Estimation in Causal Graphical Models

Alireza Daneshkhah

A thesis submitted for the degree of Doctor of Philosophy

Department of Statistics

University of Warwick

Coventry CV4 7AL

February 2004

Contents

List of Figures

DEDICATION

To my parents

for letting me pursue my dream

for so long

so far away from home

and

To my wife, Tabassom

for giving me

new dreams to pursue

Abstract

Pearl (2000), Spirtes et al (1993) and Lauritzen (2001) set up a new framework to encode the causal relationships between the random variables by a causal Bayesian network. The estimation of the conditional probabilities in a Bayesian network has received considerable attention by several investigators (e.g., Jordan (1998), Geiger and Heckerman (1997), Heckerman et al (1995)), but, this issue has not been studied in a causal Bayesian network.

In this thesis, we define the multicausal essential graph on the equivalence class of Bayesian networks in which each member of this class manifests a sort of strong type of invariance under (causal) manipulation called hypercausality. We then characterise the families of prior distributions on the parameters of the Bayesian networks which are consistent with hypercausality and show that their unmanipulated uncertain Bayesian networks must demonstrate the independence assumptions. As a result, such prior distributions satisfy a generalisation of the Geiger and Heckerman condition. In particular, when the corresponding essential graph is undirected, the mentioned class of prior distributions will reduce to the Hyper-Dirichlet family (see Chapter 6).

In the second part of this thesis, we will calculate certain local sensitivity measures and through them we are able to provide the solutions for the following questions: Is the network structure that is learned from data robust with respect to changes of the directionality of some specific arrows? Is the local conditional distributions associated with the specified node robust with respect to the changes to its prior distribution or with respect to the changes to the local conditional distribution of another node? Most importantly, is the posterior distribution associated with the parameters of any node robust with respect to the changes to the prior distribution associated with the parameters of one specific node? Finally, are the quantities mentioned above robust with respect to

the changes in the independence assumptions described in Chapter 3?

Most of the local sensitivity measures (particularly, local measures of the overall posteriors sensitivity), developed in the last decade, tend to diverge to infinity as the sample size becomes very large (Gustafson (1994) and Gustafson et al (1996)). This is in contrast to our knowledge that, starting from different priors, posteriors tend to agree as the data accumulate. Here we define a new class of metrics with more satisfactory asymptotic behaviour. The advantage of the corresponding local sensitivity measures is boundedness for large sample size.

Acknowledgements

I am extremely grateful to my supervisor Prof. Jim. Q. Smith for all his support, help, guidance, and friendship through the years of my research. His invaluable discussions and comments have deepened my insights into many problems.

I would like to express my appreciation to my family, especially my wife, my parents, and my parents-in-law for their encouragement and support.

I would like to thank my many friends and colleagues at the University of Warwick with whom I have had the pleasure of working over the years. These include Prof. S. Jacka, Dr R. Puch, Dr E. Riccomagno, and Dr E. Thonnes.

I would also like to thank all the staff in the Statistics Department at the University of Warwick, especially to Prof. J. Copas, Dr J. Hutton, Mrs P. J. Matthews, and Dr E. Shaw.

I am very grateful to the Ministry of Research and Technology of Iran for their financial support.

Declaration

I hereby declare that this thesis is based on my own work, except when otherwise stated.

Chapter 1

Introduction

Pearl's definition of causal Bayesian network (Definition 4.1) tells us how to read, from a single Bayesian network, a whole collection of new probability distributions which assert not only what will happen if we do not manipulate in the system but also what will happen if we manipulate each node in that system (Section 4.2). In other words, a causal Bayesian network does not simply explains how things are, but it can assert what will happen if the system is controlled or manipulated.

Pearl (2000) does not concentrate on the issue that, in practice, the conditional probabilities (in the factorisation of joint probability distribution associated with a Bayesian network, Equation (3.3)) in a Bayesian network usually need to be estimated. Therefore, in this thesis, we introduce a joint prior distribution on the conditional probabilities of a causal Bayesian network. More precisely, it is rational, from a Bayesian viewpoint, to ask what constraints we need to introduce on the prior distributions associated with these conditional probabilities of the Bayesian network (before imposing any manipulation) to make sure that its marginal mass function is consistent with the causal Bayesian network conditions (or consistent with the conditions introduced in Definition 4.1). It

might be expected that whether we were to learn the value of a conditional probability from some extraneous source or set this probability to some fixed value (e.g., by randomisation), it should be reasonable to substitute this value, that is, to manipulate the value of that probability to this known value and save the principle of the Bayesian network.

In this thesis, we show that the constraints required to introduce prior distributions on the conditional probabilities are independence assumptions (introduced by Geiger and Heckerman (1997), or see (Theorem 5.1)).

To link these independence assumptions and principles of causal models, we need to strengthen slightly the assumptions of factorisation invariance manipulation (see Daneshkhah and Smith (2003a)).

In Chapter 5, we introduce the hypercausal Bayesian network that asserts a set of factorisations of densities which are invariant to a class of "do" operations larger than those considered by Pearl. This can be considered as a developed version of randomised manipulation introduced by Koster (2000) and Lauritzen (2001) (see Section 5.2).

The multicausal essential graph maintains hypercausality for each member of the equivalence class of Bayesian networks represented by an essential graph (see Chapter 6). The prior density on the parameters of each Bayesian network in this class must display a generalisation of the Geiger and Heckerman condition (Theorem 6.3). The interpretation and implication of using prior distributions of these forms will be discussed in Chapters 5 and 6.

Chapter 2 consists of basic definitions, concepts and notation in graph theory which will be needed throughout this thesis.

In Chapter 3, we briefly examine learning in Bayesian networks with discrete variables (Section 3.2) and learning equivalence class of Bayesian networks (Section 3.3). However, a characterisation of these classes by the essential graphs will be presented in Section 6.2. The parameter independence (local and global) assumptions which are very useful and crucial in characterisation of a prior distribution associated with each network structure in the equivalence class of Bayesian networks are introduced in this chapter (Section 3.3) along with some examples.

We then review the Geiger and Heckerman (1997) results related to characterisation of prior distributions corresponding to parameters of equivalent Bayesian networks in Subsection 3.3.1. We emphasis that the independence assumptions play a vital role in these characterisations.

We introduce the causal models in Chapter 4. These models reveal the causal relationships between the involved variables. Pearl (1995, 2000) show that these models can be presented in terms of directed acyclic graphs. We review these causal mechanisms in Section 4.2. A causal model can be converted into a set of mathematical equations to yield an observational model. These models which are called *functional causal* will be studied in Section 4.3. As we said above, to make a connection between local and global independence assumptions with causality, we must justify the factorisation of causal Bayesian network in an appropriate way which would create randomised manipulation or more precisely contingent randomised manipulation. We introduce the randomised intervention which are originally introduced by Lauritzen (2001) and Koster (2000) (particularly, in terms of the functional causal models) in Section 4.4. Learning of causal Bayesian networks will be presented in Section 4.5. In this section, we also include some results when there are some hidden variables in the causal Bayesian models. We identify

some new unanswered questions.

In Chapter 7 local sensitivity analysis of the posterior quantities associated with a Bayesian network with respect to different sources of uncertainties will be studied. Local sensitivity analysis generally is reviewed in Section (7.2). Gustafson (1996b) calculated the local sensitivity measures of posterior quantities with respect to changes of the specific stage in hierarchical models (Subsection 7.2.2).

In Section 7.3, we study local sensitivity analysis in Bayesian networks with respect to misidentification of distributional assumptions of likelihood and prior distributions, and misidentification of independence assumptions.

A sensitivity analysis of Bayesian networks with dependent parameters is studied in Section 7.4. In this situation, we introduce the hierarchical models as prior distributions on the parameters of these networks, and then we use Gustafson's ideas to calculate the local sensitivity measures. But, in some levels of these priors, we would encounter unidentifiability of parameters. We make some suggestions to address this issue.

We examine the relationship between the local cause and the local sensitivity in Section 7.5. In Section 7.6, we study asymptotic behavior of the specific form of the local sensitivity measure, and for the wide class of priors which are compatible with some reasonable mild conditions.

In Chapter 8, we study the asymptotic behaviour of the local sensitivity measures introduced in Chapter 7. We also introduce new local sensitivity measures in terms of *credible metrics*. We will show that these local sensitivity measures display good asymptotic behaviour.

In the last chapter, we produce a summary of our results and some further work is presented.

Chapter 2

Introduction to Graphical Models

In this chapter, we briefly introduce some basic and important concepts associated with the graphical models that throughout this thesis will be frequently used. These concepts can be studied in more detail in Cowell et al (1999) and references therein.

In the next section, we briefly study conditional independence. In Section 2.2, the directed acyclic graph and the undirected graphs are introduced.

2.1 Conditional Independence

A graphical model (in a particular form, Bayesian network) is a probabilistic model based on the notion of *conditional independence* and *dependence*. Conditional independence is a fundamental notion in the analysis of interactions among multiple factors. The intuition behind the use of conditional independence is that a dependence relationship between two variables may disappear when a third variable is considered in relation with those two variables.

Dawid (1979, 1980) introduced and comprehensively studied conditional indepen-

dence. His definition of conditional independence is considered as an underlying concept in the probabilistic graphical models and clearly in this thesis.

Let $\underline{X} = (X_v, v \in V)$ be a finite set of variables. Let Y_1, Y_2, Y_3 denote any three disjoint subsets of variables in $\underline{X}$.

Definition 2.1 (Conditional Independence) Y_1 is said to be *conditionally independent* of Y_2 given Y_3, written $(Y_1 \perp\!\!\!\perp Y_2 \mid Y_3)$, if for all configurations y_1, y_2, y_3 of the variables in Y_1, Y_2, Y_3 satisfying $p(Y_3 = y_3)$, it holds that

$$p(Y_1 = y_1 \mid Y_2 = y_2, Y_3 = y_3) = p(Y_1 = y_1 \mid Y_3 = y_3).$$

The definition above can be introduced over measurable sets A and B as follows,

$$P(Y_1 \in A, Y_2 \in B \mid Y_3 = y_3) = P(Y_1 \in A \mid Y_3 = y_3) P(Y_2 \in B \mid Y_3 = y_3), \qquad (2.1)$$

An equivalent relationship can be given by

$$P(Y_1 \in A \mid Y_2 = y_2, Y_3 = y_3) = P(Y_1 \in A \mid Y_3 = y_3), \qquad (2.2)$$

Note that, the conditional statement mentioned above between Y_1, Y_2 and Y_3 leads us to this point that learning the value of Y_2 does not give any additional information about Y_1, when we know $Y_3 = y_3$. Furthermore, since the conditional independence assumptions and the individual factors often have relatively clear substantive or causal interpretations, this manner of construction facilitates explanation, whether by a statistician to her client or by an expert system to its user.

The previous definition may be seen as a factorisation criterion that tells, that the conditional probability of Y_1 given Y_2 and Y_3, is in fact a function of Y_3 alone. Note that, when Y_3 is trivial, Y_1 and Y_2 are *marginally independent*. In fact, we say that two

set of random variables Y_1 and Y_2 are *marginally independent* if their joint probability $p(y_1, y_2)$ factors like

$$p(y_1, y_2) = p(y_1)p(y_2). \tag{2.3}$$

But, if the conditioning set Y_3 is not trivial, the joint probability factorizes as follows,

$$p(y_1, y_2 \mid Y_3 = y_3) = p(y_1 \mid Y_3 = y_3)p(y_2 \mid Y_3 = y_3), \tag{2.4}$$

if and only if $Y_1 \perp\!\!\!\perp Y_2 \mid Y_3$.

For more details about the properties and axioms of the conditional independence, see Dawid (1979, 1980) and Cowell et al (1999).

2.2 Directed and Undirected Graphs

The directed acyclic graph plays a key role in the causal Bayesian networks that will be discussed in Chapter 4. Spirtes et al (1993) and Pearl (2002) represent causal Bayesian networks in terms of these graphs. Furthermore, it should be noticed that the conditional independences that are underlying a multivariate probability distribution for the variables in the domain of the problem in hand are reflected by the graphical structure of a Bayesian network, the so-called *directed acyclic graph*. In the next chapter, we will describe this sort of graph. The other component of a Bayesian network is a set of parameters, rendering a quantitative description of possible probability distributions (Section 3.2).

But to introduce Bayesian networks and causal Bayesian networks in this thesis, we need to get familiar with the basic concepts of the acyclic graph and its properties that are defined and studied in this section.

First, let us introduce a graphical model in the general form. A *graph* is a pair $G = (V, E)$ where V is the set of *vertices* (or nodes) and E is the set of *edges.* The set of edges E is a subset of the set $V \times V$ of ordered pairs of nodes. It is assumed that E contains only distinct pairs of nodes so that there exist no loops, that is, $(x, y) \in E \Rightarrow x \neq y$.

Example 2.1 The following figure illustrates a graph that supports the definition above,

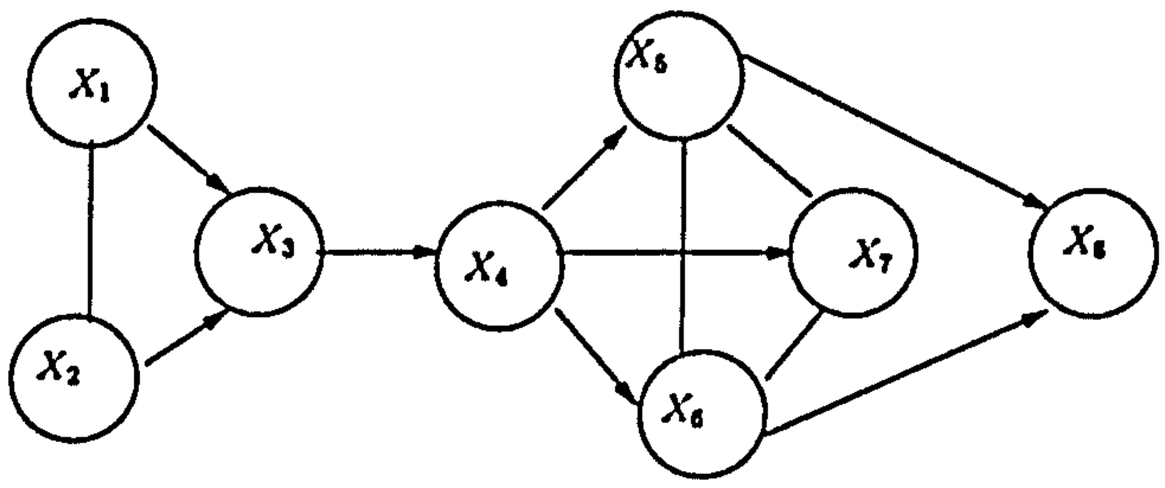


Figure 2.1: A graph with 8 vertices and 12 edges.

In Figure 2.1 the set of vertices is $V = \{X_1, \ldots X_8\}$, and the set of edges is given by $E = \{(X_1, X_2), (X_1, X_3), (X_2, X_3), (X_3, X_4), \ldots, (X_5, X_8), (X_6, X_8)\}$.

The edges in each graph can be all directed, all undirected or any combination thereof. Thus, each graph with respect to the kind of the edges between the nodes of the corresponding graph can be classified into directed graph, undirected graph or chain graph that are introduced and studied in this chapter. However, the focus of this thesis is on the directed graph or more precisely directed acyclic graph.

Given two nodes x and y, the edge between them is said to be *undirected* if and only if $(x, y) \in E$ and $(y, x) \in E$, and written $x \sim y$. If $(x, y) \in E$ but $(y, x) \notin E$, the edge is called directed, and represented by $x \rightarrow y$. Thus, x is a *parent* of y, and y is a *child* of x. The set of parents of a vertex y is denoted by $pa(y)$, and the set of children[1] of a vertex x represented by $ch(x)$.

Example 2.2 In Figure 2.1, there are 9 directed edges and 3 undirected edges. For example, the edge between (X_1, X_3) or (X_2, X_3) is directed, but the edge between X_1 and X_2 or between X_5 and X_6 are undirected. The set of parents of X_3 is $pa(X_3) = \{X_1, X_2\}$, and the set of children of X_3 is given by $ch(X_3) = \{X_4\}$.

A *cycle* of length n is a path[2] with the modification that the first and last vertex are identical $x_0 = x_n$.

Definition 2.2 (Directed acyclic graph) A directed graph $G = (V, E)$ is *acyclic* if it contains no directed cycle.
Hereafter *directed acyclic graph* is abbreviated to the term DAG (we use DAGs for the directed acyclic graphs).

Example 2.3 The following graph shown in Figure 2.2 is a DAG.

Definition 2.3 (Undirected graph) An *undirected graph* is a pair (V, E), where

[1]We need to define these sets to introduce the Markov blanket used in Chapter 7.

[2]A *path* of length n from x to y is a sequence $x = x_0, \ldots, x_n = y$ of distinct vertices such that $(x_{i-1}, x_i) \in E$ for all $i = 1, \ldots, n$. Thus a path can never cross itself and moving along a path never goes against the directions of arrows. A path from x to y is shown by $x \mapsto y$.

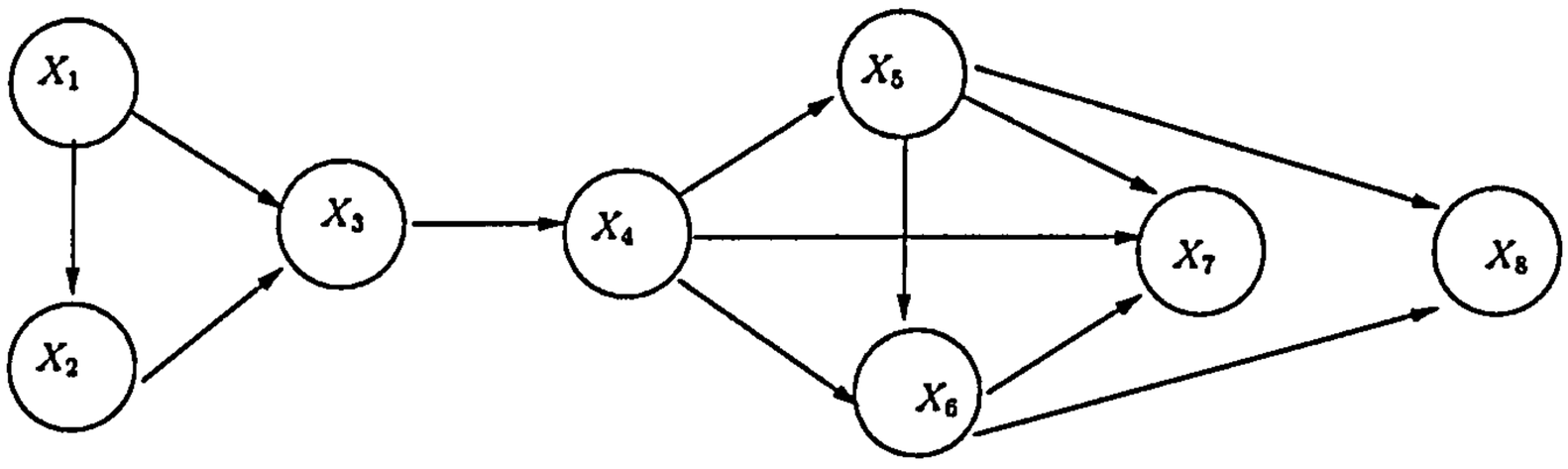


Figure 2.2: A Directed version (DAG) of the graph shown in Figure 2.1.

V is a set of vertices, and E is a set of unordered pairs $\{x, y\}$ of distinct components of V. The endpoints of the edge $e = \{x, y\}$ are the vertices x and y.

In other words, if all the edges of a graph are undirected, that graph is called *undirected.* The graphs shown in Figures 2.6 and 2.7 are the examples of undirected graphs.

To study sensitivity analysis of the Bayesian networks with the dependent parameters introduced in Section 7.4, we need to introduce neighbour(s). We introduce them in the following definition.

Definition 2.4 (Neighbours) Vertex x is a *neighbour* of vertex y if $\{x, y\}$ is in E. The set of neighbours[3] of x is introduced by $ne(x) = \{y \in V : \{x, y\} \in E\}$.

To introduce some important concepts such as clique, we must define adjacent vertices. Two vertices will be called *adjacent* if they are linked together by an undirected

[3]If $x \sim y$, then we can also say that x and y are neighbours.

edge.

The following example helps us to understand the concepts introduced above.

Example 2.4 The neighbours of X_3 in Figure 2.1 is $ne(X_3) = \{X_1, X_2, X_4\}$. But, The adjacent set of X_3 in Figure 2.1 is empty set. However, the adjacent set of X_5 is $\{X_6, X_7\}$.

Definition 2.5 (Subgraph) $G_A = (A, E_A)$ is defined by a subset $A \subseteq V$ and the induced edge set $E_A = E \cap (A \times A)$. It will be said that G_A is an *induced subgraph* of G.

An undirected graph $G = (V, E)$ is said to be *complete* if and only if every pair of vertices is adjacent. A subset of vertices is complete if it induces a complete subgraph.

Definition 2.6 (Clique) A *clique* is a maximal (with respect to $\subseteq$) complete subgraph[4].

The decomposable graphs, used in Sections 6.2 and 7.3, are factorised in terms of cliques. For further details of clique and its properties, see Cowell et al (1999).

Example 2.5 The following graph shown in Figure 2.3 is the induced subgraph of that is displayed in Figure 2.1 with the following set of vertices and edges:

$$A = \{X_4, X_5, X_6, X_7\}$$

[4]The clique in Definition 2.4 should be called *maximal clique*.

and

$$E_A = \{(X_4, X_5), (X_4, X_6), (X_4, X_7), (X_5, X_6), (X_5, X_7), (X_6, X_7)\}.$$

This subgraph is complete, because any two pairs of vertices in this subgraph are adjacent. There are 4 cliques with 3 vertices in this subgraph, for example, $\{X_4, X_5, X_7\}$ or $\{X_4, X_6, X_7\}$ creates a clique.

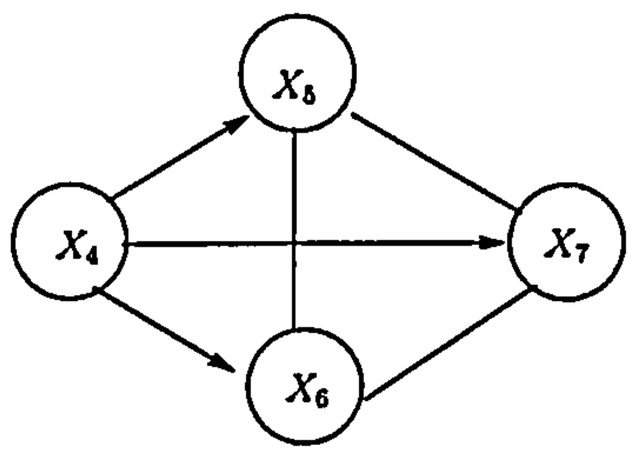


Figure 2.3: A Subgraph of the graph shown in Figure 2.1.

Definition 2.7 (Chain Graph) A graph that has no directed cycles is called a *chain graph*[5].

Example 2.6 The graphs are shown in Figures 2.1, 2.2, and 2.3 are the examples of chain graphs.

A chain graph[6] is formed by a non-empty set of chain components $\tau(G)$.

[5]Thus, undirected graphs and directed acyclic graphs are both special cases of chain graphs

[6]The essential graph introduced in Chapter 6 is a chain graph that is representing an equivalence class of Bayesian networks.

Definition 2.8 (Chain Component) The set of *chain components* of a chain graph corresponds to the set of connected components left after the removal of all directed edges in the chain graph.

Example 2.7 In Figures 2.4 and 2.5, a chain graph and its set of chain components, $\tau(G) = \{\{1,2,3,4\},\{5,6\},\{7,8\},\{9,10\},\{11\}\}$, are shown, respectively.

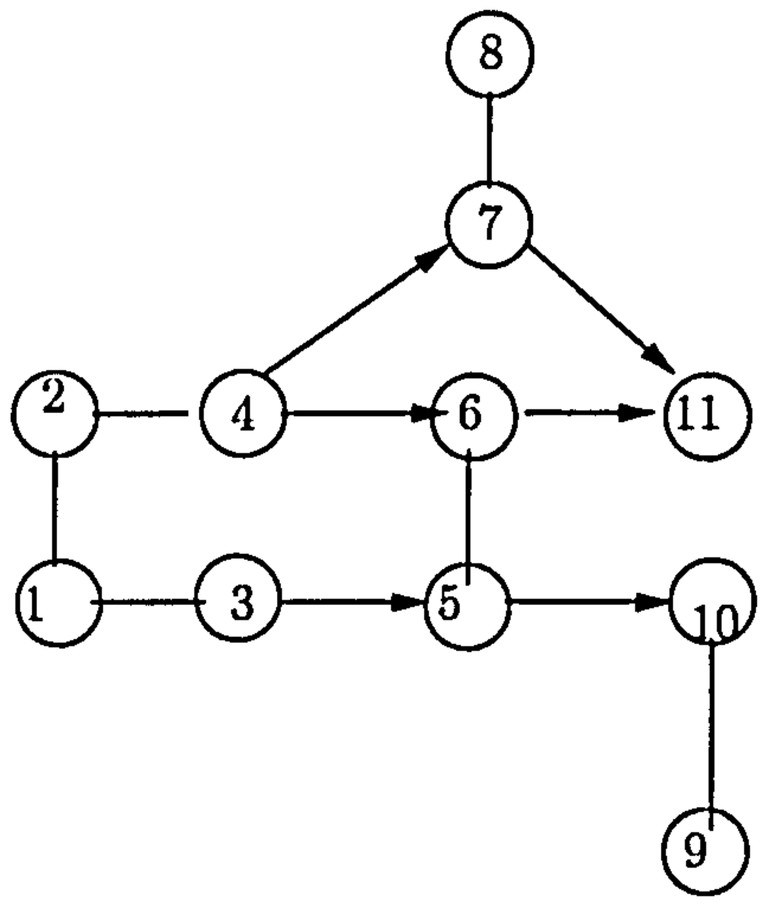


Figure 2.4: An example of a chain graph.

It should be noticed that each node of a DAG G forms a chain component of G.

Definition 2.9 (Ancestors and descendants) Given a DAG D, the set of its vertices x such that $x \mapsto y$ but not $y \mapsto x$ are the *ancestors* $an(y)$ and the descendants

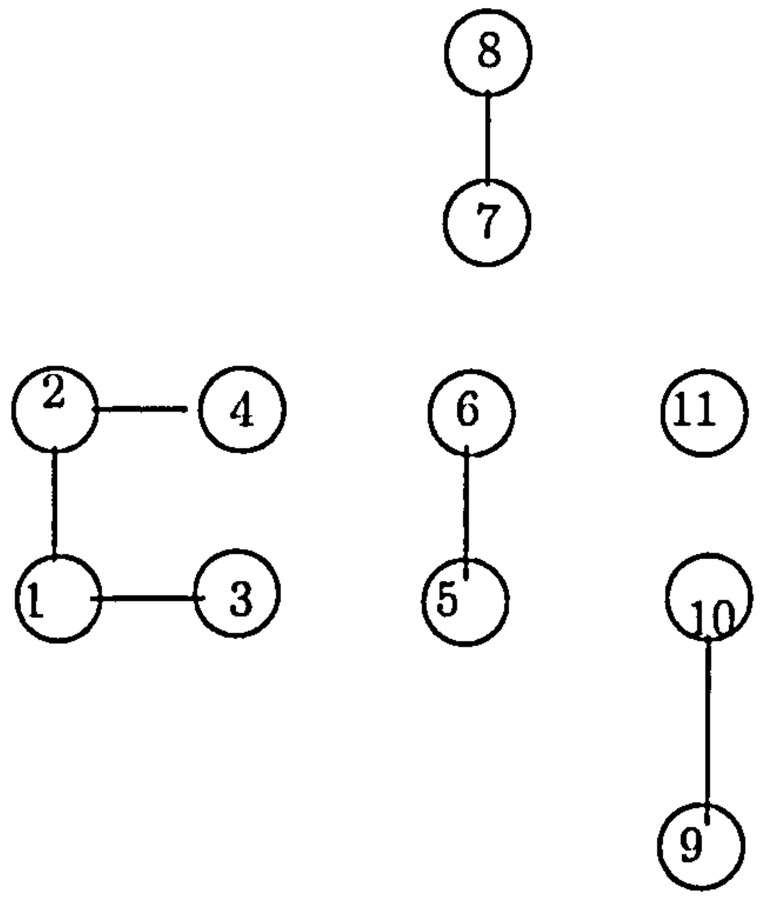


Figure 2.5: The chain components associated with the chain graph shown in Figure 2.4.

$de(x)$ of x are the vertices y such that $x \mapsto y$ but not $y \mapsto x$. The *descendants* of x are the nodes y such that there is a path from x to y but not from y to x.

It should be noticed that the definition above can be expressed in terms of the neighbours of the vertices between x and y if we define the path between x and y as follows: A path of length n is an ordered set $P = (x_0, \ldots, x_n)$ of vertices of D such that x_{i-1} is a neighbour of x_i, $i = 1, \ldots, n$, and that P does not encounter the same vertex twice.

A convenient way of characterising the set of distributions compatible with a given DAG is to list the set of conditional independencies that each such distribution must satisfy. These independencies can be read off the DAG by using a graphical (separation) criterion called *d-separation* which plays a major role to identify causal relation-

ships between variables in the causal Bayesian network (this criterion is implicitly used throughout this thesis) .

Consider three disjoint set of variables, X, Y, and Z, which are represented as nodes in a DAG G. To test whether X is independent of Y given Z in any distribution consistent with G, it is required to check whether the nodes corresponding to variables Z *block* all paths from nodes in X to nodes in Y. Blocking here can be interpreted as stopping the flow of dependency between the mentioned variables that are connected by the paths that will be introduced by the following definition presented by Pearl (1988).

Definition 2.10 (d-separation) A path P is said to be *d-separated* (or *blocked*) by a set of nodes Z if and only if

1. P contains a chain $i \rightarrow m \rightarrow j$ or a *fork* $i \leftarrow m \rightarrow j$ such that the middle node m is in Z, or

2. P contains an *inverted fork* (or *collider*) $i \rightarrow m \leftarrow j$ such that the middle node m[7] is not in Z and such that no descendant of m is in Z.

A set Z is said to be d-separate X from Y if and only if Z blocks every path from a node in X to a node in Y.

Note that, in (causal) chains $i \rightarrow m \rightarrow j$ and (causal) forks $i \leftarrow m \rightarrow j$, the two endpoint variables (i.e., i and j) are marginally dependent but will be independent of each other (i.e., blocked) once they are conditioned on the middle variable (i.e., m). But, the inverted forks act the opposite way, that means if the two extreme variables

[7]A vertex that has more than one parent is called *head-to-head* vertex.

are marginally independent, they will be dependent (i.e., connected through unblocked path) once they are conditioned on the middle node or any of its descendants.

Note that the inverted fork discussed above in the graph theory literature is usually called *immorality* (or *v-structure*) that is formed by two non-adjacent vertices with a common child.

A DAG that has no immoralities is said to be moral. A moral graph is more formally defined by Cowell et al (1999) as follows,

Definition 2.11 (Moral Graph) For a DAG D, we define the *moral graph* of D to be the undirected graph D^m obtained from D by first adding undirected edges between all pairs of vertices which have common children and are not already joined[8], and then forming the undirected version[9] of the resulting graph.

A DAG that is not moral can be moralised by *marrying* those non-adjacent parents that induce an immorality, that is, joining them with an undirected edge, and dropping directions on the rest of edges in the given DAG D. The moralised version of a DAG D is denoted by D^m.

Example 2.8 The moral graph associated with the graph shown in Figure 2.4 is presented below.

The importance of moralisation is in building the inference toolbox for a probabilistic network determined by a chain graph.

---

[8]Two vertices x and y are said to be *joined* if $(x, y) \in E$ or $(y, x) \in E$ (See Cowell et al (1999).).

[9]The *undirected version* $G^{\sim}$ of a graph G is the undirected graph obtained by replacing the directed edges of G by undirected edges.

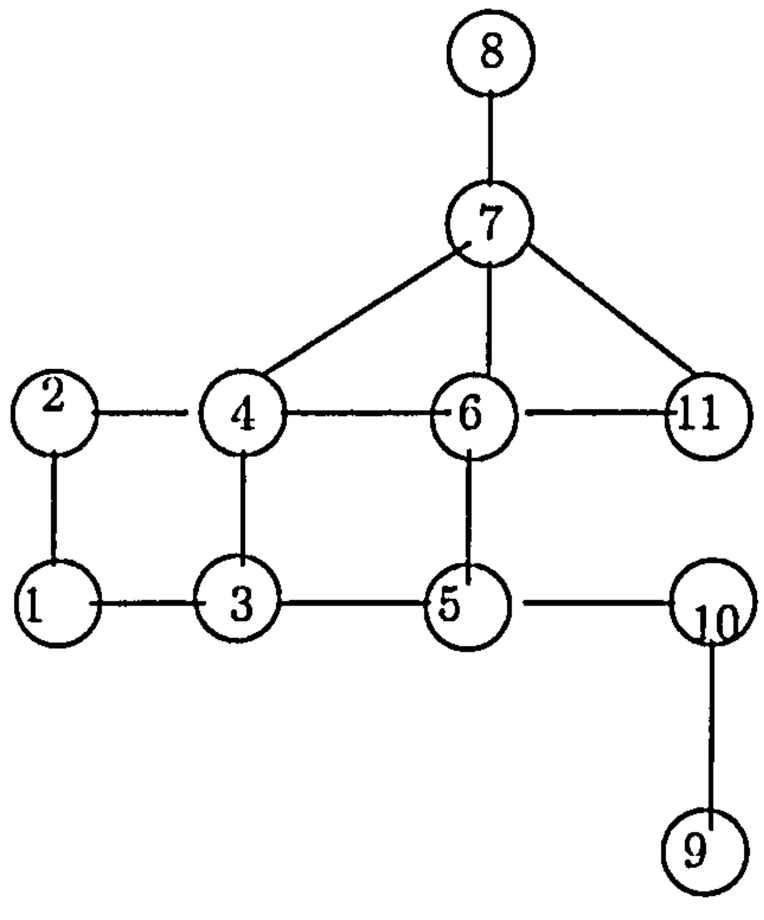


Figure 2.6: The moral graph associated with the chain graph shown in Figure 2.4.

One of the most important graphs is the decomposable graph which we briefly discuss below. We introduce the characterisation of the prior distribution associated with the probabilities for the discrete decomposable graph in Chapter 3.

Definition 2.12 (Decomposition of graph) Two induced subgraphs G_{V_1} and G_{V_2} are a *decomposition* of graph G if $V_1 \cup V_2 = V$ and $V_1 \setminus V_2 \neq \emptyset$ and $V_2 \setminus V_1 \neq \emptyset$ and $G_{V_1 \cap V_2}$ is complete and $E_1 \cup E_2 = E$.

The graph G is said to be *decomposable*, if it is complete or if there exists a decomposition of this graph into two decomposable subgraphs.
This graph will be used in Chapter 6 associated with the undirected essential graph.

Definition 2.13 (Chordal graph) A *chordal*[10] or *triangulated graph* is an undirected graph with no chordless[11] undirected cycles on more than three vertices.

The following theorem presented by Cowell et al (1999) shows the relationship between decomposability and chordality.

Theorem 2.1 The following conditions are equivalent for an undirected graph G:

1. G is decomposable;

2. G is chordal;

3. Every minimal (x, y)-separator is complete.

Example 2.9 The figure below shows examples of chordal and non-chordal undirected graphs. Note that the graph on the left hand side is also decomposable.

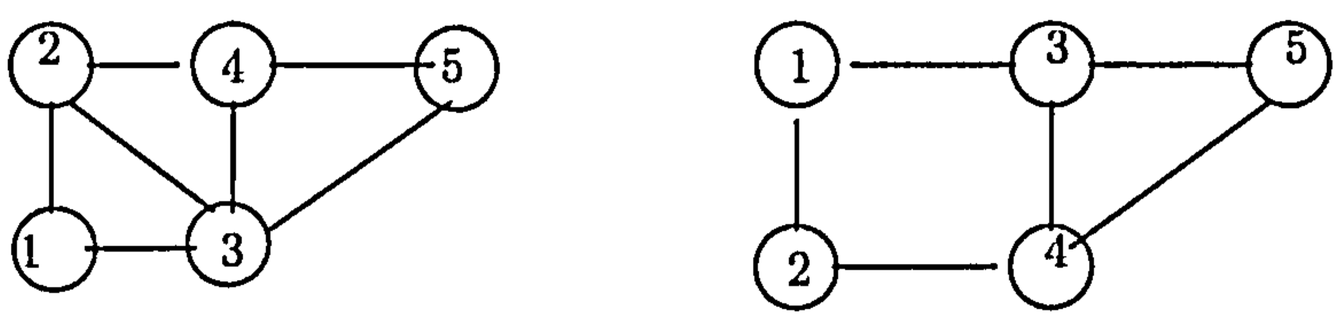


Figure 2.7: Examples of undirected chordal (LHS) and non-chordal (RHS) graphs.

[10]We use this graph to characterise the essential graph introduced in Section 6.2.

[11]A chord in a graph is an edge joining two nodes already connected by a path. A *chordless cycle* of a graph G is a graph cycle of length at least four in G that has no cycle chord (See Cowell et al (1999)).

Chapter 3

Learning Bayesian Networks

3.1 Introduction

This chapter gives a brief introduction to Bayesian networks. After defining a Bayesian network model, we focus on learning with Bayesian networks with discrete variables. We refer the interested reader to the following books and articles: Cowell et al (1999); Lauritzen (1996); Pearl (1988); Heckerman (1995) to get more comprehensive details about Bayesian networks and graphical models.

A statistical model with a large collection of variables is typically computationally complex. To reduce the dimensionality of the statistical model some notion of independence is required. Graphical models can deal with this issue. A graphical model represents a collection of random variables by a graph; each node in the graph represents a random variable and the lack of an edge between two nodes represents a conditional independence assertion. These models have been studied in details in the references above and references cited therein.

One class of graphical models that is constructed in terms of directed acyclic graphs is called *Bayesian networks*. This class will be extensively examined in this chapter and

throughout this thesis.

We study the equivalence classes of Bayesian networks in Section 3.3, and a characterisation of these classes by the essential graphs will be presented in Chapter 6. The notion of equivalence plays an important role for the learning of Bayesian networks. Two network structures are equivalent if the set of distributions that can be represented using one of the structures is identical to the set of distributions that can be represented using the other. Because equivalence is reflexive, symmetric, and transitive, the relation defines a set of equivalence classes over network structures. The main reason we consider equivalence classes of Bayesian networks is because, given the prior for any Bayesian network in a given equivalence class, we can derive the prior distribution associated with any other element in the equivalence class.

Useful assumptions that help us to define and characterise a prior distribution associated with each network structure in the equivalence class of Bayesian networks are *local and global* parameters independence. We introduce these assumptions, introduced by Spiegelhalter and Lauritzen (1990) and used by Geiger and Heckerman (1997) to characterise Dirichlet distributions associated with the parameters of Bayesian networks with two multinomial variables in Sections 3.2 and 3.3 respectively. In Chapters 5 and 6, we use these independence assumptions to elicit and characterise prior distributions associated with the parameters of the given single causal Bayesian network.

In fact, throughout this thesis, we want to establish that if a Bayesian is prepared to make bold enough assertions within a single uncertain Bayesian network then this not only introduces independence relationships between parameters, but can also characterise prior families of distributions on these parameters. We believe this is a very useful way of thinking about this class of models. We will comprehensively study these

issues in Chapters 5 and 6. In order to do this, we need to review the characterisation of prior distributions associated with parameters of equivalent Bayesian networks when independence relationships between parameters are assumed to be valid. The assumptions of local and global independence are used in this thesis in order to estimate conditional probabilities of a causal Bayesian networks and not for the purpose of selecting a Bayesian network. More precisely, it is not an assertion about a common prior to be used for the causal Bayesian network for the model selection as is more typical in, for example, Heckerman et al (1995), Cowell et al (1999) and Cooper and Yoo (1999). However, as I said earlier, the parts of their works that are technically relevant to this thesis will be presented.

3.2 Introduction to Bayesian networks

The graphical models contain a very wide class of the statistical models based on either directed acyclic graphs, undirected graphs, or a combination thereof. Among the current graphical models, Bayesian networks are certainly the most common and perhaps the most applicable.

A Bayesian network B for a set of variables $\underline{X} = \{X_1, \ldots, X_n\}$ is a pair $(G, \underline{\theta})$, where $G = (\underline{X}, E)$ is a DAG[1] (or network structure), and $\underline{\theta}$ is a set of conditional probability distributions such that $\theta_i \in \underline{\theta}$ defines the conditional probability of X_i given its parents $(pa(x_i))$ in G, that is, $\theta_i = p(x_i \mid pa(x_i), \underline{\theta})$.

The basic decomposition scheme offered by DAGs can be explained as follows. Consider a probability function p defined over n (discrete) variables with arbitrary ordering $X_1, \ldots, X_n$. The probability function p can be decomposed as a product of n conditional

[1]DAGs is an abbreviated form of directed acyclic graphs.

probability functions as follows

$$p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i \mid x_1, \ldots, x_{i-1}). \tag{3.1}$$

If x_i is independent of all other predecessors given its parents, then, we can write

$$p(x_i \mid x_1, \ldots, x_{i-1}) = p(x_i \mid pa(x_i)) \tag{3.2}$$

The Equation (3.2) assigns to each variable X_i a select set $pa(x_i)$ of preceding variables that are sufficient for determining the probability distribution of X_i. In other words, knowing the values of other preceding variables is redundant once we know the values of $pa(x_i)$.

It is therefore obvious that, every probability function satisfying Equation (3.2) must decompose into the product

$$p(x_1, \ldots, x_n) = \prod_{i} p(x_i \mid pa(x_i)) \tag{3.3}$$

where $pa(x_i) \subseteq \{x_1, \ldots, x_{i-1}\}$.

Therefore, we can say that a DAG G is a Bayesian network if and only if the probability distribution associated with the Bayesian network admits the product decomposition given in Equation (3.3) imposed by G.

We now briefly introduce learning of a Bayesian network from the data. Bayesian network learning will be done in two stages. In the first stage we learn about the network structure including introducing the nodes and the arrows between the nodes. The parameters associated with each node that can be represented as the conditional probability distributions of the nodes given their parents are learned in the second stage. Throughout this thesis, the network structure is given and we focus on learning the conditional probabilities. Consider a Bayesian network of random variables all of which

are discrete and a set P of local probability distributions associated with each variable, i.e., $p(x_i \mid pa(x_i))$. The set of parameters $\underline{\theta}$ associated with the probabilities in P can be defined as

$$\theta_{ijk} = p(X_i = x_i^k \mid pa(X_i) = pa(x_i)^j, \underline{\theta}_i), \quad i = 1, \ldots, n, \; k = 1, \ldots, l_i, j = 1, \ldots, m_i$$

where θ_{ijk} stands for the parameter associated with the level k of i^{th} variable and the level j of its parents.

Thus, $\underline{\theta}_{ij} = (\theta_{ij1}, \ldots, \theta_{ijl_i})$, $\underline{\theta}_i = \{\underline{\theta}_{ijk} \;, 1 \leq j \leq m_i, \;\; 1 \leq k \leq l_i\}$, where each component of $\underline{\theta}_i$ is positive and for fixed j, $\Sigma_{k=1}^{l_i} \theta_{ijk} = 1$, and therefore $\underline{\theta} = \{\underline{\theta}_i, 1 \leq i \leq n\}$. So Equation (3.3) becomes

$$p(\underline{x} \mid \underline{\theta}) = \prod_{i=1}^{n} p(x_i \mid pa(x_i), \underline{\theta}_i) \tag{3.4}$$

where $pa(x_i)$ denotes to the values of the parent's set of x_i.

Spiegelhalter and Lauritzen (1990) make two key assumptions which greatly simplify the related computation and subsequent analysis. The first assumption is that of *global independence* whereby the parameter vectors $\underline{\theta}_i$ are assumed mutually independent a priori. In fact, we can say that

$$p(\underline{\theta}) = \prod_{i=1}^{n} p(\underline{\theta}_i). \tag{3.5}$$

This assumption alone allows us to express the joint distribution of $\underline{x}$ and $\underline{\theta}$ as

$$p(\underline{x}, \underline{\theta}) = \prod_{i=1}^{n} p(x_i \mid pa(x_i), \underline{\theta}_i) p(\underline{\theta}_i) \tag{3.6}$$

Note that from the equation above it can be concluded that $\underline{\theta}_i$ may be considered as another parent of x_i in a general Bayesian network.

Therefore, the marginal likelihood density can be calculated as

$$p(\underline{x}) = \int p(\underline{x}, \underline{\theta}) d\underline{\theta} = \int \prod_{i=1}^{n} p(x_i \mid pa(x_i), \underline{\theta}_i) p(\underline{\theta}_i) d\underline{\theta}_i = \prod_{i=1}^{n} p(x_i \mid pa(x_i)) \tag{3.7}$$

where

$$p(x_i \mid pa(x_i)) = \int p(x_i \mid pa(x_i), \underline{\theta}_i) p(\underline{\theta}_i) d_{\underline{\theta}_i}.$$

is the expectation of the conditional probability table for x_i.

The second assumption is that of local independence whereby the parameter $\underline{\theta}_i$ decomposes into components corresponding to the levels of the factors in $pa(x_i)^j$. These components are assumed to be mutually independent a priori. Thus for the fixed $k = k^*$

$$p(\underline{\theta}_i) = \prod_{j=1}^{m_i} p(\theta_{ijk^*}) \tag{3.8}$$

Now consider a conditional probability distribution $p(x_i^k \mid pa(x_i)^{j^*}, \theta_i) = \theta_{ij^*k}$ of the particular configuration of parent nodes. The local independence assumption expressed by the equation above can be considered as follows: for the specific set of levels $pa(x_i)^{j^*}$, of $pa(x_i)$, θ_{ij^*k} parameterises the conditional probability $p(x_i^k \mid pa(x_i)^{j^*}, \underline{\theta}_i)$, and conditional on $x_i \cup pa(x_i)^{j^*}$, θ_{ij^*k} is independent of the remaining parameters $\underline{\theta}_i \setminus \theta_{ij^*k}$. The following example shows a simple Bayesian network that helps to understand better local and global independence assumptions.

