamma(\alpha_{v_i;pa_i} + M_{v_i,pa_i})}{\Gamma(\alpha_{v_i;pa_i})} \\ &\times \prod_{l=1}^{k} \prod_{pa_{i_l}} \frac{\Gamma(\alpha_{pa_{i_l}})}{\Gamma(\alpha_{pa_{i_l}} + K_{pa_{i_l}}^l)} \prod_{v_{i_l}} \frac{\Gamma(\alpha_{v_{i_l};pa_{i_l}} + K_{v_{i_l},pa_{i_l}}^l)}{\Gamma(\alpha_{v_{i_l};pa_{i_l}})} \\ &\times \prod_{l=1}^{k} \prod_{pa_{i_l}} \frac{\Gamma(\alpha_{pa_{i_l}})}{\Gamma(\alpha_{pa_{i_l}} + N_{pa_{i_l}}^l)} \prod_{v_{i_l}} \frac{\Gamma(\alpha_{v_{i_l};pa_{i_l}} + N_{v_{i_l},pa_{i_l}}^l)}{\Gamma(\alpha_{v_{i_l};pa_{i_l}})}, \end{aligned} \tag{26}$$

where

$$K_{v_{i_l},pa_{i_l}}^l = M_{v_{i_l},pa_{i_l}} - N_{v_{i_l},pa_{i_l}}^l, \; K_{pa_{i_l}}^l = \sum_{v_{i_l}} K_{v_{i_l},pa_{i_l}}^l. \tag{27}$$

A special case of ES is a series of experimental studies in which each variable in F is fixed to some value respectively. Then we use the prior given in Eq. (19) for $P(\Psi_{i_j}^j|G,\xi), j = 1, \ldots, k$, and we obtain

$$\begin{aligned} &P(\mathbb{D}_{ES}|G,do(F),\xi) \\ &= \prod_{i \notin I} \prod_{pa_i} \frac{\Gamma(\alpha_{pa_i})}{\Gamma(\alpha_{pa_i} + M_{pa_i})} \prod_{v_i} \frac{\Gamma(\alpha_{v_i;pa_i} + M_{v_i,pa_i})}{\Gamma(\alpha_{v_i;pa_i})} \\ &\times \prod_{l=1}^{k} \prod_{pa_{i_l}} \frac{\Gamma(\alpha_{pa_{i_l}})}{\Gamma(\alpha_{pa_{i_l}} + K_{pa_{i_l}}^l)} \prod_{v_{i_l}} \frac{\Gamma(\alpha_{v_{i_l};pa_{i_l}} + K_{v_{i_l},pa_{i_l}}^l)}{\Gamma(\alpha_{v_{i_l};pa_{i_l}})}. \end{aligned} \tag{28}$$

Eq. (28) has been given in [Cooper and Yoo, 1999].

Theorem 6 *The marginal likelihood $P(\mathbb{D}_{ES}|G, F, \xi)$ in (26) and $P(\mathbb{D}_{ES}|G, do(F), \xi)$ in (28) is F-transition likelihood equivalent.*

Proof: Similar to the proof of Theorem 2. $\square$

In deriving Eq.s (24), (26), and (28), we have assumed that mechanism changes occurred at different variables. The situations in which different mechanism changes happen at a same variable can be easily incorporated. For example, in experimental studies, we may set a variable to different values. For this case, Eq. (28) is still applicable while $K^{l}_{v_{i_l}, pa_{i_l}}$ as expressed in Eq. (27) should exclude all experimental data for which V_{i_l} is set to some fixed value.

In summary, to compute the marginal likelihood for dynamic data, we just need to provide an appropriate prior $P(\Xi_G|G, \xi)$ to reflect our knowledge on how those data came about. We demonstrated this method with several priors given in Eqs. (8), (23), (25) and (19).

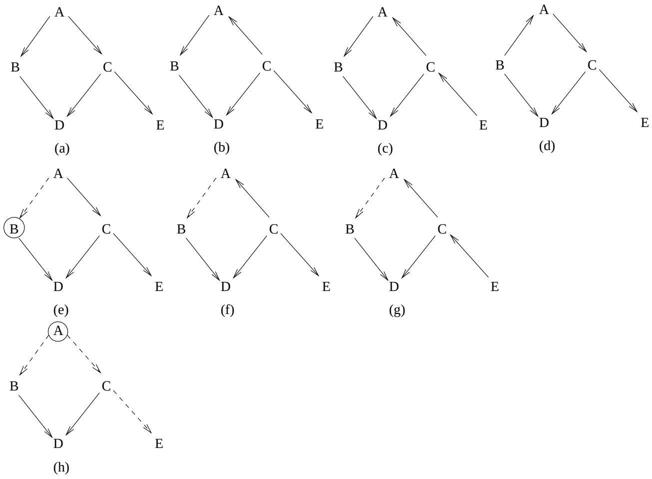

Figure 1: (a)The *Cancer* network. (a)-(d) are independence equivalent. (e)-(g) are B-transition equivalent. A mechanism change on A determines a unique causal diagram (h).

