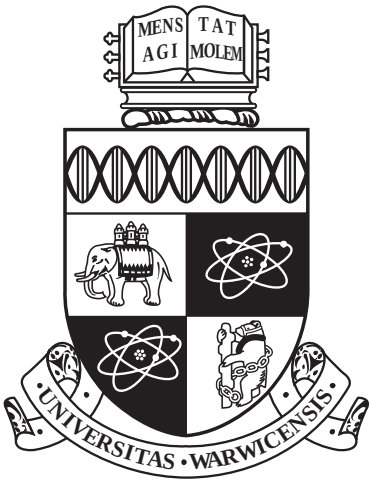

Learning and Predicting with Chain Event Graphs

by

Guy Freeman

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Statistics

September 2010

THE UNIVERSITY OF
WARWICK

Contents

List of Tables **iv**

List of Figures **v**

Acknowledgments **viii**

Declarations **x**

Abstract **xi**

Chapter 1 Introduction **1**

Chapter 2 Graphical models **13**

2.1 Introduction to graphical models 13

2.2 Introduction to Bayesian networks 16

2.3 Learning Bayesian networks 20

2.4 Causal Bayesian networks 24

2.5 Disadvantages of Bayesian network representations 27

Chapter 3 Learning chain event graphs **30**

3.1 Prerequisites 30

3.1.1 Event Trees 30

3.1.2 Chain Event Graphs 33
3.1.3 Causal trees and CEGs 36
3.2 Conjugate learning of CEGs 38
3.3 A Local Greedy Search Algorithm for finding the MAP Chain Event Graph 40
3.3.1 Preliminaries 40
3.3.2 The prior over the CEG space 43
3.3.3 The prior over the parameter space 44
3.3.4 The AHC algorithm 53
3.4 A weighted MAX-SAT algorithm for learning Chain Event Graphs . 54

Chapter 4 Dynamic graphical models **59**
4.1 Introduction to modelling time series 59
4.2 Forecasting with state-space models 60
4.3 Dynamic linear models 61
4.3.1 Multi-process Modelling 62
4.4 Steady model 65
4.5 Dynamic graphical models 70
4.5.1 Dynamic Bayesian networks 70
4.5.2 Multiregression dynamic models 71
4.5.3 Flow networks 73

Chapter 5 Dynamic chain event graphs **76**
5.1 The sampling distributions 78
5.2 The stage parameter distributions 80
5.3 The CEG distributions 84
5.4 One-step-ahead prediction 94

5.5 Causal intervention 95
5.5.1 Intervention on the CEG distribution 96
5.5.2 Intervention on T 97

Chapter 6 Analysis of exam-mark data using CEGs **100**
6.1 Learning static CEGs 100
6.1.1 Simulated data 100
6.1.2 Student exam data 102
6.1.3 AHC algorithm 102
6.1.4 Weighted MAX-SAT 105
6.2 Prediction with dynamic CEGs 106
6.2.1 Analysis of the series without intervention 107
6.2.2 Analysis of the series after intervention 111

Chapter 7 Discussion **117**

Appendix A Exam marks **123**

List of Tables

6.1 Selected stages of MAP CEG model found from data described in Section 6.1.2 using AHC. The columns respectively detail the stage number, posterior expectation of the probability vector of that stage (rounded to two decimal places), number of students passing through that stage in the dataset, number of situations from the original ET in that stage, examples of situations in that stage (shown as sequence of achieved grades 1, 2 or 3, and where 4 means that the grade is missing), and any comments or observations related to that stage. . 103

6.2 All possible stagings and their posterior probabilities at each time t for $k = 0.9$, $\rho = 0.9$, $q = 0.2$ with $P(C_1 = \{v_1,\ v_2,\ \{v_3, \ldots, v_6\},\ \{v_7, \ldots, v_{10}\}\}) = 1$ 108

List of Figures

1.1 Event tree of a student's potential progress through a hypothetical course described in Example 1. Each non-leaf node represents a juncture at which a random event will take place, with the selection of possible outcomes represented by the edges emanating from that node. Each edge distribution is defined conditional on the path passed through earlier in the tree to reach the specific node. 4

1.2 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. 8

3.1 Floret of v. This subtree represents both the random variable $X(v)$ and its state space $\mathbb{X}(v)$. 32

3.2 Simple event tree. The non-zero-probability events in the joint probability distribution of two Bernoulli random variables, A and B, with A observed before B, can be represented by this tree. Here, all four joint states are possible and hence there are four root-to-leaf paths through the nodes. 32

3.3 The CEG that reflects the three hypotheses of Example 1 36

3.4 Event tree for idle and manipulated versions of the same process . . 37

4.1 Dynamic Bayesian network of state-space model 70
4.2 Two-time-slice Bayesian network of state-space model 71
4.3 Example of a flow network 75

5.1 The Hasse diagram of the lattice of partitions of S when $|S| = 4$ . . 89

6.1 The event tree from Example 1 with the numbers representing the number of students in a simulated sample who reached each situation. 112
6.2 Sub-tree of the event tree of possible grades for the MORSE degree course at the University of Warwick. Each floret of two edges describes whether a student's marks are available for a particular module (denoted by the edge labelled A for the first module) or whether they are missing (NA). If they are available, then they are counted as grade 1 if are 70% or higher, grade 2 if they are between 50% and 69% inclusive, and grade 3 if they are below 50%. Some illustrative count data are shown on corresponding nodes. 113
6.3 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 $P(C_1 = \{v_1, v_2, \{v_3, \ldots, v_6\}, \{v_7, \ldots, v_{10}\}\}) = 1$, using the values in Table 6.2 114
6.4 Plots of probabilities that each pair of situations are in the same stage for different values of t, for the case when $k = 0.5$, $\rho = 0.25$, $q = 0.05$ with $P(C_1 = \{v_1, v_2, \{v_3, \ldots, v_6\}, \{v_7, \ldots, v_{10}\}\}) = 1$ 115

6.5 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 $P(C_1 = \{v_1, v_2, \{v_3, \ldots, v_6\}, \{v_7, \ldots, v_{10}\}\}) = 1$, and situations $v_2, v_7, v_8, v_9, v_{10}$ caused to be in the same stage at $t = 8$ 116

Acknowledgments

This thesis was typeset with LaTeX, with the help of the *warwickthesis* package put together by someone from the Physics department. In fact, I owe many people who I have never met and likely never will for the smorgasbord of open-source and Free (not just in price) software that I used during my PhD research, including the TeX stable. I transitioned from Microsoft Windows to the GNU/Linux operating system, currently in the form of the openSUSE distribution, and am eternally grateful for the opportunity to do so. To choose only a few other programs from the multitude that I don't have the time or memory to acknowledge here explicitly, I would like to thank the many authors and other contributors to GNU Emacs, gcc/g++, KDE and, most of all, the statistical programming language and interpreter R, in which I carried out the analyses of Chapter 6.

Turning to the people who have had a more direct role in helping to bring this thesis to fruition, it is only right that I first thank my supervisor, Professor Jim Q. Smith. He was and is an inspiring and thoughtful mentor who has massively influenced my statistical thinking and even my general approach to life through his effervescent enthusiasm and all-too-rare critical thinking. Thank you for putting up with my quirks so patiently.

Thank you very much also to the whole of the Statistics Department, all of whom have — to a greater or lesser extent, but it doesn't matter exactly how much in the final tally — supported me professionally and emotionally during my time here. I'll choose not to name names, but everyone knows how and how much and

why I appreciate all of them. I have made friends for life, and that is a precious gift that I won't take for granted.

Lastly, I would like to thank Sabrina, without whom I dread to imagine how my life would look: you are my foundation. And to my parents, Ron and Nava, and to my brother Eric: I owe everything to you. I hope I make you proud. It's the least I can do in thanks for everything you have done for me.

Declarations

I hereby declare that this thesis is based on my own research, except when stated otherwise. This thesis has not been submitted for a degree at another university.

Some of this work has been published, accepted for publication or submitted for publication as follows.

The material in Chapter 3, except for the part concerning the weighted MAX-SAT formulation, and some of Chapter 6, has been accepted for publication in the *Journal of Multivariate Analysis* under the title "Bayesian MAP Selection of Chain Event Graphs". The paper was co-authored with Jim Q. Smith but all the work is mine. Some of the material was also published as [Thwaites et al., 2009]. These papers are also available as CRiSM Working Papers 09-06 and 09-07 respectively.

The material in Chapter 4 is derived from a co-authored invited paper with Jim Q. Smith for the *Journal of Forecasting* under the title of "Distributional Kalman filters for Bayesian forecasting and closed form recurrences". That paper was written with Jim Q. Smith but the text in Chapter 4 is entirely my own work. It is available as CRiSM Working Paper 10-13.

The material in Chapter 5 and some of Chapter 6 has been submitted to the journal *Bayesian Analysis* under the title "Dynamic staged trees for discrete multivariate time series: forecasting, model selection and causal analysis". It is available as CRiSM Working Paper 10-14.

Abstract

Graphical models provide a very promising avenue for making sense of large, complex datasets. The most popular graphical models in use at the moment are Bayesian networks (BNs). This thesis shows, however, they are not always ideal factorisations of a system. Instead, I advocate for the use of a relatively new graphical model, the chain event graph (CEG), that is based on event trees.

Event trees directly represent graphically the event space of a system. Chain event graphs reduce their potentially huge dimensionality by taking into account identical probability distributions on some of the event tree's subtrees, with the added benefits of showing the conditional independence relationships of the system — one of the advantages of the Bayesian network representation that event trees lack — and implementation of causal hypotheses that is just as easy, and arguably more natural, than is the case with Bayesian networks, with a larger domain of implementation using purely graphical means.

The trade-off for this greater expressive power, however, is that model specification and selection are much more difficult to undertake with the larger set of possible models for a given set of variables. My thesis is the first exposition of how to learn CEGs. I demonstrate that not only is conjugate (and hence quick) learning of CEGs possible, but I characterise priors that *imply* conjugate updating based on very reasonable assumptions that also have direct Bayesian network analogues. By re-casting CEGs as partition models, I show how established partition learning algorithms can be adapted for the task of learning CEGs.

I then develop a robust yet flexible prediction machine based on CEGs for any discrete multivariate time series — the dynamic CEG model — which combines the power of CEGs, multi-process and steady modelling, lattice theory and Occam's razor. This is also an exact method that produces reliable predictions without requiring much a priori modelling. I then demonstrate how easily causal analysis can be implemented with this model class that can express a wide variety of causal hypotheses. I end with an application of these techniques to real educational data, drawing inferences that would not have been possible simply using BNs.

Chapter 1

Introduction

Very large datasets are becoming ever more common, with the ability to make sense of them becoming a major problem [Lohr, 2009]. If one uses overly simplistic models to analyse them, there is a risk of jumping to incorrect conclusions; if the models are too complex, they can at best take a very long time to compute, and at worst be opaque black boxes that have no explanatory power, cannot be quality-assured and are extremely sensitive in unpredictable ways to hyper-parameter inputs.

Graphical models provide an attractive middle way [Lauritzen, 1996]. Because of their pictorial form, graphs are excellent tools for eliciting expert opinion about a system and are transparent and communicable; because of their highly structured modular form, they can easily be operationalised for computation.

Bayesian networks (BNs) are currently one of the most widely used graphical models for representing and analysing multivariate distributions, with their explicit coding of conditional independence relationships between a system's variables [Cowell et al., 1999; Lauritzen, 1996], which is often the major knowledge domain of experts and an effective way to reduce dimensionality of a problem at a high level. However, despite their power and usefulness, it has long been known that BNs

cannot fully or efficiently represent certain common scenarios [Smith et al., 1993]. These include situations where the state space of a variable is known to depend on other variables, or where the conditional independence between variables is itself dependent on the values of other variables, called CONTEXT-SPECIFIC INDEPENDENCE in the literature [Boutilier et al., 1996]. In order to overcome such deficiencies, enhancements have been proposed to the canonical Bayesian network. Poole and Zhang [2003], for example, define CONTEXTUAL BELIEF NETWORKS. These, however, don't represent the context-specific independence relationships graphically, thus undermining the rationale for using a graphical model in the first place. Boutilier et al. [1996], meanwhile, keep the BN in place but additionally uses trees to describe the structures of the conditional probability distributions.

A new graphical model — the Chain Event Graph (CEG), first propounded by Smith and Anderson [2008] — aims to represent the context-specific independences and asymmetric sample spaces of a model explicitly and in a single graph. To this end, CEGs are based not on Bayesian networks, but on event trees (ETs) [Shafer, 1996]. Event trees are trees where nodes represent situations — i.e. scenarios in which a unit might find itself — and each node's extending edges represent possible future situations that can develop from the current one. It follows that every atom of the event space is encoded by exactly one root-to-leaf path. ETs are expressive frameworks for directly and accurately representing beliefs about a process, particularly when the model is described most naturally through how situations might unfold [Shafer, 1996]. However, as explained by Smith and Anderson [2008], ETs can contain excessive redundancy in their structure, with subtrees describing probabilistically isomorphic unfoldings of situations being represented separately. They are also unable to explicitly express a model's non-trivial conditional independence relationships. The CEG deals with these shortcomings by combin-

ing the subtrees that describe identical subprocesses so that the CEG derived from a particular ET has a simpler topology while in turn expressing more conditional independence statements than is possible through an ET.

Consider the following example, which exemplifies the types of hypotheses I plan to search over in my model selection.

Example 1. *Successful students on a one year course study components A and B, but not everyone will study the components in the same order: each student will be allocated to study either module A or B for the first 6 months and then the other component for the final 6 months. After the first 6 months each student will be examined on their allocated module and be awarded a distinction (denoted with D), a pass (P) or a fail (F), with an automatic opportunity to resit the module in the latter case. If they resit then they can pass and be allowed to proceed to the other component of their course, or fail again and be permanently withdrawn from the programme. Students who have succeeded in proceeding to the second module can again either fail, pass or be awarded a distinction. On this second round, however, there is no possibility of resitting if the component is failed. With an obvious extension of the labelling, this system can be depicted by the event tree given in Figure 1.1.*

To specify a full probability distribution for this model it is sufficient to only specify the distributions associated with the unfolding of each situation a student might reach. However, in many applications such as this one it is often natural to hypothesise a model where the distribution associated with the unfolding from one situation is assumed identical to another. Situations that are thus hypothesised to have the same transition probabilities to their children are said to be in the same *stage*. Thus in Example 1 suppose that as well as subscribing to the ET of Figure 1.1 one would want to consider the plausibility of the following three hypotheses:

1. The chances of doing well in the second component are the same whether the

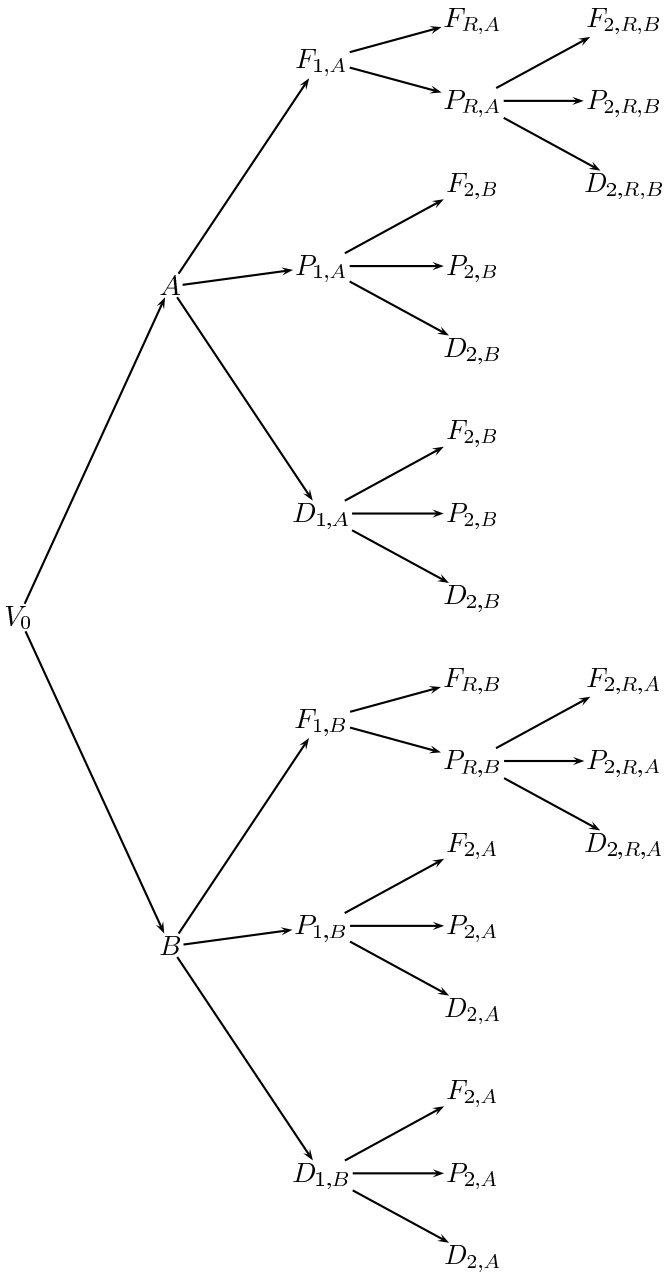

Figure 1.1: Event tree of a student's potential progress through a hypothetical course described in Example 1. Each non-leaf node represents a juncture at which a random event will take place, with the selection of possible outcomes represented by the edges emanating from that node. Each edge distribution is defined conditional on the path passed through earlier in the tree to reach the specific node.

student passed the first module the first time or after a resit.

2. The components A and B are equally hard.

3. The distribution of marks for the second component is unaffected by whether students passed or got a distinction for the first component.

Each of these hypotheses can be identified with a partitioning of the non-leaf nodes (SITUATIONS). In Figure 1.1 the set of situations is

$$S = \{V_0, A, B, P_{1,A}, P_{1,B}, D_{1,A}, D_{1,B}, F_{1,A}, F_{1,B}, P_{R,A}, P_{R,B}\}.$$

The partition C of S that encodes the above three hypotheses consists of the stages $u_1 = \{A, B\}$, $u_2 = \{F_{1,A}, F_{1,B}\}$, and $u_3 = \{P_{1,A}, P_{1,B}, P_{R,A}, P_{R,B}, D_{1,A}, D_{1,B}\}$ together with the singleton $u_0 = \{V_0\}$. Thus the second stage u_2, for example, implies that the probabilities on the edges $(F_{1,B}, F_{R,B})$ and $(F_{1,A}, F_{R,A})$ are equal, as are the probabilities on $(F_{1,B}, P_{R,B})$ and $(F_{1,A}, P_{R,A})$. Clearly the joint probability distribution of the model – whose atoms are the root to leaf paths of the tree – is determined by the conditional probabilities associated with the stages. A CEG is the graph that is constructed to encode a model that can be specified through an event tree combined with a partitioning of its situations into stages.

In the first part of this thesis I suppose that we are in a context similar to that of Example 1, where, for any possible model, with a selection of these types of hypotheses, the sample space of the problem must be consistent with a single event tree. On the basis of a sample of students' records we would want to select one of a number of these different possible CEG models, i.e. we want to find the "best" partitioning of the situations into stages. I take a Bayesian approach to this problem and choose the model with the highest posterior probability — the Maximum A Posteriori (MAP) model. This is the simplest and possibly most

common Bayesian model selection method, advocated by, for example, Bernardo and Smith [1994], Denison et al. [2002], Heckerman [1999] and Castelo [2002], the latter two specifically for models that are Bayesian networks. Because the range of possible CEG models for *any* system exceeds the set of possible BN models, however, and encode information differently from them, the algorithms for searching for MAP BNs must be adapted accordingly. In Section 6.1.1 I show how to learn a CEG from the tree in Example 1 using simulated data and the algorithm developed in Section 3.3.

My aim throughout this thesis is to ensure all calculations, at least with complete sampling, are exact, i.e. there is no need for approximate numerical techniques such as MCMC. While MCMC has vastly widened the vista of possible Bayesian analyses, it can sometimes be used as a crutch when a faster, wholly adequate exact analysis would be possible with a slight adjustment of the model. When it comes to very large datasets with a commensurately very large set of possible models, conjugate analyses can vastly speed up searches across the model space. MCMC is extremely useful for estimating parameters of models once the most appropriate choice of model has been identified, if necessary.

In the second half of the thesis I develop a class of dynamic multivariate graphical models over finite discrete state spaces based on CEGs for the purposes of prediction, where at each time point the relevant cohort of units data is represented by a different CEG. Highly multivariate discrete processes are quite common but to my knowledge have so far not been systematically studied with graphical models. These processes in the most general case tend to have the following characteristics:

1. A description is provided of the possible development histories each unit in the process can take at a given time. 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, but the range of possibilities remains fixed.

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 time cohorts, assumed here for simplicity to be equally spaced apart. 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, i.e. one of its variables is manipulated exogenously. The model then admits a "causal" extension which provides predictions of the process when subject to such a control.

I am particularly interested in making good one-step ahead predictions for such a system. I consider making good (probabilistic) predictions (or FORECASTS) to be the central goal of statistical analysis, as argued by de Finetti [1974] and Dawid [1984]. This approach will also provide, as a beneficial side-effect, the probabilities of the symmetry hypotheses through time, which can be used 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. The general points above then 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.

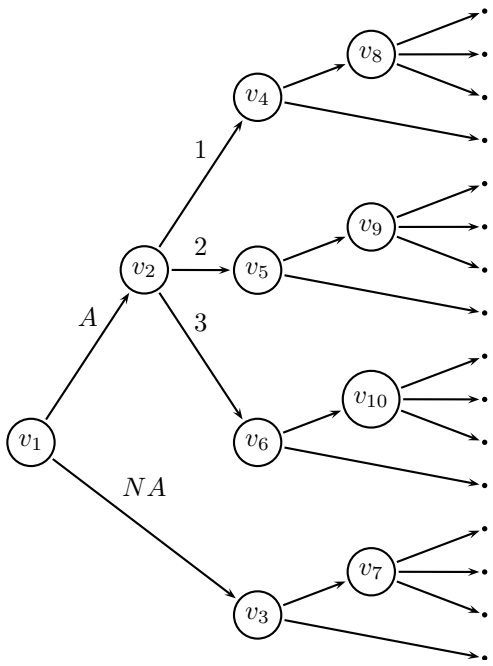

Figure 1.2: Event tree for marks for two modules in a course. Marks are discretized into 3 grades, and A and NA indicate whether the mark is recorded or missing. The 10 situations are labelled and the 16 leaf nodes are unlabelled.

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.

One example of an event tree for the marks for a course with two modules and three grades is given by Figure 1.2.

The event tree can represent any discrete event space and naturally codifies a chronological order (or partial order) in its topology, and so I base my own dynamic graphical model 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 I am interested in modelling. CEGs do, though, and so the model class developed here is based on them but extended into a more general dynamic scenario where probabilities and symmetries are allowed to change with time.

I describe the dynamics of this type of tree-structured process by a state space model incorporating a switching mechanism to neighbouring models at every time point. The earliest example of this general class, to the best of my knowledge, was studied for univariate Gaussian series [Harrison and Stevens, 1976; West and Harrison, 1997] and called Multi-process Models Class II. Frühwirth-Schnatter [2006] reviews switching models for non-Gaussian state spaces, but none of these have closed posterior forms. Here, I use a type of multi-process model which allows dynamic shifting from one symmetry partition to neighbouring ones whilst retaining conjugacy.

Various classes of discrete multivariate time series are of course well studied. Possibly the closest classes to the one considered here with associated graphical models are the models 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 $\{(T_s, E_s) : s = 1, \ldots, S\}$ 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 timing of events, usually allowed to occur within a continuous time domain. Here, in contrast, I wish to model a class of complex discrete distributions over a discrete time domain. I discuss the connections between the two model classes further in Chapter 7.

In order to take into account possible drifting on the tree parameters through time caused by unobserved background processes, one could follow the standard approach of stating a transition probability $P(\theta_t \mid \theta_{t-1}, S)$, where θ_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 almost always immediately requires the inference to be undertaken with approximating numerical methods. This is not ideal in this context for several reasons: First, in the process I consider here, conjugacy and modularity are present and it would be a shame to lose these useful properties. Secondly, because of the vastness 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 greatly speeds up computation of model goodness. Thirdly, models in this class are easier to interpret when they retain their modular and conjugate forms.

An alternative approach, which I take here, is to set a transition function

$$\mathcal{T} : P(\theta_{t-1} \mid x^{t-1}, S) \mapsto P(\theta_t \mid x^{t-1}, S) \tag{1.1}$$

where x^{t-1} are the observations up to time $t-1$. Although this approach is narrower in its scope, it is sufficient for making probabilistic predictions which is my aim here as mentioned earlier. The particular transition function I ultimately choose to use can be justified through various characterisations [Smith, 1979, 1992], encouraging

several different authors to use such transitions. I also show that it has the property of preserving the modular structure of each model in this class and works well against prior misspecification.

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 static chain event graphs by Thwaites et al. [2010]. I extend this to the dynamic model class presented here. By still retaining conjugacy and modularity when learning model probability parameters, this causal extension of the model class is particularly straightforward, allowing it to be easily used for modelling a controlled environment.

Thesis outline

The remainder of this thesis is thus structured as follows.

In Chapter 2 I review the latest theory concerning graphical models and how to learn them automatically.

In Chapter 3 I review the definitions of event trees and CEGs. I then develop the theory of how conjugate learning of CEGs is performed, and apply this theory by using the posterior probability of a CEG as its score in a model search algorithm that is derived using an analogous procedure to the model selection of BNs. I characterise the product Dirichlet distribution as a prior distribution for the CEGs' parameters under particular homogeneity conditions.

In Chapter 4 I review some theory concerning state-space and dynamic graphical models that will be relevant in developing the new dynamic graphical model based on CEGs.

In Chapter 5 I proceed to expositing the dynamic chain event graph. I formally define the necessary concepts and show how to make exact one-step ahead

predictions with the new model. I then extend the model to allow the implementation of causal analyses.

In Chapter 6 I apply all of the theory and algorithms to a simulated data for testing purposes and then to results from a real educational programme in order to make rich inferences about students' educational achievement.

I end in Chapter 7 by discussing outstanding research questions that extend from the work in this thesis.

Chapter 2

Graphical models

I begin by describing what I consider graphical models to be and why they are worthy of study and use. I then move on to discussing various statistical issues concerning the most popular contemporary graphical model: the Bayesian Network (BN). I finish by critiquing the BN and proposing a new graphical model that is more appropriate for many applications based on trees.

2.1 Introduction to graphical models

Statistical models are descriptions of stochastic systems that enable us to understand the relationships between the variables of that system. In the Bayesian paradigm, the statistical model encodes degrees of belief about various hypotheses concerning the system as probabilities, and these probabilities are updated in line with probability theory as observations of the system are made.

It is clear, therefore, that the statistical model used to describe a system and make predictions and decisions concerning that system must be chosen with great care. For very complex systems, the temptation to settle for a simple model should in general be resisted unless it can be shown that the approximation required will not

affect any analysis adversely. Excessively complex models, however, require the setting of more parameters, which leads to greater risk of model mis-specification, and also large amounts of computation which can quickly lead to intractability. What is required, as Einstein put it, is "to make the irreducible basic elements as simple and as few as possible without having to surrender the adequate representation of a single datum of experience" [Einstein, 1934]. One way to do so is to make qualitative judgements about the system, for example about any homogeneities which are believed a priori to exist between seemingly separate variables. This can reduce the dimensionality of the model as well as increase its power. To represent these statements transparently one can use a (network) graph, which characterises the model as a graphical model.

Lauritzen [1996] notes that graphical models have their origin in the early 20th century in the analysis of statistical mechanics by Gibbs [Gibbs, 1902; Borgelt and Kruse, 2002]. Nowadays graphical models are considered to be "statistical models embodying a collection of marginal and conditional independences which may be summarized by means of a graph" [Dawid and Lauritzen, 1993]. This certainly describes Bayesian networks, but I will show how the syntax of a graph can be used to describe other model properties apart from independence relationships.

My overarching aim when using graphical models is well described by Dawid [2002]:

> Seek to represent and manipulate as much as possible of the relevant structure and details of the model by purely graphical means, keeping any external information required to a minimum

As Dawid [2002] notes, "what is relevant for one purpose may be irrelevant clutter for another". I will show that Bayesian networks do not always represent the important and relevant details of a model graphically.

I begin by revising basic graph theory terminology that will be used throughout. Further details of these concepts can be found in many introductory graph theory texts, e.g. [West, 2001].

Definition 2. *A* GRAPH G *is a pair* $(V(G), E(G))$ *where* $V(G)$ *is its set of vertices (or nodes),* $E(G)$ *is its set of edges. The set of edges can be thought of as a relation on* $V(G)$*.*

When a graph is drawn, the vertices are displayed as points and the edges as curves between the appropriate points.

Definition 3. *A* DIRECTED GRAPH *(or digraph) is a graph* G *where the edges are ordered pairs of vertices. Thus the edges* $e_1 = (v_1, v_2)$ *and* $e_2 = (v_2, v_1)$ *(where* $v_1, v_2 \in V(G)$*) are distinct elements of* $E(G)$*.*

Edges in a directed graph are drawn as arrows from the first vertex to the second vertex in the ordered pair.

All graphs in this paper are directed graphs, and the following definitions assume this.

Definition 4. *In a digraph, the* CHILD *of the edge* $e = (v_1, v_2) \in E(G)$*, written* $ch(e)$*, is* v_2*. Its* PARENT $pa(e)$ *is* v_1*.*

By abuse of notation, the children of a vertex $v \in V(G)$*, written* $ch(v)$*, are defined as*

$$ch(v) = \{v' : v' \in V(G), (v, v') \in E(G)\} \tag{2.1}$$

and $pa(v)$ *is defined similarly.*

Definition 5. *A* PATH λ *between two vertices* $v_1, v_2 \in V(G)$ *is an ordered sequence of edges* $\lambda(v_1, v_2) = (e_1, \ldots, e_n)$ *where* $e_1, \ldots, e_n \in E(G)$*,* $pa(e_1) = v_1$*,* $ch(e_n) = v_2$ *and* $ch(e_k) = pa(e_{k+1})$ *for* $k = 1, \ldots, n-1$*.*

Definition 6. *The* LENGTH *of a path is the number of edges it contains, given in the above definition as n. By an abuse of notation, we say $v \in \lambda$ (where $v \in V$) if the path λ passes through v.*

Definition 7. *A* CYCLE *is a path $\lambda(v_1, v_2)$ where $v_1 = v_2$.*

Definition 8. *An* ACYCLIC GRAPH *contains no cycles.*

Definition 9. *A graph is* CONNECTED *if there exists a path in the graph between every pair of vertices, where direction of edges here can be changed if necessary.*

Definition 10. *A graph is a* COMPLETE *graph if there is an edge between every pair of nodes.*

Definition 11. *A* TREE *is a connected acyclic graph where one vertex (denoted here by v_0) has no parents and all other vertices have exactly one parent.*

Definition 12. *A* LEAF NODE *in a tree is a vertex with no children. The set of leaf nodes of a tree T is denoted here by $L(T)$.*

2.2 Introduction to Bayesian networks

The Bayesian network uses a modification of the graph theory concept of separation to represent conditional independence relationships. It can be proven that the separation properties of a Bayesian network graph match up with the conditional independence properties of a statistical model so that such a representation makes sense. I show here how this is done formally, beginning with giving the definition and formal axioms of conditional independence as defined by [Dawid, 1979]. The axioms are also called the SEMI-GRAPHOID AXIOMS after Pearl and Paz [1986].

I start by introducing a formal definition of conditional independence as given by [Dawid and Lauritzen, 1993].

Definition 13. *Let X, Y, Z be random variables on a probability space $(\Omega, \mathcal{F}, P)$. Then X is* CONDITIONALLY INDEPENDENT *of Y given Z (under P) if for any P-measurable set A in the sample space of X, $P(X \in A \mid Y, Z)$ can be expressed as a function of Z alone.*

It is clear that conditional independence is a useful modelling assumption to make if it sensible to do so, because the dimensionality of the model for any random variable can be reduced when conditioning on other variables. A special case of this phenomenon is statistical sufficiency, as explained by [Dawid, 1979; Dawid and Lauritzen, 1993], when the random variables are parameters of the model; another example of conditional independence is in linear regression where the number of explanatory variables required in the model is deemed to be sufficient to model the dependent variable.