Example 3.1 Consider the following Bayesian network with two binary random variables X and Y. The parameters associated with the probabilities of X and Y are: $\underline{\theta} = \{\theta_x, \theta_{y|X=1}, \theta_{y|X=0}\}$. Here, $X = 1$ stands for success, and $X = 0$ denote the failure of the corresponding event, likewise for Y. According to the Bayesian network in Figure 3.1, it is obvious that θ_x is independent of $\{\theta_{y|X=1}, \theta_{y|X=0}\}$ (since, there is no edge between θ_x and $\{\theta_{y|X=1}, \theta_{y|X=0}\}$), that is, $\theta_x \perp\!\!\!\perp \{\theta_{y|X=1}, \theta_{y|X=0}\}$. Therefore, the global independence assumption holds. Similarly, because no edge exists between

$\theta_{y|X=1}$ and $\theta_{y|X=0}$, we can say that the local independence assumption is also satisfied, i.e, $\theta_{y|X=1} \perp\!\!\!\perp \theta_{y|X=0}$. The prior distributions associated with these parameters, because

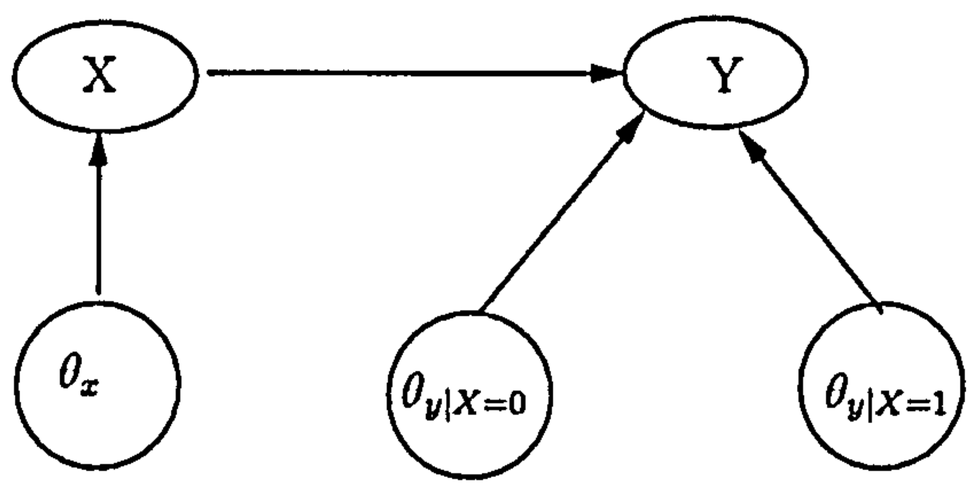


Figure 3.1: Representation of a Bayesian Network with two Binary variables.

of the existence of local and global parameter independence, factorise as

$$p(\underline{\theta}) = p(\theta_x)p(\theta_{y|X=1})p(\theta_{y|X=0})$$

Later, in this chapter, the prior distributions associated with the parameters of this Bayesian network (and its equivalent Bayesian network discussed in the next section) will be characterised. It will be shown that the Beta distributions, according to Geiger and Heckerman's results, are the most appropriate prior distributions for these parameters.

Spiegelhalter and Lauritzen (1990) indeed claimed that the independence assumptions will make the computations more feasible.

3.3 Learning Equivalence Classes of Bayesian Networks

In this section, we introduce notation and basic concepts of equivalence classes of Bayesian networks and their properties that are relevant to the topic of this thesis.

Two network structures are said to be equivalent if the set of distributions that can be represented with one of those structures is identical to the set of distributions that can be represented with the other. More formally, the definition of two equivalent Bayesian networks is given below.

Definition 3.1 (Equivalent Bayesian Networks) Two Bayesian network structures G and G' are called *equivalent* if for every Bayesian network $B = (G, \theta_G)$, there exists a Bayesian network $B' = (G', \theta_{G'})$ such that B and B' assert the same set of conditional independence statements among the variables in the domain.

Heckerman et al (1995) were interested in defining priors which were invariant to the specification of equivalence representative in its equivalence class. They wanted to choose a default prior for model selection, so that equivalent Bayesian networks (having the same likelihood) could be given the same prior. We use their results to make a connection between prior independence assumptions and causality, and to characterise prior distributions sympathetic to causal hypotheses on all Bayesian networks in the equivalence class defined by an *essential graph* (essential graph will be introduced in Chapter 6).

We use the symbol $G \approx G'$ to denote that G and G' are equivalent. Note that equivalence is reflexive, symmetric, and transitive, therefore the relation $\approx$ defines a set of equivalence class over network structures.

Definition 3.2 (Compelled Edge) A directed edge $x \rightarrow y \in E_G$ is called *compelled* in G if for any network structure (DAG) $G' \approx G$, $x \rightarrow y \in E_{G'}$.

Note that for any edge in E_G, if that edge is not compelled in G, then that edge is called *reversible* in G. In other words, there is some network structure $G' \approx G$ such that the mentioned edge has opposite direction.
The characterisation of the equivalent network structures are given by Verma and Pearl (1990) in the following theorem.

Theorem 3.1 (Verma and Pearl (1990)) Two network structures (DAGs) are equivalent if and only if they have the same skeletons and the same v-structures.

The *skeleton* of any network structure is the undirected version of the graph that is obtained by ignoring the directionality of every single edge. A v-structure in a DAG G is an ordered triple of nodes (x_1, x_2, x_3) such that G contains the arrow $x_1 \rightarrow x_2$ and $x_3 \rightarrow x_2$, and x_1 and x_3 are not adjacent in G.
A result that can be obtained from Theorem 3.1 is that for any edge participating in a v-structure in the given network structure G, if that edge is reversed in some other network structure G' then G is not equivalent with G'.
The equivalence class of Bayesian networks can be represented by *Acyclic partially directed graphs*, or *patterns*[2]. Patterns are graphs that contain both directed and undirected edges. Let $\mathcal{P}$ stands for the given pattern. Thus, the equivalence class of Bayesian networks associated with $\mathcal{P}$, denoted by $[\mathcal{P}]$, is defined as: $G \in [\mathcal{P}]$ if and only if G and $\mathcal{P}$ have the same skeleton and the same set of v-structures. Note that, from Theorem 3.1, it can be concluded that a pattern containing a directed edge for every edge par-

[2]Patterns are sometimes called PDAGs in the literature.

ticipating in v-structures, and an undirected edge for all other edges, uniquely identifies an equivalence class of Bayesian networks.

It should be noticed that although Theorem 3.1 provides a practical way to determine whether two given Bayesian networks are Markov equivalent, it does not directly give a characterisation of the entire equivalence class $[G]$ for the given Bayesian network (see Anderson et al (1997)). Furthermore, Anderson et al (1997) argued that since the number of possible orientations of all arrows that do not participate in any v-structure of a Bayesian network G grows exponentially with the number of such arrows, hence super-exponentially with the number of vertices, determination of the equivalence class $[G]$ by exhaustive enumeration of possibilities becomes computationally infeasible as the size of G increases.

Anderson et al (1997) therefore recommend the *essential* of a given DAG rather than Pearl's pattern to characterise the entire equivalence class [G]. This graph will be studied in Chapter 6.

3.3.1 Parameter Priors for Bayesian Networks

In this subsection, we present parameter priors for discrete Bayesian networks. Geiger and Heckerman (1997) showed that local and global independence assumptions for two Bayesian networks in their equivalence class were only possible if the prior of the joint probabilities associated with each of these Bayesian networks was Dirichlet. Therefore, in particular, all components including marginal and conditional probabilities of each Bayesian network were Dirichlet. Furthermore, Dawid and Lauritzen (1993) proved that the prior distribution associated with the discrete decomposable model is hyper-Dirichlet. They term a density that satisfies global independence a strong hyper-Markov law and show the importance of such laws in the analysis of decomposable graphical

models. We introduce these characterisations of prior distributions in this section and enclose some remarks on these works. The relevant results by Geiger and Heckerman (1997, 1999), and Dawid and Lauritzen (1993) for model selection will be adapted to examine the relationship between independence assumptions and causality, and in an analogous way to characterize prior distributions associated with the parameters of causal Bayesian network. In other words, it is natural to ask how priors might be set up on the uncertain probabilities in the idle system (see Chapter 5) in a way which was invariant to equivalent Bayesian networks. In particular, if we required local and global independence for every Bayesian network compatible with a given PDAG (or more precisely essential graph that will be defined in Chapter 6), the relationship between causality and these independence assumptions and a characterisation of the prior distributions can be deduced (see Chapters 5 and 6 for details).

We next introduce Geiger and Heckerman's results. To compute prior densities associated with multinomial parameters for the given complete network structure in a closed form, Geiger and Heckerman (1997) made several assumptions. These assumptions including local and global independence and parameters modularity, have been used, by Cooper and Herskovits (1992), Madigan and York (1994), Geiger and Heckerman (1997), Yoo and Cooper (1999), and Cowell et al (1999) for model selection.

To characterise prior distributions on the joint probabilities, consider the Bayesian networks shown in Figure 3.2.

Suppose X and Y are two discrete random variables with finite domain, $\{x_i\}_{i=1}^{k}$ and $\{y_j\}_{j=1}^{n}$, respectively. We define $\theta_{ij} = p(X = x_i, Y = y_j \mid \underline{\theta})$, where $\underline{\theta} = \{\theta_{ij}; i = 1, \ldots, k \quad , \quad j = 1, \ldots, n\}$. Let $\theta_{I.} = \{\theta_{i.}\}_{i=1}^{k-1}$ and $\theta_{J|i} = \{\theta_{j|i}\}_{j=1}^{n-1}$, where $\theta_{i.} = p(X = x_i) = \Sigma_{j=1}^{n}\theta_{ij}$, and $\theta_{j|i} = \frac{\theta_{ij}}{\theta_{i.}}$. Similarly, we can define $\theta_{.j}$, $\theta_{i|j}$,

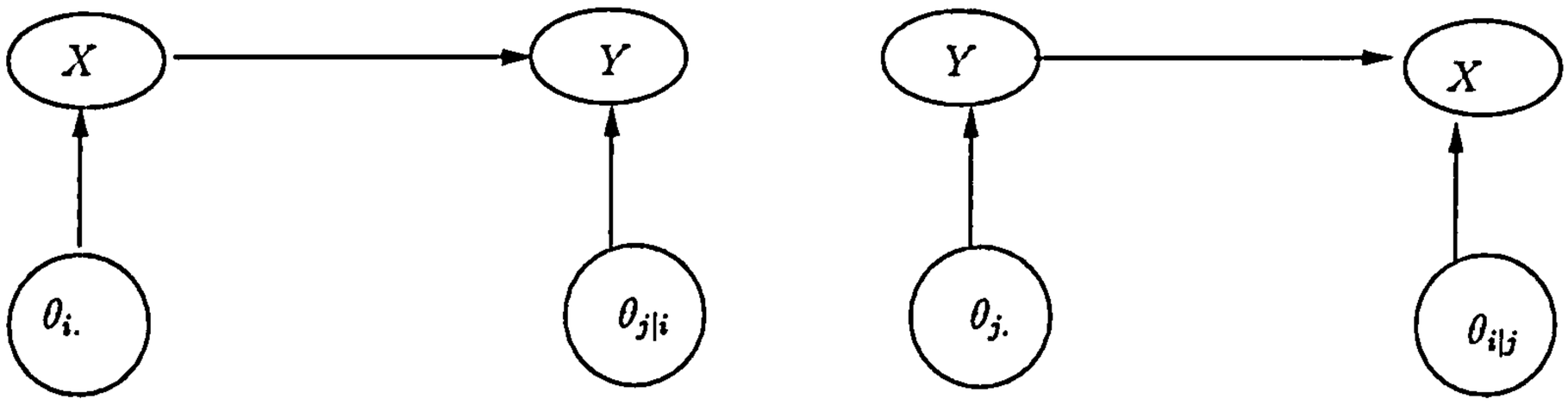


Figure 3.2: Representation of two equivalent Bayesian networks with two multinomial variables.

$\theta_{.J} = \{\theta_{.j}\}_{j=1}^{n-1}$ and $\theta_{I|j} = \{\theta_{i|j}\}_{i=1}^{k-1}$ using the obvious extension of this convention.

Geiger and Heckerman (1997) summarised their main results in the following theorem.

Theorem 3.2 Let $\{\theta_{ij}\}$; $1 \leq i \leq k$, $1 \leq j \leq n$, $\Sigma_i \Sigma_j \theta_{ij} = 1$, where n, k are integers greater than 1, be positive random variables having a positive density $p(\underline{\theta})$. If $\{\theta_{I.}, \theta_{J|1}, \ldots, \theta_{J|k}\}$ are mutually independent and $\{\theta_{.J}, \theta_{I|1}, \ldots, \theta_{I|n}\}$ are mutually independent, then $p(\underline{\theta})$ is Dirichlet.

In other words, we can say that if these independence assertions are assumed to hold, and under the assumption of strictly positive densities, then a prior Dirichlet density for $\underline{\theta}$ is the only possible choice.

Example 3.2 Let us consider the simplest case, that is, two equivalent Bayesian

networks with binary variables (k=n=2), the prior distributions associated with the joint probabilities of Bayesian network given in Figure 3.2 (LHS) and its equivalent Bayesian network given in Figure 3.2 (RHS) will be characterised respectively as

$$p(\theta_x, \theta_{y|X=1}, \theta_{y|X=0}) = \mathcal{B}(\alpha, \beta) \times \mathcal{B}(\alpha_1, \beta_1) \times \mathcal{B}(\alpha_2, \beta_2)$$

and

$$q(\theta_y, \theta_{x|Y=1}, \theta_{x|Y=0}) = \mathcal{B}(\alpha', \beta') \times \mathcal{B}(\alpha_1, \alpha_2) \times \mathcal{B}(\beta_1, \beta_2)$$

where $\alpha = \alpha_1 + \beta_1$, $\beta = \alpha_2 + \beta_2$, $\alpha' = \alpha_1 + \alpha_2$, $\beta' = \beta_1 + \beta_2$,
$\theta_x = p(X = 1 \mid \underline{\theta})$, $\theta_{y|X=1} = p(Y = 1 \mid X = 1, \underline{\theta}) \ldots, \theta_{x|Y=0} = p(X = 1 \mid Y = 0, \underline{\theta})$, and
$\underline{\theta} = \{\theta_{(X=1,Y=1)}, \cdots, \theta_{(X=0,Y=0)}\}$.

In fact, p and q can be calculated from the joint prior distribution of probabilities, i.e., $f(\underline{\theta})$, as follows

$$p(\theta_x, \theta_{y|X=1}, \theta_{y|X=0}) = \theta_x(1 - \theta_x) f(\underline{\theta})$$

and

$$q(\theta_y, \theta_{x|Y=1}, \theta_{x|Y=0}) = \theta_y(1 - \theta_y) f(\underline{\theta})$$

where $f(\underline{\theta}) = \mathcal{D}(\alpha_1, \beta_1, \alpha_2, \beta_2)$.

Therefore,

$$p(\theta_x, \theta_{y|X=1}, \theta_{y|X=0}) = \frac{\theta_x(1 - \theta_x)}{\theta_y(1 - \theta_y)} q(\theta_y, \theta_{x|Y=1}, \theta_{x|Y=0}) \tag{3.9}$$

The result above is generalised for n-variable case in the following theorem presented by Heckerman et al (1995).

Theorem 3.3 Let G_1 and G_2 be two complete network structures for U with variables ordering $(x_1, \ldots, x_n)$ and $(x_n, x_1, \ldots, x_{n-1})$, respectively. If both structures have positive multinomial parameters that obey

$$p(\underline{\theta}_{G_i}) = J_{G_i} p(\underline{\theta}_U), \quad i = 1, 2 \tag{3.10}$$

and positive densities $p(\underline{\theta}_{G_i})$ that satisfy parameter independence, then $p(\underline{\theta}_U)$, $p(\underline{\theta}_{G_1})$ and $p(\underline{\theta}_{G_2})$ are Dirichlet. Here J_{G_i} denotes to the jacobian transformation from $\underline{\theta}_U$ to $\underline{\theta}_{G_i}$.

To summarise we can say that in the discrete case, when samples are of the complete vector (and are multinomial), then under the obvious parametrisation of the Bayesian network, if its defining conditional probabilities are all believed to be mutually independent a priori (local and global independence) then under random sampling these probabilities will remain independent a posteriori. Furthermore if their prior density is believed a priori to have (a conjugate) product of independent Betas (Dirichlets in the multinomial case) then their posterior density will also be a product of independent Betas (Dirichlets). These results are straightforward to establish (see Speigelhalter and Lauritzen (1990) and Spiegelhalter et al (1993)). In fact, we can say that it is very simple to estimate posterior distributions of important quantities when sampling of full vectors, independence of prior probabilities and Dirichlet margins are all appropriate.

From the following, it can be concluded that the independence assumptions make computations of the required posterior quantities such as a posterior distribution, a posterior mean, and a posterior predictive distribution, and so on, more feasible.

In model selection we note that Cowell (1996) gave a different method for assigning Dirichlet priors to the conditional probabilities of structurally different network given a Bayesian network with the discrete domain with a Dirichlet prior. Here the main aim was to find compatible Dirichlet priors of the parameters for the Bayesian networks where the parameter prior is given for one of them and for the other one the structure is given. We can conclude that if the prior distributions of two structurally different Bayesian networks are close in some sense, then it is reasonable to have close

posterior distributions as well. But, there is an ambiguity about the predictions of the future observations. To overcome this issue, he suggests the use of an expectation of Kullback-Leibler distances over all possible future observations to assert a measure of distance between priors. Two of the most important assumptions that he considered in his method were again global and local parameter independence. However, the likelihood-equivalence mentioned above is not considered in his study.

Example 3.3 (BABIES) Spiegelhalter and Cowell (1992) studied a real example concerned with the diagnosis of congenital heart disease in the first days of life on 400 babies at Great Ormond Street Hospital for Sick Children. A Bayesian network for part of the disease spectrum is shown in Figure 3.3.

This example examines the learning procedures on very sparse data. They concluded that initial subjective judgements should not be kept in a model uncritically. So, a sequential monitoring device to enable criticism of prior judgements is introduced.
Our major interest is to introduce global independence on the parameters associated with Figure 3.3. The parameters $\underline{\theta} = \{\theta_v, v = 1, \ldots 20\}$ are globally independent of each other, if $p(\underline{\theta}) = \prod_{v=1}^{20} p(\theta_v)$. As we showed this assumption enable us to factorise the joint probability distribution of $\underline{\theta}$ and $\underline{X}$ as express in (3.7).

To make subsequent computation more feasible in this example with a relatively large set of variables, the local independence assumption should be considered. In fact by appealing to this assumption, θ_v can break into components corresponding to the different configuration of $pa(v)$. For instance, in the figure above, θ_{13} can break into 9 components according to the different configurations of "Lung parenchyma?" and "Lung blood flow?". The parameters associated with these 9 configurations are assumed to be marginally independent variables. Furthermore, this set of parameters would determine

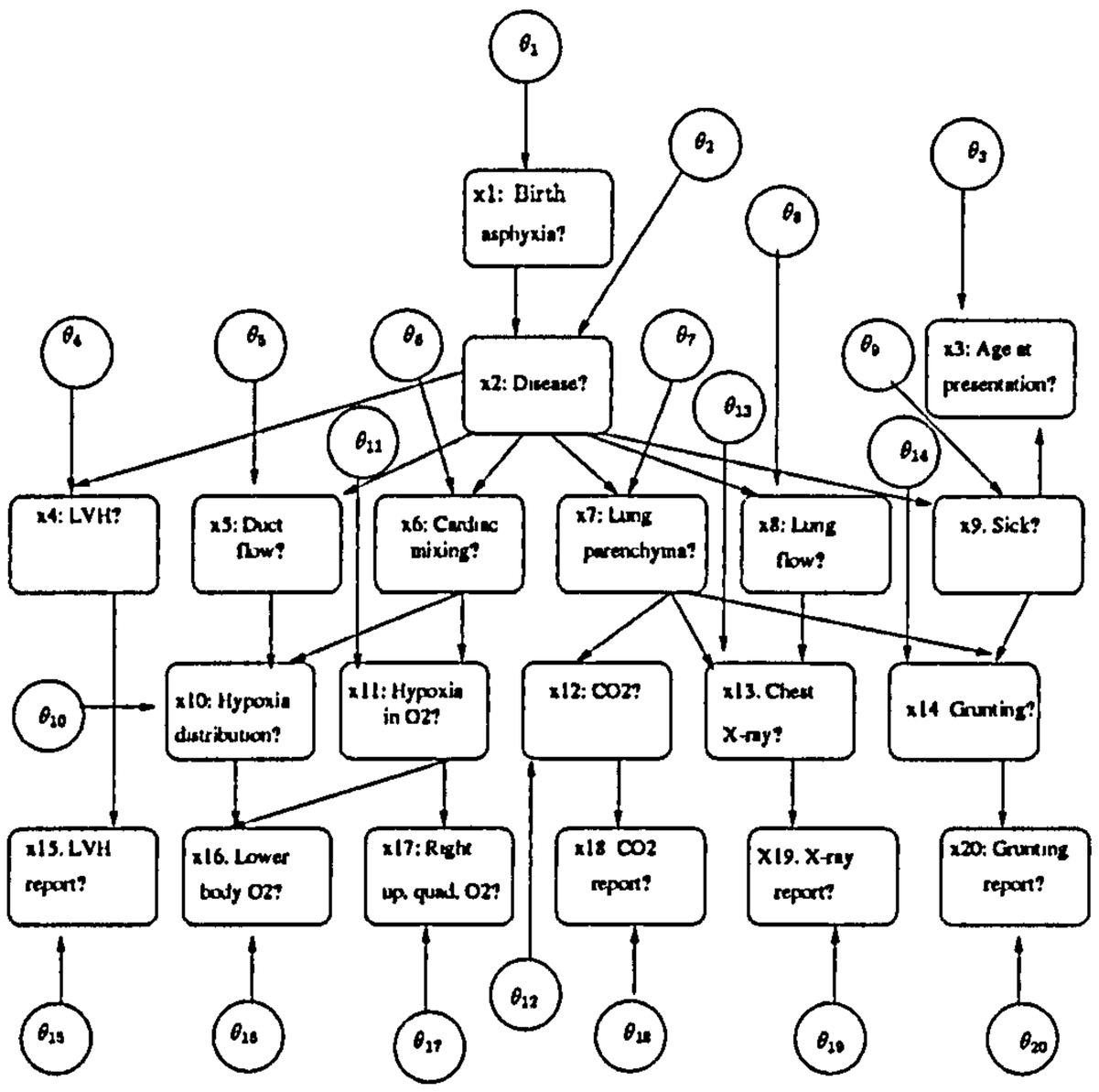


Figure 3.3: The CHILD network: Directed acyclic graph representing possible diseases that could lead to a blue baby.

the differences appearing in Chest X-ray corresponding to those configurations of its parents.

In this thesis, we do not directly study the impact of the incomplete data on the learning of causal Bayesian networks, but we briefly examine the possible issues that might arise for learning of Bayesian networks (or causal Bayesian networks) when the data are incomplete.

Example 3.4 In Example 3.2, consider an incomplete case that is complete on an ancestral set only. In this case, X can be only observed either as $X = 0$ or $X = 1$. Now, suppose that $X = 0$ is observed. Therefore the corresponding likelihood function is $L(\underline{\theta} \mid \underline{X}) = \theta_x$, and the posterior distribution is given by

$$p(\underline{\theta} \mid X = 0) = \mathcal{B}(\alpha, \beta + 1) \times \mathcal{B}(\alpha_1, \beta_1) \times \mathcal{B}(\alpha_1, \beta_1).$$

where $\underline{\theta} = (\theta_x, \theta_{y|x}, \theta_{y|\bar{x}})$.

Therefore, the posterior distribution retains the independence assumptions. Now, consider the case when Y is only observed in two possible ways. If $Y = 1$ is observed, the likelihood becomes

$$P(Y = 1 \mid \underline{\theta}) = \theta_x \theta_{y|X=1} + (1 - \theta_x)\theta_{y|X=0}$$

and cannot be separately factorised into terms involving only single parameters. The posterior density is then as follows

$$p(\underline{\theta} \mid Y = 1) \sim \mathcal{B}(\alpha + 1, \beta) \times \mathcal{B}(\alpha_1 + 1, \beta_1) \times \mathcal{B}(\alpha_2, \beta_2) +$$

$$\mathcal{B}(\alpha, \beta + 1) \times \mathcal{B}(\alpha_1, \beta_1) \times \mathcal{B}(\alpha_2 + 1, \beta_2).$$

Therefore, local independence assumptions are not satisfied here, and the posterior distribution is a mixture of the Beta distributions.

In choosing a Bayesian network, one consideration is to specify whether, and how, relationships between variables are causal. In Chapter 5, we will show that the Bayesian networks shown in Figure 3.2 (LHS) and 3.2 (RHS) are causal, in a sense to be defined precisely later in Chapter 4, if and only if their prior distributions exhibit parameter independence assumptions.

A final result concerning local and global parameter independence is given by Rusakov and Geiger (2000). They show that local parameter independence is essential in the characterization of a Dirichlet prior for a discrete Bayesian network. They determine the minimal set of assumptions (including the local independence assumption) required to have a Dirichlet prior. They also present the functional form of prior distributions that arise under the global independence assumption alone. Although these functional forms are not closed forms, they can be used in a sensitivity analysis of the prior distribution associated with the Bayesian networks corresponding to lack of local independence assumption. However, it is obvious that the computation of the sensitivity measures proposed in Chapter 7 will be complex for this purpose. But general theory with some examples of simpler cases will be studied.

Chapter 4

Causality

4.1 Introduction

In recent years, graphical models have been used to represent and manipulate complex multivariate probability distributions. Several authors, such as Spirtes et al (1993), Pearl (1995, 2000), Dawid (2000, 2002) and Lauritzen (2001), apply these representations to study the causal relationships between variables for the given system, and modeling this manipulated system by helping graphical models. There are different frameworks for causal modelling based on the directed acyclic graph that are firstly introduced by Spirtes et al (1993) and mostly developed in Pearl (2000) with contributions, criticisms and discussions by other. For example, Dawid (2000) discussed and criticised Pearl's functional models framework of causal models that is closely related to counterfactual models. In 2002, he extended and defined the causality concepts for influence diagrams. Lauritzen (2001) studied the relationship between causal models and randomised trials. As we said earlier, one of the main objectives of this thesis is to examine the relationship between causal models and contingent randomised trials. Furthermore, we make a connection between the independence assumptions discussed in Chapter 3 and causality. We present these results in Chapters 5 and 6. It should be noticed that, throughout

this thesis, we assume that a pattern has already been chosen, and we want to elaborate this model with further causal assumptions.

It has become clear that graphical models, in particular those based upon directed acyclic graphs, have natural causal interpretations and thus form a base for a language in which causal concepts can be discussed and analysed in precise terms. This chapter is dedicated to review and introduce basic concepts of causal models which are defined in terms of DAGs.

4.2 Introduction to Causal Models

The basic knowledge when reasoning under uncertainty is whether information on some events influences your belief of other events.

When we build DAG models, we can see explicitly or implicitly that these models might be used to represent various causal relationships between variables. For example, consider the following pattern of dependencies among three events: A and B are dependent, B and C are dependent, but $A \perp\!\!\!\perp C$. Pearl (2000) provide an example of three such events describe above. This example would invariably portray A and C as two independent causes and B as their common effect, that is, $A \rightarrow B \leftarrow C$. But, we would rather say that the arrows into B depict that the values of both A and C may influence the prediction of B.

Pearl (1995) expresses causal mechanisms in terms of a DAG, which he attempts to identify from data. But, it should be noticed that causation cannot be obtained directly from the interpretation of DAG as conveyors of conditional independence statements. In fact, a DAG is valid for any set of recursive independencies along with any ordering of the variables, not necessarily causal or chronological. However, the ubiquity of DAGs

in statistical applications basically develops from their causal interpretation. Thus, it is unusual for DAGs to be applied in any variable ordering other than those which respect the direction of time and causation.

There are advantages to building DAG models around causal rather than associational information. These advantages are listed below.

1. The possibility of eliciting more reliable judgements required in network (e.g., try ordering (X_a, X_l, X_b, X_s) of variables in the Bayesian network shown in Figure 4.1). The conditional independence judgments are easily accessible only when they are anchored onto causal relationships.

2. To provide a representation of action and change (remodelling).

The following example will help us to see the advantages of construction DAG model in terms of causal relationships between variables involved in the model.

Example 4.1 Let us consider a very well-known survey that has been used by several authors such as Robins (1997) and Lauritzen (2001) concerned with a large group of AIDS patients (We assume that this population is so big that we can ignore sampling error). This study consists of four binary variables. We introduce briefly the variables involved in this example as follows. Let us denote a as the label for an initial, randomised treatment, where $X_a = 1$ denotes that the patient has been treated with AZT, and $X_a = 0$ indicates placebo. After a given period it is for each patient observed whether the patient develops pneumonia, corresponding to the variable l, where X_l displays that this is the case. We assume that all patients survive up to this point. Subsequently

a secondary treatment with antibiotics is contemplated, corresponding to the variable b. For ethical reasons, all patients who have developed pneumonia are treated with antibiotics, namely, $Pr(X_b = 1 \mid X_l = 1) = 1$, whereas the treatment is randomised for the patients with $X_l = 0$. Finally after a given period it is registered whether a patient has survived up to that time, corresponding to the variable s, where $X_s = 1$ indicates that the patients has survived.

The DAG describing causal relationships between variables in this example is given in Figure 4.1. This graph is only assumed causal with respect to intervention at $\{X_a, X_b\}$. Note that the missing arrow from X_a to X_b indicates that X_b is assigned by randomisation. This point will be discussed in more details throughout this thesis.

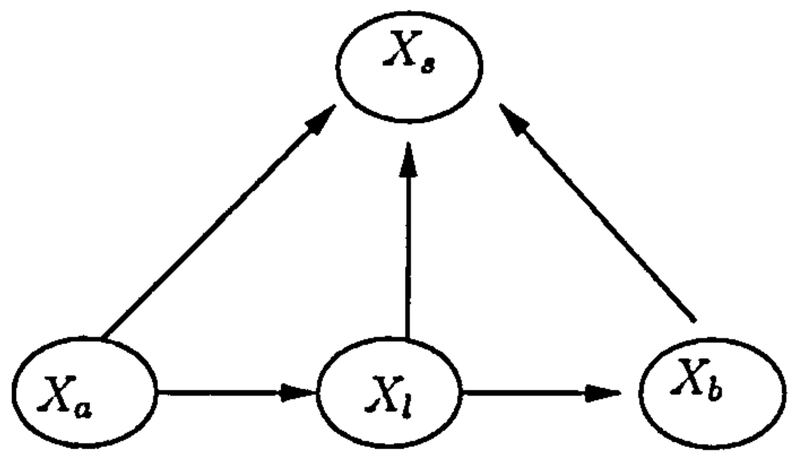


Figure 4.1: DAG model indicating causal relationships between variables in example above.

The Bayesian network shown in Figure 4.1 gives the following factorisation

$$p(x_a, x_b, x_l, x_s) = p(x_a)p(x_l \mid x_a)p(x_b \mid x_l)p(x_s \mid x_a, x_l, x_b). \tag{4.1}$$

In this example, the advantage of constructing the DAG model associated with inde-

pendence statements with the following ordering (X_a, X_l, X_b, X_s) is not only that some of these statements are more clear with respect to others but also that the conditional independence assessments are convenient and reliable only in the case when they are anchored onto more reliable and available information such as causal relationships.

The second advantage can be demonstrated by examining the following manipulation. The causal Bayesian network model corresponding to the intervention $do(X_b = 1)$, can be obtained by cutting all arrows converging into node X_b and revising $p(x_s \mid x_b, x_a, x_l)$.

We can say that deleting $p(x_b \mid x_l)$ (and corresponding arrow) indicates that, whatever relationship was between X_l and X_b before the intervention above, this relationship is no longer valid after implementing the intervention. In fact, the situation of X_b will be specified under a new mechanism due to this intervention.

The third advantage of building a DAG model around causal relationships between variables is ease of reconfiguration (See Pearl (2000)).

In other words, the mechanism of the uncertain process would be changed by an external intervention. Subsequently the character of this change, that is, the total effect of the intervention can be computed or predicated[1] by updating the probability distribution as it will be shown below (or by revising corresponding equations in the functional causal model that will be examined in Section 4.3). In fact, the connection between modularity and intervention can be considered as a function of both alteration and simulation (Pearl (2000)).

For instance, the graph and corresponding joint distribution after performing the inter-

[1]It can be said that probabilities do not predict effects of interventions.

vention $do(X_b = 1)$, corresponding to Example 4.1, are respectively given by

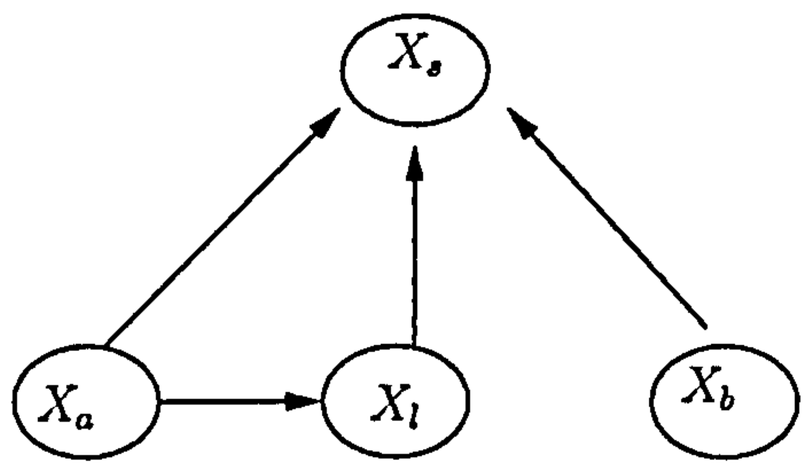


Figure 4.2: Graph indicating causal relationships between variables with respect to the intervention $do(X_b = 1)$.

$$p(x_a, x_b, x_l, x_s \mid \mathrm{do}(X_b = 1)) = p(x_a)p(x_l \mid x_a)p(x_s \mid x_a, x_l, x_b = 1) \qquad (4.1*)$$

where all terms on the right-hand side of the equation above, by virtue of autonomy, are the same as in Equation (4.1).

A system is called *idle* when its variables are observed and their values are not manipulated. Such data arise, for example, in cross sectional observational studies with no intervention. So, let $p(\underline{x})$ represent a probability mass function on a set $\underline{X}$ of discrete random variables consistent with the conditional independence relations coded in a Bayesian network G in an idle system. It is not unusual to want to make inferences about what will happen when specific variables in the system are manipulated to take certain values.

Let $p(\underline{x} \mid do(V = v))$ denote the distribution resulting from the intervention $do(V = v)$

that manipulates on a subset V of variables and forces them to take values v. Denote by p_* the set of all distributions $p(x \mid do(V = v))$, $V \subseteq \underline{X}$ including $p(\underline{x})$, which represent no intervention (i.e., $V = \emptyset$).

The following example makes clear the difference between *conditioning* and *manipulation* in Bayesian networks.

Example 4.2 Consider the Bayesian network given in the following figure. Here C represents *nuclear core activity* per hour; the *maximum temperature of cooling system* in an hour is represented by T; F indicates *failure of cooling system* in an hour. For simplicity, we will discretise the problem. So suppose, the possible values for T, F and C are respectively given by, $(0^0 - 100^0, 100^0 - 200^0, 200^0 - 300^0, 400^0+)$, (0, 1) and (Normal, Critical, Meltdown).

In the Bayesian network above, setting T to one of its values is just artificially increasing

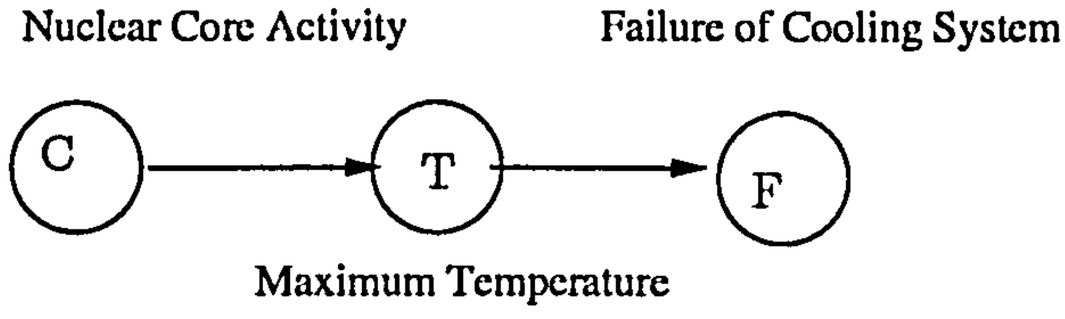


Figure 4.3: The Bayesian network for the Nuclear Activity' Example

the temperature. Clearly this will not affect core activity. However, conditioning on T (e.g., observing a high temperature at the value) is indicative of core activity. In both cases, if the failure of the cooling system is believed only to depend on T, then a natural extension of the unmanipulated Bayesian network is to make the obvious additional

assumptions that doing T will have the same effect as observing T, i.e.,

$$p(F \mid do(T = t)) = p(F \mid T = t).$$

This motivates the following definition of a causal Bayesian network on a vector of measurements, $x = (x_i, i \in \mathcal{I} = \{1, \ldots, k\})$ (Pearl (1995)).

Definition 4.1 (Causal Bayesian Network) A Bayesian network G is said to be a *Causal Bayesian Network* compatible with p_* if and only if the following three conditions hold for every $p(. \mid do(V = v)) \in p_*$:

(i) $p(\underline{x} \mid do(V = v))$ is Markov relative to G;

(ii) $p(x_i \mid do(V = v)) = 1$ for all $X_i \in V$ whenever x_i is consistent with $V = v$, and is otherwise zero;

(iii) $p(x_i \mid pa_i, do(V = v)) = p(x_i \mid pa_i)$ for all $X_i \notin V$ whenever pa_i is consistent with $V = v$, and is otherwise zero.

Causal Bayesian networks embody fierce assumptions. However there are many practical contexts, like the one above when this elaboration of a Bayesian network of an observational study is plausible. In such a context the causal Bayesian network encodes a large set of assertions efficiently in the form of a single graph. The formula in Definition 4.1 imply that the distribution $p(x \mid do(V = v))$ resulting from an intervention $do(V = v)$ will result the truncated factorisation

$$p(\underline{x} \mid do(V = v)) = \prod_{\{i \mid x_i \notin V\}} p(x_i \mid pa_i) \quad \forall\, x_i \text{ consistent with } v, \tag{4.2}$$

and

$$p(\underline{x} \mid do(V = v)) = 0, \quad x \text{ not consistent with } v.$$

When G is a causal Bayesian network with respect to p_*, the following two properties must hold.

Property 1 For all i,

$$p(x_i \mid pa_i) = p(x_i \mid do(Pa_i = pa_i))$$

Property 2 For all i and for every subset S of variables disjoint for $\{X_i, Pa(i)\}$, we have

$$p(x_i \mid do(S = s, Pa_i = pa_i)) = p(x_i \mid Pa_i = pa_i)$$

where Pa_i denote parent set of its child X_i.

Note that in Example 4.2, the intervention $do(C = Normal)$ remains invariant to changes in all mechanisms shown in this causal graph. More precisely, we can say that causal relationships remain invariant with respect to changes in the mechanism that governs the causal variables (e.g, T in the last example).

It is more desirable to model data and information therein in terms of causal relationships rather than probabilistic models. The probabilistic claims such as marginal and conditional independencies, may be useful in testing of the initial hypotheses associated with causal claims from uncontrolled observations. Furthermore, the probabilistic claims depend on the context in which those mechanisms are embedded. But, whatever judgments people express about the conditional independence statements for the given task are obtained based on the appropriate causal structure. Another point that can be made about the causal claims is that these claims are sensitive to only those mechanisms that mediate between the cause and effect.

4.3 Functional Causal Models

In this section the relationship between a graphical Markov model represented by a DAG and structural equation models in which the functional relationships may be nonlinear will be discussed. In fact, by appealing of structural equation models[2] , we can translate the causal model into a set of mathematical equations to yield an observational (statistical) model. In this statistical model, certain assumptions must be made about the form of functional equations and corresponding sampling distribution of random variables[3].

To assert causal influences, such us those determined by arrows in the DAG shown in Figure 4.1, denote independent physical mechanisms among the corresponding quantities, and these mechanisms can be exhibited by functional relationships contaminated by the source of confusions called *random disturbances.* Pearl and Verma (1991) and Pearl (2000) describe each child-parent relationship represented in terms of a directed graph G as a deterministic function given below

$$X_i = f_i(pa_{x_i}, \epsilon_i), \quad i = 1, \ldots, n \tag{4.3}$$

where pa_{x_i} stands for the parent set of variable X_i in G, and $\{\epsilon_i\}_{i=1}^{n}$ is the corresponding set of disturbances that are mutually independent and arbitrarily distributed.

---

[2]It should be noticed that the structural equation model described above can be interpreted as an asymmetrical *counterfactual relation*, and each equation within the set of equations describes a stable and independent mechanism.

[3]In most cases, the random variables are assumed to be multivariate normal and the equations are additively linear. Each variable is normally associated with parameters such as variance, covariances etc. If there is sufficient information about the variables, these parameters can be assigned values and are called fixed parameters. If there is not enough information or confidence to assign a value, the value may be estimated from the data. Such parameters are called free parameters.

Note that, often in the literature the variables X_j $(j \in pa_{x_i})$ are called causal factors of X_i, and the set of equations represented in (4.3) is called the (recursive) structural equations system that can be considered as a mathematical representation of the supposed causal theory[4] for $X_1, \ldots, X_n$.

The functions $f_1, \ldots, f_n$ can be specified completely, or partly (e.g., linear) or not at all, depending on the degree of articulation of the causal theory. In the structural equation model given in (2.3), if functions $f_1, \ldots, f_n$, disturbances (residual) variables $\epsilon_1, \ldots, \epsilon_n$, and sets $pa_{x_i} \subseteq \{1, \ldots, i-1\}$, $(i = 1, \ldots, n)$ are given, then $X_1, \ldots, X_n$ can be defined recursively by the structural equations $X_i = f_i(X_{pa_{x_i}}, \epsilon_i)$, $i = 1, \ldots, n$.
The structural equations system can be represented by a DAG $D = (\underline{X}, E)$, where

- $\underline{X} = (X_i; i = 1, \ldots, n)$ is a set of vertices representing variables.
- E is a set of arrows: $j \rightarrow i \in E$ if and only if $j \in x_{pa_i}$.

and graph D is called a causal graph for $X_1, \ldots, X_n$.
In the following example we give a structural equation representation of the DAG shown in Figure 4.1.

Example 4.3 The causal assumptions that are delivered by the model presented in the Figure 4.1 are given by the collection of the following equations,

$$X_a = f_a(\epsilon_a), \quad pa(x_a) = \emptyset$$

$$X_l = f_l(X_a, \epsilon_l), \quad pa(x_l) = \{X_a\}$$

$$X_b = f_b(X_l, \epsilon_b), \quad pa(x_b) = \{X_l\}$$

$$X_s = f_s(X_a, X_l, X_b, \epsilon_s), \quad pa(x_s) = \{X_a, X_b, X_l\}$$

[4]The causal theory is a substantive theory specifying causal mechanisms.

For instance, to represent the action $do(X_b = 1)$ in the model above, we delete the equation $X_b = f_b(X_l, \epsilon_b)$ and replace it with $X_b = 1$. The modified model will contain all the information needed for computing the effect of the action on the other variables. Furthermore, the probability function induced by the modified (structural) model will be equal to that given by Equation (4.1*) and the modified graph will coincide with that of Figure 4.2. In other words, the equation associated with X_b indicates that regardless to the values of X_l and regardless to the possible changes that could be occurred in the equations associated with X_s and X_a, if (X_l, ϵ_b) were to assume to take[5] the values (x_l, ϵ_b), X_b would then take on the value imposed by the function $f_b(x_l, \epsilon_b)$.
It should be noticed that the equation f_b is only under influence of the set of variables who are neighbours of X_b and have converging arrows into X_b.

Now, suppose $X_i = f_i(pa_{x_i}, \epsilon_i), \quad i = 1, \ldots, n$ is a recursive structural equation system with causal graph D defined above. The intervention in the structural equation systems is defined as follows

Definition 4.2 Structural equation system after the intervention $do(X_i = x_i^*)$ is the system $X_j^* = f_j^*(pa_{x_j^*}, \epsilon_j), j = 1, \ldots, n$, which is defined recursively by:

- $X_j^*, \qquad f_j^* = f_j, j < i.$
- $X_i^* = x_i^*.$
- $X_j^* = f_j^*(pa_{x_j^*}, \epsilon_j), \qquad f_j^* = f_j, j > i.$

The following theorem can be immediately concluded.

---

[5]Pearl (1995) used the functional specification as a convenient language for specifying how the resulting distribution would alter in response to external interventions.

Theorem 4.1 If $D^* = (\underline{X}^*, E^*)$ is the (causal) graph obtained from $D = (\underline{X}, E)$ by deleting all arrows $j \to i$, then distribution of $\underline{X}^* = (X_j^*, j = 1, \ldots, n)$ is Markov[6] with respect to D^*. Furthermore, it can be said that the marginal distribution of $\underline{X}^*_{V\setminus\{i\}} = (X_j^*; j \in V \setminus \{i\})$, where $V = \{1, \ldots, n\}$, is Markov with respect to $D_{V\setminus\{i\}}$, the induced subgraph obtained from D by deleting vertex i and all its adjacent arrows (both incoming and outgoing).

Now, consider the situation where some variables are simultaneously forced to get specified values. In other words, consider a system with multiple external interventions. The characterisation of structural equation system and corresponding causal Bayesian network under multiple interventions are given below.

Let $A \subseteq V$. The set of equations under the multiple manipulations, $do(X_A = x_A^*)$ is given by successive single manipulation $do(X_i = x_i^*)$, for $i \in A$. If $p(x_1, \ldots, x_n) = \prod_{i=1}^n p(x_i \mid pa(x_i))$ denote the joint density function of
$\underline{X} = (X_i; i = 1, \ldots, n)$ associated with the causal Bayesian network D, then the joint density function of this system after the manipulation $do(X_A = x_A^*)$ is given by

$$p(x_1, \ldots, x_n \mid do(X_A = x_A^*)) = \prod_{i \in A} 1_{\{x_i^*\}}(x_i) \prod_{i \in V\setminus A} p(x_i \mid pa(x_i))$$

where $1_A(x)$ stands for the indicator function that is 1 for $x \in A$ and 0 otherwise.
As a result, we can say that the distribution of $\underline{X}^* = (X_i^*; i \in V)$ is Markov with respect to D^* that is obtained from the original graph D by deleting all arrows $j \to i$ $(i \in A)$. Moreover, the marginal distribution of $\underline{X}^*_{V\setminus A} = (X_j^*; j \in V \setminus A)$ is Markov with respect

[6]The causal Markov condition for the conditional independence relationships is: A node is independent of its non-descendants (i.e., non-effects) given its parents (i.e., direct causes). The causal Markov condition permits the joint distribution of the n variables in a causal Bayesian network to be factored as in Equation (4.2).

to $D_{V\setminus A}$, the subgraph of D induced by $V \setminus A$.

