Robustness in Bayesian Networks

by

Sophia Katie Wright

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Department of Statistics

February 2019

Contents

List of Tables

List of Figures

Acknowledgments

To my supervisor Jim: thank you for introducing me to a topic which has been both fascinating and rewarding. Your endless enthusiasm and motivation has been an inspiration, and without it my research would not be the quality that it is. Throughout my entire PhD you have immersed me in practical Bayesian network (BN) modelling experiences. Thank you for your patience and for granting me the generosity of your time during our heroic meetings.

To my internal panellists Professor Simon French and Dr Simon Spencer: thank you for the invaluable feedback and recommendations. Your fresh perspectives provided a wealth of new ideas for which I am very grateful.

To the OxWaSP programme: to all those in my cohort and who taught along the way, I enjoyed the lively debate, conversations and friendships which enriched my academic life. Thanks also to the EPSRC grant EP/L016710/1 without which this research would not have been possible.

Last, but by no means least to my family: words can never express the depth of my gratitude for all your unconditional support and confidence in me. Thank you all for your words of encouragement and advice during this academic venture and my life in general.

Declarations

I declare that the material contained within this thesis is my own work, unless otherwise stated to the contrary. This work has not been submitted for any other degree; however, the following content has currently been submitted for publication:

I am co-author of the book chapter ‘Eliciting probabilistic judgements for Integrating Decision Support Systems’ in the Springer book ‘Elicitation: The science and art of structuring judgement’, henceforth cited in this thesis as Barons et al. [2018b]. This seminal book on elicitation is split into two sections with the first half dedicated to theoretical material and the second half intended to be individual chapters on applications, which is where our work appears. Our case study details the Food Security Integrated Decision Support System (IDSS) which is also the main application of this thesis; see Section 3.1 for an introductory overview. Our team developed new material which we included in the chapter; such as the link between natural language and probability distributions and a brief discussion about loops relating to short versus long dynamic arcs in Bayesian networks. My interest and focus in this collaboration was on some of the earlier material and the example, various aspects of which you will see elaborated further in this thesis.

I am a co-author on the paper ‘Assessment of the response of pollinator abundance to environmental pressures using structured expert elicitation’, henceforth cited as Barons et al. [2018a], which has been accepted for publication in the Journal of Apicultural Research. This paper documents the pollinator abundance workshop hosted by Dr. Martine Barons and Professor Jim Smith at the University of Warwick, detailed in Section 2.2.2 of this thesis. I was an active member of the team and recorded proceedings for evidence. The final results published in this paper have been used as an ongoing example in this thesis as documented in Section 3.1.3.

The new material presented in this thesis comes from two publications which are co-authored with my supervisor Professor Jim Q. Smith. The first of these papers is ‘Bayesian Networks, Total Variation and Robustness’ which we have submitted to the Journal of Bayesian Analysis [see Wright and Smith, 2018]. This paper introduces our arguments for using total variation distance to measure robustness in Bayesian networks, rather than the current common measures of Kullback-Leibler or Chan-Darwiche distances. We define new measures of total variation for use in tables and matrices, which have been presented in Section 4.3, along with bounds which help us to cap the effect of perturbations within a Bayesian network. These results are expanded further in Section 5 to determine whether simplifications commonly employed in practice are valid. The versatility of this framework enables us to use the same technology to investigate both structural changes as well as probabilistic changes. The second paper ‘Robustness, Total Variation and Dynamic Bayesian Networks’ summarises the Bayesian network results and extends the methodology to cover dynamic models too. This work constitutes Section 6 of this thesis and introduces methodology to rewrite a dynamic Bayesian network as a finite time inhomogeneous Markov Chain, enabling us to utilise the diverse Markov Chain literature. We have a draft for this paper and aim to submit in the immediate future.

Abstract

This thesis explores the robustness of large discrete Bayesian networks (BNs) when applied in decision support systems which have a pre-specified subset of target variables. We develop new methodology, underpinned by the total variation distance, to determine whether simplifications which are currently employed in the practical implementation of such systems are theoretically valid. This versatile framework enables us to study the effects of misspecification within a Bayesian network (BN), and also extend the methodology to quantify temporal effects within Dynamic BNs. Unlike current robustness analyses, our new technology can be applied throughout the construction of the BN model; enabling us to create tailored, bespoke models. For illustrative purposes we shall be applying our work to the field of Food Security and a demonstrative ecological network.

Abbreviations

2TSDBN	2 Time-Slice Dynamic Bayesian Network
ADMLC	Atmospheric Dispersion Modelling Liaison Committee (UK)
BN	Bayesian Network
COBR	Cabinet Office Briefing Room
CD	Chan-Darwiche (distance)
CPI	Consumer Price Index
CPT	Conditional Probability Table
DAG	Directed Acyclic Graph
DBN	Dynamic Bayesian Network
DSS	Decision Support System
EFSA	European Food Safety Agency
FAO	Food and Agricultural Organization
IDSS	Integrated Decision Support System
KL	Kullback-Leibler (distance)
MC	Markov Chain
OOBN	Object-Orientated Bayesian Network

SAGE	Scientific Advisory Group for Emergencies
TV	Total Variation (distance)

Chapter 1

Introduction

My first exploration into the world of Bayesian networks (BN) was a mini-project with my supervisor Professor Jim Q. Smith, creating a dynamic representation of the U.K. fruit market. This small network was an introduction to the larger Food Security application which has become the ongoing application to my work. During this project I became aware of the non-trivial task of structuring and populating such graphical models and questioned the validity of the outputs should errors be unwittingly incorporated into the system. Despite the fact that there are many people who now encourage the use of Bayesian networks in a variety of fields, and numerous facilitators who elicit the probabilities for such models, there is actually a considerable disconnect between the focus of current theoretical work and the actual practical challenges when applying such systems in practice. Although there has been substantial development for robustness of parametric models, theory of sophisticated statistical robustness analyses on graphical models is surprisingly sparse. This small project sparked the research contained within this thesis which attempts to capture the robustness within Bayesian networks, and hopefully provides tailored protocols for the creation of a robust graphical model.

One of the key elements when undertaking practical BN modelling is that there is nearly always a vector of sub-variables that is of particular interest within the system. All other variables in the system inform our learning about these target variables. Many authors encourage the construction of a BN to take into account this inherent directionality by ordering the influential variables before the target variables, which are the outputs of the system. Note that these target variables are either components of the utility function which is specified by the model user, or in large systems these target variables may be the outputs of one sub-network which are later used as inputs for the next component. By pre-specifying the target variables we are able to focus attention to a subset of the model, rather than attempting to quantify the robustness of the entire system.

Although other distance metrics have recently been incorporated into BN sensitivity analyses, we have chosen to harness the useful properties of the well-studied total variation distance. Previous work, which shall be discussed herein, has already devised bounds for the variation distance when associated with decision analyses; the types of applications we have in mind here. Using inherent properties of the total variation distance we can diminish errors within a system, and we have formalised a direct link between the total variation distance and conditional independence of variables; the underlying reasoning behind BNs.

Although many of the principles translate smoothly into the continuous domain, in this thesis we concentrate entirely on discrete BNs. Early in my research we discovered that there are many close relationships between finite Markov chains (MCs) and the BN formulation; the simplest case being the dynamic structures studied in Section 6 in this work. Whilst engaged in this technology transfer, we realised that the relationship was much more general than first realised, leading us to link time non-homogeneous Markov chains to BNs. This allows for different transition matrices between each time-slice, making the work truly general. Exploring and formalising this connection between disparate fields has enabled us to benefit from the wealth of results concerning Markov chains and variation distance, as used in convergence results of standard texts. In particular, we discovered that the type of decay relating the effect contributing nodes have on target variables, states with higher time indices, is also experienced within BNs: in the sense that perturbing variables which are distant from the target variable seemed to have little impact on a target margin. This phenomenon is widely known as a folk theorem within applied BN modelling. However, to our knowledge there has been no formal proof of this until now. This became apparent when collaborating with Dr. Tina Nane at TU Delft on BN structure selection, where amongst others, we noticed that even fairly large modifications of somewhat remote variables in the system seem to have very little impact on the target variable, making model selection difficult.

The success and versatility of BNs, has resulted in them being increasingly applied to larger systems, which has by necessity led to the creation of automatic technologies to develop robust models. All the techniques we have presented here are designed to address robustness in very large BNs. However, when writing this thesis I realised that actually annotating a very large BN with the elicitation information and the meaning of the variables made explanations rather tedious and unwieldly. Instead, I have demonstrated our results on small or moderate BNs which still illustrate the efficacy of our methodology.

We begin this thesis with a summary of BN methodology, including our interests of decision support systems and expert elicitation for populating incomplete models (Section 2). My work has been heavily influenced by my experiences and immersion in applying BNs to several different applications, the most influential of these being the Food Security system which, due to its complexity, has provided much of the incentive for this thesis material. All relevant applications and examples are detailed in Section 3. Section 4 introduces our arguments for the use of total variation distance in robustness studies, together with useful bounds and properties specifically for cases in which we have target variables. We also present original measures for using the total variation distance to compare conditional probabilities or transition matrices. Having explored the foundations we focus our attention on the robustness within a Bayesian network. Section 5 uses the new measures we have defined to calculate the maximum effects of perturbations on our target variables. We then study common structural and probabilistic simplifications, commonly seen in applied BN modelling, to deduce whether they are theoretically comprehensive in terms of robustness within the model. Our final theoretical Section, Section 6, details the formal process of writing a dynamic BN as a Markov chain to harness developments from literature studying temporal effects of misspecifications. There have been many interesting steps to this research and we feel that we have built a strong foundation from which many results could be further developed. Extended research ideas and topics that we sadly did not have time to explore, are documented in Section 7.

Chapter 2

Bayesian Networks

2.1 Introduction to Bayesian Networks

A discrete Bayesian Network (BN) $\mathcal{G}$, sometimes referred to as a directed acyclic graph (DAG), is a probabilistic graphical model allowing us to model uncertainty in a system. Commonly used in decision support systems [Constantinou et al., 2015; Leonelli and Smith, 2013], the versatility and adaptability of BNs makes them a popular choice of probabilistic graphical models in a broad range of applications such as medicine [Friedman et al., 2000; Nagarajan et al., 2013], ecology [Johnson and Mengersen, 2012; Milns et al., 2010] and wider fields [Cano et al., 2004; Johnson et al., 2012; Mascaro et al., 2014].

A BN on a random vector $X \triangleq (X_1, X_2, \ldots, X_m)$ represents a family of models that respect a set of conditional independence hypotheses. Each node in the graphical structure represents a random variable and each directed arc $X_i \rightarrow X_j$ encodes a direct (often interpreted as a causal) dependence between two variables in a set direction, as depicted in Figure 2.1. We use ancestral tree terminology to categorise relationships within the graphical structure: $Pa(i) \subseteq \{X_1, X_2, \ldots, X_{i-1}\}$ are the *parents* of X_i meaning the previously listed variables on which X_i depends, i.e. $X_i \in Pa(j)$ iff $X_i \rightarrow X_j$. Therefore, for every directed arc $X_i \rightarrow X_j$, X_i is labelled as the parent of X_j. Conversely X_j is said to be a *child* of X_i. This relationship is roughly translated to ‘the parent causes the child’. Extending this notion, we can discuss *ancestors* and *descendants* of a node when we are looking backwards or forwards in a chain of influence. For example, a node X_i is an ancestor of X_j if there is a directed path from X_i to X_j. The *ancestral graph* $A(X)$, is the graph containing the vertex X and all ancestors of the vertex X. The final term with respect to nodes that we shall refer to frequently is a *founder* node, which indicates that the node in question has no parent. Note that the absence between two nodes is structurally significant as it highlights the conditional independence between two variables. Being a DAG, a BN is defined so that it cannot contain any directed cycles: there is no path returning to a node when following the directed arrows. Later in this work we rely on the notation X_A which is the sub-vector of the nodes of X with indices in set A.

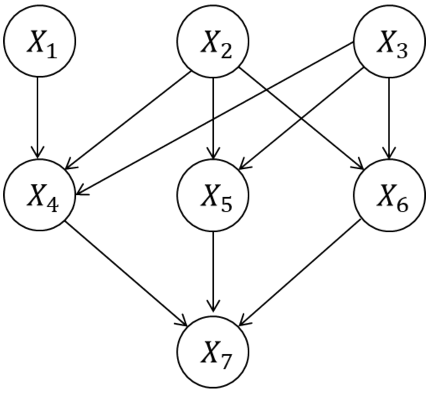
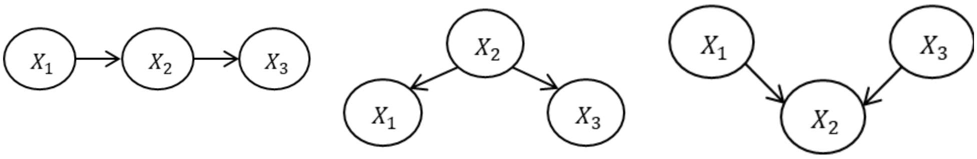


Figure 2.1: A demonstrative BN representation.

In any BN there are three main types of connections found: causal chain, common cause or common effect as represented in Figure 2.2.



(a) Causal Chain (b) Common Cause (c) Common Effect

Figure 2.2: Different structures in a BN, as represented in Korb and Nicholson [2010].

It is important to note that both causal chains and common causes (Figure 2.2(a) and Figure 2.2(b)) encapsulate the same conditional independence statement, $X_3 \perp\!\!\!\perp X_1|X_2$, whereas the common effect shown in Figure 2.2(c) (also known as a collider or v-structure) does not. The common effect actually implies that given information about the child, the parents may be conditionally dependent.

Two BNs are said to be *equivalent* if all the conditional statements underlying one can also be read from the other (and vice versa). This definition indicates that two directed graphs can be different, however the same set of conditional independence statements can be deduced from both structures (a property termed the semi-graphoid axioms, see Geiger and Pearl [1993]; Pearl [1988]). Conversely, a set of conditional independence statements does not define a unique BN representation. BNs encoding the same conditional independencies are said to be in the same *equivalence class*. Occasionally, it might be easier to elicit some graphs than others, for example those that follow a causal ordering. Andersson et al. [1997] prove that if two DAGs have the same underlying skeleton and v-structures then they are in the same equivalence class and therefore equivalent BNs.

The *Markov Blanket* is a set containing the node of interest along with its parents, children and the other parents of its children. The Markov blanket defines all the variables needed in the graphical structure to predict the node X_i or its children. Any further information not contained in the Markov Blanket is surplus.

In some algorithms we find that we need to *moralise* a BN: retain the same graphical structure as the BN $\mathcal{G}$, but iteratively take each node X_i of the graph and add an undirected arc between each pair of parents that are not already directly connected. The term moralisation furthers the family tree analogy, as we are effectively marrying the parents of each child. The *skeleton* of a graph is the exact replication of the BN $\mathcal{G}$, but replacing all directed arcs with undirected arcs and a *mixed* BN has a combination of directed and undirected arcs. Applying these definitions to the BN depicted in Figure 2.1 we can now determine the skeleton structure and the moralised graph (shown in Figure 2.3).

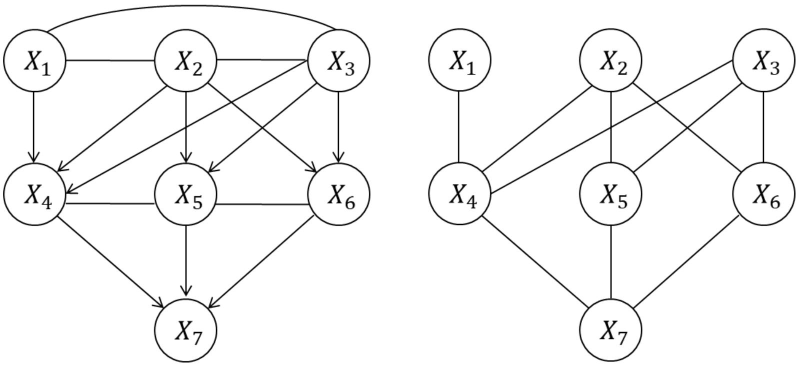


Figure 2.3: The moral graph (left) and the skeleton (right) of the BN shown in Figure 2.1.

Determining conditional independence of single nodes using v-structures is relatively simple; however deducing whether two sets of variables are independent can be trickier and to do this we use the d-separation theorem Lauritzen [1996]; Pearl [1988]. Although the d-separation theorem outlines the criteria for determining whether two sets are independent given a third conditioning set, it is difficult to compute. To check for d-separation we can follow the algorithm below:

1. Find the ancestral graph.
2. Moralise the ancestral graph.
3. Replace all directed arcs with undirected arcs.
4. Delete all given nodes and their edges.
5. If there is no path between the two sets of variables then they are independent.

It can be proved that all the conditional independence hypotheses encoded in a BN can be derived from a much smaller set of statements that link in turn to a set of nested factorisations:

$$X_i \amalg \boldsymbol{X}_{R(i)} | \boldsymbol{X}_{P_a(i)}, \tag{2.1}$$

where $R(i) \triangleq \{X_1, X_2, \ldots, X_{i-1}\} \backslash Pa(i)$.

Recall that to calculate the full joint density/mass function we can use the chain rule:

$$\begin{aligned} p(x_1, x_2, \ldots, x_m) &= p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)\ldots p(x_m|x_1, x_2, \ldots, x_{m-1}) \\ &= \prod_i p(x_i|x_1, x_2, \ldots, x_{i-1}). \end{aligned}$$

However, from Equation 2.1 note that each random variable is independent of all others conditional on its parents. This allows us to simplify the full joint density/mass function to

$$p(x_1, x_2, \ldots, x_m) = \prod_i p(x_i|Pa(i)), \tag{2.2}$$

which for the specific BN in Figure 2.1 gives

$$p(x_1, x_2, \ldots, x_7) = p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3)p(x_5|x_2, x_3)p(x_6|x_2, x_3)p(x_7|x_4, x_5, x_6).$$

An important subclass of BNs whose properties we discuss later, are those which are called decomposable. A *decomposable* BN $\mathcal{G}$ is a BN in which every parent set of each node X_i in the graph forms a complete subgraph of $\mathcal{G}$. An alternative way of specifying this is, if the moralised graph of $\mathcal{G}$ is equivalent to $\mathcal{G}$ then the graph is already decomposable. It is simple to show that any BN can (albeit inefficiently) be re-expressed as a decomposable BN constructed through a process called triangulation [Fenton and Neil, 2012; Korb and Nicholson, 2010; Lauritzen, 1996], which maintains acyclicity whilst relaxing some conditional independencies. This property, widely used for propagation algorithms, can also be used for robustness analyses. When a BN is decomposable it can be shown [see for example Lauritzen, 1996; Smith, 2010] that the joint density distribution factors in the following manner: the *cliques* $\{\boldsymbol{X}_{C_1}, \boldsymbol{X}_{C_2}, \boldsymbol{X}_{C_3}, \ldots, \boldsymbol{X}_{C_m}\}$ of a decomposable BN are its maximal connected subsets and can be ordered so that the indices respect the running intersection property starting with clique C_1. We call $\boldsymbol{X}_{S(i)}$ where $S_i = C_i \cap \cup_{k=1}^{i-1} C_k$ the *separator* of $\boldsymbol{X}_{C_i}$ from $\cup_{k=1}^{i-1} C_k$. An indexing is said to satisfy the *running intersection property* if for all $i = 1, 2, \ldots, m$ there exists some index $j < i$ such that $S_i = C_i \cap \cup_{k=1}^{i-1} C_k \subseteq C_j$. This implies that the result of intersecting a clique with all previous cliques is contained within one or more earlier cliques Lauritzen [1996]; Smith [2010]. Note the choice of C_j may not be unique.

We can select one of these orderings to depict a non-unique *junction tree* $\mathcal{J}(\mathcal{G})$. A junction tree is an undirected tree whose vertices are the cliques and where two cliques C_i and C_j are linked by an undirected edge if and only if within the chosen indexing $j = j(i)$. Note that by definition each of these edges can be labelled by a corresponding separator of $\mathcal{G}$. To simplify the development we will assume that the entries of the joint mass function are all strictly positive. This is advised from a practical point of view, by a number of authors, e.g. Korb and Nicholson [2010], when dealing with no known functional relations. Lauritzen and Spiegelhalter [1988] prove that strictly we do not need this positivity condition, but we are required to define division by zero. It is then proven [e.g. see Cowell et al., 2007; Smith, 2010] that the joint mass function of any such decomposable BN $\mathcal{G}$ respects the following formula:

$$p(\boldsymbol{x}) = \frac{p(\boldsymbol{x}_{C_1}) \cdot p(\boldsymbol{x}_{C_2}) \cdot p(\boldsymbol{x}_{C_3}) \ldots p(\boldsymbol{x}_{C_m})}{p(\boldsymbol{x}_{S_2}) \cdot p(\boldsymbol{x}_{S_3}) \ldots p(\boldsymbol{x}_{S_m})}.$$

One straightforward, but important consequence of this decomposition used herein, is that given any BN $\mathcal{G}$ and an associated junction tree $\mathcal{J}(\mathcal{G})$, then for any two cliques $\boldsymbol{X}_{C_1}, \boldsymbol{X}_{C_k}$, there is a *unique* sequence of cliques $(\boldsymbol{X}_{C_1}, \boldsymbol{X}_{C_2}, \boldsymbol{X}_{C_3}, \ldots, \boldsymbol{X}_{C_k})$ and separators $(\boldsymbol{X}_{S_2}, \boldsymbol{X}_{S_3}, \ldots, \boldsymbol{X}_{S_k})$ between $\boldsymbol{X}_{C_1}$ and $\boldsymbol{X}_{C_k}$ within $\mathcal{J}(\mathcal{G})$. This sequence forms a simple path in $\mathcal{J}(\mathcal{G})$, i.e. it corresponds to the unique path between two nodes, in which no edge is repeated. If we write $\overline{C}_k \triangleq \cup_{j=1}^{k} C_j$, and since $S_i \subseteq C_i$ we know that $\boldsymbol{x}_{S_i}$ is a subvector of $\boldsymbol{x}_{C_i}$ giving

$$\begin{aligned} p(\boldsymbol{x}_{\overline{C}_k}) &= \frac{p(\boldsymbol{x}_{C_1}).p(\boldsymbol{x}_{C_2}).p(\boldsymbol{x}_{C_3})\ldots p(\boldsymbol{x}_{C_k})}{p(\boldsymbol{x}_{S_2}).p(\boldsymbol{x}_{S_3})\ldots p(\boldsymbol{x}_{S_k})} \\ &= p(\boldsymbol{x}_{C_1}).p(\boldsymbol{x}_{C_2}|\boldsymbol{x}_{S_2}).p(\boldsymbol{x}_{C_3}|\boldsymbol{x}_{S_3})\ldots p(\boldsymbol{x}_{C_k}|\boldsymbol{x}_{S_k}). \end{aligned} \tag{2.3}$$

Lemma 2.1.1. *It follows from Equation 2.3 and the conditional independence in $\mathcal{G}$ that if $T_{1,k} \triangleq \cup_{i=3}^{k-1} S_i$ then*

$$p(\boldsymbol{x}_{C_1 \cup C_k}) = \sum_{\boldsymbol{x}_{T_{1,k}}} p(\boldsymbol{x}_{C_1}).p(\boldsymbol{x}_{S_3}|\boldsymbol{x}_{S_2}).p(\boldsymbol{x}_{S_4}|\boldsymbol{x}_{S_3})\ldots p(\boldsymbol{x}_{S_k}|\boldsymbol{x}_{S_{k-1}})p(\boldsymbol{x}_{C_k}|\boldsymbol{x}_{S_k}).$$

This is an interesting property because it links the graphical model to a non-homogeneous Markov Chain via a junction tree which has all internal cliques marginalised out. A Markov chain (MC) is simply a stochastic sequence of events underpinned by a collection of conditional independence statements. Each MC is initiated with a joint probability distribution for all variables within the system and updated temporally using transition matrices which can be categorised as homogeneous (each transition matrix between time-slices is identical) or non-homogeneous (transition matrices differ), see [Smith, 2010] for further details. Therefore we have constructed a non-homogeneous Markov Chain from the donating clique C_1 to the receiving clique C_k. This is valid because of the assumed conditional independences since each tree has a unique path between the perturbed node and the node of interest. Although this result is derived directly from the elementary properties of trees it is important, yet an often overlooked result that will be exploited later in the thesis.

2.1.1 Dynamic Bayesian Networks

A Bayesian network in its rawest form looks at a specific situation at a specific moment in time. The versatility of a BN is in its ability to fold in complex information thereby tailoring the model to create a bespoke model using the simple underlying framework discussed previously. There are many adaptations of a BN, however we will be focussing on some of the more common approaches: temporal relationships, replication within the graphical structure and decision support systems.

In the penultimate chapter of this thesis we study the robustness in BNs containing temporal conditions. This enables users to model the evolution of complex processes into the future and allows decision makers to deduce long-term effects of verdicts. A Dynamic Bayesian Network (DBN) generalises a simple BN to enable modelling over time by introducing explicit temporal relationships. Variables are now denoted by $X_{t,j}$ with $t = 1, 2, \dots, n$ indicating the time slice and $j = 1, 2, \dots, m$ being the individual variable indicator . We shall refer to a *time-slice* as a snapshot of the system at a discrete time-step which notationally translates into all nodes having the same t subscript. The most popular form of DBNs assume that the underlying conditional independence statements are identical within each time-slice, therefore allowing the topological structure of the DAG to remain unaltered through time.

In the general form of the BN introduced previously, each directed arc represented inferential relationships between variables in the same time-slice. These arcs are termed *intra-slice arcs* and satisfy the relationship $X_{t,i} \to X_{t,i'}$. *Inter-slice arcs*, also known as temporal arcs, characterise dynamic effects in the system and can be between different nodes $X_{t,i} \to X_{t',i'}$ across time-slices or the same node $X_{t,i} \to X_{t',i}$ across time-slices. For distinction, all figures within this thesis represent dynamic arcs in red and non-dynamic arcs as black. For example we can extend Figure 2.1 to be a DBN, as shown below in Figure 2.4:

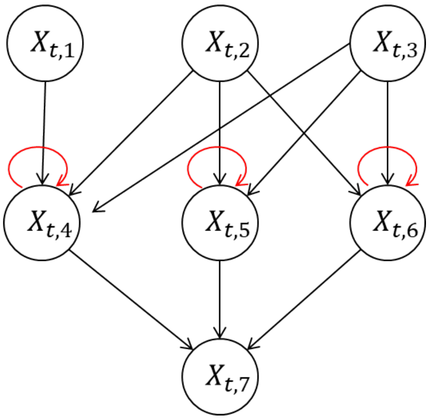


Figure 2.4: Figure 2.1 extended to a DBN by adding temporal arcs (drawn in red).

Although memory in a DBN can be infinite, it is common to assume that the Markov Property holds; each time-slice t is dependent only upon the previous time step $t-1$ and no other previous time points. DBNs which follow this condition are known as two time-slice DBNs (2TSDBN) to indicate that the systems only needs to retain its memory over the past two time-slices. As for the general BN, each node of the network must be assigned an accompanying conditional probability distribution to quantify the conditional probabilities between variables which in the discrete case will take the form of a conditional probability table (CPT), see Section 2.1.3 for more detail. However, note that each CPT for a DBN must include all intra-slice and inter-slice relationships. By construction, each DBN can be unwrapped and re-expressed as a simple BN by replicating nodes within a single time frame where necessary, this process is depicted below in Figure 2.5:

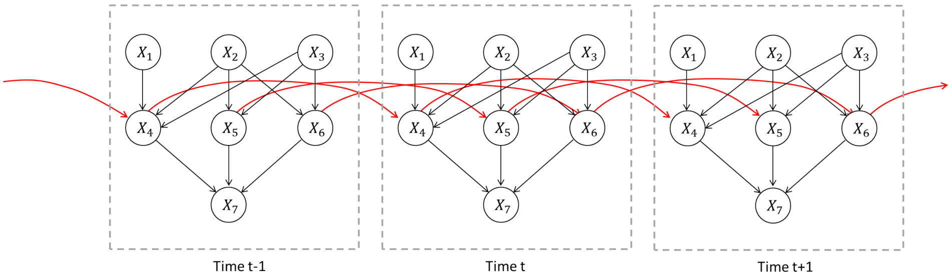


Figure 2.5: Unwrapping the DBN from Figure 2.4 to a BN.

