



University of Warwick institutional repository: <http://go.warwick.ac.uk/wrap>

This paper is made available online in accordance with publisher policies. Please scroll down to view the document itself. Please refer to the repository record for this item and our policy information available from the repository home page for further information.

To see the final version of this paper please visit the publisher's website. Access to the published version may require a subscription.

Author(s): G Freeman and JQ Smith

Article Title: Dynamic staged trees for discrete multivariate time series: forecasting, model selection and causal analysis

Year of publication: 2010

Link to published article:

<http://www2.warwick.ac.uk/fac/sci/statistics/crism/research/2010/paper10-14>

Publisher statement: None

# Using dynamic staged trees for discrete time series data: robust prediction, model selection and causal analysis

Guy Freeman

*University of Warwick*

Jim Q. Smith

*University of Warwick*

**Summary.** A new tree-based graphical model — the dynamic staged tree — is used to model discrete-valued discrete-time multivariate processes which are hypothesised to exhibit certain symmetries concerning how situations might unfold. We define and implement a one-step-ahead prediction algorithm using multi-process modelling and the power steady model. This is robust to short-term variations in the data yet sensitive to underlying system changes. We demonstrate that the whole analysis can be performed in a conjugate way so that the vast model space can be traversed quickly and results communicated transparently. We also demonstrate how to analyse causal hypotheses on this model class. Our techniques are illustrated using a simple educational example.

*Keywords:* Staged trees, Bayesian model selection, Bayes factors, forecasting, discrete time series, causal inference, power steady model, multi-process model

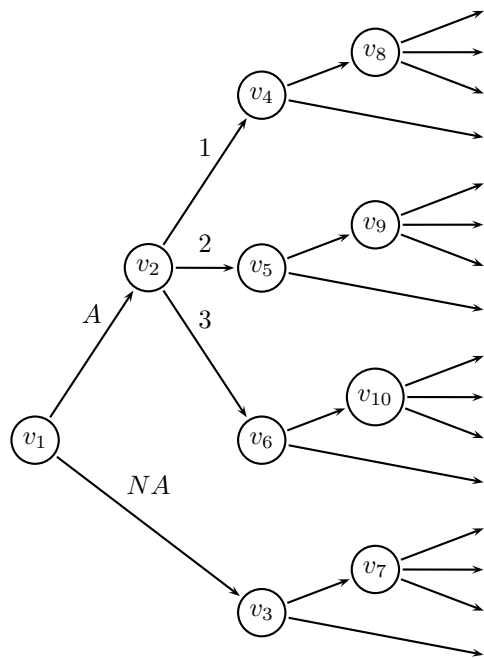
## 1. Introduction

In this paper we consider a class of dynamic multivariate models with finite discrete state spaces over the observed variables, which have the following characteristics:

- (1) A description is provided of the possible development histories each unit in a given time cohort can take. These histories could be radically different from one another in terms of length of development, the variables encountered, the state spaces of each stage of development, and so on.
- (2) There are various symmetry hypotheses for a given population of units concerning which situations in the histories have the same distributions over their immediate developments.
- (3) The units arrive in discrete equally-spaced time cohorts. The symmetries in the system are allowed to change from one time point to the next to reflect a changing environment.
- (4) The system may, at various times, be subject to local interventions. The model then admits a “causal” extension which provides predictions of the process when subject to such a control.

We are particularly interested in this paper in making good one-step ahead predictions for such a model. This will also provide the probabilities of the symmetry hypotheses through time to use as an explanatory tool.

One example of a system that fits the criteria above is a programme of study provided by an educational establishment which monitors students’ marks over time. We therefore



**Fig. 1.** Event tree for marks for two modules in a course. Marks are discretized into 3 grades, and  $A$  and  $NA$  indicate whether the mark is recorded or missing. The 10 situations are labelled and the 16 leaf nodes are unlabelled.

use this as our running example. The general points above translate into the following specific issues:

- (1) The modules of the course are always taken in a particular order (or consistent with some partial order); there might be a requirement to achieve a threshold mark before being allowed to continue onto the next module; and certain modules might have different prerequisite modules.
- (2) A student's performance on a previous module could influence the marks on a later one.
- (3) New students come in yearly cohorts. Because of any number of possible changes in any number of unobserved confounding factors the similarities in outcomes between different course histories could change for each cohort.
- (4) The administrators will be interested in predicting the effect on the mark distribution by changing the program in some way, such as changing the syllabus or lecturer for a module, changing the prerequisites for a modules, or removing a module entirely.

To graphically represent the different histories of a unit in the system one could use event trees (Shafer, 1996). One example of an event tree for the marks for a course with two modules is given by Figure 1.

The event tree can represent any discrete event space and naturally codifies a chronological order (or partial order) in its topology, and so we base our own model class on it. However, it is not sufficient for addressing the rest of our requirements by itself, particularly because it does not codify the symmetries in the system that we are interested in modelling.

Different situations having the same subsequent development can be viewed as a kind of conditional independence (Dawid, 1979) where the random variable describing the subsequent development is held to be independent of the history that led to the situation — see Studený (2005) for an overview of conditional independence structures. There are many graphical models that aim to represent conditional independence relations between the different variables of a system. Bayesian networks (BNs) (Pearl, 2000b; Cowell et al., 1999) are currently the most prominent of these models. However, they cannot easily represent in their graphical structure the asymmetry in the potential histories of a unit. Some enhancements to the canonical Bayesian network have therefore been suggested in order to take this “context-specific independence” into account, for example by Boutilier et al. (1996).

Using different semantics from BNs, Smith and Anderson (2008) defined the CHAIN EVENT GRAPH (CEG) as an enhancement of the event tree, where non-leaf nodes with the same probability distribution over their outgoing edges are linked by undirected edges, and where subtrees with identical probability distributions of their root-to-leaf paths are merged. Our model class is therefore based on CEGs, but extended into a more general dynamic scenario where probabilities are allowed to change with time.

In this paper we describe the dynamics of the process by a state space model incorporating a switching mechanism to a neighbouring model at a given time point. The earliest example of this class, to the best of our knowledge, was studied for univariate Gaussian series by Harrison and Stevens (1976) and called Multi-process Models Class II (see West and Harrison (1997) for a more recent review). Frühwirth-Schnatter (2006) reviews switching models for non-Gaussian state spaces, none of which have closed posterior forms. Here, we use a type of multi-process model which allows us to dynamically shift from one symmetry partition to another neighbouring one.

Various classes of discrete multivariate time series are of course well studied. Possibly the closest class to the one considered here are the model used in event history analysis. Event history data relates to when events of interest occur, rather than what events occur at time points of interest. Formally, an event history can be identified as a MARKED POINT PROCESS, a set  $\{(T_s, E_s) : s = 1, \dots, S\}$  of pairs of times  $T_s$  when events  $E_s$  occurred, where the times are random variables while the events of interest are fixed beforehand, although their order might be uncertain a priori (Arjas, 1989). Two graphical models developed for event history analysis are local independence graphs (Didelez, 2008) and graphical duration graphs (Gottard, 2007). While there is an overlap between event history data and the problem outlined here, it is clear that the two address quite separate concerns. In event history analyses the number of events under consideration is typically small, with the focus of analysis being the time of events, usually allowed to occur within a continuous time domain. Here, in contrast, we wish to model a complex discrete distribution over a discrete time domain.

In order to take into account possible drifting on the tree parameters through time caused by unobserved background processes, one could follow the filtering approach of stating a transition probability  $P(\theta_t | \theta_{t-1}, S)$ , where  $\theta_t$  represents the parameters on the tree at time  $t$  and  $S$  is the underlying model. The most common way to achieve this is to use a conventional state space formulation. Unfortunately, this approach immediately requires the inference to be undertaken with approximating numerical methods. This is not ideal in this context for several reasons: Firstly, in the stochastic version of the process we consider, conjugacy is retained by using product Dirichlet priors, and it would be a shame to lose this useful modular property. Secondly, because of the enormity of the model space of our domain of application it is convenient to be able to have Bayes factors calculable in closed

form, because this speeds up computation of model goodness enormously. Thirdly, models in this class are easier to interpret when they retain their modular and conjugate forms.

Another approach, which we take here, is to set a transition function

$$\mathcal{T} : P(\theta_{t-1} \mid x^{t-1}, S) \mapsto P(\theta_t \mid x^{t-1}, S) \quad (1)$$

where  $x_{t-1}$  are the observations up to time  $t-1$ . Although this approach is more restrictive in its scope, it can be justified through various characterisations (Smith, 1979, 1992) and we show below that it can have the very attractive property of preserving the conjugate structure of each model in this class, encouraging several different authors to use such transitions.

Interventions on a graphical model are covered by the causal literature (e.g. Pearl (2000b)). Causal analysis on event trees was considered by Shafer (1996) and was defined for chain event graphs by Thwaites et al. (2010). We extend this to the dynamic model class we present here. By retaining conjugacy when learning model probability parameters, this causal extension of the model class is particularly straightforward, allowing us to utilise the model for a controlled environment.

We proceed to developing our new model, the DYNAMIC STAGED TREE. In Section 2 we formally define the necessary concepts. In Section 3 we develop a multi-process model for the dynamic staged tree that can be used to make one-step ahead predictions. In Section 4 we extend the multi-process model to causal analyses on the dynamic staged tree. We end in Section 5 by applying our analyses to some real educational data.