Example 4.4 In Examples 4.1 and 4.3, the original structural equation system and corresponding causal graph are represented. The structural equation model and corresponding causal graph (Figure 4.4) with respect to the intervention $do(X_a = 1, X_b = 1)$ are given respectively as

$$X_a^* = 1$$

$$X_l^* = f_l(X_a^*, \epsilon_l)$$

$$X_b^* = 1$$

$$X_s^* = f_s(X_a^*, X_l^*, X_b^*, \epsilon_s),$$

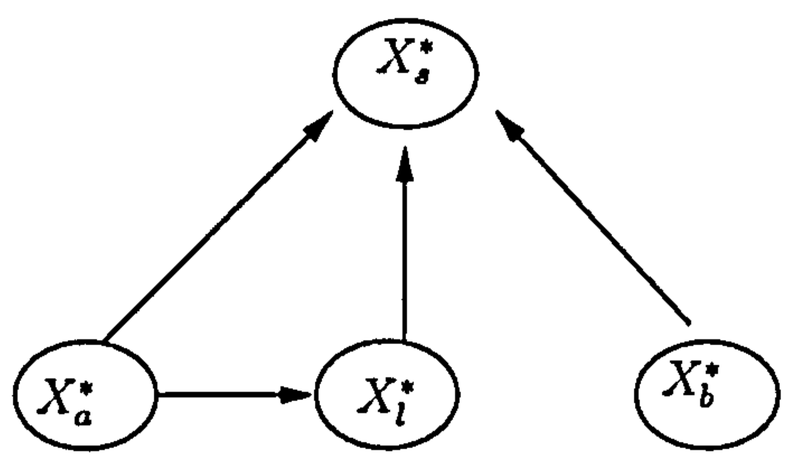


Figure 4.4: Bayesian network associated with $do(X_a = 1, X_b = 1)$.

The structural equation model associated with the multiple interventions helps us to define the randomised intervention that will be discussed in the next section and Chapter 5.

4.4 Randomised Intervention

The idea of the *randomised intervention* that was introduced by Koster (2000), Robins (1997) and Lauritzen (2001) will be presented in this section. This idea will be developed to the randomised contingent intervention in Chapter 5.

Suppose that the multiple interventions $do(X_i = x_i^*),\ i \in A$ are random and independent, that is, $p(do(X_A = x_A^*)) = \prod_{i \in A} p_i^*(x_i^*)$, where p_i^* is some probability distribution[7] on the state space of X_i. Then, the randomised intervention can be given in terms of the multiple intervention above as

Definition 4.3 (Randomised Intervention) The multiple intervention $do(X_A = x_A^*)$ is called *randomised*, if the interventions $do(X_i = x_i^*)\ \ i \in A$ are mutually independent and random, that is,

$$p(do(X_A = x_A^*)) = \prod_{i \in A} p_i^*(x_i^*)$$

for some probability density function, p_i^*, that is defined on the state space of X_i.

The joint density function of $\underline{x} = (x_1, \ldots, x_n)$ after enforcing the randomised intervention $do(X_A = x_A^*)$ is given by

$$p^*(x_1, \ldots, x_n \mid do(X_A = x_A^*)) = \prod_{i \in A} p_i^*(x_i) \prod_{i \in V \setminus A} p(x_i \mid pa_{xi}). \tag{4.4}$$

Now, the question that might be asked here is: can we combine formulas associated with the densities of idle system (system with no intervention), system with the single intervention and system with the randomised manipulation into a single framework?

[7]Note that the multiple interventions as discussed in the last section is usually considered as: $p(do(X_A = x_A^*)) = \quad_{i \in A} 1_A(x_i)$.

The *augmented causal graph* that is introduced by several authors such as Pearl (2000), Dawid (2002) can help us to find this framework. We first define the augmented causal graph given below.

Suppose that we can force X_i to take its values in its domain, $\mathcal{X}_i$. Let the decision variable F_{X_i} represent the intervention situation of variable X_i. F_{X_i} takes its values in $[\{\text{state space of } X_i\} \cup \{\phi_i^*\}]$ for $i \in A \subseteq \underline{X}$ as follows: if $F_{X_i} = \phi_i^*$, that means X_i takes its values naturally; but $F_{X_i} = \phi_i$, where $\phi_i \in \{\text{state space of } X_i\}$ (or $\phi_i \in \mathcal{X}_i$) represents a manipulation that set the value of ϕ_i to a $x_i \in \mathcal{X}_i$ or set X_i to x_i.

Now, for each $i \in A \subseteq \underline{X}$, an additional vertex F_{X_i} and an additional arrow $F_{X_i} \rightarrow X_i$ are defined for the given DAG $D = (\underline{X}, E)$. The graph that is defined over $\underline{X}$ and new vertices $\{F_{X_i}\}_{i \in A}$ is called the *augmented graph* and shown by $\hat{D} = (\underline{\hat{X}}, \hat{E})$, where $\underline{\hat{X}} = \underline{X} \cup \{F_{X_i}\}_{i \in A}$ and $\hat{E} = \underline{\hat{X}} \times \underline{\hat{X}}$. Note that all disturbance variables, $\epsilon_i,\ i = 1, \ldots n$ in the structural equation model corresponding to D and $F_{X_i},\ i \in A$ are independent.

Dawid (2002) described the conditional distribution corresponding to the augmented nodes. The conditional distribution of specific node, X_i, given pa_{x_i} and specific value of ϕ_i corresponding to the augmented node, i.e., F_{X_i}, is the same as the original distribution of X_i given pa_{x_i} if $\phi_i = \phi_i^*$; otherwise puts all its probability mass on $X_i = x_i$. In other words, for $\phi_i \in [\{\text{state space of} X_i\} \cup \{\phi_i^*\}],\ i \in A$, the joint distribution of (x_i, pa_{x_i}, ϕ_i) is given by

$$p(x_i, pa_{x_i}, \phi_i) = \begin{cases} p(x_i \mid pa_{x_i}) & \text{if } \phi_i = \phi_i^*; \\ 1_{\phi_i}(x_i) & \text{if } \phi_i \neq \phi_i^*. \end{cases} \tag{4.5}$$

Similarly, the joint density of $(x_1, \ldots, x_n, \{\phi_i\}_{i \in A})$ with respect to the randomised

intervention is given by

$$\tilde{p}(x_1,\ldots,x_n,\{\phi_i\}_{i\in A}) = \prod_{i\in A} p_i^*(\phi_i) \prod_{i\in A} p_i(x_i, pa_{x_i}, \phi_i) \prod_{i\in V\setminus A} p(x_i \mid pa_{x_i}),$$

where p_i^* is some density that is defined on $\{\phi_i\}_{i\in A}$, and $p_i(x_i, pa_{x_i}, \phi_i)$ is defined in Equation (4.5).

In fact, $\tilde{p}$ is a density which is decomposed with respect to the augmented graph $\tilde{D} = (\tilde{X}, \tilde{E})$, or $\tilde{p}$ can be considered as a density for variables $\tilde{X}_i \in \tilde{X}$, $i \in V$ and F_i, $i \in A$ satisfying the following augmented structural equation model,

- $\tilde{X}_i = h_i(pa_{\tilde{x}_i}, F_{X_i}, \epsilon_i) = \begin{cases} f_i(pa_{\tilde{x}_i}, \epsilon_i) & \text{if } F_{X_i} = \phi_i^*; \\ x_i^* & \text{if } F_{X_i} = x_i^*. \end{cases}, \quad i \in A.$
- $\tilde{X}_i = f_i(pa_{\tilde{x}_i}, \epsilon_i), \quad i \in V \setminus A.$

Example 4.5 The augmented graph associated with the causal DAG shown in Figure 4.1 is given in the following figure

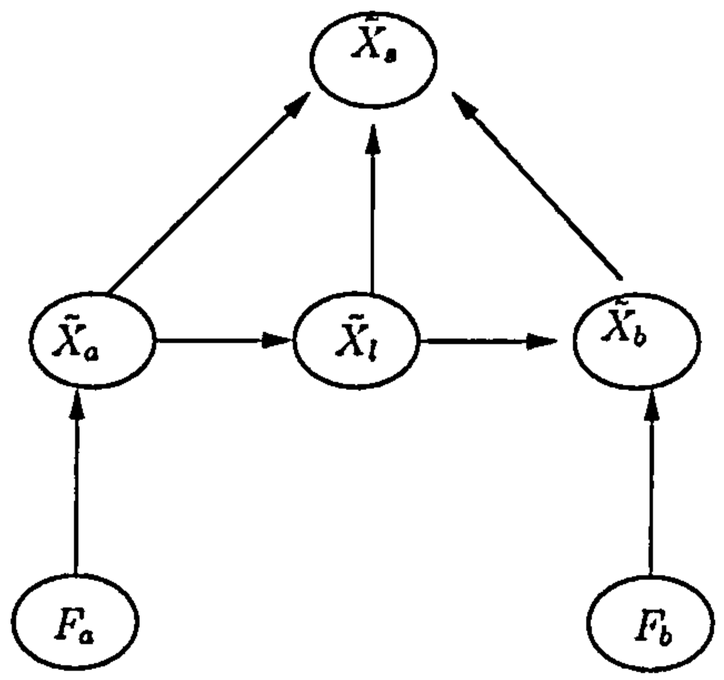


Figure 4.5: The augmented DAG corresponding to the DAG shown in Figure 4.1.

Furthermore, $p(X_s = 1 \mid do(X_a = 1, X_b = 1))$ is computed as follows

$$p(X_s = 1 \mid do(X_a = 1, X_b = 1)) =$$

$$\Sigma_{x_a}\Sigma_{x_l}\Sigma_{x_b} p(x_a, x_l, x_b, X_s = 1 \mid do(X_a = 1, X_b = 1)) =$$

$$\Sigma_{x_a}\Sigma_{x_l}\Sigma_{x_b} p(\tilde{X}_a = x_a, \tilde{X}_l = x_l, \tilde{X}_b = x_b, \tilde{X}_s = 1 \mid F_a = 1, F_b = 1) =$$

$$p(\tilde{X}_s = 1 \mid F_a = 1, F_b = 1) = \Sigma_{x_l} \tilde{p}(\tilde{X}_s = 1 \mid \tilde{X}_l = x_l, F_a = 1, F_b = 1) \times$$

$$\tilde{p}(\tilde{X}_l = x_l \mid F_a = 1, F_b = 1).$$

It can be shown that

$$\tilde{p}(\tilde{X}_s = 1 \mid \tilde{X}_l = x_l, F_a = 1, F_b = 1) = p(X_s = 1 \mid X_l = x_l, X_a = 1, X_b = 1)$$

and

$$\tilde{p}(\tilde{X}_l = x_l \mid F_a = 1, F_b = 1) = p(X_l = x_l \mid X_a = 1)$$

Therefore

$$p(X_s = 1 \mid do(X_a = 1, X_b = 1)) = \Sigma_{x_l} p(X_s = 1 \mid X_l = x_l, X_a = 1, X_b = 1) \times$$

$$p(X_l = x_l \mid X_a = 1).$$

4.5 Learning Causal Bayesian Networks

There has been considerable recent research on the topic of learning causal Bayesian networks. Although this one is not central to the material in this thesis, there are resonances and analogous which make it helpful to finish this section with a brief survey of this contentious topic.

It should be noticed that we assume that a pattern has already been chosen from data, and we want to elaborate this model with further causal assumptions. These issues

will be studied in more details in Chapters 5 and 6.

Obviously, the first difficulty in learning a causal Bayesian network from data is that many Bayesian networks are equivalent. So it is only reasonable to expect to be able to learn pattern rather than Bayesian networks (DAGs) themselves. If we want to associate the direction of causality with the direction of an edge in a DAG then, because we can only learn about patterns, many of these directions will be inaccessible, and obstructing the inference of causality. A simple example is given in the following figure.

Figure 4.6: Two network structures (DAGs) are in the same equivalent class, and the edge direction can not be inferred from observational studies alone. The causal direction can be deduced by setting the value for node X externally. If X is causal ancestor of Y, this intervention is likely to lead to a changed value of Y. If, however, Y is a causal ancestor of X, this intervention will have no effect on Y.

More formally, as we know the likelihood scores of two equivalent Bayesian network structures $\mathcal{B}$ and $\mathcal{B}'$ are the same for the likelihood computed from (2.3). As discussed earlier in this chapter, forcing a variable to take the specific value externally means that the respective node takes on this particular value with probability 1 irrespective of values of other nodes in the Bayesian network under study. Consequently, the contributions of all those nodes that are subject to intervention effectively disappear from (3.5) and decomposition (4.2) will be obtained. This modification can destroy the symmetry within an equivalence class of Bayesian networks, that is, the likelihood scores for $\mathcal{B}$ and $\mathcal{B}'$ might no longer be the same, which may resolve the ambiguity about certain edge

directions.

The second and potentially more serious difficulty is existence of possible hidden, unobserved variables. For instance, Figure 4.6 shows two network structures that explain the conditional independence between two random variables. However, a third possibility is that both observed random variables depend on a third, hidden variable, as is shown in Figure 4.7.

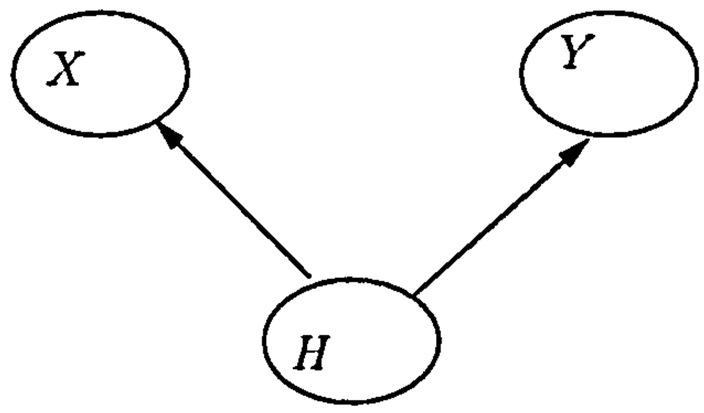


Figure 4.7: A representation of network structure between two observed variables X and Y and hidden variable H.

Another example is given in Figure 4.8, where the network on the left hand side, which only includes observed nodes, is equivalent to the network structure on the right hand side, which contains two extra hidden nodes, H_1 and H_2. By applying (3.5) to the graph with two hidden variables, it can be concluded that

$$p(X, Y, Z, H_1, H_2) =$$

$$p(X \mid H_1)p(H_1)p(Z \mid H_1, H_2)p(Y \mid H_2)p(H_2) = p(H_1 \mid X)p(X)p(Z \mid H_1, H_2)p(H_2 \mid Y)p(Y).$$

Then, marginalising over unobserved variables H_1 and H_2, gives

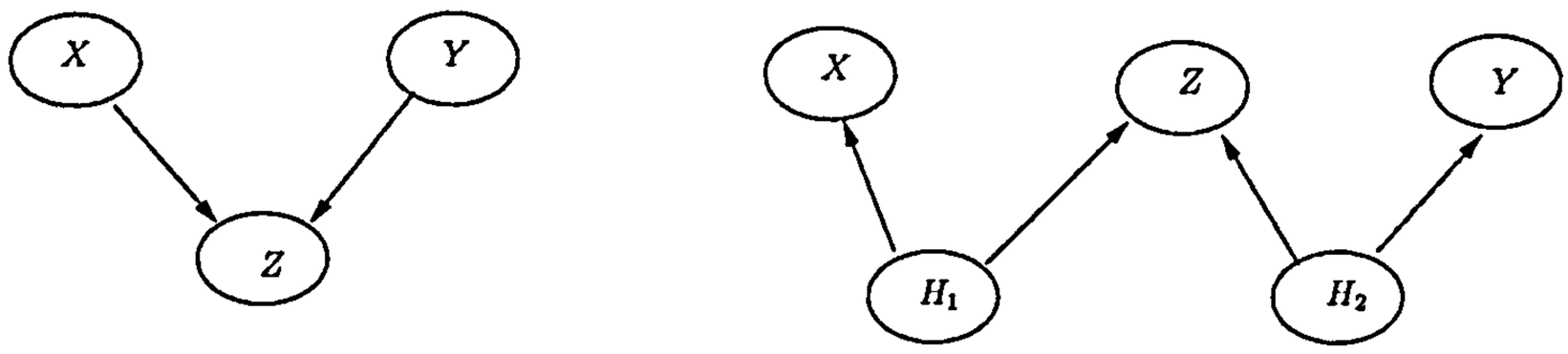


Figure 4.8: The network structure on the left without hidden variables is equivalent to the network on the right hand side, with two additional hidden variables, H_1 and H_2.

$$p(X,Y,Z) = \Sigma_{H_1}\Sigma_{H_2}p(X,Y,Z,H_1,H_2) =$$

$$p(X)p(Y)\Sigma_{H_1}\Sigma_{H_2}p(Z \mid H_1,H_2)p(H_2 \mid Y)p(H_1 \mid X) = p(X)p(Y)p(Z \mid X,Y)$$

This result is identical to the decomposition that is obtained from the Bayesian network structure on the left hand side of Figure 4.8. Thus, two network structures in this figure are identical, and it is not possible to decide whether X and Y are causal ancestors of Z, or whether all these variables are controlled by some hidden causal ancestors.

Spirtes et al (1999), in their study of a more complex case, argue that it is possible to characterise all network structures with latent variables that can result in the same set of independence relations over the observed variables, and such equivalence classes are called *partial ancestral graphs*. However, as opposed to PDAGs, it is not clear how to score a partial ancestral graph, which consists of many models with different numbers of latent variables, and this defies the Bayesian MCMC approach.

But models with latent structure also contain other implications such as identifiability

issue which will be studied in Chapter 7.

For example, in Figure 4.9, it can be concluded that,

$$Cov(X,Y) \times Cov(Y,Z) \times Cov(X,Z) \geq 0.$$

This graph is only automatically identifiable and exhibits other constraints if H is

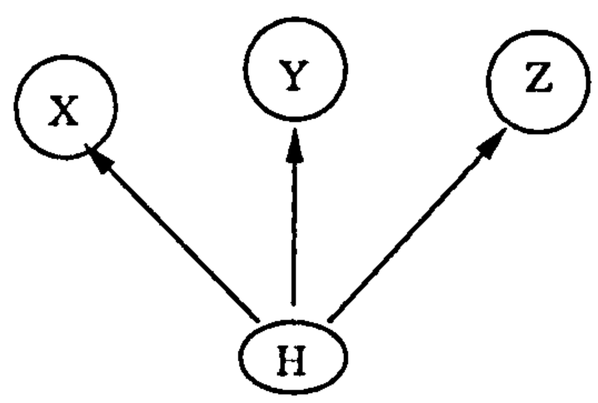


Figure 4.9: The network structure with a latent variable, H, and 3 observable variables, X, Y, and Z.

one dimensional random variable or two states (binary) discrete random variables. The identifiability of this sort of graphs were studied by Whiley and Titterington (2002) and reviewed in more details in Chapter 7.

I should emphasise that my objective in this thesis is not model selection. We actually study the estimation issues of the pattern that is already chosen from data, and from Bayesian perspective this model required some assumptions to enable a Bayesian to define prior distribution associated with the parameters of this model. In the next chapter, we first introduce the assumptions which are essential to define prior distributions of the parameters of a causal Bayesian network. In Chapter 6, we characterise these prior distributions.

Chapter 5

Hypercausality

5.1 Introduction

In this chapter we will make a connection between prior independence and causality. We also introduce the *Hypercausal Bayesian network* that asserts a set of factorisations of densities which are invariant to a class of "do" operations larger than those considered by Pearl. This requires us to develop the ideas of Koster (2000) and Lauritzen (2001) about randomised intervention. We will show that if a Bayesian network is assumed to be Hypercausal and we wish to learn about the probabilities defining it, then the prior distribution on the probabilities of the idle system[1] must exhibit local and global parameter independence.

[1]The Bayesian network without any intervention is called *idle system.*

5.2 Relationships Between Causality and Parameter Independence

Let the vector $X = (X_1, X_2, \ldots, X_k)$, of nodes of a Bayesian network G have its components X_i, $1 \leq i \leq k$, listed in an order compatible with G and their corresponding vectors of probabilities $\underline{\theta}_1, \ldots, \underline{\theta}_k$ compatibly with the partial order induced by the directed edges of the Bayesian network. Thus the parameters are listed: $\underline{\theta}_1, \ldots, \underline{\theta}_k$, where $\underline{\theta}_i = \{\theta_{i|pa_{i(l)}, 1 \leq l \leq m_i}\}$. The components of $\underline{\theta}_i$ are taken in some arbitrary but fixed order within the vector. From the familiar rules of probability, we can write the general prior distribution for our study as

$$p(\underline{\theta}) = \prod_{i=1}^{k} p(\underline{\theta}_i \mid \underline{\theta}^{i-1}), \quad \underline{\theta}^{i-1} = \{\underline{\theta}_1, \ldots, \underline{\theta}_{i-1}\}$$

Let $\underline{\theta}_A$ represent the subset of $\underline{\theta}$ whose indices are $i \in A$, A is a subset of $\{1, \ldots, k\}$. Here k is the total number of conditional probabilities needed to define G, or equivalently the number of components of $\underline{\theta}$. Let us consider each component of $\underline{\theta}_i$ as

$$\theta_{i(j)|pa_{i(l)}} = p(X_i = x_{i(j)} \mid pa_i = pa_{i(l)})$$

where $\theta_{i(j)|pa_{i(l)}}$ denotes the parameter associated with level j of the i^{th} variable and the level l of its parents. Thus,

$$\underline{\theta}_i = \{\theta_{i(j)|pa_{i(l)}},\ 1 \leq j \leq n_i,\ 1 \leq l \leq m_i\}$$

where each component of $\underline{\theta}_i$ is positive and for the fixed l (the fixed level of parent or for the components in the same strata), $\Sigma_{j=1}^{n_i} \theta_{i(j)|pa_{i(l)}} = 1$. Here n_i and m_i denote the numbers of the states of the i^{th} variable and its parents set, respectively.

The vector

$$\underline{\theta} = \{\theta_{i|pa_{i(l)}} : 1 \leq l \leq m_i, 1 \leq i \leq k\}$$

is said to exhibit *local* and *global independence* if

$$\underline{\theta}_{i|pa_{i(l)}} = (\theta_{i(1)|pa_{i(l)}}, \ldots, \theta_{i(n_i)|pa_{i(l)}})$$

are all mutually independent (Spiegelhalter and Lauritzen (1990)).

By the example below, we clarify the notations and concepts described above.

Example 5.1 Consider again the causal Bayesian network shown in Figure 5.1, where the mentioned categories of values (in Example 4.2) of maximum temperature of cooling system in an hour (T), i.e., $(0^0 - 100^0, 100^0 - 200^0, 200^0 - 300^0, 400^0+)$ can be coded to 0, 1, 2, 3, respectively. Similarly, the values of C, that is, normal, critical and meltdown are coded to 0, 1, 2 respectively. Finally, F is a binary variable that $F = 1$ indicates the failure of cooling system in an hour. Let us rename C, T and F by X_1, X_2 and X_3 respectively. Then, the parameters associated with F is given by $\underline{\theta}_3 = (\theta_{3(0)|0}, \ldots, \theta_{3(0)|3}, \theta_{3(1)|0}, \ldots, \theta_{3(1)|3})$ (such that $\Sigma_{i=0}^{1}\theta_{3(i)|l} = 1$), where, for example, $\theta_{3(1)|3} = p(F = 1 \mid T = 3)$ and so on. If $\underline{\theta}_1, \underline{\theta}_2, \underline{\theta}_3$ are mutually independent of each other, then we say the parameters associated with each node of the Bayesian network above are globally independent. Furthermore, if $(\theta_{3(0)|0}, \ldots, \theta_{3(0)|3}, \theta_{3(1)|0}, \ldots, \theta_{3(1)|3})$ are mutually independent of each other, then we say that the parameters associated with X_3 are locally independent.

5.2.1 Randomisation and Cause

Although Pearl focuses on deducing causal relationships from observational studies, traditionally *causal* effects have been more usually investigated using randomised trials in designed experiments. Thus, for example, to investigate the efficacy of a medical treatment A over an alternative treatment B, we would typically manipulate the system by randomising the allocation of the treatments. Similarly, in the example above, the failure of cooling system might be investigated by a series of randomised trials moni-

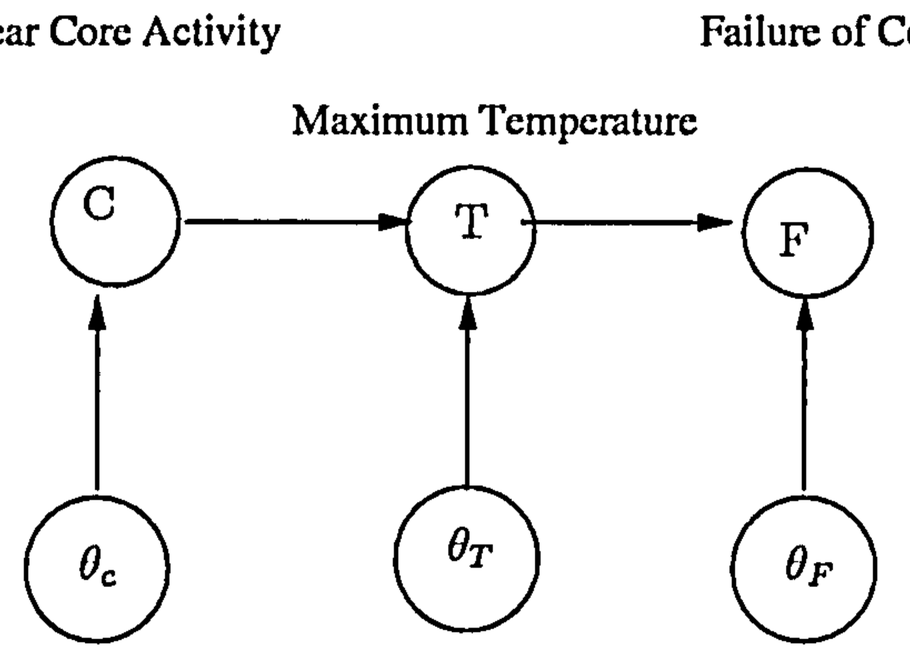


Figure 5.1: The Bayesian network for the Nuclear Activity' Example

toring failures within a range of temperatures. It is therefore natural to include such manipulations in any discussion of causation.

Definition 5.1 The contingent randomised intervention, $do(\underline{\theta}_A = \underline{\theta}^*_A)$ on a Bayesian network G whenever the contingent $pa_{i(l)}$ arises, manipulates X_i to a value $x_{i(j)}$ according to the set randomising probabilities

$$\theta^*_{i(j)|pa_{i(l)}} = p(X_i = \hat{x}_{i(j)} \mid pa_i = pa_{i(l)}) \quad \text{for each } i \in A.$$

When several interventions are employed simultaneously the effect of the manipulation is calculated in an order compatible with G.

A default choice for predicting the effect of a contingent manipulation conditional on the probability vector $\underline{\theta}$ is to use the following definition which extends, in an obvious way the definition of Spirtes et al (1993) and Pearl (2000).

Definition 5.2 A Bayesian network is said to be *Contingently Causal*, if under the contingent manipulation

$$do(X_i = \hat{x}_{i(j)} \mid pa_i = pa_{i(l)}) = do(\theta_{i(j)|pa_{i(l)}} = \theta^*_{i(j)|pa_{i(l)}}),$$

for all configurations of X consistent with $\{X_i = \hat{x}_{i(j)} \mid pa_i = pa_{i(l)}\}$ and $pa_{i(l)}$, the joint mass probability function after this manipulation of the other variables follow the formula,

$$p(\underline{x} \mid \underline{\theta}_{\hat{i}}, do(\theta_{i(j)|pa_{i(l)}} = \theta^*_{i(j)|pa_{i(l)}})) = \{\prod_{v=1,k,v\neq i} \underline{\theta}_v\} \times \theta^*_{i(j)|pa_{i(l)}}$$

For all configurations, $pa_{i(l)}$ not consistent with $\{X_i = \hat{x}_{i(j)} \mid pa_i = pa_{i(l)}\}$, we set

$$p(x \mid \underline{\theta}_{\hat{i}}, do(\theta_{i(j)|pa_{i(l)}} = \theta^*_{i(j)|pa_{i(l)}})) = 0.$$

Here we have let $\underline{\theta}_{\hat{i}}$ denote the $\underline{\theta}$ vector with the i^{th} component missing.

Clearly, if we plan to randomise over A using $\underline{\theta}^*_A$ then the randomisation should not influence other parameters in the system (See Daneshkhah and Smith (2003a)).

Definition 5.3 Call a contingent randomised intervention $do(\underline{\theta}_A = \underline{\theta}^*_A)$ on an uncertain Bayesian network[2], *Bayes faithful* if

$$p(\underline{\theta}_{\bar{A}} \mid do(\underline{\theta}_A = \underline{\theta}^*_A)) = p(\underline{\theta}_{\bar{A}})$$

where $\bar{A}$ stands for the complement of A in $\{1, \ldots, k\}$, and $p(\underline{\theta}_{\bar{A}})$ is the Bayesian's prior marginal density on $\underline{\theta}_{\bar{A}}$.

[2]If the conditional probabilities associated with each node for a given Bayesian network are unknown, this Bayesian network will then be called *uncertain*.

5.2.2 Hypercausality and Randomisation

We are now ready to define hypercausality. Effectively, this takes the *extended* Bayesian network -i.e., one that includes parameters in the Bayesian network as if they are random variables as in Figure 5.1-and demands causal consistency with this Bayesian network as well as the original Bayesian network. Let $A_u = \{1, \ldots, u\},\ 1 \leq u \leq k$. In terms of the constructions above we have the following definition.

Definition 5.4 Say an uncertain Bayesian network is a *Hypercausal Bayesian network* if it is a contingently causal Bayesian network and for all Bayes faithful contingent interventions $do(\underline{\theta}_{A_u} = \underline{\hat{\theta}}_{A_u})$, $1 \leq u \leq k$, $p(\underline{\theta}_{\bar{A}_u} \mid do(\underline{\theta}_{A_u} = \underline{\hat{\theta}}_{A_u})) = p(\underline{\theta}_{\bar{A}_u} \mid \underline{\hat{\theta}}_{A_u})$. Here $\bar{A}_u$ denotes the complement of A_u.

Pearl (2000) focused on the definition of intervention $do(X_i = \hat{x}_i)$ for Bayesian networks with known probabilities. This can be thought of as a degenerate form of contingent randomised intervention on $\hat{\theta}_i$ on a contingently causal Bayesian network as

$$\hat{\theta}_{i(j)|pa_{i(l)}} = \begin{cases} 1 & \text{if } x_{i(j)} = \hat{x}_i \text{ for all j and l} \\ 0 & \text{otherwise} \end{cases}$$

Koster (2000) has given a generalisation of this definition to the *Randomised Causal Bayesian Network*. Let us introduce his work by the following example that is chosen from Robins (1997).

Now, let us consider the following manipulation

$$\hat{\theta}_{i(j)|pa_{i(l)}} = \begin{cases} \theta_i^*(x_i) & \text{if } x_{i(j)} = \hat{x}_i \text{ for all j and l} \\ 0 & \text{otherwise} \end{cases}$$

We can easily examine that his intervention formula under contingently causal Bayesian network (and hence causal Bayesian network) and the Bayes faithfulness assumptions

coincides with ours, conditional on $\hat{\theta}_i$.

However there are examples (see Daneshkhah and Smith (2003a)) when we may want to use different randomisations for different configurations of parents of X_j and there is absolutely no reason within this framework not to extend his definition to include this case.

Before we can define an uncertain analogue of a causal Bayesian network, we first need to define unambiguously what is meant by "observing the probability θ_i". It is most natural to follow Pearl and Example 5.2 and to define this conditioning as on the 'perfect' estimate of a particular conditional probability in the Bayesian network obtained as a limiting proportion in an auxiliary sample from the same sample space.

So assume a selection mechanism, acting on a random sequence $\{X_{j(i)}[t]\}_{t\geq 1}$ of observations respecting the idle Bayesian network, which records $\{X_{j(i)}[t]\}_{t\geq 1}$ only for the m parent $pa_{j(i)}$ configuration values associated with indices $i \in A$, where m is the number of the components of A.

Call the selected subsequence $\{w_A(s)\}_{s\geq 1}$, where $w_A(s) = \{w_i(s_i) : i \in A\}$, where s_i indexes the s^{th} observation of the i^{th} variable. We can now define

$$p(\theta_{\bar{A}} \mid \theta_A) = \lim_{r\to\infty} p(\theta_{\bar{A}} \mid \{w_A(s)\}_{1\leq s\leq r})$$

where $inf\, s_i \to \infty$ as $r \to \infty$.

The proof of this statement can be obtained directly, but in a tedious way, from Borel-Cantelli lemmas. Therefore, this will give us the standard formula for the conditional probability $p(\theta_{\bar{A}} \mid \theta_A)$, from the Bayesian viewpoint.

This rather technical definition above allows us to examine more closely what assumptions on hypercausal Bayesian network model encodes. Consider the following example.

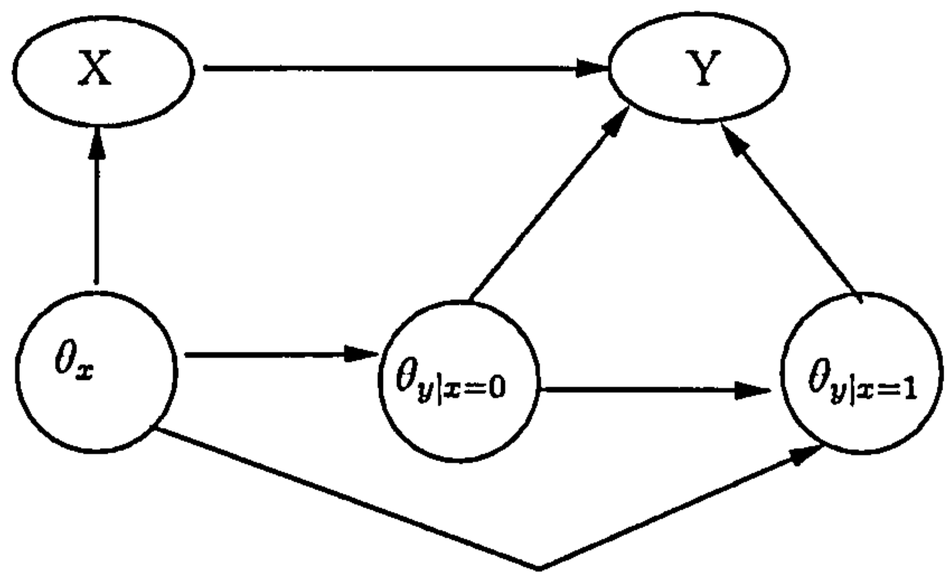


Figure 5.2: The Bayesian network with the dependent parameters of Example 5.2.

Example 5.2 Consider the extended Bayesian network shown in Figure 5.2 with two binary random variables X and Y, where the parameters θ_x , $\theta_{y|x=1}$ and $\theta_{y|x=0}$ are dependent. This dependency implies, for example, that if we learn the value of θ_x through, for example, observing an enormous sample of units where we record the X-margin only - then our assessment of the probability $\theta_{y|x=1} = p(Y = 1 \mid X = 1)$ will, in general, change. This might occur if, for example, a higher than expected value for θ_x is associated with a 'bad scenario' which we believe would lead us to expect $\theta_{y|x=1}$ would be larger than expected as well.

On the other hand, if we were to manipulate that system to take a randomised sample on X so that $p(X = 1) = \theta_x$, then by definition this *randomisation* should leave $\theta_{y|x=1}$ unchanged in the idle system. So if there is a prior dependence between θ_x and $\theta_{y|x=1}$, then we should not expect to be able to identify a randomly manipulated system from one learnt from an idle system. This model would then not be an hypercausal

Bayesian network. This argument can be extended to all combinations of values of these parameters, and suggests a close relationship between the hypercausal Bayesian network model assumption and local and global independence.

Note that the reason we needed to introduce contingent randomisation was to be able to consider the separate manipulation of all components $\underline{\theta}_i$ of the probability vector $\underline{\theta}$ to values other than zero and one. In this way we are able to perform all necessary manipulations of each of the components of the vector of probabilities in the extended Bayesian network. This heuristic argument motivates the following theorem.

Theorem 5.1 G is an hypercausal Bayesian network if and only if it exhibits local and global independence.

Proof By the definition of Bayes faithfulness on this contingently causal Bayesian network

$$p(\underline{\theta}_{\bar{A}_u} \mid do(\underline{\theta}_{A_u} = \hat{\underline{\theta}}_{A_u})) = p(\underline{\theta}_{A_u})$$

so by the definition of an hypercausal Bayesian network, equivalently, the probabilities in the idle system satisfy

$$p(\underline{\theta}_{\bar{A}_u}) = p(\underline{\theta}_{\bar{A}_u} \mid \hat{\underline{\theta}}_{A_u}), \quad 1 \leq u \leq k$$

Hence by definition of A_u $\Longleftrightarrow$

$$\underline{\theta}_k \perp\!\!\!\perp \underline{\theta}_1, \ldots, \underline{\theta}_{u-1}, \quad 1 \leq u \leq k-1 \iff \perp\!\!\!\perp_{i=1}^{k} \underline{\theta}_i, \tag{5.1}$$

$\Longleftrightarrow$ G exhibits local and global independence.

Corollary 5.1 An hypercausal Bayesian network exhibits the property that for all Bayes faithful contingent interventions on $\underline{\theta}_A, A \subseteq \{1, \ldots, k\}$,

$$p(\underline{\theta}_{\bar{A}} \mid do(\underline{\theta}_A = \hat{\underline{\theta}}_A)) = p(\underline{\theta}_{\bar{A}} \mid \hat{\underline{\theta}}_A))$$

Proof This follows trivially from equation (5.1).

When estimating models, local and global independence is assumed almost universally- see e.g., Geiger and Heckerman (1997), Cowell et al (1999) and Cooper and Yoo (1999). In so doing these implicitly assume such models are at least consistent with hypercausal Byaesian networks. The causal interpretation above is therefore one way of checking whether this assumption is appropriate in a practical situation. For example, if the root node X in the Bayesian network of Figure 5.1 concerned the failure of a component, I could ask myself whether taking a very large sample of such $X's$ in operational mode and observing their failure rate θ_x could be expected to affect the system as a whole in the same way as if we simulated simply by randomising to failure with probability θ_x artificially. Only if the answer was 'yes' should I proceed with this assumption.

It is interesting to note that whenever we move between actual and simulated scenarios we almost always implicitly make hypercausal assumptions.

5.3 Discussion

Because of the pioneering works of Robins (1986) and Spirtes et al (1993), recent studies of causality have mainly focussed on how to deduce causal relationships from observational studies. But, causal assertions have traditionally been studied through manipulating treatments in randomised trials (see Smith (2002)). Components of a hypercausal Bayesian network model are therefore usually most strongly supported by the sort of

randomised contingent experiments we discuss in this thesis. From the subjectivist viewpoint, we contend that the bold part of the hypercausal hypothesis is that the Bayesian asserts that the process in the field will behave in the same way as it did under (analogous) carefully controlled randomised trials (not vice versa). This hypothesis will usually fail when, extraneous dependencies get introduced through not being able to control conditions in the field. Hypercausal priors (which demand identity of these two cases) allows us to think about this problem the right way round.

When data is not exhaustive it is common for idle systems to exhibit extraneous dependencies not linked with science but with paucity of information. These dependencies in the idle system can be induced by misspecified priors, trial sample data sets on non-ancestral subsets of variables, selection variables and so on, quite spurious for any assessment of causality, but intrinsic for learning about the idle system.

In our opinion it is these issues which make causal deductions from uncertain idle system so prone to mislead.

Chapter 6

Essential Graphs and Multicausality

6.1 Introduction

In this section, we will review the *essential* graph that is introduced by Anderson et al (1997) to characterise the equivalence class of the Bayesian networks. Then, we introduce the *multicausal* essential graph on the equivalence class of Bayesian networks where each Bayesian network exhibits a strong form of manipulation causality called hypercausality.

6.2 Equivalent Bayesian Networks and Essential Graph

Two Bayesian networks (or more generally DAGs) are called Markov *equivalent* if they assert the same set of conditional independencies (two network structures are equivalent if the set of distribution that can be represented using one of the structures are equivalent to the set of distributions that can be represented using the other) assumptions among the variables in the domain. To make this concept clear, let us consider the Bayesian

network with three variables that is shown in Figure 6.1.

According to the (causal) Markov condition[1], the conditional independent constraint

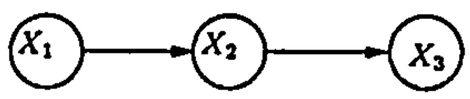
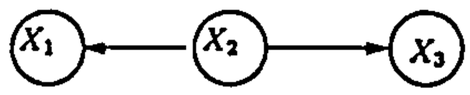


Figure 6.1: In the Bayesian network above, $X_1 \perp\!\!\!\perp X_3 \mid X_2$.

for the Bayesian network above is encoded by X_1 is independent of X_3 given X_2. There are two more different Bayesian networks over the same variable set and the same conditional independent restriction as follows



and

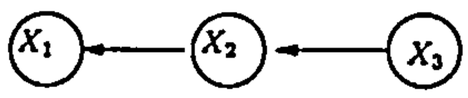


Figure 6.2: Two Bayesian networks with the different structures but the same conditional independence statement.

Therefore, we can conclude that the sets of discrete probability distributions that are Markov over the mentioned Bayesian networks above are actually similar together. Thus, these Bayesian networks are said to be *Markov equivalent*, or simply, *equivalent*.

[1]Each variable is independent of its non-descendants, given its parents.

But, the following Bayesian network is not equivalent with the Bayesian networks above.

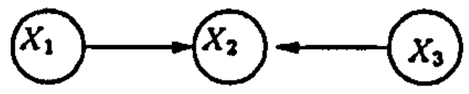


Figure 6.3: The Bayesian network with the different conditional independent restriction, $X_1 \perp\!\!\!\perp X_3$, that is not equivalent with the Bayesian networks mentioned above.

Because, the conditional constraint for this Bayesian network is: X_1 is marginally independent of X_3.

Theorem 6.1 Two Bayesian networks are equivalent if and only if they have the same *skeletons* and the same *v-structures.*

A consequence of the theorem above is that for any edge involved in a v-structure in some Bayesian network G, if that edge is reversed in some other Bayesian network G', then G and G' are not equivalent.

As we discussed in Section 3.3, Theorem 6.1 does not directly provide a characterisation of the entire equivalence class $[G]$ for the given Bayesian network. Furthermore, since the number of possible orientations of all arrows that do not involve in any v-structure (immorality) of a Bayesian network G grows exponentially with the number of such arrows. Hence, determination of the equivalence class $[G]$ by exhaustive enumeration of possibilities, rapidly becomes computationally infeasible as the size of G increases. Anderson et al (1997) introduced an alternative approach based on the graph called

essential graph. The essential graph G^* associated with a Bayesian network G is given in the following definition, introduced by Anderson et al (1997).

Definition 6.1 (Anderson et al (1997)) The essential graph G^* associated with G is the graph

$$G^* := \cup(G' \mid G' \approx G),$$

that is, G^* is the smallest graph containing every $G' \in [G]$.

It is obvious that the essential graph G^* associated with G is a graph with the same skeleton as G, but where an edge is directed (compelled) in G^* if and only if it occurs as a directed edge with the same orientation in every $G' \in [G]$. Clearly, all other edges of G^* are undirected and together with the directed edges in G^* are called the essential graph. Note that, every arrow that participates in Pearl's pattern is essential, but G may contains others essential arrows as well (see Anderson et al (1997) for details and examples).

In fact, an essential graph is a chain graph with additional characterising properties. These properties have been investigated by several authors (Chickering (1995) and Anderson et al (1997)). Anderson et al (1997) formalised the essential graph characterisation in the following theorem:

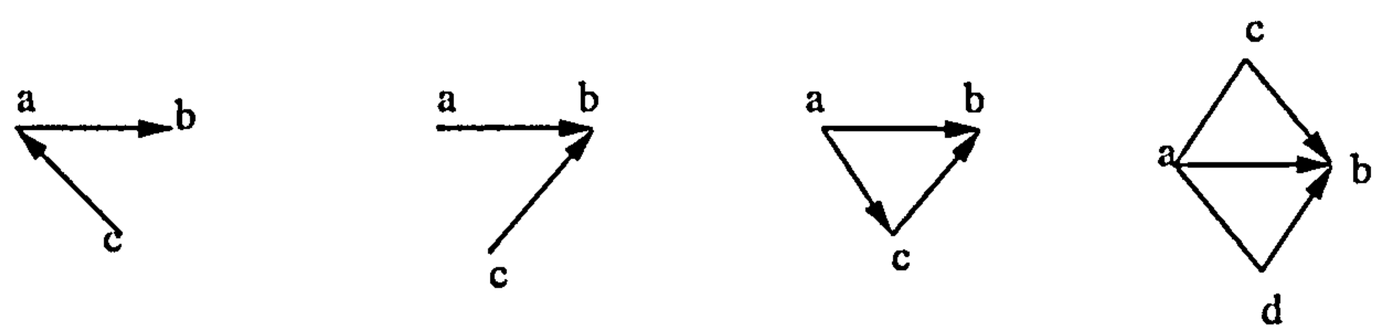
Theorem 6.2 (Essential Graph Characterisation) A graph $G = (V, E)$ is the essential graph for some Bayesian network D with vertex set V if and only if G satisfies the following four conditions:

1. G is a chain graph.

2. For each chain component $\tau \in \mathcal{T}(G)$, the undirected graph G_τ is chordal.

3. G has no induced subgraph of the form $a \rightarrow b - c$.

4. each arrow $a \rightarrow b$ in G is strongly protected, that is, it occurs in at least one of the following configurations as an induced subgraph of G:



The notation behind the last theorem is that there are directed edges which remain in the same orientation throughout all the Bayesian networks that form an equivalence class. It is said that these directed edges are essential. The characterisation mentioned in Theorem 6.2 allows us to devise and develop a polynomial-time algorithm to convert a Bayesian network into an essential graph and vice versa (see Anderson et al (1997) and Chickering (1995)). For example, the chain graph shown in Figure 6.4 is not an essential graph because the subset of vertices $\{4, 5, 6\}$ makes a subgraph that violates condition 3 of the last theorem.