Whenever a BN is used in a Decision Support System (DSS) it is crucial that we understand which node(s), or combinations of nodes, are considered to be the outputs as we are most interested in the accurate predictions of these variables. Therefore, it is essential that we perform robustness analyses with these attributes in mind, to determine the effect that any structural inaccuracies or probability misspecifications have on the results.

2.1.2 Integrated Decision Support Systems

BNs are now a widely used probabilistic modelling tool, particularly in the field of decision support. In Decision Support Systems (DSSs) a decision-maker identifies a problem for which they need coherent statistical recommendations. Due to the nature of decision support there are normally a few pre-specified variables of interest, often called the target variables, which together form the attributes of a utility function. This utility function enables the user of the system to score and rank the model recommendations with respect to their specific requirements. By pre-specifying target variables there is a dramatic decrease in size and focus within the model, enabling us to create a bespoke approach for the variables of main concern. However, decision-makers (or experts asked for probability judgements) may question the robustness of the outputs in relation to the inputs of the model, to quantify the impact on decisions if erroneous information is used.

Integrated Decision Support Systems (IDSSs) reflect the notion that it is highly unusual for one person or group to hold all the relevant information for a large complex system [Smith et al., 2015]. Additionally, if the parameters of the posterior distribution are all independent, then the system is distributive and can be separated according to relevant expert panels. This enables the distributed systems to be treated as separate objects in analogy with the objects in an object-oriented Bayesian network (OOBN). OOBNs succinctly compact any repetition within a BN by introducing the notion of classes: a collection of nodes and objects. This enables the modeller to significantly reduce the time spent populating the model by copying sub-networks and specific CPTs around the BN. The IDSS presents a shared overarching framework to coherently piece together the separate sub-networks.

The reason these issues are of interest in this thesis, is to emphasise and illustrate that the elicitation of a BN or DBN is often just part of a much bigger process. In an IDSS there are many integrated distinct components within the model, so that an output of one sub-network may be an input for another network. Therefore, we need to consider not just outputs of the entire system, but internal inputs and outputs which affect contributing systems. This directionality and contraction of interest to a select few variables is not studied in other robustness approaches. This entire process must be done in a defensible manner to give (via the BN) the reasons for the judgements. Furthermore, it demonstrates how robust are the distributions of these *particular* inputs (as needed by the composite system) to domain mispecifications.

2.1.3 Applying a BN in practice

It is now acknowledged as best practice [Cowell et al., 1999; Laskey and Mahoney, 2000; Smith, 2010] that BNs are set up in two distinct stages. Firstly the structure of the BN, as expressed by its Directed Acyclic Graph (DAG), is either directly elicited from domain experts or when sufficient supporting data exists, learned from the data using a model search algorithm with default priors on the hyperparameters, see Boneh [2010] and Korb and Nicholson [2010]. The graph can then be interrogated through natural language to examine whether the consequences of its embedded conditional independences make sense within the domain. Once this graphical framework has been found to have no obvious faults i.e. be requisite [see Phillips, 1984], the graph would be extended into a full probabilistic model. In the case of a discrete BN, this second stage involves eliciting or estimating the entries of its conditional probability tables (CPTs), using priors on probabilities informed by expert judgements. These CPTs give the numerical prespecification of all the conditional probabilities needed to generate the full joint probability mass function over the whole space and hence a fully specified probability model.

Note that in this thesis, our CPTs will have rows denoting parent level combinations and columns denoting the levels of the child. Using fundamental properties of conditional probabilities we can state that rows must sum to one. In some older texts, the conditional probability table is written as the transpose of our CPT. As an example, suppose we have three variables X_1, X_2, X_3 all of which are binary, with a structure as shown in Figure 2.6 then the CPT of X_3 would be written as in Table 2.1:

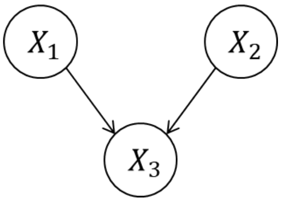


Figure 2.6: 3-node binary BN.

		$P(X_3 \mid X_1, X_2)$	
X_1	X_2	0	1
1	1	p_1	$\overline{p_1}$
1	0	p_2	$\overline{p_2}$
0	1	p_3	$\overline{p_3}$
0	0	p_4	$\overline{p_4}$

Table 2.1: The CPT of X_3.

When engaging in this two stage process it is essential for the analyst to be aware of precisely which inputs of the process might be critical to the inferences made through the BN, see Albrecht et al. [2014]. This then enables most attention in an elicitation (or statistical estimation of the graph) to be focused on ensuring these critical features are specified as accurately as possible. This is especially important when elicitation or estimation is resource limited, as is usually the case in practice. The client (or decision-maker of the model) can then optimise their allocation of resources and so spend most time eliciting those elements of the model whose misspecification might most influence the required outputs of the model.

Occasionally the modeller has access to a complete training data set from which the most promising explanatory BN can be selected, whose associated $p(\mathbf{x})$ respecting Equation 2.2 appears to best fit the data. There are many ways to do this including using software packages such as 'bnlearn' in R [see Scutari and Denis, 2014]. However, when applying such a model selection method in practice, even for low dimensional BNs, it is common to find many models that score similarly well. Even if we know this model to be true, as in a simulation exercise or even moderately sized problems, it has been demonstrated that the best model is only close to the generating process unless the training data set is absolutely enormous, see Cussens [2011]. Naturally there are also the usual statistical errors associated with the representivity of the actual data set we use, even if sampling is performed at random. Yet again, it is rare for a single data generating model to be unequivocally identified. Considering the robustness in the fitted model is therefore a fundamental element of any ensuing statistical analysis.

The second method of creating a BN is by performing a direct elicitation from an expert. Here, having listed the variables in an order (which might be) compatible with the order in which those measurements may occur, the expert is asked for each X_i, $i = 1, 2, \ldots, m$ of the previously listed variables which might be relevant to forecasting the current variable. Building on this qualitative framework, hopefully faithful to the expert's actual judgements, we then proceed to embellish the graph by supplementing the structure with the specification of the corresponding CPTs $\big\{p(x_i|\boldsymbol{x}_{P_a(i)}) : i = 1, 2, 3, \ldots, m\big\}$. These probabilities will be subject to elicitation error, although following a structured elicitation protocol (as discussed further in Section 2.2.1) aims to mitigate this specification error [EFSA, 2014; Korb and Nicholson, 2010; Smith, 2010]. An understanding of the robustness to perturbations of both the hypothesised graphical framework and also the entries in the CPTs of any inferential assumptions taken here will clearly be critical to a good statistical analysis.

It is important to understand that formulating and compiling a BN is an iterative process with many assumptions and some subjective decisions. This process has been highly discussed in the knowledge engineering literature Korb and Nicholson [2010]; Laskey [1995]. Our main focus in this work is decision support systems and the robustness within these models.

2.1.4 DBN Software

There are numerous statistical software packages that allow for the construction of BNs and DBNs, for example Netica, R, Analytica, Hugin, Uninet and GeNIe to name a few. A comprehensive comparative table is presented in Korb and Nicholson [2010], outlining benefits and drawbacks to many of these options and also including beginner's tutorials.

Throughout this work we have chosen to use mainly Netica in our Food Security application to enable systematic comparisons across networks already created. Netica is a free user-friendly, well-documented software which requires no coding experience so is the most widely used software by domain statisticians. Software such as this is ideal when creating BNs for third parties or decision makers who wish to continue adapting and updating the model with little statistical experience. However, the 'bnlearn' package in R has also been utilised for examples.

2.2 General Practice

2.2.1 Subjective Probability and Elicitation Protocols

In the digital age of the 21st Century we are increasingly awed by the idea of 'big data' and its ability to be harnessed in a multitude of scenarios to solve any problem. Unfortunately, in decision support systems it is rare to have all the necessary information to fully specify the model of interest at the outset. This deficiency can arise from a number of origins: for example, collecting the information is too expensive (either financially or laboriously); there has been little interest in the data until recently (no availability); data sources may be flawed; or more commonly, the data may not be observable at this moment in time due to the uniqueness of the scenario in question. As decision support systems are created with an inherent time-constraint, the policy or decision maker has a pre-specified deadline by which to announce their verdict. This time restriction can limit the options available to the modeller in terms of sourcing the required data for the completion of the model. In circumstances such as this, we can turn to the field of subjective probability in which people are asked for their own personal judgements or beliefs about the likelihood of an event occurring.

Subjectivity is a key underlying element of expert judgement, a topic which has been studied for many years [see for example De Finetti, 1974; French, 1985;

O'Hagan et al., 2006]. Subjective probabilities are the beliefs of a specific individual quantifying their uncertainty of an event, at the time of statement. This definition enables two individuals to maintain different probabilities for the same scenario, and also allows an individual to revise their beliefs if they learn new information pertinent to the question at a different time.

There are two main types of elicitation related to building a BN: structural elicitation and probabilistic elicitation. Structural elicitation is concerned with finding a suitable graphical structure to formally represent the variables and relationships within the complex process; Probabilistic elicitation is populating the structure you have previously elicited by asking relevant domain experts for uncertainty quantities to fill the conditional probability tables. In either elicitation procedure there is ample psychology literature [Kahneman et al., 1982; Kahneman and Tversky, 1973; Nickerson, 1998] which forewarns facilitators of behavioural biases which may adversely affect or nullify the results obtained. There is abundant advice to facilitators and experts to help combat these biases and how they may proceed with caution [see Burgman, 2016; Dias et al., 2018; French et al., 2009; O'Hagan et al., 2006]. It is important to remember that each BN will be different and require a flexible, iterative approach to create a robust and meaningful model. A detailed glossary of such biases can be found in EFSA [2014]; Fenton and Neil [2012]; O'Hagan et al. [2006]; however, we outline a few common problems below:

- **Anchoring and Adjustment:** Anchoring refers to an expert inadvertently becoming attached to a quantity stated in a previous question, or in the wording of the current question. Adjustment is the name given when this anchor is taken and modified slightly, rather than the expert initiating their own impartial reasoning.

- **Availability:** Participants are prone to using recent memories and experiences to guide their probability statements and therefore present an event as being more probable if they have a stronger recollection of it happening. Note that the strength of a memory can be affected by time since origination or the emotional attachment connected to the content.

- **Group Effects:** If consulting a panel of experts simultaneously, the group dynamic can affect responses. For example, younger/inexperienced experts may concede to older/experienced attendees, or the most charismatic/loudest person may override quieter panellists.

- **Range-Frequency Compromise:** In questions where categories are implied or stated, the probabilities quoted by experts can be influenced by an increasing number of categories.

There are many decisions involved in eliciting expert judgements which will be reproducible and coherent. In the preparatory stages, the modeller chooses which quantities to elicit; the experts from whom we wish to elicit the information; the size and make-up of the group of experts from whom we wish to elicit the information; whether to conduct the elicitation face-to-face or remotely using questionnaires; which questions to ask during the elicitation; and of course, how to aggregate the findings. The European Food Safety Authority (EFSA) has published a report EFSA [2014] that summarises the currently perceived best practice for eliciting expert judgements and incorporating them into probabilistic models. Happily, this application area coincides with our running example of Food Security.

Although fundamentally similar, there are a range of proposed elicitation processes which advocate anything between three to seven step procedures [see for example Clemen and Reilly, 2001; Phillips, 1999; Walls and Quigley, 2001]. Garthwaite et al. [2005] constructed a four step process to extract personal beliefs from people, which we discuss here due to its similarities to the IDEA protocol;

- Preparation - deduce the number of people required; determine the relevant domain experts; train the experts; choose the most appropriate protocol; organise calibration questions, etc.
- Perform the elicitation to obtain specific summaries of the expert's belief distribution.
- Fit a suitable probability distribution to the obtained summaries.
- Validate the model by iterating the process to determine whether the outcome is satisfactory or whether adjustments need to be made.

They also note that generally, poor results are produced when people are asked to provide a variance. Instead of requesting a mean and variance, it is more common to request a mean and a credible interval. Although there are multiple variations of this procedure [see O'Hagan et al., 2006], the foundation steps are very similar. In order to obtain probabilities which are reproducible and reliable, it is recommended that facilitators follow an established elicitation protocol such as (but not limited to) the Sheffield Protocol [Gosling, 2018; O'Hagan et al., 2006], the Delphi Process

[Dalkey and Helmer, 1963; EFSA, 2014], Cooke's Method [Cooke, 1991] or the IDEA protocol [Hanea et al., 2017; Hemming et al., 2018] all of which have been compared and discussed in numerous texts already highlighted. We shall briefly cover those which are most pertinent to our Food Security application below.

Cooke's Classical Method

Cooke's Classical method does not usually allow any interaction between experts: each individual attends an interview and gives a personal probability density function to the facilitator. The distributions of all experts are then combined using a mathematical aggregation formula to obtain one combined probability density function for each variable of interest. The one-on-one dynamic mitigates any group bias which may otherwise occur; however, note that the choice of mathematical aggregation formula and expert weight analysis is subjective.

In Cooke's Classical method the facilitator uses seed questions (otherwise known as calibration questions) to determine the performance of each expert. The facilitator knows the answer to a seed question, however the expert does not: usually accepted but unpublished journal articles are recommended for material. This allows the expert to use their domain knowledge to provide their subjective opinion. It is therefore imperative that the seed question be as similar to the variables of interest as possible for comparative purposes; there is also an implicit assumption that future performance is based on past performance.

IDEA Protocol

The IDEA protocol follows a standard elicitation procedure: pre-processing, elicitation and post-analysis as synthesised in Figure 2.7 below [as presented in Hanea et al., 2017].

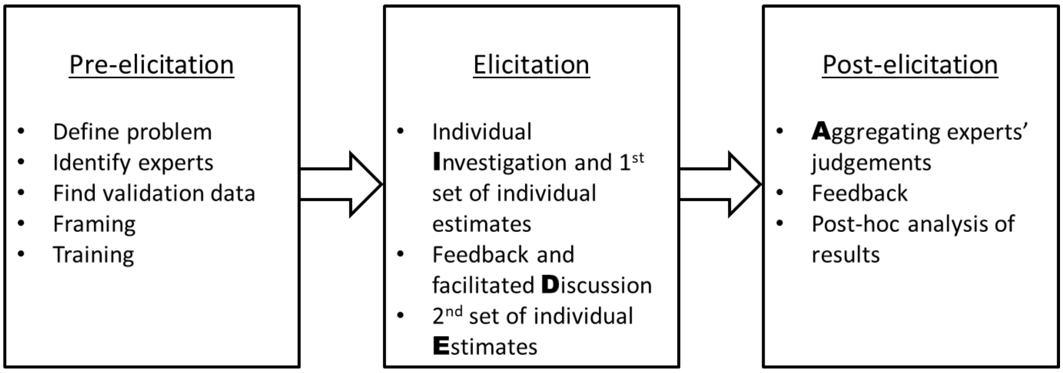

Pre-elicitation
• Define problem
• Identify experts
• Find validation data
• Framing
• Training
Elicitation
• Individual Investigation and 1st set of individual estimates
• Feedback and facilitated Discussion
• 2nd set of individual Estimates
Post-elicitation
• Aggregating experts' judgements
• Feedback
• Post-hoc analysis of results

Figure 2.7: Outline of the steps associated with the IDEA protocol as shown in Hanea et al. [2017].

The IDEA acronym denotes the four internal stages which involve the experts: Investigate, Discuss, Estimate and Aggregate. Initially experts are asked for individual assessments to the questions of interest following a strict ‘upper bound’, ‘lower bound’, ‘best estimate’, ordering in an effort to combat anchoring and adjustment bias. After personal beliefs have been collected, experts participate in a facilitated discussion (D step) to share information, relevant literature or experience. Once completed, experts are invited to revise their estimates in private if they feel that the discussion has altered their beliefs. This step is once again undertaken privately so that experts feel no pressure to edit their initial estimates if they do not feel that the discussion has given them any new information. The final step of the protocol is for the facilitator to mathematically aggregate the experts’ judgements into one final assessment.

As a relatively new elicitation protocol, the IDEA method aims to mitigate perceived flaws in older comparable procedures and uses updated psychology literature to create a more adaptive and rigorous protocol. This is reflected in the similarities to older methods: the discussion rounds are iterative (as in the Delphi method) and they encourage personal beliefs which are aggregated mathematically as in Cooke’s Classical method; however, they introduce the ability to provide final revised individual assessments. As in other methods, experts contributions are weighted using their calibration question responses.

Plotting the mean and intervals provided by experts we can deduce the following characteristics [Aspinall, 2006; Aspinall and Cooke, 2013; Burgman, 2016]:

- **Information:** This is an assessment of the confidence of an expert in their beliefs: a tighter interval bound is more informative than a wide bound.
- **Accuracy:** An expert's estimate is accurate when the probability of the true event lies within the stated bounds.
- **Bias:** When an expert consistently overstates or understates judges the probability of an event occurring.

Ideally we would like each expert to be statistically accurate and highly informative when quantifying their uncertainty. Realistically this is a rare occurrence. Experts are knowledgeable about their subject field, however an elicitation is more interested in the experts' ability to accurately assess and quantify uncertainty. Deducing the performance of experts in an elicitation is strongly advised and is a crucial step for mathematical aggregation; taking a set of personal beliefs and combining them in a statistically sound manner in order to form one uncertainty assessment. When deciding how to perform the mathematical aggregation, each expert can be weighted according to their performance. The facilitator is able to use global weights, a performance weight calculated across all calibration questions, or an item weight in which the weight is calculated for one specific seed question. Alternatively, there is an option to weight experts equally rather than by performance.

2.2.2 My Elicitation Experiences

During the course of my postgraduate studies, I have been fortunate to act as a recorder for four elicitation workshops; two of which have been related to the large Food Security project detailed in Section 3.1. Each separate occasion has enabled me to observe the interaction between facilitator and experts, critique the different elicitation protocols, scrutinise any robustness issues which naturally occur throughout such processes, along with any methods undertaken by facilitators to mitigate these problems. Being involved in such projects from infancy to fruition has awarded me a unique insight into the creation of a Bayesian network and has heavily influenced the work produced in this thesis. Using the information I have inferred from such practices, I have been able to tailor my robustness analyses for suitability and practicality. Below I briefly summarise each elicitation workshop attended, and review any significant robustness matters.

ADMLC Workshop

In November 2015, I was invited to record an elicitation jointly co-ordinated between Professor Simon French and the U.K. Atmospheric Dispersion Modelling Liaison Committee (ADMLC). The primary aim of the workshop was “communicating geographical uncertainty during a radiological emergency” and participants were members of U.K. organisations who potentially could be invited to sit on a COBR (Cabinet Office Briefing Room) committee in the event that the U.K. suffered a nuclear crisis. The thirteen experts present were from a range of disciplines and Chatham House Rules were followed: meaning that all information garnered in the meeting could be shared externally only if anonymised first.

This workshop was my initial involvement in any practical elicitation event and my role as recorder was to faithfully represent the discussions and dialogues in the room. It is important to emphasise that this work was not related to BNs and did not involve any probability elicitation. However, the hosts were trying to understand what visual/scientific aids a decision support system may require to communicate results faithfully, as well as the uses of a DSS after creation. The opportunity to gain an overview into DSSs was extremely thought provoking and I acquired some useful information such as: the diverse range of opinions expressed; the fact that most discussion was about the structure of the process and what the attributes variables were; the difficulties in getting participants to agree on one recommendation: some issues were critical whilst others were deemed unimportant either by a single individual or consensus: there are a broad range of people associated in creating a DSS.

Opening discussions welcomed experts and introduced all participants, discerning relevant expertise and domain knowledge. A detailed description of the problem was given and the goals of the exercise were re-emphasised before any elicitations commenced. Firstly, experts were invited to give their individual reflections; specifically their initial reaction to the problem, any information required to make a more informed recommendation and their own advice to COBR under the given hypothetical scenario. A mock SAGE (Scientific Advisory Group for Emergencies) meeting was then simulated to coherently review the situation, drawing attention to important scientific findings and outlining the group advice to be forwarded to COBR. The co-ordinators were not concerned with the actual advice stated in this meeting, but rather with the scientific information provided to COBR (or lack thereof) and the format in which this information was reported. Finally, experts were again asked for individual assessments on communication formats discussed and personal recommendations for improvement. Discussions were facilitated by Professor Simon French and Professor Jim Smith and were conducted either as round-table discussions or post-it note discussions (in which each participant writes single line topic suggestions on paper and the facilitator clusters similar responses to use as a starting point for further debate).

Pollinator Abundance Workshop

Unlike the ADMLC workshop, the following two elicitation experiences in which I was involved were heavily immersed in BN creation for components of the Food Security system (introduced in greater depth in Section 3.1). As in the ADMLC meeting, I recorded the Pollinator workshop (and subsequent CPI workshop) and am a co-author on the subsequent apicultural paper detailing the process and results, Barons et al. [2018a].

In April 2016 the Pollinator Abundance elicitation was hosted by Dr. Martine Barons (University of Warwick) and facilitated by Dr. Anca Hanea (Univeristy of Melbourne), one of the co-creators of the IDEA protocol. The aim of the workshop was to elicit three conditional probability tables concerning pollinator abundance of three key species dependent on weather, pest and environmental accessibility. The six focus variables formed a small subset of a sub-network within the Food Security IDSS Barons et al. [2018b] and were chosen due to the lack of observational data in the apicultural domain and their importance regarding other variables such as U.K. crop production.

Preliminary domain knowledge and relevant literature insights were obtained through a pre-workshop meeting with a key expert. This interchange led to proposed redefining of select variables, changing ambiguous definitions and setting appropriate quantitative thresholds: also of benefit was receiving crucial recommendations for potential calibration questions. These calibration questions were taken from unpublished (but accepted) papers in apicultural journals. Twelve experts attended the three day workshop, so that all three sections of the IDEA protocol could be fully explored: individual assessments of probabilities, discussion of context and reasoning, final individual assessments if changes were necessary. For each question, participants were required to give their lowest plausible, highest plausible and best guess estimates, with this specific ordering aimed at combating anchoring bias. Calibration questions were emailed to experts after the workshop and a group skype meeting was facilitated to discuss answers, thus ensuring that the three step protocol was enacted albeit remotely. Note that for the duration of the workshop we had no information regarding weights for individual experts. Therefore discussions began with an overview of anonymised individual responses alongside the unweighted group response. For a full discussion of proceedings along with probabilities and final weighted outcomes see our paper Barons et al. [2018a].

The opportunity to liaise with experts prior to the commencement of a workshop allowed the team (facilitators, problem-owner and recorders) to quickly focus on the important variables and features of the model. By asking simple questions in terms of best and worst case scenarios it was possible to deduce whether relationships between variables were strong enough to warrant inclusion, and also guided the model creator to select the optimal number of states per variable. This kind of qualitative discussion is a rough robustness analysis for those who are experienced in creating BNs and is very pertinent to the work discussed in this thesis which formalises and guides such procedures. Inexperienced BN users may question which variables are most important to accurately elicit, how many states to include for each variable and how many ancestors to include in the inferential or temporal chain it is necessary to include in a model. As a preliminary exercise, it is highly recommended that users discuss the model with an expert prior to elicitations. Our work in later chapters aims to provide formal justification to many of these questions, in scenarios when there are pre-specified target variables.

It is common in most elicitation exercises for experts to query their own expertise and they can often fixate on the accuracy of their stated probabilities. To reassure participants that subjective opinions are meaningful and relevant, each workshop opens with a detailed introduction covering the aims of the project; the procedure; basic statistical properties; psychological biases which may occur; the calibration process; assigning weights to individual experts and how results will be used in future work. To allay fears, results provided in this thesis could be used in discussion with experts about the robustness in BNs to structural errors and misspecifications. By presenting the process of bounding the effect errors have on the variables of interest, we illustrate that the model user can determine the impact on the utility function if certain stated probabilities were perturbed.

CPI Workshop

In June 2018 Dr. Martine Barons hosted a workshop with facilitator Willy Aspinall to ascertain the conditional probabilities between household food security status dependent on household disposable income and the cost of food. Questions were aimed at discovering the potential impact on food prices under specified Brexit deal scenarios or oil price fluctuations. Each food group was selected from the CPI (Consumer Price Index) basket of goods as used by most U.K. statistical agencies.

The workshop followed Cooke's Classical method in the following stages: seed questions; presenting the problem; questions of interest; plenary discussion. Note that the workshop opening differed significantly from the IDEA protocol workshop by considering the seed questions first; by gaining seed responses this then enabled the facilitator to weight all experts prior to the questions of interest being asked. Experts were then given a short presentation on the methodology, highlighting the difference between global weights and individual weights, and the importance of giving accurate or precise subjective probabilities. Next, were the questions of interest in which experts were asked to complete their individual assessments with no discussion or sharing of information; however, if clarity was needed for a definition or specific context then this was provided to all. The final stage of the workshop was communicating each anonymised individual assessments, an equal weight response and a performance weighted response with the use of graphs. This visual representation of the responses sparked facilitated discussion, with experts sharing knowledge, expertise and experience. However, unlike in the IDEA protocol, experts were not permitted to change or revise their earlier estimates in light of this discussion. At the conclusion of the workshop I felt an opportunity had been missed in not allowing experts to consider any reviews to their initial responses following these perceptive dialogues.

It is interesting to note that occasionally in these workshops we obtain a two-case scenario in which the experts form two opposing schools of thought. This is normally evident from the graphical representation of assessments. However, by combining the experts we conglomerate the probabilities into one assessment which can be within one of the schools of thought or somewhere in between both, dependent on the expert weights. In scenarios such as this we can utilise our robustness checks to determine the impact on the variables of interest of using the combined probability or either of the disjoint schools, and feedback this uncertainty to the experts and model owner. This will enable the user of the model to make informed decisions, founded upon reliable and robust models.

2.3 Discussion

Although we can find a requisite structure for a BN, even the topology of a BN is not typically known with any certainty. Individuals are uncertain which variables in the system are crucial and which have little effect on the target variables. Different parties will have a variety of different opinions and forming a structure to satisfy all panels involved in the creation of a system can be a daunting task. Usually a broad BN template is introduced, around which ideas can be shared.

When it comes to specifying the necessary probabilities in CPTs, as well as the actual uncertainty inherent in the expert's quoted value, results are often aggregated in a variety of ways to provide a single value. The aggregated probability really *cannot* be treated as known. At best the calculated aggregated values are surrogates for a *set* of values which are deemed close, in some sense, to the true quantity. Sensitivity analyses, or robustness studies are therefore a crucial component of creating a justifiable and usable BN model. In Bayesian robustness studies, variation distances should appear to be a clear choice since; models are comparable, or a close match, if their outputs and recommended decisions are similar. In this thesis I systematically develop a method which is able to perform this in a flexible manner throughout the creation process.

Chapter 3

Applications and Examples

3.1 Food Security IDSS

Since the 2008 and 2011 global food price spikes, there has been a recent surge in the desire for models that provide decision support for crisis management in the food sector. The issue of food security and food scarcity affects all nations, and government agencies require transparent and scientifically sound decision support systems to aid policy creation. During the last decade the issue of food security has steadily become a pressing concern for many governments across the globe. Although many wealthy nations such as the U.K. are currently classed as having a high level of food security, the vital challenge of ensuring that the world's population has adequate food resources is not restricted to third world countries.