## 2. Concepts and definitions

### 2.1. Event tree

We begin the paper with some definitions; see Smith and Anderson (2008) for more details about these concepts.

Let  $T = (V(T), E(T))$  be a directed tree where  $V(T)$  is its node set and  $E(T)$  its edge set. Let  $L(T)$  be the set of LEAF NODES and  $S(T) = \{v : v \in V(T) \setminus L(T)\}$  be the set of SITUATIONS of  $T$ . Let  $\lambda(v, v')$  be the path from node  $v \in S(T)$  to node  $v' \in V(T)$  (if it exists), and let  $\Lambda(v, T) = \{\lambda(v, v') : v' \in L(T)\}$ , the set of paths from  $v$  to a leaf node. Let  $\mathbb{X} = \Lambda(v_0, T)$ , where  $v_0$  is the root node of  $T$ , so that  $\mathbb{X}$  is the set of root-to-leaf paths of  $T$ . Each path  $X \in \mathbb{X}$  is an ATOMIC EVENT, corresponding to a possible unfolding of events through time by using the partial ordering induced by the paths.

Let  $\mathbb{X}(v)$  denote the set of children of  $v \in S(T)$ . In an EVENT TREE (ET), each situation  $v \in S(T)$  has an associated random variable  $X(v)$  with sample space  $\mathbb{X}(v)$ , defined conditional on having reached  $v$ . The distribution of  $X(v)$  is determined by the PRIMITIVE PROBABILITIES  $\theta(v) = \{\theta(v, v') = p(X(v) = v') : v' \in \mathbb{X}(v)\}$ .

With random variables on the same path being mutually independent, the joint probability of events on a path can be calculated by multiplying the appropriate primitive probabilities together. Each primitive probability  $\theta(v, v')$  is a colour for the directed edge  $e = (v, v')$ , so that we let  $\pi(e) := \theta(v, v')$ .

### 2.2. Staged trees

Starting with an event tree  $T$ , define a FLORET of  $v \in S(T)$  as

$$\mathcal{F}(v) = (V(\mathcal{F}(v)), E(\mathcal{F}(v))) \quad (2)$$

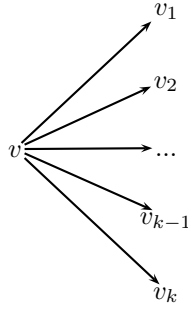


Figure 1: A floret of  $v \in S(T)$ . This subtree represents both the random variable  $X(v)$  and its state space  $\mathbb{X}(v)$ .

**Fig. 2.** A floret of  $v \in S(T)$ . This subtree represents both the random variable  $X(v)$  and its state space  $\mathbb{X}(v)$ .

where  $V(\mathcal{F}(v)) = v \cup \mathbb{X}(v)$  and  $E(\mathcal{F}(v)) = \{e \in E(T) : e = (v, v') : v' \in \mathbb{X}(v)\}$ . The floret of a vertex  $v$  is thus a sub-tree consisting of  $v$ , its children, and the edges connecting  $v$  and its children, as shown in Figure 2. This represents, as defined in section 2.1, the random variable  $X(v)$  and its sample space  $\mathbb{X}(v)$ .

Two situations  $v, v' \in S(T)$  are said to be in the same **STAGE**  $u$  if and only if  $X(v)$  and  $X(v')$  have the same distribution under a bijection

$$\psi_u(v, v') : \mathbb{X}(v) \rightarrow \mathbb{X}(v') \quad (3)$$

It follows that one necessary condition for  $v$  and  $v'$  to be in the same stage is that  $|\mathbb{X}(v)| = |\mathbb{X}(v')|$ , i.e.  $v$  and  $v'$  have the same number of children. In particular,  $\psi_u(v, v)$  is the identity function for any stage  $u$  that contains  $v$ , including  $u = \{v\}$ .

The set of stages (or **STAGING**) of  $T$  is written  $U(T)$ . It is clear that  $U(T)$  is a partition of  $S(T)$ . The set  $S(T)$  itself can be thought of as the **TRIVIAL STAGING**.

Finally, a **STAGED TREE**  $ST(T, U(T))$  is constructed from  $T$  by letting  $V(ST) = V(T)$  and  $E(ST) = E_d(ST) \cup E_u(ST)$ , where  $E_d(ST)$  and  $E_u(ST)$  are constructed as follows:

- $E_d(ST)$  is identical to  $E(T)$  except that two edges  $(v, v^*)$ ,  $(v', v'^*)$  are given the same colour if and only if  $v^* \mapsto v'^*$  under some  $\psi_u(v, v')$  as defined above;
- $E_u(ST)$ : for every  $v, v' \in S(T)$ , an undirected edge between  $v$  and  $v'$  is drawn if and only if  $X(v)$  and  $X(v')$  have the same distribution.

It is easily shown (Smith and Anderson, 2008) that BNs over finite discrete random variables are an important but small subclass of staged trees.

### 3. Prediction with dynamic staged trees

Let  $T$  be an event tree whose topology is known and fixed in time, but with an uncertain and possibly dynamic probability distribution over its structure. Let the set of situations of  $T$ ,  $S(T)$ , be denoted by  $S = \{v_1, \dots, v_{|S|}\}$ .

At each time point  $t = 1, \dots, \tau$ , we wish to predict  $x_t(v, v')$  for all  $v' \in \mathbb{X}(v)$  for all  $v \in S$ , where  $x_t(v, v')$  is the number of times  $X(v) = v'$  at time  $t$ . Let  $\mathbf{x}_t = (x_t(v))_{v \in S}$

where  $x_t(v) = (x_t(v, v'))_{v' \in \mathbb{X}(v)}$ . Then at every time  $t$  we need to construct a probability distribution over the possible values of  $\mathbf{x}_t$  conditional on all previous observations  $\mathbf{x}^{t-1} = (\mathbf{x}_1, \dots, \mathbf{x}_{t-1})$ . The marginal joint distribution  $P(\mathbf{x}^t)$  over time of the full data set can be represented as a product of the one-step ahead predictive probabilities  $P(\mathbf{x}_t | \mathbf{x}^{t-1})$ . Bayes factors associated with different models can then be expressed as a function of these quantities. It is interesting to note that this factorisation corresponds to the prequential likelihood described by Dawid (1984) used for comparing probabilistic forecasting systems.

The probability distribution of  $\mathbf{x}_t | \mathbf{x}_{t-1}$  can be written parametrically as a function of  $\theta_t$ , the values of  $\theta(v)$  for all  $v \in S$  at time  $t$ , so that

$$P(\mathbf{x}_t | \mathbf{x}^{t-1}) = \int_{\Theta_t} P(\mathbf{x}_t | \theta_t, \mathbf{x}^{t-1}) P(\theta_t | \mathbf{x}^{t-1}) d\theta_t \quad (4)$$

$\theta_t$  is unknown in the general case. One way to specify the distribution of  $\theta_t$  is to assume the process can be described by a DYNAMIC STAGED TREE. We define a dynamic staged tree to be an event tree where at each time point  $t = 1, \dots, \tau$  (where  $\tau$  can be finite or infinite) an independent sampling over  $\mathbb{X}$  occurs but with a possibly different staging  $U_t(T)$ .

If  $v, v' \in S(T)$  are in the same stage  $u$  in a partition  $U$  at time  $t$  then we assume that

$$\theta_t(v) = \theta_t(v') \triangleq \theta_t(u) \quad (5)$$

With these assumptions, the distribution of  $\theta_t$  under a staging  $U_t$  can be written as the product of the distribution of each stage's parameters:

$$P(\theta_t | U_t, \mathbf{x}^{t-1}) = \prod_{u \in U_t} P(\theta_t(u) | U_t, \mathbf{x}^{t-1}) \quad (6)$$

Therefore equation (4) can be written as

$$\begin{aligned} P(\mathbf{x}_t | \mathbf{x}^{t-1}) &= \sum_{U_t \in \mathcal{U}} \int_{\Theta_t} P(\mathbf{x}_t | \theta_t, U_t, \mathbf{x}^{t-1}) P(\theta_t | U_t, \mathbf{x}^{t-1}) P(U_t | \mathbf{x}^{t-1}) d\theta_t \quad (7) \\ &= \sum_{U_t \in \mathcal{U}} \int_{\Theta_t} \left( P(\mathbf{x}_t | \theta_t, U_t, \mathbf{x}^{t-1}) P(U_t | \mathbf{x}^{t-1}) \prod_{u \in U_t} P(\theta_t(u) | U_t, \mathbf{x}^{t-1}) \right) d\theta_t \quad (8) \end{aligned}$$

So to carry out a one-step ahead forecast on the system three probability distributions must be specified: the sampling distribution  $P(\mathbf{x}_t | \mathbf{x}^{t-1}, \theta_t, U_t)$ , the stage parameter distributions  $P(\theta_t(u) | U_t, \mathbf{x}^{t-1})$ , and the staging distributions  $P(U_t | \mathbf{x}^{t-1})$ . We show below how this can be achieved for each term in turn.

### 3.1. The sampling distributions

Under complete sampling the distribution of  $X(v)$  for any situation  $v \in S$  is conditionally independent of any other quantity given  $\theta(v)$ . In particular, this means that the distributions of  $X(v)$  and  $X(v')$  for two situations  $v, v' \in S, v \neq v'$ , are assumed to be independent conditional on  $\theta(v), \theta(v')$ .