There are some new works to characterise the essential graph. For example, Studeny (2002) presented an alternative characterisation of essential graphs. His characterisation is based on a special operation of legal merging of components and leads to an algorithm

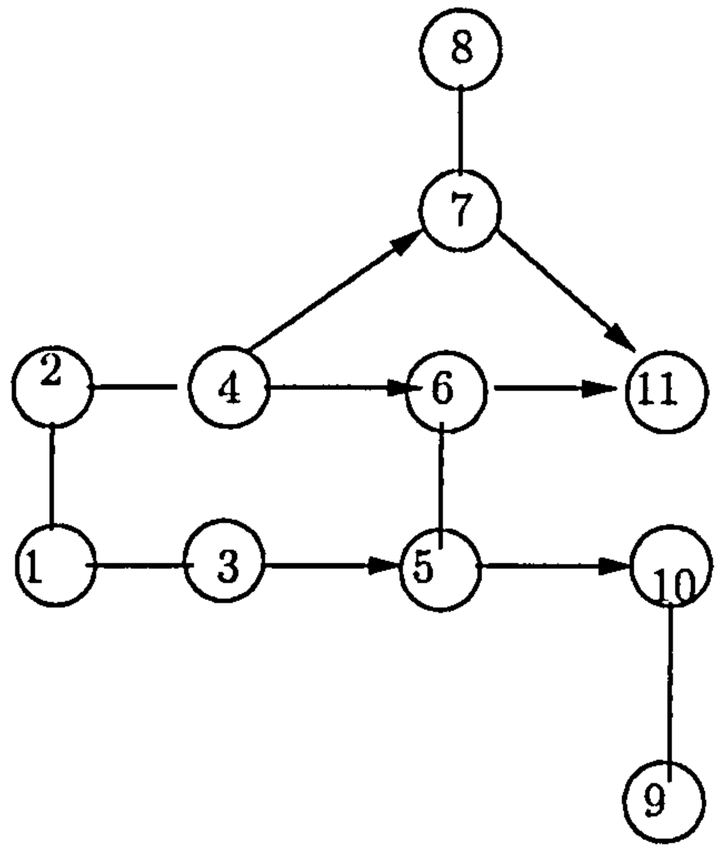


Figure 6.4: A chain graph that is not essential.

for converting a Bayesian network into the respective essential graph. He claims his algorithm avoids indicating essential (compelled) arrows. However, all the work that we develop in this chapter is based on the Anderson et al (1997) characterisation.

6.2.1 The Multicausal Essential Graph

As discussed above two Bayesian networks are Markov equivalent if and only if their mixed pattern (Verma and Pearl (1990, 1992)) or their mixed essential graphs (Anderson et al (1997)) agree. It follows that Bayesian networks with the same essential graphs will be indistinguishable from each other and from an observational study of an idle system. Some of the earliest algorithms for deducing causality from Bayesian networks (e.g., Verma and Pearl (1990, 1992) and Chickering (1995)) were based on the configurations of directed edges in essential graphs empirically fitted to exhaustive data sets from

a cross sectional experiment. The directed edges of these mixed graphs deduced from cross sectional data allowed them to make causal assertions about what might happen when the system was manipulated.

Now suppose, on the basis of observations of an analogous idle system to the one under study, a Bayesian is confident in asserting a particular essential graph H as a valid hypothesis for another analogous unmanipulated system. The components of conditional probabilities of this new system are, however, uncertain and the researcher wants to make prior assumptions which are consistent with *every* hypercausal Bayesian network consistent with the given essential graph. How can this be achieved?

In this part we define a concept of multicausality which asserts hypercausality for all Bayesian networks in the equivalence class of an essential graph. We show that the prior density on these probability parameters must satisfy a generalisation of the Geiger and Heckerman condition (see Theorem 2 in Geiger and Heckerman (1997)). In the special case when the essential graph is undirected, this family degenerates into the Hyper-Dirichlet family (see Dawid and Lauritzen (1993)). We conclude the section by discussing the interpretation and implication of using priors of this form in Bayesian networks.

Now, let us define the multicausal essential graph as follows,

Definition 6.2 An uncertain essential graph, $\mathcal{P}$ is called *multicausal* if the prior distribution on the uncertain probabilities of every Bayesian network in the equivalence class corresponding to $\mathcal{P}$, ($[\mathcal{P}]$), is consistent with an Hypercausal Bayesian network.

An important subclass of essential graphs is one where all its components are undirected. This class is called the *decomposable* equivalence class. It is natural to ask how priors might be set up on the uncertain probabilities in the system in a way which is invariant to equivalent Bayesian networks. In particular, what is the family of priors which exhibit local and global independence for every Bayesian network compatible with a given essential graph? In 1993, Dawid and Lauritzen proved that, for decomposable essential graphs, the Hyper-Dirichlet family of distributions preserved global independence. Later, Geiger and Heckerman (1997) proved a much stronger result for a two node essential graph. They demonstrated that the two Bayesian networks in this equivalence class both exhibited local and global independence if and only if the joint prior distribution on its two nodes was Dirichlet.

Explicitly, let $\{\psi_{ij}, 1 \leq i \leq k, 1 \leq j \leq n\}$, be positive random variables that sum to unity and denote the multinomial parameters in the two way probability table on the two equivalent Bayesian networks with two nodes and one edge. Denote the Dirichlet, $\mathcal{D}(\alpha)$, density on $\underline{\theta}$ by

$$p(\underline{\theta}) = \frac{\Gamma(\alpha)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \prod_{i=1}^{k} \theta_i^{\alpha_i - 1}$$

where $\alpha = \Sigma_{i=1}^{k} \alpha_i$.

Geiger and Heckerman (1997) proved that for the priors on both Bayesian networks to exhibit local and global independence, it is necessary and sufficient that, using the obvious notation, the prior distributions on ψ_{ij}, $\theta_{i.}$, $\theta_{j|i}$, $\theta_{.j}$ and $\theta_{i|j}$ are all Dirichlet. Furthermore the parameters of their respective densities $\mathcal{D}(\alpha_{ij})$, $\mathcal{D}(\alpha_{1.}, \ldots, \alpha_{k.})$, $\mathcal{D}(\alpha_{j|1}, \ldots, \alpha_{j|k})$, $\mathcal{D}(\alpha_{.1}, \ldots, \alpha_{.n})$ and $\mathcal{D}(\alpha_{i|1}, \ldots, \alpha_{i|n})$, would need to satisfy the further linear equations, $\alpha_{i.} = \Sigma_{j=1}^{n} \alpha_{ij}$, $\alpha_{.j} = \Sigma_{i=1}^{k} \alpha_{ij}$ and $\alpha_{i|j} = \alpha_{j|i} = \alpha_{ij}$.

Let us make this Dirichlet characterisation clear by a simple example on the two-binary-variable domain. We may assert that the database is a multinomial sample from the joint space $\underline{X} = \{X, Y\}$ with parameters $\underline{\theta} = \{\theta_{xy}, \theta_{x\bar{y}}, \theta_{xy}, \theta_{xy}\}$, where $\theta_{xy} = p(X = x, Y = y \mid \underline{\theta})$. We can consider two equivalent Bayesian networks as follows on these two variables:

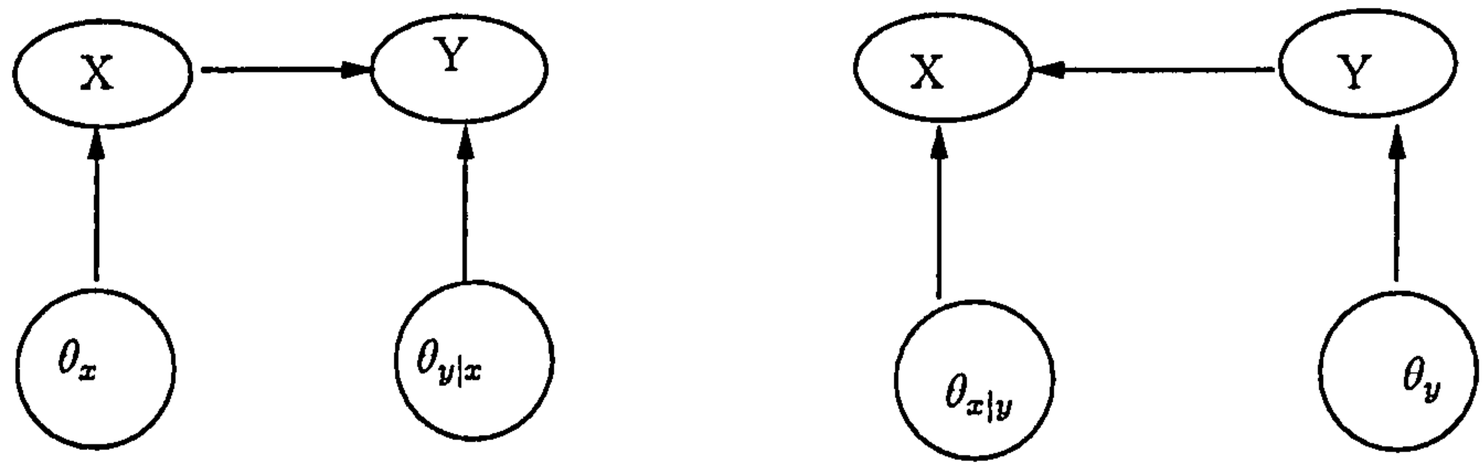


Figure 6.5: Two equivalent Bayesian network structures for a two-binary-variable domain

Given parameter independence[2], parameter modularity[3], likelihood equivalence [4], it turns out that we can compute the prior distribution for any network structure men-

[2]For each network structure the parameters associated with one node are independent of the parameters associated with other nodes (It is called *global independence*). If the parameters associated within a node given one instance of its parents are independent of the parameters of that node given other instances of its parent nodes, then we say parameters associated within that node are locally independent

[3]If a node has the same parents in two distinct networks structures, then the distribution of the parameters associated with this node are identical in both structures

[4]This assumption is obvious, because we can write $\theta_{xy} = \theta_x \theta_{y|x} = \theta_y \theta_{x|y}$.

tioned above from the given prior distribution on $\underline{\theta}$.

Now suppose we are given a density for the parameters of the joint space $p_{x\rightarrow y}(\theta_{xy}, \theta_{\bar{x}y}, \theta_{x\bar{y}})$. From this density, we can construct the parameter densities for the equivalent Bayesian networks mentioned above (shown in Figure 6.5). The parameters associated with the network structure $x \rightarrow y$, $(\theta_x, \theta_{y|x}, \theta_{y|\bar{x}})$ are related to the parameters of the joint space by the following relations:

$$\theta_{xy} = \theta_x\theta_{y|x} \quad \theta_{\bar{x}y} = (1-\theta_x)\theta_{y|\bar{x}} \quad \theta_{x\bar{y}} = \theta_x(1-\theta_{y|x})$$

Thus, we may obtain $p_{x\rightarrow y}(\theta_x, \theta_{y|x}, \theta_{y|\bar{x}})$ from the given density by the following equation,

$$p_{x\rightarrow y}(\theta_x, \theta_{y|x}, \theta_{y|\bar{x}}) = J_{x\rightarrow y}p_{x\rightarrow y}(\theta_{xy}, \theta_{\bar{x}y}, \theta_{x\bar{y}}) \tag{6.1}$$

where $J_{x\rightarrow y} = \theta_x(1-\theta_x)$ is the Jacobian of the transformation (see Geiger and Heckerman (1997) and Heckerman et al (1995) for more details).

By likelihood equivalence, that is, $p_{x\rightarrow y}(\theta_{xy}, \theta_{\bar{x}y}, \theta_{x\bar{y}}) = p_{x\leftarrow y}(\theta_{xy}, \theta_{\bar{x}y}, \theta_{x\bar{y}})$, we can similarly compute the prior distribution for the network structure $x \leftarrow y$ using the Jacobian $J_{x\leftarrow y} = \theta_y(1-\theta_y)$ from the joint prior distribution, $p_{x\leftarrow y}(\theta_{xy}, \theta_{\bar{x}y}, \theta_{x\bar{y}})$, or from the prior density on the network structure $x \rightarrow y$ as follows,

$$p_{x\leftarrow y}(\theta_y, \theta_{x|y}, \theta_{x|\bar{y}}) = \frac{J_{x\leftarrow y}}{J_{x\rightarrow y}}p_{x\rightarrow y}(\theta_x, \theta_{y|x}, \theta_{y|\bar{x}}).$$

If $p(\theta_{xy}, \theta_{x\bar{y}}, \theta_{\bar{x}y}) = \mathcal{D}(\alpha_1, \alpha_2, \alpha_3, \alpha_4)$, then we can obtain that

$$p_{x\rightarrow y}(\theta_x, \theta_{y|x}, \theta_{y|\bar{x}}) = Beta(\alpha_1+\alpha_2, \alpha_3+\alpha_4) \times Beta(\alpha_1, \alpha_2) \times Beta(\alpha_3, \alpha_4)$$

and

$$p_{x\leftarrow y}(\theta_y, \theta_{x|y}, \theta_{x|\bar{y}}) = Beta(\alpha_1+\alpha_3, \alpha_2+\alpha_4) \times Beta(\alpha_1, \alpha_3) \times Beta(\alpha_2, \alpha_4)$$

as it is discussed in the general case above.

We have seen above that there is a direct link between local and global independence and hypercausal hypotheses. Such classes of local and global independence statements must therefore correspond to several simultaneous causal hypotheses made by the Bayesian within the equivalence class of Bayesian networks in the posited essential graph.

Although the Geiger and Heckerman condition above is only valid for graphs with two variables, it is straightforward, using induction, to extend the Geiger and Heckerman condition to characterise complete (and hence undirected) essential graphs H, all of whose Bayesian networks exhibit local and global independence: see the following lemma.

Lemma 6.1 If the prior distributions of all Bayesian networks consistent with a complete essential graph H with n variables $(X_1, \ldots, X_n)$ exhibit local and global independence, then the prior density on the joint probabilities $\underline{\theta}^{(n)}$ of $(X_1, \ldots, X_n)$ in H must be Dirichlet, $n \geq 2$.

Proof Go by induction on the number k of variables in H. The assertion is clearly true for $k = 2$ by the Geiger and Heckerman condition. Suppose it is true for $k = n - 1$ and write $X^{(r)} = \{X_1, \ldots, X_r\}$, $2 \leq r \leq n$. The essential graph H' of Figure 6.6 is valid, where, by the inductive hypothesis, the prior density on the joint probabilities $\underline{\theta}^{(n-1)}$ of $X^{(n-1)}$ is Dirichlet. Since the Bayesian network which introduces X_n first and the one that introduces X_n last both exhibit local and global independence, it follows that the two Bayesian networks associated with H' also exhibit local and global independence in their probabilities. The Geiger and Heckerman result now allows us to assert that $\underline{\theta}_n$ is Dirichlet. The well known properties of the Dirichlet now allow us to complete the inductive step.

Figure 6.6: The essential graph H'

Example 6.1 There are several Bayesian networks consistent with the following essential graph that is shown in Figure 6.7. The edge between X_1 and X_2 could be in either directed, and there are 6 configurations of directions of arrows on the triangle of nodes (X_5, X_6, X_7). So this essential graph has an associated equivalence class of 12 Bayesian networks. For example, let us consider two Bayesian networks associated with undirected edge between nodes X_5 and X_6 regardless of the direction of other undirected edges in this essential graph. Since we assume multicausality, these two Bayesian networks (one with edge $X_5 \longrightarrow X_6$ and another one with edge $X_5 \longleftarrow X_6$) exhibit local and global independence. We now note that if we condition on any value of X_4 (the shared parent of X_5 and X_6), the conditional prior distributions on X_5 and X_6 exhibit local and global independence whichever way around we condition X_5 and X_6. Therefore, the Geiger and Heckerman condition must hold on X_5 and X_6 conditional on each value of X_4. Indeed if we have this condition then it is easily checked that with local and global independence on other nodes we must have multicausality.

We can claim the same thing on the Bayesian networks that are consistent with the essential graph shown in Figure 6.7: one with edge $X_1 \longrightarrow X_2$ and another one with $X_1 \longleftarrow X_2$ (without any shared parents).

This motivates the following lemma and theorem.

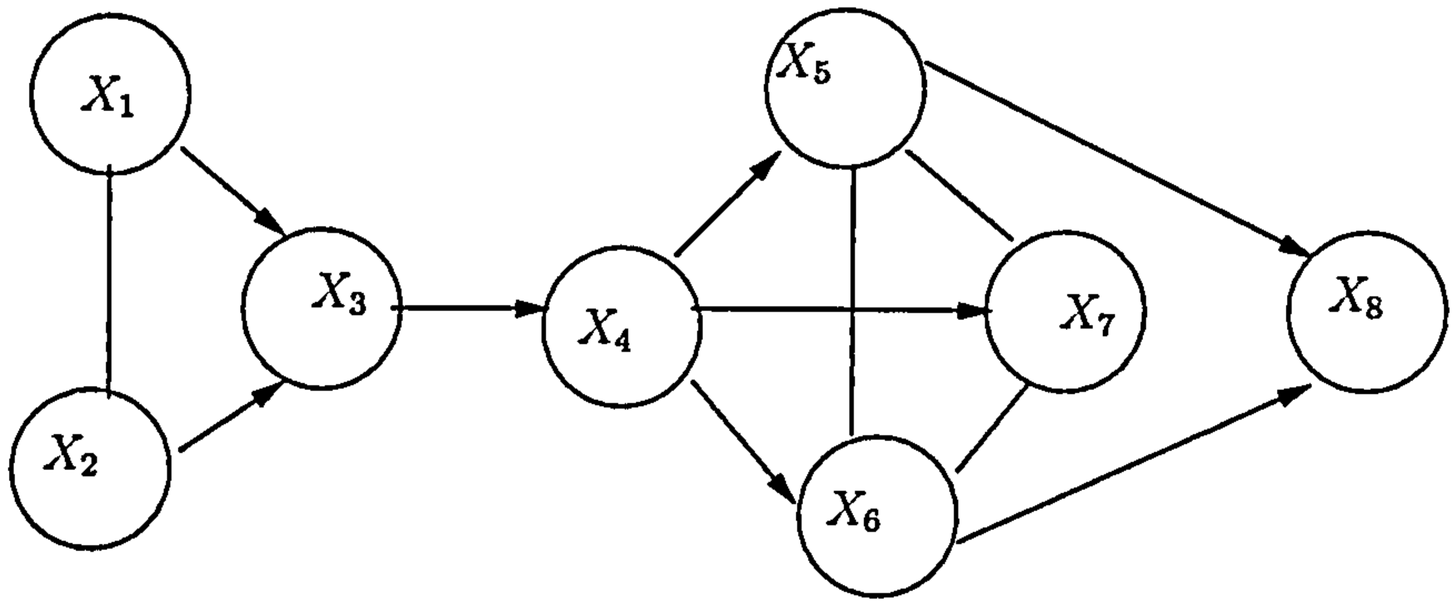


Figure 6.7: The multicausal essential graph of Example 6.1.

Lemma 6.2 The nodes of an undirected component subgraph of an essential graph H all share the same (directed) parents in H.

Proof This is a direct consequences of Lemma (1) in Chickering (1995).

Definition 6.3 The *undirected cliques* of an essential graph H are the maximally connected subsets of nodes/variables of H, all connected to one another by undirected edges.

Theorem 6.3 let H be an essential graph. Then H is multicausal if and only if, for any particular Bayesian network G in the equivalence class defined by H,

(i) The probability vector $\underline{\theta}$ of G exhibits local and global independence.

(ii) The densities of the joint probabilities of the undirected cliques of H are Dirichlet conditional on each value of their (shared and directed) parent configurations in H.

Proof Suppose H is multicausal. Condition (i) is implied by the definition. Choose an undirected clique C in H and condition on a value of its (directed) parent set-which by Lemma 6.2 is shared by nodes in C. Given this value of the parents, by the definition of H, the complete essential graph on the nodes in C is valid, and the probabilities on these nodes exhibit local and global independence for all Bayesian networks consistent with it. Condition (ii) now follows directly from Lemma 6.1.

On the other hand, suppose H satisfies condition (i) and (ii) and let G' be any Bayesian network associated with H. According to Lemma 3.2 in Anderson et al (1997), for any two Bayesian networks G and G' associated with H, there exists a finite sequence $G = D_1, \ldots, D_n = G'$ of Bayesian networks, all in the equivalence class described by H, and where D_i and D_{i+1}, $1 \leq i \leq n-1$, differ in exactly one edge direction. It is therefore sufficient to prove that if G exhibits local and global independence and condition (ii) is met then G', differing by only one edge, also exhibits local and global independence. So assume that G' is obtained from G by reversing the edge (X_i, X_j) in G. Since both G and G' are associated with H, (X_i, X_j) must be an undirected edge in H and so lie in one of its undirected cliques. By Lemma 6.2, (X_i, X_j) share (directed) parents. Note the elementary property that if the joint probabilities of $X \in C$ have a Dirichlet density then the joint probabilities of $X \in C_1 \subseteq C$ also have a Dirichlet density. So by (ii) conditional on any value of their shared parents the density of the joint probabilities of (X_i, X_j) is Dirichlet. It follows from the Geiger and Heckerman condition that G' must also exhibit local and global independence. This completes the theorem.

Suppose that all the variables $\{X_v\}_{v\in V}$ are discrete-valued, that is, they take values in finite sets $\{\Omega_v\}_{v\in V}$. The model $M(\mathcal{G})$ then considered for this set of variables is a decomposable graph[5] (See Dawid and Lauritaen (1993)). Let Ω denote the set of possible configurations of $\underline{X} = \{X_v;\ v \in V\}$:

$$\Omega = \prod_{v\in V} \Omega_v$$

Then, an arbitrary distribution for $\underline{\theta}$ in $M(\mathcal{G})$ is determined by the clique marginal probability tables $\theta_{\mathcal{C}} = \{\theta_C;\ C \in \mathcal{C}\}$ as

$$\theta(i) = \frac{\prod_{C\in\mathcal{C}} \theta_C(i_C)}{\prod_{S\in\mathcal{S}} \theta_S(i_S)}, \qquad i \in \Omega$$

where $\mathcal{C}$ is the set of cliques of $\mathcal{G}$ and $\mathcal{S}$ is the system of separators in a perfect ordering of the cliques. Note that the same set S may appear several times in the expression. For $S = C \cap D$ where C and D are cliques, θ_S can be calculated by marginalisation either from θ_C or from θ_D.

For each clique $C \in \mathcal{C}$, let

$$\alpha_C = \{\alpha_C(i_C);\ \ i_C \in \Omega_C\}$$

be a given table of arbitrary positive numbers and let $\mathcal{D}(\alpha_C)$ denote the Dirichlet distribution for θ_C with the following density

$$\pi(\theta_C \mid \alpha_C) \propto \prod_{i_C\in\Omega_C} \theta_C(i_C)^{\alpha_C(i_C)-1}$$

on the set where $\Sigma_{i_C}\theta_C(i_C) = 1$ and $\alpha_C(i_C) > 0$.

Now let us suppose that the collection of specifications $\mathcal{D}(\alpha_C),\ C \in \mathcal{C}$ are constructed in such a way that for any two cliques C and D in $\mathcal{C}$ we have:

$$\alpha_C(i_{C\cap D}) = \alpha_D(i_{C\cap D}) \tag{6.2}$$

[5]It should be noticed that the essential graph associated with this graph is undirected.

that is, if the cliques C and D overlap, then the parameters α_C and α_D are such that each implies the same marginal distribution for $\theta_{C\cap D}$.

Definition 6.4 For the perfect ordering of cliques $\alpha_{\mathcal{C}} = \{\alpha_C\}_{C\in\mathcal{C}}$ there exists a unique Dirichlet distribution for $\underline{\theta} = \{\theta_C\}_{C\in\mathcal{C}}$, that is called Hyper Dirichlet and denoted by $\mathcal{HD}(\alpha_{\mathcal{C}})$, which is hyper Markov over $M(\mathcal{G})$ and the distribution on each clique $C \in \mathcal{C}$ is $\mathcal{D}(\alpha_C)$.

Note that, it has been shown in Theorem 3.9 in Dawid and Lauritzen (1993) that there exists a unique "Hyper-Dirichlet" distribution for $\underline{\theta}$ over $M(\mathcal{G})$ such that θ_C has the marginal density $\mathcal{D}(\alpha_C)$ for all $C \in \mathcal{C}$ [6].

We can now state a corollary of Theorem 6.2.

Corollary 6.1 If H is undirected then it is multicausal if and only if the clique margins have a Hyper-Dirichlet distribution.

Proof If H is undirected then the Bayesian network is decomposable. Because all the clique margins are (consistently) Dirichlet, the result now follows directly from the definition of the Hyper-Dirichlet distribution above (for further details, see Daneshkhah and Smith (2003b)).

[6] In practice, one would construct a hyper-Dirichlet distribution by first identifying a perfect ordering of the cliques, for example, $\{C_1, \ldots, C_n\}$. Place a Dirichlet distribution $\mathcal{D}(\alpha_{C_1})$ on θ_{C_1}, next place a Dirichlet distribution $\mathcal{D}(\alpha_{C_2})$ on θ_{C_2}, with parameters constructed by Equation (6.2) and realizations constrained so that $\theta_{C_1\cap C_2}$ is identical for θ_{C_1} and θ_{C_2}. For each subsequent clique C_i, place a Dirichlet distribution on θ_{C_i} such that the parameters and the realizations of that distribution are consistent with those specified for the previous cliques.

Moving to multicausal models, we have established that if a Bayesian is prepared to make bold enough causal assertions within a single uncertain Bayesian network then this not only introduces independence relationships between parameters, but can also characterise prior families of distributions on these parameters. We believe this is a very helpful way of thinking about this class of models. Note that this multicausal essential graph characterisation concerns a single hypothesised model. It is not an assertion about a common prior to be used for causal Bayesian network for the model selection as is more typical in, for example see Geiger and Heckerman (1997), Cowell et al (1999) and Cooper and Yoo (1999) and references therein.

6.3 Discussion

How strong an assumption is multicausality, in learnt Bayesian networks? Although this assumption is almost universally made in practice, we would argue that this corresponds to a very unusual circumstance and demands very specific structures on prior information before it is valid. To illustrate this, consider supplementing the Bayesian network shown in Figure 6.5 with experimental evidence observing Y after we have randomised on X and conditioned on this value. The essential graph is $X - Y$ so that the hyperessential prior is just a Hyper-Dirichlet prior which sets the joint density of $\psi = (\psi_{00}, \psi_{01}, \psi_{10}, \psi_{11})$, where $\psi_{ij} = p(X = i, Y = j), \quad i, j = 0, 1$, having a Dirichlet density $p(\psi \mid \alpha)$ given by

$$p(\psi \mid \alpha) = K\psi_{00}^{\alpha_{00}-1}\psi_{01}^{\alpha_{01}-1}\psi_{10}^{\alpha_{10}-1}\psi_{11}^{\alpha_{11}-1}$$

where K denotes to the normalising constant. Furthermore, θ_x and $\underline{\theta}_{y|x}$ are distributed respectively as

$$p(\theta_x) = K_1(1-\theta_x)^{\alpha_{0\cdot}-1}\theta_x^{\alpha_{1\cdot}-1}$$

and

$$p(\underline{\theta}_{y|x}) = K_2 \theta_{y|x}^{\alpha_{11}-1} \theta_{y|\bar{x}}^{\alpha_{01}-1} \theta_{\bar{y}|x}^{\alpha_{10}-1} \theta_{\bar{y}|\bar{x}}^{\alpha_{00}-1}$$

similarly, we have

$$p(\theta_y) = K_3 (1-\theta_y)^{\alpha_{.0}-1} \theta_y^{\alpha_{.1}-1}$$

and

$$p(\underline{\theta}_{x|y}) = K_4 \theta_{x|y}^{\alpha_{11}-1} \theta_{x|\bar{y}}^{\alpha_{10}-1} \theta_{\bar{x}|y}^{\alpha_{01}-1} \theta_{\bar{x}|\bar{y}}^{\alpha_{00}-1}$$

where $\alpha_{i.} = \alpha_{i0} + \alpha_{i1}$, $\alpha_{.j} = \alpha_{0j} + \alpha_{1j}$, $i,j = 0,1$, K_l, $l = 1,\ldots,4$, denotes to the normalising constants, and $\theta_x = \psi_{11} + \psi_{10}$, $\theta_{y|x} = \frac{\psi_{11}}{\theta_x}$, $\theta_{y|\bar{x}} = \frac{\psi_{01}}{\theta_{\bar{x}}}$, $\theta_y = \psi_{11} + \psi_{01}$, $\theta_{x|y} = \frac{\psi_{11}}{\theta_y}$, and $\theta_{x|\bar{y}} = \frac{\psi_{10}}{\theta_{\bar{y}}}$.

Note that, in the distributions above, $\theta_x \perp\!\!\!\perp \theta_{y|x}$ and $\theta_y \perp\!\!\!\perp \theta_{x|y}$. This supplementing experimental evidence updates $\theta_{y|x}$ retains θ_x and keeps $\theta_x \perp\!\!\!\perp \theta_{y|x}$. However, because the Geiger and Heckerman condition is no longer true a posteriori, θ_y is no longer independent of $\theta_{x|y}$, the posterior density is not Hyperessential and the new prior is not multicausal. So information other than data equivalent to the direct observation in the idle system of joint margins of (X,Y) will prevent us identifying the manipulated and the idle system, implicit in the multicausal assumptions.

Note that the data affects the estimation of the idle margin of (X,Y) and this prevents the model from being an hypercausal Bayesian network. As we discussed above, it is therefore the idle system drifting away from the more scientific manipulated system-not vice versa !

Estimation of the probabilities in the idle model has introduced dependencies in the marginal Bayesian network due solely to the estimation process and spurious with respect to the mechanisms in the model. In particular, once the data set used to estimate the probabilities is large and exhaustive the model probabilities will be (almost) known and the model therefore (almost) hypercausal. These types of limiting results

and approximate Hypercausality are discussed in more detail in Daneshkhah and Smith (2003b).

It should be noticed that there are unique functional relationships between the hyper-parameters of the Dirichlet distributions characterised by Geiger and Heckerman and represented for the causal Bayesian network in this chapter. These relationships can help us to find out the given Bayesian network is a causal one. As we earlier discussed, in the Bayesian network, $X \rightarrow Y$, we typically know more about the X than the Y. This implies that the hyperparameter associated with θ_x is always larger than hyper-parameters associated with $\theta_{y|x}$. For example, consider fixed, but small values for $\alpha_{0.}$ and $\alpha_{1.}$. Then, we will obviously have small values for α_{00}, α_{01}, α_{10}, α_{11}, and for $\alpha_{.0}$ and $\alpha_{.1}$. Therefore, the Geiger and Heckerman condition is valid. But, if we consider a Dirichlet prior with small values on $\alpha_{0.}$ and $\alpha_{1.}$, and large values on α_{00} and α_{01} or, on α_{10} and α_{11} (such that the mentioned functional relationships between hyperparameters is no longer valid), the Geiger and Heckerman condition cannot be valid.

Furthermore, the importance of the parameter independence assumptions for the given causal Bayesian network can be indicated by using covariance matrix between the parameters. We examine the covariance matrix between $(\theta_x, \theta_{y|x}, \theta_{y|\bar{x}})$. If $(\theta_x, \theta_{y|x}, \theta_{y|\bar{x}})$ are mutually independent and jointly have Dirichlet distribution, then the mentioned covariance matrix will be

$$\Sigma_{(\theta_x, \underline{\theta}_{y|x})} = \begin{pmatrix} \frac{\alpha_{0.}\alpha_{1.}}{(\alpha_{0.}+\alpha_{1.})^2(\alpha_{0.}+\alpha_{1.}+1)} & 0 & \\ 0 & \frac{\alpha_{00}\alpha_{01}}{\alpha_{0.}^2(\alpha_{0.}+1)} & 0 \\ 0 & 0 & \frac{\alpha_{10}\alpha_{11}}{\alpha_{1.}^2(\alpha_{1.}+1)} \end{pmatrix}$$

As we can see, considering any sort of constraints on the parameters will give a covariance matrix with different structure such that the previously existed relationships between hyperparameters is no longer valid. As a result, Geiger and Heckerman's characterisation of the prior distribution is not also valid, and we will face more complexity

when defining prior distributions and computing posterior quantities.

This feature will be exacerbated by the fact that evidence about effects -perhaps symptoms- (in this case Y) is often more accessible than evidence about causes -perhaps diseases- (in this case X). So a spurious causal directionality in the Bayesian network from the idle system might thus be deduced just because of the form of the data or information we have, and is in this case reversed!

To conclude: hypercausal Bayesian networks will tend to be rare and multicausal Bayesian networks even more so. In observational studies they can be expected to give many spurious indications of causal direction when parameters are being estimated. However the causal framework developed in this paper at least provides a vehicle through which to seriously discuss some of the more basic inferential consequences of priors assuming local and global independence on the probabilities in a Bayesian network.

Chapter 7

The Robustness of the Bayesian Networks

7.1 Introduction

As we know, in a Bayesian network, a joint probability distribution over a set of random variables can be represented by a set of local conditional probability distributions. To specify a Bayesian network, one defines a directed acyclic graph which encodes conditional independencies among the variables. Finally, one specifies the local conditional distributions and prior distributions associated with parameters of each node given its parents.

The network structure and conditional distributions can be learned from data (e.g., Cooper and Herskovits (1992), Heckerman et al (1995), and Spiegelhalter and Lauritzen (1990). Note that all of these authors assumed the essential statement, local and global independence, on the parameters of the corresponding Bayesian network to make learning issues more feasible.) or, more commonly in systems applications, specified by experts.

Commonly, a Bayesian network is used to answer queries about the conditional distribution of a target variable (or a set of variables) given specific values of a subset of variables (usually called *evidence* variables).

Some questions arise here: is the network structure that is learned from data robust with respect to changes of the directionality of some specific arrows? Is the local conditional distribution associated with the specified node robust with respect to the changes to its prior distribution or to the changes to the local conditional distribution of another node? Most importantly, is the posterior distribution associated with the parameters of any node robust with respect to the changes to the prior distribution associated with the parameters of one specific node? Finally, are the quantities mentioned above robust with respect to the changes in the independence assumptions described in the last chapters?

Sensitivity analysis is concerned with understanding how changes in the model inputs influence the outputs. So the robustness of the output of a Bayesian network can be investigated by performing a sensitivity analysis of the network inputs. For a Bayesian network, more specifically, a sensitivity analysis serves to yield insight in the relation between the various parameters, assumptions, or probability assessments, of the network and its output.

One common approach to sensitivity analysis is to define reasonable ranges for each of the model parameters, vary each parameter from its lowest to highest reasonable values while holding the other variables fixed, and examine the resultant changes in the target value. This approach is known in the Bayesian literatures as global sensitivity analysis. But, in this thesis we will be more interested to see how sensitive are posterior quantities or the conditional probabilities with respect to small changes in the prior

distributions, or some essential assumptions such as independence assumptions. Such issues are addresses through what is called a local sensitivity analysis.

Local sensitivity analysis of standard general Bayesian problem is studied by Gustafson (1994), Gustafson et al (1996), Ruggeri and Wasserman (1994) and will be reviewed in Section 7.2. Gustafson (1996b) studied the local sensitivity measure for hierarchical models. He calculated the local sensitivity measure of the posterior quantities associated with the parameters of the specific stage[1](stage of inference) with respect to small changes of prior distribution associated with the parameters of the another stage (stage of uncertain specification). He made some quantitative and qualitative conclusions that will be presented in Section 7.2. In this chapter with the similar terminology we want to calculate the local sensitivity measure for the posterior quantities of the specific node in a Bayesian network with respect to minor changes in the prior distribution associated with the parameters of the another node. We introduce the hierarchical prior distribution for the Bayesian network with dependent prior distributions. But, when we have a large sample from a single population we find that we are immediately faced with identifiability problems. However, identifiability can be retrieved by taking several samples from different populations, all known with respect to the same structure of Bayesian network. This study will be presented in Section 7.4.

In Section 7.6 , we report a paper by Sivaganesan (1996) which examines the asymptotic behaviour of the specific local sensitivity measure for some ϵ-contamination classes of prior distributions under some mild conditions. The applications of these results in Bayesian networks will be given in this section. Furthermore, we are interested to estimate posterior distribution associated with the multinomial parameters of a given

[1]It should be noticed that *stage* in a hierarchical models is not a universal term. We present our meaning about stage in Section 7.3.

discrete Bayesian network from its equivalent Bayesian network. We have studied the asymptotic behaviour of the Hellinger distance between posterior distributions of the parameters of these two Bayesian networks in a working paper, and we have shown that this distance does not tend to zero (We have approximated this distance by the Laplace approximation, however, and have shown that this approximated distance tends to zero under some conditions.). This is another motivation to study the asymptotic behaviour of the local sensitivity measures described in this section and represented in terms of the Hellinger distance.

In Chapter 8, we study the asymptotic behaviour of the local sensitivity measures with respect to an arbitrary class of prior distributions. To show that the local sensitivity measures introduced in this chapter for large enough sample size tends to zero, we introduce a new class of metrics which we could examine prior to posterior convergence and sensitivity issues in a Bayesian model. We use this new metric to study the asymptotic behaviour of the local sensitivity measures derived for the several purposes for the Bayesian networks.

In Section 7.5, we present some results regarding the local sensitivity measures with respect to the changes in the causal Bayesian networks. We show that when the parameters are independent, the local sensitivity measures of any posterior quantity of arbitrary node with respect to small changes of the another node will be zero. However, when the parameters are not independent, we can not generally conclude this result, and this leads us to suggest a construct called approximate causality.

In this thesis, we study only local sensitivity analysis in Bayesian networks with discrete variables. However, we note that these results can be extended in a relatively straightforward way to Bayesian network with continuous variables.

7.2 Introduction to the Bayesian Robustness

Robust Bayesian analysis is the study of sensitivity of Bayesian answers to uncertain inputs such as sampling model, prior distribution, or loss function, or any combination of them. Bayesian sensitivity studies have recently seen an explosion of interest and literature on this subject. There are several reasons for this interest: foundational motivation, practical Bayesian motivation, and acceptance of Bayesian analysis (For details and examples for each reason see Berger (1984, 1990, 1994), Wasserman (1992)).

Sensitivity analysis can be divided into two categories, *global* and *local* sensitivity. The common approach to assessing sensitivity is to measure the size of the class of posteriors (or perhaps just a particular posterior quantity) that arises from a specified class of priors. This is referred to as global sensitivity analysis. The fact that global analyses often entail a large and complex computational problem has lead to a recent explosion of interest in local sensitivity analyses (see Gustafon (1996a), Gustafson et al (1996) and references therein). The idea of a local analysis is to examine the rate at which the posterior changes, relative to the prior. In this section, we introduce several methods of evaluating the local sensitivity of posterior quantities.

7.2.1 Local Sensitivity Analysis

How much does a small change in the prior (or likelihood) affect our inferences? In this section, we review one class of methods that attempt to answer this question. This class of methods is represented by local sensitivity measure. Some reasonable candidates turn out to be inappropriate for a number of reasons (see Gustafson et al (1996) for reasons).

The main focus of local sensitivity analysis is in multivariate analysis. One of the most important questions in this case is: how sensitive is the posterior marginal den-

sity for one parameter when the prior distribution associated with another parameter changes? In this chapter, we want to answer the adapted question for a Bayesian network.

Suppose a unique prior π is elicited, but small changes in the concentration of the prior distribution, caused errors in the elicitation process. Hence, measures which are '*functionally close*' to π are considered and the behaviour of the posterior functionals, under infinitesimal departures from π, are studied.

Now, we want to introduce some local sensitivity measures which will be used in this chapter.

The first local sensitivity measure that will be considered in this section is Fréchet derivative, briefly described below.

For the fixed prior distribution $P \in \Gamma$, the posterior expectation for measures in

$$\Delta_{\mathcal{M}}(P) = \{P + \delta : \delta \in \mathcal{M}\}$$

is calculated, where Γ is the class of all probability measures over parameter space, Θ, and $\mathcal{M}$ is a subset of Δ consisting of all signed measures δ with $\delta(\Theta) = 0$. Δ is a normed, linear space as $\Delta = \{\delta \geq 0 : \|\delta\| < \infty\}$ with norm given by $\|\delta\| = d(\delta, 0)$, where d is total variation metric. In particular, the distance between two probability measures $P, Q \in \Gamma$ is defined by a function $\mathbf{d} : \Gamma \times \Gamma \longrightarrow [0, \infty)$ such that $\mathbf{d}(P, Q) = \rho(P - Q) = \rho(\delta)$. Where P and Q denote the prior distribution associated with the prior densities (mass functions) p and q respectively, and ρ-function satisfies in the following conditions: (i) $\rho(\delta) = \rho(-\delta)$, (ii) $\rho(c\delta) = |c|\rho(\delta)$, for some constant c, (iii) $\rho(\delta_1 + \delta_2) \leq \rho(\delta_1) + \rho(\delta_2)$ (see Gustafson (1996a, b)).

Let $\dot{T}_g^P$ be the Fréchet derivative of the non-linear operator $T_g : \Delta \longrightarrow R$, that is, a

linear map on Δ satisfying

$$\dot{T}_g^P(\delta) = T_g(P+\delta) - T_g(P) + o(\|\delta\|). \tag{7.1}$$

$\dot{T}_g^P$ measures how a small change in P affects the posterior expectation. In other words, roughly stated, inference about the function[2] of interest, $g(\theta)$ is robust when T_g is not changing rapidly at zero; that is, when T_g^P has a small derivative at zero. The norm (the restricted norm) of $\dot{T}_g^P$ over $\mathcal{M}$ is defined by

$$\|\dot{T}_g^P\|_{\mathcal{M}} = sup_{\delta\in\mathcal{M}} \frac{|\dot{T}_g^P|}{\|\delta\|} \tag{7.2}$$

which is considered as a sensitivity measure of the posterior when the prior varies in $\mathcal{M}$. If Δ is considered as the set of all signed measures $\delta = \epsilon(Q-P)$, then

$$\Delta_{\mathcal{M}}(P) = \{(1-\epsilon)P + \epsilon Q, Q \in \Gamma\} = \{P + \epsilon(Q-P)\}$$

will be an ϵ-contamination class of priors.

Notice that the restricted norm defined in Equation (7.1) is not normalised. For computational reason the following norm as a distance between two distribution functions is usually used (see Gustafson (1996 a, b), Gustafson, et al (1996) and Fernholz (1983)). The local sensitivity measure of P in the direction of Q is defined by

$$S(P,Q;x) = \lim_{\epsilon\to 0} \frac{d(P^x, Q_\epsilon^x)}{d(P,Q_\epsilon)} \tag{7.3}$$

Where Q_ϵ denotes the perturbation of P in direction Q. In addition to the linear perturbation(ϵ-contamination class), recently the geometric perturbation that is given by $d_{Q_\epsilon} \propto [\frac{dQ}{dP}]^\epsilon dP$ has been used.

The ϕ-divergence distance between two probability measures is induced very useful class

[2]It should be noticed that for a function g on the parameter space, the posterior expectation of $g(\theta)$ can be thought of as a functional of the prior.

of the different distances such as Kullback-Liebler, Hellinger distance, Chi-squared distance, etc. It is given by

$$D_\phi(P,Q) = \int \phi(\frac{d_P}{d_Q}) d_P$$

where ϕ is a smooth convex function such that $\phi(1) = 0$. The properties of these distances under linear and geometric perturbations are studied by Dey and Birmiwal (1994) and Dey, et al (1996). We shall briefly discuss some of these useful results.

The local sensitivity measures based on the ϕ-divergence (with the same definition as above) with respect to linear perturbation of the prior, likelihood, and both of them, are given receptively by

$$S(P,Q;x) = \frac{Var_{P^x(f)}(\frac{d_Q}{d_P})}{Var_P(\frac{d_Q}{d_P})} = \Big(\frac{m_Q(x,f)}{m_P(x,f)}\Big)^2 \frac{Var_{P^x(f)}(\frac{d_{Q^x}(f)}{d_{P^x}(f)})}{Var_P(\frac{d_Q}{d_P})},$$

$$S(f,g;x) = \frac{Var_{P^x(f)}(\frac{g}{f})}{Var_f(\frac{g}{f})} = \Big(\frac{m_P(x,g)}{m_P(x,f)}\Big)^2 \frac{Var_{P^x(f)}(\frac{d_{P^x}(g)}{d_{P^x}(f)})}{Var_P(\frac{g}{f})},$$

$$S(f,g;P,Q;x) = \frac{Var_{P^x(f)}(\frac{g}{f})}{Var_P(\frac{d_P}{d_Q})} = \Big(\frac{m_P(x,g)}{m_P(x,f)}\Big)^2 \frac{Var_{P^x(f)}(\frac{d_{P^x}(g)}{d_{P^x}(f)})}{Var_P(\frac{d_Q}{d_P})}.$$

Similarly, the local sensitivity measures under geometric perturbation are, respectively,

$$S(P,Q;x) = \frac{Var_{P^x(f)}(log\frac{d_Q}{d_P})}{Var_P(log\frac{d_Q}{d_P})}, \tag{7.4}$$

$$S(f,g;x) = \frac{Var_{P^x(f)}(log\frac{g}{f})}{Var_P(log\frac{g}{f})},$$

$$S(f,g;P,Q;x) = \frac{Var_{P^x(f)}(log\frac{g}{f})}{Var_P(log\frac{d_Q}{d_P})}.$$

where, similarly, f is the elicited likelihood and g is the contaminated likelihood which belongs to a certain class $\mathcal{L}$.

The following points can be deduced from the results above:

- S(P, Q; x) is free from the choice of the ϕ-functions;
- the local sensitivity measures above depend on specific variance ratios and the compatible Bayes factors;
- this measure is actually the Fréchet derivative (see Fernholz (1983), Huber (1981), and Basu (1996) for details);
- the influence of the observation x can be assessed based on the measures above;
- the relative sensitivity with respect to two different perturbed classes of priors (likelihood) can be achieved by the restricted norms;
- the Fréchet derivative with respect to Kolomogrove and Levy metric (Cuevas and Sanz (1988) and Basu (1996)) is investigated for the measures above. But, the study of the local sensitivity measures in terms of Hellinger distance is also promising. Note that Beran (1977 a, b) considered this metric to study robustness with respect to location in classical approaches.

The next sensitivity measure is suitable for assessing sensitivity to prior marginals. A perturbation to a prior marginal can be considered as a perturbation to the entire prior. Assume that we have parameterised $\underline{\theta}$ so that it can be partitioned as $\underline{\theta} = (\underline{\phi}, \underline{\psi})$, where we would like to assess sensitivity to the prior marginal $P(\underline{\phi})$ (or denoted by $P_{\underline{\phi}}$) while keeping the prior conditional $P(\underline{\psi} \mid \underline{\phi})$ fixed. That is, the prior density is $dQ(\phi)dP(\underline{\psi} \mid \phi)$, but we replace $dQ(\phi)$ with $dQ(\phi) + \delta$. In analogy to the methods introduced above, we should specify the size of variation of the posterior expectation of $g(\underline{\theta})$ under the perturbed prior. The posterior expectation of $g(\underline{\theta})$ is given by

$$T_g(Q) = \frac{\int g(\underline{\theta}) L(\underline{\theta} \mid x) dP(\underline{\psi} \mid \underline{\phi}) dQ(\underline{\phi})}{\int L(\underline{\theta} \mid x) dP(\underline{\psi} \mid \underline{\phi}) dQ(\underline{\phi})}$$

where $L(\underline{\theta} \mid x)$ stands for the likelihood function.