Standard independence holds when Z in the above definition is the empty set.

Now I introduce the semi-graphoid axioms. Let W, X, Y, Z be four disjoint subsets of a set U and let $\perp\!\!\!\perp$ and $\mid$ form a ternary relation $R \subseteq U^3$ of subsets of U, where I write $X \perp\!\!\!\perp Y | Z$ if $(X, Y, Z) \in R$, for example. It is also possible to write $X \perp\!\!\!\perp Y$ if $(X, Y, \emptyset) \in R$. R then satisfies the semi-graphoid axioms if, as given by [Borgelt and Kruse, 2002],

Symmetry $(X \perp\!\!\!\perp Y \mid Z) \implies (Y \perp\!\!\!\perp X \mid Z)$

Decomposition $((W \cup X) \perp\!\!\!\perp Y \mid Z) \implies (W \perp\!\!\!\perp Y \mid Z)$ and $(X \perp\!\!\!\perp Y \mid Z)$

Weak union $((W \cup X) \perp\!\!\!\perp Y \mid Z) \implies (X \perp\!\!\!\perp Y \mid Z \cup W)$

Construction $(X \perp\!\!\!\perp Y \mid (Z \cup W))$ and $(W \perp\!\!\!\perp Y \mid Z) \implies ((W \cup X) \perp\!\!\!\perp Y \mid Z)$

It can be proven that conditional independence between random variables satisfies the semi-graphoid axioms [Castillo et al., 1997; Borgelt and Kruse, 2002],

and therefore we can write $X \perp\!\!\!\perp Y \mid Z$ to represent the statement that X is conditionally independent of Y given Z.

I now show how a Bayesian network can be used to graphically represent all of the conditional independence statements of a model.

Definition 14. *A* BAYESIAN NETWORK *for the model with set of random variables* $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ *on a probability space* $(\Omega, \mathcal{F}, P)$ *is a directed acyclic graph* $G = (V, E)$ *where*

1. *each node* $V_i \in V$ *corresponds to exactly one variable* $X_i \in \boldsymbol{X}$*, and*
2. *if* $P(\boldsymbol{X})$ *can be written as* $\prod_{i=1}^{n} P(X_i \mid Q_i)$*, where* $Q_i \subseteq \{X_1, \ldots, X_{i-1}\}$ *(with the exception of* $Q_1 = \emptyset$*), then* $pa(V_i) = V(Q_i)$*, where* $V(Q_i)$ *are the nodes corresponding to the random variables in* Q_i*.*

From here on in, I refer to the vertices representing random variables or sets or collections of random variables by the random variables themselves, except in cases where there might be possible confusion.

Note that a complete Bayesian network can always be drawn for a model with a finite number of random variables, as $P(X_1, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_1, \ldots, X_{i-1})$ is always true. Note therefore that there may also be more than one possible Bayesian network representation of a model; in particular, any complete directed acyclic graph with $V = V(\boldsymbol{X})$ is always a valid Bayesian network. Adding edges to a valid BN always creates another valid BN, as long as the resulting graph remains directed and acyclic.

It is clear the Bayesian network representation of a model explicitly encodes some conditional independence statements of the model. Specifically, it can be immediately read from the graph that

$$X_i \perp\!\!\!\perp (\{X_1, \ldots, X_{i-1}\} \setminus Q_i) \mid Q_i \tag{2.2}$$

for $i = 2, \ldots, n$ purely from its topology. However, more conditional independence statements can be inferred from the graph using a property of the graph that also satisfies the semi-graphoid axioms. This property is D-SEPARATION, first defined by Verma and Pearl [1988] and subsequently re-defined by Lauritzen [1996] in a more useful and operational way, where for three disjoint subsets $A, B, S \subset V$, S is said to d-separate A and B if S blocks all paths between all vertices in A and all vertices in B on a transformed version of the original BN. The transformation is as follows:

1. Delete all vertices from the BN that are neither part of A, B or S, nor have a path from themselves to another vertex in A, B or S. Delete all edges which had one of the deleted vertices at one of their ends. This is the ANCESTRAL graph of the BN.

2. For every pair of nodes that have a common child that are not connected create an edge between them. This is the MORALISED graph (because "unmarried" parent nodes are made to "marry").

3. Ignore the directions of arrows on edges for determining whether paths are blocked. This is the SKELETON graph.

Then it can be proven that for any BN set up as above,

$$S \text{ d-separates } A \text{ from } B \Longrightarrow A \perp\!\!\!\perp B \mid S \tag{2.3}$$

Thus by stating a few qualitative statements about how some variables are not relevant in determining the distributions of other variables if the values of yet another group of variables is known, many other conditional independence statements of the model can be inferred.

2.3 Learning Bayesian networks

In many scenarios, the modeller might not have complete certainty over the conditional independence relationships which hold between the variables of the system under consideration, or equivalently the Bayesian network which best represents the model. In this case, the Bayesian approach is to consider the structure itself as a random variable with a probability distribution of its form set a priori, and then updated using Bayes' theorem in the light of new data. This procedure has been described as LEARNING the Bayesian network by the artificial intelligence community, e.g. in [Heckerman, 1999] and can be considered as another form of model selection.

However, the procedure is in practice rarely so simple. The major obstacle in carrying it out is that the size of the set of possible Bayesian networks grows in size super-exponentially with respect to the size of the set of random variables [Cooper and Herskovits, 1992]. This means that setting a proper subjective prior distribution over the set of possible Bayesian networks for any practical situation is generally intractably difficult, as is setting the parameter priors and likelihoods for each possible BN.

There are some approaches advocated in the literature, however, that seek to minimise this difficulty by utilising some reasonable simplifying assumptions. I discuss the assumptions which relate to discrete variables in particular which is my focus in this thesis.

The initial set of assumptions deals with the probability model for the data implied by each Bayesian network. Let B be the random variable representing the Bayesian network which holds. Then

$$P(\boldsymbol{X} \mid \boldsymbol{\theta}_B, B) = \prod_{i=1}^{n} P(X_i \mid Q_i, \theta_{Bi}, B) \tag{2.4}$$

where $\boldsymbol{\theta}_B = \{\theta_{B1}, \ldots, \theta_{Bn}\}$ is the set of parameter vectors θ_{Bi} for each distribution $P(X_i \mid Q_i, \theta_{Bi}, B)$. Then the prior probability distribution of $\boldsymbol{\theta}_B|B$ is set by assuming PARAMETER INDEPENDENCE [Spiegelhalter and Lauritzen, 1990], so that

$$P(\boldsymbol{\theta}_B \mid B) = \prod_{i=1}^{n} \prod_{j=1}^{q_i} P(\theta_{Bij} \mid B) \tag{2.5}$$

where θ_{Bij} is the parameter vector of the probabilities $P(X_i \mid Q_i = q_j, B)$ and q_i is the number of possible values of Q_i. Note that I am assuming, in line with my relevance assumptions, that the value of θ_{Bij} does not rely on the parts of B not related to X_i and its parents, a property called LIKELIHOOD MODULARITY. If θ_{Bij} is distributed as $\mathrm{Dir}(\boldsymbol{\alpha}_{Bij})$, then the updating of $P(\theta_{Bij} \mid B, \boldsymbol{X})$ is conjugate:

$$\theta_{Bij} \mid B, \boldsymbol{X} \sim \mathrm{Dir}(\boldsymbol{\alpha}_{Bij} + \boldsymbol{N}_{ij}) \tag{2.6}$$

where $\boldsymbol{N}_{ij}$ represents the vector of counts N_{ijk} when $Q_i = \boldsymbol{q}_j$ and $X_i = x_{ik}$, where k indexes the possible values of X_i.

While parameter independence simplifies the setting and updating of $P(\boldsymbol{\theta} \mid B)$ for each possible BN B, it still requires the setting of each $P(\theta_{Bij} \mid B)$ for each B, and still does not address the setting of $P(B)$.

In order to simplify the setting of $P(\theta_{Bi} \mid B)$ — the priors for the parameters of variable X_i in a BN B — for all variables X_i for each possible BN B, one can make the assumption of PRIOR MODULARITY. This states that if two Bayesian networks B_1 and B_2 have identical parent variables Q_i for some variable X_i, then $P(\theta_{Bi} \mid B_1) = P(\theta_{Bi} \mid B_2)$, i.e. the prior on the parameters that determine the distribution of X_i are equal for both BNs. The subscript B will therefore be dropped henceforth as now only the parent set of a variable X is necessary to determine the prior distribution of its parameters.

Under the assumptions of prior and likelihood modularities, it is the case (as shown in [Heckerman and Geiger, 1995]) that in order to set parameter priors for each possible BN it is sufficient to set parameter priors only for the complete Bayesian networks. Parameter priors for incomplete networks are then derived from equivalent local structures in the corresponding complete network.

This can still be intractable, and so there is one more level of simplification possible. Assume that under any B the parameter vectors θ_{ij} are mutually independent of one another for any X_i for any values of its parents $Q_i = q_j$ as above, and that for any two Markov equivalent BNs B_1, B_2 (i.e. those which encode the same sets of conditional independence relations on $\boldsymbol{X}$, as can be determined using the methods of [Verma and Pearl, 1990] or [Chickering, 1995]) it is assumed that $P(\boldsymbol{X} \mid B_1) = P(\boldsymbol{X} \mid B_2)$ (called HYPOTHESIS EQUIVALENCE by [Heckerman et al., 1995]). Geiger and Heckerman [1997] showed that in this case that all θ_{ij} must have a Dirichlet distribution. Therefore to specify the parameter priors for any network B one needs only to specify the hyperparameters of the Dirichlet distribution of the joint distribution of $\boldsymbol{X}$ on a complete network.

The setting of $P(B)$ is comparatively simple. Apart from the obvious choices of a uniform prior over all possible B or a subset of all possible B, another possible qualitative characterisation is to consider the probability for the inclusion of each edge in a BN with a fixed order of variables [Buntine, 1991], and further still if the edges are considered exchangeable, i.e. all of the edges have a probability p of existing, then only one probability assessment — that of p — is needed.

With the parameters set as above and assuming Dirichlet priors, $P(\boldsymbol{X} \mid B)$ will be a closed formula for each B as discovered by Cooper and Herskovits [1992];

Heckerman et al. [1995]:

$$P(\boldsymbol{X} \mid B) = \prod_{i=1}^{n} \prod_{j=1}^{|\boldsymbol{q}_i|} \left[\frac{\Gamma(\alpha_{ij.})}{\Gamma(\alpha_{ij.} + x_{ij.})} \prod_{k=1}^{|\boldsymbol{x}_i|} \frac{\Gamma(\alpha_{ijk} + x_{ijk})}{\Gamma(\alpha_{ijk})} \right] \tag{2.7}$$

where $|\boldsymbol{x}_i|$ are the number of possible values of X_i, $x_{ij.} = \sum_k x_{ijk}$, x_{ijk} is the number of times $X_i = x_{ik}$ when $Q_i = q_j$, and $\alpha_{ij.} = \sum_k \alpha_{ijk}$. $P(B \mid \boldsymbol{X})$ can then be easily calculated from Bayes' theorem for each B if $P(B)$ is a fixed quantity a priori.

However, when there are a large number of possible BNs B, this might not be practical. To predict new data $\boldsymbol{X}^*$ from the system after having observed $\boldsymbol{X}$, it is necessary to calculate

$$P(\boldsymbol{X}^* \mid \boldsymbol{X}) = \sum_{B \in \boldsymbol{B}} P(\boldsymbol{X}^* \mid B) P(B \mid \boldsymbol{X}). \tag{2.8}$$

This is called MODEL AVERAGING [Hoeting et al., 1999]. For a large set of possible BNs $\boldsymbol{B}$, it would be impractical to calculate $P(\boldsymbol{X}^* \mid B)$ and $P(B \mid \boldsymbol{X})$ for each B. There are a number of approximations to the full solution which could still give good predictions while reducing the computational effort required [Hoeting et al., 1999].

If the aim is to provide a good "explanatory" network for the system, then trying to find the most probable BN (MAP, or Maximum A Posteriori BN) can be done more efficiently, if not necessarily optimally, than just calculating $P(B \mid \boldsymbol{X})$ for every possible B, by SEARCHING the model space. There have been many strategies suggested for this search, including greedy search, greedy search with restarts, best-first search, and Monte Carlo methods, all discussed by Heckerman [1999], and more recently weighted MAX-SAT solving [Cussens, 2008].

One relevant consequence of the model set-up described above which leads to equation (2.7) is that the goodness of a BN, defined here as its posterior probability,

can be calculated as the product of purely LOCAL properties of the network, where local here relates to individual nodes and their parents. This means that if two BNs differ only in one parent set Q_i of some variable X_i, the difference in scores will result only from that local difference. This allows for efficient LOCAL SEARCH ALGORITHMS for searching the model space. A simple local greedy search starts with one possible BN, then calculates the score for a BN which differs only in having an edge reversed, an edge added or an edge deleted (subject to the resulting network being acyclic) by only re-calculating the relevant local score, and chooses the BN which has the higher posterior probability. Because only the local differences in the graphs have to be taken into account, the search proceeds more quickly.

The search algorithms to find the MAP BN can also be used to find more than one high-scoring network so that $P(\boldsymbol{X}^* \mid \boldsymbol{X})$ can be approximated as

$$P(\boldsymbol{X}^* \mid \boldsymbol{X}) \approx \sum_{B \in \tilde{\boldsymbol{B}}} P(\boldsymbol{X}^* \mid B) P(B \mid \boldsymbol{X}) \tag{2.9}$$

where $\tilde{\boldsymbol{B}}$ is the set of highest-scoring networks found during the model search, where the size of the set can be chosen as high as desired.

2.4 Causal Bayesian networks

Efforts have been made to use Bayesian networks not only to incorporate beliefs about the conditional independence relations between the variables in a system, but also CAUSAL RELATIONS between them, most prominently by Pearl [Pearl, 2000b]. I briefly review how this is done and how it has led to work on learning these causal relations.

As mentioned in the last section, different Bayesian networks can represent equivalent conditional independence statements. However, when drawing a Bayesian

network of a system, there can be a conscious or unconscious desire to somehow represent certain "causal" relations between the variables. One way to describe these causal hypotheses is to consider how the system changes under external intervention. If a variable A is a cause of another B, then directly changing A will change the probability distribution of B. Pearl [Pearl, 2000b] represents the probability distribution of B after intervening in the value of A as $P(B \mid do(A))$, in order to distinguish this distribution from the one of B after merely observing A, $P(B \mid A)$. There is no reason why in general $P(B \mid do(A))$ should be related to $P(B \mid A)$, but in many cases there is a presumed relationship that can be incorporated into a model.

A CAUSAL BAYESIAN NETWORK (CBN) [Pearl, 1995] sets strict constraints on this relationship. A CBN is a BN that, as well as describing the conditional independence statements that are satisfied by the joint probability distribution over the model's variables, asserts certain beliefs about the probability distribution over the variables resulting from an exogenous manipulation of any subset of them. The exact nature of these beliefs is described in the following definition.

Definition 15. *A causal Bayesian network is a Bayesian network that additionally holds the following properties when some subset of the variables $\boldsymbol{X}_I \subseteq \boldsymbol{X}$ is intervened upon to take the vector of values $\boldsymbol{x}_I$:*

1. *The probability distribution of each $X_I \in \boldsymbol{X}_I$ becomes degenerate, so that $P(X_I = x_I) = 1$ when x_I is the relevant value from $\boldsymbol{x}_I$, and 0 otherwise*

2. *The probability distributions of all other variables $X_i \notin \boldsymbol{X}_I$ conditional on their parent variables Q_i stay unchanged.*

The effect of an intervention, therefore, is to only change the parts of the probability distribution associated with the intervened variables in the factorisa-

tion of the joint probability distribution described by a BN. Note that now BNs which were describing identical conditional independence statements have different replacement probability distributions under identical interventions.

There have been attempts to learn CBNs from data, e.g. by Heckerman [1995], Cooper and Yoo [1999], and Spirtes et al. [2001]. The approach advocated by the first two papers cited works by either considering, in addition to the random variables under investigation, whether those variables were merely observed or actively manipulated for each data point, which essentially expands the event space. This is equivalent to re-drawing the CBN as a BN with additional nodes indicating whether manipulation or mere observation led to other nodes' values, as advocated by Dawid [2002] and called an AUGMENTED DAG by him. This BN can then be learnt in the same way as discussed earlier.

Spirtes et al. [2001], meanwhile, along with others such as Glymour and Cooper [1999] and Neapolitan and Jiang [2006], claim to have algorithms to learn CBNs, and thus causal relations between variables, merely from observational data. This methodology is called CAUSAL DISCOVERY. The validity of this approach has been disputed by a number of authorities, including Humphreys and Freedman [1996], Cartwright [2007] and Dawid [2010], along the lines that, as Cartwright [1994] put it, "No causes in, no causes out" — in other words, without making causal assumptions, i.e. without explicitly stating how the idle and manipulated systems relate to one another, it is not possible to learn about manipulated systems from idle systems. I therefore do not pursue this approach further in this thesis, instead only making causal inferences when I am willing to make causal assumptions, which will only happen if data under controlled interventions are available.

2.5 Disadvantages of Bayesian network representations

Despite their obvious strengths in allowing for the reduction in the dimensionality of models' joint probability distributions and in providing a transparent framework for causal inference as described above, BNs are not optimal graphical models in all situations. The biggest problems with their use occur under two scenarios, which are not necessarily mutually exclusive:

1. when the model event space is not a simple product space, i.e. the state spaces of some of the random variables in the system are radically different — or even non-existent — depending on the values of other system variables; and

2. when conditional independence statements are true only under certain values of other variables.

Neither of these scenarios can be discerned directly from the BN. Consider the situation in Figure 1.2. The event space is clearly asymmetric because if the first module's marks are unavailable then they have no grade. Additionally, it might be the case, for example, that students who get grades 2 or 3 in the first module perform in an identical way on the second module, but student who perform the best in the first module by getting the highest grade perform completely differently. These features will not be exhibited by the structure on a BN unless special care is taken.

These blind spots of BNs are not unknown in the literature. For example, Spiegelhalter and Lauritzen [1990] already noted with regard to the second property in 1990 that "a systematic approach to the manipulation of such relevance links would be an important development". This property was termed CONTEXT-SPECIFIC INDEPENDENCE [Boutilier et al., 1996] and various approaches were tried to deal with it in the BN representation. For example, Boutilier et al. [1996], in an early attempt,

kept the BN in place but additionally used trees to describe the probability distributions of each variable, and then proceeded to re-arrange the BNs using these trees, including having multiple nodes for a single random variable in order to represent some of the context-specific independences in a BN format. Jaeger [2004] defined PROBABILITY DECISION GRAPHS (PDGs) that can represent certain context-specific independences, but PDGs cannot represent some conditional independence relations that can be represented by BNs; for example, as admitted by Jaeger [2004], the BN with nodes X_1, X_2, X_3, X_4 and edges $(X_1, X_2), (X_1, X_3), (X_2, X_4), (X_3, X_4)$ cannot be represented as a PDG. More recently, [Poole and Zhang, 2003] defined CONTEXTUAL BELIEF NETWORKS, but these are basically BNs with the extra contextual information not represented graphically.

One final approach is that of Bayesian multinets [Geiger and Heckerman, 1996], where context-specific independence is termed ASYMMETRIC INDEPENDENCE. Bayesian multinets are essentially collections of different BNs over the same set of random variables, one BN drawn for each collection of values of one of the variables (called the HYPOTHESIS VARIABLE) that makes the BN of the system different from all the others. While this solves the problem of representing context-specific independence graphically and hence explicitly, there is still a lot of redundancy in the representation due to needing to draw a BN for each of the variable values of each hypothesis. This problem only gets worse if more than one hypothesis variable is proposed. There is also no acknowledgement of how to deal with sparse conditional probability tables efficiently.

In the next chapter I re-introduce the Chain Event Graph (CEG), a tree-based rather than BN-based graphical model. It will be shown that it can represent all conditional independence statements that BNs of the same system can; that it can make explicit the asymmetries in state spaces of random variables in different

contexts; that it can graphically represent context-specific conditional independence relationships; that it allows conjugate inference and learning; and that it allows a larger class of external manipulations in the system than a BN, thereby extending the range of possible causal analyses.

Chapter 3

Learning chain event graphs

Finding that the BN is not always the optimal graphical model for modelling certain systems and processes, this chapter suggests the advantages of using a graphical model based on event trees — the chain event graph — and develops a totally new mechanism to learn CEGs from data, from characterised priors to intelligent learning algorithms.

3.1 Prerequisites

3.1.1 Event Trees

Trees, defined in Definition 11, can be used as an intuitive representation of discrete stochastic processes. They were used in the first ever expositions of mathematical probability by the likes of Huygens in the 17th century [Edwards, 1982]. Influence diagrams, which can be thought of as Bayesian networks with decision and utility nodes, were historically actually derived from decision trees [Shachter, 1986] as a simpler, if sometimes necessarily over-simplified, representation of decision problems. Developing tree-based graphical models is therefore only re-balancing a historical

anomaly. Finally, event trees have a perfect match between their topology and the sample space Ω of the Kolmogorov probability triple $(\Omega, \mathcal{F}, P)$ of a probability model, ensuring that no aspect of the model is ignored in the graphical representation, while Bayesian networks focus on random variables which are real-valued functions of events.

I start by defining event trees formally.

Let $T = (V(T), E(T))$ be a directed tree where $V(T)$ is its node set and $E(T)$ its edge set.

Definition 16. *The set of* SITUATIONS *of T, $S(T)$, is the set of non-leaf nodes $\{v : v \in V(T) \setminus L(T)\}$, where $L(T)$ is the set of leaf nodes of T.*

Let $\mathbb{X}$ be the set of root-to-leaf paths of T, so that $\mathbb{X} = \{\lambda(v_0, v) : v \in L(T)\}$ (recall that v_0 is the root node). $\mathbb{X}$ represents the event space of the model, with every root-to-leaf path an atom of the event space.

In an event tree, each situation $v \in S(T)$ has an associated random variable $X(v)$ defined conditional on having reached v. The state space of $X(v)$ is denoted as $\mathbb{X}(v)$, represented in the tree by $ch(v)$.

Definition 17. *The distribution of $X(v)$ is determined by the* PRIMITIVE PROBABILITIES *$\{\pi(v'|v) = P(X(v) = v') : v' \in \mathbb{X}(v)\}$.*

The probability of an event $\lambda \in \mathbb{X}$ can therefore be calculated by multiplying the primitive probabilities along the path. Conversely, primitive probabilities can be inferred from the probabilities for the events in $\mathbb{X}$.

Definition 18. *The* FLORET *of $v \in S(T)$ is*

$$\mathcal{F}(v) = \left(V\left(\mathcal{F}(v)\right), E\left(\mathcal{F}(v)\right)\right)$$

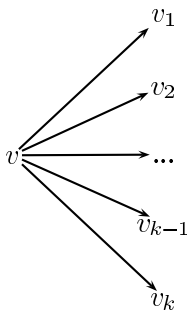

Figure 3.1: Floret of v. This subtree represents both the random variable $X(v)$ and its state space $\mathbb{X}(v)$.

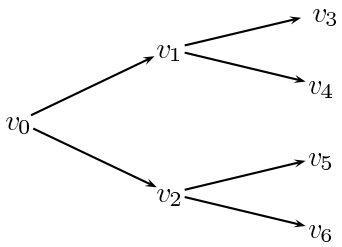

Figure 3.2: Simple event tree. The non-zero-probability events in the joint probability distribution of two Bernoulli random variables, A and B, with A observed before B, can be represented by this tree. Here, all four joint states are possible and hence there are four root-to-leaf paths through the nodes.

where $V(\mathcal{F}(v)) = \{v\} \cup \{v' \in V(T) : (v, v') \in E(T)\}$ *and* $E(\mathcal{F}(v)) = \{e \in E(T) : e = (v, v')\}$.

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 3.1. The floret represents the situation v, the associated random variable $X(v)$ and its sample space $\mathbb{X}(v)$.

Example 19. *Figure 3.2 shows a tree for two Bernoulli random variables, A and B, with A occurring before B. In an education setting A could be the indicator variable of a student passing one module, and B the indicator variable for a subsequent module.*

Here we have random variables $X(v_0) = A$, $X(v_1) = B|(A = 0)$ and $X(v_2) = B|(A = 1)$, and primitive probabilities $\pi(v_1|v_0) = p(A = 0)$, $\pi(v_3|v_1) = p(B = 0|A = 0)$ and so on for every other edge. Path probabilities can be found by multiplying primitive probabilities along a path, e.g. $p(A = 0, B = 0) = p(A = 0)p(B = 0|A = 0) = \pi(v_1|v_0)\pi(v_3|v_1)$ as (v_0, v_1) and (v_1, v_3) are on the path between v_0 and v_3.

3.1.2 Chain Event Graphs

Starting with an event tree T, we extend the definition with three new concepts to form the CEG — STAGES, EDGE COLOURS and POSITIONS – similarly to the approach of [Smith and Anderson, 2008] and [Thwaites et al., 2010].

One of the redundancies that can be eliminated from an ET is that of two situations, v and v' say, which have identical associated edge probabilities despite being defined by different conditioning paths. We say these two situations are in (or at) the same STAGE. This concept is formally defined below.

Definition 20. *Two situations $v, v' \in S(T)$ are in the same stage u if and only if $X(v)$ and $X(v')$ have the same distribution under a bijection*

$$\psi_u(v, v') : \mathbb{X}(v) \to \mathbb{X}(v') \tag{3.1}$$

Definition 20 means that every pair of situations in a stage have a bijection between their sample spaces that identifies which pairs of outcomes have equivalent probabilities.

The set of stages of an event tree T (also called its STAGING) is written $J(T)$. This set partitions the set of situations $S(T)$, due to the associated set of bijections $\{\psi_u(v, v') : v, v' \in u, u \in J(T)\}$ forming an equivalence relation on $S(T)$.

Definition 21. *Any two edges $(v, v^*), (v', v'^*) \in E(T)$ have the same colour if and*

only if $v, v' \in u \in J(T)$ *and* $\psi_u(v, v')(v^*) = v'^*$*, i.e.* v^* *and* v'^* *are considered to have equal probabilities of being reached from* v *and* v' *respectively.*

The edge colours make it clear, when drawn, which edges represent the same primitive probabilities and hence which situations are in the same stage. An alternative approach is to indicate which situations are in the same stage is to draw undirected edges between them, as in [Smith and Anderson, 2008; Thwaites et al., 2010].

Sometimes two situations have even more in common than the distribution over their respective variables: the entire subtrees with the two situations as roots share the same distribution over their paths. These two situations are said to be in the same POSITION. I define this concept formally.

Definition 22. *Two situations* $v, v' \in S(T)$ *are in the same* POSITION w *if and only if there exists a bijection*

$$\phi_w(v, v') : \Lambda(v, T) \to \Lambda(v', T)$$

where $\Lambda(v, T)$ *is the set of paths in* T *from* v *to a leaf node of* T*, such that for every path* $\lambda(v) \in \Lambda(v, T)$*, the ordered sequence of colours in* $\lambda(v)$ *equals the ordered sequence of colours in* $\lambda(v') := \phi_w(v, T)(\lambda(v)) \in \Lambda(v', T)$

I denote the set of positions as $K(T)$. It is clear that $J(T)$ is a partition of $K(T)$, as situations in the same position are in the same stage. $K(T)$ is therefore a finer partition of $S(T)$ than $J(T)$.

Now the CEG can finally be constructed by taking the staged tree $U(T)$ of an event tree and merging situations that are in the same position.

Definition 23. *The* CHAIN EVENT GRAPH *(CEG)* $C(T)$ *of an event tree* T *is the coloured directed graph with vertex set* $V(C)$ *and edge set* $E(C)$ *where*

- *$V(C) = K(T) \cup w_\infty$, so that each non-leaf node in the CEG represents one position and w_∞ represents the set of leaf nodes.*
- *Each edge in $E(C)$ exists for one of the following two reasons.*
  - *For $w, w' \in V(C) \backslash w_\infty$, there is an edge $(w, w') \in E(C)$ if and only if there exist situations $v, v' \in S(T)$ such that $v \in w$, $v' \in w'$ and $(v, v') \in E(T)$.*
  - *For $w \in V(C) \setminus w_\infty$, there is an edge $(w, w_\infty) \in E(C)$ if and only if there exist situations $v \in S(T)$ and $v' \in L(T)$ such that $v \in w$ and $(v, v') \in E(T)$.*
- *The edge $(w, w') \in E(C)$ has the same colour as $(v, v') \in E(T)$ where $v \in w$, $v' \in w'$.*

An example of a CEG that could be constructed from the event tree in Figure 1.1 is shown in Figure 3.3. It can immediately be seen that the CEG is a more compact representation of the probability distribution over the system than the event tree, but without discarding any information reflected by the tree. The non-leaf nodes in Figure 3.3 are positions representing the three hypotheses described in Chapter 1. For example, w_1 is the position reached after knowing what the first module is; if modules A and B are equally hard then the mark distributions are equivalent whether A or B is taken first, and hence the subtrees with A and B as root nodes will have identical distributions. The other positions can be identified with the hypotheses of Example 1 similarly.

It is worth noting that for a finite number of discrete variables that the set of possible CEG models over those variables is a strict superset of the set of possible BN models. While a probability model that can be described by a BN will look different when described by a CEG it will still be the same model. The conditional independence statements described by a BN can always be represented

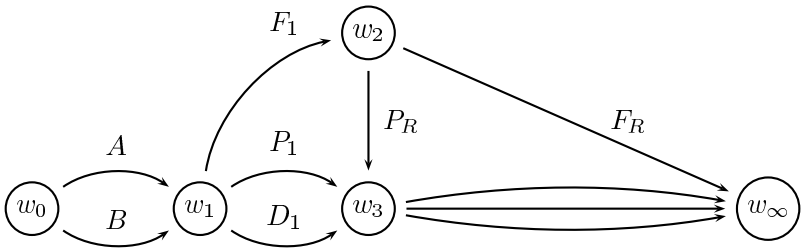

Figure 3.3: The CEG that reflects the three hypotheses of Example 1

by a CEG through stages and positions as is shown in [Smith and Anderson, 2008] and [Thwaites, 2008]. This is because the CEG works on the level of the event space of the probability model while the BN considers only random variables.

3.1.3 Causal trees and CEGs

There is another aspect to event trees (and hence CEGs) that make their use in modelling extremely appealing: their powerful expressiveness in describing causal hypotheses and learning about the effect of external interventions in the system from observational data. Due to reflecting the event space more finely, the range and realism of the possible causal analyses is better than for a Bayesian network of the same system. The intuitiveness of using trees for modelling causal hypotheses was argued forcefully by [Shafer, 1996].

A modeller can learn about some of the probabilities on edges downstream of a variable intervened upon even if observing only data from the idle, unmanipulated system, if he or she is willing to assume that the probability distributions are identical in the two systems. This is true vice versa as well, allowing inferences from an experiment to be valid for the general population. This inference is clearly generalisable to more than one type of control, different demographics, etc., as long as it is represented on the tree. This is simply not possible with a BN, at least not

through manipulation of the graphical structure itself, because the edges of a BN do not represent parts of the event space but rather the conditional independence structure of the system.

For example, consider the event tree in Figure 3.4 (inspired by an example in [Smith, 2010]) which shows the two possible developments of a process conditional on whether it is left undisturbed ("idle") or controlled.

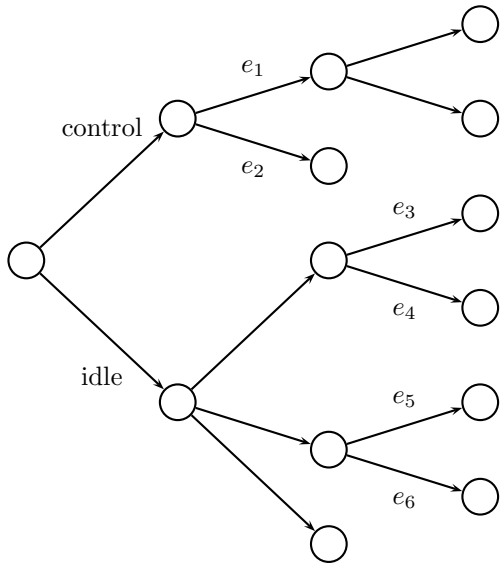

Figure 3.4: Event tree for idle and manipulated versions of the same process

In Figure 3.4, the probabilities of e_1, e_2 might be considered equal to e_3, e_4 respectively, i.e. the associated variables become independent of whether they are in the controlled or idle system, but e_5, e_6 might still be considered to be independent. This cannot be considered graphically with a BN.