In 2008, the world suffered a crisis of an unprecedented increase of food prices which saw the Food and Agriculture Organization (FAO) of the United Nations record a cereal food price index of 232.1 compared to a value of 85.8 in 2000 [see FAO, 2014]. Another unanticipated spike occurred in 2011, creating further crises in several countries. Since these emergencies, many wealthy nations have begun to undertake an exploration into the complex study of food scarcity in an attempt to tackle the problem, which is linked to the ever expanding world population. Although definitions are similar worldwide, the U.K.'s Department for Environment, Food and Rural Affairs (DEFRA) uses that provided by the United Nations FAO: "food security is when all people, at all times, have physical and economical access to sufficient, safe and nutritious food to meet their dietary needs and food preferences for an active and healthy life" [House of Commons - Environment, Food and Rural Affairs Committee, 2014]. Within this statement are demands that future generations are similarly provided for; implying that food should be produced in an environmentally sustainable manner and, if possible, in a way which is also resistant to catastrophes.

There have been numerous attributes linked to the 2008 and 2011 food price spikes which include, but are not limited to: weather factors affecting crop yields, rising oil costs impacting cost of production and transportation of goods, policy measures hindering the flow of food across nation's borders and as well as any combination of the above [Meacham, 2014]. Due to the dynamic nature of these influences, the system is highly volatile implying that the food industry is susceptible to another food price spike at any time in the immediate future. Governments wish to determine both if and when another price spike could occur, with the intention of planning for such a scenario. Decision support systems aim to assist decision makers by modelling the underlying structure and mechanisms in the system of interest and recommending appropriate utility functions and strategies to secure robust and validated advice. The food security model focussed on within this thesis is an ongoing project developed alongside U.K. local councils to determine which policies should be funded to obtain the optimum solution for helping those in, or close to, food poverty in their county. Due to the complex variety of environmental, economic and social influences, as well as the dynamic nature of food production and pricing, this model becomes quite complex. However, it fulfils the criteria to be represented as a Bayesian network [see Barons et al., 2018b].

3.1.1 CPI Food Group BNs

My initial exposure to the field of Food Security in BNs was through a mini-project modelling the availability of fruit within the U.K. market, see Figure 3.1. The Consumer Price Index (CPI) is a device introduced by the U.K. Office for National Statistics to measure the fluctuations in typical expenditure of an average household. A 'shopping basket' of goods and services is categorised into multiple components and aids in monitoring price inflation. Individual BNs representing categories from the CPI basket had previously been devised, and my contribution of a fruit BN was intended to be systematically combined with the other models to create an IDSS allowing temporal effects in the CPI prices to be quantified. Amalgamating this information with data on household income would hopefully enable decision makers to determine all those who could be classed as being close to the threshold of food poverty.

Previous Masters students at the University of Warwick have created BNs for: the U.K. sugar supply [Barons et al., 2014]; the U.K. fish supply [Jones, 2015]; the U.K. meat industry [Murins, 2015]; a feasibility study into the applicability of BNs to the apple industry [Fenlon]. Work has also been undertaken [Collins, 2016] to compare these distinct models and establish whether there is a generic template underlying all the food industries, which would help to construct other food group BNs, such as dairy products, vegetables, non-alcoholic beverages and so forth, at a much faster rate.

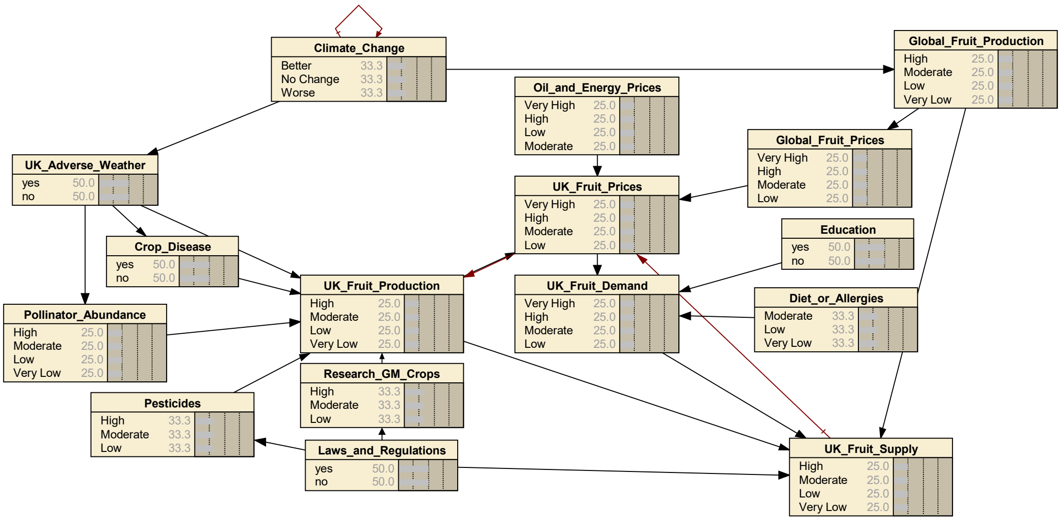


Figure 3.1: A DBN representing the U.K. Fruit Market as created for my OxWaSP mini-project - structure only.

3.1.2 Overarching Food Security IDSS

The overarching IDSS for the Food Security BN model has been discussed in great depth in Barons et al. [2018b] and encompasses the entire process underlying food prices and availability within the U.K. Figure 3.2 maps all relationships between features, with red arcs denoting dynamic effects. However, it is important to note that each node in this IDSS represents a sub-network, giving us a hierarchical DBN.

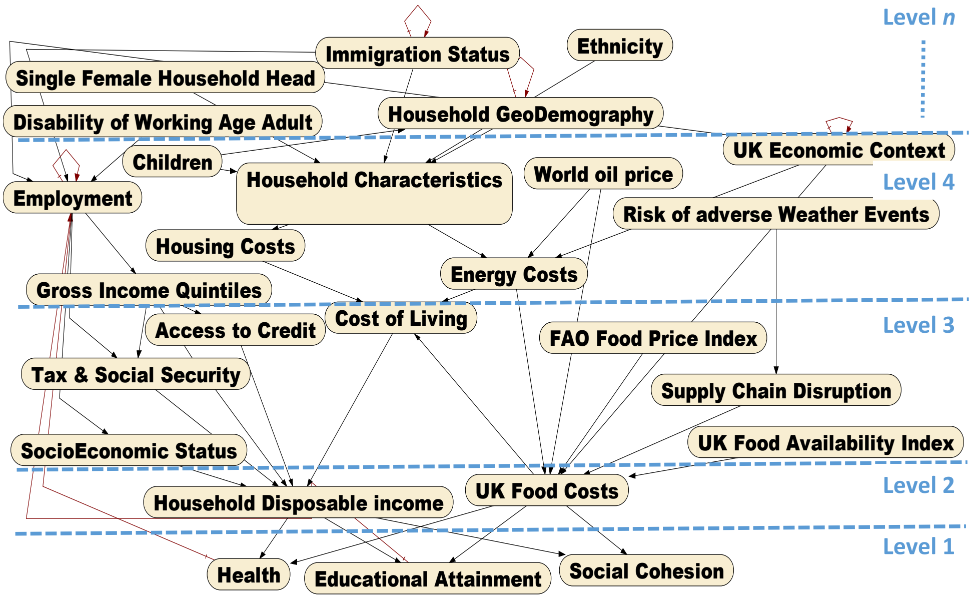


Figure 3.2: Food Security IDSS, red arcs indicating dynamic relationships, from Barons et al. [2018b].

Smith et al. [2015] observe that in large complex systems, such as Food Security applications, there is no single person who holds all the information required to form the model. To remedy this problem they introduced the concept of Integrated Decision Support Systems (as discussed in Section 2.1.2). These enable the user to segment the model into conditionally independent components and then distribute each section to the relevant panel of domain experts, for populating the whole system. Once completed, a coherent method is given for reformulating the method into one framework, to deduce scores for different situations.

Smith et al. [2015] also present an application of the same methodology to the problem of food poverty in Warwickshire County Council, including a discussion on the scoring rule drawn from the council to help deduce which policies to potentially enforce in the locality for maximum effect. Suppose Y_1 denote measures of educational attainment, Y_2 measures of health, Y_3 denote measures of social unrest and Y_4 denote cost. Let $y = (y_1; y_2; y_3; y_4)$ and define the utility function as

$$U(y) = \sum_{i \in [3]} k_i(1 - exp(-c_i y_i)) + k_4(a + by_4),$$

where the parameters $(a; b; c_1; c_2; c_3)$ were obtained from the relevant experts. There are now four attributes of interest which we label the target variables in any subsequent analyses of this model. This concentration of attention on a small subset of the variables enables users to tailor their robustness and sensitivity analyses to determine what effect errors in the system have on only the aforementioned target variables, rather than all variables. This is not always the case. In some fields, such as paternity cases, every variable (representing an individual person) in the BN could be a target node. Therefore robustness needs to be studied for the whole system and we tend to find that the more variables included in the model, the less robust the system. In the types of scenarios we are dealing with in this thesis, the focus of robustness will be robustness as it applies to a specific set of target variables. Note that in Figure 3.2 the nodes of interest are grouped within Level 1; parents are then labelled Level 2; and so forth helping to highlight those attributes which have direct relationships to these target variables.

3.1.3 Pollinator DBN

The overarching IDSS discussed above is a DBN model with four specific nodes of interest. Each component of this model can be broken down into more detailed subnetworks. For example, in Level 2, the variable 'U.K. Food Costs' depends on the availability of food, production costs and so on. Specifically, interest may focus on access to nutritious and healthy necessities such as fruit and vegetables which rely heavily on crop yield. Recent news bulletins have highlighted the impact that a dwindling bee population has on the abundance of crops, hence pollinator abundance is a critical feature which must be included when ascertaining the relationships effecting crop production. A sub-subnetwork of the overarching IDSS is that of the pollinator abundance DBN shown in Figure 3.3. In the OOBN jargon, these subnetworks and sub-subnetworks can be thought of as objects within the large overarching IDSS.

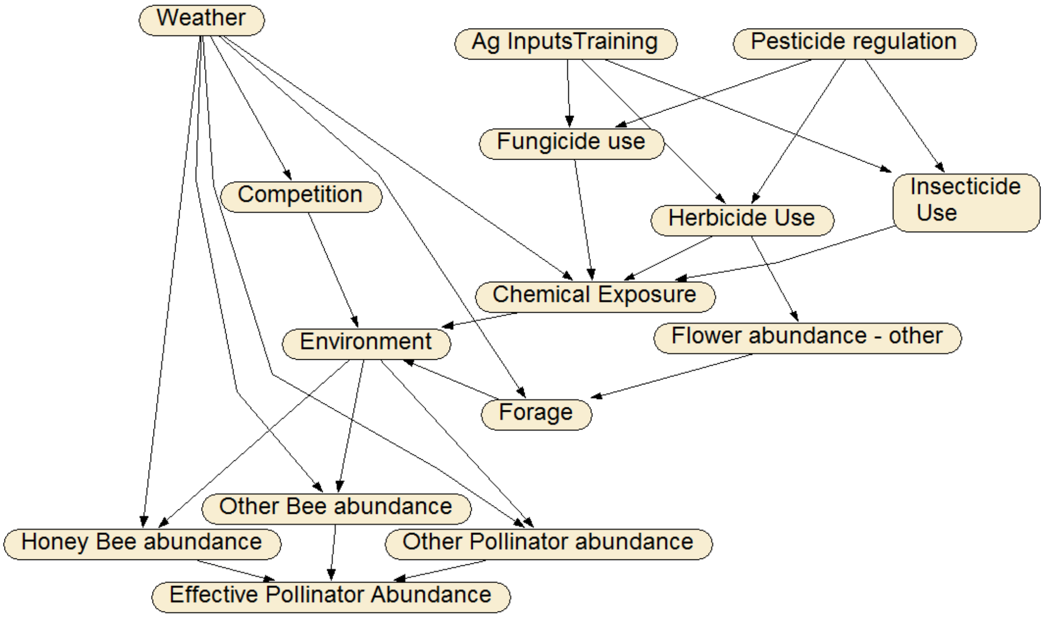


Figure 3.3: Fragment of the pollinator abundance BN sub-subnetwork, from Barons et al. [2018b].

A six-node subset of this network has been presented to experts, resulting in the elicitation of three conditional probability tables as documented in Barons et al. [2018a] and summarised in Table 3.1. These nodes were chosen due to the lack of quantitative information available in the apicultural domain and the relationships elicited are shown in Figure 3.4. For further details of the elicitation process see Section 2.2.2 of this thesis.

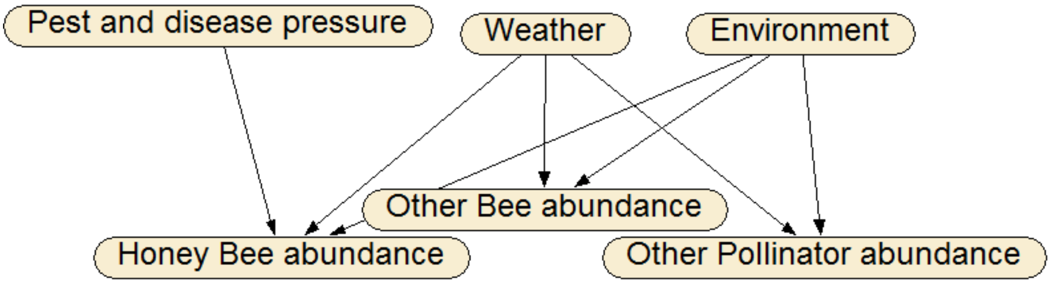


Figure 3.4: The six-node subset of the Pollinator BN elicited from experts. Results and picture from Barons et al. [2018a]

The subjective probabilities were mathematically aggregated using calibration responses as shown below, as presented in Barons et al. [2018a]:

Environment	Weather	Varroa Control	Probability abundance is good		
			Honey Bees	Other Bees	Hover Flies
Supportive	Average	Good	$0.77(0.57, 0.89)$	$0.73(0.48, 0.87)$	$0.71(0.48, 0.87)$
Supportive	Average	Poor	$0.27(0.16, 0.45)$	$0.73(0.49, 0.87)$	$0.71(0.48, 0.87)$
Supportive	Unusual	Good	$0.52(0.29, 0.76)$	$0.47(0.29, 0.73)$	$0.51(0.32, 0.71)$
Supportive	Unusual	Poor	$0.24(0.13, 0.44)$	$0.47(0.29, 0.73)$	$0.51(0.32, 0.71)$
Unsupportive	Average	Good	$0.38(0.21, 0.59)$	$0.21(0.11, 0.42)$	$0.25(0.12, 0.43)$
Unsupportive	Average	Poor	$0.14(0.07, 0.29)$	$0.21(0.11, 0.42)$	$0.25(0.12, 0.43)$
Unsupportive	Unusual	Good	$0.33(0.15, 0.51)$	$0.18(0.07, 0.41)$	$0.17(0.06, 0.37)$
Unsupportive	Unusual	Poor	$0.11(0.05, 0.23)$	$0.18(0.07, 0.41)$	$0.17(0.06, 0.37)$

Table 3.1: Best estimate (lowest, highest estimates) elicitation results as presented in Barons et al. [2018a].

3.2 Native Fish Abundance BN

To illustrate our robustness methods a well-known ecological BN, the Native Fish example, is analysed as introduced in Nicholson et al. [2010] and discussed further in Nicholson and Flores [2011]. It is important to note that this BN has been designed specifically for demonstration purposes, notably introducing non-statisticians to BNs, and is therefore a simplified version of a much more complicated process containing other domain information robustly expressed through a BN. However, due to the transparency of variable meanings, relationships and the manageable size of the BN it is large enough to illustrate our proposed methodology without becoming too complex. Within the initial report there are a number of variations of this model presented, only Version 2 is considered here due to the discrete nodes and the manageable size: a 10-node discrete network accompanied by full CPT tables, where each variable is either binary or three-state.

This BN is used to model the impact on a specific species of native fish abundance within a specific river, taking into consideration factors such as pesticide usage on surrounding fields, quantity of rainfall and so forth. The variables are briefly described in Table 3.2. For a more detailed discussion on assumptions, choice of node levels and CPT specification see the original technical report Nicholson et al. [2010].

Variable	Description	States
ENSO	El Niño Southern Oscillation changes to South Pacific sea conditions	{El Niño, Neutral, La Niña}
Pesticide Use	Quantity of pesticide currently used in the river location	{High, Low}
Drought Conditions	Historical indicator of drought status in area over past years	{Yes, No}
Annual Rainfall	Year-to-date quantity of rainfall	{Below Average, Average, Above Average}
Crop Yield	Amount of crops produced locally	{High, Low}
Irrigation	River water redirected to crop fields	{Yes, No}
Tree Condition	Health of trees on bank of river	{Good, Damaged, Dead}
Pesticide in River	The concentration of pesticide in the river water	{High, Low}
River Flow	Quality and speed of river flow	{Good, Poor}
Native Fish Abundance	Quantity of native fish residing in river	{High, Medium, Low}

Table 3.2: Node Names, Descriptions and Levels for the Version 2 Native Fish example as presented originally in Nicholson et al. [2010].

The structure of the BN can be seen below in Figure 3.5 and our node of interest is ‘Native Fish Abundance’; therefore interest focuses on structural and probabilistic variations altering the value of this variable.

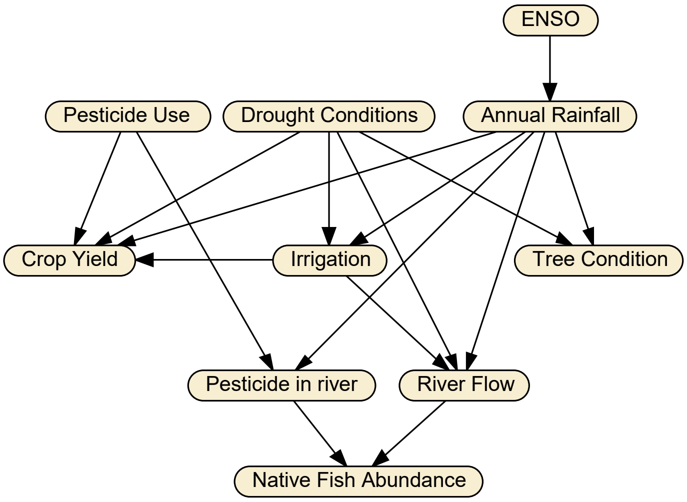


Figure 3.5: Native Fish Version 2 BN structure, from Nicholson et al. [2010].

CPTs for each variable are taken directly from Nicholson et al. [2010] and are included below for completeness:

Pesticide Use	
High	Low
0.9	0.1

Drought	
Yes	No
0.25	0.75

ENSO		
El Niño	Neutral	La Niña
0.2	0.6	0.2

ENSO	**P(Rainfall|ENSO)**		
	Below avg.	Average	Above avg.
El Niño	0.75	0.15	0.10
Neutral	0.10	0.80	0.10
La Niña	0.10	0.15	0.75

Pesticide in River	**River Flow**	**P(FishAbundance| PesticideInRiver, RiverFlow)**		
		High	Medium	Low
High	Good	0.20	0.40	0.40
High	Poor	0.01	0.10	0.89
Low	Good	0.80	0.15	0.05
Low	Poor	0.05	0.15	0.80

Drought Conditions	**Annual Rainfall**	**P(TreeCondition| Drought, Rainfall)**		
		Good	Damaged	Dead
Yes	Below avg.	0.20	0.60	0.20
Yes	Average	0.25	0.60	0.15
Yes	Above avg.	0.30	0.60	0.10
No	Below avg.	0.70	0.25	0.05
No	Average	0.80	0.18	0.02
No	Above avg.	0.90	0.09	0.01

Pesticide Use	**Annual Rainfall**	**P(PesticideInRiver| PesticideUse, Rainfall)**	
		High	Low
High	Below avg.	0.30	0.70
High	Average	0.60	0.40
High	Above avg.	0.80	0.20
Low	Below avg.	0.10	0.90
Low	Average	0.20	0.80
Low	Above avg.	0.30	0.70

	Annual	P(Irrigation|	
		Drought, Rainfall)	
Drought	**Rainfall**	Yes	No
Yes	Below avg.	0.01	0.99
Yes	Average	0.10	0.90
Yes	Above avg.	0.25	0.75
No	Below avg.	0.95	0.05
No	Average	0.50	0.50
No	Above avg.	0.20	0.80

Annual		**Pesticide**		**P(RiverFlow|**	
				Drought, Rainfall, Irrigation)	
Rainfall	**Drought**	**Use**	**Irrigation**	High	Low
Below avg.	Yes	High	Yes	0.20	0.80
Below avg.	Yes	High	No	0.02	0.98
Below avg.	Yes	Low	Yes	0.15	0.85
Below avg.	Yes	Low	No	0.01	0.99
Below avg.	No	High	Yes	0.50	0.50
Below avg.	No	High	No	0.20	0.80
Below avg.	No	Low	Yes	0.40	0.60
Below avg.	No	Low	No	0.15	0.85
Average	Yes	High	Yes	0.30	0.70
Average	Yes	High	No	0.10	0.90
Average	Yes	Low	Yes	0.25	0.75
Average	Yes	Low	No	0.05	0.95
Average	No	High	Yes	0.80	0.20
Average	No	High	No	0.50	0.50
Average	No	Low	Yes	0.60	0.40
Average	No	Low	No	0.30	0.70
Above avg.	Yes	High	Yes	0.80	0.20
Above avg.	Yes	High	No	0.50	0.50
Above avg.	Yes	Low	Yes	0.60	0.40
Above avg.	Yes	Low	No	0.30	0.70
Above avg.	No	High	Yes	0.99	0.01
Above avg.	No	High	No	0.95	0.05
Above avg.	No	Low	Yes	0.95	0.05
Above avg.	No	Low	No	0.70	.30

Drought	**Annual Rainfall**	**Irrigation**	**P(RiverFlow| Drought, Rainfall, Irrigation)**	
			Good	Poor
Yes	Below avg.	Yes	0.01	0.99
Yes	Below avg.	No	0.05	0.95
Yes	Average	Yes	0.05	0.95
Yes	Average	No	0.15	0.85
Yes	Above avg.	Yes	0.70	0.30
Yes	Above avg.	No	0.80	0.20
No	Below avg.	Yes	0.30	0.70
No	Below avg.	No	0.40	0.60
No	Average	Yes	0.50	0.50
No	Average	No	0.60	0.40
No	Above avg.	Yes	0.90	0.10
No	Above avg.	No	0.99	0.01

Although designed for demonstration purposes only, note that the probabilities in these CPTs are typical of those we usually see in an elicitation scenario; many of the entries are close to neither 0 nor 1. Some rows of the CPTs seem similar (for example row 2 and row 3 in the River Flow CPT are identical) and it is likely therefore, that the user might query whether it would be appropriate to simplify the process in some manner, such as deleting an arc or amalgamating levels. Recall that in an elicitation procedure with multiple experts contributing beliefs, the probabilities in the CPT are an aggregated value representing the approximate beliefs of the group. We therefore have uncertainty relating to the discrepancy between each individuals' stated quantity and the aggregated quantity. Additionally, we have the actual uncertainty of the experts.

Chapter 4

Total Variation Distance

Robustness studies aim to distinguish between models which can change drastically when underlying conditions or information is altered. In this thesis we are interested only in the effect on a pre-specified subset of variables, termed the target variables, rather than the whole joint density/mass function. To compare the outputs of a model pre and post perturbation, distance measures can be calculated to quantify the difference between the joint distributions of the target variables.

There are a multitude of different statistical distances which measure the discrepancy between two probability distributions. Consideration of the application often leads to a choice between fewer recommended alternatives. Recall that a distance metric by definition must fulfil four criteria:

- it must always be non-negative (i.e. $d(p,q) \geq 0$),
- the triangle inequality holds (i.e. $d(p,q) \leq d(p,r) + d(r,q)$),
- the distance can only be zero if the two densities are equivalent (i.e. $d(p,q) = 0 \Rightarrow p = q$)
- and it must be symmetric (i.e. $d(p,q) = d(q,p)$).

A general guide for choosing between competing measures has been written by Gibbs and Su [2002]. Recommendations for selecting a distance measure would begin by critiquing models previously created in relevant fields, for example recent research regarding discrete BNs have employed the Kullback-Leibler distance [see Gómez-Villegas et al., 2013; Leonelli et al., 2017] or the Chan-Darwiche distance [see Chan and Darwiche, 2005, 2004] which we discuss below. Alternatively, we can opt for the metric which provides the most logical interpretation for the intended application, for example in an expert elicitation setting we must ensure that non-statisticians can readily understand the definition and their uses. Finally, much has been written about the links between different statistical distances using bounds, therefore allowing us to use alternative metrics (which may be easier to compute) in order to tighten a bound on our chosen metric.

Although prior work has been done in the BN field using a number of alternative distance measures, we argue that for our needs, the total variation distance has some simple, yet elegant properties which make it a more appropriate choice. In this section we briefly introduce and review the total variation distance, highlighting those properties of most importance to our later work.

4.1 Definition

Assume $\boldsymbol{X} \triangleq (X_1, X_2, \ldots, X_m)$ is a vector of finite discrete random variables taking values $\boldsymbol{x} = (x_1, x_2, \ldots, x_m) \in \mathbb{X}_1 \times \mathbb{X}_2 \times \cdots \times \mathbb{X}_m$. Let $\boldsymbol{X}_A$ (taking values $\boldsymbol{x}_A \in \mathbb{X}_{i_1} \times \mathbb{X}_{i_2} \times \cdots \times \mathbb{X}_{i_{r(A)}} \triangleq \mathbb{X}_A$) denote the subvector of $\boldsymbol{X}$ comprising those components with indices $i \in A$, where $A = \left\{i_1, i_2, \ldots, i_{r(A)}\right\}$ denotes a subset of $\{1, 2, \ldots, m\}$. Let p_A, q_A denote a hypothesised and an alternative joint mass function on $\boldsymbol{X}_A$. Let $\mathbb{P}_A(E), \mathbb{Q}_A(E)$ denote the probability with respect to the mass functions p_A, q_A respectively of the set $E = E_{i_1} \times E_{i_2} \times \cdots \times E_{i_{r(A)}}$ where $E_{i_j} \subseteq \mathbb{X}_{i_j}$, $j = 1, 2, \ldots r(A)$. Note that most of the theory of BNs concerns inference and propagation on such events: see e.g. Smith [2010].

Definition 4.1.1. For the discrete case, the *(total) variation distance*, $d_V(p_A, q_A)$, is defined by

$$d_V(p_A, q_A) \triangleq \frac{1}{2} \sum_{\boldsymbol{x}_{\boldsymbol{A}} \in \mathbb{X}_A} |p_A - q_A|$$

The total variation distance is widely used by probabilists as it is relatively straightforward to prove results due to the definition formula and consequently there is a plethora of information about its properties. An important aspect of variation distance when applied in decision support systems, is that for a fixed bounded utility function U, disparities in its expectations under various decisions are bounded by linear functions of the total variation distance, see Smith [2010] for more detail and proofs. Therefore we can deduce that if there is a small variation distance between two probability distributions then there will be a corresponding small discrepancy between the expected utilities.