Hampel et al (1986) represent a very simple local sensitivity measure in terms of the directional derivative of T at $P_{\underline{\phi}}$ in direction Q. This derivative can be defined under weak conditions, and can be expressed in an influence function representation as follows

$$\frac{\partial}{\partial_\epsilon} T((1-\epsilon)P_{\underline{\phi}} + \epsilon Q)|_{\epsilon=0} = \int IF_P(z) d[Q - P_{\underline{\phi}}](z) \tag{7.5}$$

where IF_P is given by

$$IF_P(z) = E^x(g(\underline{\theta}) - E^x(g(\underline{\theta})) \mid \underline{\phi} = z)[\frac{dP^x_{\underline{\phi}}}{dP_{\underline{\phi}}}(z)] \tag{7.6}$$

Let E^x and P^x denote the posterior expectation and posterior distribution respectively under the prior P. Note that the IF_P term is useful, because it does not depend on the direction of Q, and so represents the local sensitivity to perturbations in all directions. Equation (7.5) only defines the influence function up to an additive constant. This function can be standardized by requiring $\int IF_P(z) dP_\phi(z) = 0$. However, Equation (7.6) is already in standardised form.

This measure is used by Gustafson (1996a, b) to compute the local sensitivity measure for hierarchical models. In the next section, we use the adapted version of this measure to study local sensitivity analysis for some misspecifications in Bayesian networks.

7.2.2 Some Aspects of Bayesian Robustness in Hierarchical Models

In this part, we briefly examine local sensitivity analysis with respect to small perturbations at various stages of a hierarchically specified prior. First, the general behaviour of sensitivity analysis across levels of the hierarchy is briefly examined. Then, the technical matters of the local sensitivity measures associated with these models introduced

by Gustafson (1996 b) will be presented.

The basic idea to use (Bayesian) *hierarchical* models (random-effects models in analysis of variance, random coefficient regression models, Kalman filter theory, and Markov Chain, are some examples that can be expressed by hierarchical models) is to specify a joint distribution for data and parameters, through a succession of conditional distributions. In particular, the conditional distribution of a data vector $\underline{x}$ given a parameter vector $\underline{\theta}_1$ is specified, followed by the distribution of $\underline{\theta}_1$ given a second parameter vector $\underline{\theta}_2$, and so on. At some point the specification terminates, with the distribution of $\underline{\theta}_{k+1}$ taken to be degenerate, that is, the conditional distribution for $\underline{\theta}_{k+1} \mid \underline{\theta}_k$ is specified, where $\underline{\theta}_{k+1}$ is a known hyperparameter. After collecting the data, all statistical inference is based on the posterior distribution of the entire parameter $(\underline{\theta}_1, \ldots, \underline{\theta}_k)$ given data vector, $\underline{x}$ and hyperparameter, $\underline{\theta}_{k+1}$.

In 1981, Goel and DeGroot showed that, for many measures of information, the gain in information decreases as one moves to higher levels of hyperparameters. They show that there is more information in the observation about first level of hierarchy than the other levels. In other words, for many of the widely used information measures (e.g., the expected information, the Entropy function, Kullback-Leibler information, Renyi information divergence, etc), the information about the hyperparameters decreases as one moves to higher levels away from the data. Goel (1983) presents the same result for a wide class of information measure, based on the ϕ-divergence distances.

Gustafson (1996 b) presented some techniques for assessing local sensitivity with respect to uncertainties in prior marginals. He also extended these techniques to study the local sensitivity analysis associated with hierarchical models. We briefly present his

results in this section.

Let S_{ij} be a measure of sensitivity of the $\underline{\theta}_j$ (posterior marginal) with respect to uncertainty about the prior specification for $\underline{\theta}_i \mid \underline{\theta}_{i+1}$ (prior conditional). One might postulate that S_{ij} decreases as $|i-j|$ increases. That means, sensitivity will be lower when there are many stages interceding between the stage of uncertain specification and the stage of inference. This turns out to be only partly true. Let us consider two possible hierarchical models, p and q, which differ only in $\underline{\theta}_i \mid \underline{\theta}_{i+1}$. By the following lemma introduced by Gustafson (1996b), the result claimed above could be concluded (for more details, see Gustafson (1996 b), page 64.).

Lemma 7.1 Assume that $f(x_1, x_2)$ and $g(x_1, x_2)$ each have a joint density on the sample space $(\mathcal{X}_1, \mathcal{X}_2)$ with respect to some dominating measure. If $f(x_1 \mid x_2) = g(x_1 \mid x_2)$, then $d_{TV}(f(x_1), g(x_1)) \leq d_{TV}(f(x_2), g(x_2))$, with equality if and only if $f(x_2) = g(x_2)$.

Where d_{TV} denote the total variation distance. The total variation distance is a metric on probability distributions (P, Q) on a common σ-algebra $\mathcal{A}$ on a sample space Ω, and is defined as follows,

$$d_{TV}(P, Q) = \sup_{A \in \mathcal{A}} |P(A) - Q(A)|.$$

Let $d(j) = d_{TV}(p(\theta_j \mid x, \theta_{k+1}), q(\theta_j \mid x, \theta_{k+1}))$, for fixed j and x. Gustafson proved that $d(j)$ is a strictly unimodal function, which is maximised at either $j = i$ or $j = i + 1$.

Hence for perturbations at a particular stage of the hierarchy, sensitivity falls off as the level of inference moves away from the level of perturbation, in either direction. This is in agreement with the postulated behaviour discussed above.

We will present a similar lemma in terms of the Hellinger distance in the next sub-

section, that is more useful to study the sensitivity analysis of Bayesian networks. The ideas of sensitivity analysis of Bayesian networks in many situations are derived in the analogous way that is introduced for sensitivity analysis of hierarchical models.

We also adapt another version of this lemma (Lemma 8.1) that is appropriate to study asymptotic behaviour of a new sensitivity measure introduced in Chapter 8. A new proof is also given in Section 8.3.

Gustafson (1996 a) argued that it is much harder to make general statements when the stage of inference is fixed and the stage of perturbation varies. In the special case for the normal model with known variances, it looks possible to make some progress. He considers the prior distribution for hierarchical normal model as follows:

$$\theta_i \mid \theta_{i+1} = \theta_{i+1} + \epsilon_i,$$

where $\epsilon_0, \ldots, \epsilon_k$ are independently normally distributed, with respective variances $\sigma_0^2, \ldots, \sigma_k^2$. The last stage of hierarchy, i.e, θ_{k+1} is known and $\epsilon_{k+1} = 0$.
He calculated the local sensitivity measure of the posterior mean of θ_j with respect to small perturbation of the prior on ϵ_i. In other words, the location structure is held fixed and unchanged under perturbation, and the noise distribution is just perturbed. He then showed that when j is fixed and i varies over $\{1, \ldots, j-1\}$, the sensitivity measure and σ_i^2 share the same ordering. This is also true when i is changing over $\{j, \ldots, k\}$. It can therefore be concluded that the separation $|i-j|$ does not play a role in the sensitivity ordering. Furthermore, we can conclude that the stages of higher prior variances are more influential on inference at a particular stage. An important consequence is that inference will be sensitive to improper priors, a finding that agrees with Pericchi and Nazaret (1988).

7.3 Sensitivity Analysis in Bayesian Networks

In this section the sensitivity of posterior quantities of Bayesian networks with respect to the following perturbations is investigated, using a local method described above. These perturbations are: (i) perturbation with respect to a prior distribution associated with a node or with respect to joint prior distribution associated with the subset of parameters of the given Bayesian network, (ii) perturbation with respect to independence assumptions between parameters of two nodes including global independence and local independence assumptions between configurations of each node. Furthermore, we can assess the affect of removing or adding one edge (or more) between two nodes in terms of well-known distances such as Hellinger distance.

Before we start to present a sensitivity analysis of Bayesian networks with respect to the aforementioned perturbations, we should introduce the Hellinger distance and give the reasons why we prefer to use this distance here.

To define Hellinger distance, let us denote $P(\theta)$ and $Q(\theta)$ as two probability distributions with respective densities $p(\theta)$ and $q(\theta)$ over a random variable (vector) θ - which in our applications will be considered as a collection of (conditional) probabilities - then Hellinger distance between these densities is given by

$$H^2(p(\theta), q(\theta)) = \int_\theta (\sqrt{p(\theta)} - \sqrt{q(\theta)})^2 d\theta = 2[1 - \int_\theta (p(\theta)q(\theta))^{\frac{1}{2}} d\theta]$$

where $0 \leq H^2(p(\theta), q(\theta)) \leq 2$. For more details see Smith (1995).

For many purposes we favour the Hellinger metric (or its functionally equivalent symmetric Chernov separation) to apply in the computation of distance between posterior distributions associated with the graphical models, particularly with Bayesian networks. The first reason for this choice is that, unlike its competitors of the variation

and Kullback-Leibler separation measure, distances between such densities can often be written in closed form for the distributions we have in mind, like products of independent Dirichlet (Beta) distributions, so the algebraic examination of sensitivity is fairly straightforward. Secondly, unlike the Kullback-Leibler separation it is a proper metric so our intuition about distances cannot be distorted when using it. Thirdly it is shared most of the advantages and properties of both these alternatives, being topologically equivalent to the variation metric and is such that convergence in Kullback-Leibler implies convergence in Hellinger distance.

To begin with, we study the sensitivity analysis of the mentioned aspects above over some special Bayesian networks, and go on to generalise this study for general Bayesian networks.

It should be noticed that the structure of a hierarchical model is not the same as a Bayesian network. But, in Section 7.4, we will show that a Bayesian network can be considered as a two stages hierarchical model with one latent variable. Therefore, Gustafson's ideas might be useful for Bayesian networks to answer the following questions: Do the quantities associated with the target variables change by altering the inference variables? Can we find out that the target variable is more sensitive with respect to the changes of which variable? Does the posterior marginal of the specific variable (or variables) alter by perturbing the prior distribution associated with some other specific variable?.

First, let us consider a Bayesian network with the following decomposition of the joint probability distribution for which the global independence assumption is valid.

$$p(\underline{x}, \underline{\theta}) = \prod_{v \in V} p(x_v \mid pa(v), \theta_v) p(\theta_v)$$

where $\underline{x} = (x_v, v \in V)$ and the prior distribution is defined as $p(\underline{\theta}) = \prod_{v \in V} p(\theta_v)$.

In this Bayesian network the local sensitivity measure of the posterior marginal distribution of the specific parameter, for instance, θ_j with respect to small perturbation of the prior distribution associated with any other node, for example, θ_i, $i \neq j$, is zero.

This is true since

$$p(\theta_j \mid \underline{x}) =$$

$$\frac{p(\theta_j)p(x_j \mid pa(j), \theta_j)\{\prod_{v \in V \setminus (i,j)} \int p(\theta_v)p(x_v \mid pa(v), \theta_v)d\theta_v\}\{\int p(\theta_i)p(x_i \mid pa(i), \theta_i)d\theta_i\}}{\{\prod_{v \in V \setminus i} \int p(\theta_v)p(x_v \mid pa(v), \theta_v)d\theta_v\}\{\int p(\theta_i)p(x_i \mid pa(i), \theta_i)d\theta_i\}}$$

equals to $q(\theta_j \mid \underline{x})$ that is obtained by replacing $p(\theta_i)$ by $q(\theta_i)$ in $p(\theta_j \mid \underline{x})$, where $q(\theta_i)$ is a perturbed prior distribution of $p(\theta_i)$.

Therefore we can say that, subject to existence of the global parameter independence in a Bayesian network, perturbing the prior distribution of one node does not change the posterior quantity of other nodes. However, when the node of uncertain prior specification and the node of inference are the same, the local sensitivity measure in terms of the Hellinger distance (or total variation distance) can be calculated as

$$S_i^H(\underline{x}) = \frac{H^2(p(\theta_i \mid \underline{x}), q(\theta_i \mid \underline{x}))}{H^2(p(\theta_i), q(\theta_i))}$$

or

$$S_i^{TV} = \frac{d_{TV}(p(\theta_i \mid \underline{x}), q(\theta_i \mid \underline{x}))}{d_{TV}p(\theta_i), q(\theta_i))}$$

This study when the global independence assumption is not valid would be more complicated. For example, consider the Bayesian networks shown in Figure 7.2 with the following factorisation of joint probability distribution,

$$p(\underline{x}, \underline{\theta}) = \prod_{v=1}^{k} p(x_v \mid x_{v-1}, \theta_v)p(\theta_v \mid \theta_{v-1}) \tag{7.7}$$

We denote the prior densities of two Bayesian networks with the same structure but

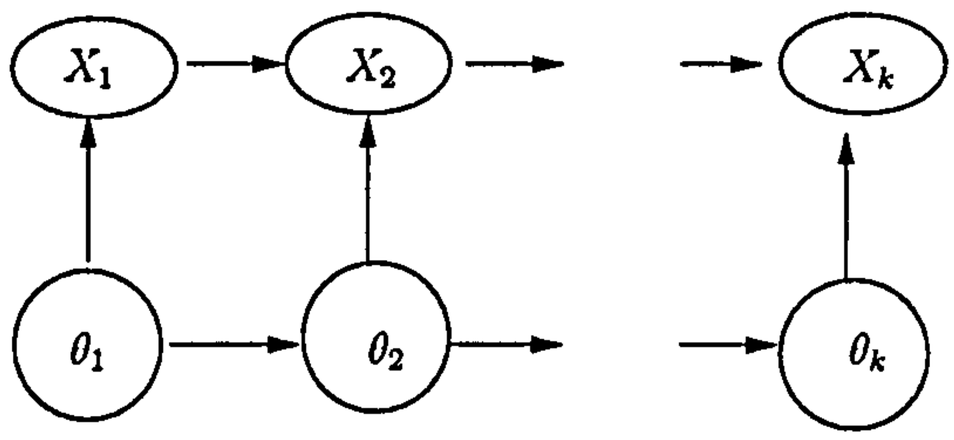


Figure 7.1: The network structure with the dependent parameters.

with different distributions on $\theta_i \mid \theta_{i-1}$, by p and q. To examine local sensitivity analysis for this Bayesian network, we use Lemma 7.1. The modified version of Lemma 7.1 for the Hellinger distance is given in the following lemma.

Lemma 7.2 Consider two different density functions, $f(x_1, x_2)$ and $g(x_1, x_2)$ on (X_1, X_2) defined on the sample space $(\mathcal{X}_1, \mathcal{X}_2)$, with respect to some dominating measure. If $f(x_1 \mid x_2) = g(x_1 \mid x_2)$, then $H^2(f(x_1), g(x_1)) \leq \sqrt{2}H^2(f(x_2), g(x_2))$, with equality if and only if $f(x_2) = g(x_2)$.

Proof As we know, the Hellinger distance is defined as

$$H^2(f(x_1), g(x_1)) = 2[1 - \int f(x_1)^{\frac{1}{2}} g(x_1)^{\frac{1}{2}} dx_1]$$

From the following inequalities[3]

$$H^2(f(x_1), g(x_1)) \leq d_{TV}(f(x_1), g(x_1)) \leq \sqrt{2}H(f(x_1), g(x_1)) \tag{7.8}$$

[3]It should be noticed that these inequalities are valid for any two densities and with the same dominating measure. These inequalities imply that not only H is a metric but it is topologically

, and the results mentioned in Lemma 7.1, we can write

$$H^2(f(x_1), g(x_1)) \leq d_{TV}(f(x_1), g(x_1)) \leq d_{TV}(f(x_2), g(x_2)) \leq \sqrt{2}H(f(x_2), g(x_2)) \quad (7.9)$$

By combining (7.8) and (7.9), the proof is complete, that is,

$$H^2(f(x_1), g(x_1)) \leq \sqrt{2}H(f(x_2), g(x_2)).$$

We denote $S_{ij}(x)$ the local sensitivity measure[4] of the posterior marginal, θ_j with respect to uncertainty about the conditional prior specification of $\theta_i \mid \theta_{i-1}$. The following result can be obtained from Lemma 7.1.

Corollary 7.1 Let $d_{TV}(j) = d_{TV}(p(\theta_j \mid x), q(\theta_j \mid x))$, for fixed i. Then $d_{TV}(j)$ is a strictly unimodal function[5], which is maximised at either $j = i$ or $j = i + 1$.

Proof First consider the case in which $1 \leq j < i$, since the distribution of $\{\theta_{j-1} \mid \theta_j, \underline{x}\}$ is determined by

$$p(\theta_{j-1} \mid \theta_j, \underline{x}) = \frac{\int \{\prod_{v \in V} p(\theta_v \mid \theta_{v-1}) p(x_v \mid x_{v-1}, \theta_v) \prod_{v \in V \setminus (j, j-1)} d\theta_v\}}{\int \{\prod_{v \in V} p(\theta_v \mid \theta_{v-1}) p(x_v \mid x_{v-1}, \theta_v) \prod_{v \in V \setminus j} d\theta_v\}} = q(\theta_{j-1} \mid \theta_j, \underline{x})$$

---

equivalent to d_{TV} - small in d_{TV} is equivalent to small in H. In particular, if H is bounded away from zero we can conclude that so is d_{TV} and hence betting schemes can be devised to penalise the use of an approximator. Furthermore, since d_{TV} is complete over distributions, H must also be (See Smith (1995)).

[4]Note that, $S_{ij}(\underline{x})$ can be calculated by any of the sensitivity measures that are introduced in Section 7.2.

[5]It should be noticed that the general ϕ-divergence between two densities $p(\theta_j \mid x)$ and $q(\theta_j \mid x)$ is given as

$$D_\phi(p(\theta_j \mid x), q(\theta_j \mid x)) = \quad p(\theta_j \mid x)\phi(\frac{q(\theta_j \mid x)}{p(\theta_j \mid x)})$$

, where we assume that ϕ is a convex function with a bounded third derivative. There are several well-known ϕ-divergence measures. For example, $\phi(x) = \frac{1}{2}|x - 1|$ defines the variational distance of L_1 norm and $\phi(x) = (\sqrt{x} - 1)^2$ gives Hellinger distance. Therefore, the results mentioned in Corollary 7.1 for the variational distance can be obtained for the Hellinger distance. See Dey et al (1996).

Therefore, by the lemma above, we would say

$$d_{TV}(p(\theta_{j-1} \mid \underline{x}), q(\theta_{j-1} \mid \underline{x})) < d_{TV}(p(\theta_j \mid \underline{x}), q(\theta_j \mid \underline{x})).$$

Similarly, for the case $i-1 < j$, we have $p(\theta_j \mid \theta_{j-1}, \underline{x}) = q(\theta_j \mid \theta_{j-1}, \underline{x})$, and by applying Lemma 7.1, we have

$$d_{TV}(p(\theta_j \mid \underline{x}), q(\theta_j \mid \underline{x})) < d_{TV}(p(\theta_{j-1} \mid \underline{x}), q(\theta_{j-1} \mid \underline{x})).$$

Therefore, the local sensitivity measure of the posterior marginal, θ_j (i.e., $S_{ij}(\underline{x})$), should be more sensitive with respect to uncertainty about the prior specification for the conditional prior $\theta_j \mid \theta_{j-1}$ or $\theta_{j-1} \mid \theta_{j-2}$. This result agrees with this point that $\forall 1 \leq j \leq k$, $\theta_j \perp\!\!\!\perp \theta_{j-2} \mid \theta_{j-1}$. That means the posterior marginal θ_j would be sensitive with respect to perturbations on $p(\theta_j \mid \theta_{j-1})$ or $p(\theta_{j-1} \mid \theta_{j-2})$. It should be noticed that if the orientation between nodes in the DAG shown in Figure 7.1 is changed, then the result above will no longer be valid. We will discuss this issue later in this chapter. Furthermore, one can pick up the possible relationship between the result above and causality interpretation in terms of external intervention that will be studied in this chapter as well.

Now, let us consider general Bayesian networks including Bayesian networks with dependent prior distributions. First, we introduce similar notation to that introduced in Chapter 5.

Let the vector $\underline{X} = (X_1, \ldots, X_k)$, of nodes of a Bayesian network have its components X_i, $1 \leq i \leq k$, listed in an order compatible with G and their corresponding vectors of probabilities $\underline{\theta}_1, \ldots, \underline{\theta}_k$ compatibly with the partial order induced by the directed edges of the Bayesian network. The general prior distribution can be then defined as

$$p(\underline{\theta}) = \prod_{i=1}^{k} p(\underline{\theta}_i \mid \underline{\theta}^{i-1}), \quad \underline{\theta}^{i-1} = \{\underline{\theta}_1, \ldots, \underline{\theta}_{i-1}\}$$

Each component of $\underline{\theta}_i$ is given by

$$\theta_{i(j)|pa_{i(l)}} = p(X_i = x_{i(j)} \mid pa_i = pa_{i(l)})$$

where $\theta_{i(j)|pa_{i(l)}}$ denotes the parameter associated with the level j of i^{th} variable and the level l of its parents. Moreover, the joint mass function is defined as follows

$$p(\underline{x} \mid \underline{\theta}) = \prod_{v=1}^{k} \underline{\theta}_v$$

For more details, see Chapter 4.

Similarly, to assess the local sensitivity measures associated with the posterior marginal distribution of $\underline{\theta}_j$ with respect to small perturbation of prior distribution of $\underline{\theta}_i, i \neq j$, Equation (7.4) can be used.

The posterior marginal of $\underline{\theta}_j$ is calculated as

$$p(\underline{\theta}_j \mid \underline{x}) = \frac{\int \prod_{v=1}^{k} \underline{\theta}_v \{\prod_{v \in V \setminus \{i,j\}} p(\underline{\theta}_v \mid \underline{\theta}^{v-1})\} p(\underline{\theta}_j \mid \underline{\theta}^{j-1}) p(\underline{\theta}_i \mid \underline{\theta}^{i}) \prod_{v \in V \setminus \{j\}} d\underline{\theta}_v}{\int \prod_{v=1}^{k} \underline{\theta}_v \{\prod_{v \in V \setminus \{i,j\}} p(\underline{\theta}_v \mid \underline{\theta}^{v-1})\} p(\underline{\theta}_j \mid \underline{\theta}^{j-1}) p(\underline{\theta}_i \mid \underline{\theta}^{i}) \prod_{v \in V} d\underline{\theta}_v}$$

$$= \frac{\underline{\theta}_j p(\underline{\theta}_j \mid \underline{\theta}^{j-1}) \int \prod_{v \in V \setminus \{j\}} \underline{\theta}_v \{\prod_{v \in V \setminus \{i,j\}} p(\underline{\theta}_v \mid \underline{\theta}^{v-1})\} p(\underline{\theta}_i \mid \underline{\theta}^{i}) \prod_{v \in V \setminus \{j\}} d\underline{\theta}_v}{\int \prod_{v=1}^{k} \underline{\theta}_v \{\prod_{v \in V \setminus \{i,j\}} p(\underline{\theta}_v \mid \underline{\theta}^{v-1})\} p(\underline{\theta}_j \mid \underline{\theta}^{j-1}) p(\underline{\theta}_i \mid \underline{\theta}^{i}) \prod_{v \in V} d\underline{\theta}_v}$$

If we perturb the prior distribution $p(\underline{\theta}_i \mid \underline{\theta}^{i-1})$ by

$$q_\epsilon(\underline{\theta}_i \mid \underline{\theta}^{i-1}) = (1 - \epsilon) p(\underline{\theta}_i \mid \underline{\theta}^{i-1}) + \epsilon q(\underline{\theta}_i \mid \underline{\theta}^{i-1})$$

a different distribution will be obtained for $q_\epsilon(\underline{\theta}_j \mid \underline{x})$. Because, we cannot factorise the integrals in the denominator and numerator of the fraction above into separated terms to enable us to cancel common terms. However, one can easily show that if there is global independence assumption between parameters, then $q_\epsilon(\underline{\theta}_j \mid \underline{x}) = p(\underline{\theta}_j \mid \underline{x})$ and consequently the Hellinger distance between these two distributions is zero. That means,

in the presence of the global independence assumption, perturbing a prior distribution associated with parameters of one node will not change the posterior marginal of another node. It is the same result that we have obtained earlier.

As we mentioned in Cororllary 7.1, the local sensitivity measure would decrease as the node under inference moves away from the node under perturbation in any direction. Furthermore, this general framework enables us to study sensitivity analysis with respect to the existence of the local independence assumption between parameters of one specified node. In this situation, the prior distribution associated with the parameters of X_i and given the level l of its parents is given by

$$p(\underline{\theta}_i) = \prod_{v=1}^{n_i} p(\theta_{i(v)|pa_{i(l)}})$$

where n_i denote the number of the states of i^{th} variable. But, the prior distribution for the same set of parameters without the local independence assumption is given by

$$p(\underline{\theta}_i) = \prod_{v=1}^{n_i} p(\theta_{i(v)|pa_{i(l)}} \mid \theta^{v-1}_{i|pa_{i(l)}}), \quad \theta^{v-1}_{i|pa_{i(l)}} = \{\theta_{i(1)|pa_{i(l)}}, \ldots, \theta_{i(v-1)|pa_{i(l)}}\}.$$

Obviously, the computation of the Hellinger distance between the posterior distributions associated with these prior distributions would be dependent on the structure of the Bayesian network under study (and the structure between parameters in the case of lack of independence assumptions). This computation would be usually infeasible. But, we could approximate it by numerical methods such as MCMC.

Now, we want to assess the local sensitivity measure of the posterior quantities with respect to variations of the prior distributions or the mentioned assumptions for the Bayesian networks with the discrete variables. In this case, it should be easier to use the sensitivity measure presented in Equations (7.5) and (7.6). To ensure that the results obtained above are valid, we will calculate this measure under two assumptions associated with the parameters: (i) global parameter independence; and (ii) dependent parameters.

First, we consider the Bayesian network with the following factorisation,

$$p(\underline{X}, \underline{\theta}) = \{\prod_{v=1}^{k} p(X_v \mid pa(x_v), \theta_v)\} \times \prod_{v=1}^{k} p(\theta_v)$$

where $p(\underline{\theta}) = \prod_{v=1}^{k} p(\theta_v)$ (i.e., the global independence assumption exists). To assess the sensitivity of posterior quantities of the specific node (nodes) with respect to perturbations of prior marginals of the other nodes, $\underline{\theta}$ should be partitioned as $\underline{\theta} = (\theta_v, \underline{\phi})$, where $\underline{\phi} = \underline{\theta} \setminus \theta_v$.

As described in the last section, to evaluate the local sensitivity measure of the posterior quantities with respect to the variations of the prior marginal, θ_v with the fixed conditional prior distribution of $\underline{\phi} \mid \theta_v$, we can use the following local sensitivity measure

$$\|\dot{T}^g(p(\theta_v))\| = \{\int (IF_p(z))^2 p_{\theta_v}(z)dz\}^{\frac{1}{2}} \tag{7.10}$$

where $IF_p(z) = E^x(g(\underline{\theta}) - E(g(\underline{\theta}) \mid x) \mid \theta_v = z)\{\frac{p_{\theta_v}(z|x)}{p_{\theta_v}(z)}\}$.

Now, let us suppose each variable X_v is distributed as a multinomial distribution with states $\{x_{vt} : 1 \leq t \leq m_v\}$, and the corresponding parameters are defined as $\theta_v = p(x_v \mid pa(x_v), \underline{\theta})$ for $v = 1, \ldots, k$. To define the baseline prior distribution on each node (and the whole Bayesian network), we need to make some assumptions to make computations more feasible. The first assumption that we made already is global parameter independence. The second assumption is local parameter independence. Therefore, for a fixed configuration $pa(x_v) = pa^*(x_v)$, the marginal prior of θ_v is distributed as a Dirichlet distribution (see Spiegelhalter and Lauritzen (1990) and Geiger and Heckerman (1997)),

$$\theta_v^* = (\theta_{v1}^*, \ldots, \theta_{vm_v}^*) \sim \mathcal{D}(\alpha_{v1}^*, \ldots, \alpha_{vm_v}^*)$$

where $\theta_{vt}^* = p(X_v = x_{vt} \mid pa(x_v) = pa^*(x_v), \underline{\theta})$.

Thus, the local sensitivity measure for the purpose mentioned above and with the complete data set can be calculated by (7.10).

The following example would help to see how the local sensitivity measure defined above will be calculated for a Bayesian network that is consistent with the assumptions described above.

Example 7.1 Let us consider the Bayesian network with three nodes that is shown in Figure 7.2.

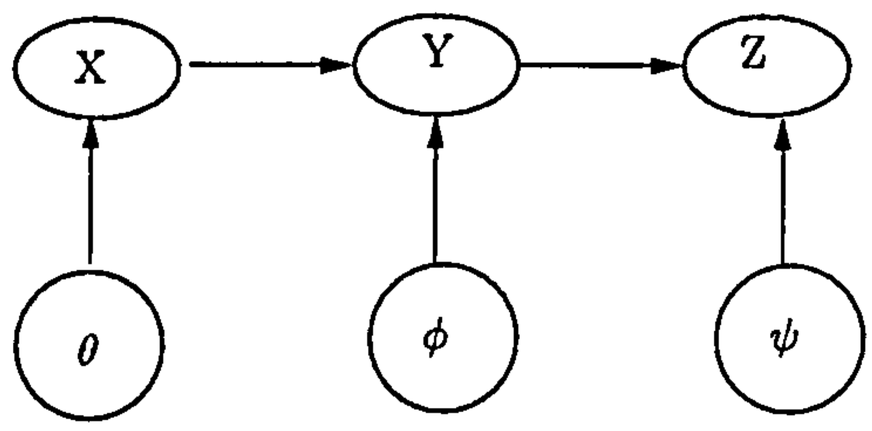


Figure 7.2: The representation of Bayesian network associated with the second sensitivity example

We assume each variable can takes three values, and the corresponding parameters are defined as follows

$$\theta_i = p(X = x_i \mid \underline{\mu}) \quad i = 1, 2, 3$$

$$\phi_{ij} = p(Y = y_i \mid X = x_j, \underline{\mu}) \quad i, j = 1, 2, 3$$

$$\psi_{ij} = p(Z = z_i \mid Y = Y_j, \underline{\mu}) \quad i, j = 1, 2, 3$$

where $\underline{\mu} = (\theta, \phi, \psi)$.

By assuming the local and global parameters independence, the prior distributions associated with the parameters are Dirichlets with the following characterisations

$$p(\theta) = \mathcal{D}(\alpha_1, \alpha_2, \alpha_3),$$

$$p(\phi_i) = \mathcal{D}(\lambda_{1i}, \lambda_{2i}, \lambda_{3i}) \quad i = 1, 2, 3,$$

$$p(\psi_i) = \mathcal{D}(\tau_{1i}, \tau_{2i}, \tau_{3i}) \quad i = 1, 2, 3,$$

If we observe the complete data set, for example, $d_1 = (X = x_1, Y = y_2, Z = z_1)$, then the likelihood function will be

$$p(X = x_1, Y = y_2, Z = z_1) = \theta_1 \phi_{21} \psi_{12}$$

The local sensitivity measure of posterior expectation of , for example, $g(\underline{\mu}) = \frac{\theta_2 \phi_{23} \psi_{13}}{\theta_1 \phi_{31} \psi_{12}}$ with respect to small perturbation of $p(\phi_1)$, is given by,

$$\|\dot{T}^g(p(\phi_1))\| = \{\int E^{d_1}(g(\underline{\mu}) - E(g(\underline{\mu}) \mid d_1) \mid \phi_1 = \hat{\phi}_1)\{\frac{p_{\phi_1}(\hat{\phi}_1 \mid d_1)}{p_{\phi_1}(z)}\}^2 p_{\phi_1}(\hat{\phi}_1) d\hat{\phi}_1\}^{\frac{1}{2}}$$

where $\hat{\phi}_1 \in (0, 1)$ and $p(\phi_1 \mid d_1) = \mathcal{D}(\lambda_{11}, \lambda_{21} + 1, \lambda_{31})$.

The local sensitivity measure above can be computed as follows,

$$\|\dot{T}^g(p(\phi_1))\| = \{\frac{\alpha_2 \lambda_{23} \tau_{13} \tau_{.2} \lambda_{.1}}{\alpha_1 \lambda_{21} \tau_{12} \tau_{.3} \lambda_{.3}}\}^2 + \{\frac{\alpha_2 \lambda_{23} \tau_{13} \tau_{.2} \lambda_{.1}^2 (\tau_{.1})}{\alpha_1 \lambda_{21}^2 \tau_{12} \tau_{.3} \lambda_{.3} (\tau_{31})}\}^2 \times \{\frac{\lambda_{21}(\lambda_{21} + 1)}{(\lambda_{.1} + 1)\lambda_{.1}}\}$$

$$-2 \times \{\frac{\alpha_2 \lambda_{23} \tau_{13} \tau_{.2}}{\alpha_1 \tau_{12} \tau_{.3} \lambda_{.3}}\} \times \{\frac{\alpha_2 \lambda_{23} \tau_{13} \tau_{.2} \lambda_{.1} (\tau_{.1})}{\alpha_1 \lambda_{21} \tau_{12} \tau_{.3} \lambda_{.3} (\tau_{31})}\} \times \{\frac{\lambda_{21}}{\lambda_{.1}}\}$$

As a numerical example, let hyperparameters take the following values,

$(\alpha_1 = 4, \alpha_2 = 6, \alpha_3 = 3)$, $(\lambda_{11} = 3, \lambda_{21} = 4, \lambda_{31} = 6)$, $(\lambda_{12} = 4, \lambda_{22} = 5, \lambda_{32} = 1)$, $(\lambda_{13} = 2, \lambda_{23} = 1, \lambda_{33} = 5)$, $(\tau_{11} = 5, \tau_{21} = 3, \tau_{31} = 3)$, $(\tau_{12} = 6, \tau_{22} = 4, \tau_{32} = 8)$, $(\tau_{13} = 2, \tau_{23} = 5, \tau_{33} = 3)$,

then the local sensitivity measure will be $\|\dot{T}^g(p(\phi_1))\| = 0.3577$.

For the noninformative prior distribution $p(\phi_1) = \mathcal{D}(\lambda_{11} = 1, \lambda_{21} = 1, \lambda_{31} = 1)$, the

norm above will be, $\|\dot{T}^g(p(\phi_1))\| = 1.9841$.

Note that if the function of interest is in terms of θ_i's only, then the local sensitivity measure[6] of $g(\theta_i)$ with respect to small perturbations to ϕ_i's or ψ_i's will be zero. For example, in the last example if we chose $g(\underline{\mu}) = \theta_1$, then

$$\|\dot{T}^g(p(\phi_1))\| = 0$$

The result above is similar with the result that is obtained for the case in which the parameters were globally independent of each other. Therefore, we can say that the posterior quantities associated with the specified nodes will remain unchanged with respect to perturbation of the prior distributions on the other nodes. The methodology for evaluating the local sensitivity measure is similar when the parameters are not independent of each other but the computation in this case becomes tedious. The numerical methods (e.g., MCMC) to calculate the influence function, $IF(z)$, and the posterior marginal distributions would be useful (see Gustafson(1996a) for the similar work in hierarchical models).

Example 7.2 In this example, we present the local sensitivity analysis for the general discrete *directed acyclic graph* model. Let us consider this DAG model with the multinomial likelihood function as follows

$$p(\underline{x} \mid \underline{\theta}) = \prod_{v=1}^{k} p(x_v \mid x_{pa(v)}, \underline{\theta}_v)$$

---

[6]The small values of the local sensitivity measure given the hyperparameters indicate some degree of robustness with respect to the choice of the perturbed prior. The inference of interest will be insensitive with respect to the choice of the priors if the local sensitivity measure is almost zero. However, how small the local sensitivity measure should be to conclude that the inference is not sensitive with respect to the choice of priors depends on the context under study.

By assuming local and global parameters independence, the prior distribution associated with parameters are given by

$$p(\underline{\theta}) = \prod_{i=1}^{k} p(\underline{\theta}_i) \quad \text{(global independence)}$$

and

$$\forall i = 1, \ldots, k, \quad p(\underline{\theta}_i) = \prod_{v=1}^{n_i} p(\theta_{i(v)|pa_i(l)}), \quad \text{(local independence)}$$

where $pa_i(l)$ denote the level l of the parent configurations of X_i, and n_i stands for the number of the states of i^{th} variables.

It is usual to assign a Dirichlet distribution to the parameters associated with each node as follows

$$\underline{\theta}_i \sim \mathcal{D}_l(\alpha_{i1}, \ldots, \alpha_{in_i}), \quad i = 1, \ldots, k, \quad l = 1, \ldots, m_i$$

and therefore the prior distribution associated with $\underline{\theta}$ is a Dirichlet product as follows

$$p(\underline{\theta}) = \prod_{i=1}^{k} \prod_{l=1}^{m_i} \mathcal{D}_l(\alpha_{i1}, \ldots, \alpha_{in_i})$$

The likelihood function of $\underline{\theta}$ given data, $\underline{x}$ can be written as

$$p(\underline{x} \mid \underline{\theta}) = \prod_{i=1}^{k} \prod_{l=1}^{m_i} \theta_{i(v)|pa_{i(l)}}^{x_{pa(i)}}$$

Now, suppose we are uncertain about the prior distribution associated with $\underline{\theta}$ or a specific element of $\underline{\theta}$. Without loss of generality, we wish to study local sensitivity analysis with respect to uncertainly in $p(\underline{\theta}_1)$. We can represent this uncertainty by the following linear perturbation,

$$q_\epsilon = (1 - \epsilon)\mathcal{D}_1(\alpha_{11}, \ldots, \alpha_{1n_1}) + \epsilon\mathcal{D}_1(\beta_{11}, \ldots, \beta_{1n_1})$$

The local sensitivity measure under this linear perturbation based on the φ-divergence (including the Hellinger distance) is given by

$$S(p, q; \underline{x}) = \{\frac{m_q(\underline{x})}{m_p(\underline{x})}\}^2 \frac{Var_{p(\underline{\theta}_1|\underline{x})}(\frac{q(\underline{\theta}_1|\underline{x})}{p(\underline{\theta}_1|\underline{x})})}{Var_{p(\underline{\theta}_1)}(\frac{q(\underline{\theta}_1)}{p(\underline{\theta}_1)})} \tag{7.11}$$

where $p = \mathcal{D}_1(\alpha_{11}, \ldots, \alpha_{1n_1})$, $q = \mathcal{D}_1(\beta_{11}, \ldots, \beta_{1n_1})$, $\underline{x} = (x_1, \ldots, x_2)$ denote the data, $p(. \mid \underline{x})$ and $q(., \mid \underline{x})$ denote the posterior distributions with respective prior densities p and q, $m_p(\underline{x}) = \int_{\theta_1} p(\underline{x} \mid \underline{\theta}_1) p(\underline{\theta}_1) d\underline{\theta}_1$ and similarly $m_q(\underline{x})$ can be defined.
It can be easily shown that

$$p(\underline{\theta}_1 \mid \underline{x}) = \mathcal{D}_1(\alpha_{11} + x_{11}, \ldots, \alpha_{1n_1} + x_{1n_1}), \quad q(\underline{\theta}_1 \mid \underline{x}) = \mathcal{D}_1(\beta_{11} + x_{11}, \ldots, \beta_{1n_1} + x_{1n_1})$$

$$m_p(\underline{x}) = \frac{\Gamma(\alpha_1)}{\prod_{i=1}^{n_1} \Gamma(\alpha_{1i})} \frac{\prod_{i=1}^{n_1} \Gamma(\alpha_{1i} + x_{1i})}{\Gamma(\alpha_1 + N_1)}, \quad m_q(\underline{x}) = \frac{\Gamma(\beta_1)}{\prod_{i=1}^{n_1} \Gamma(\beta_{1i})} \frac{\prod_{i=1}^{n_1} \Gamma(\beta_{1i} + x_{1i})}{\Gamma(\beta_1 + N_1)}$$

where $N_1 = \Sigma_{i=1}^{n_1} x_{1i}$.

Therefore, the local sensitivity measure in Equation (7.11) becomes

$$S(p, q; \underline{x}) = \{\frac{\Gamma(\beta_1)\Gamma(\alpha_1 + N_1) \prod_{i=1}^{n_1} \Gamma(\alpha_{1i})\Gamma(\beta_{1i} + x_{1i})}{\Gamma(\alpha_1)\Gamma(\beta_1 + N_1) \prod_{i=1}^{n_1} \Gamma(\beta_{1i})\Gamma(\alpha_{1i} + x_{1i})}\}^2 \times$$

$$\frac{\{\frac{\Gamma^2(\beta_1+N_1) \prod_{i=1}^{n_1} \Gamma(\alpha_{1i}+x_{1i})\Gamma(2\beta_{1i}-\alpha_{1i}+x_{1i})}{\Gamma(\alpha_1+N_1)\Gamma(2\beta_1-\alpha_1+N_1) \prod_{i=1}^{n_1}(\Gamma(\beta_{1i}+x_{1i}))^2} - 1\}}{\{\frac{\Gamma^2(\beta_1) \prod_{i=1}^{n_1} \Gamma(\alpha_{1i})\Gamma(2\beta_{1i}-\alpha_{1i})}{\Gamma(\alpha_1)\Gamma(2\beta_1-\alpha_1)(\prod_{i=1}^{n_1} \Gamma(\beta_{1i}))^2} - 1\}}$$

The smaller values of $S(p, q, \underline{x})$ indicating some degree of robustness with respect to the choice of the prior (see Dey et al (1996) for more detail).

Example 7.3 Here, we would like to study sensitivity analysis with respect to uncertainty of the directionality of an arrow between two nodes in the equivalence class of Bayesian networks. For simplicity, let us consider the equivalence class of Bayesian networks with two binary random variables. This class contains two Bayesian networks shown in the following figure. Let assume, we are uncertain about the direction of the arrow between X and Y. That is, it is not clear that the direction of the mentioned arrow is either $X \longrightarrow Y$ or $X \longleftarrow Y$.

To evaluate the local sensitivity measure, the following perturbed prior distribution is considered,

$$p_\epsilon(\theta_x, \theta_{y|x}, \theta_{y|\bar{x}}) = (1 - \epsilon)p(\theta_x, \theta_{y|x}, \theta_{y|\bar{x}}) + \epsilon q(\theta_y, \theta_{x|y}, \theta_{x|\bar{y}})$$

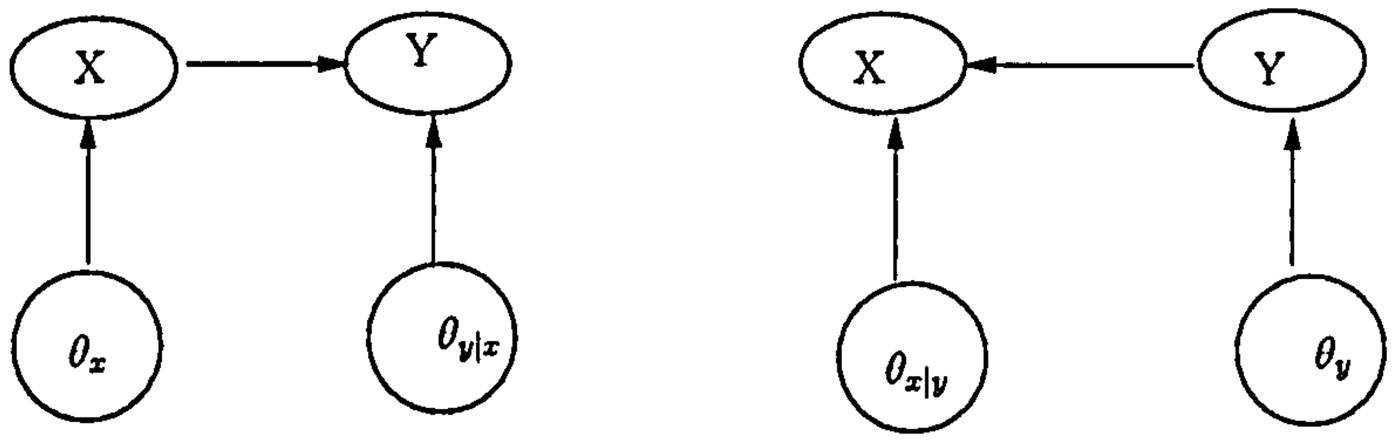


Figure 7.3: The representation of the Bayesian network with the discrete domain and its Markov equivalent.

where $p(.)$ denote the prior distribution of the parameters associated with $X \longrightarrow Y$ and $q(.)$ denote the prior distribution of the parameters of $Y \longrightarrow X$. In the perturbed prior distribution above, the base prior distribution is p and the contaminated part of the perturbed prior distribution is denoted by q.

The local sensitivity measure for this purpose is given by

$$S(p,q;\underline{x}) = (\frac{m_q(\underline{x})}{m_p(\underline{x})})^2 \frac{E_{q(.|\underline{x})}(\frac{q(\theta_x,\theta_{x|y},\theta_{x|\bar{y}}|\underline{x})}{p(\theta_x,\theta_{y|x},\theta_{y|\bar{x}}|\underline{x})}) - 1}{E_{q(.)}(\frac{q(\theta_x,\theta_{x|y},\theta_{x|\bar{y}})}{p(\theta_x,\theta_{y|x},\theta_{y|\bar{x}})}) - 1}$$

Geiger and Heckerman (1997) showed that the prior distributions associated with the parameters on these two equivalent Bayesian networks are Dirichlet as follows

$$p(\theta_x,\theta_{y|x},\theta_{y|\bar{x}}) = K\theta_x^{(\alpha_1+\alpha_2)-1}\theta_{\bar{x}}^{(\alpha_3+\alpha_4)-1}\theta_{y|x}^{\alpha_1-1}\theta_{\bar{y}|x}^{\alpha_2-1}\theta_{y|\bar{x}}^{\alpha_3-1}\theta_{\bar{y}|\bar{x}}^{\alpha_4-1}$$

and

$$p(\theta_y,\theta_{x|y},\theta_{x|\bar{y}}) = K\theta_y^{(\alpha_1+\alpha_3)-1}\theta_{\bar{y}}^{(\alpha_2+\alpha_4)-1}\theta_{x|y}^{\alpha_1-1}\theta_{\bar{x}|y}^{\alpha_3-1}\theta_{x|\bar{y}}^{\alpha_2-1}\theta_{\bar{x}|\bar{y}}^{\alpha_4-1}$$

where K stands for the normalised constant (see Chapter 3 for details).