In a CEG, the edges will either be coloured the same or merged, making explicit the model assumptions involved and in the latter case doing so efficiently. The range of manipulations possible on a CEG is explored further in [Thwaites et al., 2010].

In Section 5.5 I will show how to implement different interventions in a

dynamic version of the CEG where the same edge is considered exogenously to the structure to be equivalent in both the idle and the manipulated versions of the system, and in Section 6.2.2 I demonstrate its use with a real dataset.

3.2 Conjugate learning of CEGs

It turns out that one convenient property of CEGs is that conjugate updating of the model parameters is possible in a closely analogous fashion to that on a BN as described in Section 2.3. Conjugacy is a crucial part of the model selection algorithm that will be described in Section 3.3, because it leads to closed form expressions for the posterior probabilities of candidate CEGs, which in turn makes it possible to search the often very large model space quickly to find optimal models. The CEG model class will in general be bigger than the BN class for the same random variables, so that a model search will generally take longer but with the benefit that a richer model class is being considered. I demonstrate here how a conjugate analysis on a CEG proceeds.

Let a CEG C have set of stages $J(C) = \{u_1, \ldots, u_k\}$, and let each stage u_i have k_i outgoing edges (labelled $e_1, \ldots, e_{k_i}$) with associated probability vector $\boldsymbol{\pi}_i = (\pi_{i1}, \pi_{i2}, \ldots, \pi_{ik_i})'$ (where $\sum_{j=1}^{k_i} \pi_{ij} = 1$ and $\pi_{ij} > 0$ for $j \in \{1, \ldots, k\}$).

Then under complete sampling, the likelihood of the CEG can be decomposed into a product of the likelihood of each probability vector, i.e.

$$p(\boldsymbol{x}|\boldsymbol{\pi}, C) = \prod_{i=1}^{k} p_i(\boldsymbol{x}_i|\boldsymbol{\pi}_i, C) \tag{3.2}$$

where $\boldsymbol{\pi} = \{\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \ldots, \boldsymbol{\pi}_k\}$, and $\boldsymbol{x} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ is the complete sample data such that each $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ik_n})'$ is the vector of the sample data of the edges (or equivalence class of edges under ψ_u) taken by the units in the sample that start in

stage u_i.

With independence between the units conditional on $\boldsymbol{\pi}$ (i.e. the units are exchangeable)

$$p_i(\boldsymbol{x}_i \mid \boldsymbol{\pi}_i, C) = \prod_{j=1}^{k_i} \pi_{ij}^{x_i^{(j)}} \tag{3.3}$$

where $x_i^{(j)}$ is the number of units which take the jth edge.

Thus, just as for the analogous situation with BNs, the likelihood of a random sample also separates over components of $\boldsymbol{\pi}$. With BNs, a common modelling assumption is of local and global independence of the probability parameters [Spiegelhalter and Lauritzen, 1990]; the corresponding assumption here is that the parameters $\boldsymbol{\pi}_1$,$\boldsymbol{\pi}_2$,...,$\boldsymbol{\pi}_k$ of $\boldsymbol{\pi}$ are all mutually independent a priori. It will then follow, with the separable likelihood, that they will also be independent a posteriori.

If the probabilities $\boldsymbol{\pi}_i$ are a priori assigned a Dirichlet distribution, $\mathrm{Dir}(\boldsymbol{\alpha}_i)$, where $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{ik_i})'$, then for values of $\boldsymbol{\pi}$ where $\sum_{j=1}^{k_i} \pi_{ij} = 1$ and $\pi_{ij} > 0$ for $1 \leq j \leq k_i$, the density of $\boldsymbol{\pi}_i$, $q_i(\boldsymbol{\pi}_i|C)$, can be written

$$q_i(\boldsymbol{\pi}_i|C) = \frac{\Gamma(\alpha_{i1} + \ldots + \alpha_{ik_i})}{\Gamma(\alpha_{i1}) \ldots \Gamma(\alpha_{ik_i})} \prod_{j=1}^{k_i} \pi_{ij}^{\alpha_{ij}-1}$$

where $\Gamma(z) = \int_0^\infty t^{z-1} e^{-t} dt$ is called the Gamma function. It then follows that $\boldsymbol{\pi}_i|\boldsymbol{x}$ $(= \boldsymbol{\pi}_i|\boldsymbol{x}_i)$ also has a Dirichlet distribution, $\mathrm{Dir}(\boldsymbol{\alpha}_i^*)$, a posteriori, where $\boldsymbol{\alpha}_i^* = (\alpha_{i1}^*, \ldots, \alpha_{ik_i}^*)'$, $\alpha_{ij}^* = \alpha_{ij} + x_i^{(j)}$ for $1 \leq j \leq k_i, 1 \leq i \leq k$.

The marginal likelihood of this model, $p(\boldsymbol{x}|C)$, can be written down exactly and is a function of the prior and posterior Dirichlet parameters:

$$p(\boldsymbol{x}|C) = \prod_{i=1}^{k} \left[\frac{\Gamma(\sum_j \alpha_{ij})}{\Gamma(\sum_j \alpha_{ij}^*)} \prod_{j=1}^{k_i} \frac{\Gamma(\alpha_{ij}^*)}{\Gamma(\alpha_{ij})} \right] \tag{3.4}$$

The logarithm of the marginal likelihood, a computationally more useful quantity,

is therefore a linear combination of functions of α_{ij} and α^*_{ij}. Explicitly,

$$\log p(\boldsymbol{x}|C) = \sum_{i=1}^{k} [s(\boldsymbol{\alpha}_i) - s(\boldsymbol{\alpha}^*_i)] + \sum_{i=1}^{k} [t(\boldsymbol{\alpha}^*_i) - t(\boldsymbol{\alpha}_i)] \tag{3.5}$$

where for any vector $\mathbf{c} = (c_1, c_2, \ldots, c_n)'$,

$$s(\mathbf{c}) = \log \Gamma(\sum_{v=1}^{n} c_v) \text{ and } t(\mathbf{c}) = \sum_{v=1}^{n} \log \Gamma(c_v) \tag{3.6}$$

The posterior probability of a CEG C after observing $\boldsymbol{x}$, $q(C|\boldsymbol{x})$, can therefore be calculated using Bayes' Theorem, given a prior probability $q(C)$, as:

$$\log q(C|\boldsymbol{x}) = \log p(\boldsymbol{x}|C) + \log q(C) + K \tag{3.7}$$

for some value K which does not depend on C. This is the SCORE that will be used when searching over the candidate set of CEGs for the model that best describes the data.

3.3 A Local Greedy Search Algorithm for finding the MAP Chain Event Graph

3.3.1 Preliminaries

With $\log q(C|\boldsymbol{x})$ — the log marginal posterior probability of a CEG model C — as a CEG's score, searching for the highest-scoring CEG in the set of all candidate models $\boldsymbol{C}$ becomes equivalent to trying to find the Maximum A Posteriori (MAP) model [Bernardo and Smith, 1994]. The intuitive approach for searching $\boldsymbol{C}$ — calculating

$\log q(C|\boldsymbol{x})$ for every $C \in \boldsymbol{C}$ and choosing

$$C^* := \max_C q(C|\boldsymbol{x}) = \max_C \log q(C|\boldsymbol{x}) \tag{3.8}$$

— is infeasible for any but the most trivial problems. I describe in this section an algorithm for efficiently searching the model space for the MAP CEG by reformulating the model search problem as a clustering problem.

As mentioned in Section 3.1.2, every CEG that can be formed from a given event tree can be identified exactly with a partition of the event tree's nodes into stages. The coarsest partition C_∞ has all nodes with k outgoing edges in the tree in the same stage u_k, for all needed k; the finest partition C_0, in contrast, has each situation in its own stage, except for the trivial cases of those nodes with only one outgoing edge. Defined this way, the search for the highest-scoring CEG is equivalent to searching for the highest-scoring clustering of stages.

Various Bayesian clustering algorithms exist [Lau and Green, 2007], including many involving MCMC [Richardson and Green, 1997]. I show here how to implement an Bayesian agglomerative hierarchical clustering (AHC) exact algorithm related to that of Heard et al. [2006]. The AHC algorithm here is a local search algorithm that begins with the finest partition of the nodes of the underlying ET model (called C_0 above and henceforth) and seeks at each step to find the two nodes that will yield the highest-scoring CEG if combined.

Some optional steps can be taken to simplify the search further, which I will implement here. The first of these involves the calculation of the scores of the proposed models in the algorithm. By assuming that the probability distributions of stages that are formed from the same nodes of the underlying ET are equal in all CEGs, i.e. $p_i(\boldsymbol{x}_i \mid \boldsymbol{\pi}_i, C_1) = p_i(\boldsymbol{x}_i \mid \boldsymbol{\pi}_i, C_2)$ when $u_i \in J(C_1), J(C_2)$, it becomes more efficient to calculate the differences of model scores, i.e. the logarithms of the relevant

Bayes factors, than to calculate the two individual model scores separately. This is because if the stagings $J(C_1)$ and $J(C_2)$ differ only in that stages $u_{1a}, u_{1b} \in C_1$ are combined into $u_{2c} \in C_2$, with all other stages unchanged, then the calculation of the logarithm of their posterior Bayes factor, i.e. the calculation of $\log \frac{q(C_1|\boldsymbol{x})}{q(C_2|\boldsymbol{x})}$, depends only on the stages involved. Using the notation of Equation (3.6), this is done as follows.

$$\log \frac{q(C_1|\boldsymbol{x})}{q(C_2|\boldsymbol{x})} = \log q(C_1|\boldsymbol{x}) - \log q(C_2|\boldsymbol{x}) \tag{3.9}$$

$$= \log q(C_1) - \log q(C_2) + \log p(\boldsymbol{x}|C_1) - \log p(\boldsymbol{x}|C_2) \tag{3.10}$$

$$\begin{aligned} = \log q(C_1) - \log q(C_2) &+ \sum_i \left[s(\boldsymbol{\alpha}_{1i}) - s(\boldsymbol{\alpha}^*_{1i})\right] + \sum_i \left[t(\boldsymbol{\alpha}^*_{1i}) - t(\boldsymbol{\alpha}_{1i})\right] \\ &- \sum_j \left[s(\boldsymbol{\alpha}_{2j}) - s(\boldsymbol{\alpha}^*_{2j})\right] - \sum_j \left[t(\boldsymbol{\alpha}^*_{2j}) - t(\boldsymbol{\alpha}_{2j})\right] \end{aligned} \tag{3.11}$$

$$\begin{aligned} = \log q(C_1) - \log q(C_2) &+ s(\boldsymbol{\alpha}_{1a}) - s(\boldsymbol{\alpha}^*_{1a}) + t(\boldsymbol{\alpha}^*_{1a}) - t(\boldsymbol{\alpha}_{1a}) \\ &+ s(\boldsymbol{\alpha}_{1b}) - s(\boldsymbol{\alpha}^*_{1b}) + t(\boldsymbol{\alpha}^*_{1b}) - t(\boldsymbol{\alpha}_{1b}) \\ &- s(\boldsymbol{\alpha}_{2c}) + s(\boldsymbol{\alpha}^*_{2c}) - t(\boldsymbol{\alpha}^*_{2c}) + t(\boldsymbol{\alpha}_{2c}) \end{aligned} \tag{3.12}$$

where $\boldsymbol{\alpha}_{ab}$ is the vector of hyperparameters of the Dirichlet distribution of the parameter prior for stage u_b of CEG C_a, $a = 1, 2$.

Using the trivial result that for *any* three distinct CEGs $C_1, C_2, C_3 \in \boldsymbol{C}$

$$\log q(C_3|\boldsymbol{x}) - \log q(C_2|\boldsymbol{x}) = \left[\log q(C_3|\boldsymbol{x}) - \log q(C_1|\boldsymbol{x})\right] - \left[\log q(C_2|\boldsymbol{x}) - \log q(C_1|\boldsymbol{x})\right],$$

it can be seen that comparing two proposal CEGs (here C_2 and C_3) from the current CEG (here C_1) can be done equivalently by comparing their individual log Bayes factors against the current CEG with each other, which as shown above requires

fewer calculations.

The calculation of the score for each CEG C, as shown by Equation (3.7), shows that it is formed of two components: the prior probability of the CEG being the true model and the marginal likelihood of the data. These must therefore be set before the algorithm can be run, and it is here that the other simplifications are made.

3.3.2 The prior over the CEG space

For any practical problem $\boldsymbol{C}$, the set of all possible CEGs for a given ET, is likely to be a very large set, making setting a value for $q(C)$ for all $C \in \boldsymbol{C}$ an intractable task. An obvious way to set a non-informative or exploratory prior is to choose the uniform prior, so that $q(C) = \frac{1}{|\boldsymbol{C}|}$. This has the advantages of being simple to set and of eliminating the $\log q(C_1) - \log q(C_2)$ term in Equation (3.12).

A more sophisticated approach is to consider which potential clusters are more or less likely a priori, according to structural or causal beliefs, and to exploit the modular nature of CEGs by stating that the prior log Bayes factor of a CEG relative to C_0 is the sum of the prior log Bayes factors of the individual clusters relative to their components completely unclustered, and that these priors are modular across CEGs. In other words, the prior probability of every stage is independent of which other stages are in the CEG. This approach makes it simple to elicit priors over $\boldsymbol{C}$ from a lay expert, by requiring the elicitation only of the prior probability of each possible stage.

A particular computational benefit of this approach is when the prior Bayes factor of any CEG C with C_0 is believed to be zero, because one or more of its clusters is considered to be impossible. This is equivalent in the algorithm to not including the CEG in its search at all, as though it was never in $\boldsymbol{C}$ in the first place,

with the obvious simplification of the search following.

3.3.3 The prior over the parameter space

Just as when attempting to set $q(C)$, the size of most CEGs in practical situations leads to intractability of setting $p(\boldsymbol{x}|C)$ for each CEG C individually. However, the task is again made possible by exploiting the structure of a CEG with judicious modelling assumptions.

Assuming independence between the likelihoods of the stages for every CEG, so that $p(\boldsymbol{x}|\boldsymbol{\pi}, C)$ is as determined by Equation (3.3), and the fact that $p(\boldsymbol{x}|C) = \int p(\boldsymbol{x}|\boldsymbol{\pi}, C)p(\boldsymbol{\pi}|C)d\boldsymbol{\pi}$, it is clear that to set the marginal likelihood for each CEG is equivalent to setting the prior over the CEG's parameters, i.e. setting $p(\boldsymbol{\pi}|C)$ for each C. With the two further structural assumptions that the stage priors are independent for all CEGs (so that $p(\boldsymbol{\pi}|C) = \prod_{i=1}^{k} p(\boldsymbol{\pi}_i|C)$) and that equivalent stages in different CEGs have the same prior distributions on their probability vectors (i.e. $p(\boldsymbol{\pi}_i|C_1) = p(\boldsymbol{\pi}_i|C_2)$ for all $C_1, C_2 \in \boldsymbol{C}$) it can be seen that the problem of setting $p(\boldsymbol{x}|\boldsymbol{\pi}, C)$ is reduced to setting the parameter priors of each non-trivial floret in C_0 ($p(\boldsymbol{\pi}_i|C_0), i = 1, \ldots, k$) and the parameter priors of every stage that can be formed from the stages of C_0.

The usual prior put on the probability parameters of finite discrete BNs is the product Dirichlet distribution. In [Geiger and Heckerman, 1997] the surprising result was found that a product Dirichlet prior is inevitable if local and global independence are assumed to hold over all Markov equivalent BNs of at least two variables. In the following I will show that a new characterisation can be made for CEGs given the assumptions in the previous paragraph. I will first show that the floret parameters in C_0 must have Dirichlet priors under certain conditions, and then that all CEGs formed by clustering the florets in C_0 must also have Dirichlet priors on the stage

parameters with hyperparameters that are functions of the hyperparameters of the priors under C_0 of the constituent situations. One such characterisation of C_0 is given by Theorem 24 using a concept of "rates" of units along the paths. By rates here I mean the relative expected probabilities of the paths as well as the overall strength of belief in those probabilities.

Theorem 24. *If it is assumed a priori that the rates at which units take the root-to-leaf paths in C_0 are independent ("path independence") then the non-trivial florets of C_0 have Dirichlet priors on their probability vectors.*

The proof of Theorem 24 is based on well-known results concerning properties of the Gamma and Dirichlet distributions, which I review below. I then re-state and prove Theorem 24 as Theorem 28.

Lemma 25. *Let $\gamma_j \sim \text{Gamma}(\alpha_j, \beta), j = 1, \ldots, n$ where $\alpha_j > 0$ for $j \in \{1, \ldots, n\}$, $\beta > 0$ and assume $\perp\!\!\!\perp_{i \in \{1 \ldots n\}} \gamma_i$. Furthermore, let $\theta_j = \frac{\gamma_j}{\gamma}$ for $j \in \{1, \ldots, n\}$, where $\gamma = \sum_{i=1}^{n} \gamma_i$.*

Then $\theta := (\theta_i)_{i=\{1,\ldots,n\}} \sim \text{Dir}\,(\alpha_1, \ldots, \alpha_n)$.

Proof. Kotz et al. [2000]. □

Lemma 26. *Let $I[j] \subseteq \{1, \ldots, n\}$, $\gamma(I[j]) = \sum_{i \in I[j]} \gamma_i$ and $\theta(I[j]) = \sum_{i \in I[j]} \theta_i$.*

Then for any partition $I = \{I[1], \ldots, I[k]\}$ of $\{1, \ldots, n\}$,

$$\theta(I) = (\theta(I[1]), \theta(I[2]), \ldots, \theta(I[k])) \sim \text{Dir}\,(\alpha(I[1]), \ldots, \alpha(I[k]))$$

where $\alpha(I[j]) = \sum_{i \in I[j]} \alpha_i$.

Proof. For any $I[j] \subseteq \{1, \ldots, n\}$,

1. $\perp\!\!\!\perp_{i \in I[j]} \gamma_i$

2. $\gamma(I[j]) \sim \text{Gamma}\left(\alpha(I[j]), \beta\right)$ (a well-known result; see, for example, Weatherburn [1949])

3. for any partition $I = \{I[1], \ldots, I[k]\}$ of $\{1, \ldots, n\}$, $\underset{i \in \{1,\ldots,k\}}{\perp\!\!\!\perp} \gamma(I[j])$

Therefore, as

$$\theta(I[j]) = \sum_{i \in I[j]} \theta_i = \sum_{i \in I[j]} \frac{\gamma_i}{\gamma} = \frac{\gamma(I[j])}{\gamma}, \quad j = 1, \ldots, k \tag{3.13}$$

and $\gamma = \sum_{i=1}^{k} \gamma(I[i])$, the result follows from Lemma 25. □

Lemma 27. *For any $I[j] \subseteq \{1, \ldots, n\}$ where $|I[j]| \geq 2$,*

$$\theta_{I[j]} = \left(\frac{\theta_i}{\theta(I[j])}\right)_{i \in I[j]} \sim \text{Dir}\left((\alpha_i)_{i \in I[j]}\right)$$

Proof. Wilks [1962]. □

Theorem 28. *Let the rates of units along the root-to-leaf paths $\lambda_i \in \mathbb{X}, i \in \{1, \ldots, |\mathbb{X}|\}$ of an event tree T have independent Gamma distributions with the same scale parameter, i.e. $\gamma_i = \gamma(\lambda_i) \sim \text{Gamma}(\alpha_i, \beta), i \in \{1, \ldots, |\mathbb{X}|\}$ and $\underset{i \in \{1,\ldots,|\mathbb{X}|\}}{\perp\!\!\!\perp} \gamma_i$. Then the distribution on each floret in the tree will be Dirichlet.*

Proof. Consider a floret $\mathcal{F}$ with root node v and edge set $\{e_1, \ldots, e_l\}$. The rate for each edge e_i, $\gamma(e_i)$, is equal to

$$\gamma(e_i) = \sum_{\lambda_j \in \Lambda(e_i)} \gamma(\lambda_j) \tag{3.14}$$

where $\Lambda(e_i)$ is the set of root-to-leaf paths that contain e_i, so that $\gamma(e_i) \sim \text{Gamma}(\alpha(e_i), \beta)$ when $\underset{i \in \{1,\ldots,l\}}{\perp\!\!\!\perp} \gamma(e_i)$ as proven by Weatherburn [1949].

Let $I = \{I[\mathcal{F}], I[\overline{\mathcal{F}}]\}$ partition $\mathbb{X}$, where $I[\mathcal{F}] = \{\Lambda(e_1), \ldots, \Lambda(e_l)\}$ and $I[\overline{\mathcal{F}}] = I \setminus I[\mathcal{F}]$. Then by Lemma 27, the probability vector on $\mathcal{F}$ is Dirichlet, where

$$\theta_{I[\mathcal{F}]} \sim \text{Dir}\left((\alpha(e_i))_{i \in \{1,\ldots,l\}}\right)$$

□

$p(\boldsymbol{\pi}_i|C_0)$ is thus entirely determined by rates $\gamma(\lambda)$ on the root-to-leaf paths $\lambda \in \Lambda(v_0, C_0)$ of C_0. This is similar to the "equivalent sample sizes" method of assessing prior uncertainty of Dirichlet hyperparameters in BNs as discussed in Section 2 of [Heckerman, 1999]. This treats the parameters of the prior as having been learnt from hypothetical observed data and an uninformative prior [Steck, 2008]. Here, however, the equivalent sample size is across the entire joint distribution of the model, while in [Heckerman, 1999], [Steck, 2008] and the rest of the BN search literature it applies to each conditional probability distribution separately. Lemma 26 shows that the parameter of the Dirichlet distribution of $p(\boldsymbol{\pi}_i|C_0)$ corresponding to each edge equals the sum of the rates of the root-to-leaf paths passing through that edge.

Another way to characterise all non-trivial situations in C_0 as having Dirichlet priors on their parameter spaces is to use the characterisation of the Dirichlet distribution first proven by Geiger and Heckerman [1997], repeated here as Theorem 29.

Theorem 29. *Let $\{\theta_{ij}\}, 1 \leq i \leq k, 1 \leq j \leq n, \sum_{ij} \theta_{ij} = 1$, where k and n are integers greater than 1, be positive random variables having a strictly positive pdf $f(\{\theta_{ij}\})$. Define $\theta_{i.} = \sum_{j=1}^{n} \theta_{ij}$, $\theta_{I.} = \{\theta_{i.}\}_{i=1}^{k-1}$, $\theta_{j|i} = \theta_{ij}/\sum_j \theta_{ij}$, and $\theta_{J|i} = \{\theta_{j|i}\}_{j=1}^{n-1}$.*

Then if $\{\theta_{I.}, \theta_{J|1}, \ldots, \theta_{J|k}\}$ are mutually independent, $f(\{\theta_{ij}\})$ is Dirichlet.

Proof. Theorem 2 of Geiger and Heckerman [Geiger and Heckerman, 1997]. □

This theorem is used for CEGs as follows.

Corollary 30. *If C_0 has a composite number m of root-to-leaf paths and all Markov equivalent CEGs have independent floret distributions then the vector of probabilities on the root-to-leaf paths of C_0 must have a Dirichlet prior. This means in particular that, from the properties of the Dirichlet distribution, the floret of each situation with at least two outgoing edges has a Dirichlet prior on its edges.*

Proof. Construct an event tree C_0' with m root-to-leaf paths, where the floret of the root node v_0' has k edges and each of the florets extending from the children of v_0' have n edges terminating in leaf nodes, where $m = kn, k \geq 2, n \geq 2$. This will always be possible with a composite m. C_0' describes the same atomic events as C_0 with a different decomposition.

Let the random variable associated with the root floret of C_0' be X, and let the random variable associated with each of the other florets be $Y|X = i, i = 1, \ldots, k$. Let $\theta_{ij} = P(X = i, Y = j)$. Then by the definition of event trees, $P(\theta_{ij} > 0) > 0$ for $1 \leq i \leq k, 1 \leq j \leq n$, and $\sum \theta_{ij} = 1$. By the notation of Theorem 29, $\theta_{i.} = P(X = i)$ and $\theta_{j|i} = P(Y = j|X = i)$.

By hypothesis the floret distributions of C_0' are independent. Therefore the condition of Theorem 29 holds and hence $f(\theta_{ij})$ is Dirichlet. From the equivalence of the atomic events, the probability distribution over the root-to-leaf path probabilities of C_0 is also Dirichlet, and so by Lemma 27, all non-trivial florets of C_0 therefore have Dirichlet priors on their probability vectors. □

To show that the stage parameters of all CEGs in $\boldsymbol{C}$ have Dirichlet priors when assuming stage prior equivalence, an inductive approach will be taken. Because of the assumption of consistency – that two identically composed stages in different

CEGs have identical priors on their parameter space – then for any given CEG C whose stages all have independent Dirichlet priors on their parameters spaces, another CEG C^* formed by clustering two stages u_{1c}, u_{2c} from C into one stage u_{c^*} will have independent Dirichlet priors on all its stages apart from u_{c^*}. It is thus only required to show that $\boldsymbol{\pi}_{c^*}$ has a Dirichlet prior. I prove this result for a class of CEGs called REGULAR CEGs.

Definition 31. *A stage u is* REGULAR *if and only if every path $\lambda \in \Lambda(v_0, C)$ contains either one situation in u or none of the situations in u.*

Definition 32. *A CEG is* REGULAR *if and only if every stage $u \in J(C)$ is regular.*

Theorem 33. *Let C be a regular CEG, and let C^* be the CEG that is formed from C by setting two of its stages u_{1c} and u_{2c} as being in the same stage u_{c^*}, where u_{c^*} is a regular stage, with all other attributes of the CEG unchanged from C.*

If all stages in C have Dirichlet priors, then assuming that equal stages in different CEGs have equivalent priors, all stages in C^* have Dirichlet priors.

Proof. Without loss of generality, let all situations in u_{1c} and u_{2c} have s children each, and let the total number of situations in u_{1c} and u_{2c} be r. Thus there are r situations in u_{c^*}, each with s children. By the assumption of prior consistency across stages, all other stages in C^* have Dirichlet priors on their parameter spaces, so it is only required to prove that u_{c^*} also has a Dirichlet prior.

Consider the CEG C' formed as follows: Let the root node of C', v_0, have 2 children, v_1 and v'. Let v' be a leaf node, and let v_1 have r children, $\{v_1(1), \ldots, v_1(r)\}$, which are equivalent to the situations in u_{c^*}, including the property that they are in the same stage $u_{c'}$. Lastly, let the children of $\{v_1(1), \ldots, v_1(r)\}$, written as $\{v_1(i,j) : i = 1, \ldots, r, j = 1, \ldots, s\}$, be leaf nodes in C'.

By construction, the prior for $u_{c'}$ is the same as that for u_{c^*}.

Now construct another CEG $C^{*\prime}$ from C' by reversing the order of the stages v_1 and $u_{c'}$. The new CEG has root node v_0 with the same distribution as $v_0 \in C'$. v_0 now has two children v' – the same as before – and v_2, which has s children $\{v_2(1), \ldots, v_2(s)\}$ in the same stage. Each node $v_2(i), i = 1, \ldots, s$ has r children $v_2(i,1), \ldots, v_2(i,r)$, all of which are leaf nodes.

The two CEGs $C^{*\prime}$ and C' describe equivalent probability distributions, as it is clear that $P(v_1(i,j)) = P(v_2(j,i)), i = 1, \ldots, r, j = 1, \ldots, s$, where $P(v_1(i,j))$ is the probability of reaching the leaf node $v_1(i,j)$ from the root node under C_1, and similarly for $v_2(j,i)$. The probabilities on the floret of v_2 are thus equal to the probabilities of the situations in the stage of $u_{c'}$, and hence u_{c^*}. Because v_2 is a stage with only one situation, Theorem 24 implies that it has a Dirichlet prior. Therefore u_{c^*} has a Dirichlet prior. □

An alternative justification for assigning a Dirichlet prior to any stage that is formed by clustering situations with Dirichlet priors on their probability distributions which does not depend on assuming equivalency of probability distributions between CEGs derived from different event trees can be obtained by assuming a property analogous to that of "parameter modularity" for BNs [Heckerman, 1995]. This property states that the distribution over structures common to two CEGs should be identical. It is defined in the CEG context as follows.

Definition 34. *Let u be a stage in a CEG C composed of the situations $v_1, \ldots, v_n$ from C_0, each of which has m children $v_{i1}, \ldots, v_{im}, i = 1, \ldots, n$ such that v_{ij} are the same colour for all i for each j. Then u has the property of* MARGIN EQUIVALENCY

if

$$\pi_{uj} = P(v_{1j} \text{ or } v_{2j} \text{ or } \ldots \text{ or } v_{nj} | v_1 \text{ or } v_2 \text{ or } \ldots \text{ or } v_n) \tag{3.15}$$

$$= \frac{\sum_{i=1}^{n} P(v_{ij})}{\sum_{i=1}^{n} P(v_i)} \tag{3.16}$$

is the same for both C and C_0 for $j = 1, \ldots, m$.

Definition 35. *C has margin equivalency if all of its stages have margin equivalency.*

The alternative characterisation can then be stated and proven as follows.

Theorem 36. *Let u_c be a stage as defined in Definition 34 with $m \geq 2$. Then assuming independent priors between the situations for the associated finest-partition CEG C_0 of C, $\boldsymbol{\pi}_{v_i} \sim \text{Dir}(\boldsymbol{\alpha}_i)$ where $\boldsymbol{\alpha}_i = (\alpha_{i1}, \ldots, \alpha_{im})$ for each v_i, $i = 1, \ldots, n$. Furthermore, for both C and C_0, $\boldsymbol{\pi}_u \sim \text{Dir}(\boldsymbol{\alpha}_u)$, where $\boldsymbol{\alpha}_u = (\sum_i \alpha_{i1}, \ldots, \sum_i \alpha_{im})$.*

Proof. From Theorem 28 or Corollary 30, every non-trivial floret in C_0 has a Dirichlet prior on its edges, which includes in this case the situations $v_1, \ldots, v_n$.

Let $\gamma_{ij} = \gamma \pi_{ij}$ for $i = 1, \ldots, n$, $j = 1, \ldots, m$ where $\gamma \sim \text{Gamma}(\sum_{i,j} \alpha_{ij}, \beta)$ and $\pi_{ij} = P(v_i = v_{ij})$, where $v_{ij} \in ch(v_i)$. Then it is well-known that $\gamma_{ij} \sim \text{Gamma}(\alpha_{ij}, \beta)$ for all $1 \leq i \leq n, 1 \leq j \leq m$ for some $\beta > 0$ and that $\perp\!\!\!\perp_j \gamma_{ij}$. As $\perp\!\!\!\perp_i \boldsymbol{\pi}_{v_i}$, $\perp\!\!\!\perp_{ij} \gamma_{ij}$. By Lemma 26 therefore, where $I[j]$ there is the set of edges $\{e_{ij} = e(v_i, v_{ij}), i = 1, \ldots, n\}$ for $j = 1, \ldots, m$,

$$\boldsymbol{\pi}_u \sim \text{Dir}(\sum_i \alpha_{i1}, \ldots, \sum_i \alpha_{im}). \tag{3.17}$$

By margin equivalency, $\boldsymbol{\pi}_u$ must be set the same way for C. □

Note that the posterior of $\boldsymbol{\pi}_u$ for a stage u that is composed of the C_0 situations $v_1, \ldots, v_n$ is thus $\boldsymbol{\pi}_u \mid \boldsymbol{x} \sim \text{Dir}(\boldsymbol{\alpha}_u^*)$ where $\boldsymbol{\alpha}_u^* = \boldsymbol{\alpha}_u + \boldsymbol{x}_u = \sum_{i=1}^{n} \boldsymbol{\alpha}_{v_i} +$

$\sum_{i=1}^{n} \boldsymbol{x}_{v_i}$, where $\boldsymbol{\alpha}_{v_i}$ is the vector of hyperparameters of the distribution of $\boldsymbol{\theta}_{v_i}$ under C_0 and $\boldsymbol{x}_{v_i}$ is the vector of counts on the floret of v_i. Equation (3.12), therefore, becomes

$$\begin{aligned}\log \frac{q(C_1|\boldsymbol{x})}{q(C_2|\boldsymbol{x})} = \log q(C_1) - \log q(C_2) + s(\boldsymbol{\alpha}_{1a}) - s(\boldsymbol{\alpha}^*_{1a}) + t(\boldsymbol{\alpha}^*_{1a}) - t(\boldsymbol{\alpha}_{1a}) \\ + s(\boldsymbol{\alpha}_{1b}) - s(\boldsymbol{\alpha}^*_{1b}) + t(\boldsymbol{\alpha}^*_{1b}) - t(\boldsymbol{\alpha}_{1b}) - s(\boldsymbol{\alpha}_{1a} + \boldsymbol{\alpha}_{1b}) \\ + s(\boldsymbol{\alpha}^*_{1a} + \boldsymbol{\alpha}^*_{1b}) - t(\boldsymbol{\alpha}^*_{1a} + \boldsymbol{\alpha}^*_{1b}) + t(\boldsymbol{\alpha}_{1a} + \boldsymbol{\alpha}_{1b}) \quad (3.18)\end{aligned}$$

Setting priors on the paths rather than the florets also ensures that the distribution of the probabilities of the atomic events remain the same under different tree representations of the event space.