4.2 General Properties

4.2.1 Links to other distance metrics

A metric related to the total variation distance is the Hellinger distance:

Definition 4.2.1. The *Hellinger distance*, $d_H(p_A, q_A)$ is defined as

$$d_H(p_a, q_A) \triangleq \sqrt{1 - \sum_{x \in \mathbb{X}_A} \sqrt{p_A q_A}}.$$

For all mass functions p and q it is shown [for example see Reiss, 1989] that

$$d_H^2(p,q) \leq d_V(p,q) \leq \sqrt{2} d_H(p,q).$$

These inequalities demonstrate that when p and q are close in variation distance then they are close in Hellinger distance and vice versa. This allows us to call these two distances *topologically equivalent.* These inequality bounds linking total variation and Hellinger are important because it is not always possible to write explicit closed expressions for the total variation distance, whereas for Hellinger distances we can.

Another popular distance measure is the Kullback-Leibler measure:

Definition 4.2.2. The *Kullback-Leibler* divergence, $d_{KL}(p,q)$ is discretely defined as

$$d_{KL}(p,q) \triangleq \sum_i p(i) \frac{p(i)}{q(i)}.$$

Note that although often called a distance, the Kullback-Leibler (KL) measure does not satisfy the symmetry condition. KL is commonly seen in practice because, similar to the Hellinger distance, it has a closed form property. Leonelli et al. [2017] review sensitivity analyses in discrete BNs under single or multiple simultaneous parameter changes, noting that KL is a popular choice when distributions are assumed to be Gaussian. They also argue that the KL divergence can often become difficult to compute because it belongs to the ϕ-divergence family. They advocate an alternative measure which is the final measure we discuss in this thesis: the Chan-Darwiche measure:

Definition 4.2.3. The *Chan-Darwiche* distance, $d_{CD}(p,q)$ is discretely defined as

$$d_{CD}(p,q) \triangleq \log \max_i \frac{q(i)}{p(i)} - \log \min_i \frac{q(i)}{p(i)},$$

with $\frac{0}{0}$ defined as 1.

Although the Kullback-Leibler and Chan-Darwiche distance have both been applied in sensitivity analyses of BNs, their suitability can be contended. Firstly, both of these measures are defined in terms of log probabilities or ratios of probabilities. This implies that the modeller will suffer heavy penalties if they cannot accurately specify very small probabilities. For example suppose we have the following three probability distributions:

$$\begin{aligned}\pi_1 &= (0.2, 0.8 - 10^{-8}, 10^{-8}),\\ \pi_2 &= (0.2, 0.8 - 10^{-2}, 10^{-2}),\\ \pi_3 &= (0.5, 0.5 - 10^{-2}, 10^{-2}).\end{aligned}$$

Then we can calculate

$$\begin{array}{llr} d_V(\pi_1, \pi_2) & = 0.00999999, & d_{CD}(\pi_1, \pi_2) = 13.82809,\\ d_V(\pi_2, \pi_3) & = 0.3, & d_{CD}(\pi_2, \pi_3) = 1.393918.\end{array}$$

The variation distance therefore agrees with our intuitive deduction that π_2 is more similar to π_1 than π_3, whereas the Chan-Darwiche measure has inflated the small distance between almost identical distributions. This illustrates that the Chan-Darwiche is not a suitable distance measure when dealing with small probabilities in conditional probability tables. When using BNs within decision support systems the modeller must ensure that the BN is transparent and understandable to the user or the decision-maker will not inherit the model. Unlike the Kullback-Leibler and the Chan-Darwiche measures, which are defined using log probabilities or ratios of probabilities, the total variation distance is a more natural measure of discrepancy which is simple to explain and demonstrate to non-statisticians. Being able to anticipate whether two probability distributions are similar or disparate can speed up elicitation and CPT population: as discussed later in this thesis, see Section 5.

One extremely important property of variation distances is that if p_A and q_A are close, then this not only ensures that expected utilities associated with different decisions are close, but that whenever they are far apart some expected utilities functions will be far apart too, Smith and Daneshkhah [2010]. This joint property is not shared by competing divergences such as Chan-Darwiche or Kullback-Leibler. This critical, but often overlooked property implies that by choosing to approximate a specified Bayesian network with another different BN (in a way that minimises variation distance) will be one of the best approximations if we wish to ensure that expected utility evaluations are as close as possible. We use this idea many times throughout this work.

4.2.2 Convexity of Total Variation

A useful and well-known property of total variation is its convexity under mixing, however to prove this result we require some simple bounds first. To our knowledge the properties of total variation distance presented in this section have not yet been explored, and therefore all lemmas and proofs are our own work.

Let $\mathbf{p}, \mathbf{q}_0, \mathbf{q}_1$ be three vectors of mass functions and define

$$\mathbf{q}_\alpha \triangleq (1-\alpha)\mathbf{q}_0 + \alpha\mathbf{q}_1,$$

for $0 \leq \alpha \leq 1$.

Lemma 4.2.1.

$$d_V(\mathbf{p}, \mathbf{q}_\alpha) \leq (1-\alpha) d_V(\mathbf{p}, \mathbf{q}_0) + \alpha d_V(\mathbf{p}, \mathbf{q}_1).$$

Proof. Note that for $0 \leq \alpha \leq 1$,

$$\mathbf{p} = (1-\alpha)\mathbf{p} + \alpha\mathbf{p}$$

so that

$$\begin{aligned} 2d_V(\mathbf{p}, \mathbf{q}_\alpha) &= \sum_{j=1}^{n} |p_j - (1-\alpha)q_{0j} - \alpha q_{1j}| \\ &= \sum_{j=1}^{n} |(1-\alpha)p_j + \alpha p_j - (1-\alpha)q_{0j} - \alpha q_{1j}| \\ &= \sum_{j=1}^{n} |(1-\alpha)(p_j - q_{0j}) + \alpha(p_j - q_{1j})| \,. \end{aligned}$$

Next note that if $(p_j - q_{0j}), (p_j - q_{1j})$ are the same sign then

$$|(1-\alpha)(p_j - q_{0j}) + \alpha(p_j - q_{1j})| = (1-\alpha)\,|p_j - q_{0j}| + \alpha\,|p_j - q_{1j}| \,,$$

whilst if $(p_j - q_{0j}), (p_j - q_{1j})$ are of different sign then

$$\begin{aligned}\left|(1-\alpha)(p_j - q_{0j}) + \alpha(p_j - q_{1j})\right| &\le \max\left\{(1-\alpha)\left|p_j - q_{0j}\right|, \alpha\left|p_j - q_{1j}\right|\right\} \\ &\le (1-\alpha)\left|p_j - q_{0j}\right| + \alpha\left|p_j - q_{1j}\right|.\end{aligned}$$

It follows that

$$\begin{aligned}2d_V(\mathbf{p}, \mathbf{q}_\alpha) &\le \sum_{j=1}^{n}(1-\alpha)\left|p_j - q_{0j}\right| + \alpha\left|p_j - q_{1j}\right| \\ &= (1-\alpha)\sum_{j=1}^{n}\left|p_j - q_{0j}\right| + \alpha\sum_{j=1}^{n}\left|p_j - q_{1j}\right| \\ &= 2\left\{(1-\alpha)d_V(\mathbf{p}, \mathbf{q}_0) + \alpha d_V(\mathbf{p}, \mathbf{q}_1)\right\},\end{aligned}$$

proving the result. □

Hence, the variation distance between a mass function **p** and the mixture of two others, is less than the mixture of the mass function **p** and the two extremal distributions $\mathbf{q}_0, \mathbf{q}_1$. This suggests that distances between two mass functions of receiving variable functions, is no greater than the distance between the two donating mass functions, if propagated under the same CPT.

We can now extend this result so that it applies to any finite mixtures: let $\pi = (\pi_1, \pi_2, \ldots, \pi_n)$ with $\sum_{i=1}^{n} \pi_i = 1$, $\mathbf{q}_i = (q_{i1}, q_{i2}, \ldots, q_{in})$ and define

$$\mathbf{q}_\pi \triangleq \sum_{i=1}^{n} \pi_i \mathbf{q}_i.$$

Lemma 4.2.2.

$$d_V(\mathbf{p}, \mathbf{q}_\pi) \le \sum_{i=1}^{n} \pi_i d_V(\mathbf{p}, \mathbf{q}_i).$$

Proof. We shall proceed using induction. For the case $n = 2$ we have $\mathbf{q}_\pi = \pi_1 q_1 + \pi_2 q_2$ with $\pi_1 + \pi_2 = 1$ indicating that $\mathbf{q}_\pi = (1 - \pi_2)q_1 + \pi_2 q_2$. By applying the previous lemma we know that

$$\begin{aligned}d_V(p, \mathbf{q}_\pi) &\le (1-\pi_2)d_V(p, q_1) + \pi_2 d_V(p, q_2) \\ &= \pi_1 d_V(p, q_1) + \pi_2 d_V(p, q_2) \\ &= \sum_{i=1}^{2} \pi_i d_V(p, q_i).\end{aligned}$$

So this statement certainly holds true for $n = 2$. Assume that it is true for $n = k$. Then note that for the case $n = k + 1$:

$$\begin{aligned}
\mathbf{q}_\pi = \sum_{i=1}^{k+1} \pi_i \mathbf{q}_i &= \pi_1 \mathbf{q}_1 + \pi_2 \mathbf{q}_2 + \ldots + \pi_k \mathbf{q}_k + \pi_{k+1} \mathbf{q}_{k+1} \\
&= (\pi_1 + \pi_2 + \ldots + \pi_k)\mathbf{q}_0 + \pi_{k+1}\mathbf{q}_{k+1} \\
&= (1 - \pi_{k+1})\mathbf{q}_0 + \pi_{k+1}\mathbf{q}_{k+1} \\
&= (1 - \alpha)\mathbf{q}_0 + \alpha \mathbf{q}_{k+1},
\end{aligned}$$

where $\alpha \triangleq \pi_{k+1}$ and for $i = 1, 2, \ldots k$, $\pi_i' \triangleq \pi_i (1 - \pi_{k+1})^{-1}$, giving

$$\mathbf{q}_0 \triangleq \sum_{i=1}^{k} \pi_i' \mathbf{q}_i = \frac{1}{1 - \pi_{k+1}} \sum_{i=1}^{k} \pi_i q_i.$$

By Lemma 4.2.1

$$d_V(\mathbf{p}, \mathbf{q}_\pi) \leq (1 - \alpha) d_V(\mathbf{p}, \mathbf{q}_0) + \alpha d_V(\mathbf{p}, \mathbf{q}_{k+1}),$$

where by the inductive hypothesis

$$d_V(\mathbf{p}, \mathbf{q}_0) \leq \sum_{i=1}^{k} \pi_i' d_V(\mathbf{p}, \mathbf{q}_i).$$

Substituting the inductive hypothesis into our equation gives

$$\begin{aligned}
d_V(\mathbf{p}, \mathbf{q}_\pi) &\leq \sum_{i=1}^{k} (1 - \alpha) \pi_i' d_V(\mathbf{p}, \mathbf{q}_i) + \alpha d_V(\mathbf{p}, \mathbf{q}_{k+1}) \\
&= \sum_{i=1}^{k+1} \pi_i d_V(\mathbf{p}, \mathbf{q}_i)
\end{aligned}$$

as required. □

Finally, we can look at the convexity of total variation in the following sense. Let $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_n)$, $\boldsymbol{\pi'} = (\pi_1', \pi_2', \ldots, \pi_{n'}')$, $\boldsymbol{q}_i = (q_{i1}, q_{i2}, \ldots, q_{in})$, $\boldsymbol{p}_i = (p_{i1}, p_{i2}, \ldots, p_{in'})$ and define

$$\boldsymbol{q}_{\boldsymbol{\pi}} \triangleq \sum_{i=1}^{n} \pi_i \boldsymbol{q}_i, \qquad \boldsymbol{p}_{\boldsymbol{\pi}'} \triangleq \sum_{i=1}^{n'} \pi_i' \boldsymbol{p}_i,$$

then we have the following property

Lemma 4.2.3 (Convexity of total variation distance)**.**

$$d_V(\boldsymbol{p}_{\boldsymbol{\pi}'}, \boldsymbol{q}_{\boldsymbol{\pi}}) \leq \sum_{i=1}^{n} \sum_{i'=1}^{n'} \pi_i \pi'_{i'} d_V(\boldsymbol{p}_{i'}, \boldsymbol{q}_i).$$

Proof. Here we simply use the symmetry of variation distance. From Lemma 4.2.2 we have that

$$d_V(\mathbf{p}_{\pi'}, \mathbf{q}_{\pi}) \leq \sum_{i=1}^{n} \pi_i d_V(\mathbf{p}_{\pi'}, \mathbf{q}_i), \tag{4.1}$$

but for $i = 1, 2, \ldots, n$

$$d_V(\mathbf{p}_{\pi'}, \mathbf{q}_i) = d_V(\mathbf{q}_i, \mathbf{p}_{\pi'}) \leq \sum_{i'=1}^{n'} \pi'_i d_V(\mathbf{q}_i, \mathbf{p}_{i'}) = \sum_{i'=1}^{n'} \pi'_i d_V(\mathbf{p}_{i'}, \mathbf{q}_i). \tag{4.2}$$

Substituting Inequality 4.2 into Inequality 4.2 gives us our result. □

In particular we can now say that if we know extremal distributions are small, then convex linear combinations of these are all small: for example in the process of marginalisation. This enables us to prove a number of useful results concerning the contraction of error under learning in a BN: see below.

4.3 Introducing the Diameter

As variation distance is a probability measure it can only be applied to probability distributions; however, we are interested in comparing two matrices, for example alternative CPTs. In this section we define several new objects which will be pivotal in later development of these ideas.

Let P and Q, with rows $\boldsymbol{p}_i, \boldsymbol{q}_i$ respectively (for $i = 1, 2, \ldots, n$), be two $n \times n'$ CPT matrices of a random vector $\boldsymbol{Y}$ taking n' levels, given another random vector $\boldsymbol{X}$ taking n levels. For a BN, $\boldsymbol{Y}$ will typically be a random variable whilst $\boldsymbol{X}$ will be the vector of its parents; nevertheless when studying junction trees it is also helpful to consider cases when $\boldsymbol{Y}$ is a vector.

There is a natural variation distance we can now define between P and Q:

Definition 4.3.1. Let the variation distance $d_V^+(P, Q)$, $(d_V^-(P, Q))$, between conditional probability tables P and Q be defined by

$$d_V^+(P,Q) \triangleq \max_{1\leq i\leq n} d_V(\boldsymbol{p}_i,\boldsymbol{q}_i), \qquad d_V^-(P,Q) \triangleq \min_{1\leq i\leq n} d_V(\boldsymbol{p}_i,\boldsymbol{q}_i).$$

Example 4.3.1. As an example let us look at the CPT for the node 'Tree Condition' in the Native Fish example introduced in Section 3.2. In the original work (Nicholson et al. [2010]), the CPT is given as:

Drought Conditions	Annual Rainfall	$P(TreeCondition \mid Drought, Rainfall)$		
		Good	Damaged	Dead
yes	below average	0.2	0.6	0.2
yes	average	0.25	0.6	0.15
yes	above average	0.3	0.6	0.1
no	below average	0.7	0.25	0.05
no	average	0.8	0.18	0.02
no	above average	0.9	0.09	0.01

We now simplify this CPT to matrix form, denoted by P. Let us assume that this CPT was elicited from experts who disagree on a couple of probabilities resulting in an alternate CPT, Q:

$$P = \begin{pmatrix} 0.20 & 0.60 & 0.20 \\ 0.25 & 0.60 & 0.15 \\ 0.30 & 0.60 & 0.10 \\ 0.70 & 0.25 & 0.05 \\ 0.80 & 0.18 & 0.02 \\ 0.90 & 0.09 & 0.01 \end{pmatrix}, \qquad Q = \begin{pmatrix} 0.20 & 0.60 & 0.20 \\ 0.30 & 0.50 & 0.20 \\ 0.30 & 0.60 & 0.10 \\ 0.65 & 0.25 & 0.10 \\ 0.80 & 0.18 & 0.02 \\ 0.90 & 0.10 & 0.00 \end{pmatrix}.$$

We can now compute $d_V^+(P,Q) = \max\{0, 0.1, 0, 0.05, 0, 0.01\} = 0.1$ and immediately deduce that $d_V^-(P,Q) = 0$ because some rows of P are equivalent to rows of Q.

Note that if $\boldsymbol{\rho}(P)$ and $\boldsymbol{\rho}(Q)$ are the vectors of marginal mass functions of $\boldsymbol{Y}$ and $\boldsymbol{\pi}$ is a margin on $\boldsymbol{X}$ then for all possible margins $\boldsymbol{\pi}$

$$d_V(\boldsymbol{\rho}(P), \boldsymbol{\rho}(Q)) \leq d_V^+(P,Q),$$

where

$$d_V(\boldsymbol{\rho}(P), \boldsymbol{\rho}(Q)) = d_V^+(P,Q)$$

whenever $\boldsymbol{\pi}$ puts all its mass on atoms indexed by i^+ where

$$i^+ \triangleq \arg \max_{1 \leq i \leq n} d_V(\boldsymbol{p}_i, \boldsymbol{q}_i).$$

Similarly if $\boldsymbol{\rho}(P), \boldsymbol{\rho}(Q)$ are the vectors of marginal mass functions of Y and $\boldsymbol{\pi}$ is a margin on X then for all possible margins $\boldsymbol{\pi}$

$$d_V\left(\boldsymbol{\rho}(P), \boldsymbol{\rho}(Q)\right) \geq d_V^-\left(P, Q\right),$$

where

$$d_V\left(\boldsymbol{\rho}(P), \boldsymbol{\rho}(Q)\right) = d_V^-\left(P, Q\right)$$

whenever $\boldsymbol{\pi}$ puts all its mass on an atom or atoms indexed by i^- where

$$i^- \triangleq \arg \min_{1 \leq i \leq n} d_V(\boldsymbol{p}_i, \boldsymbol{q}_i).$$

Thus we have that for all possible margins $\boldsymbol{\pi}$

$$d_V^-\left(P, Q\right) \leq d_V\left(\boldsymbol{\rho}(P), \boldsymbol{\rho}(Q)\right) \leq d_V^+\left(P, Q\right).$$

This therefore gives rather coarse, but quick bounds requiring only comparisons of the pairs of individual rows of the perturbed CPT.

Previously, we pointed out that when eliciting a BN we first elicit hypotheses of conditional independence and only then enhance this with a full probability specification through the numerical values in its CPTs. So we next consider robustness measures associated with small deviations from conditional independence. The definition we present below is, to our knowledge, a new construction using variation distance on CPTs to determine the measure of dependence between variables.

Definition 4.3.2. The *diameter*, $d^+(P)$, the *I-local diameter* $d^{I+}(P)$ and the *dis-*

tinctiveness $d^-(P)$ of a stochastic matrix $P \triangleq \{p_{ij}\}$ are respectively defined as

$$\begin{aligned}
d^+(P) &= \frac{1}{2}\max_{1\le i,i'\le n}\left\{\sum_{j=1}^{n'}\left|p_{ij}-p_{i'j}\right|\right\}, \\
d^{I+}(P) &= \frac{1}{2}\max_{i,i'\in I}\left\{\sum_{j=1}^{n'}\left|p_{ij}-p_{i'j}\right|\right\}, \\
d^-(P) &= \frac{1}{2}\min_{1\le i\ne i'\le n}\left\{\sum_{j=1}^{n'}\left|p_{ij}-p_{i'j}\right|\right\}.
\end{aligned}$$

Example 4.3.2. We can now calculate the internal diameter of the CPT for the node 'Tree Condition' which is conditional upon the values of 'Drought' and 'Rainfall' as given in Example 4.3.1. We can directly calculate $d^+(P) = 0.7$ which is the distance between row 1 and row 6.

Using all the CPTs as given in Nicholson and Flores [2011] we can calculate the diameter of each node, as summarised in Table 4.1. We note that these values of the diameter are typical of the values we have obtained in other elicitation exercises.

Node	**Diameter**	**Node**	**Diameter**
Annual Rainfall	0.65	Crop Yield	0.98
River Flow	0.98	Irrigation	0.94
Pesticide in River	0.7	Tree Condition	0.7
Native Fish Abundance	0.84		

Table 4.1: Diameters of each CPT in the Native Fish BN.

For the elicitation workshop detailed in Barons et al. [2018a], three CPTs from Figure 3.3 were elicited from experts to determine pollinator abundance under specific conditions regarding pest control, environment suitability and weather anomalies (see Table 3.1). The diameter of each of these CPTs can easily be calculated to obtain the values in Table 4.2:

Honey Bee Abundance	0.66
Other Bee Abundance	0.55
Other Pollinator Abundance	0.54

Table 4.2: Diameters of each CPT in Pollinator sub-network BN.

The diameter of a conditional probability table P, can be thought of as a measure of the dependence of $\boldsymbol{Y}$ on $\boldsymbol{X}$. When $\boldsymbol{Y} \amalg \boldsymbol{X}$, all rows of P will be equal and so $d^{+}(P) = 0$. It is easy to check that whenever some non-trivial function $\tau(\boldsymbol{Y})$ of $\boldsymbol{Y}$ can be written as a deterministic function of $\boldsymbol{X}$ then $d^{+}(P) = 1$, its maximum value. Hence, if there is only a weak relationship between $\boldsymbol{Y}$ and $\boldsymbol{X}$, in the sense that changing the different levels of $\boldsymbol{X}$ impacts only slightly on the conditional mass function of $\boldsymbol{Y}$, then $d^{+}(P) \bumpeq 0$.

The I-local diameter has the same property; this time it is conditional on $\boldsymbol{X}$ taking values only in the set of levels I, useful when comparing the efficacy of deleting a parent in a BN or when combining a collection of rows of the CPT/levels of $\boldsymbol{X}$ into a single entry: see following Section.

The distinctiveness, $d^{-}(P)$, is a measure of how *little* different levels of $\boldsymbol{X}$ can impact on $\boldsymbol{Y}$. Thus if $d^{-}(P) = 0$ then two rows of P must be identical. So there is a non invertible function $\tau(\boldsymbol{X})$ of $\boldsymbol{X}$ which forecasts $\boldsymbol{Y}$ as well as $\boldsymbol{X}$, i.e. for which $\boldsymbol{Y} \amalg \boldsymbol{X} | \tau(\boldsymbol{X})$. On the other hand when $d^{-}(P) = 1$ every pair of possible levels i, i' of X for some non-trivial function $\tau_{ii'}(\boldsymbol{Y})$ of $\boldsymbol{Y}$, can be written as a deterministic function of these two values and so critically distinguishes two different features of the response vector.

Definition 4.3.3. Let the superbound, $d_V^{*}(P, Q)$, between stochastic matrices P and Q be defined by

$$d_V^{*}(P, Q) \triangleq \max_{1 \leq i, i' \leq n} d_V(\boldsymbol{p}_i, \boldsymbol{q}_{i'}) \leq 1.$$

Here we compare variation distances between each row of P and possibly different rows of Q before selecting the largest difference. Note by definition and the triangle inequality that

$$d_V^{+}(P, Q) \leq d_V^{*}(P, Q) \leq \max\left\{d_V^{+}(P, Q) + \max\left\{d^{+}(P), d^{+}(Q)\right\}, 1\right\}. \tag{4.3}$$

Example 4.3.3. Let us compare the two alternative CPTs, P and Q, for the 'Tree Condition' node as introduced in Example 4.3.1. The value of $d_V^*(P,Q)$ can be calculated directly from the total variation distance between every possible pairwise combination of rows in P and Q. For this example $d_V^*(P,Q) = 0.7$ corresponding to $d_V(\boldsymbol{p}_1, \boldsymbol{q}_6)$ and $d_V(\boldsymbol{p}_6, \boldsymbol{q}_1)$.

Let $P_{A|B}$, $Q_{A|B}$ represent respectively, the conditional probability mass functions of $\boldsymbol{X}_A|\ \boldsymbol{X}_B$ under the hypothesis and alternative given $\boldsymbol{X}_B$, where without loss we can assume that A and B are disjoint. Notice that these can be seen as CPTs whose rows correspond to the different values of the vector $\boldsymbol{x}_B$. Then under our definitions of transition matrices above whenever $\boldsymbol{X}_A \amalg\ \boldsymbol{X}_B$

$$d_V^-(P_{A|B}, Q_{A|B}) = d_V^+(P_{A|B}, Q_{A|B}) = d_V^*(P_{A|B}, Q_{A|B}) = d_V(p_A, q_A).$$

This arises as $\boldsymbol{X}_A \amalg \boldsymbol{X}_B$ implies that all rows in the CPT matrix are equal to each other and so equal to the corresponding margin on $\boldsymbol{X}_A$. Thus we see that standard analyses that elicit irrelevances or independences, translate here to equations on variation distance. We will see later in Section 5 that this enables us to study the implications of models where the embedded conditional independences are only approximately true.

Definition 4.3.4. The *stochastic variation matrix* $D^+(P) = \{D^+(i,j)\}_{1\leq i,j\leq n}$ is the $n \times n$ symmetric matrix whose entries are the variation distances between the different rows of the matrix P.

Note that by definition the leading diagonal of $D^+(P)$ is $\mathbf{0}$ because $d_V(p,p) = 0$. Additionally, $D^+(P)$ will have a maximum entry of $d^+(P)$.

4.3.1 Diameter Bounds

We now present some basic results about diameters of transition matrices between two vectors of random variables under various marginalisations and conditioning of the subvectors. These bounds are particularly helpful when moving from a BN to a junction tree.

Let $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2)$ and $\boldsymbol{Y} = (\boldsymbol{Y}_1, \boldsymbol{Y}_2)$. Let $P_{\boldsymbol{Y}|\boldsymbol{X}}$ be the transition matrix associated with the conditional distribution of $\boldsymbol{Y}|\boldsymbol{X}$ and let $P_{\boldsymbol{Y}|\boldsymbol{X}_1}$ be the same conditional distribution $\boldsymbol{Y}|\boldsymbol{X}_1$, but now with $\boldsymbol{X}_2$ marginalised out. Denote their respective diameters by $d^+(P_{\boldsymbol{Y}|\boldsymbol{X}}), (d^+(P_{\boldsymbol{Y}|\boldsymbol{X}_1}))$.

Lemma 4.3.1.

$$d^+(P_{\boldsymbol{Y}|\boldsymbol{X}_1}) \leq d^+(P_{\boldsymbol{Y}|\boldsymbol{X}})$$

Proof. This is immediate since each of the rows of $P_{\boldsymbol{Y}|\boldsymbol{X}_1}$ is a weighted average (the weights on row labelled $\boldsymbol{x}_1$ corresponding to the masses on $\boldsymbol{X}_2|\boldsymbol{X}_1 = \boldsymbol{x}_1$). □

Note that this bound is tight in the sense that it is attained for a particular distribution on $\boldsymbol{X}_2|\boldsymbol{X}_1 = \boldsymbol{x}_1$. Suppose $d^+(P_{\boldsymbol{Y}|\boldsymbol{X}})$ is attained when we compare the row $(\boldsymbol{x}_1, \boldsymbol{x}_2)$ with $(\boldsymbol{x}'_1, \boldsymbol{x}'_2)$ and

$$P\left(\boldsymbol{X}_2 = \boldsymbol{x}_2|\boldsymbol{X}_1 = \boldsymbol{x}_1\right) = 1 \text{ and } P\left(\boldsymbol{X}_2 = \boldsymbol{x}'_2|\boldsymbol{X}_1 = \boldsymbol{x}'_1\right) = 1$$

then it is easy to check that $d^+(P_{\boldsymbol{Y}|\boldsymbol{X}_1}) \leq d^+(P_{\boldsymbol{Y}|\boldsymbol{X}})$.