Therefore,

$$E_{q(.)}(\frac{q(\theta_x, \theta_{x|y}, \theta_{x|\bar{y}})}{p(\theta_x, \theta_{y|x}, \theta_{y|\bar{x}})}) = \int \frac{(\theta_x\theta_{y|x} + \theta_{\bar{x}}\theta_{y|\bar{x}})^2(\theta_x\theta_{\bar{y}|x} + \theta_{\bar{x}}\theta_{\bar{y}|\bar{x}})^2}{\theta_x^2\theta_{\bar{x}}^2} p(\theta_x, \theta_{y|x}, \theta_{y|\bar{x}}) =$$

$$\frac{(\alpha_1+1)(\alpha_2+1)\alpha_1\alpha_2}{(\alpha_1+\alpha_2+3)(\alpha_1+\alpha_2+2)(\alpha_3+\alpha_4-1)(\alpha_3+\alpha_4-2)} +$$

$$\frac{(\alpha_1+1)(\alpha_4+1)\alpha_1\alpha_4}{(\alpha_1+\alpha_2+1)(\alpha_1+\alpha_2)(\alpha_3+\alpha_4+1)(\alpha_3+\alpha_4)}$$

$$+\frac{2(\alpha_1+1)\alpha_4\alpha_1\alpha_2}{(\alpha_1+\alpha_2+2)(\alpha_1+\alpha_2+1)(\alpha_3+\alpha_4-1)(\alpha_3+\alpha_4)}$$

$$+\frac{(\alpha_3+1)(\alpha_4+1)\alpha_3\alpha_4}{(\alpha_1+\alpha_2-1)(\alpha_1+\alpha_2-2)(\alpha_3+\alpha_4+3)(\alpha_3+\alpha_4+2)} +$$

$$\frac{(\alpha_2+1)(\alpha_3+1)\alpha_2\alpha_3}{(\alpha_1+\alpha_2+1)(\alpha_1+\alpha_2)(\alpha_3+\alpha_4+1)(\alpha_3+\alpha_4)} +$$

$$\frac{2(\alpha_3+1)\alpha_3\alpha_4\alpha_2}{(\alpha_1+\alpha_2-1)(\alpha_1+\alpha_2)(\alpha_3+\alpha_4+2)(\alpha_3+\alpha_4+1)} +$$

$$\frac{2(\alpha_2+1)\alpha_1\alpha_3\alpha_2}{(\alpha_1+\alpha_2+2)(\alpha_1+\alpha_2+1)(\alpha_3+\alpha_4-1)(\alpha_3+\alpha_4)} +$$

$$\frac{2(\alpha_4+1)\alpha_4\alpha_1\alpha_3}{(\alpha_1+\alpha_2-1)(\alpha_1+\alpha_2)(\alpha_3+\alpha_4+1)(\alpha_3+\alpha_4+2)} +$$

$$\frac{4\alpha_1\alpha_2\alpha_3\alpha_4}{(\alpha_1+\alpha_2+1)(\alpha_1+\alpha_2)(\alpha_3+\alpha_4+1)(\alpha_3+\alpha_4)} \tag{7.12}$$

Similarly, we can compute $E_{q(.|\underline{x})}(\frac{q(\theta_x,\theta_{x|y},\theta_{x|\bar{y}})}{p(\theta_x,\theta_{y|x},\theta_{y|\bar{x}})})$ by replacing α_i's by $\alpha_i + x_i$ in each term in Equation (7.12). The equation associated with $E_{q(.|\underline{x})}(\frac{q(\theta_x,\theta_{x|y},\theta_{x|\bar{y}})}{p(\theta_x,\theta_{y|x},\theta_{y|\bar{x}})})$ is called (7.12*). Furthermore, it can be shown that

$$\frac{m_q(\underline{x})}{m_p(\underline{x})} = 1$$

Thus, $S(p, q; \underline{x})$ will be computed by plugging Equations (7.12) and (7.12*) into $E_{q(.)}(\frac{q(\theta_x,\theta_{x|y},\theta_{x|\bar{y}})}{p(\theta_x,\theta_{y|x},\theta_{y|\bar{x}})})$ and $E_{q(.|\underline{x})}(\frac{q(\theta_x,\theta_{x|y},\theta_{x|\bar{y}})}{p(\theta_x,\theta_{y|x},\theta_{y|\bar{x}})})$ respectively.

We can also evaluate the local sensitivity measure with respect to uncertainty of the local independence assumption of the specific node in a Bayesian network.

Example 7.4 As an example, let us consider the Bayesian network shown in the left hand side of Figure 7.4 above. For the purpose above, the following perturbed prior is suggested,

$$p_\epsilon(\theta_x, \theta_{y|x}, \theta_{y|\bar{x}}) = (1-\epsilon)p(\theta_x, \theta_{y|x}, \theta_{y|\bar{x}}) + \epsilon q(\theta_x, \theta_y)$$

where $q(\theta_x, \theta_y) = p(\theta_x)p(\theta_y)$, and $p(\theta_y) = Beta(\alpha_1 + \alpha_3, \alpha_2 + \alpha_4)$.

Therefore, the required local sensitivity measure can be easily calculated as follows

$$S(p, q; \underline{x}) = \lim_{\epsilon \to 0} \frac{H(p^*_\epsilon(\theta_x, \theta_{y|x}, \theta_{y|\bar{x}} \mid \underline{x}), p^*(\theta_x, \theta_{y|x}, \theta_{y|\bar{x}} \mid \underline{x}))}{H(p_\epsilon(\theta_x, \theta_{y|x}, \theta_{y|\bar{x}}), p(\theta_x, \theta_{y|x}, \theta_{y|\bar{x}}))} =$$

$$\{\frac{m(\underline{x} \mid q)}{m(\underline{x} \mid p)}\}^2 \times \frac{var_{p^*}(\frac{q^*(\theta_x, \theta_{y|x}, \theta_{y|\bar{x}} \mid \underline{x})}{p^*(\theta_x, \theta_{y|x}, \theta_{y|\bar{x}} \mid \underline{x})})}{var_p(\frac{q(\theta_x, \theta_{y|x}, \theta_{y|\bar{x}})}{p(\theta_x, \theta_{y|x}, \theta_{y|\bar{x}})})}$$

The small values of $S(p, q; \underline{x})$ means the inference in terms of posterior distribution of parameters of those nodes is not sensitive to the local independence assumption.

Example 7.5 In this example, we want to assess local sensitivity analysis with respect to uncertainty about the prior distribution associated with the parameters on the cliques in the Bayesian networks shown in Figure 7.4. This graph consists of three cliques: $C_1 = \{X_1, X_2, X_3\}$, $C_2 = \{X_2, X_3, X_4\}$, $C_3 = \{X_4, X_5, X_6\}$, $S_2 = \{X_2, X_3\}$, $S_3 = \{X_4\}$. Note that $S_2 \cap S_3 = \emptyset$.

According to Figure 7.4, we can conclude that

$$\theta_1 \perp\!\!\!\perp (\theta_2, \theta_3) \perp\!\!\!\perp \theta_4 \perp\!\!\!\perp (\theta_5, \theta_6)$$

However, (θ_2, θ_3) and (θ_5, θ_6) are dependent on each other.

The likelihood function of the Bayesian network shown in Figure 7.4 can be written as

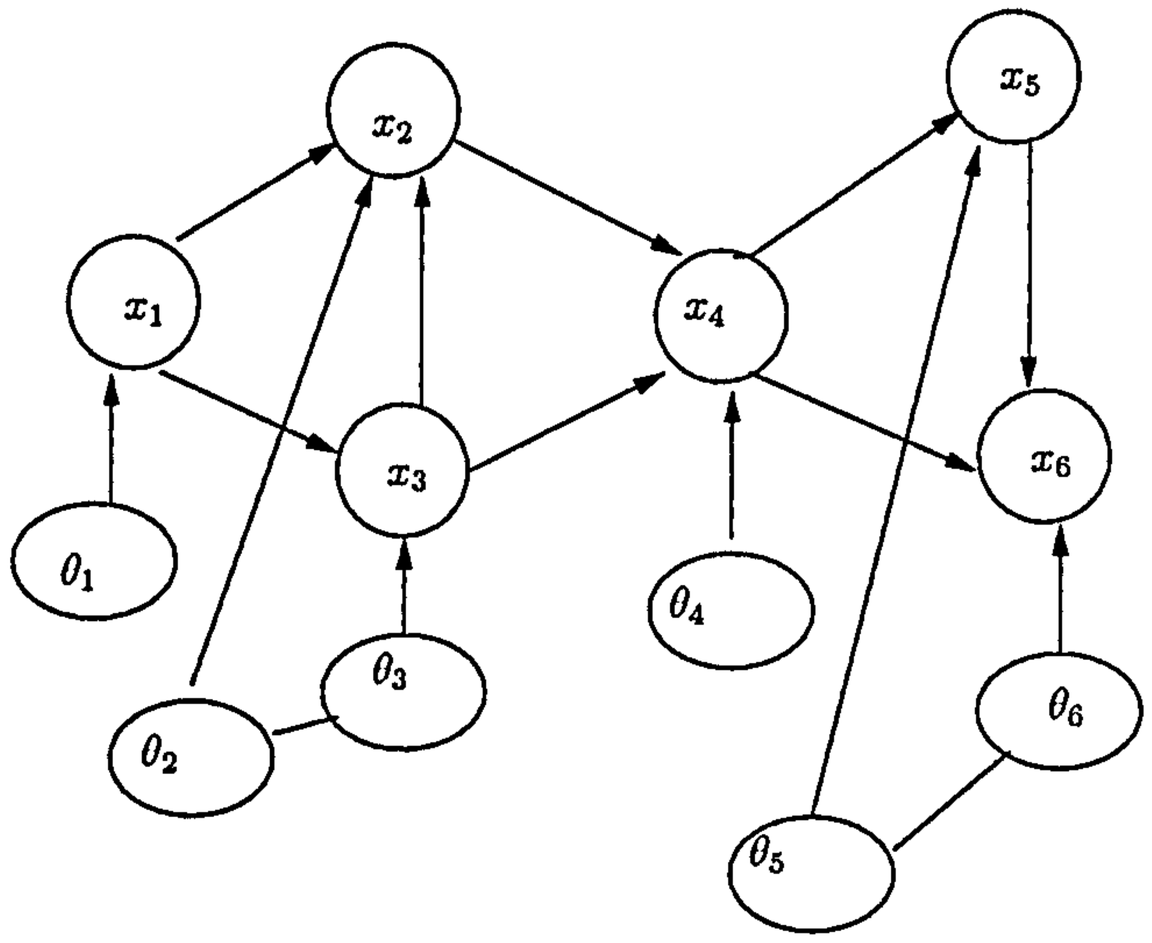


Figure 7.4: The representation of a Bayesian network with three cliques and disjoint separators.

follows

$$p(\theta_1, x_1, \theta_2, \theta_3, x_2, x_3, \theta_4, x_4, \theta_5, \theta_6, x_5, x_6) =$$

$$p(\theta_1, x_1, \theta_2, \theta_3, x_2, x_3) p(\theta_4, x_4 \mid x_2, x_3) p(\theta_5, \theta_6, x_5, x_6 \mid x_4)$$

Let us assume that we are uncertain about the prior distribution associated with $(\theta_1, \theta_2, \theta_3)$. The following geometric perturbed prior distribution is appropriate for this purpose.

$$q_\epsilon(.) = \frac{p^\epsilon(.) q^{1-\epsilon}(.)}{\int_\theta p^\epsilon(.) q^{1-\epsilon}(.)}$$

where

$$q(\theta_1, x_1, \theta_2, \theta_3, x_2, x_3, \theta_4, x_4, \theta_5, \theta_6, x_5, x_6) =$$

$$q(\theta_1, x_1, \theta_2, \theta_3, x_2, x_3) p(\theta_4, x_4 \mid x_2, x_3) p(\theta_5, \theta_6, x_5, x_6 \mid x_4)$$

However, we can similarly define the perturbed prior distribution associated with uncertainty in θ_4 or (θ_5, θ_6).

The perturbed prior distribution, q_ϵ, is defined as follows

$$q_\epsilon(.) \propto \int_{(x_1,x_2,x_3)} \{p^\epsilon(\theta_1, x_1, \theta_2, \theta_3, x_2, x_3) q^{1-\epsilon}(\theta_1, x_1, \theta_2, \theta_3, x_2, x_3) dx_1 dx_2 dx_3\} \int_{x_4} \{p(\theta_4, x_4 \mid x_2, x_3) d$$

$$\int_{(x_5,x_6)} \{p(\theta_5, \theta_6, x_5, x_6 \mid x_4) dx_5 dx_6\}$$

or

$$\log q_\epsilon(.) \propto \log \int_{(x_1,x_2,x_3)} \{p^\epsilon(\theta_1, x_1, \theta_2, \theta_3, x_2, x_3) q^{1-\epsilon}(\theta_1, x_1, \theta_2, \theta_3, x_2, x_3) dx_1 dx_2 dx_3\} +$$

$$\log \int_{x_4} \{p(\theta_4, x_4 \mid x_2, x_3) dx_4\} + \log \int_{(x_5,x_6)} \{p(\theta_5, \theta_6, x_5, x_6 \mid x_4) dx_5 dx_6\}$$

We can compute the local sensitivity measure for the purpose above under geometric (or linear) perturbation by the sensitivity measures presented in Equation (7.3) or (7.4). However, the computation of these measures should be infeasible and the numerical methods, such as MCMC, are required.

This example helps us to define prior distribution for Bayesian networks with dependent parameters in terms of hierarchical prior distributions. We will study sensitivity analysis of these Bayesian networks in the next section.

Example 7.6 (Forensic Science) In this example, we present an application of the local sensitivity analysis to study robustness of Bayesian network used in Forensic science with respect to uncertainty in the prior distributions associated with each variable. Now, we give the details of the Bayesian network, shown in Figure 7.5, associated with the assessment of Forensic Fibre Evidence in the discussion below.

The forensic scientist compares evidence through the likelihood ratio of the probability of evidence given the prosecution proposition $Pr(E \mid C)$ versus the probability of

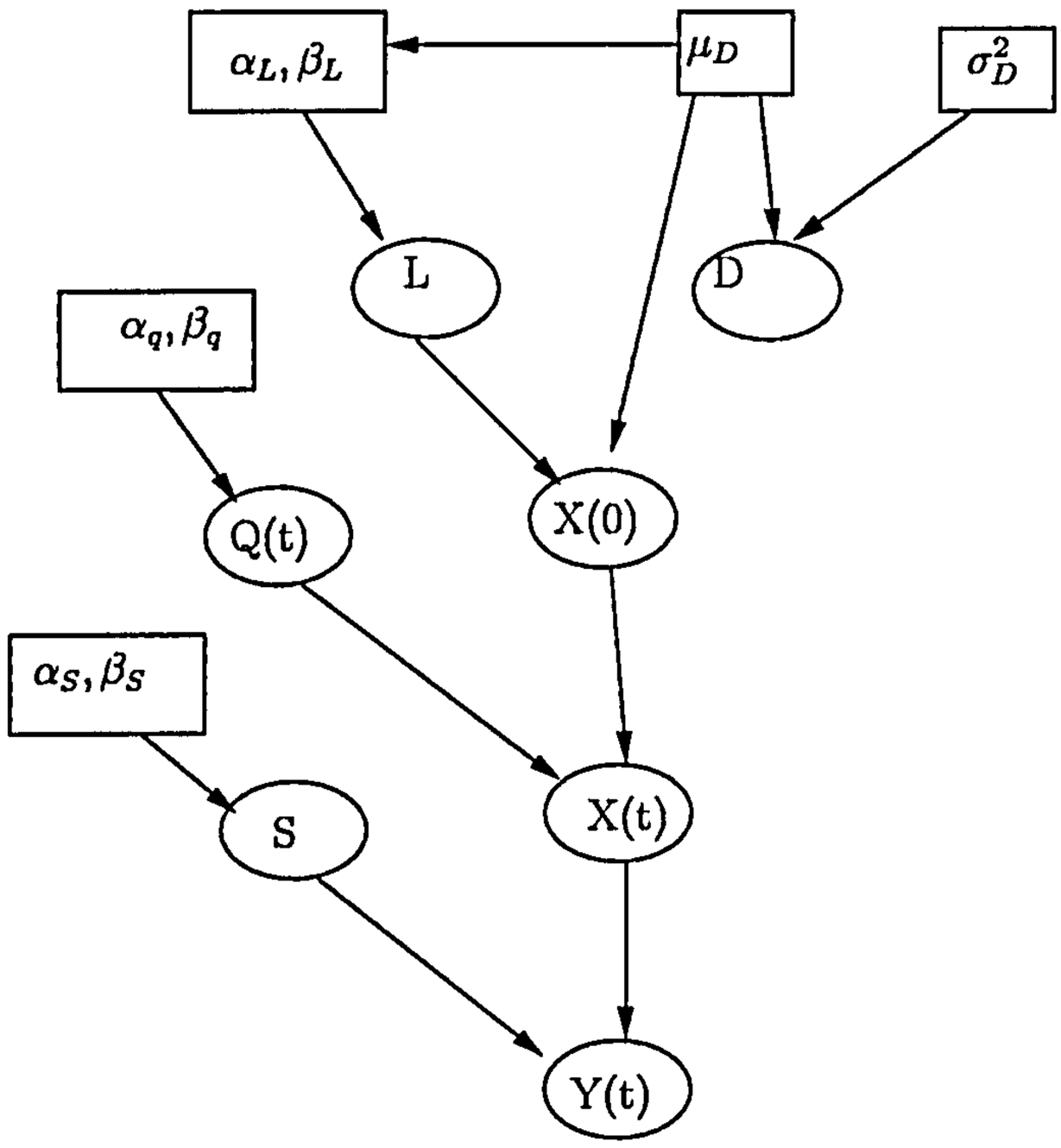


Figure 7.5: Transfer, persistence and recovery Bayesian network.

evidence[7] given the defence proposition $Pr(E \mid \bar{C})$. More precisely, the likelihood ratio is defined as follows:

$$LR = \frac{Pr(\text{Evidence} \mid C)}{Pr(\text{Evidence} \mid \bar{C})}$$

where C denotes the prosecution proposition and is defined as

C: The suspect wore the mask found at the crime scene at the time when the crime occurred,

[7]The evidence in this case study is the number of retrieved fibres.

and

$\bar{C}$: The suspect is unconnected to the incident.

Puch and Smith (2002) presented a Bayesian network (shown in Figure 7.5) for computing a distribution for the number Y_t of fibres retrieved from the suspect's head hair that were originally transferred from the musk.

They suggested the following distributions for the variables L and X_0:

$$L \sim Gam(\alpha_L, \beta_L), \qquad X_0 \mid L \sim Pois(L)$$

where L models the average number of fibres that are transferred to the offender's head hair and X_0 models the actual number of transferred fibres. For details of the parameters α_L, β_L, see Puch and Smith (2002).

The variables $X(0)$, $Q(t)$ and $X(t)$ model persistence. The following distributions are considered for these variables:

$$Q(t) \sim Beta(\alpha_q, \beta_q), \quad X(t) \sim Bin(X(0), \mu_D Q(t))$$

where $Q(t)$ models the proportion of fibres that persisted in the suspect's head up to time t without considering physical and head disturbance. The variable $X(t)$ counts the number of fibres that persisted with success rate $\mu_D Q(t)$ from the initially transferred fibres $X(0)$.

Finally, the variables $X(t)$, S and $Y(t)$ construct a model for fibre retrieval. They present the following distributions for the mentioned variables in this stage:

$$S \sim Beta(\alpha_s, \beta_s), \qquad Y(t) \sim Bin(X(t), S)$$

where S models the proportion of fibres that are retrieved in the laboratory. The variable $Y(t)$ counts the number of fibres that are actually retrieved from the offender's hair given that the proportion of recoverable fibres is S and that the number of fibres on the

offender's hair is $X(t)$.

The aim of this network is to compute the marginal distribution of $Y(t)$. Puch and Smith (2002) provide a software to calculate this distribution required in the likelihood ratio mentioned above.

There are some sources of uncertainties to compute the marginal distribution of $Y(t)$ which the likelihood ratio might be influenced by the changes of these uncertainties. These uncertainties are: the choice of hyper-parameters, the independence assumptions between parameters, the distributional assumptions considered for the variables and parameters.

To check whether the likelihood ratio might be influenced by changing these uncertainties, we can use the sensitivity measures introduced in this chapter. Unfortunately, the computation of these measures for this network is not feasible, and the numerical methods are required to calculate the marginal distribution of $Y(t)$ and the corresponding likelihood ratio with respect to perturbation of the qualitative and quantative assumptions mentioned above.

As we said above, the provided software enables user to compute the marginal distribution of $Y(t)$ for the specified hyperparameters. Therefore, the sensitivity analysis can be implemented by changing the values of those hyperparameters and assessing the changes in the marginal distribution of $Y(t)$ and the likelihood ratio by using this software. However, it most useful to do such sensitivity analysis within the context of the particular type of perturbations that one might find in the given practical situations.

7.4 The Sensitivity Analysis of the Bayesian Networks with Dependent Parameters

In most of the last section, we studied the local sensitivity analysis of the posterior quantities to perturbations of prior distributions for the Bayesian networks under local and global independence assumptions on the parameters. Now, let us consider Bayesian networks where parameters are not locally and globally independent of each other. In this situation, the dependent structure between parameters can be defined by hierarchical models. However, the dependency between two parameters can also be introduced using a *latent variable.* It should be noticed that the hierarchical models can be considered as a DAG (see Guihenneuc-Jouyaux et al (1998) and Kirby-Spiegelhalter (1994)).

When we introduce the hierarchical prior distribution for the Bayesian network with the dependent prior distributions, *identifiability* issues need to be addressed. We give some suggestions to overcome unidentifiability. Finally, sensitivity analysis of these Bayesian networks will be assessed.

7.4.1 Bayesian Identifiability for Hierarchical Models

In hierarchical models, stagewise specification often introduces random effects, yielding an overall parametric model of high dimension. Typically, for at least some of the parameters, there is a sense that the data provide little information (i.e., these parameters are weakly identified) and hence that the model is weakly identified. In particular, suppose that the Bayesian model is denoted by likelihood $L(\underline{\theta} \mid x)$, prior $\pi(\underline{\theta})$ and where $\underline{\theta}$ is partitioned as $\underline{\theta} = (\underline{\theta}_1, \underline{\phi})$. If

$$p(\underline{\phi} \mid \underline{\theta}_1, \underline{x}) = p(\underline{\phi} \mid \underline{\theta}_1) \tag{7.13}$$

then we say that $\underline{\phi}$ is not identifiable. This means that, if observing data $\underline{x}$ does not increase our prior knowledge about $\underline{\phi}$ given $\underline{\theta}_1$, then $\underline{\phi}$ is not identified by the data. Unidentifiability[8]occurs in the most rudimentary hierarchical specification, $p(\underline{x} \mid \underline{\theta}_1)p(\underline{\theta}_1 \mid \underline{\phi})p(\underline{\phi})$. In addition, because

$$p(\underline{\phi} \mid \underline{\theta}_1, \underline{x}) \propto L(\underline{\theta}_1, \underline{\phi} \mid \underline{x})p(\underline{\phi} \mid \underline{\theta}_1)p(\underline{\theta}_1) \tag{7.14}$$

$\underline{\phi}$ is not identifiable if and only if the likelihood is free of $\underline{\phi}$. Hence the formal definition of Bayesian unidentifiability (Dawid (1979)) is equivalent to a lack of identifiability in the likelihood. This hierarchical model is shown in Figure 7.6.

It should be noticed that the data $\underline{x}$ are conditionally uninformative for $\underline{\phi}$ given $\underline{\theta}_1$ (Since Equation (7.13) holds for this hierarchical model.).

Settimi and Smith (1999, 2002) showed that for the graph given in Figure 7.7, if $\theta_1 \perp\!\!\!\perp \theta_2 \mid \phi$, and there is only knowledge on the margin (θ_1, θ_2) obtained by data, $\underline{x} = (x_1, x_2)$, then the corresponding model for the probabilities is unidentifiable[9] (see Settimi and Smith (2000), Croft and Smith (2003) for more details and more examples of unidentifiable systems).

The important message here is that considerable caution should be considered in analysing data, which contain the systematically missing observation, by hierarchical models.

---

[8]More formally, a subset Θ_0 of the parameter space Θ is unidentifiable for some data $\mathcal{D}$ if for all parameters θ and θ' in Θ_0 the probability distributions are such that $p(\mathcal{D} \mid \theta) = p(\mathcal{D} \mid \theta')$. Therefore, the posterior inference on the unidentifiable parameters will be extremely dependent on the prior settings in the model as it is shown above.

[9]Note that, in the paper by Settimi and Smith (2000), θ_1 and ϕ are discrete. But the phenomena still holds if they are continuous.

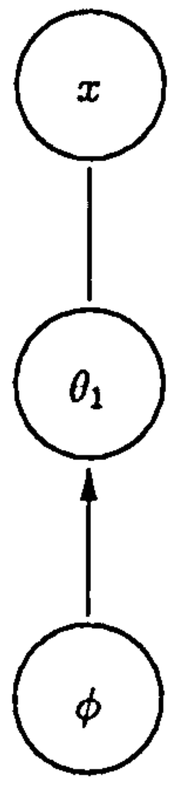


Figure 7.6: In the hierarchical model with the structure shown above, ϕ is not identifiable.

In 2002, Whiley and Titterington consider the identifiability of the *naive* Bayesian network with a binary, unobservable, root node and binary observable nodes. They define a model is identifiable if and only if the matrix associated with the transformation between the model parameters and the parameters of the observable variables is of full rank[10]. They show that this matrix is of full rank for the naive Bayesian network with a binary, unobservable, root node and 3 observable binary node. However, the parameter space for this model also contains a number of non-identifiable models (See Whiley and Titterington (2002) for some examples of these models.).

To make sure to have an identifiable model, one could impose the order (restriction on

[10]This actually is on agreement with the result represented by Geiger an Heckerman (1997) concerned with the characterisation of Dirichlet distribution

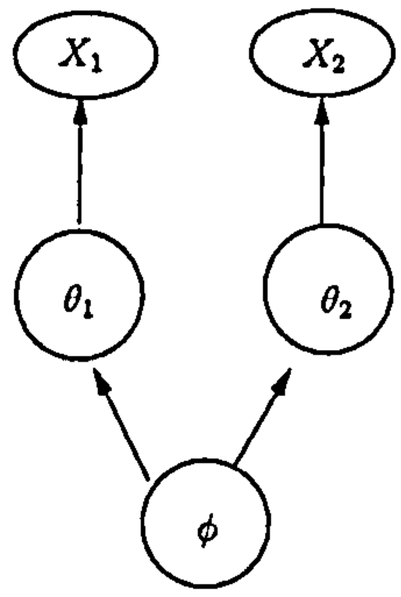


Figure 7.7: The hierarchical model with unidentified parameters mentioned above.

the order) on the hyperparameters associated with the variable (or node) that make the model identifiable (for example, Richardson and Green (1997) imposed the restriction on the order of the eigenvalues of the covariance matrix for the Gaussian Regression models.). It should be noticed that this approach is recommended for the models with Gaussian distributions.

Now, let us consider the Bayesian network with hierarchical prior with three stages as is shown in Figure 7.8.

The stages of this hierarchical prior are: $\theta_i \mid \phi_i$, $\phi_i \mid \psi$ and ψ as the known stage. The relationships between parameters mentioned above can be considered as

$$\theta_1 = \phi_1 + \epsilon_1, \quad \theta_2 = \alpha_1\phi_1 + \alpha_2\phi_2 + \epsilon_2 \quad \theta_3 = \phi_2 + \epsilon_3,$$

$$\phi_i = \psi + \epsilon_i', \quad i = 1, 2$$

where $\alpha_1, \alpha_2 \geq 0$ and $\alpha_1 + \alpha_2 = 1$.

Let assume the normal distributions for the variables above as follows,

$$\epsilon_i \sim N(\underline{0}, \Sigma_i), \quad i = 1, 2, 3, \quad \epsilon_i' \sim N(\underline{0}, \Upsilon_i), \quad i = 1, 2,$$

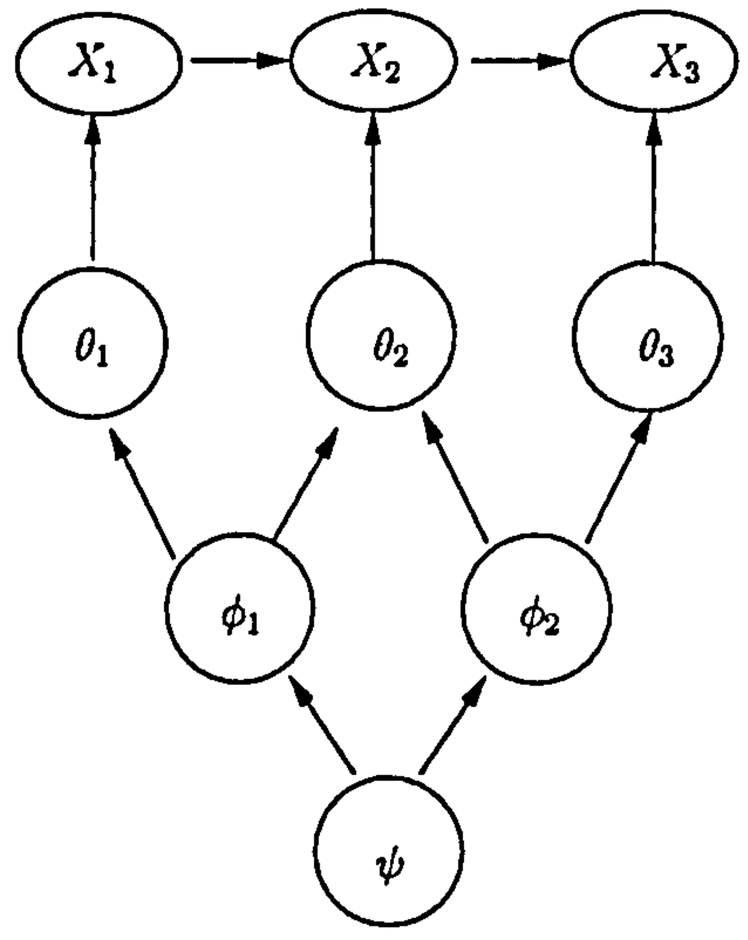


Figure 7.8: The representation of Bayesian network with three stages hierarchical prior distribution.

$$\theta_i \mid \phi_i \sim N(\phi_i, \Sigma_i), \quad i = 1,2,3, \quad \phi_i \mid \psi \sim N(\psi, \Upsilon_i), \quad i = 1,2 \quad \text{and} \quad \psi = \psi^*,$$

where Σ_i, Υ_i denote the covariance matrices of ϵ_i and ϵ_i' respectively, and ψ^* is a fixed value of ψ.

According to Lindley (1972, p. 46), "... unidentifiability causes no real difficulty in a Bayesian approach"; suitable proper priors (prior distributions) ensure proper posteriors for the model unknowns, hence model estimability[11]. Note that, in practice, ignorance or mathematical convenience often lead us to choose rather vague priors for at least some parameters (as Lindley and Smith (1972) considered vague priors on the parameters of the second stage in hierarchical model).

[11]It should be noticed that it also means that any inference will then be very sensitive to the prior specification used.

It should be noticed that unidentifiability causes no real difficulty in a Bayesian approach if we have multiple sample or repeated measurements in the sense that is discussed below.

Furthermore, note that we do not consider the identifiability as above. In fact, we believe that identifiability will be an issue on a Bayesian model if unidentifiable parameters can affect the distribution of future observables. Otherwise the ways we set the priors will not make a predictive difference.

Therefore, the posterior distribution associated with the parameters in the hierarchical model shown in Figure 7.8 in terms of one sample would not be identifiable. In this situation, we need a large number of samples (repeated measurements) of $d_i = (X_1^{(i)}, X_2^{(i)}, X_3^{(i)}; 1 \leq i \leq m)$, where m denote the number of repeated samples that is required to estimate or update $\underline{\theta}$ in identifiable way. In fact, when we have a large sample from a single population, we will find that we are immediately faced with identifiability problems. However, this issue can be solved by choosing several samples from different populations, all known with respect to the Bayesian networks with the same structure. We can formalise this discussion as follows. Let us consider the variables of the hierarchical model shown above as $\{x_{(i,j)} : 1 \leq i \leq n, 1 \leq j \leq m\}$, where i indexes observations of process j, and m different process (samples) are described by this model. According to each process, we have an estimation for $\underline{\theta}$, and hence the following set for m processes: $\{\underline{\hat{\theta}}(j) : 1 \leq j \leq m\}$. Then, we can get consistent estimate of $\underline{\theta}$ (or logistic form of that) from these m points. Furthermore, we can consistently estimate the distribution of $\underline{\theta}$, and hence the covariance matrix of $(\theta_1, \theta_2, \theta_3)$ given the observation collected from m processes mentioned above and fixed orientation of $\underline{\phi}$.

Now, let us define the hierarchical prior in the general form for the Bayesian network with k variables, $\underline{X} = (X_1, \ldots, X_k)$. If the parameters associated with these variables

are not independent, we can consider the hierarchical prior distribution with at most k-stages as

$$p(\theta_1,\ldots,\theta_k) = \int (\prod_{i=1}^{k} p(\theta_i \mid \underline{\theta}_i^{(1)})) \times \prod_{i=1}^{k-1} p(\theta_i^{(1)} \mid \underline{\theta}_i^{(2)}) \times \ldots \times \prod_{i=1}^{k-j} p(\theta_i^{(j)} \mid \underline{\theta}_i^{(j-1)}) \ldots \times p(\theta^{(k)})$$

where $\underline{\theta}_i^{(j)} \subseteq (\theta_1^{(j)},\ldots,\theta_{k-j}^{(j)})$ can be considered as the parents of the parameters at the j^{th} level, $\underline{\theta}_l^{(j)} \cap \underline{\theta}_m^{(j)} \neq \emptyset$ for $l \neq m$, $1 \leq l,m \leq k$ and $\theta^{(k)}$ denote the last and known stage. In fact, if

$$p(\underline{x},\underline{\theta}) = \prod_{i=1} p(x_i \mid pa(x_i),\underline{\theta}) \times p(\underline{\theta})$$

then, the prior $p(\underline{\theta})$ can be factorised as the hierarchical model in the following decomposition:

$$\forall \theta_i \in \underline{\theta},\ \theta_i \sim p_1(\theta_i \mid \underline{\theta}_i^{(1)}), \underline{\theta}_i^{(1)} \sim p_2(\underline{\theta}_i^{(1)} \mid \underline{\theta}_i^{(2)}),\ldots,\underline{\theta}_i^{(k-1)} \sim p_k(\underline{\theta}_i^{(k-1)} \mid \underline{\theta}_i^{(k)}), \underline{\theta}_i^{(k)} \sim p_{k+1}(\underline{\theta}_i^{(k)}) \tag{7.15}$$

where $\underline{\theta} = (\theta_1,\ldots,\theta_k)$.

Note that, the hierarchical prior above has some advantages. First, if the hyperparameters $\underline{\theta}_i^{(1)},\ldots,\underline{\theta}_i^{(k)}$ are of no interest for the inference (about $\underline{\theta}$), it is equivalent to consider the simpler hierarchical model with two stages as follows

$$\underline{\theta} \mid \underline{\theta}_i^{(1)} \sim p(\underline{\theta} \mid \underline{\theta}_i^{(1)})$$

and

$$\underline{\theta}_i^{(1)} \sim p(\underline{\theta}_i^{(1)}) = \int \prod_{i=1}^{k-2} p(\theta_i^{(1)}) \mid \underline{\theta}_i^{(2)}) \times \ldots \times \prod_{i=1}^{k-j} p(\theta_i^{(j)} \mid \underline{\theta}_i^{(j-1)}) \ldots \times p(\theta^{(k)}).$$

The second advantage of hierarchical models that makes the computation of the corresponding estimators easier is given by the following lemma that is originally introduced

by Robert (2001):

Lemma 7.3 For the hierarchical model given by

$$x \sim p(x \mid \theta), \quad \theta \sim p_1(\theta \mid \theta_1), \ldots, \quad \theta_n \sim p_{n+1}(\theta_n),$$

the full conditional of θ_i given x and the θ_j's $(j \neq i)$ satisfies

$$p(\theta_i \mid x, , \theta, \theta_1, \ldots, \theta_k) = p(\theta_i \mid \theta_{i-1}, \theta_{i+1}).$$

It means that the conditional distributions in a hierarchical model only involve by local hyperparameters. So, this lemma could be useful to study the sensitivity analysis of the Bayesian networks with the dependent parameters.

By using the *local Markov property* on the moral graph[12] $\mathcal{D}^m$, we can say that

$$\alpha \perp\!\!\!\perp V \setminus \alpha \mid bl(\alpha) \tag{7.16}$$

where $bl(\alpha)$ is the so-called *Markov blanket*[13] of α. It can be found directly from the original DAG $\mathcal{D}$ as the set of α's parents, children, and children's parents:

$$bl(\alpha) = pa(\alpha) \cup ch(\alpha) \cup \{\beta : ch(\beta) \cap ch(\alpha) \neq \emptyset\}$$

The same lemma can be presented for the Bayesian network with hierarchical prior as follows,

Lemma 7.4 The conditional distribution of $\theta_i^{(l)}$ given $\underline{x}$ and the $\theta_j^{(m)}$'s $(j \neq i)$ associated with the Bayesian network described above with the hierarchical prior distribution is given by

$$p(\theta_i^{(l)} \mid \underline{x}, \theta_1, \ldots, \theta_k, \theta_1^{(1)}, \ldots, \theta_1^{(k-1)}, \ldots, \theta_{k-1}^{(1)}, \theta_{k-1}^{(2)}, \theta_k^k) = p(\theta_i^{(l)} \mid bl(\theta_i^{(l)}))$$

[12]The moral graph is obtained from the original DAG $\mathcal{D}$, by adding undirected edges between all pairs of parents of each vertex which are not already joined, and then making all edges undirected.

[13]The Markov blanket is the set of neighbours of α in the moral graph $\mathcal{D}^m$.

Proof. As is discussed above, we can conclude that

$$\theta_i^{(l)} \perp\!\!\!\perp (\underline{x}, \theta_1, \ldots, \theta_k, \theta_1^{(1)}, \ldots, \theta_1^{(k-1)}, \ldots, \theta_{k-1}^{(1)}, \theta_{k-1}^{(2)}, \theta_k^k) \setminus \theta_i^{(l)} \quad | \, bl(\theta_i^l),$$

Therefore, by using the following conditional independence statement

$$A \perp\!\!\!\perp B \; | \, C \Leftrightarrow Pr(A \mid B, C) = Pr(A \mid C)$$

the proof is trivial.

Example 7.7 Consider the following Bayesian network, shown in Figure 7.9, with the hierarchical prior distribution

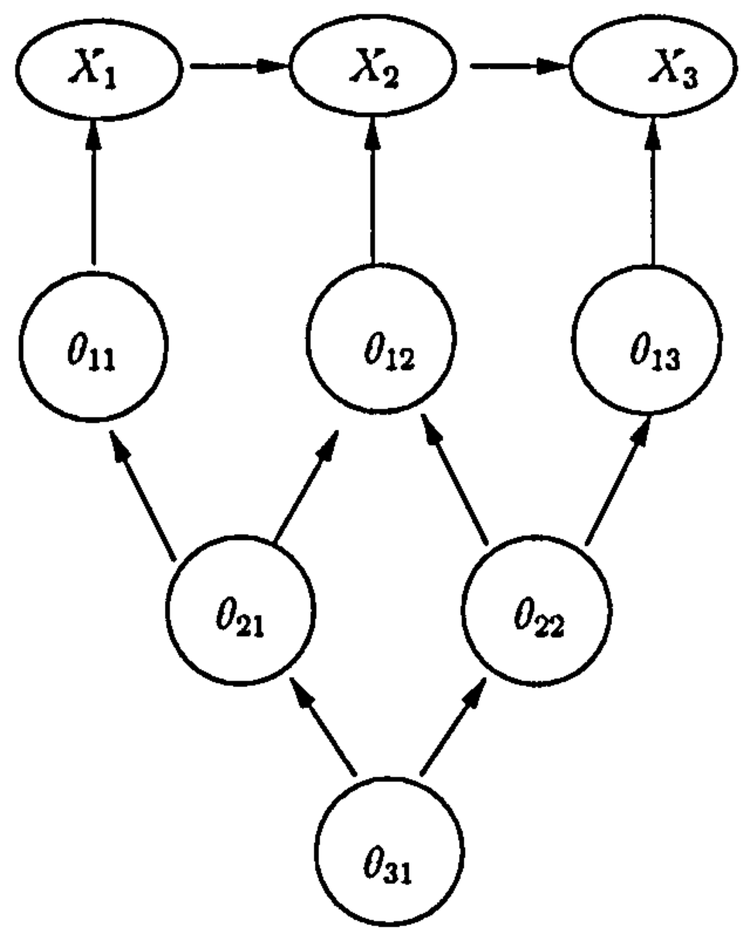


Figure 7.9: The Bayesian network representation with to the hierarchical prior distribution.

According to Lemma 7.4, the posterior quantities of θ_{12} are influenced by the neighbours of θ_{12}, $\{X_1, X_2, \theta_{21}, \theta_{22}\}$, determined in the corresponding moral graph and shown in Figure 7.10.

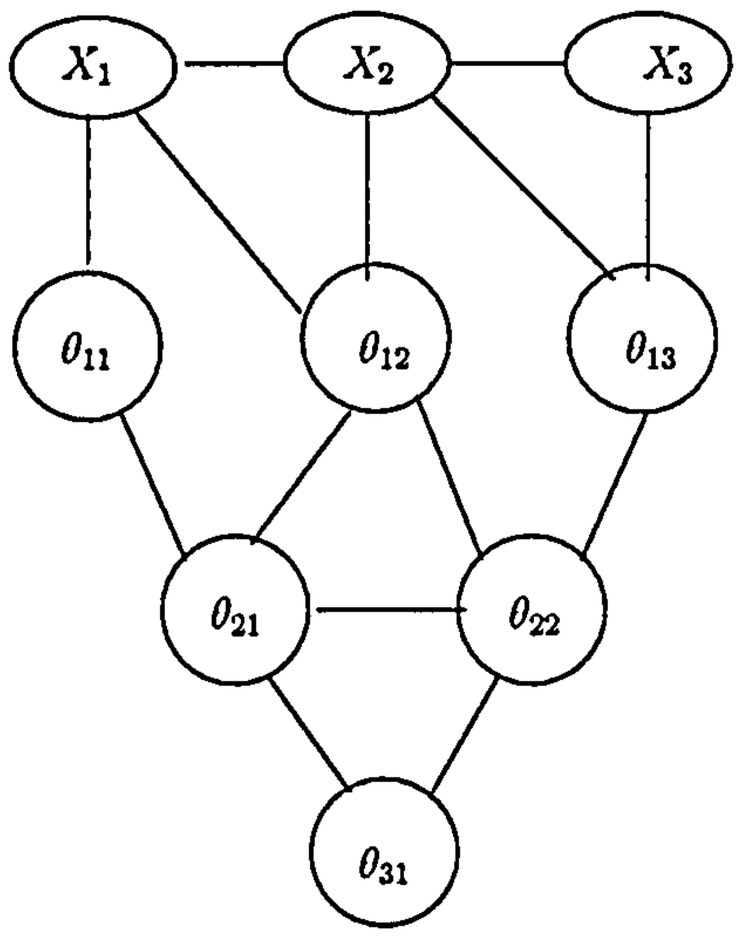


Figure 7.10: The moral graph corresponding the Bayesian network represented in Figure 7.9.

Therefore, we can conclude that the posterior quantities of θ_{12} could be sensitive with respect to small perturbations of θ_{12} neighbours' distributions, and are not influenced by the rest of variables. Let us write the parameters involved in this hierarchical prior as θ_{ij}, $i = 1, \ldots, k$, $j = 1, \ldots, (k - i + 1)$, where k denotes the number of nodes. By this notation, we can say that the local sensitivity measure of the posterior quantities of θ_{ij} is influenced by small perturbations of the prior distributions associated with $\{\theta_{(i-1)j}, \theta_{(i-1)(j+1)}, \theta_{i(j-1)}, \theta_{i(j+1)}, \theta_{(i+1)(j-1)}, \theta_{(i+1)j}\}$.

To compute the local sensitivity measure with respect to those perturbations mentioned above, the following measure that is similar to the one used by Gustafson (1996b) will be considered. This measure for the posterior distribution of θ_{ij} with respect to the small perturbation on the prior distribution of, for example, $\theta_{i(j-1)}$ is given by

$$S_{i\{j,(j-1)\}}(\underline{x}) = \lim_{\epsilon \to 0} \frac{H^2(p(\theta_{ij} \mid \underline{x}), p_\epsilon(\theta_{ij} \mid \underline{x}))}{H^2(p(\theta_{ij}), p_\epsilon(\theta_{ij}))}$$

where $p_\epsilon(\theta_{ij} \mid \underline{x})$ denote the posterior distribution of θ_{ij} associated with the perturbed prior distribution, $p_\epsilon(\theta_{i(j-1)})$.

As an example, let us consider the local sensitivity measure of the posterior distribution of θ_{12} with respect to perturbation on the prior distribution of θ_{21}. The Hellinger distance required to calculate this measure is given by

$$H^2(p(\theta_{12} \mid \theta_{21}, \theta_{22}, x_1, x_2), p_\epsilon^{\theta_{21}}(\theta_{12} \mid \theta_{21}, \theta_{22}, x_1, x_2)) =$$

$$2[1 - \int p(\theta_{12} \mid \theta_{21}, \theta_{22}, x_1, x_2)^{\frac{1}{2}} p_\epsilon^{\theta_{21}}(\theta_{12} \mid \theta_{21}, \theta_{22}, x_1, x_2)^{\frac{1}{2}}]$$

The integral part of equation above is calculated as

$$I = \int p(\theta_{12} \mid \theta_{21}, \theta_{22}, x_1, x_2)^{\frac{1}{2}} p_\epsilon^{\theta_{21}}(\theta_{12} \mid \theta_{21}, \theta_{22}, x_1, x_2)^{\frac{1}{2}} = \int \frac{A}{B_1^{\frac{1}{2}} B_2^{\frac{1}{2}}}$$

where

$$A = p(x_1)p(x_2 \mid x_1, \theta_{12})p(\theta_{12} \mid \theta_{21}, \theta_{22})p^{\frac{1}{2}}(\theta_{21})p_\epsilon^{\frac{1}{2}}(\theta_{21})p(\theta_{22}),$$

$$B_1 = p(x_1)p(x_2 \mid x_1, \theta_{12})p(\theta_{12} \mid \theta_{21}, \theta_{22})p(\theta_{21})p(\theta_{22}),$$

and

$$B_2 = p(x_1)p(x_2 \mid x_1, \theta_{12})p(\theta_{12} \mid \theta_{21}, \theta_{22})p_\epsilon(\theta_{21})p(\theta_{22})$$

The integral above is usually hard to calculate in the analytical way (even if the conjugate priors are chosen.). Therefore, the numerical methods such as MCMC can be

implemented to calculate these integrals.

7.5 The Relationship Between a Manipulated Bayesian network and Sensitivity Measures

In this section, we will study the possible relationship between a local cause that is defined in the causal Bayesian networks by forcing a node or a set of nodes to get specific values and the local perturbation that we use to study the local sensitivity analysis throughout this chapter. In the other words, we want to answer the following question: Could the *local cause* be defined in terms of the local sensitivity for the given graphical model?

Before we mention the general result, it is instructive to start with examining a simple Bayesian network with two binary variables as shown in Figure 7.11.

In Section 7.2, we show that the local sensitivity measure of the posterior expectation of

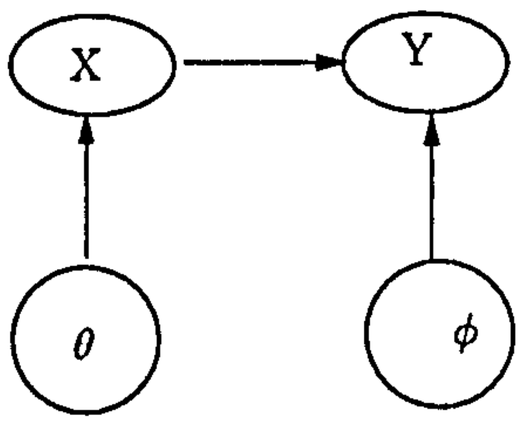


Figure 7.11: The representation of the Bayesian network with the discrete domain and its Markov equivalent.