This does not necessarily apply to  $x_t(v)$ , because the distribution of the number of samples  $N_t(v)$  from  $X(v)$  at time  $t$  is unknown in the general case. We assume here,

however, that for all situations  $v$  bar the root node  $v_0$  that  $N_t(v)$  equals the value of  $x_t(v^*, v)$ , where  $v^*$  is the situation such that  $v \in \mathbb{X}(v^*)$ , i.e. where  $v^*$  is the parent node of  $v$ . We discuss the setting of  $N_t(v_0)$  shortly.

We can therefore write  $P(\mathbf{x}_t \mid \boldsymbol{\theta}_t, U_t, \mathbf{x}^{t-1})$  as

$$P(\mathbf{x}_t \mid \boldsymbol{\theta}_t, U_t, \mathbf{x}^{t-1}) = \sum_{N_t(v_0)} P(\mathbf{x}_t \mid N_t(v_0), \boldsymbol{\theta}_t, U_t, \mathbf{x}^{t-1}) P(N_t(v_0) \mid \boldsymbol{\theta}_t, U_t, \mathbf{x}^{t-1}) \quad (9)$$

$$= \sum_{N_t(v_0)} \left( \left[ \prod_{v \in S} P(x_t(v) \mid \boldsymbol{\theta}_t(v), x_t(v^*, v)) \right] P(N_t(v_0) \mid \boldsymbol{\theta}_t, U_t, \mathbf{x}^{t-1}) \right) \quad (10)$$

$$= \sum_{N_t(v_0)} \left( \left[ \prod_{v \in S} \mathbb{I}_{\{\sum x_t(v, v') = x_t(v^*, v)\}} \prod_{v' \in \mathbb{X}(v)} \theta_t(v, v')^{x_t(v, v')} \right] P(N_t(v_0) \mid \boldsymbol{\theta}_t, U_t, \mathbf{x}^{t-1}) \right) \quad (11)$$

where  $\mathbb{I}_A$  is the indicator variable for an event  $A$  and  $x_t(v^*, v_0)$  is understood to mean  $N_t(v_0)$ .

The modelling of the distribution of  $N_t(v_0)$  depends on the details of the system under consideration. One common scenario is when  $N_t(v_0)$  is believed to be independent of all other system parameters apart from, at most, values of  $N_s(v_0)$  for  $s < t$ . One approach in this case is to model  $N_t(v_0)$  as a Poisson variable with parameter  $\lambda$ , where  $\lambda$  can either be constant or itself given a conjugate prior of Gamma( $\alpha_\lambda, \beta_\lambda$ ) at time 1.

When  $N_t(v_0)$  is known, equation (11) becomes

$$P(\mathbf{x}_t \mid \boldsymbol{\theta}_t, U_t, \mathbf{x}^{t-1}) = \prod_{v \in S} \left[ \mathbb{I}_{\{\sum x_t(v, v') = x_t(v^*, v)\}} \prod_{v' \in \mathbb{X}(v)} \theta_t(v, v')^{x_t(v, v')} \right] \quad (12)$$

where  $x_t(v^*, v_0)$  should again be read as  $N_t(v_0)$ .

### 3.2. The stage parameter distributions

As with every aspect of the model, the specification of the probability distribution over the floret parameters for each possible stage should be tailored to the scenario at hand. In many cases, however, it is possible to characterise the distribution from some common qualitative modelling assumptions.

Consider first the trivial staging  $U_t = S$ . It is shown in Freeman and Smith (2009) that if it is assumed that the relative rates of the root-to-leaf paths are independent, then the additional assumption of mutual independence of the floret distributions implies that each non-trivial floret's distribution must be Dirichlet. Therefore, denoting its set of hyperparameters as  $\alpha_t(v) = (\alpha_t(v, v'))_{v' \in \mathbb{X}(v)}$ , the density of  $\boldsymbol{\theta}_t(v) \mid U_t = S, \mathbf{x}^{t-1}$  for a non-trivial floret  $v \in S$  is

$$f_{\boldsymbol{\theta}_t(v)}(\boldsymbol{\theta}_t(v) \mid U_t = S, \mathbf{x}^{t-1}) = \Gamma \left( \sum_{v' \in \mathbb{X}(v)} \alpha_t(v, v') \right) \prod_{v' \in \mathbb{X}(v)} \frac{\theta_t(v, v')^{\alpha_t(v, v') - 1}}{\Gamma(\alpha_t(v, v'))} \quad (13)$$

for  $\sum_{v' \in \mathbb{X}(v)} \theta_t(v, v') = 1$  and  $\alpha_t(v, v') > 0$  for all  $v' \in \mathbb{X}(v)$ , and 0 otherwise.



Now consider a staging  $U$  that is not a trivial partition of  $S$ . In Freeman and Smith (2009) we show that requiring MARGIN EQUIVALENCY to hold for its stages  $u \in U$  characterises the prior on the floret distributions. A stage  $u$  has margin equivalency when

$$P(X(u) \mid \theta, U) = P(X(u) \mid \theta, S). \quad (14)$$

where  $X(u)$  is the random variable with sample space  $\bigcup_{v' \in v_u} \{v' \cup \{\bigcup_{v \in u} \psi(v_u, v)(v')\}\}$ , i.e. the edge equivalence classes under a stage. This property is analogous to that of parameter modularity for Bayesian networks (Heckerman, 1999). With the distribution for florets in  $S$  given above, this implies that the prior probability of  $\theta_t(u) \mid U_t = U, \mathbf{x}^{t-1}$  has a Dirichlet distribution too, with hyperparameters that are sums of the corresponding hyperparameters under  $S$  of the constituent florets:

$$f_{\theta_t(u)}(\theta_t(u) \mid U_t = U, \mathbf{x}^{t-1}) = \Gamma \left( \sum_{v' \in \mathbb{X}(v_u)} \bar{\alpha}_t(u, v') \right) \prod_{v' \in \mathbb{X}(v_u)} \frac{\theta_t(u, v')^{\bar{\alpha}_t(u, v') - 1}}{\Gamma(\bar{\alpha}_t(u, v'))} \quad (15)$$

where  $v_u$  is any situation in  $u$ ,  $\theta_t(u, v')$  are the elements of the vector  $\theta_t(u)$  and  $\bar{\alpha}_t(u, v') = \sum_{v: v \in u} \alpha_t(v, \psi_u(v_u, v)(v'))$ . Informally, equation (15) says that the hyperparameter vector for all of the floret distributions of the situations in stage  $u$  is equal to the sum of the hyperparameter vectors of the floret distributions under  $S$ .

With margin equivalency and independence between the floret distributions under  $S$ , the floret distributions under different stagings for stages composed of the same situations will always be the same. Therefore the probability distributions for a stage's parameters (13) and (15) depend only the composition of the stage and not on the rest of the staging. This property is useful since it allows us to discuss the characteristics of stage clusters of variable groups without reference to the partition in which they appear. This makes individual models much simpler to explain. It also reduces the computational complexity in calculating (13) and (15) because they will not be dependent on staging.

As every  $\theta_t(u)$  is conditionally independent of all other quantities given  $\alpha_t(u) = (\alpha_t(v))_{v \in u}$ , setting  $P(\theta_t \mid U_t, \mathbf{x}^{t-1})$  only requires the setting of  $\alpha_t(v)$  for each  $v \in S$  for every  $t$ . This task can be simplified further by assuming

$$f_{t+1, v}(\theta) = \mathcal{T}(f_{t, v}^*(\theta)) \quad (16)$$

for some function  $\mathcal{T}$  for all  $t > 1$ , where  $f_{t, v}(\theta)$  is the density of  $\theta_t(v) \mid \mathbf{x}^{t-1}, U_t = S$  as given in equation (13), and  $f_{t, v}^*(\theta)$  is the density of  $\theta_t(v) \mid \mathbf{x}^t, U_t = S$ , so that for every  $v \in S$  only  $\alpha_1(v)$  needs to be set.

The simplest choice of  $\mathcal{T}$  is the identity functional, so that  $f_{t+1, v}(\theta) = f_{t, v}^*(\theta)$  for  $t > 1$ . With  $f_{t, v}(\theta)$  as given in equation (13) and  $P(x_t(v) \mid \theta_t(v)) \propto \prod_{v' \in \mathbb{X}(v)} \theta_t(v, v')^{x_t(v, v')}$  as given by equation (11), Bayes' theorem requires

$$f_{\theta_t(v)}^*(\theta_t(v) \mid \mathbf{x}^t) = \Gamma \left( \sum_{v' \in \mathbb{X}(v)} \alpha_t^*(v, v') \right) \prod_{v' \in \mathbb{X}(v)} \frac{\theta_t(v, v')^{\alpha_t^*(v, v') - 1}}{\Gamma(\alpha_t^*(v, v'))} \quad (17)$$

where  $\alpha_t^*(v, v') = \alpha_t(v, v') + x_t(v, v')$ , and so

$$\alpha_{t+1}(v) = \alpha_t(v) + x_t(v). \quad (18)$$

As equation (18) is true for all  $t > 1$ ,  $\alpha_t(v)$  can be written as a function of only  $\alpha_1(v)$  and  $x^{t-1}(v)$ ,

$$\alpha_t(v) = \alpha_1(v) + \sum_{\tau=1}^{t-1} x_\tau(v) \tag{19}$$

for all  $v \in S$ .