Lemma 4.3.2. *Using the obvious notation, for any two joint probability mass functions* $p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}), p'_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})$ *over* $\boldsymbol{X}, \boldsymbol{Y}$

$$d_V(p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}), p'_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})) \leq \inf\left\{d_V(p_{\boldsymbol{X}}(\boldsymbol{x}), p'_{\boldsymbol{X}}(\boldsymbol{x})) + \sup_{\boldsymbol{x}} d_V(p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}), p'_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})), 1\right\}.$$

Proof. Note

$$\begin{aligned} 2d_V(p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}), p'_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})) &\triangleq \sum_{x,y} \left|p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}) - p'_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})\right| \\ &\triangleq \sum_{x,y} \left|p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) p_{\boldsymbol{X}}(\boldsymbol{x}) - p'_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) p'_{\boldsymbol{X}}(\boldsymbol{x})\right| \end{aligned}$$

Let $r(\boldsymbol{y}|\boldsymbol{x}) \triangleq p'_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}) - p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})$. Then

$$\begin{aligned} & 2d_V(p_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}), p'_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})) \\ = & \sum_{x,y} \left|p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})\left(p_{\boldsymbol{X}}(\boldsymbol{x}) - p'_{\boldsymbol{X}}(\boldsymbol{x})\right) + r(\boldsymbol{y}|\boldsymbol{x}) p'_{\boldsymbol{X}}(\boldsymbol{x})\right| \\ \leq & \sum_{x} \left\{\left|p_{\boldsymbol{X}}(\boldsymbol{x}) - p'_{\boldsymbol{X}}(\boldsymbol{x})\right| \left(\sum_{y} p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})\right)\right\} + \sum_{x,y} p'_{\boldsymbol{X}}(\boldsymbol{x}) \left|r(\boldsymbol{y}|\boldsymbol{x})\right| \\ = & \sum_{x} \left|p_{\boldsymbol{X}}(\boldsymbol{x}) - p'_{\boldsymbol{X}}(\boldsymbol{x})\right| + \sum_{x} \left\{p'_{\boldsymbol{X}}(\boldsymbol{x}) \sum_{y} \left|r(\boldsymbol{y}|\boldsymbol{x})\right|\right\} \\ \leq & \sum_{x} \left|p_{\boldsymbol{X}}(\boldsymbol{x}) - p'_{\boldsymbol{X}}(\boldsymbol{x})\right| + \sum_{y} \sup \left|r(\boldsymbol{y}|\boldsymbol{x})\right| \\ \triangleq & \ 2d_V(p_{\boldsymbol{X}}(\boldsymbol{x}), p'_{\boldsymbol{X}}(\boldsymbol{x})) + 2 \sup_{\boldsymbol{x}} d_V(p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x}), p'_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}|\boldsymbol{x})). \end{aligned}$$

The result follows, since total variation ($d_V(\cdot, \cdot)$) is by definition bounded by 1.

This simple result leads to another useful result. □

Lemma 4.3.3. $d^+(P_{\boldsymbol{Y}|\boldsymbol{X}}) \leq \inf \left\{ d^+(P_{\boldsymbol{Y}_1|\boldsymbol{X}}) + d^+(P_{\boldsymbol{Y}_2|\boldsymbol{X},\boldsymbol{Y}_1}), 1 \right\}$

Proof. From the Lemma above we have for all values of $\boldsymbol{x}, \boldsymbol{x}'$

$$\begin{aligned}
& d_V(p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}_1, \boldsymbol{y}_2|\boldsymbol{x}), p_{\boldsymbol{Y}|\boldsymbol{X}}(\boldsymbol{y}_1, \boldsymbol{y}_2|\boldsymbol{x}')) \\
\leq \; & d_V(p_{\boldsymbol{Y}_1|\boldsymbol{X}=x}(\boldsymbol{y}_1|\boldsymbol{x}), p_{\boldsymbol{Y}_1|\boldsymbol{X}=x}(\boldsymbol{y}_1|\boldsymbol{x}')) \\
& + \sup_{(\boldsymbol{y}_1,\boldsymbol{x}),(\boldsymbol{y}_1',\boldsymbol{x}')} d_V(p_{\boldsymbol{Y}_2|\boldsymbol{Y}_1,\boldsymbol{X}}(\boldsymbol{y}_2|\boldsymbol{y}_1, \boldsymbol{x}), p_{\boldsymbol{Y}_2|\boldsymbol{Y}_1,\boldsymbol{X}}(\boldsymbol{y}_2|\boldsymbol{y}_1, \boldsymbol{x}')).
\end{aligned}$$

However, by definition

$$\begin{aligned}
d^+(P_{\boldsymbol{Y}_1|\boldsymbol{X}}) &= \sup_{\boldsymbol{x},\boldsymbol{x}'} \left\{ d_V(p_{\boldsymbol{Y}_1|\boldsymbol{X}=x}(\boldsymbol{y}_1|\boldsymbol{x}), p_{\boldsymbol{Y}_1|\boldsymbol{X}=x}(\boldsymbol{y}_1|\boldsymbol{x}') \right\} \\
d^+(P_{\boldsymbol{Y}_2|\boldsymbol{X},\boldsymbol{Y}_1}) &= \sup_{(\boldsymbol{y}_1,\boldsymbol{x}),(\boldsymbol{y}_1',\boldsymbol{x}')} \left\{ d_V(p_{\boldsymbol{Y}_2|\boldsymbol{Y}_1,\boldsymbol{X}}(\boldsymbol{y}_2|\boldsymbol{y}_1, \boldsymbol{x}), p_{\boldsymbol{Y}_2|\boldsymbol{Y}_1,\boldsymbol{X}}(\boldsymbol{y}_2|\boldsymbol{y}_1, \boldsymbol{x}') \right\}.
\end{aligned}$$

Hence, the result follows. □

4.4 Discussion

In this section we have argued that the total variation distance is a much more suitable metric for the study of robustness in BNs than the alternatives currently in employment. Although difficult to write explicitly, the properties of total variation (such as its convexity) combined with a pre-specified set of target variables enable BN modellers to confine attention to a subset of the model.

We have introduced new definitions to quantify conditional independence relationships between variables in a BN, underpinned by the total variation distance. These simple, yet powerful measurements are later employed to quantify inferential errors in ancestral chains and structural misspecifications of a given model, see Section 5.

Chapter 5

Robustness in BNs

Both for purposes of elicitation and inference, an important component of modelling with BNs is to develop a meaningful understanding of how sensitive the critical features of a particular BN might be to the misspecification of its inputs. Specifically, we are interested in the sensitivity of the chosen target variables to errors in the topology of the BN or row probability perturbations within given CPTs. When time resources are limited, then such issues impact directly on the permissible level of complexity of the BN model as well as the quantity and quality of the elicitations undertaken to ascertain various components. Currently most sensitivity analyses are performed once the whole BN has been compiled and are based on Kullback-Leibler information measures or Chan-Darwiche distances. Instead, we argue that robustness methods based on the familiar measure of total variation distance, provide simple and more useful bounds on robustness in BNs. These bounds also have the added bonus of being both formally justifiable and transparent, which is essential in order for a BN model to be inherited by the decision maker.

When engaging in the two stage process of structurally engineering and probabilistically evaluating the model, the analyst needs to be fully aware of precisely which inputs might be critical to the inferences made through the BN, see Albrecht et al. [2014]. One critical element of an elicitation or statistical estimation of the graph is to ensure these critical features are specified as accurately as possible. This is particularly important when elicitation or estimation is resource limited, as is usually the case in practice. The client can then optimise their allocation of resources to concentrate on eliciting those elements of the model whose misspecification might most influence the required outputs.

To this end the practitioner, prompted by the functionality of various software, is progressively being encouraged to develop awareness of the robustness in a chosen model (to its inputs) by performing a one-at-a-time numerical *sensitivity* analysis of the preliminary BN. Here, various different forms of numerical contaminations of the model are investigated, where effects are usually measured in terms of mutual information or Kullback-Leibler divergence, see Albrecht et al. [2014]; Friedman et al. [1997]; Nicholson and Jitnah [1998]; Zaragoza et al. [2011]. This type of study is obviously extremely useful, but on the other hand it has drawbacks. Firstly, it relies on the chosen enacted perturbations covering the whole space to discover all those features within the system whose impact on the final analysis is most dramatic. This becomes more challenging as models become even moderately large. Furthermore, even if such a search is performed systematically, the impacts, as measured by mutual information, are not directly relevant to the impact on ensuing decisions, see below for further clarification. Secondly, such an analysis must perforce be performed after the model has been fully specified. This means that the whole probability model is needed before the sensitivity analysis can be performed. One recent attempt to provide such assessments after the structural elicitation phase, but before the probabilistic population is through the use of distance weighted sensitivity measures [see Albrecht et al., 2014]. However, these do not fit well with mutual information measures and have a level of arbitrariness in the choice of weight function needed to use this model.

Over recent years more formal and systematic *robustness* analyses have appeared. Robustness in probability models has been studied by statisticians for many decades, and specific methodology for Bayesian networks has also been recently developed: Coupé and van der Gaag [2002], Gómez-Villegas et al. [2013], Laskey [1995], O'Neill [2009], Renooij [2010]. These fall into two main streams: local robustness studies and global studies. In the former, a chosen probability model is perturbed using a finite parametrised modification. The latter, termed global analyses, do not rely on perturbations lying within a given parametric family [O'Neill, 2009; Smith and Daneshkhah, 2010]. Here instead, an appropriate divergence measure is applied to first specify an appropriate neighbourhood system around each model. Bounds are then calculated for the maximum deviation in the inference that could be achieved by a model in this neighbourhood. If this deviation is small then the model is deemed to be robust [Gustafson and Wasserman, 1995; Smith and Rigat, 2012]. Both types of robustness analysis have been applied to BNs, for example see Smith and Daneshkhah [2010]. In this work we focus solely on global robustness studies and apply these mostly to discrete marginal tables.

Thus far, global robustness studies in BNs have in the main, centred around the analysis of how robust a model might be to perturbations with respect to Kullback-Leibler (KL) or Chan-Darwiche divergences. Both of these divergence measures benefit from some helpful technical properties which allow various measures of dependence to be written as explicit formulae. For details of some of these see Chan and Darwiche [2005], Gómez-Villegas et al. [2013], Leonelli et al. [2017]. These measures are specified in terms of the closeness of log probabilities in the KL case or equivalently ratios of probabilities in the Chan-Darwiche instance, therefore both have the disadvantage that they depend very heavily on the modeller being able to specify small probabilities accurately. It is well documented that it is precisely these small probabilities that typically exhibit the largest elicitation error [see O'Hagan et al., 2006; Smith, 2010]. Furthermore, when BNs are learned from data, any associated small probabilities are difficult to reliably estimate from data, because almost by definition we will see very few of these events in any training set we use to estimate a model.

In many circumstances (especially in decision analysis) the misspecification of the probabilities of improbable events has only a small impact on the required outputs of a decision analysis: see results below for confirmation (Section 5.1). For the purposes of the two stage process described above, the Kullback-Leibler and Chan-Darwiche divergence measures are hardly ideal as they can be highly sensitive to small misspecification.

We demonstrate that an alternative robustness measure based on a more conventional divergence measure (widely used in probability theory and stochastic analysis), which is the total variation distance, has some serious practical and theoretical advantages over its main alternatives. Although it is often difficult to derive *explicit* formulae for the impacts of deviation in variation, it is nevertheless straightforward to *effectively bound* such deviations in variation distance. Deviation in variation corresponds much more closely to the types of error we would envisage experiencing within either an elicitation exercise or through misestimation. Perhaps the most important aspect is that variation distance is a fundamental property, so that for a fixed bounded utility function U, disparities in its expectations under various decisions (induced by an approximation) are simply bounded by linear functions of the total variation in the probability distributions of the attributes of U, Smith [2010]. Note that in a BN, these attributes will typically constitute a small subset of the totality of its variables, so small variation distances (between probability mass functions) on these small subsets translate *directly* into small effects in the pertinent expected utilities. Conversely, large deviations translate into large effects that may have a greater impact on some specification of a utility. These are properties not shared in totality by the two alternative measures discussed above. The study of the robustness in BNs using variation distances (where we focus on the impact of misspecification on the small subset of attribute variables) can therefore be extremely useful.

The bounds we present in this chapter often depend upon a univariate measure of strength which we introduced in Section 4.3; the diameter. Recall that this is a function of the variation distance which measures the dependence along the edge of a BN and is reasonably quick to calculate. Combining this measure of independence with the junction tree representation of our given BN enables us to determine the sequential effect of misspecification of one margin (or conditional probability) on another within a simple path. We demonstrate this impact can be bounded explicitly in terms of a simple function of the extreme entries of the CPTs within the BN. These results have the useful spin-off that CPTs do not necessarily need to be fully elicited before the robustness analysis can take place. In Section 5.2 we show how these explicit measures of robustness can be applied both to determine the effect of approximating simplifications on the topology of the BN and to decide the number of levels into which to categorise each variable. We demonstrate how (using total variation), sensitivity analyses can be performed in a harmonious composite way that directly bounds the impact on decision making of various types of expedient approximations. Finally, in Section 5.5 we provide some guidelines to best employ our results in practice and discuss some enhancements of our strategy. We demonstrate that such methods can be used within the process of building a BN, allowing the creator to adapt priorities as the model is being formulated.

5.1 Variation and Mixtures of CPTs

With our new definitions of diameter we can now use total variation distance to prove the following result that enables us to track a perturbation in a marginal probability through a given BN:

Theorem 5.1.1. *Let $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ be two possible margins of vectors $\boldsymbol{X}$ and $\boldsymbol{Y}$ of random variables and suppose that $P(\boldsymbol{Y}|\boldsymbol{X})$ is the (shared) CPT of the concatenated* *levels of the conditional $\boldsymbol{Y}|\boldsymbol{X}$ and that $\boldsymbol{\rho}_1$ and $\boldsymbol{\rho}_2$ are the margins of Y. Then*

$$d_V(\boldsymbol{\rho}_1, \boldsymbol{\rho}_2) \leq d^+(P(\boldsymbol{Y}|\boldsymbol{X}))d_V(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2).$$

Proof. For $i = 1, 2, \ldots, n$ let $\pi_1 \wedge \pi_2 \triangleq \min_i(\pi_1(i), \pi_2(i))$ and let us define

$$\pi_{1\wedge 2} = \frac{\pi_1 \wedge \pi_2}{\sum(\pi_1 \wedge \pi_2)} = \frac{\pi_1 \wedge \pi_2}{\beta}.$$

So by definition $1 - \beta = d_V(\pi_1, \pi_2)$.

Let

$$\begin{aligned}\pi_1' &= d_V(\pi_1, (\pi_1 \wedge \pi_2)),\\ \pi_2' &= d_V(\pi_2, (\pi_1 \wedge \pi_2)).\end{aligned}$$

However, we would like π_1' and π_2' to be densities, so let:

$$\begin{aligned}\pi_1^* &= \frac{\pi_1'}{\sum \pi_1'} = \frac{\pi_1'}{1-\beta},\\ \pi_2^* &= \frac{\pi_2'}{\sum \pi_2'} = \frac{\pi_2'}{1-\beta}.\end{aligned}$$

We know that $\pi_1 = (\pi_1 \wedge \pi_2) + \pi_1'$ and a similar statement holds for π_2, however we can now rewrite π_1 and π_2 as a sum of two densities:

$$\begin{aligned}\pi_1 &= \beta\pi_{1\wedge 2} + (1-\beta)\pi_1^*,\\ \pi_2 &= \beta\pi_{1\wedge 2} + (1-\beta)\pi_2^*.\end{aligned}$$

Note that

$$\begin{aligned}d_V(\rho_1, \rho_2) &= d_V(\pi_1 P, \pi_2 P)\\ &= d_V\Big(\left[\beta\pi_{1\wedge 2} + (1-\beta)\pi_1^*\right]P, \left[\beta\pi_{1\wedge 2} + (1-\beta)\pi_2^*\right]P\Big)\\ &\leq (1-\beta)d_V(\pi_1^* P, \pi_2^* P).\end{aligned}$$

Now consider $\pi_1^* P, \pi_2^* P$ as different mixture mass functions on Y given that X takes each of its n values, i.e. the different rows $\mathbf{p}_i$ of P. Let $\pi_1^* = (\pi_{11}^*, \pi_{12}^*, \ldots, \pi_{1n}^*)$, $\pi_2^* = (\pi_{21}^*, \pi_{22}^*, \ldots, \pi_{2n}^*)$ and the rows of P be $\mathbf{p}_i = (p_{i1}, p_{i2}, \ldots, p_{in})$. Note that by simple matrix multiplication on the entries we can now write

$$\begin{aligned}\pi_1^*P &= \left(\left(\pi_{11}^*p_{11} + \pi_{12}^*p_{21} + \ldots + \pi_{1n}p_{n1}\right), \left(\pi_{11}^*p_{12} + \pi_{12}^*p_{22} + \ldots + \pi_{1n}p_{n2}\right), \ldots, \right.\\ &\qquad\qquad \left.\left(\pi_{11}^*p_{1n} + \pi_{12}^*p_{2n} + \ldots + \pi_{1n}p_{nn}\right)\right)\\ &= \pi_{11}^*\left(p_{11}, p_{12}, \ldots, p_{1n}\right) + \pi_{12}^*\left(p_{21}, p_{22}, \ldots, p_{2n}\right) + \ldots + \pi_{in}\left(p_{n1}, p_{n2}, \ldots, p_{nn}\right)\\ &= \pi_{11}^*\mathbf{p}_1 + \pi_{12}^*\mathbf{p}_2 + \ldots + \pi_{1n}^*\mathbf{p}_n\\ &= \sum_{i=1}^{n} \pi_{1i}^*\mathbf{p}_i.\end{aligned}$$

Similarly,

$$\pi_2^*P = \sum_{i=1}^{n} \pi_{2i}^*\mathbf{p}_i$$

therefore we can employ Lemma 4.2.3:

$$\begin{aligned}d_V(\rho_1, \rho_2) &\leq (1-\beta)d_V(\pi_1^*P, \pi_2^*P)\\ &\leq (1-\beta)\sum_{i=1}^{n}\sum_{i'=1}^{n} \pi_{1i}^*\pi_{2i'}^* d_V(\mathbf{p}_{i'}, \mathbf{p}_i)\\ &= (1-\beta)\pi_1 D^+(P)\pi_2^T.\end{aligned}$$

Therefore, by definition of the diameter $d^+(P)$

$$d_V(\rho_1, \rho_2) \leq (1-\beta)\sum_{i=1}^{n}\sum_{i'=1}^{n} \pi_{1i}^*\pi_{2i'}^* d^+(P)$$

$$\begin{aligned}&= (1-\beta)d^+(P)\\ &= d^+(P)d_V(\pi_1, \pi_2).\end{aligned}$$

□

This property is particularly useful since it demonstrates that if $P(\boldsymbol{Y}|\boldsymbol{X})$ has been specified accurately but that the margin $\boldsymbol{\pi}_1$ is uncertain then our marginal beliefs about $\boldsymbol{Y}$ are no more uncertain than those about $\boldsymbol{X}$, because by definition $d^+(P(\boldsymbol{Y}|\boldsymbol{X})) \leq 1$. More importantly we have a bound on how much our uncertainty, measured in terms of total variation distance, reduces with respect to $d^+(P(\boldsymbol{Y}|\boldsymbol{X}))$ – how far $\boldsymbol{Y}$ is from independence of $\boldsymbol{X}$. For example, if $\boldsymbol{Y}$ contained the set of attributes of a utility function whilst $\boldsymbol{X}$ were a vector of variables simply informing $\boldsymbol{Y}$, then the theorem above gives us a bound of the effect a misspecification of the margin of $\boldsymbol{X}$ might have on the margin of interest $\boldsymbol{Y}$. One fairly simple consequence of this theorem is that if $\boldsymbol{X} \amalg \boldsymbol{Y}$ then $d^{+}(P(\boldsymbol{Y}|\boldsymbol{X})) = 0$; therefore, misspecification of the mass function of $\boldsymbol{X}$ will have no effect on the $\boldsymbol{Y}$ margin. We use this result more subtly in later bounds by transferring our BN into a junction tree representation in order to deduce sequential effects in long paths by using products of specific CPT diameters.

Example 5.1.1. Let us once again look at the CPT of 'Tree Condition', P, which had a binary parent 'Drought' and a three-state parent 'Rainfall'. The joint distribution can be calculated from CPTs as $\boldsymbol{\pi}_1 = (0.05, 0.175, 0.025, 0.15, 0.525, 0.075)$; however, suppose we were wondering whether $\boldsymbol{\pi}_2 = (0.05, 0.275, 0.03, 0.15, 0.4, 0.095)$ was a more appropriate margin. We have previously determined that $d^{+}(P) = 0.7$ and can calculate $d_V(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2) = 0.125$ allowing us to directly apply Theorem 5.1.1:

$$\begin{aligned} d_V(\boldsymbol{\rho}_1, \boldsymbol{\rho}_2) &\leq d^{+}(P(\boldsymbol{Y}|\boldsymbol{X})) d_V(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2) = 0.7 \times 0.125 \\ &= 0.0875 \end{aligned}$$

We can of course calculate this margin exactly as $d_V(\boldsymbol{\rho}_1, \boldsymbol{\rho}_2) = 0.0555$. However, if we knew only the extreme entries of P then we could still calculate our bound which is of the right order of magnitude: a property we have found to be typical of the types of CPTs we commonly elicit.

5.1.1 Global Bound Approximations

Now that we have looked at the simple case of perturbing a margin only, we move on to the study cases when not only a margin $\boldsymbol{\pi}_1$ of $\boldsymbol{X}$ is perturbed to $\boldsymbol{\pi}_2$, but also the conditional mass functions of $\boldsymbol{Y}|\boldsymbol{X}$ are simultaneously perturbed. Below we formulate variation bounds on the consequent perturbation on the margins $\boldsymbol{\rho}_1, \boldsymbol{\rho}_2$ of $\boldsymbol{Y}$.

From the proof of Theorem 5.1.1 we noted that we can write:

$$\begin{aligned} \boldsymbol{\pi}_1 &= (1-\beta)\boldsymbol{\pi}_1^* + \beta\boldsymbol{\pi}_{1\wedge 2}, \\ \boldsymbol{\pi}_2 &= (1-\beta)\boldsymbol{\pi}_2^* + \beta\boldsymbol{\pi}_{1\wedge 2}, \end{aligned}$$

where $(1-\beta) = d_V(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2)$ and where without loss we can assume the mixing process is shared by the two mass functions, so points are drawn either from $\boldsymbol{\pi}_{1\wedge 2}$ or alternatively something drawn from either $\boldsymbol{\pi}_1^*$ or $\boldsymbol{\pi}_2^*$. Using the same argument as for when $P_1 = P_2$

$$\begin{aligned}
d_V\left(\boldsymbol{\rho}_1, \boldsymbol{\rho}_2\right) &= d_V\left(\boldsymbol{\pi}_1 P_1, \boldsymbol{\pi}_2 P_2\right) \\
&= d_V\left(\left((1-\beta)\boldsymbol{\pi}_1^* + \beta\boldsymbol{\pi}_{1\wedge 2}\right) P_1, \left((1-\beta)\boldsymbol{\pi}_2^* + \beta\boldsymbol{\pi}_{1\wedge 2}\right) P_2\right) \\
&\leq \beta d_V\left(\boldsymbol{\pi}_{1\wedge 2} P_1, \boldsymbol{\pi}_{1\wedge 2} P_2\right) + (1-\beta) d_V\left(\boldsymbol{\pi}_1^* P_1, \boldsymbol{\pi}_2^* P_2\right) \\
&\leq \beta d_V^+\left(P_1, P_2\right) + (1-\beta) d_V^*\left(P_1, P_2\right).
\end{aligned}$$

We can then show

$$d_V\left(\boldsymbol{\rho}_1, \boldsymbol{\rho}_2\right) \leq d_V^+\left(P_1, P_2\right) + d_V\left(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2\right) d_V^*\left(P_1, P_2\right). \tag{5.1}$$

Note in particular that

$$d_V\left(\boldsymbol{\rho}_1, \boldsymbol{\rho}_2\right) \leq \lambda d_V\left(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2\right),$$

where

$$\lambda \triangleq \frac{d_V^+\left(P_1, P_2\right)}{d_V\left(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2\right)} + d_V^*\left(P_1, P_2\right),$$

and if

$$\lambda < 1 \Leftrightarrow d_V\left(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2\right) > \frac{d_V^+\left(P_1, P_2\right)}{1 - d_V^*\left(P_1, P_2\right)}.$$

Using Equation 4.3 we have that

$$d_V\left(\boldsymbol{\rho}_1, \boldsymbol{\rho}_2\right) \leq \left\{1 + d_V\left(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2\right)\right\} d_V^+\left(P_1, P_2\right) + d_V\left(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2\right) \max\left\{d(P_1), d(P_2)\right\}. \tag{5.2}$$

5.2 Approximations of the CPTs in a known BN

Suppose all clients are content that the conditional independences in a given BN are valid. Without changing the random variables in the system, we are now interested in finding ways of approximating the graphical model and refining initial probability estimates within this given BN.

The simplest BN to consider is the decomposable BN, i.e. a BN in which all parents of a given child node are married. When a graph is decomposable then we can always find a (non-unique) junction tree which is usually used to guide probability propagation. Here we shall be using the junction tree to inform robustness judgements instead.

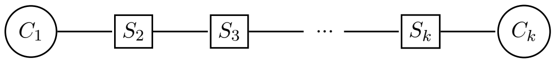
5.2.1 Diminishing tree propagated approximation error

The following result explains that when using standard propagation algorithms to update a clique margin C_1, the effect on subsequent clique margins becomes progressively weaker as the updated cliques become increasingly more remote from C_1. Furthermore the extent of this deviation can be measured, in the sense that it can be bounded above. This enables us to bound the potential error of the focus variable distributions, which may be induced by misspecifications within the structure or even CPT inaccuracies. This is particularly useful when we elicit a large BN and want to know how far away from target nodes we need to elicit the corresponding CPTs accurately.

Theorem 5.2.1. *Let $(C_1, C_2, C_3, \ldots, C_k)$ be the minimal sequence of cliques with associated separators $(S_2, S_3, \ldots, S_k)$. Let each undirected edge of the marginalised junction tree be denoted by δ_i for $i = 1, 2, \ldots, k$; the diameter of the conditional probability table between the two sequential nodes. For example $\delta_1 = d^+(P(S_2|C_1)), \delta_2 = d^+(P(S_3|S_2)), \ldots, \delta_k = d^+(P(C_k|S_k))$. Then*

$$d_V(p_{C_k}(\boldsymbol{x}_{C_k}), q_{C_k}(\boldsymbol{x}_{C_k})) \leq d_V(p_{C_1}(\boldsymbol{x}_{C_1}), q_{C_1}(\boldsymbol{x}_{C_1})) \prod_{i=1}^{k} \delta_i.$$

Proof. By Lemma 2.1.1 we can rewrite our junction tree to marginalise over internal cliques leaving us with the graphical structure:



By successive application of Theorem 5.1.1:

$$\begin{aligned}
d_V(p_{C_k}(\boldsymbol{x}_{C_k}), q_{C_k}(\boldsymbol{x}_{C_k})) &\leq d^+(P(C_k|S_k)) d_V(p_{S_k}(\boldsymbol{x}_{S_k}), q_{S_k}(\boldsymbol{x}_{S_k})) \\
&\leq d^+(P(C_k|S_k)) d^+(P(S_k|S_{k-1})) d_V(p_{S_{k-1}}(\boldsymbol{x}_{S_{k-1}}), q_{S_{k-1}}(\boldsymbol{x}_{S_{k-1}})) \\
&\leq d^+(P(C_k|S_k)) d^+(P(S_k|S_{k-1})) \ldots \\
&\qquad \ldots d^+(P(S_3|S_2)) d^+(P(S_2|C_1)) d_V(p_{C_1}(\boldsymbol{x}_{C_1}), q_{C_1}(\boldsymbol{x}_{C_1})) \\
&= \left(\prod_{i=1}^{k} \delta_i \right) d_V(p_{C_1}(\boldsymbol{x}_{C_1}), q_{C_1}(\boldsymbol{x}_{C_1})).
\end{aligned}$$

□

Next we define the impact of one clique upon another in order to ascertain the diminishing effect of errors downstream in the causal chain.