8 Experimental Results

We tested the BDe_TP score with data generated from a known network, the *Cancer* Bayesian network.[1] We assumed a uniform prior distribution over all possible network structures. We used the parameters: $\alpha_{v_i;pa_i} = 1/r_i q_i$, where r_i is the number of states of V_i and q_i is the number of states of Pa_i, which satisfies the likelihood-equivalence requirement [Heckerman *et al.*, 1995].

A mechanism change at a variable V_i is simulated as follows. Consider parameters in $\vec{\theta}_{pa_i}$. If $\theta_{v_{i1};pa_i} \leq 0.5$ then let $\theta'_{v_{i1};pa_i} = \theta_{v_{i1};pa_i} + \delta$, else let $\theta'_{v_{i1};pa_i} = \theta_{v_{i1};pa_i} - \delta$, where δ is a parameter for adjusting the change magnitude. The rest of the parameters in $\vec{\theta}_{pa_i}$ are changed in proportional to their original values as: $\theta'_{v_{ij};pa_i} = \alpha\theta_{v_{ij};pa_i}, j = 2, \ldots, r_i$, where $\alpha = (1 - \theta'_{v_{i1};pa_i})/(1 - \theta_{v_{i1};pa_i})$. When we simulate a mechanism change at V_i, we change parameters in $\vec{\theta}_{pa_i}$ as above for each $pa_i \in Dm(Pa_i)$.

The *Cancer* network is shown in Figure 1(a). It has only 5 nodes, hence we can exhaustively go through all 29,281 possible structures to compute the Bayesian average of any hypothesis of interest and to find the diagrams with the maximum posterior probabilities. We computed the probability of each edge in the true *Cancer* network as in Eq. (5), and compared the results given by the BDe_TP metric (15) with that by the BDe metric (18). We experimented with δ values of 0.1 and 0.5, and focal variables B and A respectively, and generated a TP dataset $\mathbb{D}_{TP} = \{D^0, D^1\}$ for each case by first generating 2000 cases from the original network as D^0, then simulating a mechanism change, and finally generating another 2000 cases as D^1.

The results are shown in Table 1 for the first N cases in the dataset (N from D^0 and N from D^1). When using the BDe metric, the Cancer network and its independence-equivalent diagrams of Figure 1(b)-(d) obtain the maximum score when the sample size is large enough, and they obtain a much larger posterior than all other structures. $P(A \to B|\mathbb{D})$ goes to 0.75 because three of the four diagrams of Figure 1(a)-(d) have the edge $A \to B$ and we assumed a uniform distribution over structures. For the same reason, with the BDe metric, $P(A \to C|\mathbb{D})$ goes to 1/2, $P(B \to D|\mathbb{D})$ and $P(C \to D|\mathbb{D})$ goes to 1, and $P(C \to E|\mathbb{D})$ goes to 3/4. When using the BDe_TP metric and B as the focal variable, the posterior over structures concentrated sharply around the three B-transition equivalent diagrams of Figure 1(e)-(g) when the sample size is large. Hence with the increasing sample size, $P(A \to B|\mathbb{D})$ goes to 1, $P(A \to C|\mathbb{D})$ goes to 1/3, and $P(C \to E|\mathbb{D})$ goes to 2/3. With A as the focal variable, the BDe_TP score concentrated sharply around the unique *Cancer* network (see Figure 1(h)) for large sample size, and the posteriors of all five edges go to 1.

9 Conclusion

We have demonstrated, using simulated data, that the use of information about local changes may improve the power of discovery up to the theoretical limits set by statistical indistinguishability. The major ad-

[1] We used the version downloaded from the web site of Norsys Software Corporation, http://www.norsys.com.

Table 1: The posteriors of edges in the *Cancer* network.

$\delta = 0.1, B$ as the focal variable.