Letting  $\mathcal{T}$  be the identity functional reflects a modelling assumption that the underlying probabilities associated with each stage do not evolve from year to year for a given staged tree. Sometimes this will be too strong an assumption to make. In this case, a weaker set of assumptions are needed which will represent the fact that there is an “information drift” between the time points. This will also guard against spurious jumps in the model probabilities from expected model drift.

One way to characterise  $\mathcal{T}$  to meet this need is provided by the POWER STEADY MODEL (Smith, 1979, 1981, 1992). It was shown by Smith (1979) that if, loosely speaking, it is assumed that the Bayes decision under a step loss function would stay the same over time if no more information was gathered about the system but that the expected loss of the decision increases due to increasing uncertainty, then it is required that

$$f_{t+1,v}(\theta) \propto (f_{t,v}^*(\theta))^k \tag{20}$$

for some  $0 < k \leq 1$ .

This transition function has a number of appealing properties. Firstly, when applied to the joint distribution  $P(\theta_t | U_t, \mathbf{x}^{t-1})$  of the  $\theta$  under any staging  $U_t$ , the floret independence structure is kept intact.

Secondly, it can be shown that use of the steady model guards against misspecified priors, making predictions more robust. Let the LOCAL DE ROBERTIS MEASURE  $DR_A$  be defined as follows (Smith and Daneshkhah, 2010):

$$d_A^L(f, g) = \sup_{\theta, \phi \in A} \{(\log f(\theta) - \log g(\theta)) - (\log f(\phi) - \log g(\phi))\} \tag{21}$$

for any  $A \in \Theta$ . Smith and Daneshkhah (2010) show that the local de Robertis measure is a separation measure where its separations do not change under Bayesian updating (Smith and Daneshkhah, 2010). It therefore represents artifacts of the model that cannot be changed by observation. It can easily be shown that, where  $f^* \propto f^k$  and similarly for  $g$ ,

$$d_A^L(f^*, g^*) = k(d_A^L(f, g)), \tag{22}$$

Thus using the steady model brings distributions closer together when  $0 < k \leq 1$ . In this sense steady models tend to be robust against initial prior misspecification, if we see  $f$  as the prior used in the analysis and  $g$  as the “true” prior. See Smith and Rigat (2008) for further details.

A similar result can be shown for Kullback-Leibler (KL) distances (Kullback and Leibler, 1951): recall that for two densities  $f$  and  $g$  the KL distance is given by

$$d_{KL}(f; g) = \int (\log f(\theta) - \log g(\theta))g(\theta)d\theta$$

and that the entropy  $H$  of a density is given by

$$H(f) = - \int f(\theta) \log f(\theta)d\theta$$

Let  $f_1, f_2$  be any two densities such that  $H(f_1) = H(f_2)$ . Then

$$d_{KL}(p_{t+1}; f_1) - d_{KL}(p_{t+1}; f_2) = k(d_{KL}(p_t; f_1) - d_{KL}(p_t; f_2)) \quad (23)$$

where  $p_t$  is the density of the stage parameters at time  $t$ , both after observing  $\mathbf{x}^t$ , so that  $p_{t+1} \propto (p_t)^k$ . Equation (23) says that the distance between the density of the stage parameters under any staging and two arbitrary densities with the same entropy decreases by a fixed proportion at each time step, again indicating a robustness to prior mis-specification.

Thirdly, with  $\alpha_t^*(v) = \alpha_t(v) + x_t(v)$ , equation (20) implies that  $\theta_{t+1}(v)$  is still distributed Dirichlet if  $\theta_t(v)$  is Dirichlet but with the hyperparameters of the distribution now given by the values

$$\alpha_{t+1}(v, v') = k\alpha_t(v, v') + kx_t(v, v') - k + 1 \quad (24)$$

Solving this recurrence relation for a constant  $k$  yields

$$\alpha_t(v, v') = k^{t-1}(\alpha_1(v, v') - 1) + \sum_{\tau=1}^{t-1} k^{t-\tau} x_\tau(v, v') + 1 \quad (25)$$

which heuristically can be seen as weighting recent observations more heavily for the setting of the latest prior.

Each situation can have its own  $k$ ,  $k(v)$ , and it might be desired that this  $k(v)$  be different for different  $t$ , for example when an external intervention in the system occurs.

We note that the use of the power steady model has a long history with Dirichlet distributions (e.g. in Smith (1979); Queen et al. (1994); Cowell et al. (1999)) and more generally (e.g. Ibrahim and Chen (2000); Rigat and Smith (2009)), and has also been used in Bayesian forecasting under the alternative name of exponential forgetting (Raftery et al., 2010). Here we use the power steady model as a justifiable and conjugate method for making inference about tree models whose floret probabilities evolve.

### 3.3. *The staging distributions*

We have allowed for drift over time in the values of probabilities associated with the conditional independence structure implicit in a staged tree model. However, we also want to allow for the possibility that the tree stagings themselves — and not just their parameters — evolve in time. It is unfeasible and usually unnecessary to model all possible changes of this partition space; in most applications it is appropriate to assume that changes in stage structure will be small in number and occur locally.

We therefore propose a dynamic model for the staged trees analogous to the Class 2 Multi-process Models used for dynamic linear models (DLMs) (Harrison and Stevens, 1976; West and Harrison, 1997). This was developed for the case where “no single [model] adequately describes what might happen to the process in the next time interval” (West and Harrison, 1997). We describe the C2MPM here as it applies to the staged tree setting.

Let  $\mathbf{U}$  be the set of all possible stagings of  $T$ , and for each  $U \in \mathbf{U}$  and  $t > 1$  let  $\pi_t(U) = P(U_t = U \mid \mathbf{x}^{t-1})$ . There is obviously a large class of possible specifications for  $\pi_t(U)$ , but we note the three “practically important possibilities” mentioned by (West and Harrison, 1997):

- (1) Fixed model selection probabilities, such that

$$\pi_t(U) = \pi(U) \quad \text{for all } t \geq 1 \quad (26)$$

Here, one needs to only specify one prior over  $\mathbf{U}$ . This prior remains fixed through time and isn't changed by observations.

- (2) First-order Markov probabilities, where fixed transition probabilities between the models

$$\pi(U | U') = P(U_t = U | U_{t-1} = U') \quad (27)$$

are specified a priori, so that

$$\pi_t(U) = \sum_{U' \in \mathbf{U}} \pi(U | U') P(U_{t-1} = U' | \mathbf{x}^{t-1}) \quad (28)$$

Some initial prior distribution over the staged tree space,  $\pi_1(C)$ , would need to be set.

- (3) Higher-order Markov probabilities, where the probabilities of the stagings at time  $t$  additionally depend on the stagings at  $t - 2$ ,  $t - 3$ , etc.

While the first possible modelling strategy, of fixed staging probabilities, is much the simpler one (and the option used by West and Harrison for exposition), the second and third strategies are often going to be more accurate reflections of experts' beliefs. We show here how to implement first-order Markov transitions between stagings.

A common assumption will be that  $\pi(U | U')$  is larger the "closer"  $U$  is to  $U'$  in some sense, so that the underlying process is unlikely to change too much over a short period of time. If  $\pi(U | U') = 0$  for some  $U \in \mathbf{U}$ , this has the advantage of reducing the number of terms in equation (28).

We therefore require a metric on  $\mathbf{U}$  and then let  $\pi(U | U')$  be a function of this metric. Any intuitive metric on general sets of partitions can be used, e.g. that of Meilă (2007). A simple metric that we use here can be derived from the Hasse diagram of the lattice of partitions of  $S$  under the relation "finer than" (see Stanley (1997) for a detailed overview of such lattice terminology). The Hasse diagram for  $|S| = 4$  is shown in Figure 3.

The length of the shortest path between two partitions on the Hasse diagram is a metric on the partition space of  $S$ , and we call it  $\ell$  here. A distance of  $\ell = 1$  represents the division of a stage or the merging of two stages. One possible way to set  $\pi(U | U')$  based on this metric is

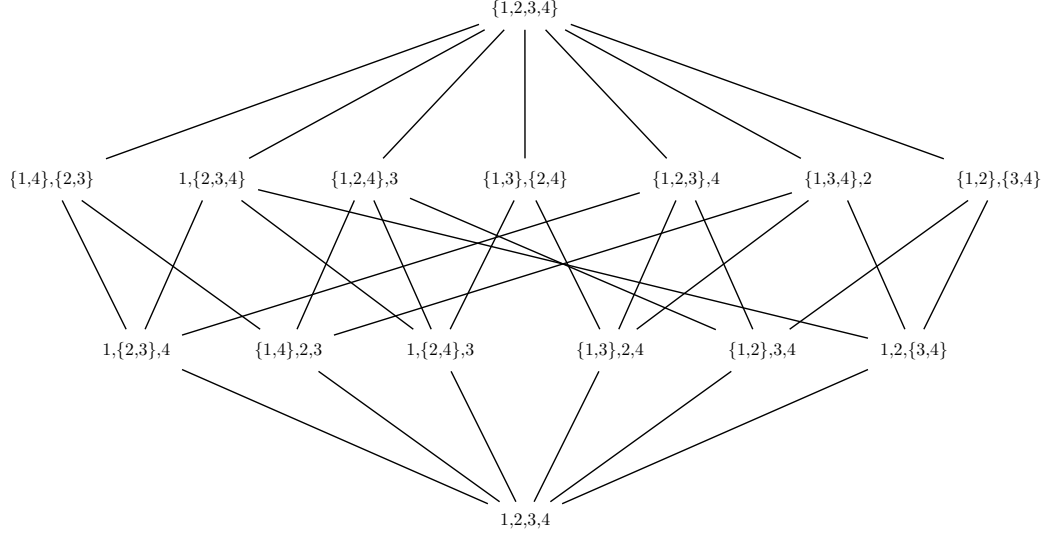
$$\pi(U | U') = \begin{cases} \rho & \text{if } U = U' \\ |B_\epsilon(U')|^{-1}(1 - \rho) & \text{if } 0 < \ell(U, U') \leq \epsilon, \\ 0 & \text{otherwise} \end{cases} \quad (29)$$

where  $B_\epsilon(U') = \{U \in \mathbf{U} : \ell(U, U') \leq \epsilon, U \neq U', U \in \mathbf{U}\}$  is an  $\epsilon$ -ball of stagings around  $U'$  under the  $\ell$  metric. This represents a belief that the underlying symmetry process changes only locally and slowly. If more radical changes in the symmetry process are taking place due to external intervention in the system then the methodology in Section 4.1 can be deployed.