Definition 5.2.1. Define the *impact* $I(C_k|C_1)$ of C_1 on C_k to be $\prod_{i=1}^{k} \delta_i$.

The impact of one clique on another is a simple measure of the maximum possible influence the misspecifation of one set of clique probabilities could have on another as measured by a bound on the variation distance. Note that in general we can label the edges of a junction tree (which are also labelled by a separator between adjacent cliques) C_i and C_j by two diameters $\delta_{i\to j}$ and $\delta_{j\to i}$ one measuring the impact of i on j and the other the impact of j on i. Note that these two impacts are not necessarily equal, and are often very different. However, in the contexts we consider here (where our primary interest concerns the robustness in the margins of an identified subset of attributes) we usually need to focus on propagation in a single direction. Furthermore, if construction of the BN is consistent with a conjectured causal directionality in mind, then this directionality often tends to have the attributes at the end of the causal chain. Ergo, the diameters we need can often be calculated directly from the diameter of the elicited CPTs of the BN.

Example 5.2.1. The separators of the Native Fish example are too simple to effectively demonstrate the impact quantity defined above. As a substitute suppose we have the following BN:

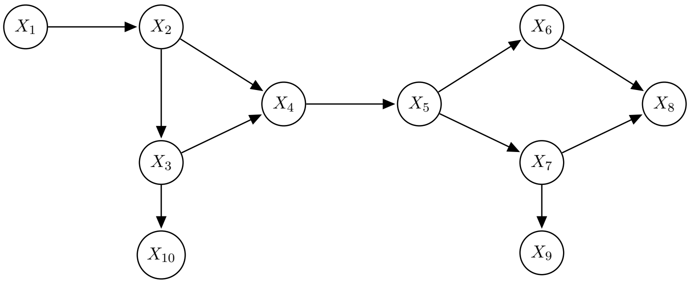


Figure 5.1: BN example to determine Impact of cliques.

Let us label the cliques of Figure 5.1 to satisfy the running intersection property:

$$C_1 = \{X_1, X_2\}, \quad C_2 = \{X_2, X_3, X_4\}, \quad C_3 = \{X_4, X_5\},$$
$$C_4 = \{X_5, X_6, X_7\}, \quad C_5 = \{X_6, X_7, X_8\}, \quad C_6 = \{X_3, X_{10}\}, \quad C_7 = \{X_7, X_9\}$$

Giving us separators:

$$S_2 = \{X_2\}, \quad S_3 = \{X_4\}, \quad S_4 = \{X_5\}, \quad S_5 = \{X_6, X_7\}$$
$$S_6 = C_2 \cap C_6 = \{X_3\}, \quad S_7 = C_5 \cap C_7 = \{X_7\}$$

Suppose we wish to determine the effect on X_9 if we perturb X_1. First draw the ancestral graph of nodes X_1 and X_9, shown below in Figure 5.2:

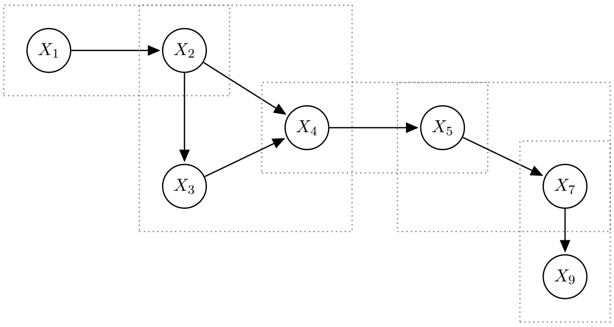


Figure 5.2: Ancestral graph of Figure 5.1 for nodes X_1 and X_9, each dashed box represents a clique.

Giving us the following equation for the impact (which is simply the product of diameters of each separator conditional on the previous previous separators):

$$\begin{aligned} I(X_9|C_1) &= p(X_2|X_1)p(X_4|X_2)p(X_5|X_4)p(X_7|X_5)p(X_9|X_7) \\ &\leq d^+(X_2)d^+(X_4)d^+(X_5)d^+(X_7)d^+(X_9) \end{aligned}$$

Extending this further, we can determine the impact on cliques X_6 and X_7 simultaneously, if we perturb both X_1 and X_2. Following the same steps of creating cliques and separators for the ancestral graph of these nodes, the impact is given as:

$$I(X_6, X_7|X_1, X_2) = p(X_2|X_1)p(X_4|X_2)p(X_5|X_4)p(X_6, X_7|X_5).$$

This can be written in terms of the original BN CPTs using Lemma 4.3.3, as some separators contain more than one node:

$$\begin{aligned} I(X_6, X_7|X_1, X_2) &\leq d^+(X_2)d^+(X_4)d^+(X_5)\Big[\inf\{d^+(X_6|X_5) + d^+(X_7|, X_6, X_5), 1\}\Big] \\ &\leq d^+(X_2)d^+(X_4)d^+(X_5)\Big[\inf\{d^+(X_6|X_5) + d^+(X_7|, X_5), 1\}\Big] \end{aligned}$$

There are various practical corollaries to the simple theorem above:

Corollary 5.2.1.1. *If $\mathcal{G}$ is decomposable and C_i lies on the minimal sequence between C_1 and C_k and if all attributes are in C_k then the probabilities of C_i have a higher influence on C_k than those of C_1.*

Note that we can *always* embed a BN in a decomposable BN by triangulisation e.g. see Cowell et al. [1999] and Smith [2010]. Therefore these bounds can be applied to *any* BN so long as it has been pre-processed to ensure that it is decomposable. We recommend following the construction below to ensure that your BN is in a suitable format to apply Theorem 5.2.1:

- Begin with a BN $\mathcal{G}$, the diameters of whose CPTs have been provisionally elicited.
- Identify a donating variable or complete vector $\boldsymbol{X}_i$ of $\mathcal{G}$ and the vector of focus $\boldsymbol{X}_k$.
- Find the ancestral set of $\boldsymbol{X}_i, \boldsymbol{X}_k$ in $\mathcal{G}$.
- Construct the ancestral graph, A, which has variables $(X_1, X_2, \ldots, X_n)$ where the order of these vertices are chosen compatible with $\mathcal{G}$.
- Create a triangularised version, A^*, of A and find its junction tree J. Denote the clique containing $\boldsymbol{X}_i$ as C_1 and the clique containing $\boldsymbol{X}_k$, C_k.
- Find the single path J^* starting from clique C_1 to C_k labelling the cliques in order $C_1, C_2, \ldots, C_k$.
- Remove all variables that are not in one of these cliques.

Note that these influences provide a very useful tool for prioritisation of the elicitation in a BN. For example if we can obtain estimates of influence across a junction tree (either from direct elicitation of diameter or alternatively after having performed a preliminary coarse elicitation of the corresponding CPTs) then we can use these influences to identify which of those CPTs to refine. For example suppose all attributes consistently lie in a single clique. We can then follow the simple guidelines:

- We first refine the elicitation of the CPTs whose attributes and parents lie in this clique,
- We then elicit the CPTs associated with parents/separators with the most influence,

- Using the method above, we decide on the basis of influence whether to refine the CPTs associated with other parents or parents of parents.

5.3 Approximations associated with a general BN

In a junction tree each vector has just a single parent within a given compatible ordering. Of course in the case of a BN this is no longer necessarily true. We would still like to find the bound on the impact measure of one variable on another and so annotate each of its directed edges with a value between zero and one which reflects this. The result below gives us a way of coding this impact in a useful way.

Suppose $\boldsymbol{Y}$, taking values $\boldsymbol{y} \in \mathbb{Y}$, is potentially dependent on k vectors $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \dots \boldsymbol{X}_k)$ taking values $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_k) \in \mathbb{X} = \mathbb{X}_1 \times \mathbb{X}_2 \times \cdots \times \mathbb{X}_k$. For $j = 1, 2, \dots, k$ let $\boldsymbol{x}_{\widehat{j}} \in \mathbb{X}_{\widehat{j}} \triangleq \mathbb{X}_{\{1,2,\dots,k\}\backslash\{j\}}$ be a vector of values of other variables $\boldsymbol{X}_{\widehat{j}}$. Let the CPT of $\boldsymbol{Y}$ given $\boldsymbol{X}$ be P so that its diameter is given by

$$d^{+}(P) = \frac{1}{2} \max_{\boldsymbol{x},\boldsymbol{x}' \in \mathbb{X}} \left\{ \sum_{\boldsymbol{y} \in \mathbb{Y}} \left| p_{xy} - p_{x'y} \right| \right\}.$$

Definition 5.3.1. Let the diameter $d_j^{+}(P)$ of $\boldsymbol{Y}$ to $\boldsymbol{X}_j$ be defined by

$$d_j^{+}(P) = \frac{1}{2} \max_{\boldsymbol{x}_{\widehat{j}} \in \mathbb{X}_{\widehat{j}}} \max_{\boldsymbol{x}_j, \boldsymbol{x}'_j \in \mathbb{X}_j} \left\{ \sum_{\boldsymbol{y} \in \mathbb{Y}} \left| p_{xy} - p_{x'y} \right| \right\}.$$

Hence the diameter $d_j^{+}(P)$ is the maximum extra effect varying the value of $\boldsymbol{x}_j$ can have on the distribution of $\boldsymbol{Y}$ for any fixed value $\boldsymbol{x}_{\widehat{j}} \in \mathbb{X}_{\widehat{j}}$ of the other variables. Notice in particular that

$$\boldsymbol{Y} \amalg \boldsymbol{X}_j | \boldsymbol{X}_{\widehat{j}} \Leftrightarrow d_j^{+}(P) = 0.$$

Thus in a formal sense $d_j^{+}(P)$ is a measure of the extent by which this conditional independence is violated as well as quantifying the added value of knowing $\boldsymbol{X}_j$ once we know the value of $\boldsymbol{X}_{\widehat{j}}$. We now have the simple but pleasing additive relationship between $d^{+}(P)$ and $d_j^{+}(P)$, $j = 1, 2, \dots, k$.

Theorem 5.3.1. *With the notation above*

$$d^+(P) \leq \sum_{j=1}^{k} d_j^+(P).$$

Proof. Here we simply use the triangle inequality to bound $d^+(P)$, changing the entries of the conditioning variables $\boldsymbol{X}_j$ one at a time. So if $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k)$, $\boldsymbol{x}' = (\boldsymbol{x}_1', \boldsymbol{x}_2', \ldots, \boldsymbol{x}_k')$, $\boldsymbol{x}(0) = \boldsymbol{x}, \boldsymbol{x}(1) = (\boldsymbol{x}_1^*, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_k)$, $\boldsymbol{x}(2) = (\boldsymbol{x}_1', \boldsymbol{x}_2^*, \mathrm{x}_3 \ldots, \boldsymbol{x}_k)$, …, $\boldsymbol{x}(k) = (\boldsymbol{x}_1', \boldsymbol{x}_2', \ldots, \boldsymbol{x}_k^*)$, then

$$\begin{aligned}
\max_{\boldsymbol{x},\boldsymbol{x}' \in \mathbb{X}} \left\{ \sum_{j=1}^{n'} \left| p_{xy} - p_{x'y} \right| \right\} &\leq \sum_{i=1}^{k} \max_{\boldsymbol{x}(i),\boldsymbol{x}(i-1) \in \mathbb{X}} \left\{ \sum_{j=1}^{n'} \left| p_{\boldsymbol{x}(i)y} - p_{\boldsymbol{x}(i-1)y} \right| \right\} \\
&= \sum_{i=1}^{k} \max_{\boldsymbol{x}_{\widehat{j}} \in \mathbb{X}_{\widehat{j}}} \max_{\boldsymbol{x}_j, \boldsymbol{x}_j^* \in \mathbb{X}_j} \left\{ \sum_{j=1}^{n'} \left| p_{xy} - p_{x^*y} \right| \right\}.
\end{aligned}$$

since by definition of $\boldsymbol{x}$ and $\boldsymbol{x}'$, the adjacent CPTs appearing in the sum above, differ only in the i^{th} entry. □

5.4 Robustness to approximation by simplification

Often a BN is chosen to be sparser than it would be were we to have more information or time. This happens for a variety of reasons. For example, when eliciting a BN we often ask for the list of the *most important* variables on which a specific variable X might depend; by definition this phrase implies that variables not included in the list could be expected to have only a small influence on X. This restriction is imposed because it is difficult for a client to think clearly about the interrelationships between more than a handful of variables. Increasing the number of different joint levels on the conditioning variables increases the entries in the CPT that need to be elicited.

If data is used to inform the model choice then a severe penalty is often imposed (or is implicit) to ensure the selection of models with smaller size parameter spaces - which in this context usually implies sparser associated graphs. Finally, for reasons of suitability of implementation, it is quite common for a search of candidate BNs to include only those graphs whose nodes have no more than a fixed number of parents: the limit often set to be two or three, e.g. see Cussens [2011].

It is therefore very important to properly understand the implications of these potential over-simplifications on the robustness in the BN model. We present some corollaries on the use of these results.

5.4.1 Edge Deletion

When constructing BN systems we seek the model which best describes the underlying physical process; realistically however we are often limited by restrictions on resources and time which lead us to instead develop a model. The size of a BN can grow exponentially by adding more variables, relationships or even states, but eliciting these probabilities becomes problematic. Experts tend to struggle with conceptually visualising complex relationships between more than a couple of variables. Therefore, a frequent simplification of a model can be to delete unnecessary edges that represent very weak influences, without impacting on the robustness in the system.

The methodology we have introduced in earlier sections can be directly applied to this problem in order to quantify the cost of removing a certain variable from the parent set of another variable.

Example 5.4.1. Return once again to the CPT of 'Tree Condition' given its binary parent 'Drought' and three-state parent 'Annual Rainfall', as in Example 4.3.1. Denote the rows of P by $\boldsymbol{p}_i$ for $i = 1, 2, \dots, 6$. If we are interested in the effects of deleting the arc Drought $\longrightarrow$ Tree Condition, then we can use Definition 5.3.1 directly:

$$\max_{\{yes,no\}} \max_{\{Below,Avg.,Above\}} \{d_V(\boldsymbol{p}_i, \boldsymbol{p}_{i'})\} = 0.6$$

Alternatively, deleting arc Annual Rainfall $\longrightarrow$ Tree Condition gives us

$$\max_{\{Below,Avg.,Above\}} \max_{\{yes,no\}} \{d_V(\boldsymbol{p}_i, \boldsymbol{p}_{i'})\} = 0.2$$

Figure 5.3 presents the full BN with each edge annotated with the quantitative effect on the child if we delete the parent.

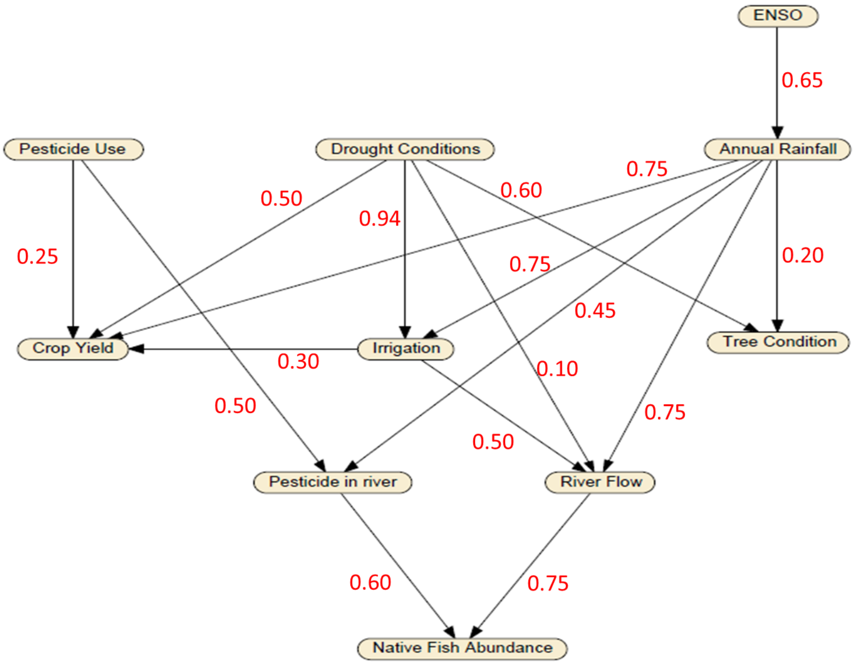


Figure 5.3: Visual representation of edge deletion effect in the Native Fish BN.

The higher the value of $d_j^+(P)$ the larger the effect of deleting the arc $\boldsymbol{X}_j \longrightarrow \boldsymbol{Y}$ since it indicates that the corresponding rows of other parent responses are dissimilar. Of course this model is already well designed and so we find that most edges need to be there. However, the possible exception is the edge from Drought Conditions to River Flow. In practical models devised early in an analysis, as in large BNs, we often find that many edges can be omitted without strong effects on the outputs of the system.

This definition scales up easily to handle parent sets of discrete BNs with numerous nodes and large numbers of states due to the variation distance being a simple and transparent piece of arithmetic.

5.4.2 Level Amalgamation

One practical issue experience by BN modellers, is choosing the number of levels to assign to each random variable within the system; even when using given data the selection of appropriate bins is not straightforward. Obviously there is a compromise here. The finer the division of levels, the more nuanced the BN can be. On the other hand, the fewer the number of levels, the easier it will be to faithfully elicit or efficiently estimate the probabilities within a BN. One advantage of using the variation approach for robustness is that such considerations can be taken under the same technical umbrella as other necessary approximations. We simply evaluate the potential that such simplifications might have on the distribution of the attributes of the problem, just as in considering whether or not to keep a weak edge in the system.

When deciding to amalgamate levels within a specified random variable, we ensure that the interpretation of the states can still be understood and quantified by experts. When those variables are ordinal and have a monotonic relationship with its neighbours, we recommend that consideration is given only to amalgamating consecutive node levels. For example, the node of Annual Rainfall in Example 4.3.1 had levels Below Average, Average, Above Average and if we were to combine Below Average with Above Average we would not have a cohesive state, which the experts may find very difficult to quantify.

The second step in level amalgamation is combining the chosen states. We recommend taking a vector average of the chosen CPT rows, selecting one of the rows or take a weighted average of rows. The simplest case would be to take the row average because the convexity of variation distance tells us that this will enable us to avoid increasing the diameter of the original CPT and has the added benefit of not requiring additional information (which would otherwise be required for a weighted row average).

Occasionally the modeller or expert will have an intuitive feel for which states should be combined, possibly from past experience or relevant data. However, sometimes this may not be obvious. In the latter case we can calculate the variation distance between the considered states and combine the closest states first, then find the next closest state and add to the amalgamation iteratively until the combination appears to imply significant variation distance from the original.

Example 5.4.2. For example, let us once again consider the 'Tree Condition' CPT from 4.3.1. To reduce the three-state node Annual Rainfall to a binary node we could either combine Below Average with Average (case i) or Above Average with Average (case ii). To decide we compare the variation distance between rows $\boldsymbol{p}_i$:

$$\begin{aligned} &\text{Case i:} && \max\{d_V(\boldsymbol{p}_1,\boldsymbol{p}_2), d_V(\boldsymbol{p}_4,\boldsymbol{p}_5)\} = 0.1, \\ &\text{Case ii:} && \max\{d_V(\boldsymbol{p}_2,\boldsymbol{p}_3), d_V(\boldsymbol{p}_5,\boldsymbol{p}_6)\} = 0.1. \end{aligned}$$

In this instance we can arbitrarily choose between the two, so opting for case (i) to form the amalgamated state 'Average or Below' using a simple average of relevant rows to obtain P':

$$P = \begin{pmatrix} 0.20 & 0.60 & 0.20 \\ 0.25 & 0.60 & 0.15 \\ 0.30 & 0.60 & 0.10 \\ 0.70 & 0.25 & 0.05 \\ 0.80 & 0.18 & 0.02 \\ 0.90 & 0.09 & 0.01 \end{pmatrix}, \qquad P' = \begin{pmatrix} 0.225 & 0.60 & 0.175 \\ 0.30 & 0.60 & 0.10 \\ 0.725 & 0.215 & 0.06 \\ 0.90 & 0.09 & 0.01 \end{pmatrix}.$$

Calculating that $d_V^+(P, P') = 0.075$, results such as Theorem 5.1.1 show that the effect of using this amalgamated CPT compared to the original is small.

5.5 Some principled strategies for BN creation

Obviously the evaluation criteria we indicate here can be embedded into a formal protocol. However, there are many considerations that a user has to consider before undertaking model construction: transparency of the model, computational issues, elicitation constraints and so forth, which vary in importance depending on the context of the model building. So setting a bound on any effects or perturbations against differing approaches is often best undertaken more informally. Nevertheless, we acknowledge that the framework we have presented here is sufficiently formal to admit generalisation and this is work that we plan to undertake next.

To implement our techniques as efficiently as possible we recommend two differing approaches tailored to the specific circumstances of the modeller. Firstly, there are occasions when we have obtained provisional information from one expert who can recommend nodes, levels, interactions and provisional CPTs before undertaking a more formal elicitation conference with multiple experts. Such was the case in the pollinator example discussed in Barons et al. [2018a] and Barons et al. [2018b]. In this particular scenario, we can design the analysis by using the bounds on the preliminary values stated by the expert as discussed earlier. We recommend starting by eliciting attributes and nodes of interest prior to working systematically backwards along the chain of influence to discover parent nodes and conditional independences. To determine the efficacy of including certain variables in the model, perform variation measures on preliminary CPT values. Of course, after the full elicitation has taken place the robustness analyses suggested above can be repeated for a final sensitivity analysis.

In situations when we begin modelling with no such preliminary information, it may be wisest to attempt to elicit the diameter of each CPT directly. This can then be bounded and decisions undertaken on whether to include certain variables in the chain or not. To elicit the diameter directly we need to ascertain the largest differences between rows of a CPT which correspond to requesting the “best case scenario” probabilities and the “worst case scenario” probabilities before calculating the variation distance between the two. Recall that if the diameter is zero then the two variables are said to be conditionally independent. So if we elicit a small diameter then we can omit any interactions which are not strong enough to have a marked effect on the outputs of the model. This course of action allows a facilitator to elicit the least number of probabilities to gain an insight into the importance of each variable, and if more information is required then a full elicitation on the CPT can follow. We note that this methodology is consistent with the common practice of first eliciting extreme row probability vectors of a CPT and then using these as a benchmark to complete the table of intermediate rows. Although not often documented, experienced BN modellers tend to elicit best and worse case scenarios for configurations of parents and then complete the intermediate entries using these extremes as benchmarks.

Following this procedure, we continuously appraise and compare each possible simplification against the potential accuracy of an analysis, weighted against the issues provided by a simpler representation of a model. We can see through these examples that in many cases the effects of various simplifications are often very small, and approximations based on these simplifications are justified from a pragmatic point of view. We also note that some of the best approximations to use are often not the ones currently undertaken in practice. For example we often find that using an approximation which deletes an edge can cause significant changes, whilst allowing dependence on subsets of levels only, performs much better.

5.6 Discussion

We have demonstrated here how the properties of variation distance can be harnessed to study the robustness in a discrete BN, if certain target variables are known a priori to be those of primary interest. Although all our illustrative examples in this thesis have perforce been of moderate size, our methods become ever more useful as the number of nodes in the BN increases. Even when this number is huge, we can show it is possible to identify a priori which features of the full joint distribution will have the strongest impact on the target variables of interest, and therefore employ effective and expedient approximations to make inferences which are both accurate and feasible for the task at hand. In such models, since the simple paths between learned variables and attribute variables is typically much longer, it is possible to formally demonstrate that some remote variables are just not worth eliciting directly, but should be marginalised out of the system.

The approach we have introduced in this thesis relies on the well-studied variation distance which naturally embeds conditional independence relationships between variables. We have therefore devised a seamless way of looking at perturbed versions of a BN in a manner which enables us to apply the same devices to generic effects, be these perturbations associated with edge deletion, the effect on descendants of learning certain variables, misquoting probabilities within CPTs or changing the number of levels for nodes. All of these different alterations can be compared on an equal footing whereas previous work usually depended on model selection of BNs, using methodology such as Bayes Factors, which focused on a particular perturbation at a time. We note however that many of the techniques, such as the use of heatmaps demonstrated in Albrecht et al. [2014], to visually communicate the bounding effect of one node on a target node can be straightforward to adapt and reapplied to this domain.

There is of course much work to still be undertaken in this field, beginning with refining the bounds we have developed here. The robustness studies we have introduced can also be applied to context-specific BNs where we have a natural trade-off between the number of probabilities to elicit and the robustness in the model. Typically if we elicit fewer probabilities for a context-specific model, we can show weakening of the robustness of the system due to the constraint of forcing inputs to be the same. Similarly within this work we have had no space to consider the robustness of the choice of probability distribution on the entries of the CPTs of a BN. Smith and Daneshkhah [2010] studies BN robustness associated with the inputs of the distribution in terms of the local DeRobertis distance. In Smith and Rigat [2012] it is shown that the DeRobertis distance can be bounded by the variation distance. Therefore a fairly straightforward extension to the variation bounds we have presented here can be developed by carefully combining our results with the DeRobertis distance to provide a comprehensive robustness analysis when necessary. Essentially we can show that with sufficient data and global independence assumptions the most robust CPTs are the ones whose probabilities are best known.

Finally, recent theoretical advances have suggested that if a Bayesian accepts that their model is only approximate, the M-open scenario [Bernardo and Smith, 1994], then Bayesian learning using Bayes Rule may not be optimal and that other updating rules based on divergences other than the KL divergence should be considered. There are exciting new possibilities of combining this technology with the robustness methods described here when that divergence is defined as the variation distance, Jewson et al. [2018].

Our ideas also apply directly to the Dynamic BN where the robustness in the system can be far more important when the dynamic nature of the problem makes the model much more complex. Throughout this thesis, for simplicity, we have considered only robustness as it applies to finite discrete BNs. However the whole technology we describe here, translates seamlessly into tools for examining continuous and mixed Bayesian networks. Using the variation distance on these highly structured and complex Markov Processes using the approach demonstrated here can help us to determine the robustness in DBNs to dynamic effects. Work in this more general setting has already begun in the following section, soon to be submitted for publication. Thus we have demonstrated that this is actually a very fruitful way of addressing robustness within this family of graphical models.

Chapter 6

Robustness of Temporal Dynamics in DBNs

6.1 Robustness in DBNs

The two time slice Bayesian network (2TSDBN) is now a well used tool for the analysis of multivariate finitely discrete time series, especially when embellished by additional context specific hypotheses. To begin the process an underlying collection of Markov hypotheses are first elicited or selected with the help of a directed graph. These involve not only hypotheses associated with the one step transitions of the Markov process (inter-slice arcs) but also hypotheses concerning the relationships between the variables lying on a particular time slice (intra-slice arcs). This process has been extensively described in texts such as Jensen and Nielsen [2007], Korb and Nicholson [2010] and Smith [2010] and is supported by software as discussed in Section 2.1.4.

We need to customise any elicited DBN by using context specific information that embellishes the framework into a full probabilistic description. For this we require to know which of the conditional probability tables (CPTs) in the system can be regarded as the same, before specifying the probabilities within those CPTs to provide a full probability model of the process. This then provides a family of finite Markov processes whose transition matrices define the possible developments of the process from one time slice to the next. The structure of the DBN, its symmetries and also its conditional probabilities will all need to be either elicited from experts or alternatively estimated from data. Both of these approaches can be fragile and are often very time consuming in any practical application; therefore, methods of tailoring the process by ranking essential conditional probabilities are critical to this task enabling the user to develop an awareness of those features within the model that might have a strong impact on the desired outputs.