$g(\theta)$ with respect to the small perturbation to the prior distribution of ϕ is zero, subject to existence of global parameter independence. As we have learned (see Chapter 4), if a Bayesian network is causal, then the parameters associated with this Bayesian network must exhibit the local and global independence parameters. Therefore, the result obtained above for the causal Bayesian network is valid.

The joint distribution of the causal Bayesian network shown in Figure 7.13 is factorised as

$$p(x, y, \theta, \phi) = p(x \mid \theta)p(y \mid x, \phi)p(\theta)p(\phi)$$

Now, if we $do\{\phi = \hat{\phi}\}$, then the representation above becomes

$$p(x, y, \theta, \phi \mid do\{\phi = \hat{\phi}\}) = p(x \mid \theta)p(y \mid x, \hat{\phi})p(\theta)$$

Therefore, we can conclude that

$$E(g(\theta) \mid \varepsilon = (x, y), do\{\phi = \hat{\phi}\}) = E^{p_\epsilon(\phi)}(g(\theta) \mid \varepsilon) = E^{p(\phi)}(g(\theta) \mid \varepsilon)$$

That means, if the parameters in the causal Bayesian network are globally independent, then the inference in term of the posterior expectation of $g(\theta)$ with respect to small perturbation to $p(\phi)$ is robust. Equivalently, by the assumption above, doing $\phi = \hat{\phi}$ will not affect distribution of θ, and therefore, it could be concluded that the local sensitivity of posterior expectation of $g(\theta)$ with respect to small perturbation to the distribution of ϕ is zero. It should be noticed that doing $\phi = \hat{\phi}$ can be expressed as the local perturbation to $p(\phi)$. Therefore, we can say that

$$\theta \perp\!\!\!\perp \phi \Rightarrow \|\dot{T}^g(p(\phi))\| = 0 \Leftrightarrow do\{\phi = \hat{\phi}\} \text{ does not affect the distribution of } \theta$$

This motivates the following theorem.

Theorem 7.1 If the parameters in the causal Bayesian network are globally independent, then, doing $(\theta_i = \hat{\theta}_i)$ does not have any effect on θ_j, $j \neq i$, and equivalently, the local sensitivity of the posterior quantities of the function of θ_j $(j \neq i)$ with respect to small perturbation to θ_i is equal to zero.

Proof The joint distribution of $\underline{X} = (X_v : v \in V)$ in the causal Bayesian network which the parameters are a priori independent of each other is given by

$$p(\underline{x}, \underline{\theta}) = \prod_{v \in V} p(x_v \mid pa_{x_v}, \theta_v) p(\theta_v)$$

Thus

$$p(\underline{x}, \underline{\theta} \mid do\{\theta_i = \hat{\theta}_i\}) = \prod_{v \in V \setminus \{i\}} [p(x_v \mid pa_{x_v}, \theta_v) p(\theta_v)] \times p(x_i \mid pa_{x_i}, \hat{\theta}_i)$$

and

$$\hat{\mu} = E(g(\theta_j) \mid \underline{X} = x, do\{\theta_i = \hat{\theta}_i\}) =$$

$$\frac{\int p(x_j \mid pa(j), \theta_j) p(\theta_j) g(\theta_j) d\theta_j}{\int p(x_j \mid pa(j), \theta_j) p(\theta_j) d\theta_j} = E(g(\theta_j) \mid X = x) = \mu$$

The local sensitivity of the posterior expectation of $g(\theta_j)$ with respect to small perturbation to $p(\theta_i)$ (e.g; $p_\epsilon(\theta_i) = (1 - \epsilon)p(\theta_i) + \epsilon q(\theta_i)$) is given by the following norm

$$\|\dot{T}^g(p(\phi))\| = [\int \{E^x(g(\theta_j) - \mu \mid \theta_i = z) \frac{p_{\theta_i}(z \mid x)}{p_{\theta_i}(z)}\}^2 p_{\theta_i}(z)]^{\frac{1}{2}}$$

Since the existence of global parameters independence, the norm above will be equal to zero, that is, $\|\dot{T}^g(p(\phi))\| = 0$.

Note that, the same result can be obtained when the roles of θ_i and θ_j are reversed.

These results motivate us to explore the causal Bayesian networks with weak form of dependency between parameters which is called, *approximate causal Bayesian networks*, and the causal relationships between the corresponding variables called *approximate causality*.

As we showed above the parameters in a causal Bayesian network must exhibit local and global independence. However, we speculate that there could be systems (particularly, Bayesian networks) with some sort of weak dependency between their parameters which under some external manipulation, this dependency could be destroyed or could be modelled in terms of combination of prior distributions with independence assumptions. We could explore this dependency between parameters by the local sensitivity measures mentioned above.

This is an exciting possible elaboration of the idea of causality which can be universally applied to any given prior distribution. However, results associated with such systems are beyond the scope this thesis and will be studied later.

7.6 Asymptotic Behaviour of the Specific Local Sensitivity Measure

It is very natural to expect that if the sample size becomes large enough, then the local sensitivity measures introduced in this chapter must become very small or more precisely tend to zero.

Many authors including Gustafson (1994, 1996a), Gustafson et al (1996), Gustafson and Wasserman (1995) showed that this is generally not true (in the next chapter, we discuss this topic in more details). However, Sivaganesan (1996) claimed that, for a specific form of local sensitivity measure, and for the class of priors which satisfy some mild conditions, local sensitivity measures (not as general as the local sensitivity measures introduced in this chapter) converge to zero asymptotically. This result is presented in this section, and we will apply this approach to study the asymptotic behaviour of the local sensitivity measure for very special classes of Bayesian networks.

It should be noticed that the main aim of this section is to study the asymptotic behaviour of the local sensitivity measure as an example. One can realise that this measure is very specific and shows quite good behaviour for some specific classes of prior distributions under some strong conditions.

In the next chapter, we present a more general approach by defining a new class of metrics called *credibility metrics* whose asymptotic behaviour are very suitable and when they are used as local sensitivity measures, we do not need to restrict the classes of prior distributions or assume some strong conditions as the other authors including Sivaganesan (1996) did.

Now, let $g(\underline{\theta})$ be a function of interest, which is assumed to be differentiable. As described above, the local sensitivity measure for the purpose above can be written as

$$S(p,q) = \lim_{\epsilon \to 0} \frac{d_2(E_{q_\epsilon}(g(\underline{\theta}) \mid \underline{x}), E_p(g(\underline{\theta}) \mid \underline{x}))}{d_1(q_\epsilon(\underline{\theta}), p(\underline{\theta}))} \tag{7.17}$$

where d_1 and d_2 stand for arbitrary distances between priors and posterior quantities, and $q_\epsilon(\underline{\theta})$ is considered as a perturbed density in the following class of prior distribution.

$$\Gamma = \{q_\epsilon(\underline{\theta}) = (1-\epsilon)p(\underline{\theta}) + \epsilon q(\underline{\theta}); q \in \mathbf{Q}\}$$

where Γ is called ϵ-contamination class of prior distributions, and $\mathbf{Q}$ denote the perturbed class of densities $q(\underline{\theta})$ that is defined over the support of $\underline{\theta}$.

As a diagnostic the supremum of the quantity in Equation (7.17) might be used over a perturbed class of prior densities. If both d_1 and d_2 are total variation distances, then under some regularity conditions,

$$S(p,\mathbf{Q}) = \sup_{q \in \mathbf{Q}} S(p,q) = \frac{\sup_{q \in \mathbf{Q}} \int p(\underline{x} \mid \underline{\theta}) |g(\underline{\theta}) - E_p(g(\underline{\theta} \mid \underline{x}))| q(\underline{\theta}) d\underline{\theta}}{m_p(\underline{x})} \tag{7.18}$$

where $m_p(\underline{x}) = \int p(\underline{x} \mid \underline{\theta}) p(\underline{\theta}) d\underline{\theta}$.

The measure above over all priors, denoted by $\mathbf{Q}_A$, becomes

$$S(p, \mathbf{Q}_A) = \sup_{q \in \mathbf{Q}} S(p, q) = \frac{\sup_{\theta} \{p(\underline{x} \mid \underline{\theta}) \{|g(\underline{\theta}) - E_p(g(\underline{\theta} \mid \underline{x}))|\}\}}{m_p(\underline{x})} \tag{7.19}$$

Gustafson et al (1996) show that, under mild regularity conditions, the local sensitivity measure introduced by Equation (7.17) increases at rate $n^{\frac{k}{2}}$ where k is the dimension of the parameter space. This surprising result is discussed further in Section 8.2. Thus, if we used this as a diagnostic we would conclude that the posterior becomes increasingly sensitive to the posterior as the sample size increase.

Sivaganesan (1996) studied the asymptotic behaviour of the specific form of certain local sensitivity measure (in fact, he considered the local sensitivity measure with the linear derivative) given in Equation (7.18). He shows that this measure converges to zero under some conditions. He considered an uniformly bounded class, $\mathbf{Q}$, of densities, $q(\underline{\theta})$, on the support of parameter space. That is, there is a $M < \infty$ such that $\sup_{\underline{\theta}} q(\underline{\theta}) < M$ for all $q \in \mathbf{Q}$. The regularity conditions required for the convergence of the posterior quantities with respect to each $q_\epsilon \in \Gamma$ are assumed to be valid. In particular, it is assumed that $p(\underline{\theta})$ is bounded, has bounded continuous derivatives, and $p(\underline{\theta}_0) > 0$ (most of these conditions are not generally valid). Furthermore, it is assumed that $\underline{\theta} \to \underline{\theta}_0$, and

$$\hat{I} = -\frac{1}{n} l''(\hat{\underline{\theta}}) \to I(\underline{\theta}_0),$$

where $\hat{\underline{\theta}}$ denote the maximum likelihood estimator, $\underline{\theta}_0$ is the true value of $\underline{\theta}$, $I(.)$ denote the expected Fisher information matrix, $l(.) = \log p(. \mid \underline{\theta})$ and

$$l''(\hat{\underline{\theta}}) = \left(\frac{\partial^2 l(\underline{\theta})}{\partial \theta_i \partial \theta_j}\Big|_{\underline{\theta}=\hat{\underline{\theta}}}\right).$$

Then, he shows that if $q \in Q$ have uniformly bounded densities (note that the class of priors considered by Gustafson et al (1996) was unbounded), $S(p, \mathbf{Q})$ converges at rate

$n^{-\frac{1}{2}}$. When $q \in \mathbf{Q}$ has uniformly bounded densities with uniformly bounded derivatives, $S(p, \mathbf{Q})$ is convergent at rate n^{-1}.

Furthermore, let $\mathbf{Q}$ be given by

$$\mathbf{Q} = \{q : \ L(\underline{\theta}) \leq q(\underline{\theta}) \leq U(\underline{\theta})\}$$

where L and U are bounded and continuous. Then

$$\sqrt{n}S(p, \mathbf{Q}) \rightarrow \sqrt{\frac{I(\underline{\theta}_0)}{2\pi}} \frac{[U(\underline{\theta}_0) - L(\underline{\theta}_0)]}{p(\underline{\theta}_0)}.$$

Example 7.8 Consider the Bayesian network shown in Figure 7.2 and studied earlier. Let consider the baseline prior $p(\underline{\mu})$ as follows

$$p(\underline{\mu}) = p(\theta) \prod_{i=1}^{3} p(\phi_i)p(\psi_i) \tag{7.20}$$

where $p(\theta) = \mathcal{D}(\alpha_1, \alpha_2, \alpha_3)$, $p(\phi_i) = \mathcal{D}(\lambda_{1i}, \lambda_{2i}, \lambda_{3i})$, $i = 1, 2, 3$, and $p(\psi_i) = \mathcal{D}(\tau_{1i}, \tau_{2i}, \tau_{3i})$, $i = 1, 2, 3$.

We want to assess the local sensitivity measure of the posterior expectation of $g(\underline{\mu}) = \theta_2\phi_{13}\psi_{23}$ with respect to small perturbation of $p(\phi_1) = P(\phi_{11}, \phi_{12}, \phi_{31})$. The following class of perturbed priors, Q, will be considered,

$$\mathbf{Q}_{\mathcal{D}} = \{q : q(\underline{\mu}) = \mathcal{D}(\alpha_1, \alpha_2, \alpha_3) \prod_{i=2}^{3} \mathcal{D}(\lambda_{1i}, \lambda_{2i}, \lambda_{3i}) \prod_{i=1}^{3} \mathcal{D}(\tau_{1i}, \tau_{2i}, \tau_{3i})\mathcal{D}(\beta_{11}, \beta_{12}, \beta_{13});$$

$$\beta_{1i} \geq 1, \lambda_{1i}, \lambda_{2i}, \lambda_{3i} \geq 1, \quad i = 1, 2, 3\}$$

It can be shown that every density in the class above is bounded with the bounded continuous derivatives.

The local sensitivity measure for the purpose above is given by

$$S(p, \mathbf{Q}_{\mathcal{D}}) = \frac{\sup_{q \in \mathbf{Q}_{\mathcal{D}}} \int p(\underline{x} \mid \underline{\mu})(g(\underline{\mu}) - E_p(g(\underline{\mu} \mid \underline{x})))q(\underline{\mu})d\underline{\mu}}{m_p(\underline{x})}$$

where

$$p(\underline{x} \mid \underline{\mu}) = \frac{\Gamma(n)}{\prod_{i,j,l} \Gamma(x_{ijk})} \prod_{i=1}^{3} \theta_i^{x_{i..}} \prod_{i,j=1}^{3} \phi_{ij}^{x_{ij.}} \prod_{j,l=1}^{3} \psi_{jl}^{x_{.jl}}$$

where $x_{i..} = \Sigma_{j=1}^{3}\Sigma_{l=1}^{3} x_{ijk}$, $x_{ij.} = \Sigma_{l=1}^{3} x_{ijl}$, $x_{.jl} = \Sigma_{i=1}^{3} x_{ijl}$, $n = \Sigma_{i=1}^{3}\Sigma_{j=1}^{3}\Sigma_{l=1}^{3} x_{ijl}$, and

$$m_p(\underline{x}) = \frac{\Gamma(\alpha)\Gamma(n)\prod_{i=1}^{3}\Gamma(\alpha_i + x_{i..})}{\Gamma(\alpha+n)\prod_{i=1}^{3}\Gamma(\alpha_i)\Gamma(x_{i..})} \times \frac{\prod_{i=1}^{3}\prod_{j=1}^{3}\Gamma(\lambda_j)\Gamma(x_{.j.})\Gamma(\lambda_{ij}+x_{ij.})}{\prod_{i=1}^{3}\prod_{j=1}^{3}\Gamma(\lambda_{ij})\Gamma(\lambda_j + x_{.j.})\Gamma(x_{ij.})} \times$$

$$\frac{\prod_{j=1}^{3}\prod_{l=1}^{3}\Gamma(\tau_l)\Gamma(x_{..l})\Gamma(\tau_{jl}+x_{.jl})}{\prod_{j=1}^{3}\prod_{l=1}^{3}\Gamma(\tau_{jl})\Gamma(\tau_l + x_{..l})\Gamma(x_{.jl})}.$$

By using Stirling's approximation it can be shown that the local sensitivity above converges to zero at rate n^{-k}. Where k denote the dimension of parameter space and n stands for the sample size. In the model presented in this example, $k = 14$.

Another perturbed class that can be consider is given below:

$$\mathbf{Q}_{DR} = \{q : \quad L(\underline{\mu}) \leq q(\underline{\mu}) \leq U(\underline{\mu})\}$$

where

$$L(\underline{\mu}) = \mathcal{D}(\upsilon_1, \upsilon_2, \upsilon_3)\prod_{i=1}^{} \mathcal{D}(\omega_{1i}, \omega_{2i}, \omega_{3i})\mathcal{D}(\nu_{1i}, \nu_{2i}, \nu_{3i})$$

and

$$U(\underline{\mu}) = \mathcal{D}(\rho_1, \rho_2, \rho_3)\prod_{i=1}^{3} \mathcal{D}(\varrho_{1i}, \varrho_{2i}, \varrho_{3i})\mathcal{D}(\sigma_{1i}, \sigma_{2i}, \sigma_{3i})$$

where for all $i, j = 1, \ldots, 3$, $\upsilon_i \leq \rho_i$, $\omega_{ji} \leq \varrho_{ji}$ and $\nu_{ji} \leq \sigma_{ji}$.

Therefore

$$\sqrt{n}S(p, \mathbf{Q}_{DR}; \hat{\underline{\mu}}) \rightarrow \sqrt{\frac{I(\underline{\mu}_0)}{2\pi}}\frac{[U(\underline{\mu}_0) - L(\underline{\mu}_0)]}{p(\underline{\mu}_0)}$$

where $\underline{\mu}_0$ denote the true value of $\underline{\mu}$, and $\hat{\underline{\mu}} = (\hat{\theta}_i, \hat{\phi}_{ij}, \hat{\psi}_{jl})$ stands for the maximum likelihood estimations, where

$$\hat{\theta}_i = \frac{x_{i..}}{n}, \quad \hat{\phi}_{ij} = \frac{x_{ij.}}{x_{.j.}}, and \;\; \hat{\psi}_{jl} = \frac{x_{.jl}}{x_{..l}}$$

Note that, the asymptotic results above are obtained under very strong conditions on the perturbed class of priors. In fact, these results are valid for the restricted classes of priors. We give more general results in terms of new class of metric and under weaker

conditions on the prior distributions and likelihood function in the next chapter.

7.6.1 Discussion

Gustafson et al (1996) pointed out that the asymptotic behaviour of the local sensitivity measures introduced in this chapter was very surprising. These measures displayed very high sensitivity of the posterior to prior, despite the increase in the sample size. By using Sivaganesan's notation, they showed that there are (ϵ-contamination) classes such that the local sensitivity measure represented in Equation (7.18) does not tend to zero, as $n \to \infty$. The problem might have been the measures of sensitivity and/or the width of the class of the priors. But, Sivaganesan believed that the problem is in the classes of priors and it may be solved if we give up some priors. However, he removed point masses priors from the contaminating priors but, as pointed out in Gustafson et al (1996), that is not enough to claim that the local sensitivity measures converge to zero.

Ruggeri (1996) claimed that if Sivaganesan considered more regular contaminating prior distributions, i.e. asking for uniformly bounded derivatives, besides uniformly bounded densities[14], he would find a larger rate of convergence.

It seems reasonable to ask more requirements on the multi-dimensional priors which could be ended up with some specific classes of priors.

Thus, the asymptotic results represented above are obtained under very strong conditions on some contaminating priors and for a specific local sensitivity measure. We present more general results in terms of a new class of metrics and under weaker conditions on the prior distributions and likelihood function in the next chapter.

[14]By considering bounds on derivatives of higher orders, larger rates might be given.

Chapter 8

Bayesian Convergence and Sensitivity under Credibility Metrics

8.1 Introduction

The local sensitivity analysis, as studied in the last chapter, is recognised for its computational simplicity, and its potential use in multi-dimensional and similar complex problems where global robustness investigation may be difficult. The major drawback of this approach is about the asymptotic bahaviour. It is reasonable that in most cases the influence of prior distribution on the posterior quantities becomes less important as the sample size tends to infinity. Gustafson (1994) reported that for the most classes of priors considered in the literature, the global robustness measures such as size of ranges, tends to zero as the sample size goes to infinity. Therefore, it is reasonable to expect that the local sensitivity measures must converge to zero asymptotically.

Gustafson et al (1996) and Gustafson (1994) showed that for most general classes of

prior distributions the local sensitivity measures do not tend to zero, and even diverge for some multidimensional classes of priors.

To overcome this issue, Gustafson et al (1996) restricted the class of contaminated priors to a particular parametric family and imposed some mild conditions on each element of this new class of priors. This approach is useful for some classes of prior distributions.

Furthermore, as we discussed in the last chapter, Sivaganesan (1996) indicated that a specific local sensitivity measure for some classes of prior distributions which satisfy some mild conditions would go to zero as the sample size tends to infinity. However, he reported these results for one-dimensional parameter space, and the mild conditions are not satisfied for most classes of prior distributions (see Section 7.6 for further discussions).

In this chapter, we suggest defining the local sensitivity measures introduced in the last chapter in terms of a new class of metrics with which to examine prior to posterior convergence (LeCam and Yang (1990) and Schervish (1996)) and sensitivity issues (Gustafson et al (1996)) in a Bayesian model.
We then focus on the posterior predictive distribution instead of posterior distribution. Although the computation of a posterior predictive distribution is more difficult, we believe that using posterior predictive distribution instead of posterior distribution for the local sensitivity measures produces more stable results.

The methods are illustrated using estimated graphical models and various new asymptotic results are derived. In the first section of this chapter, we examine prior to

posterior convergence in terms of the total variation distance (and Hellinger distance). In Section 8.2, we show that the prior to posterior convergence in terms of the standard metrics introduced by Schervish (1996), and in terms of the Fréchet derivative used as a local sensitivity measure by Gustafson et al (1996) are not appropriate. Section 8.3 is dedicated to introducing a new and more general class of metrics called *credible metrics* whose asymptotic behaviour are very suitable and when used as a local sensitivity measure, does not require us to restrict ourselves to a special class of priors. We also study the asymptotic behaviour of this metric by looking at the predictive distributions in this section. We believe that more stability can be obtained by using predictive distributions. In Section 8.4, we study the asymptotic behaviour of the credibility metrics with respect to posterior distributions. Finally, we use these results to study the asymptotic behaviour of the local sensitivity measures derived for the several purposes in Bayesian networks.

8.2 Introduction to the Bayesian Convergency

In this section, we first examine some preliminary notations and concepts concerned with Bayesian convergence. We then examine the asymptotic behaviour of the local sensitivity measures studied in the last chapter.

Let $\mathcal{P}$ be the set of all probability measures on the parameter space and given a prior density p we denote $p(. \mid x)$ the corresponding posterior density represented as

$$p(\theta \mid x) = \frac{L(\theta \mid x)p(\theta)}{\int L(\theta \mid x)p(\theta)d\theta}.$$

Let $T : \mathcal{P} \rightarrow \mathcal{T}$ denote some quantity of interest. For example, we might take $T(P) = P(\theta \mid x)$ and $\mathcal{T} = \mathcal{P}$. In other words, the quantity of interest is the whole posterior distribution. Another example is $T_g(P) = \int g(\theta)p(\theta \mid x)d\theta$, the posterior expectation of $g(\theta)$. In this case $\mathcal{T}$ is the range of g. We might also be interested in the predictive

distribution of a new observation x. In this case $T(P)$ is a probability distribution with the following density

$$f(. \mid x) = \int f(. \mid \theta)p(\theta \mid x)d\theta. \tag{8.1}$$

In the last chapter, we examined the sensitivity of a prior P in the direction of another prior Q, using the sensitivity measure

$$S(P, Q) = \lim_{\epsilon \to 0} \frac{d_2(T(Q_\epsilon), T(P))}{d_1(Q_\epsilon, P)} \tag{8.2}$$

where Q_ϵ is the perturbed prior distribution and can be either linear or geometric perturbation of P, and d_1 and d_2 denote any distances such as total variation distance or Hellinger distance. Unless otherwise stated, throughout this chapter we will assume that $d_1(.,.)$ and $d_2(.,.)$ mentioned in Equation (8.2) are both the total variation metrics.

Note that if $\Gamma \subseteq \mathcal{P}$ then we can define $S(P, \Gamma) = \sup_{Q \in \Gamma} S(P, Q)$. We will study the asymptotic behaviour of this sensitivity measure under the credible metric which is defined in the next section.

In Chapter 7, it was shown that under mild regularity conditions $S(P, \mathcal{P})$ ($S(P, \Gamma)$ for many classes of Γ) increases at rate $n^{\frac{k}{2}}$, where k is the dimension of the parameter space. Therefore, if we use this quantity as a diagnostic we will conclude that the posterior becomes increasingly sensitive to the prior as the sample size becomes very larger.

This is because, $\mathcal{P}$ comprises many unreasonable priors (e.g., priors which put all the mass at one point, or priors that have very noisy behaviour at their tails), Gustafson et al (1996) initially conjectured that this issue would be solved by restricting the class of priors to a subset Γ of $\mathcal{P}$. But, they showed that this issue still remains as long as P is an interior point of Γ with respect to the density ratio metric that is introduced by DeRobertis (1978) and DeRobertis and Hartigan (1981) as follows

$$\delta(P, Q) = \log \sup_A \frac{P(A)Q(A^c)}{Q(A)P(A^c)}$$

where A can be any subset of the parameter space.

This is a very severe constraint on any prior family (precluding mixtures with distribution measures for example), but despite this, the type of divergence discussed by Gustafson and Wasserman (1995) will still occur under this prior family constraint.

Gustafson et al (1996) consider the parametric priors as the restricted class of priors, and show that the diagnostic measure under this class of priors produces better asymptotic behaviour (see Gustafson et al (1996) for details and examples). But this is rather unsatisfactory, because sensitivity then depends on a prior lying in a particular parametric family: which is exactly the sort of dependence we want to avoid.

They obtain the similar asymptotic behaviour when d_1 and d_2 are the ϕ-divergence distances, and the geometric perturbation is used.

8.3 A New Class of Metrics

In this section, we shall define a new class of metrics which exhibit better asymptotic behaviour in the study of Bayesian convergence and sensitivity analysis. For this purpose, we should first introduce some well known results for the total variation distance.

Let $d(P,Q)$ be a metric on probability distributions (P,Q) on a common $\sigma-$ algebra $\mathbb{C}$ on a sample space Ω. Unless otherwise stated throughout this paper we will assume that $d(.,.)$ is the (total) variation metric, so that

$$d(P,Q) = \sup_{C \epsilon \mathbb{C}} |P(C) - Q(C)| \tag{8.3}$$

The properties of the variation metric are well studied (Zolotarev (1983), Diaconis and Freedman (1986), Reiss (1989), LeCam and Yang (1990), Rachev (1991), Smith (1995), and Gibbs and Su (2002)). This is because of its intimate links with the distance between betting preferences which commonly define, at least implicitly, Bayesian prior

distributions. Two well known properties that we will use extensively in this section are:

1. If P and Q have the same dominating measure, with respective densities p and q then

$$d(P,Q) = \frac{1}{2}\int_{\theta\epsilon\Omega} |p(\theta) - q(\theta)|d\theta \tag{8.4}$$

2. The metric is invariant to transformations f in the following sense. If the transformation $f : \Omega \to \Omega', \theta \mapsto \theta'$ is bijective and measurable and $(P_\theta, Q_\theta), (P_{\theta'}, Q_{\theta'})$ are two probability measures on θ and $\theta' = f(\theta)$, then

$$d(P_\theta, Q_\theta) = d(P_{\theta'}, Q_{\theta'}) \tag{8.5}$$

There is also a well known result that, for a fixed known family of sample distributions, the variation distance between two predictive distributions is no larger than the distance between their prior distributions. For completeness we state this result as a lemma.

Lemma 8.1 If $P_{\theta,X}$ and $Q_{\theta,X}$ have respective densities, (with common conditional density $p(x|\theta)$) of $X \in \Omega_x$ and $\theta \in \Omega_\theta$,

$$p(\theta, x) = p(\theta)p(x|\theta) \tag{8.6}$$

$$q(\theta, x) = q(\theta)p(x|\theta) \tag{8.7}$$

where $p(\theta)$ is the density of P_θ, the margin of $P_{\theta,X}$ on θ and $q(\theta)$ is the density of Q_θ, the margin of $Q_{\theta,X}$ on θ then the total variation density $d(.,.)$ satisfies

$$d(P_X, Q_X) \leq d(P_\theta, Q_\theta) \tag{8.8}$$

Proof

$$d(P_{\theta,X}, Q_{\theta,X}) = \sup_{C\epsilon\mathbb{C}} |P_{\theta,X}(C) - Q_{\theta,X}(C)| \tag{8.9}$$

So trivially, taking $C = C_\theta \times \Omega_X$ and $C = \Omega_\theta \times C_X$

$$d(P_\theta, Q_\theta) \leq d(P_{\theta,X}, Q_{\theta,X}), \tag{8.10}$$

$$d(P_X, Q_X) \leq d(P_{\theta,X}, Q_{\theta,X}) \tag{8.11}$$

Finally note that

$$2|P_{\theta,X}(C) - Q_{\theta,X}(C)| = \int_{\theta,x\epsilon C} |p(\theta,x) - q(\theta,x)| d\theta dx$$

$$= \int_{\theta,x\epsilon C} |p(\theta) - q(\theta)| p(x|\theta) d\theta dx$$

$$= \int_{\theta\epsilon C'} |p(\theta) - q(\theta)| \int_{x\in\Omega_x} \{p(x|\theta)dx\} d\theta$$

where $C' = \{\theta : \theta \in \cup_{x\in\Omega_x} C \cap \{X = x\}\} \subseteq \Omega_\theta$

$$= \int_{\theta\epsilon C'} |p(\theta) - q(\theta)| d\theta = 2|P_\theta(C') - Q_\theta(C')|$$

So

$$\sup_{C\epsilon\mathbb{C}} |P_{\theta,X}(C) - Q_{\theta,X}(C)| \leq \sup_{C\epsilon\mathbb{C}} |P_{\theta,X}(C) - Q_{\theta,X}(C)|$$

The result follows.

Thus, in particular, if θ separates future observations from the past then by ensuring two posterior distributions are close is enough to ensures that two predictive distributions are also close.

Example 8.1 Suppose $X_1, X_2, ..., X_k,$ are independent identically distributed standard Gaussian $N(0,1)$ variables. Write $X^{(n)} = \{X_1, X_2, ..., X_n\}$. Note that $S_i(n) = n^{-\frac{1}{2}} \sum_{j=n(i-1)+1}^{ni} X_j$ is such that, for two different prior densities $p_j(\theta)$, $j = 1, 2$ on θ, then note that for all $n > 0$ and $j = 1, 2$ we have that

$$p_j(\theta|X^{(n)} = x^{(n)}) = p_j(\theta|S_1(n) = s_1(n))$$

We also have that

$$p_j(s_2(n)|x^{(n)}) = \int_{\theta\in\Omega_\theta} p_j(s_2(n)|\theta)p_j(\theta|x^{(n)})d\theta$$

$$= \int_{\theta\in\Omega_\theta} p_j(s_2(n)|\theta)p_j(\theta|s_1(n))d\theta$$

$$= \int_{\theta'\in\Omega_\theta} p_j(s_2(n)|\theta')p_j(\theta'|s_1(n))d\theta'$$

So since we have

$$p_j(s_2(n)|\theta') = p_j(s_2(1)|\theta) \sim N(\theta', 1)$$

$$p_j(s_2(n)|x^{(n)}) = \int_{\theta'\in\Omega_\theta} p_j(s_2(1)|\theta')p_j(\theta'|s_1(n))d\theta'$$

where $\theta' = n^{\frac{1}{2}}\theta$ (Note here we replace θ by $\theta' = n^{\frac{1}{2}}\theta$ when we calculate this integral, to get the exact analogue, but note that the variation distances between the priors of θ and θ' are the same because of scale invariance.). Hence, in one sense the problem of predictive densities does not appear to depend on the number of observations observed n. In particular

$$d(p_1(s_2(n)|x^{(n)}), p_2(s_2(n)|x^{(n)}))$$

does not depend on n. Since, from the above, for all $n > 0$

$$d(p_1(\theta|x^{(n)}), p_2(\theta|x^{(n)})) \geq d(p_1(s_2(n)|x^{(n)}), p_2(s_2(n)|x^{(n)}))$$

$$= d(p_1(x_2|x_1), p_2(x_2|x_1))$$

As Gustafson and Wasserman (1995) point out, this distance cannot converge as $n \to \infty$. This looks counterintuitive, since we know that, whatever the prior for θ, in this circumstance, given $x^{(n)}$, $n^{-\frac{1}{2}}(\theta - \overline{x})$ tends to a Gaussian $N(0,1)$ density, so the posterior densities are close to one another, spiking near $\overline{x}$. That is until we remember that the variation metric is scale invariant, and we need to see the difference between the posterior densities appropriately magnified up onto the region to which θ converges.

But if we have some way of fixing the scale of the deviation, then this is not so. For example,

$$d(p_1(x_{n+1}|x^{(n)}), p_2(x_{n+1}|x^{(n)})) = \int |p(x_{n+1}|\theta)\{p_1(\theta|x^{(n)}) - p_2(\theta|x^{(n)})\}|d\theta$$

$$\leq \int_{\theta \in B(\overline{x},\delta)} p(x_{n+1}|\theta)|p_1(\theta|x^{(n)}) - p_2(\theta|x^{(n)})|d\theta + \int_{\theta \notin B(\overline{x},\delta)} p(x_{n+1}|\theta)(p_1(\theta|x^{(n)}) + p_2(\theta|x^{(n)}))d\theta$$

$$\leq \sup\{p(x_{n+1}|\theta) : \theta \in B(\overline{x},\delta)\}\mu(B(\overline{x},\delta)) + \sup\{p(x_{n+1}|\theta) : \theta \notin B(\overline{x},\delta)\}2\eta(\delta)$$

$$< (2\pi)^{-1}\{\mu(B(\overline{x},\delta)) + 2\eta(\delta)\} \to 0 \quad \text{as} \quad n \to \infty$$

so one step ahead prediction of the next observation certainly converges.

It should be noticed that we could link these results into Bayesian predictive inference, if we consider the distribution of θ as a way of communication our betting preferences about future observables. These predictions will be stable if prediction about θ are stable. So stability in terms of this (and in many other metrics) is consistent with ideas about Bayesian Sufficiency etc. However, this work is under study, but using posterior predictive distributions or making inference in terms of Bayesian predictive systems have been supported by several researchers.

In the closest work to this thesis, Cowell (1996) investigated computation of compatible priors for the given network structure. Let P denote a probability model (for discrete variables) with distribution, $P(\underline{X} \mid \underline{\theta})$ and prior distribution, $P(\underline{\theta})$. For computational reasons, we may need to approximate $P(\underline{\theta})$ by another prior distribution, $Q(\underline{\theta})$, which is simpler to work with from a computational perspective. However, the approximated prior must capture as closely as possible the predictive properties of $P(\underline{\theta})$. That can be achieved by minimising the Kullback-Leibler divergence between the predictive distributions of P, Q (the similar approach has been suggested in terms of Hellinger distance in the work in progress: Smith and Daneshkhah (2004), where we are looking for an alternative model with equivalent network structure with the given Bayesian network such that the direction of some nodes are not causally matter for the enough large data

set). Then, this approach[1] was used for matching hyper-Dirichlet priors of conditional probabilities for Bayesian networks with discrete variables, where prior distribution for one network (P) is given, but only the structure of the other network (Q) is known.

Dawid (1997) argued that it is more reasonable to make inference in terms of predictive distribution about the quantities that can indeed be observed than in terms of statements about unknown parameters of probability distributions.

Suppose a sequence $\underline{X} = (X_1, X_2, \ldots)$ of uncertain quantities, in turn, can be observed. To make an inference about unknown parameters of the distribution of $\underline{X}$, a sequence of forecasts for the (X_i) can be easily made. After observing the values $\underline{x}_n = (x_1, \ldots, x_n)$ of $\underline{X}_n = (X_1, \ldots, X_n)$, we can make a forecast for the next quantity X_{n+1}. Then, we observe x_{n+1} and compare it with its forecast in an appropriate approach, and the whole process then repeated with $n+1$ replacing n. Dawid (1984) called this procedure, *prequential forecasting*. He showed that the prequential procedure is consistent and this consistency is attainable by using Bayesian forecasting system and statistical forecasting system (both systems are behaving very well for any value of uncertain parameters).

8.4 Credible Metrics Between Posterior Distributions

In this section, we study the asymptotic behaviour of the new metric called credible metricsintroduced above for posterior distributions. The credible metrics are more conventionally based on posterior distributions. But, these measures are more stable in the sense that they at least do not diverge as we obtain more data.

[1]He also used of matching moment and minimising the expected posterior of Kullback-Leibler divergence. This approach is also applied to reduce a mixture of conjugate priors to a smaller mixture (see Cowell (1996, 1998) for details).

First, we present some more notations and definition in this section. Then, we present some results which approve the claim above.

Let $P \mid A$ $[Q \mid A]$ denote the conditional probability of $P[Q]$ given an event $A \in \mathbb{C}$, $P(A) > 0$ and define

$$d_{A[p]}(P, Q) = d(P \mid A, Q \mid A) \tag{8.12}$$

Note that this is a pseudometric (i.e. all the metric axioms hold other than $d_{A[p]}(P, Q) = 0 \Rightarrow P = Q$).

Call a set $\mathbb{A}$ of events P-conditioning, if

$$\{\Omega\} \subseteq \mathbb{A} \subseteq \mathbb{C}^+[P]$$

where $\mathbb{C}^+[P] = \{C \in \mathbb{C} : P(C) > 0\}$.

For any P-conditioning set $\mathbb{A}$ denote

$$d_{\mathbb{A}}(P, Q) = \sup\{d_{A[p]}(P, Q) : A[P] \in \mathbb{A}\} \tag{8.13}$$

Finally denote by $\mathbb{P}(P)^+$ the set of probability measures with the same support as P.

Lemma 8.2 If $\mathbb{A}$ is P-conditioning then $d_{\mathbb{A}}(.,.)$ is a metric on $\mathbb{P}(P)^+$.

Proof Let $P, Q, R \in \mathbb{P}(P)^+$. Since $d(.,.)$ is a metric on $\mathbb{P}(P)^+$, we can then conclude that $P \neq Q$ and $d_{\mathbb{A}}(P, Q) \geq d(P, Q) > 0$. Furthermore, since $d_{A[P]}(P, Q)$ is a pseudometric for all $A[P] \in \mathbb{A}$, we have both $d_{\mathbb{A}}(P, P) = \sup\{d_{A[P]}(P, P) : A[P] \in \mathbb{A}\} = 0$, and $d_{\mathbb{A}}(P, Q) = d_{\mathbb{A}}(Q, P)$.

Finally, again since $d_{A[P]}(P, Q)$ is a pseudometric for all $A[P] \in \mathbb{A}$,

$$d_{\mathbb{A}}(P, Q) = \sup\{d_{A[P]}(P, Q) : A[P] \in \mathbb{A}\}$$

$$\leq \sup\{d_{A[P]}(P,R) + d_{A[P]}(R,Q) : A[P] \in \mathbb{A}\}$$

$$\leq \sup\{d_{A[P]}(P,R) : A[P] \in \mathbb{A}\} + \sup\{d_{A[P]}(R,Q) : A[P] \in \mathbb{A}\}$$

$$\leq d_{\mathbb{\Lambda}}(P,R) + d_{\mathbb{\Lambda}}(R,Q) \tag{8.14}$$

Note that this result does not rely on $d(.,.)$ being the variation metric. In particular, it works with the Hellinger metric as well.

8.4.1 Further Properties of the New Metric

To help understand the nature of this new class of metric, we make a few remarks here as some lemmas below.

Lemma 8.3 If P is discrete and $\mathbb{A}$ contains all two point sets $\{i,j\}$, then the $d_{\mathbb{\Lambda}}(P,Q)$ neighbourhoods of P are contained in DeRobertis (1978) density ratio spheres

$$\wedge_s(p;\epsilon) = \{Q : \sup_{i,j} |\log(p_i) - \log(q_i) - \log(p_j) + \log(q_j)| \leq \epsilon\} \tag{8.15}$$

Proof Suppose, without loss of generality that

$$\rho = \frac{p_i}{p_j} \geq \frac{q_i}{q_j}$$

Then

$$d_{\{i,j\}}(P,Q) = \frac{p_i}{p_i+p_j} - \frac{q_i}{q_i+q_j} = \frac{p_iq_j - p_jq_i}{(p_i+p_j)(q_i+q_j)} = \frac{\rho}{1+\rho}\frac{c-1}{c+\rho}$$

where $c = \frac{p_iq_j}{p_jq_i} = \exp\{|\log p_i - \log q_i - \log p_j + \log q_j|\} \geq 1$.

Clearly $d_{\{i,j\}}(P,Q)$ is increasing in $c \geq 1$. Therefore, the result follows.

So, for discrete variables, the topology defined by such a metric is at least as refined as topology defined by density ratio spheres.

We briefly study the relationship between this metric and the difference of logarithms

of two densities associated with two Bayesian networks in Section 8.5.

Now assume that $\mathbb{A} = \mathbb{C}$. Note that, for any set $A \in \mathbb{C}$

$$2d_A(P,Q) = \int_A |\frac{p(\theta)}{P(A)} - \frac{q(\theta)}{Q(A)}| d\theta = \frac{1}{P(A)} \int_A p(\theta) |\frac{P(A)}{Q(A)} \exp\{t(\theta)\} - 1| d\theta$$

$$\leq \sup_{\theta \in A} |\frac{P(A)}{Q(A)} \exp\{t(\theta)\} - 1|$$

where $t(\theta) = |\log p(\theta) - \log q(\theta)|$.

Now assume $|t(\theta)| \leq \tau$, and note that

$$e^{-\tau} \leq \frac{Q(A)}{P(A)} = \frac{\int_A q(\theta) d\theta}{\int_A p(\theta) d\theta} = \frac{\int_A \exp\{t(\theta)\} p(\theta) d\theta}{\int_A p(\theta) d\theta} \leq e^{\tau}$$

which implies

$$2d_A(P,Q) \leq \exp\{2\tau\} - 1$$

Thus we have proved the following lemma.

Lemma 8.4 Suppose probability measures P and Q have respective densities p and q with respect to the same dominating measures, and strictly positive on their shared support. Then if, for all $\epsilon > 0$, there exist (small) values of $\tau(\epsilon) > 0$, if $\theta \in A$, $A \in \mathbb{C}$

$$|\log p(\theta) - \log q(\theta)| \leq \tau$$

then

$$d_A(P,Q) \leq \epsilon.$$

It is clear therefore that although these metrics are much fiercer than the variation metric, the open set around P are rich, provided that $\mathbb{A}$ does not contain sets which are too improbable. In fact, we have a partial converse of this result.

Lemma 8.5 Suppose probability measures P and Q ($P \neq Q$) have respective continuous densities p, q with respect to the same dominating measures, non-zero on their shared support. For all $\tau > 0$, write

$$A_U(\tau) = \{\theta : \log p(\theta) - \log q(\theta) \geq \tau\}$$

$$A_L(\tau) = \{\theta : \log p(\theta) - \log q(\theta) \leq -\tau\}$$

$$A_M(\tau) = \{\theta : |\log p(\theta) - \log q(\theta)| < \tau\}$$

Suppose there exists a value of $\eta > 0$ such that, for all $\tau < \eta$

$$min\{P(A_U(\tau)), P(A_L(\tau))\} > 0$$

Then for all $c > 0$ there exits a value $\tau > 0$ and a set $C(\tau) \subset \Omega$, $P(C) > 0$

$$d_C(P, Q) \geq (1 - e^{-\tau}).$$

Proof First note that

$$P(A_U(\tau)) - Q(A_U(\tau)) = \int_{A_U(\tau)} (p(\theta) - q(\theta))d\theta \geq (1 - e^{-\tau})P(A_U)$$

and

$$Q(A_L(\tau)) - P(A_L(\tau)) = \int_{A_L(\tau)} (q(\theta) - p(\theta))d\theta \geq (e^{\tau} - 1)P(A_L)$$

Now if $\mu(A_U(\tau)) = 0$, then

$$d(P, Q) = \tfrac{1}{2} \int_{\theta \in \Omega} |p(\theta) - q(\theta)|d\theta$$

$$\leq \frac{1}{2} \int_{\theta \in \Omega} (q(\theta) - p(\theta))d\theta + \int_{A_M} |p(\theta) - q(\theta)|d\theta \leq \tau$$

Similarly, if $\mu(A_L(\tau)) = 0$

$$d(P, Q) \leq \frac{1}{2} \int_{\theta \in \Omega} (p(\theta) - q(\theta))d\theta + \int_{A_M} |p(\theta) - q(\theta)|d\theta \leq \tau$$

Therefore, for all τ

$$min\{\mu(A_U(\tau)), \mu(A_L(\tau))\} = 0$$

then $P = Q$ in contradiction to our hypothesis. So, provided τ is small enough, say $\delta < \eta$, then

$$min\{\mu(A_U(\tau)), \mu(A_L(\tau))\} > 0$$

which, since p is strictly positive in turn implies

$$min\{P(A_U(\tau)), P(A_L(\tau))\} > 0$$

It follows from the above that both A_U and A_L are such that $P(A_U(\tau)) - Q(A_U(\tau)) > 0$ and $P(A_L(\tau)) - Q(A_L(\tau)) > 0$.

If $P(A_U) - Q(A_U) \geq Q(A_L) - P(A_L)$, then one can choose any subset B_U of A_U such that $P(B_U) - Q(B_U) = Q(A_L) - P(A_L)$. This is clearly possible if P and Q are continues. On the other hand if $P(A_L) - Q(A_L) \geq Q(A_U) - P(A_U)$, then one can choose any subset B_L of A_L such that $P(B_L) - Q(B_L) = Q(A_U) - P(A_U)$.

So under the conditions above we can construct two sets B_U, B_L such that

$$B_U = \{\theta : \log p(\theta) - \log q(\theta) \geq \tau\}$$

and

$$B_L = \{\theta : \log p(\theta) - \log q(\theta) \leq -\tau\}$$

where $P(C) = Q(C)$, and $C = B_U \cup B_L$. Then, we can write

$$2d_C(P,Q) = \int_C |\tfrac{p(\theta)}{P(C)} - \tfrac{q(\theta)}{Q(C)}| d\theta =$$

$$\frac{1}{P(C)}\{\int_{B_U} (p(\theta) - q(\theta))d\theta - \int_{B_L} (p(\theta) - q(\theta))d\theta\}$$

$$= \frac{1}{Q(C)}\{Q(B_U)(e^{\tau}-1) + Q(B_L)(1-e^{-\tau})\}$$

$$\geq (1-e^{-\tau})$$

which implies

$$d_C(P,Q) \geq (1-e^{-\tau})$$

as required.

So if we set $\mathbb{A} = \mathbb{C}$, this metric is not really new. It essentially demands that the log-densities of two distributions are close everywhere. Furthermore, this is not that practical, because it demands proportionate closeness in the tails of the density and it would be unrealistic to expect such levels of subjective certainty on sets with very small probability. Sets that have large prior probability do not affect the topology of $d_{\mathbb{A}}(P,Q)$ as is demonstrated in the following lemma.

Lemma 8.6 If $P(A) > c > 0$ then for all $\epsilon > 0$ there exits a δ such that if $d(P(A),Q(A)) < \delta$, then $d_A(P(A),Q(A)) < \epsilon$.