The other term in (27),  $P(U_{t-1} = U' | \mathbf{x}^{t-1})$ , can be calculated for each  $U_{t-1}$  using Bayes' theorem:

$$P(U_{t-1} = U' | \mathbf{x}^{t-1}) \propto P(\mathbf{x}_{t-1} | U_{t-1} = U') P(U_{t-1} = U' | \mathbf{x}^{t-2}) \quad (30)$$

$$= \frac{P(\mathbf{x}_{t-1} | U_{t-1} = U') P(U_{t-1} = U' | \mathbf{x}^{t-2})}{\sum_{U' \in \mathbf{U}} P(\mathbf{x}_{t-1} | U_{t-1} = U') P(U_{t-1} = U' | \mathbf{x}^{t-2})} \quad (31)$$



**Fig. 3.** The Hasse diagram of the lattice of partitions of  $S$  when  $|S| = 4$

The  $P(U_{t-1} = U' | \mathbf{x}^{t-2})$  terms on the right-hand side of (31) will be already be available at time  $t - 1$ . The term  $P(\mathbf{x}_{t-1} | U_{t-1} = U')$ , meanwhile, can be calculated as follows, using equations (11) and (15) at time  $t - 1$ :

$$P(\mathbf{x}_{t-1} | U_{t-1} = U') = \int_{\Theta_{t-1}} P(\mathbf{x}_{t-1} | \boldsymbol{\theta}_{t-1}, U_{t-1} = U') P(\boldsymbol{\theta}_{t-1} | U_{t-1} = U') d\boldsymbol{\theta}_{t-1} \quad (32)$$

$$\propto \int_{\Theta_{t-1}} \prod_{u \in U'} \left( \Gamma \left( \sum_{v' \in \mathbb{X}(v_u)} \bar{\alpha}_{t-1}(u, v') \right) \prod_{v' \in \mathbb{X}(v_u)} \frac{\theta_{t-1}(u, v')^{\bar{\alpha}_{t-1}^*(u, v') - 1}}{\Gamma(\bar{\alpha}_{t-1}(u, v'))} \right) d\boldsymbol{\theta}_{t-1} \quad (33)$$

$$= \prod_{u \in U'} \left( \frac{\Gamma \left( \sum_{v' \in \mathbb{X}(v_u)} \bar{\alpha}_{t-1}(u, v') \right)}{\Gamma \left( \sum_{v' \in \mathbb{X}(v_u)} \bar{\alpha}_{t-1}^*(u, v') \right)} \prod_{v' \in \mathbb{X}(v_u)} \frac{\Gamma(\bar{\alpha}_{t-1}^*(u, v'))}{\Gamma(\bar{\alpha}_{t-1}(u, v'))} \right) \quad (34)$$

where  $v_u$  is any situation in  $u$ ,  $\bar{\alpha}_{t-1}^*(u, v') = \bar{x}_{t-1}(u, v') + \bar{\alpha}_{t-1}(u, v')$ , where  $\bar{x}_{t-1}(u, v') = \sum_{v: v \in u} x_{t-1}(v, \psi_u(v_u, v)(v'))$  and  $\bar{\alpha}_{t-1}$  is as defined in equation (15).

The number of terms in (28) can be reduced further by setting the values of  $P(U_{t-1} = U' | \mathbf{x}^{t-1})$  below a threshold  $q$  as zero and normalising the remaining probabilities to ensure they still sum to 1. A similar approach advocated by Madigan and Raftery (1994) as ‘‘Occam’s window’’ is to discard models  $U'$  that are not in the set

$$U_t^* = \left\{ U_t \in U : \frac{P(U_t | \mathbf{x}^t)}{\max_U P(U | \mathbf{x}^t)} \leq q \right\} \quad (35)$$

for some  $0 < q < 1$ , i.e., to only keep models where the Bayes factor between them and the most probable model a posteriori are above a certain threshold. This has the advantage of guaranteeing that at least one model will be kept.

### 3.4. One-step-ahead prediction

Equation (7) can now be written, using the foregoing, as

$$\begin{aligned}
P(\mathbf{x}_t | \mathbf{x}^{t-1}) &= \sum_{U_t \in \mathcal{U}} \int_{\Theta_t} \left( \sum_{U_{t-1} \in \mathcal{U}} \pi(U_t | U_{t-1}) P(U_{t-1} | \mathbf{x}^{t-1}) \right) \left( \sum_{N_t(v_0)} P(N_t(v_0) | \boldsymbol{\theta}_t, U_t, \mathbf{x}^{t-1}) \right. \\
&\quad \left. \cdot \prod_{u \in U_t} \mathbb{I}_A \left[ \Gamma \left( \sum_{v' \in \mathbb{X}(v_u)} \bar{\alpha}_t(u, v') \right) \cdot \prod_{v' \in \mathbb{X}(v_u)} \frac{\theta_t(u, v')^{\bar{\alpha}_t(u, v') + \bar{x}_t(u, v') - 1}}{\Gamma(\bar{\alpha}_t(u, v'))} \right] \right) d\boldsymbol{\theta}_t
\end{aligned} \tag{36}$$

where  $A$  is the event  $\forall v \in u \setminus v_0, \sum_{v'} x_t(v, v') = x_t(v^*, v)$ . If it is assumed that the distribution of  $N_t(v_0)$  depends only on  $\mathbf{x}^{t-1}$  then (36) can be further simplified to the closed-form solution

$$\begin{aligned}
P(\mathbf{x}_t | \mathbf{x}^{t-1}) &= \sum_{U_t \in \mathcal{U}} \left( \left( \sum_{U_{t-1} \in \mathcal{U}} \pi(U_t | U_{t-1}) P(U_{t-1} | \mathbf{x}^{t-1}) \right) \left( \sum_{N_t(v_0)} P(N_t(v_0) | \mathbf{x}^{t-1}) \right. \right. \\
&\quad \left. \left. \cdot \prod_{u \in U_t} \mathbb{I}_A \left[ \frac{\Gamma \left( \sum_{v' \in \mathbb{X}(v_u)} \bar{\alpha}_t(u, v') \right)}{\Gamma \left( \sum_{v' \in \mathbb{X}(v_u)} \bar{\alpha}_t^*(u, v') \right)} \prod_{v' \in \mathbb{X}(v_u)} \frac{\Gamma(\bar{\alpha}_t^*(u, v'))}{\Gamma(\bar{\alpha}_t(u, v'))} \right] \right) \right)
\end{aligned} \tag{37}$$

If  $N_t(v_0)$  is always known, then (37) can be simplified further to become

$$\begin{aligned}
P(\mathbf{x}_t | \mathbf{x}^{t-1}) &= \sum_{U_t \in \mathcal{U}} \left( \left( \sum_{U_{t-1} \in \mathcal{U}} \pi(U_t | U_{t-1}) P(U_{t-1} | \mathbf{x}^{t-1}) \right) \right. \\
&\quad \left. \cdot \prod_{u \in U_t} \mathbb{I}_A \left[ \frac{\Gamma \left( \sum_{v' \in \mathbb{X}(v_u)} \bar{\alpha}_t(u, v') \right)}{\Gamma \left( \sum_{v' \in \mathbb{X}(v_u)} \bar{\alpha}_t^*(u, v') \right)} \prod_{v' \in \mathbb{X}(v_u)} \frac{\Gamma(\bar{\alpha}_t^*(u, v'))}{\Gamma(\bar{\alpha}_t(u, v'))} \right] \right)
\end{aligned} \tag{38}$$

## 4. Causal intervention

With any forecasting system there is also an attendant need to consider the effects of external intervention in the system, including by the forecasters themselves (West and Harrison, 1989). This ensures that all relevant information is taken into account, increasing the accuracy of future forecasts.

The predicted effect of an intervention depends both on the nature of that intervention and the context in which it applies. Many interventions act only on certain local features of a model while leaving the other features of the model unchanged. We note that these types of interventions have now been extensively studied on non-dynamic BNs (Pearl, 2000b; Spirtes et al., 2001), which are called CAUSAL BAYESIAN NETWORKS (CBNs) in this context. Dynamic extensions of CBNs also exist (Queen and Smith, 1993; Eichler and Didelez, 2007; Queen and Albers, 2009). A Bayesian way of learning causal Bayesian networks is given by Cooper and Yoo (1999).