The path priors would in the first instance be set based on expert knowledge of the system at hand, possibly using the "equivalent sample size" heuristic to aid elicitation. In problems where there is no strong prior information, as with the analogous Dirichlet model selection issues for Bayesian networks [Steck and Jaakkola, 2003; Silander et al., 2007], the performance of the selection procedure is rather sensitive to the prior value put on each of the components of $\boldsymbol{\alpha}$.

Within the context of the types of problem discussed here it seems natural in the absence of information to the contrary to set all the components of this vector equal to each other a priori. This implies that for the model with no stages, C_0, we a priori believe that all the atoms — i.e. all possible root to leaf paths — are equally probable, implying that were a model with no structure true then we have no prior information to expect one path to be more likely than another.

Even if we choose to set these all equal, the equivalent sample size parameter $\alpha_{\cdot} \triangleq \mathbf{1}^T\boldsymbol{\alpha}$ — the sum of the rate parameters — has an important role in determining

the performance of the selection procedure. One default is to let $\boldsymbol{\alpha}$ be a vector of 1s. This ensures both a uniform prior over all possible combinations of path probabilities and equal expected path probabilities.

3.3.4 The AHC algorithm

The algorithm thus proceeds as follows:

1. Starting with the initial ET model, form the CEG C_0 with the finest possible partition, where all leaf nodes are placed in the terminal stage u_∞ and all nodes with only one emanating edge are placed in the same stage. Calculate $\log q(C_0|\boldsymbol{x})$ using (3.7).

2. For each pair of situations $v_i, v_j \in C_0$ with the same number of edges, calculate $\log \frac{q(C_1^*|\boldsymbol{x})}{q(C_0|\boldsymbol{x})}$ where C_1^* is the CEG formed by having v_i, v_j in the same stage and keeping all others in their own stage; do not calculate if $q(C_1^*) = 0$.

3. Let $C_1 = \max_{C_1^*}(\log \frac{q(C_1^*|\boldsymbol{x})}{q(C_0|\boldsymbol{x})})$.

4. Now calculate $\log \frac{q(C_2^*|\boldsymbol{x})}{q(C_1|\boldsymbol{x})}$ for each CEG C_2^* that can be formed from a pair of stages in C_1 except where $q(C_2^*) = 0$ a priori, and record $C_2 = \max(q(C_2^*|\boldsymbol{x}))$.

5. Continue for C_3, C_4 and so on until the coarsest partition C_∞ has been reached.

6. Select the CEG C amongst $C_0, \cdots, C_\infty$ that has the highest score $q(C \mid \boldsymbol{x})$ as the MAP model.

Note that the algorithm can also be run backwards, starting from C_∞ and splitting one cluster in two at each step. This approach has the advantage of making the identification of positions in the MAP model easier. Note the similarity in that case to backward stepwise elimination of regression models which discards a variable at each step based on model selection criteria such as BIC [Hocking, 1976].

3.4 A weighted MAX-SAT algorithm for learning Chain Event Graphs

There are two potential and related flaws with using the AHC algorithm of the last section: being a greedy search, it might find a local maximum in the CEG space, but not necessarily the global MAP CEG; and once it decides that two stages should be combined, it does not reverse this decision.

An alternative way to search for the MAP CEG is to reformulate the endeavour as a weighted Maximum Satisfiability (MAX-SAT) problem. This was a successful strategy for searching for MAP BNs [Cussens, 2008] and partitions [Liverani et al., 2010]. Algorithms for solving MAX-SAT problems, weighted and unweighted, have been worked on for decades [Hansen and Jaumard, 1990], and many are available pre-programmed in the UBCSAT package [Tompkins and Hoos, 2005]. By reformulating the MAP CEG search problem as a weighted MAX-SAT problem it is possible to utilise the algorithm-designing expertise of generations of computer scientists.

Weighted MAX-SAT is a modified form of the original SAT problem. The SAT problem has been described as follows in [Hansen and Jaumard, 1990]:

> Given a collection C of m clauses involving n logical variables [which are also called ATOMS, the name I adopt in the following], $x_1, \ldots, x_n$, determine whether or not there exists a truth assignment for C such that all clauses are simultaneously satisfied.

where a clause is a statement in logic consisting of the conjunction and disjunction of boolean variables (or their negative), and a truth assignment is a function that sets the truth values of atoms.

The MAX-SAT problem asks for the assignment in the same situation that

satisfies the maximum number of clauses. The weighted MAX-SAT problem then asks what assignment leads to the minimum sum of weights for clauses that are not satisfied by it, where each clause is now given a weight. A well-known result in propositional logic is that every collection of clauses can be transformed into conjunctive normal form (CNF), i.e. each clause is disjunctive (i.e. a pure OR statement). As the algorithms in the UBCSAT package demand that the clauses are given in CNF form, I will ensure in the following that the clauses are disjunctive.

Recall that under the assumptions detailed earlier, the log posterior probability of a CEG C, which the MAP CEG maximises over the set of possible CEGs $\boldsymbol{C}$, is as given in equation (3.7),

$$\log q(C|\boldsymbol{x}) = \log p(\boldsymbol{x}|C) + \log q(C) + K \tag{3.19}$$

where K is a constant relative to C. Recall also that the log of the marginal likelihood, $\log p(\boldsymbol{x}|C)$, can be written as the sum of functions of its stage hyperparameters.

$$\log p(\boldsymbol{x}|C) = \sum_{i=1}^{k} [s(\boldsymbol{\alpha}_i) - s(\boldsymbol{\alpha}_i^*) + t(\boldsymbol{\alpha}_i^*) - t(\boldsymbol{\alpha}_i)], \tag{3.20}$$

That $\log p(\boldsymbol{x}|C)$ is a sum of functions of its stages and that every stage's contribution would have the same value in any other possible staging is crucial for the representation of the search for the MAP CEG as a weighted MAX-SAT problem.

If the logarithm of the prior $p(C)$ is either constant relative to C — which would imply all possible C are equally probable — or also obeys these two conditions, then the search can be represented as a weighted MAX-SAT problem. An example of a suitable prior is the product prior for partitions given in [Crowley, 1997; McCullagh and Yang, 2006; Booth et al., 2008; Liverani et al., 2010]. Adapted for CEGs it takes

the form

$$p(C) = \frac{\Gamma(\lambda)\lambda^{|C|}}{\Gamma(|C_0| + \lambda)} \prod_{u \in C} \Gamma(|u|) \tag{3.21}$$

where $\lambda > 0$ is a hyperparameter not related to C, as its logarithm is separable over the stages of C:

$$\log p(C) = \log \Gamma(\lambda) - \log \Gamma(|C_0| + \lambda) + \sum_{u \in C} (\log \Gamma(|u|) + 1) \tag{3.22}$$

Therefore $\log P(C \mid \boldsymbol{x})$ is the sum of functions of the component stages of C and the value of those functions does not change in other CEGs.

The weighted MAX-SAT representation of the search for the MAP CEG can now be set up as follows.

The weighted MAX-SAT version of the search for a MAP CEG treats every possible stage that can be formed from C_0 — and therefore every stage u that can be part of some $C \in \boldsymbol{C}$ — as an atom in a propositional logic. Each atom can be true or false, representing whether the associated stage is part of the MAP CEG or not. The clauses which restrict the set of possible assignments of truth values are then chosen as follows, in order to be both disjunctive as required by the conjunctive normal form and reflective of the CEG structure:

1. As stages which share situations cannot both be "true", there will be many clauses for every situation $v \in S$ of the form

$$\overline{u_i} \vee \overline{u_j} \tag{3.23}$$

where $\vee$ indicates logical OR and $\overline{x}$ indicates logical NOT, and where $u_i \cap u_j \ni v$. There will be one of these for each pair of stages that overlaps. Each of these clauses ensures that at most only one of the constituent stages is chosen,

because each one is equivalent, by de Morgan's laws, to

$$\overline{u_i \wedge u_j} \tag{3.24}$$

where $\wedge$ indicates logical AND.

2. Clauses are also needed to ensure that for each situation one stage containing it is considered "true". Thus for each situation $v_i \in S(T)$ there will be exactly one clause of the form

$$u_{i,1} \vee u_{i,2} \vee \cdots \vee u_{i,n(i)} \tag{3.25}$$

where $\{u_{i,j} : 1 \leq j \leq n(i)\} = \{u \subseteq S : v_i \in u\}$.

3. Lastly, each stage has its own clause (known as a FACT). For reasons that will be explained in the following, each clause will be in the form $\overline{u_i}$, i.e. the stage *not* being part of the MAP CEG.

Clauses of type 1 and 2 above are hard clauses. In theory they should be given infinite weights to ensure they are satisfied. In practice this is not implementable with the UBCSAT package and so the weights will be extremely large for the same effect.

The clauses of type 3 are the soft clauses with finite weights, where the weights are a fixed linear function of the associated stage scores. As contributions to the overall weight are only given by clauses *not* satisfied, clauses of type 3 are in the form $\overline{u_i}$, so that if u_i is assigned true then its weight is contributed.

As weighted MAX-SAT aims to *minimise* the overall weight, while MAP search aims to *maximise* posterior probability, it is sufficient, if the contribution made by a stage u to the overall score of a CEG equals $s(u)$, that the weight of the associated clause equals $-s(u)$.

The optimal solution to the weighted MAX-SAT problem described above is now the MAP CEG.

There are two disadvantages to finding the MAP CEG by solving the associated weighted MAX-SAT problem rather than using the AHC algorithm directly.

First, especially for large trees, there is no guarantee that a valid staging will be found within a short period of time, let along the optimal one. The algorithm will search over many solutions which are not valid.

Second, and more prosaically, the UBCSAT package requires all clauses and their weights to be given before starting the search for an optimal assignment, which means that the stage scores must be calculated for all possible stages before the algorithm is run. For a reasonably large tree this problem can be attenuated, after judiciously ensuring that all subjectively impossible stages are not included in the problem (e.g. by assuming the CEG must be hierarchical), by only considering stages of a certain maximum size. While this would be inappropriate for some partition searches (e.g. in [Liverani et al., 2010], which was motivated by clustering genes), it is not always unsuitable for CEGs. In the educational example given in Chapter 1, for example, it might not make sense for more than a few categories of students to perform equivalently, even on the same exam.

In Chapter 6 I will compare the performances of the AHC and weighted MAX-SAT approaches in searching for a MAP CEG with real data of students' exam marks. Before that, I will spend the next two chapters on discussing how to extend CEG learning to a dynamic context.

Chapter 4

Dynamic graphical models

Data can often be time-indexed, with the time measured continuously or discretely. When the time points at which the data is observed are discrete and equally spaced then the data are said to be a TIME SERIES. In this chapter I briefly review various models for time series data, focusing in particular on graphical models for time series data, which are often called DYNAMIC GRAPHICAL MODELS.

4.1 Introduction to modelling time series

Time series data $\boldsymbol{X}$ can be partitioned by the time points at which they were observed. $\boldsymbol{X}$ can then be written as separate data sets $\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_\tau$, where each subscript denotes the associated time point. I use the conventional notation $\boldsymbol{X}^t$ henceforth to mean $\{\boldsymbol{X}_1, \boldsymbol{X}_2, \cdots, \boldsymbol{X}_t\}$.

A STATIONARY process is a time series where the joint distribution of some of its quantities does not change when shifted in time. This modelling assumption implies certain exchangeability conditions in the data, making the absolute time index less relevant. The formal definition follows.

Definition 37. *A time series* $\boldsymbol{X}$ *is stationary when*

$$P(X_{t_1} \in A_{t_1}, \cdots, X_{t_k} \in A_{t_k}) = P(X_{t_1+s} \in A_{t_1}, \cdots, X_{t_k+s} \in A_{t_k}), \tag{4.1}$$

for all possible values k, s *and* $t_1 \cdots t_k$.

I am interested in this thesis in highly multi-dimensional non-stationary processes, typically longitudinal studies of different cohorts. It is only possible to assume that the underlying system process at a particular time point has more in common with its nearer past than its distant past.

One way of modelling non-stationary time series is state-space modelling. This involves modelling observations in terms of an underlying stochastic process. This separation of the observable and the latent processes allows for a very general and hence powerful modelling technique. An excellent introduction to this topic is Durbin and Koopman [2000].

Following the compelling arguments of Dawid [1984], I am only interested in the statistical model's ability to *predict* (or forecast) observations well, and not in inferring values of underlying parameters per se.

4.2 Forecasting with state-space models

State-space models define a latent process $S_1, \cdots, S_\tau$ and the relations between these unobserved variables and the observed time series $\boldsymbol{X}_1, \cdots, \boldsymbol{X}_\tau$. Usually $\boldsymbol{X}_t$ is conditionally independent of all other variables, observables and unobservables, conditional on S_t.

In the prequential approach of Dawid [1984] all that is required from a statistical model of a time series is the quantity $P(\boldsymbol{X}_t \mid \boldsymbol{X}^{t-1})$ for all t. In the state-space

model setting this translates, given the above, to

$$P(\boldsymbol{X}_t \mid \boldsymbol{X}^{t-1}) = \int P(\boldsymbol{X}_t \mid S_t) P(S_t \mid \boldsymbol{X}^{t-1}) dS_t \tag{4.2}$$

$P(S_t \mid \boldsymbol{X}^{t-1})$, in turn, can be written as

$$P(S_t \mid \boldsymbol{X}^{t-1}) = \int P(S_t \mid S^{t-1}) P(S^{t-1} \mid \boldsymbol{X}^{t-1}) dS^{t-1} \tag{4.3}$$

if it is assumed that $S_t \perp\!\!\!\perp \boldsymbol{X}^{t-1} \mid S^{t-1}$.

$P(S^{t-1} \mid \boldsymbol{X}^{t-1})$ can be calculated from Bayes' theorem as

$$P(S^{t-1} \mid \boldsymbol{X}^{t-1}) \propto P(\boldsymbol{X}_{t-1} \mid S^{t-1}) P(S^{t-1} \mid \boldsymbol{X}^{t-2}) \tag{4.4}$$

It can be seen that state-space models admit a recursive definition which allows "on-line" prediction. At time t, $P(S^{t-1} \mid \boldsymbol{X}^{t-1})$ is available. $P(S_t \mid \boldsymbol{X}^{t-1})$ is then obtained using $P(S_t \mid S^{t-1})$ with equation (4.3). Then $P(\boldsymbol{X}_t \mid \boldsymbol{X}^{t-1})$ can be calculated using $P(\boldsymbol{X}_t \mid S_t)$. Bayes' theorem gives $P(S_t \mid \boldsymbol{X}^t)$ (this step is called FILTERING in some time series literature) and the process begins again.

4.3 Dynamic linear models

Dynamic linear models (DLMs) [Harrison and Stevens, 1976; West and Harrison, 1997; Petris et al., 2009] are the classic state-space model. They are defined as follows.

Definition 38. *A* DYNAMIC LINEAR MODEL *consists of time vectors of observations* $\boldsymbol{X}$ *and state parameters* $\boldsymbol{\theta}$ *such that at time* $t = 0$

$$\theta_0 \sim N(m_0, \sigma^2) \tag{4.5}$$

and at time $t \geq 1$

$$X_t = F_t\theta_t + v_t \tag{4.6}$$

$$\theta_t = G_t\theta_{t-1} + w_t \tag{4.7}$$

where F_t *and* G_t *are known matrices of appropriate order and* v_t, w_t *are independent multivariate-Normal variables with mean zero and variances* V_t, W_t *respectively. Equation (4.6) is conventionally called the* OBSERVATION EQUATION *while equation (4.7) is the* STATE EQUATION *or* SYSTEM EQUATION.

The DLM is therefore a state-space model with the added assumptions of linearity and Gaussianity. This allows for exact, conjugate updating of distributions when applying the recursive procedure described above, as originally exploited by the Kalman filter [Kalman, 1960]. When either linearity or Gaussianity are not plausible, conjugacy is often hard to retain. I will introduce in the next chapter a dynamic graphical model that allows for complex multi-variate distributions at each time point that also retains conjugacy. First I discuss some general time series modelling tools that will help in this task.

4.3.1 Multi-process Modelling

Even if a process is determined to be accurately represented by, say, a DLM, it is natural to have uncertainty about the underlying parameter process, e.g. because of knowledge of regime change, or external intervention [West and Harrison, 1989]. In the DLM context this corresponds to being unsure as to the exact nature of F and G in the process equations. This uncertainty can itself be modelled by introducing a new level to the standard state-space model class given above. West and Harrison [1997] call this MULTI-PROCESS MODELLING [Harrison and Stevens, 1976] in the

DLM context, and is also known in the literature as switching state-space models [Frühwirth-Schnatter, 2006].

In the West-Harrison terminology, multi-process models of the first class apply when for all t there is some M which determines the parameter values for the whole process — in the DLM this corresponds to uncertainty about F and G — but it is not known which value of M from a possible set $\mathcal{M}$ is the true one.

This can be transparently dealt with under the Bayesian paradigm as follows. A prior distribution $P(M)$ over $\mathcal{M}$ is specified before the first observations. Predictions for each X_t are calculated as a weighted average over the possible values of M (shown here for a finite $\mathcal{M}$) conditional on observations up to time $t-1$ inclusive:

$$P(X_t \mid X^{t-1}) = \sum_{M \in \mathcal{M}} P(X_t \mid M) P(M \mid X^{t-1}) \tag{4.8}$$

$$= \sum_{M \in \mathcal{M}} \int_{\Theta_t} P(X_t \mid \theta_t) P(\theta_t \mid M) P(M \mid X^{t-1}) d\theta_t \tag{4.9}$$

It can be seen that the usual assumption is that X_t is independent of M given θ_t; in other words, M is purely a description of the latent process, which in turns determines the distribution of the observable process, as before.

The distribution of M is then updated after each observation in the usual way, similarly to the filtering method described above:

$$P(M \mid X^t) \propto P(M \mid X^{t-1}) P(X_t \mid M) \tag{4.10}$$

$P(M \mid X^{t-1})$ was obtained after observing X_{t-1}, and $P(X_t \mid M)$ was calculated in equation (4.8).

Where each process is a DLM, the multi-process model of the first class is, of course, not a DLM itself, but rather a mixture of DLMs.

A usually more realistic assumption is that at each time t a different value of M holds. The dependence between the values of M at different times must then be modelled explicitly, whether the values at different times are entirely independent or highly correlated. This was named by West and Harrison [1997] a multi-process model of the second class. It is clear that this class includes multi-process models of the first class as a special case.

Now the prediction formula is updated in the following way:

$$P(X_t \mid X^{t-1}) = \sum_{M^{t-1} \in \mathcal{M}^{t-1}} \sum_{M_t \in \mathcal{M}} P(X_t \mid M_t) P(M_t \mid M^{t-1}, X^{t-1}) P(M^{t-1} \mid X^{t-1}) \tag{4.11}$$

$$= \sum_{M^{t-1} \in \mathcal{M}^{t-1}} \sum_{M_t \in \mathcal{M}} \int_{\Theta_t} P(X_t \mid \theta_t) P(\theta_t \mid M_t) P(M_t \mid X^{t-1}, M^{t-1}) P(M^{t-1} \mid X^{t-1}) d\theta_t \tag{4.12}$$

While there are is obviously a large class of possible specifications for $P(M_t \mid X^{t-1}, M^{t-1})$, the three "practically important possibilities" recommended by West and Harrison [1997] are as follows:

1. Fixed model probabilities, such that

$$P(M_t \mid X^{t-1}, M^{t-1}) = \pi(M_t) \qquad \text{for all } t \geq 1 \tag{4.13}$$

   Here one needs to only specify one prior over $\mathcal{M}$. This prior remains fixed through time and is not changed by observations.

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

$$\pi(M \mid M') = P(M_t = M \mid M_{t-1} = M') \tag{4.14}$$

are specified a priori for all $M, M' \in \mathcal{M}$, so that

$$P(M_t \mid X^{t-1}, M^{t-1}) = \sum_{M' \in \mathcal{M}} \pi(M \mid M') P(M_{t-1} = M' \mid X^{t-1}) \tag{4.15}$$

Some initial prior distribution over $\mathcal{M}$ would need to be set. These Markov transition probabilities would also not change throughout the process.

3. Higher-order Markov probabilities, where the probabilities of M_t additionally depend on the values of M at $t-2$, $t-3$, etc. as well as $t-1$.

It should be clear that multi-process models of the second class are more complicated than those of the first class with the benefit of allowing flexibility in the models to changing circumstances in the system.

In the next chapter I will introduce a multi-process model where at each time point M_t represents a possible underlying CEG, allowing for far more complicated systems to be modelled than is possible with DLMs but nonetheless retaining conjugacy.

4.4 Steady model

Any state-space model, as shown in Section 4.2, can be used to give $P(X^T)$ because

$$\prod_{t=1}^{T} P(X_t \mid X_{t-1}) P(X_1), \tag{4.16}$$

obtaining each $P(X_t \mid X^{t-1})$ by integrating out S_t from $P(X_t \mid S_t)P(S_t \mid X^{t-1})$. With $P(X_t \mid S_t)$ being given explicitly by the state-space model, there is only a need to specify $P(S_t \mid X^{t-1})$ in general. One way to do so is as a function of $P(S_{t-1} \mid X^{t-1})$, which can itself be calculated using Bayes theorem applied to

$P(S_{t-1} \mid X^{t-2})$ and $P(X_{t-1} \mid S_{t-1})$, the former being a function of $P(S_{t-2} \mid X^{t-2})$, and so on.

In the case of the DLM, the system equation (4.7)

$$\theta_t = G_t\theta_{t-1} + w_t \tag{4.17}$$

does this, because equation (4.7) is true conditional on all possible values of X^{t-1}. This clearly generalises to a possible strategy for all state-space models: assuming

$$P(S_t \mid S^{t-1}) = P(S_t \mid S^{t-1}, X^{t-1}) \tag{4.18}$$

for all X^{t-1} means that

$$P(S_t \mid X^{t-1}) = \int_{\mathcal{S}^{t-1}} P(S_t \mid S^{t-1})P(S^{t-1} \mid X^{t-1})dS^{t-1} \tag{4.19}$$

It should be clear, however, that there are some disadvantages to this approach.

First, setting $P(S_t \mid S^{t-1})$ which holds for all X^{t-1} is over-specification from a forecasting perspective, because we're only interested in how $P(S_t \mid X^{t-1})$ relates to $P(S_{t-1} \mid X^{t-1})$ for the X^{t-1} actually observed.

Second, as has already been noted, setting $P(S_t \mid S^{t-1})$ which is invariant to X^{t-1} means that we learn nothing from the data about the latent process. The prior belief put into the model endures, and any inferences will be sensitive to this belief.

Third, when the state-space model is either non-Gaussian or non-linear a loss of conjugacy of the parameters almost always follows, leading to a reliance on numerical methods and an unfortunate subsequent loss of speed or precision.

Therefore, it is worth trying an alternative approach. One that I will utilise is the power steady model [Smith, 1979, 1981, 1992], which I will refer to here simply as the steady model. This simply states that, letting $p_t(S_t \mid X^{t-1})$ be the probability density function (pdf) of $S_t \mid X^{t-1}$ and $p^*_{t-1}(S_{t-1} \mid X^{t-1})$ the pdf of $S_{t-1} \mid X^{t-1}$,

$$p_t(S_t \mid X^{t-1}) \propto \left\{p^*_{t-1}(S_{t-1} \mid X^{t-1})\right\}^k \tag{4.20}$$

for some value of $0 < k \leq 1$ where the constant of proportionality is uniquely determined to ensure $p_t(S_t \mid X^{t-1})$ is a density. The reciprocal of k is sometimes called the TEMPERATURE as it plays a similar role in physical models of gas diffusion. A similar technique used for ensuring good mixing when carrying out MCMC is called simulated annealing [Geyer and Thompson, 1995].

There are a number of justifications for the power steady model quite apart from its simplicity.

First, it satisfies some intuitive common modelling assumptions, and can be proven to do so in a formal way. These intuitive assumptions are, in decision theoretic terms, that (as described in [Smith, 1979]):

1. decisions should not change between time points in the absence of further information

2. the associated loss from making the decision should not decrease between time points

It can be proven that for a step-loss utility function, the power steady model satisfies the above criteria, and moreover is characterised by them if we also demand that truncating the distribution should leave unaffected the density in the new support except for a new constant of proportionality.

Second, when the transform (4.20) is applied to any multivariate distribution

then all of its conditional independences are retained, and in many cases the distributional family is also left intact. The former assertion can easily be shown: if S_t is a vector of univariate parameters $(S_t^{(1)}, \cdots, S_t^{(n)})$, with conditional independence relationships reflected in the factorisation of $p(S_t \mid X^t)$ (as described in Section 2.2)

$$p(S_t \mid X^t) = \prod_{i=1}^{n} p(S_t^{(i)} \mid Q_t^{(i)}, X^t) \tag{4.21}$$

where $Q_t^{(i)}$ is the minimal sufficient subset of $(S_t^{(1)}, \cdots, S_t^{(i-1)})$ to make the above equation accurate (and obviously with $Q_t^{(1)} = \emptyset$), then applying the power steady transform will yield

$$\big(p(S_t \mid X^t)\big)^k = \prod_{i=1}^{n} \Big(p(S_t^{(i)} \mid Q_t^{(i)}, X^t)\Big)^k, \tag{4.22}$$

making clear that conditional independence relationships will be left intact.

The latter assertion of distributional family invariance depends on the form of the density, but examples of distributions that retain their form after the power steady transform (called the linear expanding distributions in Smith [1979, 1981]) are the normal, Student-t, Gamma, Beta, Dirichlet, and Pareto distributions, and their product versions.

Third, it can be shown that use of the steady model guards against misspecified priors, making predictions more robust to this potential problem. I demonstrate this in two different ways:

1. Let the LOCAL DE ROBERTIS MEASURE DR_A be defined as in [Smith and Daneshkhah, 2010]:

$$d_A^L(f, g) = \sup_{\theta, \phi \in A} \{(\log f(\theta) - \log g(\theta)) - (\log f(\phi) - \log g(\phi))\} \tag{4.23}$$

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. 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,

$$d_A^L(f^*, g^*) = k(d_A^L(f, g)), \tag{4.24}$$

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.

2. 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

$$d_{KL}(p_{t+1}; f_1) - d_{KL}(p_{t+1}; f_2) = k(d_{KL}(p_t; f_1) - d_{KL}(p_t; f_2)) \tag{4.25}$$

where p_t is the density at time t and $p_{t+1} \propto (p_t)^k$. Equation (4.25) says that the K-L distance between the model density 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.

4.5 Dynamic graphical models

There has been much research in the last couple of decades on representing time series with graphical models, for the usual advantages that graphical modelling brings as discussed in Chapter 1. I review some of these models here.

4.5.1 Dynamic Bayesian networks

The most obvious way of representing discrete-time data is with a BN as usual, with one node for the value of each variable at each time point. Then the conditional independence of variables between and within time points can be represented explicitly. Such BNs are called DYNAMIC BAYESIAN NETWORKS (DBNs) [Koller and Lerner, 2000].

The first to propose this idea were Dean and Kanazawa [1988, 1989], although they did not invent the name. The state-space model whose state-space process is a Markov chain (also called a Hidden Markov Model (HMM), and which includes the DLM), for example, holds the following conditional independence properties:

$$X_t \perp\!\!\!\perp X^{t-1}, S^{t-1} \mid S_t \quad t = 1, 2, \cdots \tag{4.26}$$

$$S_t \perp\!\!\!\perp X^{t-1}, S^{t-2} \mid S_{t-1} \quad t = 1, 2, \cdots \tag{4.27}$$

These can be represented as the DBN

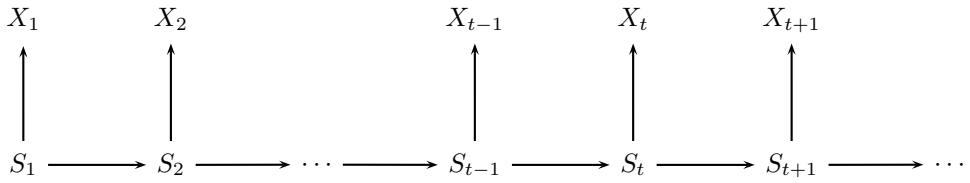

Figure 4.1: Dynamic Bayesian network of state-space model

In many instances, as in the example in Figure 4.1, the same graphical pat-

tern is established between each pair of consecutive time points, and where the whole process is a Markov chain. Thus all that is required in this case is the prior distribution $P(X_1, S_1)$ and the invariant Markov transition function $P(X_{t+1}, S_{t+1} \mid X_t, S_t)$. Instead of drawing a DBN for the whole process, it is then sufficient to draw a TWO-TIME-SLICE BAYESIAN NETWORK (2TBN), which depicts merely the relationship between all consecutive pairs of time points graphically. For example, for the DBN in Figure 4.1, the following 2TBN can be used to represent the process:

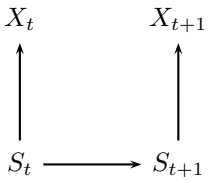

Figure 4.2: Two-time-slice Bayesian network of state-space model

In the most general case DBNs and 2TBNs will not allow closed-form updating, as seen with the special case of non-Gaussian DLMs.

4.5.2 Multiregression dynamic models

One graphical model which does allow for the exact modelling of multivariate time series is the MULTIREGRESSION DYNAMIC MODEL (MDM) [Queen and Smith, 1993; Queen and Albers, 2009]. This models the independences between separate univariate regression DLMs in a conscribed way that ensures conjugacy.

Definition 39. *A* MULTIREGRESSION DYNAMIC MODEL *(MDM) is a BN with nodes* $X_t(1), \cdots, X_t(n)$ *representing the* n *components of the* n*-dimensional observable time series* $\boldsymbol{X}_t$*,* $t = 1, \cdots, \tau$ *and the following conditional independence properties hold for* $i = 2, \cdots, n$ *for all* t*:*

1. $X_t(i) \perp\!\!\!\perp \{X_t(1), \cdots, X_t(i-1)\} \mid Q_i$

2. $X_t(i) \perp\!\!\!\perp \{X^t(1), \cdots, X^t(i-1)\} \mid \{Q_i, X^{t-1}(i)\}$

where Q_i*, as before, denotes the set of parents of* $X_t(i)$ *(which must be a subset of* $\{X_t(1), \cdots, X_t(i-1)\}$*) in the BN. Property 1 is implied by the BN, while property 2 describes the locality of the relationship of the system at time* t *with its past.*

These conditional distributions are explicitly defined in the form of DLMs, i.e., for each $X_t(i)$*, for all* t *and* i*,*

$$X_t(i) = F_t(i)\theta_t(i) + v_t(i) \tag{4.28}$$

$$\boldsymbol{\theta}_t = G_t\boldsymbol{\theta}_{t-1} + \boldsymbol{w_t} \tag{4.29}$$

but where now $F_t(i)$ *is an* s_i*-dimensional column vector (where* s_i *is the dimension of* $\theta_t(i)$*) which can be a (known) function of* $X^{t-1}(i)$ *and* $Q_i(X_t)$*,* G_t *is a block-diagonal matrix with non-zero square sub-matrices* $\{G_t(1), \cdots, G_t(n)\}$ *each respectively of dimension* s_i*,* $v_t(i)$ *has mean 0 and variance* $V_t(i)$*, and* $\boldsymbol{w}_t$ *has mean 0 and a block-diagonal covariance matrix* $W_t = blockdiag\{W_t(1), \cdots, W_t(n)\}$ *where again* $W_t(i)$ *is an* $s_i \times s_i$ *square matrix for* $i = 1, \cdots, n$*.*

Finally, $\boldsymbol{\theta}_0$ *is assigned mean* $\boldsymbol{m}_0$ *and block-diagonal covariance matrix* C_0 *structured similarly to* G_t *and* W_t*.*

Note that $v_t(i)$ and $\boldsymbol{w}_t$ are now not required to be explicitly Gaussian.