N	$P(A \to B\vert\mathbb{D})$		$P(A \to C\vert\mathbb{D})$		$P(B \to D\vert\mathbb{D})$		$P(C \to D\vert\mathbb{D})$		$P(C \to E\vert\mathbb{D})$	
	BDe_TP	BDe	BDe_TP	BDe	BDe_TP	BDe	BDe_TP	BDe	BDe_TP	BDe
100	0.138	0.419	0.103	0.0394	0.997	0.87	0.853	0.86	0.552	0.441
200	0.335	0.482	0.354	0.136	1	0.993	0.983	0.993	0.607	0.403
500	0.604	0.686	0.43	0.457	1	0.999	0.996	1	0.713	0.728
1000	0.999	0.733	0.338	0.49	1	1	1	1	0.667	0.74
2000	1	0.75	0.336	0.5	1	1	1	1	0.666	0.75
$\delta = 0.5, B$ as the focal variable.										
100	0.999	0.238	0.0325	0.0141	1	0.484	0.284	0.293	0.0733	0.239
200	1	0.289	0.212	0.0516	1	0.663	0.83	0.546	0.0476	0.0106
500	1	0.658	0.495	0.651	1	0.992	1	0.989	0.0476	0.00518
1000	1	0.726	0.342	0.547	1	1	1	1	0.645	0.538
2000	1	0.75	0.334	0.5	1	1	1	1	0.666	0.75

$\delta = 0.1, A$ as the focal variable.

N	$P(A \to B\vert\mathbb{D})$		$P(A \to C\vert\mathbb{D})$		$P(B \to D\vert\mathbb{D})$		$P(C \to D\vert\mathbb{D})$		$P(C \to E\vert\mathbb{D})$	
	BDe_TP	BDe	BDe_TP	BDe	BDe_TP	BDe	BDe_TP	BDe	BDe_TP	BDe
100	0.832	0.471	0.226	0.106	0.979	0.911	0.958	0.84	0.477	0.441
200	0.827	0.494	0.278	0.0367	0.985	0.978	0.964	0.972	0.389	0.206
500	0.997	0.747	0.961	0.505	1	1	1	1	0.697	0.736
1000	0.995	0.75	0.948	0.5	1	1	1	1	0.961	0.75
2000	1	0.75	0.99	0.5	1	1	1	1	0.986	0.75
$\delta = 0.5, A$ as the focal variable.										
100	1	0.586	0.832	0.57	0.999	0.916	0.961	0.878	0.0882	0.0171
200	1	0.676	0.992	0.642	1	0.999	1	0.999	0.47	0.113
500	1	0.746	1	0.507	1	1	1	1	0.963	0.739
1000	1	0.744	1	0.513	1	1	1	1	0.932	0.731
2000	1	0.75	1	0.5	1	1	1	1	0.994	0.75

vantage of the Bayesian treatment of local changes, vis-a-vis the purely topological approach reported in [Tian and Pearl, 2001], lies in that the Bayesian score is less sensitive to topological errors (e.g., remote descendants of focal variables that do not change). On the other hand, the Bayesian method is more computation intensive; hybrid schemes remain to be investigated.

References

[Cooper and Herskovits, 1992] G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

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

[Cooper, 1999] G.F. Cooper. An overview of the representation and discovery of causal relationships using Bayesian networks. In Glymour C. and Cooper G.F., editors, *Computation, Causation, and Discovery*, 1999. AAAI Press and MIT Press.

[Heckerman *et al.*, 1995] D. Heckerman, D. Geiger, and D.M. Chickering. Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning*, 20:197–243, 1995.

[Heckerman *et al.*, 1997] D. Heckerman, C. Meek, and G. Cooper. A Bayesian approach to causal discovery. Technical Report MSR-TR-97-05, Microsoft Research, 1997.

[Heckerman, 1995] D. Heckerman. A Bayesian approach to learning causal networks. In *Proceedings UAI*, 1995.

[Hoover, 1990] K.D. Hoover. The logic of causal inference. *Economics and Philosophy*, 6:207–234, 1990.

[Pearl and Verma, 1991] J. Pearl and T. Verma. A theory of inferred causation. In *Proceedings KR'91*.

[Spirtes *et al.*, 1993] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search*. Springer-Verlag, New York, 1993.

[Tian and Pearl, 2001] J. Tian and J. Pearl. Causal discovery from changes. Technical Report R-280, UCLA, 2001. Submitted to UAI 2001.

[Verma and Pearl, 1990] T. Verma and J. Pearl. Equivalence and synthesis of causal models. In *Proceedings UAI*, 1990.