We believe that tree-based graphical models are very useful in general for carrying causal analysis. Due to the multiple representations of each variable in the graph — one for each possible path-history on parent variables — much more refined interventions in the system can be represented (Shafer, 1996). How causal hypotheses can be represented within the framework of static CEGs has been investigated by Thwaites and Smith (2006) and Thwaites et al. (2010).

We now show how causal analysis affects the one-step ahead forecast on a dynamic staged tree given by equations (36), (37), and (38) for two different types of intervention: one on the possible stagings on a tree  $T$  and one on the structure of the tree  $T$  itself.

#### 4.1. Intervention on the staging distribution

Suppose that at time  $t$  some situations will be moved into their own stage  $\mathbf{u}$ , leaving all other stages intact. For example, in our educational example, the exams for the second module might be tailored so that performance in the first module is no longer a predictor in how well students should perform in it. The one-step ahead forecasts can then be modified in the following way to reflect this intervention.

Recall that  $\pi_{t-1}^*(U) = P(U_{t-1} = U \mid \mathbf{x}^{t-1})$ . Let  $\pi_t^\dagger(U) = P(U_t = U \mid \mathbf{x}^{t-1}, I_t)$ , where  $I_t$  is the intervention described above. Then one approach to modelling the intervention is to set  $\pi_t^\dagger(U) = \pi_{t-1}^*(U)$  for each  $U \in \mathcal{U}$  such that  $\mathbf{u} \in U$ , and set  $\pi_t^\dagger(U^\dagger) = \pi_{t-1}^*(U)$  and  $\pi_t^\dagger(U) = 0$  for  $U \in \mathcal{U}$  such that  $\mathbf{u} \notin U$ , where  $U^\dagger$  is the same as  $U$  except that  $\mathbf{u} \in U^\dagger$  and other stages that contained situations  $v \in \mathbf{u}$  are reduced accordingly. The effect of this approach is to transfer the probability mass on the stagings where  $\mathbf{u} \notin U$  to stagings where  $\mathbf{u} \in U$ .

One issue that now arises is how the distribution of  $\boldsymbol{\theta}_t \mid U_t$  is affected. In the absence of further information, a good default is to use the steady model as in the idle system but with a lower value for the steady parameter  $k$ . This indicates that past data might not be as useful in helping to make predictions in this situation as under the idle system. We note that this is analogous to setting a higher variance on evolution parameters in dynamic linear models when forecasting after interventions is required for that model class (Section 1.2.2 of West and Harrison (1997)).

#### 4.2. Intervention on $T$

Recalling the event tree pictured in Figure 1, consider the case where at time  $t$  the course directors decide to eliminate the first module on the tree from course. This means that the marks that students would have got for this module are unknown from that time onwards, and therefore all of the data at time  $t$  for this module will be concentrated on the second (“NA”) edge of the  $v_1$  floret.

This type of intervention is analogous to the *do* operator introduced for CBNs by (Pearl, 2000a), where a random variable is forced to take a particular value with probability 1. The difference with CBNs is that staged trees allow a richer set of interventions on their structure, including letting an intervention take place at specific time and situations, and not merely changing the value of a variable under all circumstances.

We assume that the probability distributions on any unmanipulated florets remain unchanged, just as for CBNs manipulations are local Pearl (2000a). We will also assume that once an intervention is made, it endures thereon. We now describe how the learning framework outlined previously can be adapted to prediction after an intervention of this type

occurs.

Without loss of generality, say that at time  $t$  an intervention  $I_t(v, v')$  at situation  $v \in S$  occurs so that  $\theta_t(v, v')$  is equal to 1 for a specific  $v' \in \mathbb{X}(v)$  and to 0 for all other  $v' \in \mathbb{X}(v)$ . By the definition of the event tree, along with the causal assumptions, all other floret distributions are technically unchanged. However, notice that the probability of reaching any node in any  $\Lambda(v^*, T)$ , the sub-tree with  $v^*$  as the root node, is zero. It follows that the tree  $T$  is equivalent to the reduced tree  $T'$  where all  $\Lambda(v^*, T)$  are deleted and only the edge  $(v, v')$  remains in the floret  $\mathcal{F}(v)$ , and so the process can henceforth be considered to take place on this reduced tree  $T'$ .

The one-step ahead forecasts can now be calculated as before with a few modifications due the set of situations  $S$  changing; call this new set  $S^\dagger$ . Firstly, the distribution over  $U^\dagger$ , the new set of possible stagings, must be set. There are several possible choices here. In the absence of any other information, a good default is to let

$$P(U_t = U^\dagger \mid \mathbf{x}^{t-1}, I_t(v, v')) = P(U_{t-1} = U \mid \mathbf{x}^{t-1}), \quad (39)$$

where  $U^\dagger$  is the staging formed from  $U$  by replacing each stage  $u \in U$  with a new stage  $u^\dagger := u \setminus \{v^\dagger\}_{v^\dagger \in S \setminus S^\dagger}$ , and by splitting the stage  $u^\dagger \in U^\dagger$  that contains the intervention node  $v$  into  $u^\dagger \setminus v$  and  $v$ .

Secondly, the distributions of the stage parameters  $\theta_t(u)$  for any  $u$  need to be reconsidered. Under the causal assumptions considered here, interventions have only local effects, so a sensible default model is to let  $f_{\theta_t(u)}(\theta_t(u) \mid U_t = U, \mathbf{x}^{t-1}, I_t(v, v'))$  be calculated as before, i.e. as given in Equation (15).

Assuming that all of the other system characteristics, e.g. the steady model and the multinomial sampling, are intact post-intervention, the one-step ahead forecast (37) is adjusted to become

$$P(\mathbf{x}_t \mid \mathbf{x}^{t-1}, I_t(v, v')) = \sum_{U_t^\dagger \in U^\dagger} \left( \left( \sum_{U_{t-1} \in U} \pi^\dagger(U_t^\dagger \mid U_{t-1}) P(U_{t-1} \mid \mathbf{x}^{t-1}) \right) \left( \sum_{N_t(v_0)} P(N_t(v_0) \mid \mathbf{x}^{t-1}) \right. \right. \\ \left. \left. \cdot \prod_{u \in U_t^\dagger} \mathbb{I}_A \left[ \frac{\Gamma \left( \sum_{v' \in \mathbb{X}(v_u)} \bar{\alpha}_t(u, v') \right)}{\Gamma \left( \sum_{v' \in \mathbb{X}(v_u)} \bar{\alpha}_t^*(u, v') \right)} \prod_{v' \in \mathbb{X}(v_u)} \frac{\Gamma(\bar{\alpha}_t^*(u, v'))}{\Gamma(\bar{\alpha}_t(u, v'))} \right] \right) \right) \quad (40)$$

where  $\pi^\dagger(U_t^\dagger \mid U_{t-1}) = \pi(U_t \mid U_{t-1})$  by the argument above.

## 5. A simple educational example

In this section we illustrate how to carry out one-step ahead predictions with dynamic staged trees using 12 years' worth of exam marks for two first-year undergraduate modules. The underlying event tree used was that shown in Figure 1.

We made the following assumptions:

- (1)  $N_t(v_0)$  was known for all values of  $t$
- (2) The distribution over the root-to-leaf paths at time  $t = 1$  under  $U_1 = S$  was Dirichlet with all path hyperparameters equal to 1



**Table 1.** All possible stagings and their posterior probabilities at each time  $t$  for  $k = 0.9$ ,  $\rho = 0.9$ ,  $q = 0.2$  with  $P(U_1 = \{v_1, v_2, \{v_3, \dots, v_6\}, \{v_7, \dots, v_{10}\}\}) = 1$

Time	$U_t$	$P(U_t   x^t)$
1	1, 2, {3,4,5,6}, {7,8,9,10}	1
2	1, 2, 3, {4,5,6}, {7,8,9,10}	0.824
	1, 2, {3,4,5,6}, {7,9,10}, 8	0.175
3	1, 2, 3, {4,5,6}, {7,8,9,10}	0.766
	1, 2, 3, {4,5,6}, {7,10}, {8,9}	0.233
4	1, 2, 3, {4,5,6}, {7,8,9,10}	0.677
	1, 2, 3, {4,5,6}, {7,10}, {8,9}	0.322
5	1, 2, 3, {4,5,6}, {7,8,9,10}	0.328
	1, 2, 3, {4,5,6}, {7,10}, {8,9}	0.671
6	1, 2, 3, {4,5,6}, {7,10}, {8,9}	1
7	1, 2, 3, {4,5,6}, {7,10}, {8,9}	0.609
	1, 2, 3, {4,5,6}, {7,10}, 8, 9	0.390
8	1, 2, 3, {4,5,6}, {7,10}, {8,9}	0.304
	1, 2, 3, {4,5,6}, {7,10}, 8, 9	0.695
9	1, 2, 3, {4,5,6}, {7,10}, 8, 9	1
10	1, 2, 3, {4,5,6}, {7,10}, 8, 9	1
11	1, 2, 3, {4,5,6}, {7,10}, 8, 9	1
12	1, 2, 3, {4,5,6}, {7,10}, 8, 9	1

- (3) For the transitions between stagings we used the  $\ell$  metric with  $\epsilon = 1$ , i.e. only transitions between models with local changes were considered possible.