It was proven by Queen and Smith [1993] that under the MDM model, if for all $i = 1, \cdots, n$

$$\theta_{t-1}(i) \perp\!\!\!\perp \{\boldsymbol{\theta}_{t-1} \setminus \theta_{t-1}(i)\} \mid \boldsymbol{X}^{t-1} \tag{4.30}$$

then

$$\theta_t(i) \perp\!\!\!\perp \{\boldsymbol{\theta}_t \setminus \theta_t(i)\} \mid \boldsymbol{X}^t \tag{4.31}$$

This says that if the components of $\boldsymbol{\theta}_{t-1}$ are mutually independent up to time $t-1$

then under the MDM $\boldsymbol{\theta}_t$ will also have mutually independent components after additionally observing X_t. Thus if all $\theta_0(i)$ are mutually independent a priori then the state parameters remain so throughout the process.

It was also proven in this case that for all $i \in \{1, \cdots, n\}$

$$\theta_t(i) \perp\!\!\!\perp X^t(i+1), \cdots, X^t(n) \mid X^t(1), \cdots, X^t(i) \tag{4.32}$$

If $\boldsymbol{v}_t$ and $\boldsymbol{w}_t$ are chosen to be Gaussian, then $X_t(i) \mid Q_i(X_t(i))$ will be Gaussian too. Each component will therefore follow the normal DLM as described in section 4.3, which implies conjugate updating as in that case. Note that interventions on individual components can also be easily implemented in the MDM, as shown in Section 4 of Queen and Albers [2009].

4.5.3 Flow networks

A flow network F is a directed graph which models the flow of units from a root node v_0 to a sink node v_s [West, 2001]. Each edge $e \in E(F)$ has an associated CAPACITY $c(e)$ and flow $f(e)$ (where $0 \leq f(e) \leq c(e)$). Flow networks also assume a CONSERVATION OF FLOW property: for every node that is not a root node or sink node (i.e., for every $v \in V(F) \setminus \{v_0, v_s\}$),

$$\sum_{e \in \{e: ch(e) = v\}} f(e) = \sum_{e \in \{e: pa(e) = v\}} f(e), \tag{4.33}$$

i.e., the sum of the flows into v must equal the sum of the flows out of v.

Flow networks are clearly an excellent graphical model for representing flows of material, such as traffic or oil supply. They have been extended to allow probabilistic forecasting by Figueroa-Quiroz [2003] in the form of DYNAMIC FLOW NETWORKS. These extend the flow networks described above by assigning each edge a

TRANSPORT TIME as well as a capacity. In addition, a modified multilevel DLM — a state-space model with a DLM between the observed variables and one level of latent variables, and then another between that level of latent variables and another, and so on — is posited for the root-to-sink path flows for the case when the flow network is strictly hierarchical, i.e. every root-to-sink path is the same length. By modelling the path flows as opposed to the flows through nodes, the conservation of flows requirement is neatly side-stepped, and the path flows can be modelled as independent; the node flows are functions of the path flows and can therefore be recovered. The DLM is different from the canonical one described in Section 4.3 because some of the information is delayed, but similar exact updates and forecasts can be undertaken. Also, just as in the original DLM case, interventions can easily be incorporated within the model class through formal Bayesian intervention.

Although flow networks can be treated directly as BNs (as in [Whitlock and Queen, 2000]), Figueroa-Quiroz [2003] showed that instead the DLM on the flow paths can be drawn as a 2TBN (as shown in Smith and Figueroa [2007]).

The original flow network, however, can in some circumstances be considered as a CEG. Consider the example flow network in Figure 4.3 (adapted from Smith and Figueroa [2007]).

In a hierarchical model such as this, where all root-to-sink paths are the same length (the condition for the dynamic flow network model in [Figueroa-Quiroz, 2003; Smith and Figueroa, 2007]), if the transport time is the same for all edges (and even if not, the flow network can be transformed into one with "phantom" nodes, as shown in Figueroa-Quiroz [2003], in which this condition is fulfilled), then the amount of material at a node at any time t is simply the sum of the amount of material at its parents at time $t-1$, and so the data can be considered in a cohort fashion. If at every node the process that decides where its material ends up next is not dependent

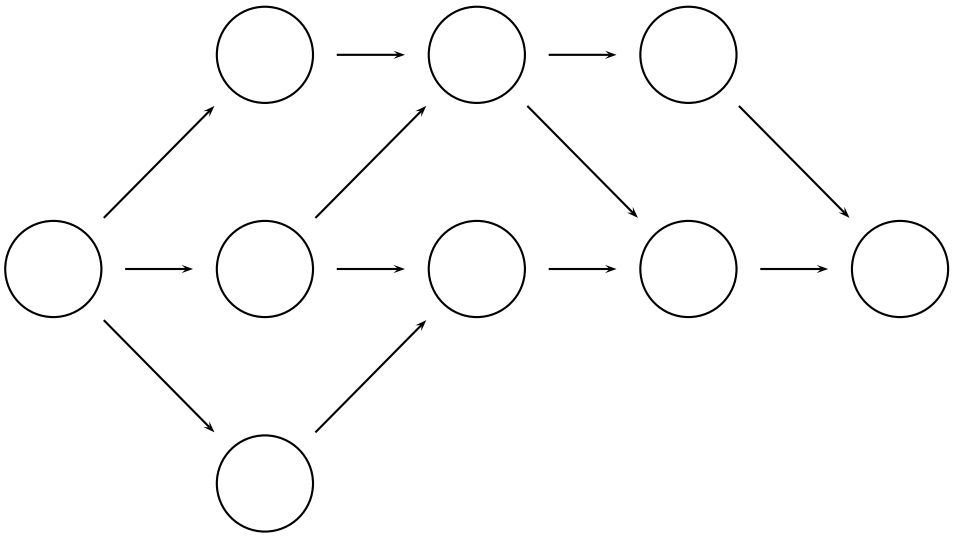

Figure 4.3: Example of a flow network

on the path that the material took to reach the node, and all units of material at a node are exchangeable, then the flow network can also be interpreted as a CEG where each node in the flow network is a position in the CEG sense. While this will not be valid in all cases, e.g. when two physical nodes have identical probabilities distributions over where their respective flows go next, it does seem more natural than interpreting the flow network as a BN.

While the dynamic flow network is very useful in the case when a DLM on paths is valid, a more general dynamic model will be shown in the next chapter which allows for conjugate analyses of non-linear and non-Gaussian multivariate variables with changes in the underlying process, and which also incorporates any needed formal intervention as needed, by at each time point modelling the data as a mixture of CEGs. This model is the DYNAMIC CHAIN EVENT GRAPH.

Chapter 5

Dynamic chain event graphs

I present in this chapter a new dynamic graphical model based on CEGs that admits a conjugate analysis and exact predictions of discrete multivariate time series without sacrificing realism.

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 possible CEGs. Let the set of situations of T, $S(T)$, be denoted by $S = \{v_1, \ldots, v_{|S|}\}$.

At each time point $t = 1, \ldots, \tau$, we wish to predict $x_t(v)$ for all $v \in S$, where $x_t(v)$ is the vector of values of $X(v)$ at time t. Let $\boldsymbol{x}_t = (x_t(v))_{v \in S}$. Then at every time t we need to construct a probability distribution over the possible values of $\boldsymbol{x}_t$ conditional on all previous observations $\boldsymbol{x}^{t-1} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_{t-1})$. The marginal joint distribution $P(\boldsymbol{x}^\tau)$ over time of the full data set can then be calculated as a product of the one-step ahead predictive probabilities $P(\boldsymbol{x}_t \mid \boldsymbol{x}^{t-1})$. Bayes factors associated with different models can then be expressed as a function of these quantities. Note that this factorisation corresponds to the prequential likelihood described by Dawid [1984] used for comparing probabilistic forecasting systems.

The probability distribution of $\boldsymbol{x}_t \mid \boldsymbol{x}^{t-1}$ can be written parametrically as a

function of $\boldsymbol{\theta}_t$, the values of $\theta(v)$ for all $v \in S$ at time t, so that

$$P(\boldsymbol{x}_t \mid \boldsymbol{x}^{t-1}) = \int_{\boldsymbol{\Theta}_t} P(\boldsymbol{x}_t \mid \boldsymbol{\theta}_t, \boldsymbol{x}^{t-1}) p(\boldsymbol{\theta}_t \mid \boldsymbol{x}^{t-1}) d\boldsymbol{\theta}_t \tag{5.1}$$

$\boldsymbol{\theta}_t$ is unknown in the general case. One way to specify the distribution of $\boldsymbol{\theta}_t$ is to assume the process can be described by a DYNAMIC CHAIN EVENT GRAPH. A dynamic chain event graph is defined to be a collection of chain event graphs with possibly different CEGs $C_t(T)$ at each time point for one fixed event tree T.

If $v, v' \in S(T)$ are in the same stage u in a CEG C_t at time t then it is assumed, given the definition of stages, that

$$\theta_t(v) = \theta_t(v') \triangleq \theta_t(u) \tag{5.2}$$

If it is assumed that $\theta_t(u_1) \perp\!\!\!\perp \theta_t(u_2)$ when $u_1 \cap u_2 = \emptyset$ for all t when $u_1, u_2 \in J(C)$ for all possible C then the distribution of $\boldsymbol{\theta}_t$ under a CEG C_t can be written as the product of the distribution of each stage's parameters:

$$p(\boldsymbol{\theta}_t \mid C_t, \boldsymbol{x}^{t-1}) = \prod_{u \in C_t} p(\theta_t(u) \mid C_t, \boldsymbol{x}^{t-1}) \tag{5.3}$$

Therefore equation (5.1) can be written as

$$P(\boldsymbol{x}_t \mid \boldsymbol{x}^{t-1}) = \sum_{C_t \in \boldsymbol{C}} \int_{\boldsymbol{\Theta}_t} P(\boldsymbol{x}_t \mid \boldsymbol{\theta}_t, C_t, \boldsymbol{x}^{t-1}) p(\boldsymbol{\theta}_t \mid C_t, \boldsymbol{x}^{t-1}) P(C_t \mid \boldsymbol{x}^{t-1}) d\boldsymbol{\theta}_t \tag{5.4}$$

$$= \sum_{C_t \in \boldsymbol{C}} \int_{\boldsymbol{\Theta}_t} \left(P(\boldsymbol{x}_t \mid \boldsymbol{\theta}_t, C_t, \boldsymbol{x}^{t-1}) P(C_t \mid \boldsymbol{x}^{t-1}) \prod_{u \in C_t} p(\theta_t(u) \mid C_t, \boldsymbol{x}^{t-1}) \right) d\boldsymbol{\theta}_t \tag{5.5}$$

To carry out a one-step ahead forecast on the system three probability distributions must therefore be specified: the sampling distribution $P(\boldsymbol{x}_t \mid \boldsymbol{x}^{t-1}, \boldsymbol{\theta}_t, C_t)$,

the stage parameter distributions $p(\theta_t(u) \mid C_t, \boldsymbol{x}^{t-1})$, and the CEG distributions $P(C_t \mid \boldsymbol{x}^{t-1})$. I show below how this can be achieved for each item in turn using techniques discussed in previous chapters and some new ideas.

5.1 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. I assume here, however, that for all situations bar the root node v_0 — i.e. for all $v \in S \setminus v_0$ — that $N_t(v)$ equals the value of $x_t^v(v^*)$, the number of times that $X(v^*) = v$ at time t, where v^* is the situation such that $v \in \mathbb{X}(v^*)$, i.e. where v^* is the parent node of v. This matches the view of the units moving along the root-to-leaf paths, similarly to a flow network. I discuss the setting of $N_t(v_0)$ shortly.

$P(\boldsymbol{x}_t \mid \boldsymbol{\theta}_t, C_t, \boldsymbol{x}^{t-1})$ can therefore be written as

$$P(\boldsymbol{x}_t \mid \boldsymbol{\theta}_t, C_t, \boldsymbol{x}^{t-1}) = \sum_{N_t(v_0)} P(\boldsymbol{x}_t \mid N_t(v_0), \boldsymbol{\theta}_t, C_t, \boldsymbol{x}^{t-1}) P(N_t(v_0) \mid \boldsymbol{\theta}_t, C_t, \boldsymbol{x}^{t-1}) \tag{5.6}$$

$$= \sum_{N_t(v_0)} \left(\left[\prod_{v \in S} P(x_t(v) \mid \theta_t(v), x_t^v(v^*)) \right] P(N_t(v_0) \mid \boldsymbol{\theta}_t, C_t, \boldsymbol{x}^{t-1}) \right) \tag{5.7}$$

$$= \sum_{N_t(v_0)} \left(\left[\prod_{v \in S} \mathbb{I}_{\{\sum x_t^{v'}(v) = x_t^v(v^*)\}} \prod_{v' \in \mathbb{X}(v)} \theta_t(v, v')^{x_t^{v'}(v)} \right] P(N_t(v_0) \mid \boldsymbol{\theta}_t, C_t, \boldsymbol{x}^{t-1}) \right) \tag{5.8}$$

where $\mathbb{I}_A$ is the indicator variable for an event A, $x_t^{v_0}(v^*)$ is abuse of notation meaning $N_t(v_0)$, and $\theta(v, v') = P(\mathbb{X}(v) = v')$.

The modelling of the distribution of $N_t(v_0)$ depends on the details of the system under consideration.

Sometimes $N_t(v_0)$ will be known in advance. For example, in the educational scenario of the example in Chapter 1, the number of students enrolling every year might be fixed.

Another 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 λ, where λ 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 (5.8) becomes

$$P(\boldsymbol{x}_t \mid \boldsymbol{\theta}_t, C_t, \boldsymbol{x}^{t-1}) = \prod_{v \in S} \left[\mathbb{I}_{\{\sum x_t^{v'}(v) = x_t^v(v^*)\}} \prod_{v' \in \mathbb{X}(v)} \theta_t(v, v')^{x_t^{v'}(v)} \right] \tag{5.9}$$

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

5.2 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 along the lines shown in Chapter 3.

Consider first the trivial CEG $C_t = C_0$. Recall that if it is assumed that the relative rates of the root-to-leaf paths are independent, each non-trivial floret's parameters must themselves be Dirichlet distributed. Therefore, denoting its collection of hyperparameters as $\alpha_t(v) = (\alpha_t(v, v'))_{v' \in \mathbb{X}(v)}$, the density of $\theta_t(v) \mid C_t = C_0, \boldsymbol{x}^{t-1}$ for a non-trivial floret $v \in C_0$ is

$$f_{\theta_t(v)}(\theta_t(v) \mid C_t = C_0, \boldsymbol{x}^{t-1}) = \Gamma\left(\sum_{v' \in \mathbb{X}(v)} \alpha_t(v, v')\right) \prod_{v' \in \mathbb{X}(v)} \frac{\theta_t(v, v')^{\alpha_t(v,v')-1}}{\Gamma(\alpha_t(v, v'))} \quad (5.10)$$

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 CEG C that is not a trivial partition of C_0. In Chapter 3 it was shown that requiring margin equivalency to hold for its stages $u \in C$ characterises the prior on the floret distributions. A stage u has margin equivalency when

$$P(X(u) \mid \theta, C) = P(X(u) \mid \theta, C_0). \quad (5.11)$$

where $X(u)$ is the random variable with sample space $\bigcup_{v' \in ch(v_u)} \{v' \cup \{\bigcup_{v \in u} \psi(v_u, v)(v')\}\}$, i.e. the edge equivalence classes under a stage, where v_u is any situation in u. With the distribution for florets in C_0 as given above, this implies that the prior proba-

bility of $\theta_t(u) \mid C_t = C, \boldsymbol{x}^{t-1}$ has a Dirichlet distribution too, with hyperparameters that are sums of the corresponding hyperparameters under C_0 of the constituent florets:

$$f_{\theta_t(u)}(\theta_t(u) \mid C_t = C, \boldsymbol{x}^{t-1}) = \Gamma\left(\sum_{v' \in \mathbb{X}(v_u)} \bar{\alpha}_t(u, v')\right) \prod_{v' \in \mathbb{X}(v_u)} \frac{\theta_t(u, v')^{\bar{\alpha}_t(u,v')-1}}{\Gamma(\bar{\alpha}_t(u, v'))} \tag{5.12}$$

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 (5.12) 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 C_0.

With margin equivalency and independence between the floret distributions under C_0, the floret distributions under different CEGs for stages composed of the same situations will always be the same. Therefore the probability distributions for a stage's parameters (5.10) and (5.12) depend only the composition of the stage and not on the rest of the CEG. This property is useful since it allows discussion of 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 (5.10) and (5.12).

Recall that $\theta_t(u)$ is conditionally independent of all other quantities given its hyperparameters $\alpha_t(u)$, which itself depends only on $\alpha_t(v)$, $v \in u$, where $\alpha_t(v)$ is the collection of hyperparameters of $\theta_t(v)$ under C_0. Therefore setting $P(\boldsymbol{\theta}_t \mid C_t, \boldsymbol{x}^{t-1})$ simply requires the setting of $\alpha_t(v)$ for each situation $v \in S$ for every t. This model can be simplified still further by relating the floret distributions between time points. This can be done, as discussed in Section 4.4, with, for example, a (power) steady

model. This relates the floret prior at a time t with its posterior at time $t-1$, i.e.,

$$f_{t+1,v}(\theta) = \mathcal{T}(f^*_{t,v}(\theta)) \tag{5.13}$$

for some function $\mathcal{T}$ for all $t > 1$, where $f_{t,v}(\theta)$ is the density of $\theta_t(v) \mid \boldsymbol{x}^{t-1}, C_t = C_0$ as given in equation (5.10), and $f^*_{t,v}(\theta)$ is the density of $\theta_t(v) \mid \boldsymbol{x}^t, C_t = C_0$. With this, only $\alpha_1(v)$ needs to be set for every $v \in S$ to specify the one-step ahead forecasting model.

The simplest choice of $\mathcal{T}$ is the identity functional, so that

$$f_{t+1,v}(\theta) = f^*_{t,v}(\theta) \tag{5.14}$$

for $t > 1$. With $f_{t,v}(\theta)$ as given in equation (5.10) and $P(x_t(v) \mid \theta_t(v))$ as given by equation (5.8), Bayes' theorem implies that $\theta_t(v)$ has a Dirichlet distribution a posteriori

$$f^*_{\theta_t(v)}(\theta_t(v) \mid C = C_0, \boldsymbol{x}^t) = \Gamma\left(\sum_{v' \in \mathbb{X}(v)} \alpha^*_t(v, v')\right) \prod_{v' \in \mathbb{X}(v)} \frac{\theta_t(v, v')^{\alpha^*_t(v,v')-1}}{\Gamma(\alpha^*_t(v, v'))} \tag{5.15}$$

where $\alpha^*_t(v, v') = \alpha_t(v, v') + x^{v'}_t(v)$, and so

$$\alpha_{t+1}(v) = \alpha^*_t(v) \tag{5.16}$$

$$= \alpha_t(v) + x_t(v) \tag{5.17}$$

As equation (5.17) is true for all $t > 1$, $\alpha_t(v)$ can be written as a function of only $\alpha_1(v)$ and $x^{t-1}(v)$,

$$\alpha_t(v) = \alpha_1(v) + \sum_{\tau=1}^{t-1} x_\tau(v) \tag{5.18}$$

for all $v \in S$.

Letting $\mathcal{T}$ be the identity functional as above reflects a modelling assumption that the underlying probabilities associated with each stage do not evolve for any CEG. 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] discussed in the previous chapter. 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

$$f_{t+1,v}(\theta) \propto (f^*_{t,v}(\theta))^k \tag{5.19}$$

for some $0 < k \leq 1$. It also has the advantage here of preserving the Dirichlet distributions of the floret priors.

With $\alpha^*_t(v) = \alpha_t(v) + x_t(v)$, equation (5.19) 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

$$\alpha_{t+1}(v, v') = k\alpha_t(v, v') + kx_t(v, v') - k + 1 \tag{5.20}$$

Solving this recurrence relation for a constant k yields

$$\alpha_t(v, v') = k^{t-1}(\alpha_1(v, v') - 1) + \sum_{\tau=1}^{t-1} k^{t-\tau} x_\tau(v, v') + 1 \tag{5.21}$$

which heuristically can be seen as weighting recent observations more heavily for the setting of the latest prior, corresponding to the popular exponential-weighted moving average method of estimating parameters in classical time series models.

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 at $v \in S$ then a smaller value of $k(v)$ can be used to indicate increased uncertainty about its new value, just as West and Harrison [1997] do for DLM parameters.

I 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 I use the power steady model as a justifiable conjugate method for making inference about tree models whose floret probabilities evolve.

5.3 The CEG distributions

We have allowed in the previous section for drift over time in the values of probabilities associated with the conditional independence structure implicit in a dynamic CEG model. However, it is necessary to allow in most applications for the possibility that the underlying CEG itself — and not just its parameters — evolves in time. It is unfeasible and usually unnecessary to model all possible changes over the partition space; in most applications it is appropriate to assume that changes in stage structure will be small in number and occur locally.

I therefore propose a dynamic model for the CEGs analogous to the Class 2 Multi-process Models used for dynamic linear models (DLMs) [Harrison and Stevens, 1976; West and Harrison, 1997] discussed in the previous chapter. This was devel-

oped for the case where "no single [model] adequately describes what might happen to the process in the next time interval" [West and Harrison, 1997].

Let $\boldsymbol{C}$ be the set of all possible CEGs of T, and for each $C \in \boldsymbol{C}$ and $t > 1$ let $\pi_t(C) = P(C_t = C \mid \boldsymbol{x}^{t-1})$.

Recall the three modelling strategies proposed by West and Harrison [1997] when using the C2MPM of fixed model probabilities, a first-order Markov transition between the models, or a higher-order Markov transition. While the first possible modelling strategy, of fixed model probabilities, is much the simpler one, the second and third strategies are often going to be more accurate reflections of experts' beliefs. I show here how to implement the second strategy of first-order Markov transitions between CEGs.

At the first time point, $t = 1$, the marginal distribution of the observations $\boldsymbol{x}_1$ can be calculated as follows:

$$P(\boldsymbol{x}_1) = \sum_{C \in \boldsymbol{C}} \pi_1(C) P(\boldsymbol{x}_1 \mid C_1 = C) \tag{5.22}$$

At $t = 2$, after having observed $\boldsymbol{x}_1$,

$$P(\boldsymbol{x}_2 \mid \boldsymbol{x}_1) = \sum_{C \in \boldsymbol{C}} P(\boldsymbol{x}_2 \mid C_2 = C) \pi_2(C) \tag{5.23}$$

$$= \sum_{C \in \boldsymbol{C}} \left[P(\boldsymbol{x}_2 \mid C_2 = C) \sum_{C' \in \boldsymbol{C}} \pi(C \mid C') \pi_1^*(C') \right] \tag{5.24}$$

where $\pi(C \mid C')$ is the fixed transition probability $P(C_t = C \mid C_{t-1} = C')$ for any t,

and

$$\pi_1^*(C') = P(C_1 = C' \mid \boldsymbol{x}_1) \tag{5.25}$$

$$\propto P(\boldsymbol{x}_1 \mid C_1 = C')\pi_1(C'), \tag{5.26}$$

with all the terms on the right-hand side of equation (5.26) available from (5.22). So for all times $t > 1$,

$$P(\boldsymbol{x}_t \mid \boldsymbol{x}^{t-1}) = \sum_{C \in \boldsymbol{C}} \left[P(\boldsymbol{x}_t \mid C_t = C) \sum_{C' \in \boldsymbol{C}} \pi(C \mid C')\pi_{t-1}^*(C') \right] \tag{5.27}$$

with $\pi_{t-1}^*(C')$ available from the previous time point $t-1$. This is the recursive property of state-space models as discussed in Section 4.2.

A common assumption will be that $\pi(C \mid C')$ is larger the "closer" C is to C' in some sense, so that the underlying process is unlikely to change too dramatically over a short period of time in the idle system. If $\pi(C \mid C') = 0$ for some $C \in \boldsymbol{C}$, this has the advantage of reducing the number of terms in equations (5.24) and (5.27). This is particular attractive when calculating $P(\boldsymbol{x} \mid C)$ is very expensive for each CEG C, as is the case for CEGs with a large number of stages or where some stages have large sample spaces.

One way to represent this "closeness" is through a metric over $\boldsymbol{C}$. Meilă [2007] derived a metric for general partition spaces called the "variation of information" metric. It is defined as follows for any two partitions C and C' of a set S:

$$VI(C, C') = H(C) + H(C') - 2I(C, C') \tag{5.28}$$

where

$$H(C) = -\sum_{u \in C} P(u) \log P(u) \tag{5.29}$$

$$I(C, C') = \sum_{u \in C} \sum_{u' \in C'} P(u, u') \log \frac{P(u, u')}{P(u)P(u')} \tag{5.30}$$

and where $P(u) = \frac{|u|}{|S|}$, $P(u, u') = \frac{|u \cap u'|}{|S|}$. The variation of information metric can be justified using information theory.

Recalling that $\boldsymbol{C}$ is a subset of the set of partitions of S, we can therefore set $\pi(C \mid C')$ as a function of $VI(C, C')$. One intuitive way of doing so is to let

$$\pi(C \mid C') = \begin{cases} \rho & \text{if } C = C' \\ |B_\epsilon(C')|^{-1}(1 - \rho) & \text{if } 0 < VI(C, C') \leq \epsilon \\ 0 & \text{otherwise} \end{cases} \tag{5.31}$$

where $0 < \rho < 1$, and $B_\epsilon(C') = \{C \in \boldsymbol{C} : VI(C, C') \leq \epsilon, C \neq C'\}$. This implies that only CEGs in a small neighbourhood around C' are considered and they have an equal chance of occurring. The parameter ρ — the probability of the staging remaining unchanged — determines the conservatism of the process.

The choice of ϵ can be characterised by considering the value of $VI(C, C')$ for some common transformations. For example, when C is obtained from C' by splitting one of the latter's stages, say u' into $u_1, \ldots, u_m$, $VI(C, C')$ in this case was calculated by Meilă [2007] to be

$$VI(C, C') = \frac{|u'| \log |u'|}{|S|} - \frac{1}{|S|} \sum_{l=1}^{m} |u_l| \log |u_l| \tag{5.32}$$

If $|u'| = m \leq |S|$ and $|u_l| = 1$ (so that in this case $u_l \in S$) for $l \in \{1, \ldots, m\}$, then

$$VI(C, C') = \frac{m \log m}{|S|} \tag{5.33}$$

This is also the value of $VI(C, C')$ if C is formed from C' by the reverse of this process, thanks to the symmetry of VI due to its being a metric. So a simple choice for ϵ can be $\frac{m \log m}{|S|}$ for some value of $1 \leq m \leq |S|$, not necessarily an integer. Having $m = |S|$ (i.e., $\epsilon = \log |S|$) would be equivalent to not ruling out any CEG.

If more radical changes in the CEG process are taking place due to external intervention in the system then the methodology in Section 5.5.1 can be deployed.

The VI metric has the disadvantage of its not being immediately clear what its value is between two arbitrary CEGs, making it hard to select only "close" CEGs in an algorithm without calculating its value for all CEGs.

A more intuitive and implementable metric that can be used 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 $|S| = 4$, as an example, is shown in Figure 5.1.

The length of the shortest path between two partitions on the Hasse diagram is a metric on the partition space of S, and I call it ℓ here. A distance of $\ell = 1$ represents the division of a stage or the merging of two stages. One way to set $\pi(C \mid C')$ based on this metric is to do so in a similar way as with the VI metric,

$$\pi(C \mid C') = \begin{cases} \rho & \text{if } C = C' \\ |B_\epsilon(C')|^{-1}(1 - \rho) & \text{if } 0 < \ell(C, C') \leq \epsilon, \\ 0 & \text{otherwise} \end{cases} \tag{5.34}$$

where $B_\epsilon(C') = \{C \in \boldsymbol{C} : \ell(C, C') \leq \epsilon, C \neq C'\}$ is an ϵ-ball of CEGs around C'

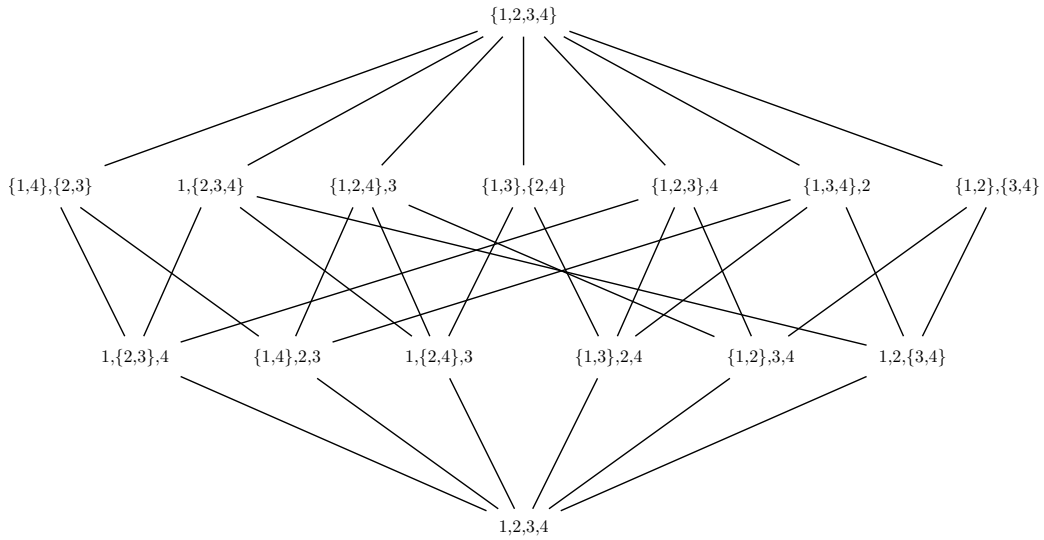

Figure 5.1: The Hasse diagram of the lattice of partitions of S when $|S| = 4$

under the ℓ metric.

The advantage of using this metric ℓ instead of VI is that generating $B_\epsilon(C')$ is much simpler under the former metric. Under VI, it is not clear how to generate general neighbourhoods of a partition C' in the scheme above without calculating $VI(C, C')$ for all $C \in \boldsymbol{C}$, which for even moderately large $|S|$ could quickly become unfeasible. Restricting $\boldsymbol{C}$ further in some way could eliminate this difference, however. Ultimately $\pi(C \mid C')$ must be set according to the statistical needs of the model.