As detailed in Section 2.1.1, DBNs are simply a graphical representation of a collection of conditional independence statements, which is also the underlying foundation of Markov Chains. In this section we present a novel approach allowing us to write a DBN in terms of a time inhomogeneous Markov Chain; an initial distribution and a transition matrix. This drastically reduces attention from multiple CPTs to a single transition matrix which enables more succinct sensitivity analyses to be performed. We combine some established results for time inhomogeneous finite Markov Chains with our new variation distance methodology to provide a robustness study customised to the needs of DBNs. The results enable us to highlight where more information needs to be incorporated within a model and to prioritise the elicitation on those features of the model that are most critical to the accuracy of the outputs.

In this section we review some well established DBN results and recap notation before discussing the link between DBNs and finite Markov Chains. We present a formal approach for rewriting a DBN as a finite Markov Chain (MC). In Section 4 we discussed some simple properties of the total variation distance, including the useful convexity property, before introducing our new definitions to measure conditional independence relationships. Here, we apply our new methodology to bound the errors in a DBN associated with the impacts of probability misspecifications over time.

6.2 Finite Markov Chains and DBNs

We shall now present and discuss some fundamental properties of Markov Chain theory which can be found in most introductory MC texts [see Brémaud, 1999; Cox and Miller, 1965; Moran, 1968; Seneta, 2006, for introductory texts]. The following results are well established, however the are included for completeness and will be used later.

DBNs are fundamentally a collection of independence statements relating variables over time, that is they are essentially time inhomogeneous Markov Chains (when the conditional probability tables are not indexed by time). This link between

DBNs and MCs enables us to introduce a new method to rewrite our probabilistic graphical model as a Markov Chain: in terms of an initial probability vector and a transition matrix denoting the evolution of the system between time-slices.

When we have a fully populated DBN this corresponds to a fully numeric transition matrix for which the eigenvalues can be easily found and so enables us to calculate the equilibrium distribution with relative ease, using statistical software for larger matrices. We can also determine the rate at which a Markov Chain approaches the stationary distribution using the Perron-Frobenius Theorem which is described in detail in Brémaud [1999] and for the full proof of the theorem see Seneta [2006]. We know that when we have a stochastic matrix, such as our transition matrices, the largest eigenvalue is $\lambda_1 = 1$. However, it is the second largest eigenvalue λ_2 which is most important for our context since the convergence of our matrix will be driven by the absolute value of the second largest eigenvalue of our matrix. As we converge to the equilibrium distribution at a rate equivalent to the second largest eigenvalue, it is important to quantify and bound this value.

In our context, convergence thus depends on the differences of the size of the second largest (non-unity) eigenvalue of the stochastic matrix. Although this theorem is helpful in determining the type of effects that we might see, the actual magnitude of the second eigenvalue is difficult to appreciate or quantify in practice. Later we show that there is a surrogate measure, defined through the use of variation distance that is easier to quantify, which can be calculated and its practical implications are easier to see.

Unfortunately, it is often the case in DBNs that we have missing data within our CPTs which would require us to look at the algebraic class of transition matrices. In these instances, finding the eigenvalues can be extremely complicated for large matrices. We therefore study robustness directly by using variation distance in order to understand exactly how the model is responding to misspecifications or errors in the system.

6.2.1 Rewriting a DBN as a MC

As we have already indicated, due to the similarities underpinning the DBN and MC structures we can present a new method in which any discrete DBN can be written as a time inhomogeneous Markov Chain.

Consider a 2TSDBN in which each time slice has identical graphical and CPT structure, with nodes denoted by $X_{t,j}$: t represents the time slice and j represents the node indicator. Let the parent set of node $X_{t,j}$ be written as $\mathrm{pa}(X_{t,j})$. Note that the parents of a node may be within the same time-slice or in the previous time slice if it is a dynamic relationship.

Let each random variable $X_{t,j}$ take values $x_j \in \mathbb{X}_{t,j}$, and let $\#(\mathbb{X}_{t,j})$ be the total number of levels of the node $X_{t,j}$. As an example, suppose that $X_{1,1}$ is binary, therefore $\mathbb{X}_{1,1} = \{0, 1\}$ and $\#(\mathbb{X}_{1,1}) = 2$. We can now define Y to be the set of all possible outcomes of all variables in one single time slice, which is the set of all sequences such that $(X_{t,1}, X_{t,2}, \ldots, X_{t,n}) \in \mathbb{X}_{t,1} \times \mathbb{X}_{t,2} \times \ldots \times \mathbb{X}_{t,n}$. Let $S_{t,j}$ be the set of all possible outcomes of the parent set of node $X_{t,j}$.

For simplicity we shall denote the conditional probabilities as

$$P(x_{t,j} = l|s_{t,j}) = \alpha_{t,j;l|s_i}$$

where $l \in \mathbb{X}_{t,j}$ is the level of the random variable, and $s_{t,j} \in S_{t,j}$ is a specific sequence from $S_{t,j}$. We need to specify $\sum_{i=1}^{n} (|X_{t,i}| - 1) (|S_{t,i}|)$ unique probabilities for use within the transition matrix. All other probabilities can be calculated using the rule that $\sum_k P(A_k|B) = 1$. Traditionally in Markov Chain theory, the transition matrix is denoted by P; however, whenever it relates to the transition matrix of a Graphical model we shall instead use G (not to be confused with the DAG indicator $\mathcal{G}$).

Lemma 6.2.1. *The transition matrix G of a specific DBN is a $\prod_j |X_{t,j}| \times \prod_j |X_{t,j}|$ matrix with entries given by*

$$g_{tj} = \prod_{j=1}^{n} p(x_{t,j}|pa(x_{t,j})) = \prod_{j=1}^{n} \alpha_{t,j,l|s_{t,j}}.$$

Therefore each entry of G is a product of n probabilities, where n is the number of nodes in each time slice.

Each time slice in a 2TSDBN is conditionally independent of all previous time slices, given the immediate prior time slice. Therefore for time slice $t+1$, we only require information from slice t, any additional information from $t-1, t-2, \ldots$ is unnecessary. This automatically implies that the 2TSDBN can be written as a transition matrix in which entries of G are directly obtained from the factorisation of the joint density. Each entry of G is a monomial in terms of the CPTS of the original DBN.

6.2.2 2TSDBN Simple Example

Suppose we have a simple 2TSDBN with two binary nodes in each time step. Therefore $n = 2$ and we have nodes $X_{1,1}, X_{1,2}, X_{2,1}, X_{2,2}$. Suppose we have the following structure:

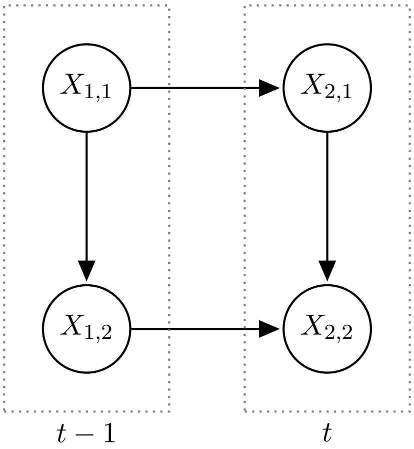


Figure 6.1: A simple structure 2TSDBN with binary nodes.

The joint distribution is simply

$$p(\mathbf{x_{t-1}}, \mathbf{x_t}) = p(x_{1,1})p(x_{1,2}|x_{1,1})p(x_{2,1}|x_{1,1})p(x_{2,2}|x_{1,2}, x_{2,1}),$$

where $\mathbf{x_{t-1}} = (x_{1,1}, x_{1,2})$ and $\mathbf{x_t} = (x_{2,1}, x_{2,2})$.

This allows us to write

$$\begin{aligned}
\text{pa}(X_{1,1}) &= \emptyset, & S_{1,1} &= \emptyset,\\
\text{pa}(X_{1,2}) &= \{X_{1,1}\}, & S_{1,2} &= \{0, 1\},\\
\text{pa}(X_{2,1}) &= \{X_{1,1}\}, & S_{2,1} &= \{0, 1\},\\
\text{pa}(X_{2,2}) &= \{X_{1,2}, X_{2,1}\}, & S_{2,2} &= \{11, 10, 01, 00\}.
\end{aligned}$$

For conciseness let us denote our conditional probabilities as

$$\begin{aligned}
p(x_{2,1} = 1|x_{1,1} = 0) &= \alpha_{2,1;1|0},\\
p(x_{2,1} = 1|x_{1,1} = 1) &= \alpha_{2,1;1|1},\\
p(x_{2,2} = 1|x_{1,2} = 0, x_{2,1} = 0) &= \alpha_{2,2;1|0,0},\\
p(x_{2,2} = 1|x_{1,2} = 0, x_{2,1} = 1) &= \alpha_{2,2;1|0,1},\\
p(x_{2,2} = 1|x_{1,2} = 1, x_{2,1} = 0) &= \alpha_{2,2;1|1,0},\\
p(x_{2,2} = 1|x_{1,2} = 1, x_{2,1} = 1) &= \alpha_{2,2;1|1,1}.
\end{aligned}$$

Entries of G are given from the equation $p(\mathbf{x_t}, \mathbf{x_{t-1}}) = p(x_{2,1}|x_{1,1})p(x_{2,2}|x_{1,2}, x_{2,1})$, leading to the transition matrix:

$$\begin{array}{c} \begin{array}{cc} x_{1,1} & x_{1,2} \end{array} \\ \begin{bmatrix} 1 & 1 \\ 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \end{array} \longrightarrow \underbrace{\begin{array}{c} \begin{array}{cccc} x_{2,1}{=}1,x_{2,2}{=}1 & x_{2,1}{=}1,x_{2,2}{=}0 & x_{2,1}{=}0,x_{2,2}{=}1 & x_{2,1}{=}0,x_{2,2}{=}0 \end{array} \\ \begin{bmatrix} \alpha_{2,1;1|1}\alpha_{2,2;1|1,1} & \alpha_{2,1;1|1}(1-\alpha_{2,2;1|1,1}) & (1-\alpha_{2,1;1|1})\alpha_{2,2;1|1,0} & (1-\alpha_{2,1;1|1})(1-\alpha_{2,2;1|1,0}) \\ \alpha_{2,1;1|1}\alpha_{2,2;1|0,1} & \alpha_{2,1;1|1}(1-\alpha_{2,2;1|0,1}) & (1-\alpha_{2,1;1|1})\alpha_{2,2;1|0,0} & (1-\alpha_{2,1;1|1})(1-\alpha_{2,2;1|0,0}) \\ \alpha_{2,1;1|0}\alpha_{2,2;1|1,1} & \alpha_{2,1;1|0}(1-\alpha_{2,2;1|1,1}) & (1-\alpha_{2,1;1|0})\alpha_{2,2;1|1,0} & (1-\alpha_{2,1;1|0})(1-\alpha_{2,2;1|1,0}) \\ \alpha_{2,1;1|0}\alpha_{2,2;1|0,1} & \alpha_{2,1;1|0}(1-\alpha_{2,2;1|0,1}) & (1-\alpha_{2,1;1|0})\alpha_{2,2;1|0,0} & (1-\alpha_{2,1;1|0})(1-\alpha_{2,2;1|0,0}) \end{bmatrix} \end{array}}_{\text{G - the family of transition matrices of this partcular BN}}$$

We have calculated the entries in G using the following formula from the definition of conditional probability:

$$\begin{aligned}
p(\mathbf{x_t}|\mathbf{x_{t-1}}) &= p(x_{2,1}, x_{2,2}|x_{1,1}, x_{1,2})\\
&= \frac{p(x_{1,1}, x_{1,2}, x_{2,1}, x_{2,2})}{p(x_{1,1}, x_{1,2})}\\
&= \frac{p(x_{1,1})p(x_{1,2}|x_{1,1})p(x_{2,1}|x_{1,1})p(x_{2,2}|x_{1,2}, x_{2,1})}{p(x_{1,1}, x_{1,2})}\\
&= \frac{p(x_{1,1})p(x_{1,2}|x_{1,1})p(x_{2,1}|x_{1,1})p(x_{2,2}|x_{1,2}, x_{2,1})}{p(x_{1,1})p(x_{1,2}|x_{1,1})}\\
&= p(x_{2,1}|x_{1,1})p(x_{2,2}|x_{1,2}, x_{2,1}).
\end{aligned}$$

Therefore the last entry in the matrix G is equivalent to finding the transition from $(x_{1,1} = 0, x_{1,2} = 0)$ to $(x_{2,1} = 0, x_{2,2} = 0)$, which is calculating the value

$$\begin{aligned}
p(x_{2,1} = 0, x_{2,2} = 0|x_{1,1} = 0, x_{1,2} = 0) &= p(x_{2,1} = 0|x_{1,1} = 0)p(x_{2,2} = 0|x_{1,2} = 0, x_{2,1} = 0)\\
&= (1 - \alpha_{2,1;1|0})(1 - \alpha_{2,2;1|0,0}).
\end{aligned}$$

Example 6.2.1. Suppose we take the 2TSDBN structure shown in Figure 6.1 and assume that all nodes are binary. Let

$$\begin{aligned}
\alpha_{2,1;1|0} &= 0.2, & \alpha_{2,2;1|0,0} &= 0.5,\\
\alpha_{2,1;1|1} &= 0.3, & \alpha_{2,2;1|0,1} &= 0.3,\\
& & \alpha_{2,2;1|1,0} &= 0.7,\\
& & \alpha_{2,2;1|1,1} &= 0.4,
\end{aligned}$$

then we obtain the transition matrix:

$$G_1 = \begin{pmatrix} 0.12 & 0.18 & 0.49 & 0.21 \\ 0.09 & 0.21 & 0.35 & 0.35 \\ 0.08 & 0.12 & 0.56 & 0.24 \\ 0.06 & 0.14 & 0.40 & 0.40 \end{pmatrix}$$

with eigenvalues equal to $(\lambda_1 = 1, \lambda_2 = 0.17881527, \lambda_3 = 0.1, \lambda_4 = 0.01118473)$ and equilibrium distribution $\pi_\infty = (0.07854406, 0.1436782, 0.4760536, 0.3017241)$.

For a moderate DBN, the dimensions of the transition matrix rapidly grow as we add more variables and more levels for each variable. Due to the non-stationary nature of some variables in the time-slice we can actually reduce the size of the transition matrix in scenarios where we have a subset of nodes which are unaffected by time:

Corollary 6.2.1.1. *The transition matrix, G, of a specific DBN for which we have a subset of stationary variables in A can be reduced from the size stated in Lemma 6.2.1 to a $\prod_j |X_{k,j}| \times \prod_j |X_{i,j}|$ matrix, where $k \in A$.*

This corollary implies that although we require all nodes to initialise the system, only nodes with dynamic relationships drive the system and therefore it is these nodes which are most important for robustness.

Example 6.2.2. The six-node subset of the Pollinator BN shown in Figure 3.4, was originally formulated as a component of a DBN (see Figure 6.2), with temporal arcs showing the year-on-year relationship of pollinator abundance within each species due to population dynamics:

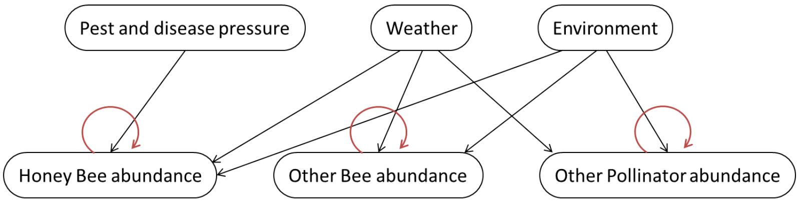


Figure 6.2: A DBN of the Pollinator subnetwork with temporal arcs shown in red, as shown in Barons et al. [2018a].

As all variables are binary in this DBN, the transition matrix of the DBN shown in Figure 6.2 is a $2^6 \times 2^6$ matrix. However, using Corollary 6.2.1.1 above we can significantly reduce the dimensions because there are three stationary variables (Pest and disease pressure, Weather and Environment). TO explicitly show the application of Corollary 6.2.1.1, let us redraw Figure 6.2 as a BN with simplified node labels:

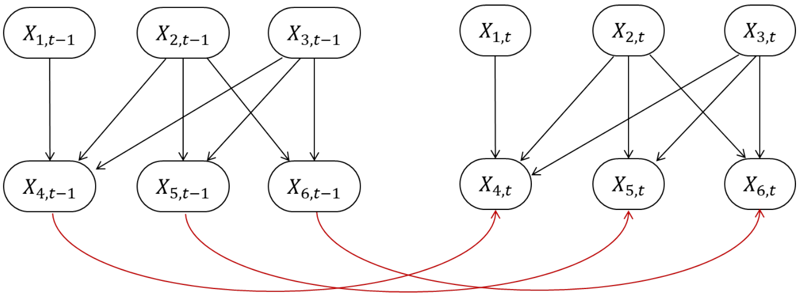


Figure 6.3: Unwrapping the Pollinator DBN from Figure 6.2 into a BN representation with simplified node labels.

Therefore when calculating the entries of G we can simplify the conditional formula as follows:

$$\begin{aligned} p(\mathbf{x}_t|\mathbf{x}_{t-1}) =& p(x_{1,t})p(x_{2,t})p(x_{3,t})p(x_{4,t}|x_{1,t}, x_{2,t}, x_{3,t}, x_{4,t-1}) \\ & p(x_{5,t}|x_{2,t}, x_{3,t}, x_{5,t-1})p(x_{6,t}|x_{2,t}, x_{3,t}, x_{6,t-1}). \end{aligned} \tag{6.1}$$

Using this information, we know that there will be repetition in the rows of the matrix, i.e. any row with the same $(x_{4,t-1}, x_{5,t-1}, x_{6,t-1})$ combination will have identical row entries because the initial starting values of nodes $(x_{1,t-1}, x_{2,t-1}, x_{3,t-1})$ do not effect the G entries, as shown in Equation 6.1. As all nodes are binary, suppose they take values $(0,1)$, then the following two rows in G are equivalent:

$$\begin{aligned}
&p(x_{1,t-1}=1, x_{2,t-1}=0, x_{3,t-1}=0, x_{4,t-1}=1, x_{5,t-1}=0, x_{6,t-1}=1)\\
=&p(x_{1,t-1}=1, x_{2,t-1}=1, x_{3,t-1}=1, x_{4,t-1}=1, x_{5,t-1}=0, x_{6,t-1}=1).
\end{aligned}$$

Therefore all eight rows relating to the initial values $p(x_{1,t-1}, x_{2,t-1}, x_{3,t-1}, x_{4,t-1} = 1, x_{5,t-1} = 0, x_{6,t-1} = 1)$ will be equivalent and can be reduced to a single row by looking solely at combinations between the non-stationary variables. This leaves us with a $2^3 \times 2^6$ matrix instead of the original $2^6 \times 2^6$.

Corollary 6.2.1.1 is an important result as it indicates that we do not need to elicit nodes within a time-slice which are not dynamic because the memory of the system forgets these values over time. It is therefore recommended that elicitation prioritisation is given to nodes with dynamic relationships.

6.3 Diameter and Stochastic Matrices

The new measures introduced in Section 4.3 underpinned by the total variation distance can be applied without edit to DBNs. For BNs we compared CPTs, whereas for DBNs we compare transition matrices between individual time-slices of the DBN. Therefore, the interpretation of the diameter and other measures differ slightly. For example, when $n = n'$ and $G_\infty = \lim_{n\to\infty} G^n$ is the equilibrium stochastic matrix, then all rows in the transition matrix are equivalent (to the equilibrium distribution), thus $d^+(G_\infty) = 0$ and all entries in the matrix $D^+(G_\infty)$ will also be zero. The diameter of a transition matrix can be thought of as a measure of how far away a stochastic matrix is from its equilibrium distribution. The smaller the value of the diameter (for a transition matrix), the closer we are to equilibrium.

Example 6.3.1. Using G_1 from Example 6.2.1 we can calculate

$$d^+(G_1) = 0.21 \qquad \text{and} \qquad D^+(G_1) = \begin{pmatrix} 0.00 & 0.17 & 0.10 & 0.19\\ 0.17 & 0.00 & 0.21 & 0.10\\ 0.10 & 0.21 & 0.00 & 0.18\\ 0.19 & 0.10 & 0.18 & 0.00 \end{pmatrix}.$$

Earlier in Theorem 5.1.1 we proved that we could bound the distance of a receiving mass function by the diameter of the transition matrix multiplied by the distance between the donating mass functions, both of which can be calculated with relative ease. This result can be generalised further to determine the distance between receiving mass functions over a period of more than one time step.

Corollary 6.3.0.1. *When $n = n'$;* $\quad d_V(\rho_1(t), \rho_2(t)) \leq \{d^+(G)\}^t \, d_V(\pi_1, \pi_2)$.

Proof. By Theorem 5.1.1

$$\begin{aligned}
\text{at time } t = 1: \qquad & d_V(\rho_1(1), \rho_2(1)) \leq d^+(G) d_V(\pi_1, \pi_2), \\
\text{at time } t = 2: \qquad & d_V(\rho_1(2), \rho_2(2)) \leq d^+(G) d_V(\rho_1(1), \rho_2(2)), \\
& \qquad\qquad\qquad\quad \leq d^+(G)^2 d_V(\pi_1, \pi_2) \\
\vdots \qquad & \\
\text{at time } t = s: \qquad & d_V(\rho_1(s), \rho_2(s)) \leq d^+(G) d_V(\rho_1(s-1).\rho_2(s-1)) \\
& \qquad\qquad\qquad\quad \leq d^+(G)^s d_V(\pi_1, \pi_2).
\end{aligned}$$

Therefore by successive substitution we obtain the result

$$d_V(\rho_1(t), \rho_2(t)) \leq \{d^+(G)\}^t d_V(\pi_1, \pi_2)$$

as desired. □

This result is extremely useful because we can now determine the sequential effects on time-slices that a slight misspecification of a probability can cause throughout the DBN.

Corollary 6.3.0.2. *When G is a square stochastic matrix $d^+(G^t) \leq \{d^+(G)\}^t$, for $t = 1, 2, \ldots, n$.*

Proof. The entries of the matrix G^t are simply weighted averages of the rows of G^{t-1} indicating that the distance between rows must be shrinking because each row must sum to 1. □

6.4 Time homogeneous forecasts

When $n' = n$ we are often interested in a t-step ahead margin $d_V\left(\boldsymbol{\pi}_1 G_1^t, \boldsymbol{\pi}_2 G_2^t\right)$. There are then two types of bound that can be calculated. The first simply iterates using the bounds from Equation 5.1 and Equation 5.2. For example if $\boldsymbol{\rho}_1(t) =$

$\boldsymbol{\pi}_1 G_1^t, \boldsymbol{\rho}_2(t) = \boldsymbol{\pi}_2 G_2^t$ then from Equation 5.1

$$\begin{aligned}
&d_V\left(\boldsymbol{\rho}_1(t), \boldsymbol{\rho}_2(t)\right)\\
&\quad\le d_V^+\left(P_1, P_2\right) + d_V\left(\boldsymbol{\rho}_1(t-1), \boldsymbol{\rho}_2(t-1)\right) d_V^*\left(P_1, P_2\right)\\
&\quad\le (1 + d_V^*\left(P_1, P_2\right)) d_V^+\left(P_1, P_2\right) + d_V\left(\boldsymbol{\rho}_1(t-2), \boldsymbol{\rho}_2(t-2)\right) \left\{d_V^*\left(P_1, P_2\right)\right\}^2\\
&\quad\vdots\\
&\quad\le \left(\sum_{s=1}^{t} \left\{d_V^*\left(P_1, P_2\right)\right\}^{s-1}\right) d_V^+\left(P_1, P_2\right) + d_V\left(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2\right) \left\{d_V^*\left(P_1, P_2\right)\right\}^t
\end{aligned}$$

Thus,

$$d_V\left(\boldsymbol{\rho}_1(t), \boldsymbol{\rho}_2(t)\right) \le \left(\frac{1 - \left\{d_V^*\left(G_1, G_2\right)\right\}^t}{1 - d_V^*\left(G_1, G_2\right)}\right) d_V^+\left(G_1, G_2\right) + d_V(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2) \left\{d_V^*\left(G_1, G_2\right)\right\}^t .$$

Note that this bound can be straightforwardly evaluated from the given matrices G_1, G_2 and the two donating margins $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2$. Also, that this can be rewritten in the form

$$d_V\left(\boldsymbol{\rho}_1(t), \boldsymbol{\rho}_2(t)\right) \le \left(\frac{d_V^+\left(G_1, G_2\right)}{1 - d_V^*\left(G_1, G_2\right)}\right) + \left\{d_V(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2) - \frac{d_V^+\left(G_1, G_2\right)}{1 - d_V^*\left(G_1, G_2\right)}\right\} \left\{d_V^*\left(G_1, G_2\right)\right\}^t .$$

So as a function of the length of the future time step t this is a constant and a term exponentially decaying with t. Note that for a time inhomogeneous stochastic matrix of known form similar recurrences can be derived simply by substitution.

6.4.1 Recursions using equilibrium distributions

The second t-step ahead margin is a completely different recursion which can be derived using the triangle inequality and the vectors of the two equilibrium mass functions $\boldsymbol{\pi}_{1\infty}, \boldsymbol{\pi}_{2\infty}$ of the two stochastic matrices $G_{1\infty}, G_{2\infty}$. We observe that as all the rows of $G_{1\infty}$ and all the rows of $G_{2\infty}$ are duplicated

$$d_V\left(\boldsymbol{\pi}_1 G_{1\infty}, \boldsymbol{\pi}_2 G_{2\infty}\right) = d_V\left(\boldsymbol{\pi}_{1\infty}, \boldsymbol{\pi}_{2\infty}\right).$$

So

$$\begin{aligned}
d_V\left(\boldsymbol{\rho}_1(t), \boldsymbol{\rho}_2(t)\right) &= d_V\left(\boldsymbol{\pi}_1 G_1^t, \boldsymbol{\pi}_2 G_2^t\right)\\
&\le d_V\left(\boldsymbol{\pi}_1 G_1^t, \boldsymbol{\pi}_1 G_{1\infty}\right) + d_V\left(\boldsymbol{\pi}_1 G_{1\infty}, \boldsymbol{\pi}_2 G_{2\infty}\right) + d_V\left(\boldsymbol{\pi}_2 G_{2\infty}, \boldsymbol{\pi}_2 G_2^t\right)\\
&= d_V\left(\boldsymbol{\pi}_{1\infty}, \boldsymbol{\pi}_{2\infty}\right) + d_V\left(\boldsymbol{\pi}_1 G_1^t, \boldsymbol{\pi}_1 G_{1\infty}\right) + d_V\left(\boldsymbol{\pi}_2 G_{2\infty}, \boldsymbol{\pi}_2 G_2^t\right).
\end{aligned}$$

Now note that by definition of an equilibrium mass function, for $i = 1, 2$

$$\boldsymbol{\pi}_i G_{i\infty} = \boldsymbol{\pi}_{i\infty} = \boldsymbol{\pi}_{i\infty} G_i^t.$$

Thus

$$\begin{aligned} d_V\left(\boldsymbol{\rho}_1(t), \boldsymbol{\rho}_2(t)\right) &= d_V\left(\boldsymbol{\pi}_{1\infty}, \boldsymbol{\pi}_{2\infty}\right) + d_V\left(\boldsymbol{\pi}_1 G_1^t, \boldsymbol{\pi}_{1\infty} G_1^t\right) + d_V\left(\boldsymbol{\pi}_{2\infty} G_2^t, \boldsymbol{\pi}_2 G_2^t\right) \\ &\leq d_V\left(\boldsymbol{\pi}_{1\infty}, \boldsymbol{\pi}_{2\infty}\right) + d_V\left(\boldsymbol{\pi}_1, \boldsymbol{\pi}_{1\infty}\right) d(G_1^t) + d_V\left(\boldsymbol{\pi}_{2\infty}, \boldsymbol{\pi}_2\right) d^+(G_1^t), \end{aligned}$$

where from previous results we know that, for $i = 1, 2$; $d(G_i^t) \leq \{d^+(G_i)\}^t$

$$d_V\left(\boldsymbol{\rho}_1(t), \boldsymbol{\rho}_2(t)\right) \leq d_V\left(\boldsymbol{\pi}_{1\infty}, \boldsymbol{\pi}_{2\infty}\right) + d_V\left(\boldsymbol{\pi}_1, \boldsymbol{\pi}_{1\infty}\right)\left\{d^+(G_1)\right\}^t + d_V\left(\boldsymbol{\pi}_{2\infty}, \boldsymbol{\pi}_2\right)\left\{d^+(G_2)\right\}^t.$$

In particular

$$d_V\left(\boldsymbol{\rho}_1(t), \boldsymbol{\rho}_2(t)\right) \leq d_V\left(\boldsymbol{\pi}_{1\infty}, \boldsymbol{\pi}_{2\infty}\right) + 2\kappa^t$$

where $\kappa \triangleq \max\{d^+(G_1), d^+(G_2)\}^t$. So as in the other bound, this bound is a constant and a term exponentially decaying with t. Note that this bound requires us to calculate $\boldsymbol{\pi}_{1\infty}, \boldsymbol{\pi}_{2\infty}$.