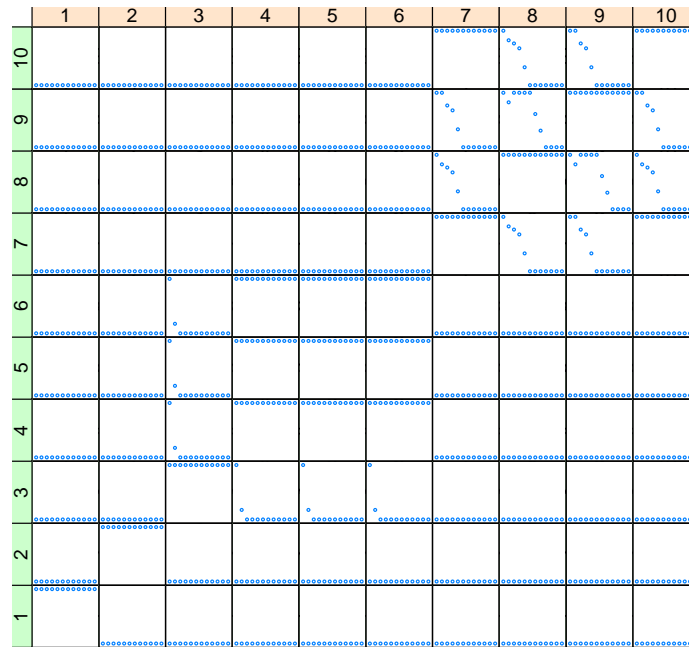
We present here the posterior probabilities  $P(U_t | \mathbf{x}^t)$  for the stagings after  $t = 1$  for each time  $t$  for different hyperparameter values, when analysed with and without an external intervention.

In a full analysis this application could be run over a distribution of the hyperparameters  $k$  (the steady model parameter),  $\rho$  (the probability of the underlying model not changing) and  $q$  (the Occam's window threshold), perhaps after taking account of an elicited prior over their possible values. However, to illustrate the efficacy of our methods rather than learn these hyperparameters it is better to hold them fixed so that we can better focus on the impact of various structured assumptions we learn about.

### 5.1. Analysis of the series without intervention

In Table 5.1 we present  $P(U_t | \mathbf{x}^t)$  for  $t = 1 \dots 12$  for the model where  $U_1 = \{v_1, v_2, \{v_3, \dots, v_6\}, \{v_7, \dots, v_{10}\}\}$  with probability 1 and  $k = 0.9$ ,  $\rho = 0.9$  and  $q = 0.2$ . The latter two parameter values ensure that few new models will be kept in the analysis, as the high value of  $\rho$  gives a low prior probability on transitions between stagings and this value of  $q$  makes the Occam's window set of equation (35) small. This speeds up the computation of the forecasts at the expense of possibly worse predictions through fewer stagings being included in the model averaging.

An alternative way of presenting this information is to plot how  $P(v_i, v_j \in u)$ , the probability that situations  $v_i, v_j$  are in the same stage, changes for increasing values of  $t$ . Figure 4 shows this for the information in Table 5.1.

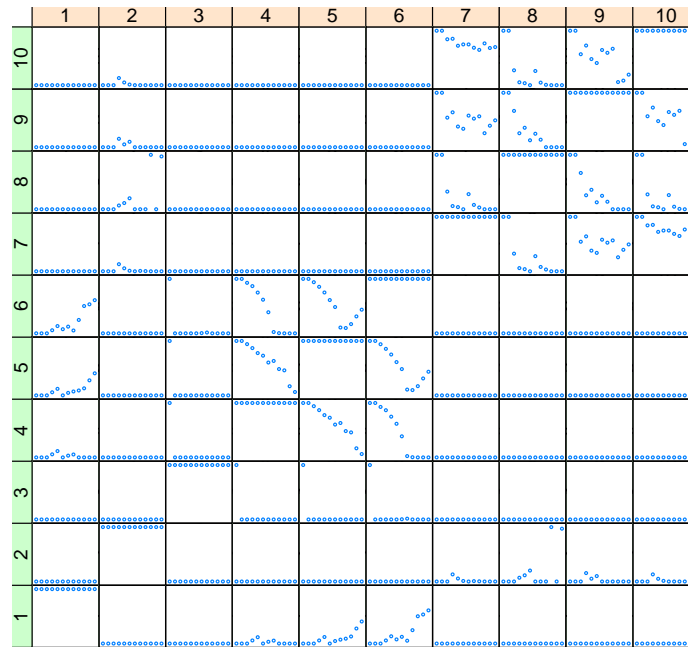


**Fig. 4.** Plots of probabilities that each pair of situations are in the same stage for different values of  $t$ , for the case when  $k = 0.9, \rho = 0.9, q = 0.2$  with  $P(U_1 = \{v_1, v_2, \{v_3, \dots, v_6\}, \{v_7, \dots, v_{10}\}\}) = 1$ , using the values in Table 5.1

To illustrate how the level of detail in the staging distribution changes as a function the hyperparameters, we carried out the analysis again with radically different values: we set  $k = 0.5$  (so that floret distributions are flattened more quickly and therefore past observations more heavily discounted),  $\rho = 0.25$  (so that the probability of moving between stagings is more likely), and  $q = 0.05$  (so that stagings with poorer Bayes factors relative to the most likely are kept in the analysis) with  $P(U_1 = \{v_1, v_2, \{v_3, \dots, v_6\}, \{v_7, \dots, v_{10}\}\}) = 1$  still assumed for consistency. The resulting matrix plot of probabilities of situations being in the same stage against time is as shown in Figure 5.

It can be seen from the latter figure that the analysis with the new hyperparameter values gives much the same qualitative description of the system as the more conservative hyperparameters at greater computational expense, with the pay-off of greater detail.

Some interesting characteristics of the system can be discerned from this initial exploratory analysis of this system. With regard to the situations concerning whether marks are available for a module or not,  $\theta(v_3)$  — the probability distribution for the second module’s marks being available given that the mark in the first module is itself missing — does not appear to be related to the others at any time point. Until  $t = 7$ ,  $v_4, v_5$  and  $v_6$ , the situations representing the probability of marks being missing in the second module after gaining a high, medium or low mark respectively in the first module, had high but falling probabilities of being in the same stage, implying that independence of the second module’s marks being missing from skill in the first module kept decreasing from a high point. At  $t = 8$ , in contrast, these probabilities are much lower, although the probability distributions of marks being missing after gaining a medium or low mark in the first module are



**Fig. 5.** Plots of probabilities that each pair of situations are in the same stage for different values of  $t$ , for the case when  $k = 0.5$ ,  $\rho = 0.25$ ,  $q = 0.05$  with  $P(U_1 = \{v_1, v_2, \{v_3, \dots, v_6\}, \{v_7, \dots, v_{10}\}\}) = 1$

deemed to become slightly more likely to be the same after that, with students performing well in the first module continuing to have a very different probability distribution for the missingness of their second module marks. We investigate a possible causal hypothesis that might explain what might have changed at  $t = 8$  in the next section.

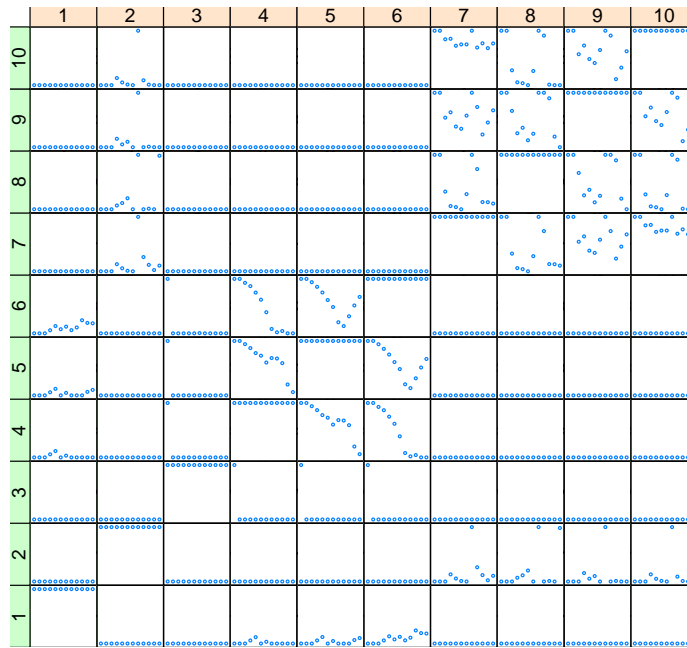
Another notable finding is that  $v_7$  and  $v_{10}$  — the situations concerning marks in the second module after getting a poor grade or having a missing mark in the first module, respectively — are always strongly related. It therefore appears that the second module marks of students who did poorly in the first module should be used to predict the second module performance of students whose first module marks are missing.

It is worth noting again that these detailed homogeneities would not have been as easily identifiable if the model class was restricted to Bayesian networks.

### 5.2. Analysis of the series after intervention

We also carried out an analysis with the latter parameters after a hypothesised causal intervention: we assumed that at  $t = 8$  the situations for the grades ( $v_2, v_7, v_8, v_9, v_{10}$ ) were put into the same stage. This could have happened, for example, because the modules were re-defined to be very similar in difficulty for students with different skills. The resulting matrix of probabilities of situations being in the same stage through time is shown in Figure 6.