The other term in equation (5.27), $P(C_{t-1} = C' \mid \boldsymbol{x}^{t-1})$, can be calculated for each C_{t-1} using Bayes' theorem:

$$P(C_{t-1} = C' \mid \boldsymbol{x}^{t-1}) \propto P(\boldsymbol{x}_{t-1} \mid C_{t-1} = C')P(C_{t-1} = C' \mid \boldsymbol{x}^{t-2}) \tag{5.35}$$

$$= \frac{P(\boldsymbol{x}_{t-1} \mid C_{t-1} = C')P(C_{t-1} = C' \mid \boldsymbol{x}^{t-2})}{\sum_{C' \in \boldsymbol{C}} P(\boldsymbol{x}_{t-1} \mid C_{t-1} = C')P(C_{t-1} = C' \mid \boldsymbol{x}^{t-2})} \tag{5.36}$$

The $P(C_{t-1} = C' \mid \boldsymbol{x}^{t-2})$ terms on the right-hand side of (5.36) will be already be

available at time $t-1$. The term $P(\boldsymbol{x}_{t-1} \mid C_{t-1} = C')$, meanwhile, can be calculated as follows, using equations (5.8) and (5.12) at time $t-1$ (assuming $N_t(v_0)$ is known):

$$P(\boldsymbol{x}_{t-1} \mid C_{t-1} = C') = \int_{\boldsymbol{\Theta}_{t-1}} P(\boldsymbol{x}_{t-1} \mid \boldsymbol{\theta}_{t-1}, C_{t-1} = C') P(\boldsymbol{\theta}_{t-1} \mid C_{t-1} = C') d\boldsymbol{\theta}_{t-1} \tag{5.37}$$

$$\propto \int_{\boldsymbol{\Theta}_{t-1}} \prod_{u \in C'} \left(\Gamma\left(\sum_{v' \in \mathbb{X}(v_u)} \bar{\alpha}_{t-1}(u, v') \right) \prod_{v' \in \mathbb{X}(v_u)} \frac{\theta_{t-1}(u, v')^{\bar{\alpha}^*_{t-1}(u,v')-1}}{\Gamma(\bar{\alpha}_{t-1}(u, v'))} \right) d\boldsymbol{\theta}_{t-1} \tag{5.38}$$

$$= \prod_{u \in C'} \left(\frac{\Gamma\left(\sum_{v' \in \mathbb{X}(v_u)} \bar{\alpha}_{t-1}(u, v') \right)}{\Gamma\left(\sum_{v' \in \mathbb{X}(v_u)} \bar{\alpha}^*_{t-1}(u, v') \right)} \prod_{v' \in \mathbb{X}(v_u)} \frac{\Gamma(\bar{\alpha}^*_{t-1}(u, v'))}{\Gamma(\bar{\alpha}_{t-1}(u, v'))} \right) \tag{5.39}$$

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 (5.12). Note the similarity to equation (3.4).

The number of terms when calculating equation (5.27) can be reduced further by setting the values of $P(C_{t-1} = C' \mid \boldsymbol{x}^{t-1})$ that are below a threshold q as zero and normalising the remaining probabilities to ensure they still sum to 1. This will guard against calculating $P(C' \mid C)$, $P(\boldsymbol{x}_t \mid C)$ and $P(\boldsymbol{x}_t \mid C')$ for $C' \in B_\epsilon(C)$ for any CEG C that is considered unlikely a posteriori at time $t-1$. A similar approach advocated by Madigan and Raftery [1994] as "Occam's window" is to discard models C' that are not in the set

$$\boldsymbol{C}^*_t = \left\{ C_t \in \boldsymbol{C} : \frac{P(C_t \mid \boldsymbol{x}^t)}{\max_C P(C \mid \boldsymbol{x}^t)} \leq q \right\} \tag{5.40}$$

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.

One last way to consider for easing the calculations is to reduce the number of CEGs under consideration that have overly similar marginal likelihood functions, because these will give similar predictions and hence it is redundant to consider them all separately. A rigorous method of determining the similarity between densities that satisfies desirable properties is to consider their F-DIVERGENCE [Ali and Silvey, 1966]. This is a class of functions defined for two probability distributions P_1, P_2 over the same sample space (as long as they are absolutely continuous with respect to each other over the sample space) as follows.

Define the f-divergence between densities P_1 and P_2 to be

$$\text{fdiv}(P_1, P_2) = f[E_1(g(\phi))] \tag{5.41}$$

where ϕ is the Radon-Nikodym derivative of P_2 relative to P_1, g is a continuous convex function, E_1 denotes expectation with respect to P_1, and f is a non-decreasing function on $\mathbb{R}$.

In the context here this translates into calculating, for any time t

$$\begin{aligned}\text{fdiv}(&P(\boldsymbol{x}_t \mid C_t = C, \boldsymbol{x}^{t-1}), P(\boldsymbol{x}_t \mid C_t = C', \boldsymbol{x}^{t-1})) \\ &= f\left(\sum_{\boldsymbol{x}_t} P(\boldsymbol{x}_t \mid C_t = C, \boldsymbol{x}^{t-1}) \cdot g\left(\frac{P(\boldsymbol{x}_t \mid C_t = C', \boldsymbol{x}^{t-1})}{P(\boldsymbol{x}_t \mid C_t = C, \boldsymbol{x}^{t-1})}\right)\right)\end{aligned} \tag{5.42}$$

There are many choices of g present in the literature. One of the most famous examples is the Kullback-Leibler distance [Kullback and Leibler, 1951] where $g(\phi)$ is $-\log\phi$. I illustrate here $g(\phi) = (\sqrt{\phi} - 1)^2$, known in the literature when $f(x) = \frac{1}{2}x$ as the HELLINGER DISTANCE (and by Ali and Silvey [1966] as Kolmogorov [1963]'s measure of distance). The Hellinger distance between the marginal likelihoods at

time t of two CEGs C and C' is

$$\mathrm{hd}(C, C') = \frac{1}{2} \sum_{\boldsymbol{x}_t} P(\boldsymbol{x}_t \mid C_t = C, \boldsymbol{x}^{t-1}) \cdot g\left(\frac{P(\boldsymbol{x}_t \mid C_t = C', \boldsymbol{x}^{t-1})}{P(\boldsymbol{x}_t \mid C_t = C, \boldsymbol{x}^{t-1})} \right) \tag{5.43}$$

$$= 1 - \sum_{\boldsymbol{x}_t} \sqrt{P(\boldsymbol{x}_t \mid C_t = C', \boldsymbol{x}^{t-1}) P(\boldsymbol{x}_t \mid C_t = C, \boldsymbol{x}^{t-1})} \tag{5.44}$$

Equation (5.44) cannot be calculated exactly. However, it can be related to the Hellinger distance between the distributions of the tree parameters under the two CEGs, a quantity that *can* be calculated exactly.

Let p_1 denote the density $p_1(\boldsymbol{\theta}_t \mid \boldsymbol{x}^{t-1}, C)$ and p_2 the equivalent density under C'. Furthermore, let $p_1^\dagger$ denote the density $p_1^\dagger(\boldsymbol{\theta}_t, \boldsymbol{x}_t \mid \boldsymbol{x}^{t-1}, C)$ and similarly for $p_2^\dagger$. Then

$$p_1^\dagger = p_1 P(\boldsymbol{x}_t \mid \boldsymbol{\theta}_t, C) \tag{5.45}$$

and similarly for $p_2^\dagger$.

Then

$$1 - \mathrm{hd}(p_1^\dagger, p_2^\dagger) = \int_{\boldsymbol{\Theta}_t} \sum_{\boldsymbol{x}_t} \left(p_1 P(\boldsymbol{x}_t \mid \boldsymbol{\theta}_t, C)\right)^{\frac{1}{2}} \left(p_2 P(\boldsymbol{x}_t \mid \boldsymbol{\theta}_t, C')\right)^{\frac{1}{2}} d\boldsymbol{\theta}_t \tag{5.46}$$

$$= \int_{\boldsymbol{\Theta}_t} p_1^{\frac{1}{2}} p_2^{\frac{1}{2}} d\boldsymbol{\theta}_t \tag{5.47}$$

because $P(\boldsymbol{x}_t \mid \boldsymbol{\theta}_t, C) = P(\boldsymbol{x}_t \mid \boldsymbol{\theta}_t, C')$, and so

$$\mathrm{hd}(p_1^\dagger, p_2^\dagger) = \mathrm{hd}(p_1, p_2) \tag{5.48}$$

Now let p_1^* denote $p_1^*(\boldsymbol{\theta}_t \mid \boldsymbol{x}^t, C)$, and similarly for p_2^* with C'. Then it is also true that

$$p_1^\dagger = p_1^* P(\boldsymbol{x}_t \mid \boldsymbol{x}^{t-1}, C) \tag{5.49}$$

and similarly for p_2. Therefore

$$1 - \mathrm{hd}(p_1^\dagger, p_2^\dagger) = \sum_{\boldsymbol{x}_t} \int_{\boldsymbol{\Theta}_t} \left(p_1^* P(\boldsymbol{x}_t \mid \boldsymbol{x}^{t-1}, C)\right)^{\frac{1}{2}} \left(p_2^* P(\boldsymbol{x}_t \mid \boldsymbol{x}^{t-1}, C')\right)^{\frac{1}{2}} d\boldsymbol{\theta}_t \tag{5.50}$$

$$= \sum_{\boldsymbol{x}_t} P(\boldsymbol{x}_t \mid \boldsymbol{x}^{t-1}, C)^{\frac{1}{2}} P(\boldsymbol{x}_t \mid \boldsymbol{x}^{t-1}, C')^{\frac{1}{2}} \int_{\boldsymbol{\Theta}_t} (p_1^* p_2^*)^{\frac{1}{2}} \, d\boldsymbol{\theta}_t \tag{5.51}$$

$$\leq 1 - \mathrm{hd}(C, C') \tag{5.52}$$

as by Schwarz's inequality

$$\int_{\boldsymbol{\Theta}_t} (p_1^* p_2^*)^{\frac{1}{2}} \, d\boldsymbol{\theta}_t \leq \left(\int_{\boldsymbol{\Theta}_t} p_1^* d\boldsymbol{\theta}_t\right)^{\frac{1}{2}} \left(\int_{\boldsymbol{\Theta}_t} p_2^* d\boldsymbol{\theta}_t\right)^{\frac{1}{2}} = 1 \tag{5.53}$$

In fact $\mathrm{hd}(p_1, p_2)$ does not strictly exist because if the underlying stagings are different then p_1 and p_2 are not absolutely continuous with respect to each other. However, the Hellinger distance of the marginal densities of each floret under the different stagings can be calculated, as each will be Dirichlet distributed with the same number of parameters. The Hellinger distance between two Dirichlet densities can be calculated as in [Rauber et al., 2008]. The Hellinger distance between two marginal likelihoods for different CEGs can probably be related to these marginal distances, but the derivation is beyond the scope of this thesis.

So if for two CEGs C, C' where $\pi_t(C), \pi_t(C') > 0$ their Hellinger distance $\mathrm{hd}(C, C')$ is bounded above by some threshold h as calculated above then the consideration of C and C' can be "merged" by changing $\pi_t(C)$ to $\pi_t(C) + \pi_t(C')$ and $\pi_t(C')$ to 0. The sum over $\boldsymbol{C}$ in equation (5.4) will then take place over fewer terms.

5.4 One-step-ahead prediction

Equation (5.4) can now be written, using the foregoing, as

$$P(\boldsymbol{x}_t \mid \boldsymbol{x}^{t-1}) = \sum_{C_t \in \boldsymbol{C}} \int_{\boldsymbol{\Theta}_t} \left(\sum_{C_{t-1} \in \boldsymbol{C}} \pi(C_t \mid C_{t-1}) P(C_{t-1} \mid \boldsymbol{x}^{t-1}) \right) \left(\sum_{N_t(v_0)} P(N_t(v_0) \mid \boldsymbol{\theta}_t, C_t, \boldsymbol{x}^{t-1}) \right.$$
$$\left. \cdot \prod_{u \in C_t} \mathbb{I}_A \left[\Gamma\Big(\sum_{v' \in \mathbb{X}(v_u)} \bar{\alpha}_t(u, v') \Big) \cdot \prod_{v' \in \mathbb{X}(v_u)} \frac{\theta_t(u, v')^{\bar{\alpha}_t(u,v') + \bar{x}_t(u,v') - 1}}{\Gamma(\bar{\alpha}_t(u, v'))} \right] \right) d\boldsymbol{\theta}_t \tag{5.54}$$

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 $\boldsymbol{x}^{t-1}$ then (5.54) can be further simplified to the closed-form solution

$$P(\boldsymbol{x}_t \mid \boldsymbol{x}^{t-1}) = \sum_{C_t \in \boldsymbol{C}} \left(\left(\sum_{C_{t-1} \in \boldsymbol{C}} \pi(C_t \mid C_{t-1}) P(C_{t-1} \mid \boldsymbol{x}^{t-1}) \right) \left(\sum_{N_t(v_0)} P(N_t(v_0) \mid \boldsymbol{x}^{t-1}) \right) \right.$$
$$\left. \left. \cdot \prod_{u \in C_t} \mathbb{I}_A \left[\frac{\Gamma\left(\sum_{v' \in \mathbb{X}(v_u)} \bar{\alpha}_t(u, v') \right)}{\Gamma\left(\sum_{v' \in \mathbb{X}(v_u)} \bar{\alpha}_t^*(u, v') \right)} \prod_{v' \in \mathbb{X}(v_u)} \frac{\Gamma(\bar{\alpha}_t^*(u, v'))}{\Gamma(\bar{\alpha}_t(u, v'))} \right] \right) \right) \tag{5.55}$$

If $N_t(v_0)$ is always known in advance, then (5.55) can be simplified further to become

$$P(\boldsymbol{x}_t \mid \boldsymbol{x}^{t-1}) = \sum_{C_t \in \boldsymbol{C}} \left(\left(\sum_{C_{t-1} \in \boldsymbol{C}} \pi(C_t \mid C_{t-1}) P(C_{t-1} \mid \boldsymbol{x}^{t-1}) \right) \right.$$
$$\left. \cdot \prod_{u \in C_t} \mathbb{I}_A \left[\frac{\Gamma\left(\sum_{v' \in \mathbb{X}(v_u)} \bar{\alpha}_t(u, v') \right)}{\Gamma\left(\sum_{v' \in \mathbb{X}(v_u)} \bar{\alpha}_t^*(u, v') \right)} \prod_{v' \in \mathbb{X}(v_u)} \frac{\Gamma(\bar{\alpha}_t^*(u, v'))}{\Gamma(\bar{\alpha}_t(u, v'))} \right] \right) \tag{5.56}$$

This quantity can be computed with the following algorithm at each time

t, incorporating the techniques mentioned earlier. The quantities not associated directly with $\boldsymbol{x}_t$ can be calculated first:

1. For each C_{t-1}, calculate $P(C_{t-1} \mid \boldsymbol{x}_{t-1})$ using equation 5.36
2. Discard C_{t-1} for which $\frac{\max P(C_{t-1}|\boldsymbol{x}_{t-1})}{P(C_{t-1}|\boldsymbol{x}_{t-1})} \leq q$ for some threshold q, and normalise probabilities of remaining C_{t-1}.
3. For each remaining C_{t-1}, find C_t in $B_\epsilon(C_{t-1})$ and calculate $\pi(C_t \mid C_{t-1})$ under VI or ℓ metric
4. Calculate $P(C_t \mid \boldsymbol{x}^{t-1}) = \sum_{C_{t-1} \in \boldsymbol{C}} \pi(C_t \mid C_{t-1}) P(C_{t-1} \mid \boldsymbol{x}^{t-1})$
5. For C, C' where $\mathrm{hd}(C, C') < h$, change $P(C \mid \boldsymbol{x}^{t-1})$ to $P(C \mid \boldsymbol{x}^{t-1}) + P(C' \mid \boldsymbol{x}^{t-1})$ and $P(C' \mid \boldsymbol{x}^{t-1})$ to 0.

Now for each value of $\boldsymbol{x}_t$ of interest,

1. If necessary, calculate $P(N_t(v_0) \mid \boldsymbol{x}^{t-1})$
2. For each C_t such that $P(C_t \mid \boldsymbol{x}^{t-1}) > 0$, calculate for each $u \in C_t$

$$\frac{\Gamma\left(\sum_{v' \in \mathbb{X}(v_u)} \bar{\alpha}_t(u, v')\right)}{\Gamma\left(\sum_{v' \in \mathbb{X}(v_u)} \bar{\alpha}_t^*(u, v')\right)} \prod_{v' \in \mathbb{X}(v_u)} \frac{\Gamma(\bar{\alpha}_t^*(u, v'))}{\Gamma(\bar{\alpha}_t(u, v'))} \tag{5.57}$$

   where $\alpha_t(v, v') = k^{t-1}(\alpha_1(v, v') - 1) + \sum_{\tau=1}^{t-1} k^{t-\tau} x_\tau + 1$ if using the steady model with a constant k.

3. Substitute all the calculated quantities into equation (5.55).

5.5 Causal intervention

With many forecasting systems there is also an attendant need to consider the effects of external intervention in the system, including by the forecasters themselves

[Harrison and Stevens, 1976; 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. These types of interventions have now been extensively studied on CBNs [Pearl, 2000b; Spirtes et al., 2001] as discussed in Section 2.4. Dynamic extensions of CBNs also exist [Queen and Smith, 1993; Eichler and Didelez, 2007; Queen and Albers, 2009].

As discussed in Section 3.1.3 I believe that tree-based graphical models are very useful in general for carrying out causal analysis, as 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].

I will now show how causal analysis affects the one-step ahead forecast on a dynamic CEG given by equation (5.56) for two different types of intervention not possible on BNs: one on the possible CEGs on a tree T and one on the topology of the tree T itself.

5.5.1 Intervention on the CEG distribution

Suppose that at time t it is determined that some situations will be moved into their own stage $u^\dagger$, leaving all other stages intact. For example, in the educational example of Figure 1.2, 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 $\pi^*_{t-1}(C) = P(C_{t-1} = C \mid \boldsymbol{x}^{t-1})$. Let $\pi^\dagger_t(C) = P(C_t = C \mid \boldsymbol{x}^{t-1}, I_t)$, where I_t is the intervention described above. Then one approach to modelling the intervention is to set $\pi^\dagger_t(C) = \pi^*_{t-1}(C)$ for each $C \in \boldsymbol{C}$ such that $\boldsymbol{u} \in C$, and set $\pi^\dagger_t(C^\dagger) = \pi^*_{t-1}(C)$ and $\pi^\dagger_t(C) = 0$ for $C \in \boldsymbol{C}$ such that $\boldsymbol{u} \notin C$, where $C^\dagger$ is the same as C except that $\boldsymbol{u} \in C^\dagger$ and other stages that contained situations $v \in \boldsymbol{u}$ are reduced accordingly. The effect of this approach is to transfer the probability massed on the CEGs where $\boldsymbol{u} \notin C$ to CEGs where $\boldsymbol{u} \in C$.

One issue that now arises is how the distribution of $\boldsymbol{\theta}_t \mid C_t$ 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. 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]).

5.5.2 Intervention on T

Recalling the event tree pictured in Figure 1.2, consider the case where at time t the course directors decide to eliminate the first module on the tree from the course. This means that the marks that students would have gotten 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 CEGs 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.

I assume that the probability distributions on any unmanipulated florets remain unchanged, just as for CBNs manipulations are local [Pearl, 2000a]. I will also assume here that once an intervention is made, it endures thereon. I 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 $\Lambda(v^*, T)$ for $v^* \in \mathbb{X}(v) \setminus v'$ is now 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 $\mathcal{F}(v)$. 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 $S^\dagger$. First, the distribution over $\boldsymbol{C}^\dagger$, the new set of possible CEGs, must be set. There are several possible choices here. In the absence of any other information, a good default is to let

$$P(C_t = C^\dagger \mid \boldsymbol{x}^{t-1}, I_t(v, v')) = P(C_{t-1} = C \mid \boldsymbol{x}^{t-1}), \tag{5.58}$$

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

Second, the distributions of the stage parameters $\theta_t(u)$ need to be reconsid-

ered. Under the causal assumptions considered here, interventions have only local effects, so a sensible default model is to let $f_{\theta_t(u)}(\theta_t(u) \mid C_t = C, \boldsymbol{x}^{t-1}, I_t(v, v'))$ be calculated as before, i.e. as given in equation (5.12), except of course for $\theta_t(v)$.

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 (5.55) is adjusted to become

$$P(\boldsymbol{x}_t \mid \boldsymbol{x}^{t-1}, I_t(v, v')) = \sum_{C_t^\dagger \in \boldsymbol{C}^\dagger} \left(\left(\sum_{C_{t-1} \in \boldsymbol{C}} \pi^\dagger(C_t^\dagger \mid C_{t-1}) P(C_{t-1} \mid \boldsymbol{x}^{t-1}) \right) \left(\sum_{N_t(v_0)} P(N_t(v_0) \mid \boldsymbol{x}^{t-1}) \right. \right.$$
$$\left. \left. \cdot \prod_{u \in C_t^\dagger} \mathbb{I}_A \left[\frac{\Gamma\left(\sum_{v' \in \mathbb{X}(v_u)} \bar{\alpha}_t(u, v')\right)}{\Gamma\left(\sum_{v' \in \mathbb{X}(v_u)} \bar{\alpha}_t^*(u, v')\right)} \prod_{v' \in \mathbb{X}(v_u)} \frac{\Gamma(\bar{\alpha}_t^*(u, v'))}{\Gamma(\bar{\alpha}_t(u, v'))} \right] \right) \right) \tag{5.59}$$

where $\pi^\dagger(C_t^\dagger \mid C_{t-1}) = \pi(C_t \mid C_{t-1})$ by the argument above, and using the same modelling approximations as before.

Chapter 6

Analysis of exam-mark data using CEGs

The theory and algorithms developed in the thesis up to this point are intended to be used to model real multivariate systems. I show here how implementations of the algorithms perform with real and simulated exam-mark data based on the examples of Chapter 1.

6.1 Learning static CEGs

6.1.1 Simulated data

To demonstrate the efficacy of the AHC algorithm described in Section 3.3 I tested the algorithm using simulated data on the event tree shown in Figure 1.1. I generated the data from a distribution on the tree described by the CEG in Figure 3.3. This CEG corresponds to the three hypotheses described after Example 1, repeated here for convenience:

1. The chances of doing well in the second component are the same whether the

student passed first time or after a resit.

2. The components A and B are equally hard.

3. The distribution of marks for the second component is unaffected by whether students passed or got a distinction for the first component.

Figure 6.1 (on page 112) shows the number of students in the sample who reached each situation. It can be seen that there is naturally a conservation of "flow" at each situation node reflecting that the root-to-leaf paths are the fundamental events of the probability model.

For illustration purposes I set a uniform prior on the CEG priors and a Dir($\mathbf{1}$) uninformative prior distribution on the root-to-leaf paths of C_0. The priors on the floret parameters for any candidate CEG can be calculated from the path priors using the methods of Section 3.3.3.

Recall that at every step of the AHC algorithm that every possible pair of situations is considered for merging. Consider first the merging of two of the situations with two outgoing edges, $F_{1,A}$ and $F_{1,B}$. Under the prior assumptions described in the previous paragraph each of these two florets will have Beta(1,3) priors on its edge probabilities because one edge on each leads directly to a leaf node and the other is on three root-to-leaf paths. The combined stage will therefore have a Beta(2,6) prior on its parameters assuming that the two terminal edges (i.e. the edges $(F_{1,A}, F_{R,A})$ and $(F_{1,B}, F_{R,B})$) are considered equivalent. Using equation (3.12) the log Bayes factor of the posterior probabilities of the CEGs in this case is calculated to be 1.85 in favour of the merged CEG.

Carrying out similar calculations for all the pairs of situations with three edges, it is decided to merge the nodes $P_{1,A}$ and $P_{1,B}$ because the log Bayes factor for the resulting CEG is 3.76 in favour of the merged CEG. Applying the algorithm

to the updated set of nodes and iterating as required, the CEG in Figure 3.3 (shown on page 36) that generated the data was found to be the MAP CEG, validating the AHC algorithm in this instance.

6.1.2 Student exam data

I applied both of the learning algorithms of Chapter 3 — AHC and weighted MAX-SAT — to a real dataset in order to test their efficacy in a real-life situation and to identify remaining issues with their usage as well to make inferences about the education system under investigation. The dataset I used was an appropriately disguised set of marks taken over a 12-year period from four core modules of the MORSE degree course taught at the University of Warwick. A part of the event tree used as the underlying model for the first two modules is shown in Figure 6.2 (on page 113) along with a few illustrative data points. This is a large enough example to illustrate the richness of inference possible with CEG search.

6.1.3 AHC algorithm

For simplicity, the prior distributions on the candidate models and on the root-to-leaf paths for the trivial CEG C_0 were both chosen to be uniform distributions, in the latter case by again assuming $\alpha_i = 1$ for each root-to-leaf path λ_i.

An R program implementing the algorithm found that the MAP CEG model was not C_0, i.e. that there were some non-trivial stages. In total, in fact, 170 situations were clustered into 32 stages. Some of the more interesting stages of this model are described in Table 6.1.

From inspecting the membership of stages it is possible to identify various situations which were discovered to share distributions. For example, students who reach one of the two situations in stage 7 — specifically, the marks for the second

Stage	Probability vector	Students	Situations	Locations	Comments
7	(0.47, 0.44, 0.08)	685	2	1; 1,1,1	High achievers
11	(0.22, 0.43, 0.35)	412	6	3; 1,2; 3,1; 1,1,3	Middling students
13	(0.33, 0.33, 0.33)	16	18	4; 4,2; 4,3	No students appeared in 17 of these situations
17	(0.07, 0.27, 0.66)	86	4	1,3; 3,2; 3,2,4	Struggling students
27	(0.19, 0.56, 0.25)	464	7	1,1,4; 1,2,2; 1,3,2; 1,4,2	More likely to get grade 2 than stage 11
28	(0.11, 0.51, 0.38)	436	6	1,2,3; 3,1,3; 1,2,4	More likely to get grade 3 than stage 27

Table 6.1: Selected stages of MAP CEG model found from data described in Section 6.1.2 using AHC. The columns respectively detail the stage number, posterior expectation of the probability vector of that stage (rounded to two decimal places), number of students passing through that stage in the dataset, number of situations from the original ET in that stage, examples of situations in that stage (shown as sequence of achieved grades 1, 2 or 3, and where 4 means that the grade is missing), and any comments or observations related to that stage.

module after getting the highest grade in the first module or the marks for the fourth module after getting the highest grade in the first three modules — have an expected probability of 0.47 in getting a high mark, an expected probability of 0.44 of getting a middling grade, and an expected probability of only 0.08 of achieving the lowest grade. From being in a stage of their own, it can be deduced that students in these situations have qualitatively different prospects from students in any other situations. In contrast, students who reach one of the four situations in stage 17 have an expected probability of 0.66 of getting the lowest grade. It is instructive that the CEG search found that, by examining stage 17, students getting the top grade in the first module but then only getting the lowest grade in the second module perform identically in the third module to students who only got the lowest grade in the first module and then got the middling grade for the second module.

It is also interesting to note that the 18 situations which had no or almost no students in the data are clustered into one stage. In the absence of prior information distinguishing the situations, I believe this is a positive feature of the algorithm. First, it reduces the dimensionality of the problem relatively painlessly, making the representation of the problem more parsimonious. Second, even if the clustering is ultimately incorrect, due to the very small chance, a posteriori, that many students will traverse these situations in a non-uniform way, the expected loss due to incorrect predictions under any reasonable utility function will be minimal.

It is worth considering at this point how this data-set would traditionally be analysed and contrast it with the method here. One common approach would be to model the events — in this case students' complete exam records — as Poisson-distributed, and hence to use a log-linear model. This models the expected frequencies in a multi-way contingency table using a generalised linear model with a log link function.

The trouble with automatically using a log-linear model or some other regression model for such data is that the assumptions required for the analysis to be valid are generally more restrictive than those for modelling them with a CEG. For example, a log-linear model for Poisson distributed data requires that the contribution to the expected (log) cell frequencies from the factors be linear. It also doesn't easily allow for the sort of complex dependence structures that were found with the CEG search method. With the CEG learning approach one starts only with the event tree and possibilities for situations to have equivalent probability distributions, ensuring the results are more likely to be valid by not assuming too much.

6.1.4 Weighted MAX-SAT

I also undertook a search for a MAP CEG for the data above using the weighted MAX-SAT approach of Section 3.4 under the same assumptions. Due to computer memory restrictions caused by requiring all stage scores to be calculated and stored a priori — a problem discussed in Section 3.4 — it was necessary to restrict the maximum stage size. I ran the algorithm with maximum stage sizes of 2 and 4.

With a maximum stage size of 2, the MAP CEG found had 143 stages, which means there were 27 stages with 2 situations. As each stage's weight was equal to its contribution to the log-likelihood in this application, the sum of the weights was equal to -1 multiplied by the log-likelihood of the CEG. The sum of the weights of the MAP CEG after 10^6 steps was 3951.46, which means the likelihood of the CEG was $\exp(-3951.46)$. As the log-likelihood of C_0 is -3953.40, this indicates that the MAP CEG from the set of CEGs with a maximum stage size of 2 barely fits the data better than C_0. Setting the algorithm to search the set for longer (10^8 steps) yielded only a CEG with log-likelihood of -3947.78. This indicates either that there

is no CEG with maximum stage size of 2 for this dataset that fits the data much better than C_0, or that the algorithm is not able to find that CEG. Either way, this reflects badly on this restriction.

With a maximum stage size of 4, the set of CEGs searched over is clearly bigger than when the maximum stage size is restricted to 2, but the problem this time is that the number of clauses grows super-exponentially. While the event tree, with 170 situations — 85 each with two and three outgoing edges respectively — requires 614,380 clauses in CNF form to describe the problem in weighted MAX-SAT form when the maximum stage size is 2 [using the typology of Section 3.4, these are made up of 170 hard clauses (of type 1) for each situation, 7310 weighted clauses (of type 3) for the possible stages, and $170 \times \binom{85}{2} = 606,900$ hard clauses (of type 2) ensuring that stages that overlap cannot both be chosen], for the case where the maximum number of situations per stage is limited to 4 the number of possible stages is 204,850, the total number of clauses is 1,083,813,258 and the text file containing them is 25GB. Attempting to run the algorithm therefore failed because of memory constraints. Considering that the MAP CEG found with the AHC algorithm contained stages with up to 18 situations, this strategy is clearly not viable.

Some modifications of the usage of weighted MAX-SAT to find MAP CEGs, including combining its usage with AHC, will be discussed in Chapter 7.

6.2 Prediction with dynamic CEGs

In this section I illustrate how to carry out one-step ahead predictions with dynamic CEGs for the 12 years' worth of exam marks used in the last section for two of the undergraduate modules. The underlying event tree used was again that shown in Figure 1.2, so that there are 10 situations, 5 with two edges each describing

availability of marks and another 5 with three edges each for grades.

I made the following assumptions:

1. $N_t(v_0)$, the number of students every year, was known for all values of t
2. The distribution over the root-to-leaf paths at time $t = 1$ under $C_1 = S$ was Dirichlet with all path hyperparameters equal to 1 a priori
3. For the transitions between stagings I used the ℓ metric with $\epsilon = 1$, i.e. only transitions between models that require at most one split or merge were considered possible.

I present here the posterior probabilities $P(C_t \mid \boldsymbol{x}^t)$ for the stagings after $t = 1$ for each time t for different modelling values of the hyperparameters k (the steady model parameter), ρ (the probability of the underlying model not changing) and q (the Occam's window threshold), when analysed with and without an external intervention. In a full analysis this application could be run over a distribution of the hyperparameters, perhaps after taking account of an elicited prior over their possible values. However, to illustrate the efficacy of the methods rather than learn these hyperparameters it is better to hold them fixed so that there is better focus on the impact of various structured assumptions that can be learnt about. Also, I consider ρ and q in particular to be tuning parameters which determine the desired trade-off between the speed and accuracy of the algorithm as well as reflecting real beliefs about the underlying process.

6.2.1 Analysis of the series without intervention

In Table 6.2 I present $P(C_t \mid \boldsymbol{x}^t)$ for $t = 1 \ldots 12$ for the model where $C_1 = \{v_1, v_2, \{v_3, \ldots, v_6\}, \{v_7, \ldots, v_{10}\}\}$ 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

Time	C_t	$P(C_t \mid x^t)$
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

Table 6.2: All possible stagings and their posterior probabilities at each time t for $k = 0.9$, $\rho = 0.9$, $q = 0.2$ with $P(C_1 = \{v_1, v_2, \{v_3, \ldots, v_6\}, \{v_7, \ldots, v_{10}\}\}) = 1$

analysis, as the high value of ρ gives a low prior probability on transitions between stagings and the high value of q makes the Occam's window set of equation (5.40) 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 $P_t(v_i, v_j \in u \mid \boldsymbol{x}^t)$, the a posteriori probability that situations v_i, v_j are in the same stage u at time t, changes over time. This can be calculated from

$$P_t(v_i, v_j \in u \mid \boldsymbol{x}^t) = \sum_{C \in \boldsymbol{C}} P(C_t = C \mid \boldsymbol{x}^t) \mathbb{I}(\exists u \in C : v_i, v_j \in u) \tag{6.1}$$

Figure 6.3 shows this for the information in Table 6.2.