6.4.2 Disturbed equilibria

The above t step ahead bounds are useful. Nevertheless, all too often the context appears highly unpredictable when looking to the long term future. The question is whether or not it is possible to fold this into a robustness analysis.

There is actually a very simple way of doing this. Suppose that we have elicited a model with input from a panel of experts and a consensus is to use $(\boldsymbol{\pi}_1, G_1)$. However, there is some disagreement about the fine detail of these judgements so that various experts in the panel believe that another donated margin $\boldsymbol{\pi}_2$ and another transition matrix G_2 might be more appropriate but where

$$d_V\left(\boldsymbol{\pi}_1, \boldsymbol{\pi}_2\right) < \delta, \qquad d_V^+\left(G_1, G_2\right) < \eta_+ \qquad \text{and } d_V^*\left(G_1, G_2\right) < \eta_*$$

Then for example

$$d_V\left(\boldsymbol{\rho}_1(t), \boldsymbol{\rho}_2(t)\right) \leq \left(\frac{1 - \eta_*^t}{1 - \eta_*}\right)\eta_+ + \delta\eta_*^t$$

This looks as though there is a long term consequence of the misspecification of G_1 as measured by $\left(\frac{1-\eta_*^t}{1-\eta_*}\right)\eta_+$.

Yet suppose all experts accept that there is a probability ε that at some as yet unknown future time T the whole process will change for all time to a new time inhomogeneous process governed by a sequence of as yet unknown transition matrices $\{Q(s) : s \geq T\}$.

If there is consensus that this is likely, then $(\boldsymbol{\pi}_1, G_1)$ or $(\boldsymbol{\pi}_2, G_2)$ should be replaced and instead use transitions at time s of the form

$$\begin{aligned}\overline{G}_1(s) &\triangleq (1-\varepsilon)G_1 + \varepsilon Q(s),\\ \overline{G}_2(s) &\triangleq (1-\varepsilon)G_2 + \varepsilon Q(s).\end{aligned}$$

Let us employ $(\overline{\boldsymbol{\rho}}_1(t), \overline{\boldsymbol{\rho}}_2(t))$ as the forecasters using the current settings $(\boldsymbol{\pi}_1, G_1)$ or $(\boldsymbol{\pi}_2, G_2)$ but allowing for the fact that everything could switch.

The point now is to note that *whatever the new transitions* $\{Q(s) : s \geq T\}$ no-one will want to keep their original transitions $(\boldsymbol{\pi}_1, G_1)$ or $(\boldsymbol{\pi}_2, G_2)$. All experts will change their stated judgements.

Thus if $T = 0$, an event of probability ε, then the forecasts of $(\overline{\boldsymbol{\rho}}_1(1), \overline{\boldsymbol{\rho}}_2(1))$ using $\overline{G}_1(s)$ and $\overline{G}_1(s)$ will be the same. So with probability ε, $d_v\left(\overline{\boldsymbol{\rho}}_1(1), \overline{\boldsymbol{\rho}}_2(1)\right) = 0$ and also for $d_v\left(\overline{\boldsymbol{\rho}}_1(s), \overline{\boldsymbol{\rho}}_2(s)\right) = 0$ for all $s \geq 1$. Alternatively, with probability $(1-\varepsilon)$ nothing will change and

$$d_V\left(\overline{\boldsymbol{\rho}}_1(1), \overline{\boldsymbol{\rho}}_2(1)\right) = d_V\left(\boldsymbol{\rho}_1(1), \boldsymbol{\rho}_2(1)\right).$$

So we see that

$$\begin{aligned}d_V\left(\overline{\boldsymbol{\rho}}_1(1), \overline{\boldsymbol{\rho}}_2(1)\right) &= (1-\varepsilon) d_V\left(\boldsymbol{\rho}_1(1), \boldsymbol{\rho}_2(1)\right)\\ &\leq (1-\varepsilon)\left\{\left(\frac{1-\eta_*}{1-\eta_*}\right)\eta_+ + \delta\eta_*\right\}.\end{aligned}$$

Iterating this argument we then obtain a new bound of the form

$$d_V\left(\overline{\boldsymbol{\rho}}_1(t), \overline{\boldsymbol{\rho}}_2(t)\right) \leq (1-\varepsilon)^t \left\{\left(\frac{1-\eta_*^t}{1-\eta_*}\right)\eta_+ + \delta\eta_*^t\right\}.$$

Note that here we retrieve a robustness measure which implicitly exponentially discounts the weights of the effects on utility scores obtained into the future: in an exact analogy to the exponential discounting of rewards commonly used in Markov decision theory.

6.5 Discussion

DBNs have become an increasingly popular graphical model for decision support systems, however it is very rare in practice to find data to fully populate all of the conditional probability tables in the model. Expert elicitations have been vastly studied and we highly recommend following a structured protocol to find reproducible and reliable results. However, users often find themselves in the situation where there are limited resources, but a large quantity of probabilities they are required to elicit. In these circumstances it would be optimal to proceed by focussing attention on inputs which most affect the outputs of the model. To do this we must somehow rank the importance of the attributes or else perform a suitable robustness analysis to measure the effects on the utility functions or decisions that the model outputs.

It is critical to develop an awareness of the features within an elicitation that may have a strong impact upon the outputs of the model. Once these variables or probabilities have been ascertained, the model creator can provide more time and resources to the elicitation of these values to ensure they are accurate as possible. In this paper we have provided some new distance metrics and other measures for determining the dissimilarity between matrices in order to quantify the impact of erroneous CPT probabilities on the outputs of the model.

We have illustrated that since it is possible to rewrite our DBN as a transition matrix with accompanying initial probability distribution the fundamental basis underpinning dynamic Bayesian networks and finite homogeneous Markov Chains. This succinctly whittles down our graphical model from a visual representation with a CPT on each node, to one single matrix encoding all CPT probabilities.

The measures and bounds we have provided are based upon the total variation distance, which for the discrete case is easy to calculate in theory, and is also quick to assess visually. The bounds we have found allow DBN users to tailor their model process and incorporate robustness checks into their protocols so that when feeding back results they can discuss the performance of the DBN and the stability of the recommended decisions.

Chapter 7

Conclusions

During the course of this work, we hope to have emphasised the necessity for scrutinising the robustness in a BN. To appreciate the strong and weak features of a model, users should undertake some form of analysis to deduce areas which require further attention. Commonly, sensitivity analyses have been recommended for BNs: these perturb a single parameter to determine effects on output variables. Recently, these algorithms have been extended to cope with multiple simultaneous parameter perturbations. Unfortunately, the efficacy of sensitivity analyses relies on the ability to cover the entire space of possible perturbations, which for large models is unrealistic. An additional drawback is that sensitivity analyses can only be performed once the whole model has been specified, and should issues be discovered, then the full model must be reconstructed (a time consuming and difficult task). The work presented in this thesis aims to fill this gap by creating structured robustness studies which can be performed during model creation, thereby allowing users to tailor a bespoke BN model which is robust to anomalies on the first iteration. Recall that our primary concern is BNs applied within decision support systems, meaning that we usually have a set of pre-specified target variables of interest. This contraction of focus from a large model to a subset of the total variables enables us to harness the convexity of probability measures such as the total variation distance.

Contrary to current work in the field of BN global robustness, we have argued in this thesis that the total variation is a conventional divergence measure which should be chosen given some important benefits. Recent studies have pioneered the use of the Kullback-Leibler divergence or the Chan-Darwiche measure, both of which rely on logarithmic probabilities or ratios of probabilities. Regrettably, the ability to accurately specify very small probabilities is difficult to accomplish using either expert elicitation or data. Additionally, very small discrepancies between probability distributions can be exaggerated. Although occasionally difficult to explicitly derive formulae for the impacts of total variation, it is possible to combine useful bounds that limit the maximum derivation in total variation.

We have introduced the diameter, a univariate measure of conditional independence strength, which can be elicited directly or if a conditional probability table is fully specified it can be calculated with ease. To elicit, the diameter is the difference between the probability of the worst case scenario and the best case scenario occurring. From data, it is the maximum variation distance between each pairwise combination of rows within a conditional probability table. With this quantity we can determine: whether to omit weak relationships between parents and children nodes; how many levels of a variable to include in the model; whether amalgamating levels of a variable is a justifiable simplification; and the temporal effects of perturbations over time. This measure is therefore very adaptable as it can be used for both structural and probabilistic misspecifications during the creation of the BN process.

Using our new definitions and bounds proven in this work we have proven some key results. Firstly, we have provided a formal validation for the popular folk theorem stating that it is unnecessary to include a long ancestral influence chain for a variable of interest. In fact, errors contract between each variable and its child, meaning that misspecified probabilities are only relevant if in the immediate parent set. Secondly, by eliciting a diameter we have a two-fold use from the results: whether to include an arc in the structure of the BN, and if so, whether the relationship is strong enough to warrant multiple levels or a binary case. Lastly, we have shown that if we have a decomposable BN then we can create a junction tree, allowing us to find a unique simple path between a donating and receiving variable and thus quantify the impact of perturbing one on the other. In our final section we have introduced a method of transforming a BN into a finite time non-homogeneous Markov Chain: an initial probability and a transition matrix. We can therefore directly apply our results to the dynamic BN setting and have shown that errors contract over time, and that we can cap how far back to examine in the model.

Within the three years of completing this research, we believe we have proved the necessary results to undertake a comprehensive robustness analysis of a discrete BN, in a manner which interfaces well with standard practices of elicitation. There are many interesting ideas and applications extending this work which have yet to be investigated. Below we shall briefly discuss a few of the most promising of these.

Throughout this thesis we have highlighted key works which provide results for BN model selection or model simplification, however we have been unable to compare previous methodology to our results due to time restrictions.

Although our examples have been small for simplicity, our methods can be easily scaled up to deal with vast BN structures. One element of interest of large distributive systems that we have seen in the Food Security IDSS, are the multiple modules within a system communicating with each other. Target variables are then defined to be those variables in the system which are outputs of one module, but contribute to another receiving component. This scenario has yet to be framed formally; nevertheless, the technology is analogous to the simplified context specific situation discussed in the penultimate section of this thesis.

In enormous BNs we may want to focus attention on a small component of the larger network. How can we perform this localised interest in a robust manner? Using our total variation approach we can query the distributivity of large systems to learn about a subset of nodes and their dependence relationships. By selecting pre-specified target variables, we can determine which information and relationships can be systematically rejected and also emphasise any relationships in the system which it is paramount to elicit/model accurately. Although this is a big data problem, our methodology outlined here can be extended to look at stability of smaller distributive systems.

Further investigation could be implemented to tighten the bounds provided here and tailor them for context-specific BNs such as object-oriented BNs. In modern technology associated with large systems, another form of simplification is to assume that certain substructures within the BN are repeated. A 2TSDBN is a very simple example of this, where we assume structural and some CPT equivalences across time-slices. If we assume that there are similarities within modules, then there is a natural trade-off: we can either elicit all CPTs independently which requires considerable time allocation, or we can assume similarities and therefore elicit the one repeated CPT with greater accuracy in a similar amount of time. As with any compromise the choice is not clear, and we question how to decide when CPTs are sufficiently comparable with one another to be considered the same object. There is very little literature covering this conundrum; however, experienced BN modellers have their own intuition about such scenarios. We have provided the means in this thesis to answer questions such as this; nevertheless, we still need to embed this material into formal protocols outlining appropriate assumptions for context-specific settings.

The methodology provided herein has been applied to discrete systems although the underlying foundation enables the process to be effortlessly expanded to continuous or mixed BNs. This thesis does not touch upon the topic of robustness to parameter estimation, but we note that it is closely linked to variation distance of the parameter distribution. These parameter probabilities are continuous random variables and therefore the analysis must first be reproduced for continuous random variables before it can be adopted in this mixed setting.

Now that I have experience of eliciting probabilities for graphical models using two methods (the IDEA protocol and Cooke's method), I am more aware of the ambiguities in estimations and dissention between experts that can arise. These are not only about the setting of certain probabilities within a BN, but the controversies of one BN structural compared to another. There are many other well-known elicitation methods not detailed in this work, each of which has its own advantages. In most methods, as well as eliciting the best estimates we collect information about the uncertainty of each expert. These confidence bounds, expert weightings and other outputs of the method are typically forgotten after the probabilities have been set in the CPTs, but we wonder whether we can somehow use this extra information to inform the variation distance bounds.

Although BNs are the chosen graphical model for this thesis, they are not the only choice of Markov graph. There are many other graphical classes to which our technology could be applied, so long as there is a directed element in the model. Efforts at the University of Warwick Statistics Department are already focussing at a technology transfer of my robustness results associated with BNs and applying these to graphs such as Chain Event Graphs.

Clearly there are many ideas and applications of our methodology that can be undertaken to make a comprehensive suite of BN robustness analyses for any given context. In this thesis we have provided a solid foundation and framework, setting the scene for further research in the future.

Bibliography

D. W. Albrecht, A. E. Nicholson, and C. Whittle. Structural sensitivity for the knowledge engineering of Bayesian networks. In L. C. van der Gaag and A. J. Feelders, editors, *Proceedings of the Seventh European Workshop on Probabilistic Graphical Models (PGM 2014)*, volume 8754, pages 1–16. Springer, Cham, 2014.

S. A. Andersson, D. Madigan, and M. D. Perlman. A characterization of Markov equivalence classes for acyclic digraphs. *The Annals of Statistics*, 25(2):505–541, 1997.

W. P. Aspinall. Structured elicitation of expert judgment for probabilistic hazard and risk assessment in volcanic eruptions. *Statistics in Volcanology*, 1:15–30, 2006.

W. P. Aspinall and R. M. Cooke. Quantifying scientic uncertainty from expert judgement elicitation. In *Risk and Uncertainty Assessment for Natural Hazards*, chapter 4, pages 64–99. Cambridge University Press, 2013.

M. J. Barons, X. Zhong, and J. Q. Smith. Dynamic Bayesian networks for decision support and sugar food security. *CRiSM Report, University of Warwick*, 14(18), 2014.

M. J. Barons, A. M. Hanea, S. K. Wright, K. C. R. Baldock, L. Wilfert, D. Chandler, S. Datta, J. Fannon, C. Hartfield, A. Lucas, J. Ollerton, S. G. Potts, and N. L. Carreck. Assessment of the response of pollinator abundance to environmental pressures using structured expert elicitation. *Journal of Apicultural Research*, 57 (5):593–604, 2018a.

M. J. Barons, S. K. Wright, and J. Q. Smith. Eliciting probabilistic judgements for integrating decision support systems. In Dias L. C., A. Morton, and J. Quigley, editors, *Elicitation: The science and art of structuring judgement*, chapter 17, pages 445–478. Springer, 2018b.

J. M. Bernardo and A. F. M. Smith. *Bayesian Theory.* Wiley, Chichester, 1994.

T. Boneh. *Ontology and Bayesian decision networks for supporting the meteorological forecasting process.* PhD thesis, Clayton School for Information Technology, Monash University, 2010.

P. Brémaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation and Queues.* Springer-Verlag, New York, 1999.

M. A. Burgman. *Trusting judgements: how to get the best out of experts.* Cambridge University Press, 2016.

R. Cano, C. Sordo, and J. M. Gutiérrez. Applications of Bayesian networks in meteorology. In *Advances in Bayesian Networks*, pages 309–328. Springer, Berlin, Heidelberg, 2004.

H. Chan and A. Darwiche. Sensitivity analysis in bayesian networks: From single to multiple parameters. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, pages 67–75. AUAI Press, 2004.

H. Chan and A. Darwiche. A distance measure for bounding probabilistic belief change. *International Journal of Approximate Reasoning*, 38(2):149–174, 2005.

R. T. Clemen and T. Reilly. *Making Hard Decisions with Decision Tools.* Duxbury Press, 2001.

J. Collins. Murpa report: Bayesian networks and food security. Technical report, Monash University, 2016.

A. C. Constantinou, M. Freestone, W. Marsh, N. Fenton, and J. Coid. Risk assessment and risk management of violent reoffending among prisoners. *Expert Systems with Applications*, 42(21):7511–7529, 2015.

R. M. Cooke. *Experts in uncertainty: opinion and subjective probability in science.* Oxford University Press, 1991.

V. M. Coupé and L. C. van der Gaag. Properties of sensitivity analysis of Bayesian belief networks. *Annals of Mathematics and Artificial Intelligence*, 36(4):323–356, 2002.

R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic Networks and Expert Systems.* Springer-Verlag, New York, 1999.

R. G. Cowell, R. J. Verrall, and Y. K. Yoon. Modeling operational risk with Bayesian networks. *Journal of Risk and Insurance*, 74(4):795–827, 2007.

D. R. Cox and H. D. Miller. *The Theory of Stochastic Processes.* Methuen & Co. Ltd, London, 1965.

J. Cussens. Bayesian network learning with cutting planes. In F. Cozman and A. Pfeffer, editors, *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence*, pages 153–160, 2011.

N. Dalkey and O. Helmer. An experimental application of the Delphi method to the use of experts. *Management Science*, 9(3):458–467, 1963.

B. De Finetti. *Theory of Probability.* Wiley, Chichester, 1974.

L. C. Dias, A. Morton, and J. Quigley, editors. *Elicitation: The science and art of structuring judgement.* Springer, 2018.

EFSA. Guidance on expert knowledge elicitation in food and feed safety risk assessment. *EFSA Journal*, 12(6):3734, 2014.

FAO. Food outlook: Biannual report on global food markets. Technical report, FAO of the United Nations, Trade and Markets Division, 2014.

J. Fenlon. Feasibility study on the use of Bayesian networks for food security [draft]. Technical report, University of Warwick.

N. Fenton and M. Neil. *Risk assessment and decision analysis with Bayesian networks.* CRC Press, 2012.

S. French. Group consensus probability distributions: a critical survey (with discussion). In J. M. Bernardo, M. H. DeGroot, D. V. Lindley, and A. F. M. Smith, editors, *Bayesian Statistics 2*, pages 183–201. North-Holland, Amsterdam, 1985.

S. French, J. Maule, and N. Papamichail. *Decision behaviour, analysis and support.* Cambridge University Press, 2009.

N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29(2-3):131–163, 1997.

N. Friedman, M. Linial, I. Nachman, and D. Pe'er. Using Bayesian networks to analyze expression data. *Journal of Computational Biology*, 7(3-4):601–620, 2000.

P. H. Garthwaite, J. B. Kadane, and A. O'Hagan. Statistical methods for eliciting probability distributions. *Journal of the American Statistical Association*, 100 (470):680–701, 2005.

D. Geiger and J. Pearl. Logical and algorithmic properties of conditional independence and graphical models. *The Annals of Statistics*, 21(4):2001–2021, 1993.

A. L. Gibbs and F. E. Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

Gómez-Villegas, M. A., P. Main, and R. Susi. The effect of block parameter perturbations in gaussian Bayesian networks: sensitivity and robustness. *Information Sciences*, 222:439–458, 2013.

J. P. Gosling. SHELF: the Sheffield elicitation framework. In Dias L. C., A. Morton, and J. Quigley, editors, *Elicitation: The science and art of structuring judgement*, chapter 4, pages 61–93. Springer, 2018.

P. Gustafson and L. Wasserman. Local sensitivity diagnostics for Bayesian inference. *The Annals of Statistics*, 23(6):2153–2167, 1995.

A. M. Hanea, M. F. McBride, M. A. Burgman, B. C. Wintle, F. Fidler, L. Flander, C. R. Twardy, B. Manning, and S. Mascaro. I$_{\text{nvestigate}}$ d$_{\text{iscuss}}$ e$_{\text{stimate}}$ a$_{\text{ggregate}}$ for structured expert judgement. *International Journal of Forecasting*, 33:267–279, 2017.

V. Hemming, M. A. Burgman, A. M. Hanea, M. F. McBride, and B. C. Wintle. A practical guide to structured expert elicitation using the idea protocol. *Methods in Ecology and Evolution*, 9:169–180, 2018.

House of Commons - Environment, Food and Rural Affairs Committee. Food security: HC 243, second technical report of session 2014-15. Technical report, TSO (The Stationery Office), 2014.

F. V. Jensen and T. D. Nielsen. *Bayesian Networks and Decision Graphs. (Second Edition)*. Springer Science & Business Media, 2007.

J. Jewson, J. Q. Smith, and C. Holmes. Principles of Bayesian inference using general divergence criteria. *Entropy*, 20(6):442, 2018.

S. Johnson and K. Mengersen. Integrated Bayesian network framework for modeling complex ecological issues. *Integrated Environmental Assessment and Management*, 8(3):480–490, 2012.

S. Johnson, S. LowChoy, and K. Mengersen. Integrating Bayesian networks and geographic information systems: Good practice examples. *Integrated environmental assessment and management*, 8(3):473–479, 2012.

D. Jones. MURPA 2015 report, UK fish supply DBN. Technical report, Monash University, 2015.

D. Kahneman and A. Tversky. On the psychology of prediction. *Psychological Review*, 80(4):237, 1973.

D. Kahneman, P. Slovic, and A. Tversky, editors. *Judgment under uncertainty: heuristics and biases.* Cambridge University Press, 1982.

K. B. Korb and A. E. Nicholson. *Bayesian artificial intelligence, second edition.* CRC press, 2010.

K. B. Laskey. Sensitivity analysis for probability assessments in Bayesian networks. In *IEEE Transactions on Systems, Man, and Cybernetics*, volume 25, pages 901–909, 1995.

K. B. Laskey and S. M. Mahoney. Network engineering for agile belief network models. *IEEE Transactions on Knowledge and Data Engineering*, 12(4):487–498, 2000.

S. L. Lauritzen. *Graphical models.* Clarendon Press, Oxford, 1996.

S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 157–224, 1988.

M. Leonelli and J. Q. Smith. Dynamic uncertainty handling for coherent decision making in nuclear emergency response. In *Proceedings of the Winter Meeting of the American Nuclear Society*, 2013.

M. Leonelli, C. Görgen, and J. Q. Smith. Sensitivity analysis in multilinear probabilistic models. *Information Sciences*, 411:84–97, 2017.

S. Mascaro, A. E. Nicholson, and K. B. Korb. Anomaly detection in vessel tracks using Bayesian networks. *International Journal of Approximate Reasoning*, 55(1): 84–98, 2014.

T. Meacham. Food price spikes and global food markets. *Insight*, 3, 2014.

I. Milns, C. M. Beale, and V. A. Smith. Revealing ecological networks using Bayesian network inference algorithms. *Ecology*, 91(7):1892–1899, 2010.

P. A. P Moran. *An introduction to probability theory.* Clarendon Press, Oxford, 1968.

D. Murins. Dynamic Bayesian networks for decision support and meat food security. Technical report, University of Warwick, 2015.

R. Nagarajan, M. Scutari, and S. Lèbre. *Bayesian networks in R with applications in systems biology.* New York: Springer-Verlag, 2013.

A. Nicholson, O. Woodberry, and C. Twardy. The "native fish" Bayesian networks. Technical report, Technical Report 2010/3, Bayesian Intelligence., 2010.

A. E. Nicholson and M. J. Flores. Combining state and transition models with dynamic Bayesian networks. *Ecological Modelling*, 222(3):555–566, 2011.

A. E. Nicholson and N. Jitnah. Using mutual information to determine relevance in Bayesian networks. In *Pacific Rim International Conference on Artificial Intelligence*, pages 399–410. Springer, Berlin, Heidelberg, 1998.

R. S. Nickerson. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology*, 2(2):175, 1998.

A. O'Hagan, C. E. Buck, A. Daneshkhah, J. R. Eiser, P. H. Garthwaite, D. J. Jenkinson, J. E. Oakley, and T. Rakow. *Uncertain judgements: eliciting experts' probabilities.* John Wiley & Sons, 2006.

B. O'Neill. Importance sampling for Bayesian sensitivity analysis. *International Journal of Approximate Reasoning*, 50(2):270–278, 2009.

J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference.* Morgan Kaufmann Publishers Inc., 1988.

L. D. Phillips. A theory of requisite decision models. *Acta Psychologica*, 56:29–48, 1984.

L. D. Phillips. Group elicitation of probability distributions: are many heads better than one? In J. Shanteau, B. A. Mellers, and D. A. Schum, editors, *Decision Science and Technology: Reflections on the Contributions of Ward Edwards.*, pages 313–330. Springer, 1999.

R. D. Reiss. Approximation distribution of order statistics: with applications to nonparametric statistics. 1989.

S. Renooij. Bayesian network sensitivity to arc-removal. In *Proceedings of the Fifth European Workshop on Probabilistic Graphical Models, Helsinki.*, page 233, 2010.

M. Scutari and J. B. Denis. *Bayesian networks: with examples in R.* CRC press, 2014.

E. Seneta. *Non-negative matrices and Markov chains, revised printing.* Springer Science & Business Media, 2006.

J. Q. Smith. *Bayesian decision analysis: principles and practice.* Cambridge University Press, 2010.

J. Q. Smith and A. Daneshkhah. On the robustness of Bayesian networks to learning from non-conjugate sampling. *International Journal of Approximate Reasoning*, 51(5):558–572, 2010.

J. Q. Smith and F. Rigat. Isoseparation and robustness in finite parameter Bayesian inference. *Annals of the Institute of Statistical Mathematics*, 64(3):495–519, 2012.

J. Q. Smith, M. J. Barons, and M. Leonelli. Coherent frameworks for statistical inference serving integrating decision support systems. *arXiv preprint arXiv:1507.07394*, 2015.

L. Walls and J. Quigley. Building prior distributions to support Bayesian reliability growth modelling using expert judgement. *Reliability engineering and system safety*, 74(2):117–128, 2001.

S. K. Wright and J. Q. Smith. Robustness, total variation and Bayesian networks. *arXiv preprint arXiv:1811.07179*, 2018.

J. C. Zaragoza, L. E. Sucar, and E. F. Morales. A two-step method to learn multi-dimensional Bayesian network classifiers based on mutual information measures. In *Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference (FLAIRS)*, pages 644–649, 2011.