It can be seen that the probabilities are not too different from those in Figure 5, but there are increased probabilities of  $v_8, v_9$  and  $v_{10}$  being in the same stage even for  $t > 8$ , which indicates slightly higher probabilities of dependence between the second module's grades



**Fig. 6.** Plots of probabilities that each pair of situations are in the same stage for different values of  $t$ , for the case when  $k = 0.5$ ,  $\rho = 0.25$ ,  $q = 0.05$  with  $P(U_1 = \{v_1, v_2, \{v_3, \dots, v_6\}, \{v_7, \dots, v_{10}\}\}) = 1$ , and situations  $v_2, v_7, v_8, v_9, v_{10}$  caused to be in the same stage at  $t = 8$

for students who performed differently in the first module under the causal hypothesis considered here.

### 6. Discussion

We have presented in this paper a new discrete time series modelling class, the dynamic staged tree, that is intuitive to use and suitable for carrying out causal analysis.

Obviously the class of models we define here can be usefully refined. In many potential applications we would like to allow for multiple possible trees at any time point. If the general class of event trees  $\mathbf{T}$  is required, then  $P(\mathbf{x}_t | \mathbf{x}^{t-1})$  can still be calculated as outlined in this paper but with the additional step of marginalising over the  $T \in \mathbf{T}$  such that  $P(T | \mathbf{x}^{t-1}) > 0$ , assuming the number of such  $T$  is tractable. If all that is required is the subclass of  $\mathbf{T}$  which is trees that are merely different partitionings of the same set of root-to-leaf path events, then assuming that the same root-to-leaf path events on different trees have the same probability, the foret distributions on all trees can be characterised as Dirichlet by the method used here. The method of assigning probabilities over the tree space in either case, or how those probabilities change over time, would still need to be resolved. We plan to explore this class in a later paper.

Another way of enlarging the model space is to allow for uncertainty in  $\psi_u(v, v')$ , i.e. which edges of two forets are coloured identically when their root nodes are in the same stage. The type of hypothesis this could capture includes a belief in the stability in values between different random variables. In our educational example, this would translate into

believing that the probability of getting the same grade in the second module as in the first one is the same for all grades.

We note that the number of possible  $\psi_u(v, v')$  for any pair  $v, v'$  is  $|\mathbb{X}(v)|!$ . Therefore to make the model search tractable in general either the number of possibilities must be restricted using contextual information or a local neighbourhood switching function, as with the staging space model in this paper.

One blind spot of the staged tree is its ignoring the time it takes for events to occur by modelling only which events occur. In applications like the educational one above, where the time of the events is known, or where the time to an event is not relevant to the probability of it occurring — as with constant hazard function models — this does not matter. For other systems, however, the time at which events occur are an extremely important part of the underlying process, and is the type of domain where event history analysis has been applied. We are currently investigating the link between staged tree models of the type described here and event history models which explicitly acknowledge this extra source of variation.

Finally, it appears that the dynamic staged tree process can be extended to model processes defined on continuous as well as discrete variables. Converting the leaf nodes on a tree into continuous sample spaces is trivial. When other variables are continuous then analogous conjugate models can be defined which describe hierarchical clustering models. We will report on these developments in a later paper.

## References

- Arjas, E. (1989). Survival models and martingale dynamics (with discussion and reply). *Scandinavian Journal of Statistics* 16(3), 177–225. ArticleType: primary\_article / Full publication date: 1989 / Copyright © 1989 Board of the Foundation of the Scandinavian Journal of Statistics.
- Boutilier, C., N. Friedman, M. Goldszmidt, and D. Koller (1996). Context-Specific independence in bayesian networks. In *Uncertainty in Artificial Intelligence*, pp. 115–123.
- Cooper, G. and C. Yoo (1999). Causal discovery from a mixture of experimental and observational data. In *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*, San Francisco, CA, pp. 116–12. Morgan Kaufmann.
- Cowell, R. G., A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter (1999). *Probabilistic Networks and Expert Systems*. Springer.
- Dawid, A. P. (1979). Conditional independence in statistical theory. *Journal of the Royal Statistical Society. Series B (Methodological)* 41(1), 1–31.
- Dawid, A. P. (1984). Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)* 147(2), 278–292. ArticleType: primary\_article / Issue Title: The 150th Anniversary of the Royal Statistical Society / Full publication date: 1984 / Copyright © 1984 Royal Statistical Society.
- Didelez, V. (2008, February). Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 70(1), 245–264.

- Eichler, M. and V. Didelez (2007). Causal reasoning in graphical time series models. In *Proceedings of the 23rd Annual Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers.
- Freeman, G. and J. Q. Smith (2009). Bayesian MAP model selection of chain event graphs. CRiSM 09-06, University of Warwick, Coventry.
- Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. New York: Springer.
- Gottard, A. (2007, November). On the inclusion of bivariate marked point processes in graphical models. *Metrika* 66(3), 269–287.
- Harrison, P. J. and C. F. Stevens (1976). Bayesian forecasting. *Journal of the Royal Statistical Society. Series B (Methodological)* 38(3), 205–247. ArticleType: primary\_article / Full publication date: 1976 / Copyright © 1976 Royal Statistical Society.
- Heckerman, D. (1999). A tutorial on learning with Bayesian networks. In M. I. Jordan (Ed.), *Learning in Graphical Models*, pp. 301–354. MIT Press.
- Ibrahim, J. G. and M. Chen (2000, February). Power prior distributions for regression models. *Statistical Science* 15(1), 46–60. ArticleType: primary\_article / Full publication date: Feb., 2000 / Copyright © 2000 Institute of Mathematical Statistics.
- Kullback, S. and R. A. Leibler (1951, March). On information and sufficiency. *The Annals of Mathematical Statistics* 22(1), 79–86. ArticleType: primary\_article / Full publication date: Mar., 1951 / Copyright © 1951 Institute of Mathematical Statistics.
- Madigan, D. and A. E. Raftery (1994, December). Model selection and accounting for model uncertainty in graphical models using Occam’s window. *Journal of the American Statistical Association* 89(428), 1535–1546. ArticleType: primary\_article / Full publication date: Dec., 1994 / Copyright © 1994 American Statistical Association.
- Meilä, M. (2007, May). Comparing clusterings—an information based distance. *Journal of Multivariate Analysis* 98(5), 873–895.
- Pearl, J. (2000a, June). Causal inference without counterfactuals: Comment. *Journal of the American Statistical Association* 95(450), 428–431. ArticleType: primary\_article / Full publication date: Jun., 2000 / Copyright © 2000 American Statistical Association.
- Pearl, J. (2000b). *Causality*. Cambridge University Press.
- Queen, C. M. and C. J. Albers (2009, June). Intervention and causality: Forecasting traffic flows using a dynamic bayesian network. *Journal of the American Statistical Association* 104(486), 669–681.
- Queen, C. M. and J. Q. Smith (1993). Multiregression dynamic models. *Journal of the Royal Statistical Society. Series B (Methodological)* 55(4), 849–870.
- Queen, C. M., J. Q. Smith, and D. M. James (1994, September). Bayesian forecasts in markets with overlapping structures. *International Journal of Forecasting* 10(2), 209–233.

- Raftery, A. E., M. Kárný, and P. Ettlér (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics* 52(1), 52–66.
- Rigat, F. and J. Q. Smith (2009). Semi-parametric dynamic time series modelling with applications to detecting neural dynamics. *The Annals of Applied Statistics* 3(4), 1776–1804.
- Shafer, G. (1996, November). *The Art of Causal Conjecture*. Artificial Intelligence. The MIT Press.
- Smith, J. Q. (1979). A generalization of the bayesian steady forecasting model. *Journal of the Royal Statistical Society. Series B (Methodological)* 41(3), 375–387.
- Smith, J. Q. (1981). The multiparameter steady model. *Journal of the Royal Statistical Society. Series B (Methodological)* 43(2), 256–260.
- Smith, J. Q. (1992, April). A comparison of the characteristics of some bayesian forecasting models. *International Statistical Review* 60(1), 75–87.
- Smith, J. Q. and P. E. Anderson (2008, January). Conditional independence and chain event graphs. *Artificial Intelligence* 172(1), 42–68.
- Smith, J. Q. and A. Daneshkhah (2010, June). On the robustness of bayesian networks to learning from non-conjugate sampling. *International Journal of Approximate Reasoning* 51(5), 558–572.
- Smith, J. Q. and F. Rigat (2008). Iseparation and robustness in finite parameter Bayesian inference. CRiSM 07-22, University of Warwick, Coventry.
- Spirtes, P., C. N. Glymour, and R. Scheines (2001). *Causation, Prediction, and Search* (2nd ed.). MIT Press.
- Stanley, R. (1997). *Enumerative combinatorics*. Cambridge: Cambridge University Press.
- Studený, M. (2005). *Probabilistic conditional independence structures*. Information Science and Statistics. London: Springer.
- Thwaites, P., J. Q. Smith, and E. Riccomagno (2010, August). Causal analysis with chain event graphs. *Artificial Intelligence* 174(12-13), 889–909.
- Thwaites, P. E. and J. Q. Smith (2006, September). Evaluating causal effects using chain event graphs. In *Proceedings of the 3rd European Workshop on Probabilistic Graphical Models*, Prague.
- West, M. and J. Harrison (1989). Subjective intervention in formal models. *Journal of Forecasting* 8(1), 33–53.
- West, M. and J. Harrison (1997, February). *Bayesian Forecasting and Dynamic Models* (Second ed.). Springer Series in Statistics. Springer-Verlag.