It can be seen very clearly from Figure 6.3 that most situations by time

$t = 6$ are either totally independent of one another or certainly in the same stage. The stages that remain by that time that are not composed of one situation are $\{v_4, v_5, v_6\}$ which are the situations concerning the availability or missingness of grades for the second module after getting a top, middling or bottom grade respectively in the first module; and $\{v_7, v_{10}\}$, which are the situations for the florets describing the grades gained in the second module after either getting a grade 3 or not having a grade at all in the first module. The former stage indicates that whether a mark is available for the second module is independent of the grade achieved in the first one, assuming that is itself not missing; the second stage says that the grade gained in the second module is independent of whether the student did poorly in, or just has a mark missing for, the first module. Both of these inferences would have been impossible to achieve with a Bayesian network search of the same probability model: the first one demands an asymmetric sample space (because if there is no mark available then it cannot be described), while the second is a context-specific conditional independence.

The above analysis is "quick and dirty", in that very clear signals were gained from the dynamic model quickly. To illustrate how the level of detail in the CEG distribution changes as a function of the modelling hyperparameters, allowing more subtle analyses, I ran the algorithm again with radically different values: I set $k = 0.5$ (so that floret distributions are flattened more quickly and therefore past observations more heavily discounted, allowing the data to "speak for itself" more), $\rho = 0.25$ (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 nonetheless kept in the analysis) with the initial degenerate staging distribution $P(C_1 = \{v_1, v_2, \{v_3, \ldots, v_6\}, \{v_7, \ldots, v_{10}\}\}) = 1$ still assumed for consistency. The resulting matrix plot of probabilities of situations being in the same stage against

time is as shown in Figure 6.4.

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 analysis. With regard to the situations concerning the missingness of marks, $\theta(v_3)$ — the probability distribution for the second module's marks being available given that the mark in the first module is itself missing — retains the appearance of being unrelated to the floret distributions at any time point. Until $t = 7$ or so the situations v_4, v_5 and v_6, whose state spaces represent the missingness of marks for the second module after respectively gaining a high, medium or low mark in the first module, had initially high but then gradually falling probabilities of being in the same stage, implying that independence of the missingness of the second module's marks from the marks gained in the first module kept decreasing. At $t = 8$, in contrast, these probabilities become much lower, although the probability distributions of marks being missing after gaining a medium or low mark in the first module are 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. This more subtle analysis was not captured by the more conservative analysis earlier which claimed these situations were simply in the same stage with probability 1 throughout the process. I 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 gained in the second module after getting a poor grade or having a missing mark

in the first module, respectively — are always strongly related, just as in the first analysis. 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 within the system would not have been as easily identifiable if the model class was restricted to Bayesian networks.

6.2.2 Analysis of the series after intervention

I also carried out an analysis with the latter modelling hyperparameters after a hypothesised causal intervention: I 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 believed to 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 6.5.

It can be seen that the probabilities are not too different from those in Figure 6.4, 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 for students who performed differently in the first module under the causal hypothesis considered here.

It is worth noting again the ease with which this causal hypothesis or any other one implemented on the structure or the staging is implemented in the prediction algorithm.

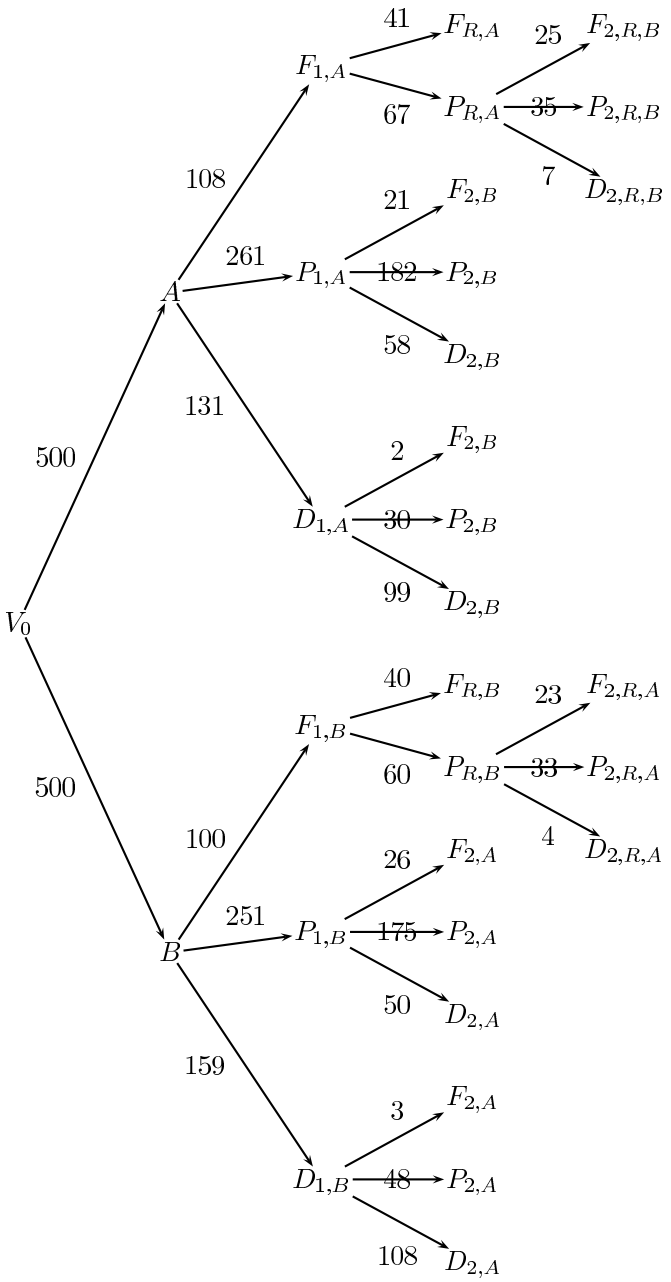

Figure 6.1: The event tree from Example 1 with the numbers representing the number of students in a simulated sample who reached each situation.

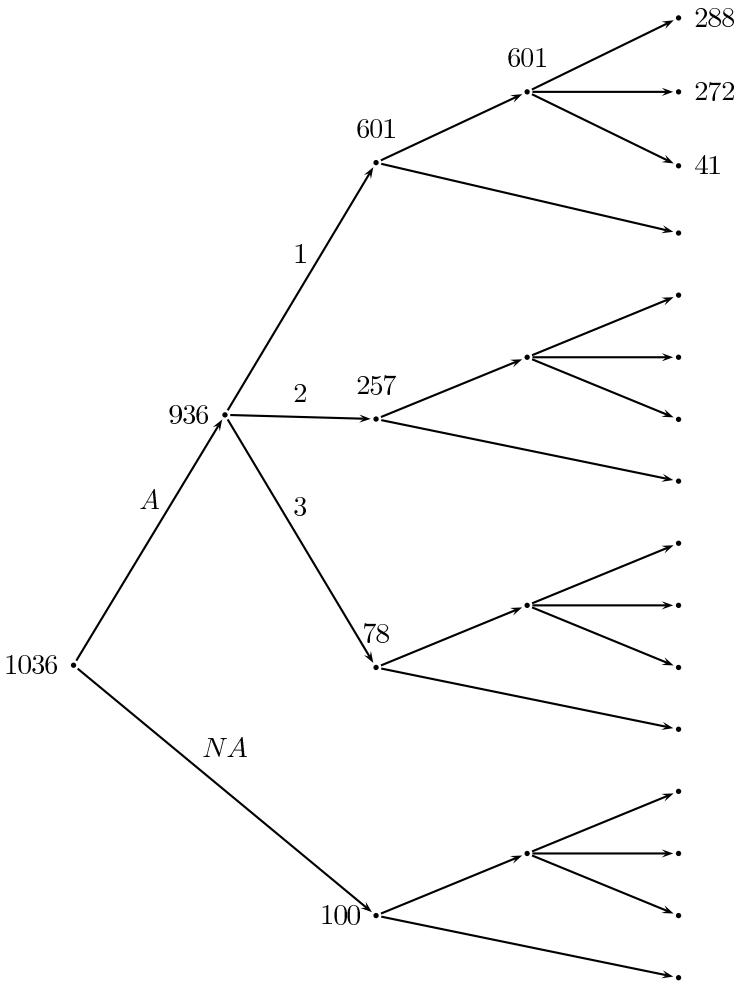

Figure 6.2: Sub-tree of the event tree of possible grades for the MORSE degree course at the University of Warwick. Each floret of two edges describes whether a student's marks are available for a particular module (denoted by the edge labelled A for the first module) or whether they are missing (NA). If they are available, then they are counted as grade 1 if are 70% or higher, grade 2 if they are between 50% and 69% inclusive, and grade 3 if they are below 50%. Some illustrative count data are shown on corresponding nodes.

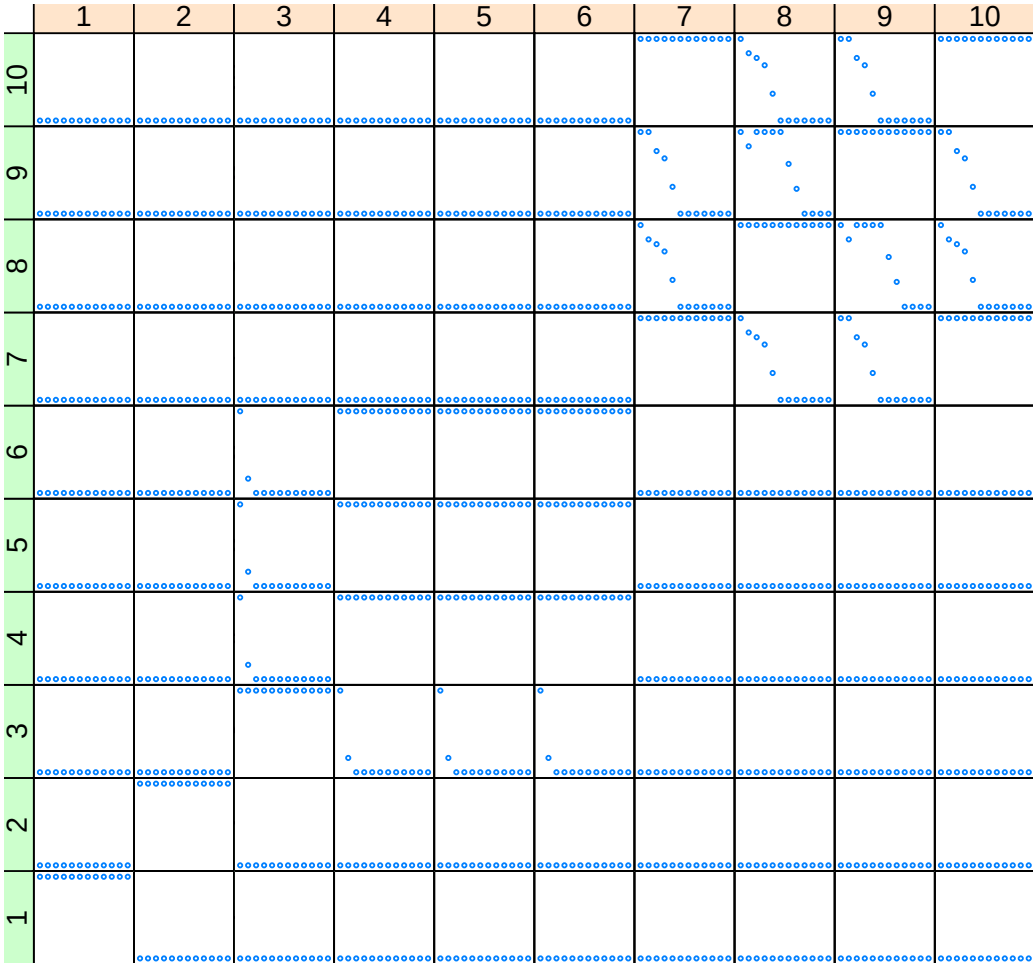

Figure 6.3: Plots of probabilities that each pair of situations are in the same stage for different values of t, for the case when $k = 0.9$, $\rho = 0.9$, $q = 0.2$ with $P(C_1 = \{v_1, v_2, \{v_3, \ldots, v_6\}, \{v_7, \ldots, v_{10}\}\}) = 1$, using the values in Table 6.2

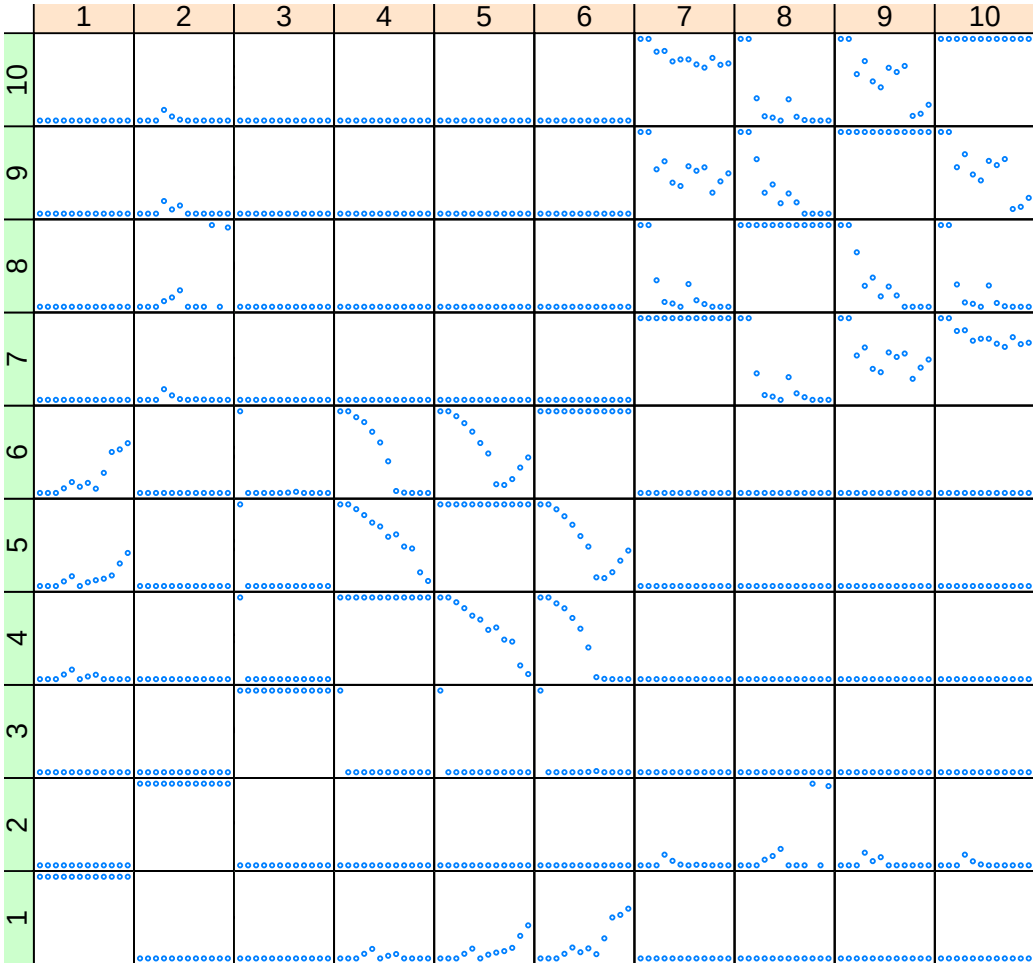

Figure 6.4: Plots of probabilities that each pair of situations are in the same stage for different values of t, for the case when $k = 0.5$, $\rho = 0.25$, $q = 0.05$ with $P(C_1 = \{v_1, v_2, \{v_3, \ldots, v_6\}, \{v_7, \ldots, v_{10}\}\}) = 1$

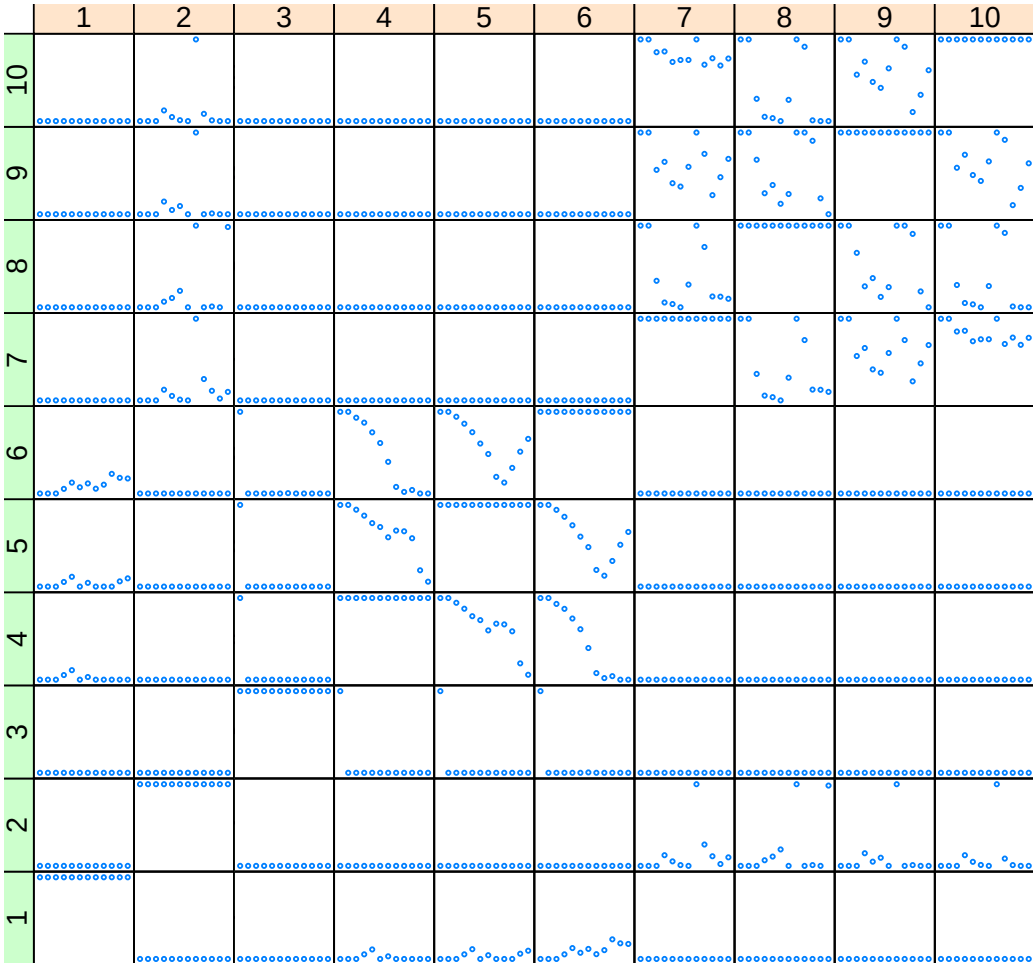

Figure 6.5: Plots of probabilities that each pair of situations are in the same stage for different values of t, for the case when $k = 0.5$, $\rho = 0.25$, $q = 0.05$ with $P(C_1 = \{v_1, v_2, \{v_3, \ldots, v_6\}, \{v_7, \ldots, v_{10}\}\}) = 1$, and situations $v_2, v_7, v_8, v_9, v_{10}$ caused to be in the same stage at $t = 8$

Chapter 7

Discussion

In this thesis I have shown that chain event graphs are not just an efficient way of storing the information contained in an event tree, but also a natural way to represent the information that is most easily elicited from a domain expert: the order in which events happen, the distributions of variables conditional on the process up to the point they are reached, and prior beliefs about the relative homogeneity or symmetry of different situations. This strength is exploited when the MAP CEG is discovered, as this can be used in a qualitative fashion to detect homogeneity between seemingly disparate situations, or when predictions need to be made, allowing flexible and robust specification of the system structure. The range of possible applications goes beyond the educational one, with forensic, biological and medical systems seeming particularly suitable with their asymmetric processes and complex independence structures.

One difficulty with model selection over CEGs is simply the expressiveness and hence relative size of the model space, which means that to be feasible for even larger problems one needs to add more contextual information to limit the size of the space. This is particularly the case if the underlying tree is allowed to em-

body different orders for when situations happen as described in the last paragraph. One possible method is to use search for a MAP BN as a coarse initialisation step and then, taking a CEG consistent with its conditional independences, refine the search using methods described here. In other contexts, it is worth remembering that to allow all possible combinations of florets into stages, as done here, would be implausible. When this is the case, the search algorithm can accommodate this information easily and therefore be carried out faster; for example, it might be decided that only situations the same distance away from the root node could be combined, which would make sense if the underlying even tree is drawn in a hierarchical manner, with the same system variables being represented in the same order along all root-to-leaf paths of the tree.

It was found in Chapter 6 that even in moderately-sized problems that weighted MAX-SAT can quickly become intractable, even with restrictions on the maximum size a stage can take, due to needing to calculate every stage score beforehand. AHC, on the other hand, while fast, might not explore the space as well as an algorithm to solve weighted MAX-SAT can. An algorithm that combines AHC and weighted MAX-SAT might thus be worth investigating. One possible approach to investigate, used successfully to search over partitions in [Liverani et al., 2010], is as follows:

1. Use AHC initially to reduce the number of stages to a manageable number

2. For each stage of the staging found with AHC, run the weighted MAX-SAT algorithm to find the optimal partition of that stage. Replace each stage with its optimal partition.

3. If nothing changed after the weighted MAX-SAT step, stop. Otherwise run AHC on the new partition, and repeat until the staging is stable.

This hybrid algorithm exploits the speed of AHC with the thoroughness of weighted MAX-SAT.

Another algorithm which has recently emerged for learning BNs uses INTEGER LINEAR PROGRAMMING [Cussens, 2010]. This involves restating the search for a MAP CEG as a problem in propositional logic just as with weighted MAX-SAT. The conditions for being to apply it here are

1. that the score of each CEG is a linear function of its stages, which is the case here;

2. that the constraints for a staging to be valid (i.e. that it be a partition of the situations) also be linear in terms of the stages chosen, whether as equalities or inequalities.

If this second condition holds — and this must be investigated — then an algorithm called an IP SOLVER can be used to solve the formulated integer programming problem.

There are a number of extensions to the theory in this thesis that look worthy of pursuit. One important modelling extension arises from uncertainty about the underlying event tree. With each different event tree of the same event space, different factorisations and conditional independence statements can be learnt from the data. A similar model search algorithm to the one described in this thesis is possible in this case after setting a prior distribution on the candidate event trees. In many potential applications it would be desirable to allow for multiple possible trees at any time point. Sometimes all that is required is the subclass of $\boldsymbol{T}$ — the general class of event trees — that consists of trees that are merely different partitions of the same fixed set of root-to-leaf path events. In that case, assuming that the same root-to-leaf path events on different trees have the same probability implies that the floret distributions on all trees can be characterised as Dirichlet by the methods

used here, with the parameters for each possible CEG characterised similarly. The method of assigning probabilities over the tree space, or how those probabilities change over time, would still need to be resolved. If a bigger subset of $\boldsymbol{T}$ is required due to uncertainty about the nature of the event space, then $P(\boldsymbol{x}_t \mid \boldsymbol{x}^{t-1})$ can still be calculated as outlined here but with the additional step of marginalising over the $T \in \boldsymbol{T}$ such that $P(T \mid \boldsymbol{x}^{t-1}) > 0$, assuming the number of such T is tractable.

In this thesis it was assumed that it was always known that the edges that were coloured the same in florets deemed to be in the same stage were those of equal value. Another way of enlarging the model space is therefore to allow for uncertainty in the function $\psi_u(v, v')$ which determines which edges are coloured identically. This would allow symmetries to emerge beyond simple conditional independence. One type of hypothesis this could capture is stability between values of different random variables. For example, consider the event tree of Figure 3.2. Stability would be described by colouring edges (v_1, v_3) and (v_2, v_6) identically, so that the probabilities $P(B = 0 \mid A = 0)$ and $P(B = 1 \mid A = 1)$ are equal. Stability is therefore a kind of independence, as $|B - A| \perp\!\!\!\perp A$. This example is called a NOISY OR GATE in computer science, and is another kind of structure that cannot be easily represented with the structure of a BN.

In the educational examples described in this thesis the assumption of stability would translate into believing that the probability of getting the same grade in two different modules is the same for all possible grades, i.e. that students who perform poorly in one module will continue to perform poorly in the next module with approximately the same proportion as that of students who do well in the second module after doing well in the first one.

It must be noted that the number of possible $\psi_u(v, v')$ for any pair of situations $v, v' \in S$ 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, like the one used in this thesis, could incorporate this feature too.

One other possible enhancement is to allow for more complex relationships between floret distributions. At the moment they must either be identical or independent. However, many applications have non-homogeneous samples. One approach which keeps the strengths of the CEG models is to analyse relevant sub-populations on separate trees, perhaps with dependence modelling of floret parameters of different sub-population trees but keeping the CEG framework within trees. Indeed, the dynamic CEG here can be a seen as an analysis of this sort, with different stagings for different cohorts and a specification of the relationship between adjacent years' parameters.

One aspect of the CEG that is worth noting is its ignoring of the time it takes for events to occur by modelling only which events occur. In applications like the educational one, where the time of the events is predetermined, 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 times 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. The incorporation of non-constant hazard functions into CEGs is thus worth investigating, perhaps through "transport times" as used with flow networks.

Finally, it appears that the static and dynamic CEGs could 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 as upstream nodes are unaffected. When other variables are continuous then analogous conjugate models can be defined which describe hierarchical clustering models, discretising continu-

ous variables intelligently. This is quite analogous to the relationship between the Dirichlet distribution and the Dirichlet process, and certainly merits further investigation.

Appendix A

Exam marks

Below I present the raw data used in Chapter 6.1, with grades for different yearly cohorts presented separately. For the dynamic analysis of Section 6.2, only the first two modules were used, while for the static case in Section 6.1 years were obviously ignored. A blank space means the grade is not available.

1994 cohort

ID	ST108grade	IB104grade	ST213grade	IB207grade
33	2	2		
34			3	2
46	1	1	2	1
61	3	2	2	1
75	3	2	3	2
78	1	2		
80	1	2	3	2
81	2	1	1	1
89	2	2	3	2

106	3	1	2	2
126	1	1	1	2
207	2	2	2	2
232	2	3		
233		2		
234				
248	2	3	2	3
260		2		
261	1	3		
287	1	1	2	1
315	1	2	2	1
403	1	1	2	1
413	1	2	2	1
420	2	2	3	2
421				
439			3	2
440	1	2		
443	1	2	2	2
448	2	2	3	3
456				
457	3	2	3	2
463	2	3	3	3
464			3	2
465	3	2		
477	1	2	2	1
485	2	2	3	2

486				
496	2	3		
497		2		
503	1	2	1	1
511	3	3	3	2
525	2	2	3	2
611	3	3		
676			3	3
677	2	2	2	2
678	2	2		
683				
684	1	2	3	2
700	2	2	3	1
701	2	1	3	3
702				
712	3	2	3	2
721				
833	3	1	3	2
880		2		
881	2	3	1	1
886	3	2	3	2
887				
890	2	1	1	1
893	2	2	2	3
897	1	1	2	2
911				

912	3	3		
946	1	2	2	2

1995 cohort

ID	ST108grade	IB104grade	ST213grade	IB207grade
17	1	1	2	1
23	2	2	3	1
25	3	2	3	3
49	1	1	3	1
50	2	2	1	1
51	2	2	3	1
60	1	1	2	1
74	2	2	3	2
112	1	2	2	2
212	1	2	2	1
221	1	1		
222	1	2	1	1
244	3	3	3	3
245		1		
270	3	3		
271		2		
282	2	2	3	1
291	1	1	1	2
310	1	1	3	1

319	1	1	2	1
401	2	2	2	1
402	2	1	2	1
416	2	2	2	1
436				
437	3	3	3	2
450	1	1	2	1
479				
480	2	2	3	2
495	2	2	2	3
510	1	1		
530	1	2	3	2
617	3	1	2	2
621			1	1
633	2	3	3	2
634				
635	3	2	3	2
699	2	3	2	3
711	2	2	2	2
714	2	2	3	2
730	3	1	2	3
734	1	1		
742	1	2	2	2
829	3	3	3	1
837	2	3	2	2
846	1	2	2	1

852	1	1	1	1
855	1	2	2	1
864	2	2	3	3
883	2	2	2	2
892	2	3	3	3
895	1	1	3	3
904	2	2	2	2
909	2	3	3	2
919	2	2	3	3
920	3	1	2	2
925	1	2	2	2
941	2	2	3	2
957	1	1	2	1

1996 cohort

ID	ST108grade	IB104grade	ST213grade	IB207grade
14	1	2	1	2
30	1	2	2	1
31				
42	1	1	2	2
48	1	1	1	2
66	1	2	1	1
77	2	2	3	3
104	2	1	2	2

194	1	2	3	3
195	1	1	2	2
220	1	1	2	1
236			2	1
241	1	2	3	1
250	1	1	2	2
252	1	1	1	1
256			2	2
262	1	2	3	2
269	1	2	2	1
303		3		
304	3	2	3	3
406				
407	2	2	2	2
408		1	3	3
411	1	1	1	1
412	2	2	2	1
419	2	1	3	1
434	1	2	2	1
435	1	2	1	1
460	1	1	2	1
492	2	2	3	2
498	1	1	2	1
500	1	2	1	2
512	1	1	1	1
615	1	2	2	2

627	1	1	1	2
651		1		
652	3	3		
653	1	2	1	1
659	1	2	2	2
665	1	1	1	1
670	1	2	1	2
674	1	2	2	2
680	1	2	2	2
719	2	2	3	3
739	1	2	2	1
741	2	2	3	2
832	1	2	3	2
845	1	2	1	2
848	2	2	3	2
858	1	1	1	2
903	1	1	2	2
917	2	1	2	2
924	1	1	1	1
943	3	1		3
962	2	2	2	2

1997 cohort

ID	ST108grade	IB104grade	ST213grade	IB207grade
6	2	2	3	3
10	1	2	1	3
13	2	2	3	2
15	2	1	2	3
22	2	2	2	3
41	1	1	1	2
68	1	1		
73	1	2	1	3
76	1	1	2	1
86	1	2	2	3
87	1	1	1	1
93	2	2	2	2
110	1	1	1	2
111	2	1	2	1
114	1	2	1	2
121				
210	1	2	1	2
216	2	2	3	3
219	1	2	1	3
246	1	1	1	1
273	1	2	3	3
277	3	2	2	2
298	2	1	3	3
299	2	2	2	2
318	3	2	2	3

446	3	1		
455	1	1	2	2
481	2	2	2	3
502	2	1	3	2
509	1	2	2	3
637	2	3		
643	2	2	2	2
648	2	3		
649			3	3
654	2	2	1	2
661	2	1		
662	2	2	3	2
664	1	1	1	2
682	2	2	3	3
690	2	2	2	1
691	1	1	1	1
718	1	2	3	2
722	1	2	2	2
744	1	1	2	1
746	1	1	3	3
825	2	2	1	3
836	3	2	2	3
847	1	1	1	1
850	2	1	3	2
857	1	2	1	2
865	1	2	3	3

871	3	3		
872				
901	2	3	3	3
902	2	2	3	2
928	2	2	2	1
929	1	2	2	3
936	1	1	2	2
949	2	1	3	3
955	1	2	2	2
960	2	1	3	2

1998 cohort

ID	ST108grade	IB104grade	ST213grade	IB207grade
4	2	1	3	2
8	3	3		
9		3		
21	1	2	2	3
35	1	2	2	2
36	2	2		
37		3		
47	2	2	3	3
57	2	1	2	2
62	3	2	3	3
65	2	1	3	3

96	1	3	3	3
97				
109	1	2	2	2
113	1	1	2	2
208	1	1	2	1
209	1	1	2	2
227	2	2	3	3
235	1	2	1	1
238	1	2	3	3
249	2	2	3	2
251	2	2		
275	2	2	3	3
280	3	1	3	2
296	2	2		
297			3	3
307	2	1	1	3
404	1	1	2	2
422	2	2	3	3
423	1	1	2	2
442	2	2	3	3
449	2	1	3	2
491	1	1	1	1
514	1	2	2	3
529	1	2	2	2
612	2	2		
613				

620	1	1	1	2
629	1	1	2	1
636	1	1	2	2
646	2	1	2	2
647	1	1	3	2
685	1	1	1	1
689	1	2	2	3
707	1	1	1	1
708	1	2	2	2
717	3	2	3	3
733	1	1	3	2
736	1	1	2	2
827	1	1	2	3
831	1	2	3	3
834	1	2	2	3
835	2	2	1	1
853	1	2	2	1
854	1	2	2	3
856	2	2	3	3
867	1	1	1	1
874	3	2	2	2
875		1		
888	2	2	3	3
891	2	1	3	2
910	1	2	2	3
914	2	2	3	3

915	1	1	2	1
930	1	1	1	1
953	1	2	3	2

1999 cohort

ID	ST108grade	IB104grade	ST213grade	IB207grade
24	2	3	3	3
55	1	1	1	1
70	2	2	3	3
175	2	1	3	2
198			3	3
199	2	1		
228	2	2	3	2
279	2	2	3	3
283	2	2		
284			3	3
335	2	2	3	2
357	2	2	3	2
358			3	2
360	3	2		
382			3	1
383	2	2		
417	1	2		
426	1	1		

427			3	2
472	3	2	3	2
478	1	1	2	2
489	2	2	3	2
493	2	2	2	2
548	3	3		
549			3	2
553	2	1	2	2
558	1	1	2	2
561	2	2	2	2
562	1	1	2	2
568	1	1	2	2
577	1	1	1	2
587			3	2
588	2	2		
591	1	1	2	2
609			3	2
610	3	2		
618	2	3	3	2
619				
655	1	2	3	3
671	2	1	3	1
675	2	1	2	2
681	1	2	3	2
697	1	2	3	1
710	2	1	3	3

723				
724	2	2	3	2
727			3	2
728	2	3		
737		2		
738	1	2	3	2
748	2	1	3	2
779	2	2	3	2
791	2	1	2	1
804	1	2	3	2
811			3	3
812	2	2		
816	3	2	3	2
840	3	2	3	3
841		3		
842	2	3	3	3
866	1	2	2	2
927	1	1	3	3
969	2	2	3	2
970				
978	1	1	2	1
986			1	1
987	1	2	3	2
994	1	1	2	1
999	1	1	1	1
1000	2	2	3	3

1006	1	1	3	2
1019	1	2	2	2
1029	1	1	2	2

2000 cohort

ID	ST108grade	IB104grade	ST213grade	IB207grade
119	1	1	1	1
125	1	1	2	2
135	1	1	1	1
138	1	3	3	3
151	1	2	1	2
154	2	3	3	3
156	1	3	3	3
167	1	1	1	1
170	1	2	1	2
186	2	3	3	3
188	2	3		
196	1	2	2	2
197	1	2	1	3
223	2	3	3	3
224	1	2	1	1
225	1	3	2	3
231	1	2	2	2
288	1	2	1	2

308	2	2	3	2
331	2	2	2	3
332	1	2	3	2
336	1	3	2	2
338	2	3	2	3
343	1	2	1	2
363	1	1	2	2
379	1	3	2	2
380	2	3	3	3
398	1	3	3	3
400	1	2	2	3
430	1	2	2	2
444	1	2	2	2
447	1	2	2	2
452	1	2	2	3
467	1	2	1	2
501	2	2	2	2
504	1	2	2	3
516	1	3	2	2
539	1	2	2	3
551	1	2	3	2
557	1	2	1	1
571	1	2	2	2
623	1	2	2	2
645	1	2	1	1
660	1	2	1	1

668	1	2	3	2
687	1	1	1	2
694	1	2	1	2
715	1	2	2	2
726	1	3	3	3
750	1	3	3	2
753	1	3	3	3
765	1	2	3	3
771	2	1	3	2
781	1	2	2	2
802	1	2	2	2
803	1	2	2	1
808	1	2	2	1
824	1	2	1	2
839	1	2	1	2
868	1	1	1	2
938	3	3	3	3
942	1	1	3	3
945	2	3	3	3
952	1	2	1	1
966	1	1	2	2
972	1	2	2	2
979	1	2	1	2
988	1	3	2	3
991	1	2	1	1
1027	1	2	2	2

1036	1	2	1	2

2001 cohort

ID	ST108grade	IB104grade	ST213grade	IB207grade
12	1	2	2	2
26	1	1	3	1
32	1	1	1	2
38	1	1	1	1
53	2	2	2	3
79	2	2	3	3
92	1	1	2	1
115	1	1	2	2
117	1	1	2	1
137	2	2	2	2
144	1	2	2	1
147	1	1	1	2
158	1	1	2	1
162	1	2	3	3
164	2	2	3	3
173	1	2	2	2
176	1	2	3	3
180	1	2	3	3
213	1	2	3	3
230	1	2	2	2

254	1	2	1	2
257	2	2	2	2
263	2	2	2	2
276	1	2	1	2
321	1	2	1	3
322	1	2	2	3
324	1	1	2	2
341	1	2	3	2
362	2	1	1	2
366	1	1	3	2
370	3	2	3	3
384	2	2	3	2
385	1	2	2	2
388			3	3
389	3	3		
410	1	2	2	2
414	2	1	3	3
433	2	2	2	2
445	1	2	3	2
474	1	2	2	3
484	2	2	2	1
507	1	2	1	2
517	2	1	1	2
519	1	3	1	1
520	2	2	3	3
521				

535	2	2	2	2
544	1	2	3	2
564	3	2	2	3
583	1	1	2	2
584	1	1	1	2
593	1	2	3	3
607	3			
608				
622	1	1	3	3
624	1	2	2	1
640	2	2	3	2
669	1	1	1	2
679	1	1	1	2
703	1	2	1	3
706	1	1	1	3
740	1	1	2	1
745	1	1	1	1
747	2	2	3	3
761	2	3	3	3
764	1	1	1	1
774	3	3		
775			3	3
789	1	1	2	1
793	1	1	3	1
795	2	3	2	3
799	1	1	1	2

800	2	2	3	2
805	1	2	2	2
810	1	2	2	2
813	1	2	3	2
843	2	1	1	2
863	1	2	2	2
873	2	1	3	3
884	1	2	2	2
894	2	1	1	1
905	1	2	2	2
921	1	1	3	3
922	1	2	3	2
947	1	2	3	2
975	1	2	3	3
984	1	2	2	2

2002 cohort

ID	ST108grade	IB104grade	ST213grade	IB207grade
2	1	1	3	2
11	1	2	3	3
28			3	3
45	1	2	1	2
54	1	1	1	2
56	2	2	2	3

82	1	2	3	2
84	1	1	2	2
94	1	1	1	2
98	1	2	2	3
100				
101	1	3	3	3
118	1	2	1	2
120	1	1	3	2
152	1	2	3	3
174	1	2		
179	1	1	3	2
189	1	1	2	3
191	1	1	1	2
202	1	2	2	3
204	1	1	3	2
226	1	1	3	2
237	1	2	1	2
239	1	2	2	2
247	1	1	3	1
278	1	2	2	2
281	1	1	2	3
286	1	1	2	2
292			3	
293	3	3	3	3
294	1	3	3	3
295	1	1	1	1

329				
330	2	2	3	2
349	1	2	2	1
361	1	1	1	1
374	1	1	2	3
386	1	1	1	1
387	1	1	1	1
390	1	1	2	1
394	1	3	2	2
424	1	2		
425			3	3
428	1	2	2	3
451	1	3	3	3
459	1	1	1	1
468	1	2	2	3
473	1	2	2	2
523	2	3	3	
524			3	3
528	2	3	3	3
582	1	2	3	3
586	1	2	2	2
592	1	1	2	2
594	1	1	1	2
601				
602	3	3	3	3
631	1	1	2	2

642	2	2	2	1
644	1	2	2	2
672	1	2	2	3
693	1	1	1	2
704	1	2	2	2
709	1	1	1	1
757	1	1	1	2
758				
759	1	1	2	2
769	1	1	2	2
784			3	
785	3	2	3	3
798	1	2	1	2
818	1	2	3	2
822	1	1	2	2
826	1	2	1	2
828	1	2	2	3
859	1	1	1	2
878	1	2	3	2
889	1	1	1	1
926	1	2	2	3
950	1	1	3	2
959	1	1	3	1
964	1	1	1	2
976	1	2	3	3
982	2	2	3	2

990	1	2	3	3
993	1	1	1	1
995				
996	2	2		
997	1	2	1	2
998	1	2	3	1
1002	1	2	2	2
1005	2	2	3	2
1022	1	1	1	2
1024	1	2	2	2
1028	1	2	2	1
1030	1	2	2	2
1031	1	2	3	2
1032	1	2	3	1

2003 cohort

ID	ST108grade	IB104grade	ST213grade	IB207grade
5	1	1	2	2
18	2	2	2	2
83	1	1	3	2
90	1	2	3	3
91	2	2	3	3
102	2	2	3	2
103				

123	2	2	3	3
124				
131	1	1	2	2
133	2	1	3	3
139	1	1	2	3
140	2	1	3	1
149	2	3	3	3
150	1	2	3	3
155	1	1	3	3
161	1	1	2	3
163	1	1	3	3
165	1	1	1	3
166	2	2	3	2
181	2	1	2	2
184	1	1	1	1
203	1	2	3	2
229	1	2	2	1
240	1	1	3	3
243	1	1	1	2
258	2	1	3	3
259	1	1	3	2
268	1	3	2	3
301				
302	2	2	3	2
311	1	1	1	2
316	3	3	3	3

317	2		3	
327	3	3		
328	2			
339	3	3	3	3
340	3	2		
342	2	3	3	3
365	3	3	3	
367	1	1	1	2
399	1	2	2	2
415	2	1	3	3
418	1	2	3	3
429	2	2	3	3
431			2	2
453	1	1	2	1
461	1	1	1	1
462	1	1	3	1
476	2	2	2	3
482	1	2	3	3
494	1	3	2	2
513	2	2	3	3
515	1	2	3	2
518	1	2	1	1
526	1	1	2	2
531	1	2	2	3
534	1	1	1	2
536	3	3	3	3

550	2	3	2	3
554	2	2		
559	1	1	2	3
563	1	1	2	2
565	1	2	3	3
572	2	2	3	3
580	1	2	3	3
585	1	1	1	2
590	1	1	2	1
595	1	1	1	1
597	1	1	1	1
600	1	2	3	3
604	1	1	3	2
605	2	2	3	3
616	1	1	2	2
641	1	2	2	2
658	1	1	2	1
667	3	1	3	3
713	2	1	2	2
720	2	3	3	3
731		3		
732	2	3	3	3
754	1	1	3	3
756	1	3		
760	2	2	3	2
766				

767	2	2	3	3
776			3	3
777	3	3		
782	1	2		
783			3	3
801	2	1	2	2
807	2	3	3	3
814	3	2	1	3
815	1	1	2	2
820	1	1	2	3
821	1	1	1	3
849	3	2	3	3
869	2			
870	3	3		
933	1	3	3	3
935	1	1	1	2
939	2	1	3	3
940	2	1	2	3
956	2	1	2	3
963	1	1	1	2
973	1	1	2	3
974	1	1	3	3
989	1	2	2	3
1003	1	1	3	2
1004	3	3	2	3
1007	1	2	2	3

1013	3	3	3	3
1015	1	1	3	2
1035	2	2	3	3

2004 cohort

ID	ST108grade	IB104grade	ST213grade	IB207grade
1	1	1	3	2
3	1	2	3	2
19	1	1	3	1
20	2	2	3	2
39				
40	2	2	3	2
43	1	2	2	2
44	1	3	1	1
58			3	3
59	2	3		
69	1	3	3	2
88	1	2	2	1
105	1	2	3	2
108	2	2	3	2
122	2	2	3	2
127	1	2	3	2
129	3	2	3	2
132	1	1	2	1

136	1	1	3	2
142	1	2	3	2
148			3	2
157	1	2		
160	1	2		2
168	1	3	3	3
169	1	1	3	1
177	1	1	3	3
178				
185	2	2	3	2
190	3	3		
193	1	1	2	2
200	1	1	2	1
201	1	3	3	2
206	2	2	2	2
211	1	1	1	2
217	1	3	2	1
242	2	3	3	2
255	1	2	2	1
264	1	1	3	2
266	3			
267				
272	1	2	3	2
285	1	3	3	2
306	1	2	2	2
309	1	3	3	1

312				
313	3			
320	1	2	3	2
323	1	2	2	2
325	1	1		
326			3	2
333	2	3	3	2
334				
344	1	2	3	2
345	1	1		
347	1	2	3	2
350	2	3	3	2
351	1	3	2	2
352	3	3	2	1
355	1	2	3	2
356	1	2	3	1
359	1	2	3	2
368				
369	3			
372	1	2	3	1
376			3	3
377	2	2		
381	1	1	2	2
392	1	1	1	1
393	1	2	3	1
395	2	3	3	2

397	1	2	2	2
438	2	1	3	2
441	1	1	3	1
458	3	3	3	3
469	1	1	2	2
470				
471	2	1	3	2
487	1	1	1	1
488	3	3	3	2
490	1	3	3	2
506	2	3	3	2
532	1	3	3	2
533	1	2	3	3
540	1	2	3	2
541	1	1	1	2
543	1	2	3	1
552	1	1	1	2
556	1	1	1	1
569	1	1	1	1
576	1	2	1	1
581	1	2	3	2
589	1	1	2	2
596	1	2	3	3
598	2		3	2
599	3	2		
614	1	1	3	2

625	1	2	1	1
626	1	2	3	2
630	1	3	3	2
638	1	2	3	1
656	1	1	2	1
657	1	1	2	2
666	1	2	2	2
692	1	1	2	2
698	1	3	3	2
705	1	2	3	2
735	1	3	2	2
763	1	2	3	2
773	1	2	2	2
778	1	2	2	3
786	1	2	1	1
796	1	2	2	2
806	1	1	2	1
817	1	2	3	2
860	1	1	3	2
861	1	1	1	2
862	1	1	2	1
877	1	2	3	3
882	2	3		
898	1	2	3	2
899	1	3	3	2
906	1	3	3	2

907			3	2
908	2	1		
913	1	2	2	2
918	1	2	2	2
932	1	1	1	1
934	2	2	3	2
951	1	1	1	1
965	2	2	3	2
977	1	1	3	2
981	1	2	1	2
983	1	1	2	1
992	1	2	3	3
1001	1	1	1	1
1009	2	3	3	2
1010	2	1	3	1
1016	2	3		
1017				
1021	1	1	1	1
1034	1	2	3	3

2005 cohort

ID	ST108grade	IB104grade	ST213grade	IB207grade
7	1	2		
16	1	2		

27	2	2
29	1	1
52	1	3
63	2	1
64	2	1
67	1	1
71	1	1
72	1	2
85	1	2
95	1	1
99	3	2
107	2	2
116	1	1
128	1	1
130	2	2
134	1	2
141	2	2
143	1	2
145	3	2
146	1	1
153	1	1
159	1	1
171	1	1
172	1	1
182	3	1
183	1	2

187	2	2
192	1	1
205	1	1
215	2	3
218	2	2
265	1	2
274	1	2
289	2	1
290	1	2
300	1	1
305	1	1
314	3	
337	1	1
346	2	1
348	1	1
353	1	1
354	1	1
364	1	1
371	1	1
373	1	1
375	1	1
378	1	1
391	1	2
396	2	1
405		1
409	2	2

432	1	2
454	1	2
466	1	1
475	1	2
483	2	1
499	1	2
505	2	2
508	1	1
522	1	1
527	2	3
537	1	1
538	1	2
542	1	2
545	1	1
546	1	1
547	1	1
555	1	2
560	1	1
566	1	2
567	1	1
570	1	1
573	2	3
574	1	1
575	1	1
578	1	2
579	2	2

603	1	1
606	2	1
628	2	1
632	1	1
639	3	3
650	1	2
673	1	2
686	1	1
695	1	1
696	2	3
716	1	1
725	1	1
729	2	1
743	2	3
749	2	2
751	2	
752	1	2
755	1	1
762	1	1
768	1	3
772	1	1
780	3	3
787	1	1
788	1	2
790	1	1
792	1	3

794	1	1
797	2	1
809	1	1
819	1	1
823	2	1
830	1	1
844	1	2
851	1	1
876	1	2
885	3	1
896	2	3
900	1	1
916	1	1
923		
931	1	1
937	1	1
948	1	1
954	3	2
958	1	1
961	1	1
967	2	2
968	1	1
971	2	3
980	2	1
985	1	1
1008	1	2

1011	2	1
1012	1	1
1014	1	3
1018	1	1
1020	1	1
1023	1	1
1025	1	1
1026	1	1
1033	1	1

No year available

ID	ST108grade	IB104grade	ST213grade	IB207grade
214				
253			3	1
663			3	3
688			1	1
770			2	3
838			3	3
879				
944			3	3

Bibliography

Ali, S. M. and Silvey, S. D. (1966). A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(1):131–142.

Arjas, E. (1989). Survival models and martingale dynamics (with discussion and reply). *Scandinavian Journal of Statistics*, 16(3):177–225.

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. Wiley Series in Probability and Statistics. Wiley, Chichester, England.

Booth, J. G., Casella, G., and Hobert, J. P. (2008). Clustering using objective functions and stochastic search. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):119–139.

Borgelt, C. and Kruse, R. (2002). *Graphical Models: Methods for Data Analysis and Mining*. Wiley, Chichester.

Boutilier, C., Friedman, N., Goldszmidt, M., and Koller, D. (1996). Context-Specific independence in Bayesian networks. In Horvitz, E. and Jensen, F. V., editors, *Proceedings of the Twelfth Conference on Uncertainty in Artificial Intelligence*, pages 115–123, Reed College, Portland, Oregon, USA. Morgan Kaufmann.

Buntine, W. (1991). Theory refinement on Bayesian networks. In *Proceedings of the Seventh Conference on Uncertainty in Artificial Intelligence*, pages 52–60.

Cartwright, N. (1994). *Nature's capacities and their measurement*. Clarendon Press, Oxford.

Cartwright, N. (2007). *Hunting Causes and Using Them: Approaches in Philosophy and Economics*. Cambridge University Press.

Castelo, R. (2002). *The discrete acyclic digraph Markov model in data mining*. PhD thesis, Faculteit Wiskunde en Informatica, Universiteit Utrecht.

Castillo, E., Gutierrez, J. M., and Hadi, A. S. (1997). *Expert Systems and Probabilistic Network Models*. Monographs in Computer Science. Springer, illustrated edition.

Chickering, D. M. (1995). A transformational characterization of equivalent Bayesian network structures. In Hanks, S. and Besnard, P., editors, *Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence*, page 87–98. Morgan Kaufmann.

Cooper, G. and Yoo, C. (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)*, pages 116–12, San Francisco, CA. Morgan Kaufmann.

Cooper, G. F. and Herskovits, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9(4):309–347.

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*. Springer.

Crowley, E. M. (1997). Product partition models for normal means. *Journal of the American Statistical Association*, 92(437):192–198.

Cussens, J. (2008). Bayesian network learning by compiling to weighted MAX-SAT. In McAllester, D. A. and Myllymäki, P., editors, *Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence*, pages 105–112, Helsinki, Finland. AUAI Press.

Cussens, J. (2010). Maximum likelihood pedigree reconstruction using integer programming. In *Proceedings of the Workshop on Constraint Based Methods for Bioinformatics*, pages 9–19, Edinburgh.

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.

Dawid, A. P. (2002). Influence diagrams for causal modelling and inference. *International Statistical Review*, 70(2):161–189.

Dawid, A. P. (2010). Beware of the DAG! In Guyon, I., Janzing, D., and Schöolkopf, B., editors, *Proceedings of the NIPS 2008 Workshop on Causality*, volume 6 of *Journal of Machine Learning Research Workshop and Conference Proceedings*, pages 59–86.

Dawid, A. P. and Lauritzen, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *The Annals of Statistics*, 21(3):1272–1317.

de Finetti, B. (1974). *Theory of Probability. a Critical Introductory Treatment*. Wiley series in probability and mathematical statistics. Wiley, London.

Dean, T. and Kanazawa, K. (1988). Probabilistic temporal reasoning. In *Proc. AAAI-88*, pages 524–528. AAAI.

Dean, T. and Kanazawa, K. (1989). A model for reasoning about persistence and causation. *Computational Intelligence*, 5(2):142–150.

Denison, D. G. T., Holmes, C. C., Mallick, B. K., and Smith, A. F. M. (2002). *Bayesian Methods for Nonlinear Classification and Regression*. Wiley Series in Probability and Statistics. Wiley.

Didelez, V. (2008). Graphical models for marked point processes based on local independence. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):245–264.

Durbin, J. and Koopman, S. J. (2000). Time series analysis of non-Gaussian observations based on state space models from both classical and Bayesian perspectives. *Journal of the Royal Statistical Society: Series B (Methodological)*, 62(1):3–56.

Edwards, A. W. F. (1982). Pascal and the problem of points. *International Statistical Review / Revue Internationale de Statistique*, 50(3):259–266.

Eichler, M. and Didelez, V. (2007). Causal reasoning in graphical time series models. In *Proceedings of the 23rd Annual Conference on Uncertainty in Artificial Intelligence*. Morgan Kaufmann Publishers.

Einstein, A. (1934). On the method of theoretical physics. *Philosophy of Science*, 1(2):163–169.

Figueroa-Quiroz, L. J. (2003). *Bayesian forecasting and intervention in dynamic flow systems*. PhD thesis, Department of Statistics University of Warwick.

Frühwirth-Schnatter, S. (2006). *Finite mixture and Markov switching models*. Springer, New York.

Geiger, D. and Heckerman, D. (1996). Knowledge representation and inference in similarity networks and Bayesian multinets. *Artificial Intelligence*, 82(1-2):45–74.

Geiger, D. and Heckerman, D. (1997). A characterization of the Dirichlet distribution through global and local parameter independence. *The Annals of Statistics*, 25(3):1344–1369.

Geyer, C. J. and Thompson, E. A. (1995). Annealing Markov chain monte carlo with applications to ancestral inference. *Journal of the American Statistical Association*, 90(431):909–920.

Gibbs, J. W. (1902). *Elementary principles in statistical mechanics: developed with especial reference to the rational foundations of thermodynamics*. Charles Scribner's Sons, New York.

Glymour, C. and Cooper, G. F. (1999). *Computation, Causation, and Discovery*. AAAI Press, illustrated edition edition.

Gottard, A. (2007). On the inclusion of bivariate marked point processes in graphical models. *Metrika*, 66(3):269–287.

Hansen, P. and Jaumard, B. (1990). Algorithms for the maximum satisfiability problem. *Computing*, 44:279–303.

Harrison, P. J. and Stevens, C. F. (1976). Bayesian forecasting. *Journal of the Royal Statistical Society. Series B (Methodological)*, 38(3):205–247.

Heard, N. A., Holmes, C. C., and Stephens, D. A. (2006). A quantitative study of gene regulation involved in the immune response of anopheline mosquitoes: An

application of Bayesian hierarchical clustering of curves. *Journal of the American Statistical Association*, 101(473):18–29.

Heckerman, D. (1995). A Bayesian approach to learning causal networks. In *Proceedings of the Proceedings of the Eleventh Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-95)*, pages 285–295, San Francisco, CA. Morgan Kaufmann.

Heckerman, D. (1999). A tutorial on learning with Bayesian networks. In Jordan, M. I., editor, *Learning in Graphical Models*, pages 301–354. MIT Press.

Heckerman, D. and Geiger, D. (1995). Likelihoods and parameter priors for Bayesian networks. Technical report MSRTR-95-54, Microsoft.

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

Hocking, R. R. (1976). A Biometrics Invited Paper. The analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49.

Hoeting, J. A., Madigan, D., Raftery, A. E., and Volinsky, C. T. (1999). Bayesian model averaging: A tutorial. *Statistical Science*, 14(4):382–401.

Humphreys, P. and Freedman, D. (1996). Review: The grand leap. *The British Journal for the Philosophy of Science*, 47(1):113–123.

Ibrahim, J. G. and Chen, M. (2000). Power prior distributions for regression models. *Statistical Science*, 15(1):46–60.

Jaeger, M. (2004). Probabilistic decision graphs-combining verification and AI tech-

niques for probabilistic inference. *International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems*, 12(SUPPLEMENT):19–42.

Kalman, R. E. (1960). A new approach to linear filtering and prediction problems. *Journal of Basic Engineering*, 82(1):35–45.

Koller, D. and Lerner, U. (2000). Sampling in factored dynamic systems. In Doucet, A., de Freitas, J. F. G., and Gordon, N., editors, *Sequential Monte Carlo Methods In Practice*. Springer-Verlag.

Kolmogorov, A. N. (1963). On the approximation of distributions of sums of independent summands by infinitely divisible distributions. *Sankhyā: The Indian Journal of Statistics, Series A*, 25(2):159–174.

Kotz, S., Balakrishnan, N., and Johnson, N. L. (2000). *Continuous Multivariate Distributions*. Wiley series in probability and statistics. Applied probability and statistics. Wiley, New York, 2nd edition.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

Lau, J. W. and Green, P. J. (2007). Bayesian Model-Based clustering procedures. *Journal of Computational and Graphical Statistics*, 16(3):526–558.

Lauritzen, S. L. (1996). *Graphical Models (Oxford Statistical Science Series)*. Oxford University Press, USA.

Liverani, S., Cussens, J., and Smith, J. (2010). Searching a multivariate partition space using MAX-SAT. In Masulli, F., Peterson, L., and Tagliaferri, R., editors, *Computational Intelligence Methods for Bioinformatics and Biostatistics*, volume 6160 of *Lecture Notes in Computer Science*, pages 240–253. Springer Berlin / Heidelberg.

Lohr, S. (2009). For today's graduate, just one word: Statistics. *The New York Times*.

Madigan, D. and Raftery, A. E. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, 89(428):1535–1546.

McCullagh, P. and Yang, J. (2006). Stochastic classification models. In *Proceedings of the International Congress of Mathematicians*, volume III, page 669–686.

Meilă, M. (2007). Comparing clusterings–an information based distance. *Journal of Multivariate Analysis*, 98(5):873–895.

Neapolitan, R. and Jiang, X. (2006). A tutorial on learning causal influence. In Holmes, D. and Jain, L., editors, *Innovations in Machine Learning*, volume 194 of *Studies in Fuzziness and Soft Computing*, pages 29–71. Springer Berlin / Heidelberg.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Pearl, J. (2000a). Causal inference without counterfactuals: Comment. *Journal of the American Statistical Association*, 95(450):428–431.

Pearl, J. (2000b). *Causality*. Cambridge University Press.

Pearl, J. and Paz, A. (1986). Graphoids: Graph-Based logic for reasoning about relevance relations or when would x tell you more about y if you already know z? In du Boulay, J. B. H., editor, *ECAI*, pages 357–363.

Petris, G., Petrone, S., and Campagnoli, P. (2009). *Dynamic Linear Models with R*. Springer New York, New York, NY.

Poole, D. and Zhang, N. L. (2003). Exploiting contextual independence in probabilistic inference. *J. Artificial Intelligence Res.*, 18:263–313.

Queen, C. M. and Albers, C. J. (2009). 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 Smith, J. Q. (1993). Multiregression dynamic models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 55(4):849–870.

Queen, C. M., Smith, J. Q., and James, D. M. (1994). Bayesian forecasts in markets with overlapping structures. *International Journal of Forecasting*, 10(2):209–233.

Raftery, A. E., Kárný, M., and Ettler, P. (2010). Online prediction under model uncertainty via dynamic model averaging: Application to a cold rolling mill. *Technometrics*, 52(1):52–66.

Rauber, T., Braun, T., and Berns, K. (2008). Probabilistic distance measures of the Dirichlet and Beta distributions. *Pattern Recognition*, 41(2):637–645.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *Journal of the Royal Statistical Society. Series B (Methodological)*, 59(4):731–792.

Rigat, F. and Smith, J. Q. (2009). Semi-parametric dynamic time series modelling with applications to detecting neural dynamics. *The Annals of Applied Statistics*, 3(4):1776–1804.

Shachter, R. D. (1986). Evaluating influence diagrams. *Operations Research*, 34(6):871–882.

Shafer, G. (1996). *The Art of Causal Conjecture*. Artificial Intelligence. The MIT Press.

Silander, T., Kontkanen, P., and Myllymäki, P. (2007). On sensitivity of the MAP Bayesian network structure to the equivalent sample size parameter. In Parr, R. and van der Gaag, L., editors, *Proceedings of the The 23rd Conference on Uncertainty in Artificial Intelligence*, pages 360–367. AUAI Press.

Smith, J. E., Holtzman, S., and Matheson, J. E. (1993). Structuring conditional relationships in influence diagrams. *Operations Research*, 41(2):280–297.

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). A comparison of the characteristics of some Bayesian forecasting models. *International Statistical Review / Revue Internationale de Statistique*, 60(1):75–87.

Smith, J. Q. (2010). *Bayesian Decision Analysis: Principles and Practice*. Cambridge University Press.

Smith, J. Q. and Anderson, P. E. (2008). Conditional independence and chain event graphs. *Artificial Intelligence*, 172(1):42–68.

Smith, J. Q. and Daneshkhah, A. (2010). 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 Figueroa, L. J. (2007). A causal algebra for dynamic flow networks. In *Advances in Probabilistic Graphical Models*, volume 214 of *Studies in Fuzziness and Soft Computing*, pages 39–54. Springer Berlin / Heidelberg.

Smith, J. Q. and Rigat, F. (2008). Isoseparation and robustness in finite parameter Bayesian inference. CRiSM 07-22, University of Warwick, Coventry.

Spiegelhalter, D. J. and Lauritzen, S. L. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks*, 20(5):579–605.

Spirtes, P., Glymour, C. N., and Scheines, R. (2001). *Causation, Prediction, and Search*. MIT Press, 2nd edition.

Stanley, R. (1997). *Enumerative combinatorics*. Cambridge University Press, Cambridge ;;New York.

Steck, H. (2008). Learning the Bayesian network structure: Dirichlet prior vs data. In McAllester, D. A. and Myllymäki, P., editors, *Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence*, pages 511–518. AUAI Press.

Steck, H. and Jaakkola, T. (2003). On the Dirichlet prior and Bayesian regularization. In Becker, S., Thrun, S., and Obermayer, K., editors, *Advances in Neural Information Processing Systems 15*, pages 697–704, Vancouver, British Columbia, Canada. MIT Press.

Thwaites, P., Smith, J. Q., and Riccomagno, E. (2010). Causal analysis with chain event graphs. *Artificial Intelligence*, 174(12-13):889–909.

Thwaites, P. A. (2008). *Chain Event Graphs: Theory and Application*. PhD thesis, University of Warwick.

Thwaites, P. A., Freeman, G., and Smith, J. Q. (2009). Chain event graph map model selection. In Dietz, J. L. G., editor, *Proceedings of the International Conference on Knowledge Engineering and Ontology Development*, pages 392–395, Funchal, Madeira, Portugal. INSTICC Press.

Thwaites, P. E. and Smith, J. Q. (2006). Evaluating causal effects using chain event graphs. In *Proceedings of the 3rd European Workshop on Probabilistic Graphical Models*, Prague.

Tompkins, D. A. D. and Hoos, H. H. (2005). UBCSAT: an implementation and experimentation environment for SLS algorithms for SAT and MAX-SAT. In Hoos, H. H. and Mitchell, D. G., editors, *Theory and Applications of Satisfiability Testing*, volume 3542 of *Lecture Notes in Computer Science*, pages 306–320. Springer Berlin / Heidelberg.

Verma, T. and Pearl, J. (1988). Causal networks: semantics and expressiveness. In Shachter, R. D., Levitt, T. S., Kanal, L. N., and Lemmer, J. F., editors, *UAI '88: Proceedings of the Fourth Annual Conference on Uncertainty in Artificial Intelligence*, pages 69–78. North-Holland.

Verma, T. and Pearl, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the Sixth Annual Conference on Uncertainty in Artificial Intelligence*, page 270.

Weatherburn, C. E. (1949). *A first course in mathematical statistics*. CUP, 2nd edition.

West, D. B. (2001). *Introduction to Graph Theory*. Pearson Education Asia Limited and China Machine Press, China.

West, M. and Harrison, J. (1989). Subjective intervention in formal models. *Journal of Forecasting*, 8(1):33–53.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer Series in Statistics. Springer-Verlag, second edition. Published: Hardcover.

Whitlock, M. E. and Queen, C. M. (2000). Modelling a traffic network with missing data. *Journal of Forecasting*, 19(7):561–574.

Wilks, S. S. (1962). *Mathematical Statistics*. Wiley, New York.