Proof

$$2d_A(P(A),Q(A)) = \int_A |\frac{p(\theta)}{P(A)} - \frac{q(\theta)}{Q(A)}|d\theta$$

$$\leq \frac{1}{P(A)}\int_A |p(\theta)-q(\theta)|d\theta + |\frac{1}{P(A)} - \frac{1}{Q(A)}|\int_A |q(\theta)|d\theta$$

$$\leq \frac{1}{P(A)}\int_A |p(\theta)-q(\theta)|d\theta + \frac{|P(A)-Q(A)|}{P(A)Q(A)}$$

$$\leq \frac{2\delta}{c} + \frac{\delta}{c(c-\delta)} = \frac{\delta(2c+1-\delta)}{c(c-\delta)}$$

Which for fixed values of c is continuous at zero and equal to zero when $\delta = 0$.

So, when considering limits, we will gain nothing over the variation metric by including

sets with higher than a threshold probability. It is the distances associated with small sets A which might contribute something new. However, when we learn through Bayes rule, typically, as our sample increases in size, the posterior densities associated with different priors will tend to concentrate round the same small open balls. It follows that there may be considerable gain by restricting our attention to the whole space together with small open balls. This provokes the following definition.

Call $d_{\mathbb{A}}(P,Q) = d^{\Delta|C}(P,Q)$ the (δ, C)-credibility metric if

$$\mathbb{A} = \{\Omega\} \cup \bigcup \{B(\theta_0;\delta) : \theta_0 \in C \subseteq \Omega, 0 < \delta \leq \Delta\} \tag{8.16}$$

where $B(\theta_0;\delta)$ is a Euclidean open ball with center at θ_0 and diameter δ, $\mu(B(\theta_0;\delta))$ is its dominating measure and $d(.,.)$ is the total variation metric.

Write $d^{\Delta}(P,Q) = d^{\Delta|\Omega}(P,Q)$. We see that, provided that the space of densities we consider is smooth enough, this metric gives the sort of limiting results we require. Furthermore, the type of smoothness conditions we need to impose seem relatively benign and plausible from a subject perspective.

Explicitly, we can write $d_{B(\theta_0;\delta)}(P,Q) = d(P \mid B(\theta_0;\delta), Q \mid B(\theta_0;\delta))$. We show that, within a set $\mathbb{A}$ a sufficiently "small" ball $B(\theta_0,\delta)$ is not active in $d_{\mathbb{A}}(P,Q)$, provided the log-densities of P and Q are defined and continuous at θ_0. So, P and Q can be very different in variation metric and still be close under this conditional metric. All we require is that both are sufficiently smooth.

Lemma 8.7 Assume that for densities p and q of P and Q respectively, for all $\omega > 0$, there exists a $\delta(\theta_0 : \omega, p) > 0$ such that, for all $\theta \in B(\theta_0;\delta)$

$$|\log p(\theta) - \log p(\theta_0)| < \omega$$

and, for all $\omega > 0$, there exists a $\delta(\theta_0; \omega, q) > 0$ such that, for all $\theta \in B(\theta_0; \delta)$

$$|\log q(\theta) - \log q(\theta_0)| < \omega$$

then

$$\frac{1}{[\mu(B)]} \int_B |\frac{p(\theta)}{p(\theta_0)} - 1| d\theta < (e^{\omega} - 1)$$

$$e^{-\omega} < \frac{p(\theta_0)\mu(B(\theta_0; \delta))}{P(B(\theta_0; \delta))} < e^{\omega}$$

$$\frac{1}{[\mu(B)]} \int_B |\frac{q(\theta)}{q(\theta_0)} - 1| d\theta < (e^{\omega} - 1)$$

$$e^{-\omega} < \frac{q(\theta_0)\mu(B(\theta_0; \delta))}{Q(B(\theta_0; \delta))} < e^{\omega}$$

Proof Note that for all $\theta \in B(\theta_0; \delta)$

$$|\log p(\theta) - \log p(\theta_0)| < \omega \Leftrightarrow |\frac{p(\theta)}{p(\theta_0)} - 1| < e^{\omega} - 1$$

so the first assertion follows. To prove the second assertion, note that

$$(\frac{P(B(\theta_0; \delta))}{p(\theta_0)\mu(B(\theta_0; \delta))} - 1) = \frac{\int_B (\frac{p(\theta)}{p(\theta_0)} - 1) d\theta}{\mu(B(\theta_0; \delta))}$$

and substituting the first result gives

$$e^{-\omega} - 1 < (\frac{P(B(\theta_0; \delta))}{p(\theta_0)\mu(B(\theta_0; \delta))} - 1) < e^{\omega} - 1$$

which rearranges to the given expression. The last two inequalities hold simply by substituting q for p.

One immediate consequence of these inequalities is that they hold if and only if the corresponding conditions hold for the posterior distribution of a shared sampling model, and the log-likelihood is smooth and continuous at θ_0. For then, for example

$$|\log p(\theta \mid x) - \log p(\theta_0 \mid x)|$$

$$= |\log p(\theta) + \log p(x \mid \theta) + \log \int p(\theta)p(x \mid \theta)d\theta$$

$$- \log \int p(\theta)p(x \mid \theta)d\theta - (\log p(\theta_0) + \log p(x \mid \theta_0))|$$

$$\leq |\log p(\theta) - \log p(\theta_0)| + |\log p(x \mid \theta) - \log p(x \mid \theta_0)|$$

so that, for all $\omega' > 0$, provided δ is chosen small enough, for all $\theta \in B(\theta_0; \delta)$

$$|\log p(x \mid \theta) - \log p(x \mid \theta_0)| < \omega'$$

and we obtain analogous inequalities for the posterior densities in $B(\theta_0; \delta)$.

This is important. It means that, with a continuity condition on the likelihood, prior closeness with respect to this metric guarantees posterior closeness. We use this fact in the next section.

Theorem 8.1 For all $\epsilon > 0$, if P and Q satisfy the continuity conditions above, there exist values of $\eta > 0$, such that if $\delta < \eta$ then

$$d_{B(\theta_0;\delta)}(P, Q) < \epsilon.$$

Proof

$$2d_{B(\theta_0;\delta)}(P, Q) = 2d(P \mid B(\theta_0; \delta), Q \mid B(\theta_0; \delta))$$

$$= \int_{\theta \in B(\theta_0;\delta)} |\frac{p(\theta)}{P(B(\theta_0; \delta))} - \frac{q(\theta)}{Q(B(\theta_0; \delta))}|d\theta \leq A(\delta) + B(\delta) + C(\delta)$$

where, whenever $\log p$, $\log q$ are continuous,

$$A(\delta) = \int_{\theta \in B(\theta_0;\delta)} |\frac{p(\theta)}{P(B(\theta_0; \delta))} - \frac{p(\theta_0)}{P(B(\theta_0; \delta))}|d\theta$$

$$= \frac{p(\theta_0)\mu(B(\theta_0; \delta))}{P(B(\theta_0; \delta))}\{[\frac{1}{\mu(B(\theta_0; \delta))}] \int_{\theta \in B(\theta_0;\delta)} |\frac{p(\theta)}{p(\theta_0)} - 1|d\theta\}$$

which by the inequalities above

$$A(\delta) < e^{\omega(\theta_0;\delta,p)}(e^{\omega(\theta_0;\delta,p)} - 1)$$

Similarly

$$C(\delta) = \int_{\theta \in B(\theta_0;\delta)} |\frac{q(\theta_0)}{Q(B(\theta_0;\delta))} - \frac{q(\theta)}{Q(B(\theta_0;\delta))}| d\theta < e^{\omega(\theta_0;\delta,q)}(e^{\omega(\theta_0;\delta,q)} - 1)$$

and

$$B(\delta) = \mu(B(\theta_0;\delta))|\frac{p(\theta_0)}{P(B(\theta_0;\delta))} - \frac{q(\theta_0)}{Q(B(\theta_0;\delta))}|$$

$$\leq |\frac{\mu(B(\theta_0;\delta))p(\theta_0)}{P(B(\theta_0;\delta))} - 1| + |\frac{\mu(B(\theta_0;\delta))q(\theta_0)}{Q(B(\theta_0;\delta))} - 1|$$

which by the inequalities above,

$$B(\delta) \leq (e^{\omega(\theta_0;\delta,p)} - 1) + (e^{\omega(\theta_0;\delta,q)} - 1).$$

Thus, for a given θ_0, and P and Q, for all $\epsilon > 0$, there is a value of δ such that

$$d_{B(\theta_0;\delta)}(P,Q) < \frac{1}{2}\{e^{\omega(\theta_0;\delta,p)}(e^{\omega(\theta_0;\delta,p)} - 1) + e^{\omega(\theta_0;\delta,q)}(e^{\omega(\theta_0;\delta,q)} - 1)$$

$$+(e^{\omega(\theta_0;\delta,p)} - 1) + (e^{\omega(\theta_0;\delta,q)} - 1)\}$$

$$= \frac{1}{2}\{+(e^{2\omega(\theta_0;\delta,p)} - 1) + (e^{2\omega(\theta_0;\delta,q)} - 1)\} = \epsilon(\theta_0, P, Q, \delta)$$

as required.

A remark and some corollaries can be obtained form the theorem above and we list them below.

Remark 8.1 We note that it would be helpful if we could find, for any fixed P, conditions for the bound $\epsilon(\theta_0, P, Q, \delta)$ not to depend on θ_0 or Q, for then we could assert that the small open sets $B(\theta_0;\delta)$ were not active in the limit in any convergence in

confidence metrics. In fact, this is relatively straightforward.

Corollary 8.1 Suppose that P has a differentiable log density with derivative $D\log p(\theta)$, bounded by M for all θ, i.e.

$$\|D\log p(\theta)\|_0 \leq M$$

Then for all distributions Q with differentiable log-densities also bounded by M. For all $\epsilon > 0$, there exists a value of η such that for all sets $B(\theta_0;\delta)$, whenever $\delta < \eta$,

$$d_{B(\theta_0;\delta)}(P,Q) < \epsilon.$$

Proof This follows immediately from the lemma above, since if $D\log p(\theta)$, $D\log q(\theta)$ are bounded then they are automatically uniformly continuous in θ_0.

It should be noticed that according to this corollary, we can write $d_{B(\theta_0;\delta)}(P,Q) = d(P \mid B(\theta_0;\delta), q \mid B(\theta_0;\delta))$. As we said before, $B(\theta_0;\delta)$ is a sufficiently small ball in $\mathbb{A}$ which is not active in (δ, C)- credibility metric, $d_{\mathbb{A}}(P,Q) = d^{\Delta|C}(P,Q)$, where $\mathbb{A}$ is defined in Equation (8.16), and δ is defined in Corollary 8.1.

This is very useful. It implies, in particular, that two strictly positive unimodal bounded prior densities with sub-exponential tails will look locally similar in the sense of this metric. Later we will use this to relate the metric above to well-known results about robust families of priors. We could link this to the Gustafson's ides to restrict the class of prior distributions into a parameterised class of priors (see Smith and Daneshkhah (2004)).

However, the result is also rather disturbing. In a practical setting, it is difficult to imagine how we might be confident in asserting the condition of this corollary, which makes strong statements about the tail behaviour of a prior density. Fortunately, the uniform continuity we require for the convergence of our metric can be obtained provided we require closeness for sets $B(\theta_0;\delta)$ for which $p(\theta_0) > c > 0$.

Corollary 8.2 Suppose that P has a continuous bounded density p at all points θ_0, such that $p(\theta_0) \geq c_p > 0$, and all distributions Q have a continuous bounded density q at all points θ_0, such that $q(\theta_0) \geq c > 0$. Suppose the sets $D_p = \{\theta_0 : p(\theta_0) \geq c_p > 0\}$ and $D_q = \{\theta_0 : q(\theta_0) \geq c_q > 0\}$ are compact. Then for all $\epsilon > 0$ there exists a value of η such that for all sets $B(\theta_0 : \delta)$, $\theta_0 \in D_p \cup D_q$, whenever $\delta < \eta$,

$$d_{B(\theta_0;\delta)}(P, Q) < \epsilon.$$

Proof The required uniform continuity is immediate from the compactness of the sets and the continuity and boundedness of p and q. So for small open sets in a credibility set, with sufficient smoothness assumptions we can expect all associated variation distances to be small a priori.

The motivation behind of this corollary is that we may be able to assert densities which are close and do not wobble too much. (For more discussions and prior to posterior analysis of these densities, see Smith and Daneshkhah (2004)).

8.4.2 Convergence and Sensitivity Under η-Credibility Metrics

The usefulness of the credibility metrics arises from the following simple observation.

Theorem 8.2 Suppose P^* and Q^* are the posterior distributions associated with P and Q respectively after we observe that $\theta \in B \in \mathbb{A}$. Then if $\mathbb{A}$ is closed under intersection, then

$$d_{\mathbb{A}}(P^*, Q^*) \leq d_{\mathbb{A}}(P, Q)$$

Proof. Since $\mathbb{A}$ is closed under intersection with B, $\mathbb{A} \mid B = \{A' \in \mathbb{A} : A' = A \cap B, A \in \mathbb{A}\} \subseteq \mathbb{A}$. Hence, because

$$d_A(P^*, Q^*) = d_A(P \mid \{\theta \in B\}, Q \mid \{\theta \in B\}) = d_{A \cap B}(P, Q)$$

the result is now immediate by definition.

Note that this means that under an extended variation metric, learning about θ directly cannot increase neighbourhoods: in particular the Fréchet derivative always reduces as zero-one information about θ arrives. This is in strong contrast to the use of the ordinary variation metric for which this is untrue in general (see Gustafson and Wasserman (1995)).

In particular, if our experiment tells us that $\theta \in B(\theta_0; \delta)$ and $\delta \to 0$ then, under the conditions of the two corollaries (8.1) and (8.2), the Fréchet derivative does not diverge, and is bounded.

There are more problems here when we learn through a sample distribution.

Lemma 8.8 Prior small credibility closeness gives rise to posterior credibility closeness with a likelihood continuous at all the relevant θ_0.

We next show that the variation distance between posterior cannot explode if we use close priors that equals with smooth priors.

Theorem 8.3 Suppose for all $\gamma > 0$ there exists a value Δ such that, for all $\delta < \Delta$ and Q such that $d_\Lambda(P, Q) < \eta$, $Q(B^c) < \gamma$ where $B = \bigcup_{i=1}^{m} B(\theta_i; \delta)$ and for all $\omega > 0$, there exists a $\Delta > 0$ such that for all $\delta < \Delta$ and all $\{i : 1 \leq i \leq m\}$

$$|\log p(\theta) - \log p(\theta_i)| < \omega$$

$$|\log p(x \mid \theta) - \log p(x \mid \theta_i)| < \omega$$

then for all $\epsilon > 0$, there exists a $\Delta > 0$ such that for all $\delta \leq \Delta$,

$$d(P \mid x, Q \mid x) < \epsilon.$$

Proof

$$d(P \mid x, Q \mid x) = \int |p(\theta \mid x) - q(\theta \mid x)| d\theta$$

$$\leq \Sigma_{i=1}^{m} \int_{B(\theta_i;\delta)} |p(\theta \mid x) - q(\theta \mid x)| d\theta + \int_{B^c} |p(\theta \mid x) - q(\theta \mid x)| d\theta$$

$$\leq \Sigma_{i=1}^{m} (I_i^1(\delta) + I_i^2(\delta) + I_i^3(\delta)) + \int_{B^c} p(\theta \mid x) d\theta + \int_{B^c} q(\theta \mid x) d\theta$$

$$\leq \Sigma_{i=1}^{m} I_i^1(\delta) + \Sigma_{i=1}^{m} I_i^2(\delta) + \Sigma_{i=1}^{m} I_i^3(\delta) + 2\gamma$$

where

$$I_i^1(\delta) = \int_{\theta \in B(\theta_i;\delta)} |\frac{p(\theta_i \mid x)}{p(\theta \mid x)} - 1| p(\theta \mid x) d\theta$$

$$\leq P(\theta \in B(\theta_i : \delta)) \sup\{|\frac{p(\theta_i \mid x)}{p(\theta \mid x)} - 1| : \theta \in B(\theta_i; \delta)\}$$

$$\leq P(\theta \in B(\theta_i; \delta))[e^{2\omega} - 1]$$

Since

$$|\frac{p(\theta_i \mid x)}{p(\theta \mid x)} - 1| = |\exp[(\log p(x \mid \theta) - \log p(x \mid \theta_i)) - (\log p(\theta) - \log p(\theta_i))] - 1|$$

$$\leq [e^{2\omega} - 1]$$

Similarly

$$I_i^3 \leq Q(\theta \in B(\theta_i; \delta))[e^{2\omega} - 1]$$

Finally

$$I_i^2(\delta) = \int_{B(\theta_i;\delta)} |p(\theta_i \mid x) - q(\theta_i \mid x)| d\theta = \mu(B(\theta_i; \delta)) |p(\theta_i \mid x) - q(\theta_i \mid x)|$$

$$= \mu(B(\theta_i : \delta)) p(x \mid \theta_i) \frac{|p(\theta_i) M_i(q) - q(\theta_i) M_i(p)|}{M_i(p) M_i(q)}$$

where

$$M_i(p) = \int_{B(\theta_i;\delta)} p(x \mid \theta) p(\theta) d\theta$$

$$M_i(q) = \int_{B(\theta_i;\delta)} p(x \mid \theta) q(\theta) d\theta,$$

and

$$I_i^2(\delta) = S(\theta_i, \delta, x) \times T(\theta_i, \delta, x)$$

where

$$S(\theta_i, \delta, x) = \frac{p(\theta_i)\mu(B(\theta_i;\delta))p(x \mid \theta_i)}{\int_{B(\theta_i;\delta)} p(x \mid \theta)p(\theta)d\theta}$$

$$T(\theta_i, \delta, x) = \frac{\int_{B(\theta_i;\delta)} p(x \mid \theta)q(\theta)|1 - \frac{p(\theta)}{p(\theta_i)}\frac{q(\theta_i)}{q(\theta)}|d\theta}{\int_{B(\theta_i;\delta)} p(x \mid \theta)q(\theta)d\theta} \leq$$

$$\sup\{|1 - \frac{p(\theta)}{p(\theta_i)}\frac{q(\theta_i)}{q(\theta)}| : \theta \in B(\theta_i;\delta)\}$$

$$= \sup\{|1 - \exp\{(\log p(\theta) - \log p(\theta_i)) - (\log q(\theta) - \log q(\theta_i))|\} : \theta \mid B(\theta_i;\delta)\} \leq \{e^{2\omega} - 1\}$$

Now, note that

$$S^{-1}(\theta_i, \delta, x) = \frac{\int_{B(\theta_i;\delta)} p(x \mid \theta)p(\theta)d\theta}{p(\theta_i)\mu(B(\theta_i;\delta))p(x \mid \theta_i)}$$

$$= \frac{\int_{B(\theta_i;\delta)} \exp\{[\log p(x \mid \theta) - \log p(x \mid \theta_i)] - [\log p(\theta) - \log p(\theta_i)]\}d\theta}{\mu(B(\theta_i;\delta))}$$

$$\geq \exp\{-2\omega\}$$

by hypothesis.

So

$$I_i^2(\delta) \leq e^{2\omega}\{e^{2\omega} - 1\}$$

So

$$d(P \mid x, Q \mid x) \leq (e^{2\omega} - 1)\{\Sigma_{i=1}^m [P(\theta \in B(\theta_i;\delta) \mid x) + Q(\theta \in B(\theta_i;\delta) \mid x) + e^{2\omega}]\} + 2\gamma$$

$$\leq (e^{2\omega} - 1)\{2m + me^{2\omega}\} + 2\gamma = \epsilon$$

By hypothesis, the function on the right hand side of the inequality above can be made as small as we like by choosing Δ small enough as it is required.

So, contrary to the assertion that it is necessary to restrict our class of prior distributions into a parametrised family of distributions, we can work with a general class of priors here. It is just needed to work with an appropriate extended variation metric above.

It should be noticed that a similar theorem can be proved for posterior predictive distributions. As we discussed in the last section, however, working with predictive distributions because of complex computation is quite difficult, but by using them, we can avoid of the priors with unstable behaviours (with too much wobble). We present similar results as above for posterior predictive distributions.

First, we should show that $d(p(z \mid x), q(z \mid x))$ is a lower bound for $d(p(\theta \mid x), q(\theta \mid x))$. That means,

$$d(p(z \mid x), q(z \mid x)) \leq d(p(\theta \mid x), q(\theta \mid x))$$

where $p(z \mid x) = \int_\theta p(z \mid \theta) p(\theta \mid x) d\theta$.

For this purpose, we use the total variation distance as follows,

$$d(p(z \mid x), q(z \mid x)) = \frac{1}{2} \int_z |p(z \mid x) - q(z \mid x)| dz = \frac{1}{2} \int_z | \int_\theta p(z \mid \theta)(p(\theta \mid x) - q(\theta \mid x)) d\theta | dz$$

$$\leq \frac{1}{2} \int_z \int_\theta |p(z \mid \theta)||p(\theta \mid x) - q(\theta \mid x)| d\theta dz$$

By some assumptions that will be mentioned in the following theorem, we are able to conclude that

$$= \frac{1}{2} \int_\theta |p(\theta \mid x) - q(\theta \mid x)| \{ \int_z |p(z \mid \theta)| dz \} d\theta = \frac{1}{2} \int_\theta |p(\theta \mid x) - q(\theta \mid x)| d\theta$$

$$= d(p(\theta \mid x), q(\theta \mid x))$$

We will then show that as $n \to \infty$ (or equivalently as $\Delta \to 0$), $d(p(z \mid x), q(z \mid x))$ would be very small (and bounded).

Theorem 8.4 Suppose the likelihood function $p(x \mid \theta)$ is bounded by M, and suppose that any prior distribution P has a differentiable log density with derivative $D \log p(\theta)$ bounded by N, i.e., there exits $N > 0$ such that for all θ, $\|D \log p(\theta)\| \leq N$, and for any other arbitrary prior distribution and for all $\gamma > 0$ there exists $\Delta > 0$ such that for all $\delta \leq \Delta$, $Q(B^c(\theta_0(x); \delta)) < \gamma$, where $\theta_0(x)$ denote an estimation such as maximum likelihood. Then, for all $\epsilon > 0$, there exists a $\Delta > 0$ such that for all $\delta \leq \Delta$

$$d(p(z \mid x), q(z \mid x)) < \epsilon$$

Proof

We can write the equation below

$$|p(z \mid x) - q(z \mid x)| = \int_\theta p(z \mid \theta)|p(\theta \mid x) - q(\theta \mid x)|d\theta$$

as the following form

$$= \int_{\theta \in B(\theta_0(x);\delta)} p(z \mid \theta)|p(\theta \mid x) - q(\theta \mid x)|d\theta + \int_{\theta \notin B(\theta_0(x);\delta)} p(z \mid \theta)|p(\theta \mid x) - q(\theta \mid x)|d\theta$$

where $B(\theta_0(x); \delta)$ is a open ball with center at $\theta_0(x)$ and diameter δ.

It can be easily concluded that

$$\int_{\theta \notin B(\theta_0(x);\delta)} p(z \mid \theta)|p(\theta \mid x) - q(\theta \mid x)|d\theta \leq 2M\gamma$$

By the hypothesis in Corollary 8.1, we can say that for all $\omega > 0$ there exists $\Delta > 0$ such that for all $\delta < \Delta$

$$|\log p(\theta) - \log q(\theta)| < \omega$$

where p and q denote the densities associated with P and Q respectively.

Therefore, by the results obtained from Theorem 8.3, the following inequality can be obtained

$$\int_{\theta \in B(\theta_0(x);\delta)} p(z \mid \theta)|p(\theta \mid x) - q(\theta \mid x)|d\theta \leq Me^{2\omega}\{e^{2\omega} - 1\}$$

Therefore,

$$|p(z \mid x) - q(z \mid x)| \leq M\{2\gamma + e^{2\omega}\{e^{2\omega} - 1\}\} = \epsilon$$

as required.

More details and discussion with some applications can be found in Smith and Daneshkhah (2004).

8.5 Credible Convergence in Estimated Bayesian Networks

In this section, we briefly present some results associated with asymptotic behaviour of credibility metric for the Bayesian networks. We present more results in full detail including the rate of convergence, approximate cause, and approximation of hypercausality in Smith and Daneshkhah (2004).

Suppose we have a general Bayesian network over $\{X_1, \ldots, X_n\}$ and defined by $\{X_{pa_i}, i = 2, \ldots, n\}$, with two different densities p and q consistent with this Bayesian network, so that

$$p(\underline{x}) = \prod_{i=1}^{n} p(x_i \mid x_{pa_i}), \quad q(\underline{x}) = \prod_{i=1}^{n} q(x_i \mid x_{pa_i})$$

Then, by the triangular inequality, we can write

$$\sup_{\underline{x}} |\log p(\underline{x}) - \log q(\underline{x})| \leq \sum_{i=1}^{n} \sup_{(x_i, x_{pa_i})} |\log p(x_i \mid x_{pa_i}) - \log q(x_i \mid x_{pa_i})|$$

So clearly, to ensure closeness of these two joint densities, it is sufficient to have closeness of each conditional term in the product presentations mentioned in the equation above.

On the other hand, we can write the DeRobertis's density ratio metric as

$$\sup_{(\underline{x},\, \underline{x}')} |\log (\frac{p(\underline{x})q(\underline{x}')}{p(\underline{x}')q(\underline{x})})| \leq \sup_{\underline{x}} |\log p(\underline{x}) - \log q(\underline{x})| + \sup_{\underline{x}'} |\log p(\underline{x}') - \log q(\underline{x}')|$$

$$\leq 2 \sup_{\underline{x}} |\log p(\underline{x}) - \log q(\underline{x})| \leq 2\Sigma_{i=1}^{n} \sup_{(x_i, x_{pa_i})} |\log p(x_i \mid x_{pa_i}) - \log q(x_i \mid x_{pa_i})|$$

That means, De Robertis's metric and $\sup_{\underline{x}} |\log p(\underline{x}) - \log q(\underline{x})|$ are topologically equivalent. Further studies can be found in Smith and Daneshkhah (2004).

8.6 Discussion and Further Work

In this chapter we define a new local sensitivity measures in terms of credibility metrics. We have shown that these metrics asymptotically behave better. We have argued that the corresponding Fréchet derivative similar to the derivatives studied by Gustafson et al (1996) does not tend to zero. However, we do have uniform boundedness under appropriate conditions. That means a close credible metric a priori will give a close credible metric a posteriori. So, we do not get the sort of divergence that Gustafson et al (1996) derived with the total variation metric.

During the study of material in Chapter 7 we discussed that general robustness measures totally suitable for a thorough investigation of robustness of Bayesian networks. The necessary analysis, begun in this chapter has rather distracted us from analysis of sensitivity in Bayesian networks. We are currently applying the results obtained in Chapter 8 to the Bayesian networks, and developing some more results specific to the Bayesian networks with respect to possible situations that we did not considered in this chapter. Some of these developing works will be briefly discussed below.

It is important to investigate how the theorems and lemmas introduced in this chapter are applicable to Bayesian networks. However, the proposed likelihood (multinomial distributions) and prior distribution (Dirichlet or product of Dirichlet's) for Bayesian networks with discrete variables would provide the conditions (especially, continuity condition) in the theorems and lemmas. But this still needs to be formally proved.

However, we have shown that the credibility metrics between two Dirichlet distributions for the small ball (as described in Theorem 8.1) around of the maximum likelihood, $\hat{\theta}$ subject to the conditions mentioned in this theorem would be very small. However, more efforts are required to investigate the validity of the remaining conditions. This is now under study in the paper in progress by Smith and Daneshkhah (2004).

Another inspiring work which still needs to be completed is related to the asymptotic behaviour of the local sensitivity measures (or closeness distances). The local sensitivity measures introduced in this thesis including the credibility metrics as the closeness distances between densities can be usually represented in terms of difference between logarithms of the densities. In many cases, this difference would ensure that the Hellinger distance (or other equivalent distances or metrics), and thereby the corresponding local sensitivity measure will at least be bounded for large enough sample sizes. We believe that if the difference between logarithms of two prior densities associated with two equivalent Bayesian networks are very close in the sense described above (or more precisely as mentioned in Lemma 8.4), the Hellinger distance between corresponding posterior distributions (or between predictive distributions, probably with more infeasibility in computation) will tend to zero as the sample size goes to infinity. Therefore, the direction of the reversible arrow(s) in the essential graph representing the equivalent class of these Bayesian networks no longer matters for large enough sample size. It is called *approximate cause* by Smith and Daneshkhah (2004).

It should be noticed that the aforementioned closeness measures require a condition of density positivity. This condition should be valid for the likelihood and prior distribution of Bayesian networks with discrete variables and the credibility metrics. The validity of this condition requires a formal investigation.

Another challenging work that we should consider here is the generalisation of Lemma 8.3 for the continuous variables. We claim that it could be done in an analogous approach in terms of histograms.

During our study on the asymptotic behaviour of the local sensitivity measures, we implicitly derived this point that using of the predictive distributions instead of posterior distributions could represent better result. In this chapter, we have made some effort regarding to this point for the local sensitivity measures derived in terms of the credible metrics. We suspect that despite the complexity of computation, the derivatives involved in the aforementioned measures do indeed achieve convergence when we use the predictive distributions.

Chapter 9

Conclusion

In this thesis we have constructed prior families of distributions on the parameters of a causal Bayesian network. We have shown that if the relationships between variables in a Bayesian network are causally asserted within an uncertain Bayesian network, then the corresponding parameters must be locally and globally independent of each other. To enter these independence assumptions into a causal Bayesian network, we need to modify the assumptions of factorisation invariance introduced by Pearl (2000). This modified set of factorisation of densities are invariant to a new class of manipulations able to assert (contingent) randomised intervention. We called the causal Bayesian network associated with this new "do" operator, hypercausal Bayesian network (or hypercausal prior for a causal Bayesian network). In fact, the definition of hypercausal Bayesian network combine one's information about the parameters of the Bayesian network with the actual relationships in the world, as represented by the structure and the parameters of the Bayesian network. Our results (in Chapters 5 and 6) concern what kinds of prior information about parameters allow these causal relationships to be identified from observational data (see Daneshkhah and Smith (2003a, b) for more detail). We then have defined the multicausal essential graph on the equivalence class of Bayesian networks

where each member of this class demonstrates hypercausality. It might be asked why someone would want to put down a prior over the parameters in an essential graph as described in Chapter 6 when the causal interpretation of the resulting parameters of interest. To deal with this confusion, it should be noticed that the multicausal essential graph characterisation, introduced in Chapter 6, concerns a single hypothesised model. It is not an assertion about a common prior to be used for causal Bayesian network for the model selection as is more typical in, for example, Cooper and Yoo (1999) and references therein.

In Chapter 7, we reviewed current robustness measures and studied the local sensitivity analysis of the Bayesian networks with respect to some source of uncertainties in inputs such as misidentification in prior distributions and independence assumptions. Our approach, in some senses, are different with ones studied by Laskey (1993) (used an ordinary linear derivative as a sensitivity measure), Cozman (1996) (used convex set of distributions to study global robustness analysis).

We have calculated the local sensitivity measures introduced by Basu (1996), Gustafson (1996a), and Wasserman (1992) (for general Bayesian models), and Gustafson (1996b) with respect to the uncertainties mentioned above for the graphical models: Bayesian networks with independent parameters; Bayesian networks with dependent parameters; decomposable graphical models. We have demonstrated that, subject to parameter independence assumptions, perturbing the prior distribution of one node does not change the posterior quantity of other nodes. When the parameters are not independent, calculation of the local sensitivity measures is not feasible in some situations. We have defined an hierarchical prior distribution on these parameters and used Gustafson's idea to calculate and interpret the local sensitivity measures. In this case, one might be

faced with the Bayesian unidentifiability issue. We have made some points that might be useful to deal with this issue. However, more studies are required here to implement.

We have made an effort to define the local cause in terms of the local sensitivity in the causal Bayesian network. We have addressed some issues that needed to be studied in Section 7.5.

In Section 7.6, and Chapter 8, we have studied the asymptotic behaviour of the local sensitivity measures introduced throughout this thesis. Unfortunately, these measure tend to infinity as the sample size becomes large enough. Gustafson et al (1996) have tried to solve this problem by restricting the class of prior distributions. This suggestion might work for some rare examples (not even proved).

They led us to construct a new general class of robustness measures in which the local sensitivity calculated in terms of these metrics exhibit more reasonable behaviour than standard ones and yet exhibit plausible levels of generality. We have shown that our measure at least bounded for the large sample size. We feel we have only scratched the surface of the study of this promising class of credibility metrics. A lot of works should be done to finished this study as we have mentioned them in detail in Chapter 8.

Bibliography

[1] Andersson, S. A., Madigan, D., and Perlman, M. D. (1997) A characterisation of Markov equivalence classes for acyclic. *The Annals of Statistics,* **24**, 505-541.

[2] Basu, S. (1996). Local sensitivity, Functional Derivatives and nonlinear posterior quantities. *Statistics and Decisions,* **14**, 405-418.

[3] Beran. R. (1977a). Robust location estimates. *The Annals of Statistics,* **5**, 431-444.

[4] Beran. R. (1977b). Minimum Hellinger distance estimates for parametric models. *The Annals of Statistics,* **5**, 445-463.

[5] Berger, J. O. (1984). The robust Bayesian viewpoint (with discussion). In *Robustness of Bayesian Analysis,* J. Kadane (Ed.), Noth-Holland, Amesterdam.

[6] Berger, J. O. (1990). Robust Bayesian analysis: sensitivity to the prior. *J. Statist. Plann. Inference,* **25**, 303-328.

[7] Berger, J. O. (1994). An overview of robust Bayesian analysis. *Test,* **3**, 5-58.

[8] Castelo, R. (2002). *The Discrete Acyclic Digraph Markov Model in Data Mining.* PhD Thesis, University of Utrecht.

[9] Chickering, D. M. (1995) A transformational characterisation of Bayesian network structures. In *Uncertainty in Artificial Intelligence*, P. Besnard and S. Hanks (Eds.). San Francisco: Morgan Kaufmann, **11**, 87–98.

[10] Cooper, G. F., and Yoo, C. (1999) Causal Discovery from a Mixture of Experimental and Observational Data. In K.B. Laskey and H. Prade (eds.),*Proceeding of the Fifteenth Conference on Uncertainty in Artificial Interlligence.* Morgan Kaufmann Publishers, San Francisco.

[11] Cooper, G. F., and Herskovits, E. (1992). A Bayesian Method for the Induction of Probabilistic Networks from Data. *Machine Learning*, **9**, 309-347.

[12] Cozman. F. (1996). Robustness analysis of Bayesian networks with global neighbourhoods. Technical Report CMU-RI-TR-96-42, Carnegie Mellon University, Pittsburgh.

[13] Cowell, R. G. (1996). On compatible priors for Bayesian networks. *IEEE Trans. Pattern Anal. Machine Intelligence*, **18**, 901-911.

[14] Cowell, R. G. (1998). Mixture reduction via predictive scores. *Statistics and Computing.*, **8**, 97-103.

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

[16] Croft, J., and Smith, J. Q. (2003). Discrete mixtures in simple Bayesian Networks with hidden variables. *J of Computational Statistics and Data Analysis* (to appear).

[17] Cuevas, A., and Sanz, P. (1988). On differentiability properties of Bayes operators. In *Bayesian Statistics 3*, J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M, Smith and M. West (Eds.). Oxford: University Press, 569-577.

[18] Daneshkhah, A. R., and Smith, J. Q. (2003a) A Relationship between randomised Manipulation and Parameter Independence. In *Bayesian Statistics 7*, J. M.

Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M, Smith and M. West (Eds.). Oxford: University Press, 477-484.

[19] Daneshkhah, A. R., and Smith, J. Q. (2003b) Multicausal prior families, randomisation and essential graphs. In *Proceedings of the First European workshop on probabilistic graphical models,* J. A. Gmez, and A. Salmern (Eds.), Cuenca, pp. 25-32. It will also be published as a refereed paper in *Advances in Bayesian Networks,* J.A. Gmez, S. Moral and A. Salmern (Eds.), Physica-Verlag.

[20] Dawid, A. P. (1979). Conditional independence in Statistical theory (with discussion). *J. Roy. Statist. Soc.*, B, **41**, 1-31.

[21] Dawid, A. P. (1980). Conditional independence for Statistical operations. *The Annals of Statistics,* **8**, 598-617.

[22] Dawid, A. P. (1984). Present position and potential developments: some personal views (with discussion). *J. Roy. Statist. Soc.*, A, **147**, 278-292.

[23] Dawid, A. P. (1997). Prequential analysis. *Encyclopaedia of Statistical Sciences.*, Update volume 1, S. Kotz, C. B. Read, and D. L. Banks (Eds.). Wiley-Interscience, 464-470.

[24] Dawid, A. P. (2000). Causal Inference without counterfactuals (with discussion). *J. Amer. Statist. Ass.*, **95**, 407-448.

[25] Dawid, A. P. (2002). Influence Diagrams for causal modelling and Inference. *Intern. Statist. Rev.*, **70**, 161-189.

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

[27] DeRobertis, L. (1978). *The Use of Partial Prior Knowledge in Bayesian Inference.* PhD thesis, Yale University, New Haven, CT.

[28] DeRobertis, L., and Hartigan, J. (1981). Bayesian inference using intervals of measures. *The Annals of Statistics* , **9**, 235-244.

[29] Dey, D. K., and Birmiwal, L. R. (1994). Robust Bayesian analysis using entropy and divergence measures. *Statist. Probab. Lett.*, **20**, 287-94.

[30] Dey, D. K., Ghosh, S. K. and Lou, K. (1996). On local sensitivity measures in Bayesian analysis (with discussion). In *Bayesian Robustness*, J. O. Berger et al, (Eds.), IMS, Lecture Notes-Monograph Series, vol. **29**, 63-80.

[31] Diaconis, P., and Freedman, D. (1986). On the consistency of Bayes estimates. *The Annals of Statistics,* **14**, 68-87.

[32] Fernholz, L. T. (1983). *von Mises calculus for Statistical Functionals.* Springer-Verlag Lecture Notes in Statistics, **19**, New York.

[33] Geiger, D. and Heckerman, D. (1997). A characterisation of the Dirichlet distribution through global and local parameter independence. *The Annals of Statistics,* **25**, 1344-1369.

[34] Geiger, D. and Heckerman, D. (1999). Parameter Priors for directed acyclic graphical models and the characterization of several probability distributions. Technical Report, MSR-TR-98-67, Microsoft Research, Advanced Technology Division.

[35] Gibbs, A. L., and Su, F. E. (2002). On choosing and Bounding Probability metrics. The paper is available from WWW URL: `http://www.math.hmc.edu/~su/papers.dir/metrics.pdf`.

[36] Goel, P. K. (1983). Information measures and Bayesian Hierarchical models. *J. Amer. Statist. Assoc.*, **78**, 408-410.

[37] Goel, P. K., and DeGroot, M. H. (1981). Information about Hyperparameters in hierarchical models. *J. Amer. Statist. Assoc.*, **76**, 140-147.

[38] Guihennеuc-Jouyaux, C., Richardson, S., and Lasserre, V. (1998). Convergence assessment in latent variable models: Application to the longitudinal modelling of a marker of HIV progression. In *Discretization and MCMC Convergence Assessment,* C. P. Robert (Ed.). Springer Verlag, New York, 147-159.

[39] Gustafson, P. (1994). *The Local sensitivity of Posterior Expectations,* unpublished PhD thesis, Department of Statistics, Carnegie Mellon University.

[40] Gustafson, P. (1996a). Local sensitivity of inferences to prior marginals. *J. Amer. Statist. Assoc.*, **91**, 774-81.

[41] Gustafson, P. (1996b). Aspects of Bayesian robustness in hierarchical models (with discussion). In *Bayesian Robustness*, J. O. Berger et al, (Eds.), IMS, Lecture Notes-Monograph Series, **29**, 63-80.

[42] Gustafson, P., Srinivasan, C. and Wasserman, L. (1996). Local sensitivity. In *Bayesian Statistics 5*, J. M. Bernardo et al, (Eds.), Oxford: Oxford University Press.

[43] Gustafson, P., and Wasserman, L. (1995). Local sensitivity Diagnostics for Bayesian inference. *The Annals of Statistics,* **23**, 2153-2167.

[44] Hampel, F. R., Rousseeuv, P. J., and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions.* John Wiley, New York.

[45] Heckerman, D. (1995). A tutorial on learning Bayesian networks. Technical Report MSR-TR-95-06, Microsoft Research.

[46] Heckerman, D., Geiger, D. and Chickering, D. (1995). Learning Bayesian networks: The combination of knowledge and statistical data. *Machine Learning,* **20**, 197-243.

[47] Huber. P. J. (1981). *Robust Statistics.* John Wiley, Ney York.

[48] Kirby, A. J., and Spiegelhalter, D. J. (1994). Statistical modeling for the Precursors of Cervical cancer. Case in *Biometry,* N. Lange (Ed.). John Wiley and sons, New York.

[49] Koster, J. T. A. (2000). Graphs, Causality and Structural Equation Models. The slides of the presentation is available from WWW URL: `http://www.knaw.nl/09public/rm/ koster.pdf`.

[50] Laskey, K. B. (1993). Sensitivity Analysis for Probability Assessments in Bayesian Networks. In *Proceedings of the Ninth Conference on Uncertainty in Artificial Intelligence,* and available from WWW URL: `http://ite.gmu.edu/~klaskey/publications.html`.

[51] Lauritzen, S. L. (1996). *Graphical Models.* Clarendon Press, Oxford.

[52] Lauritzen, S. L. (2001). Causal inference from graphical models. In *Complex Stochastic Systems,* O. E. Barndorff Nielsen, D. R. Cox, and C. Kluppelberg (Eds.). Chapman and Hall/CRC Press, London/Boca Raton, 63-107.

[53] LeCam, L., and Yang, G. (1990). *Asymptotics in Statistics.* Springer-Verlag, New York.

[54] Lindley, D. V. (1972). *Bayesian Statistics, a Review.* Philadelphia, PA: SIAM.

[55] Lindley, D. V., and Smith, A. F. M. (1972). Bayes estimates for the linear model (with discussion). *J. Roy. Statist. Soc.* B, **34**, 1-41.

[56] Madigan, D. and York, I. (1994). Bayesian graphical models for discrete data. *International Statistical Review,* **63**, 215-32.

[57] McCulloch, R. (1989). Local model influence. *J. Amer. Statist. Assoc.*, **84**, 473-78.

[58] Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, San Mateo, CA.

[59] Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, **82**, 669-710.

[60] Pearl, J. (2000). *Causality: Models; Reasoning; and Inference.*, Cambridge University Press.

[61] Pearl, J., and Verma, T. S. (1991). A Theory of Inferred Causation. In *Principles of Knowledge Representation and Reasoning: Proceeding of the Second International Conference*, J.A Allen, R. Fikes, and E. Sandewall (Eds.) San Mateo, CA: Morgan Kaufmann, 441-452.

[62] Pericchi, L. R., and Nazaret, W. (1988). On being imprecise at the higher levels of a hierarchical linera model (with discussion). In *Bayesian Statistics 3*, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.). Oxford: University Press, 361-375.

[63] Puch, R. O., and Smith. J. Q. (2002). FINDS: A training package to assess forensic fibre evidence. In *Proceedings of the second Mexican international conference on artificial intelligence,* C. A. Coello-Coello, A. De Albornoz, L. E. Sucar, and O. Cairo-Battistuti (Eds.). Merida, Mexico.

[64] Rachev, S. T. (1991). *Probability metrics and the stability of stochastic models.* John Wiley, New York.

[65] Reiss, R. D. (1989).*Approximate distributions of order statistics: with applications to nonparametric Statistics.* Springer-Verlag, New York.

[66] Richardson, S., and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. *J. Roy. Statist. Soc.*, B, **59**, 731-792.

[67] Robert, C. P. (2001). *The Bayesian Choice.* Springer-Verlag, New York.

[68] Robins, J. M. (1986). A new approach to causal inference in mortality studies with sustained exposure periods-application to control of the healthy worker survivor effect. *Mathematical Modeling,* **7**, 1393-1512.

[69] Robins, J. M. (1997). Causal Inference from complex longitudinal data. In *Latent Variable Modeling and Applications to Causality. Lecture Notes in Statistics (120),* M. Berkane (Ed.). Springer Verlag, New York, 69-117.

[70] Ruggeri, F. (1996) Discussion of ' Asymptotics of some local and global robustness measures. by S. Sivaganesan. In *Bayesian Robustness*, J. O. Berger et al, (Eds.), IMS, Lecture Notes-Monograph Series, **29**, 207-209.

[71] Ruggeri, F., and Wasserman, L. (1993). Infinitesimal sensitivity of posterior distributions. *Canad. J. Statist.*, **21**, 195-203.

[72] Rusakov, D. and Geiger, D. (2000). On parameter priors for discrete DAG models. *Technical report CIS, Technion,* **08.**

[73] Schervish, M. J. (1996). *Theory of Statistics.* Springer-Verlag, New York.

[74] Settimi, R., and Smith, J. Q. (1999). Geometry, moments and Bayesian networks with hidden variables. In *Proceedings of the 7th International Workshop on Artificial Intelligence and Statistics.* Morgan Kaufmann, 472-479.

[75] Settimi, R., and Smith, J. Q. (2000). Geometry, moments and conditional independence tress with hidden variables. *The Annals of Statistics,* **28**, 1179-1205.

[76] Settimi, R., and Smith, J. Q. (2002). On the geometry and model selection of Bayesian directed graphs with isolated hidden nodes. *Submitted to Journal of American Statistical Association.*

[77] Sivaganesan, S. (1993). Robust Bayesian diagnostics. *J. Statist. Plann. Inference*, **35**, 171-188.

[78] Sivaganesan, S. (1996). Asymptotics of some local and global robustness measures. In *Bayesian Robustness*, J. O. Berger et al, (Eds.), IMS, Lecture Notes-Monograph Series, **29**, 195-206.

[79] Smith, J. Q. (1995). Bayesian approximations and the Hellinger metric. Research Report, 729, Department of Statistics, The University of Warwick, Coventry.

[80] Smith, J. Q. (2002) Discussion of ' Chain graph model and their causal interpretation', by S. L. Lauritzen and T. Richardson. *J. R. Statist. Soc. B*, **64**.

[81] Smith, J. Q., and Daneshkhah, A. R. (2004). On Stable Bayesian Inference. Working paper in Department of Statistics, The University of Warwick.

[82] Spiegelhalter, and Cowell, R. G. (1992). Learning in probabilistic expert systems (with discussion). In *Bayesian Statistics 4*, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.). Oxford: University Press, 447-465.

[83] Spiegelhalter, D. J. and Lauritzen, S. L. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks*, **20**, 579-605.

[84] Spiegelhalter, D. J., Dawid, A. P., Lauritzen, S. L. and Cowell, R. G. (1993). Bayesian analysis in expert systems (with discussion). *Statistical Sci*, **8**, 219-247.

[85] Spirtes, P., Glymour, C., and Scheines, R. (1993). *Causality, Prediction and Search.* Springer-Verlag, New York.

[86] Spirtes, P., Glymour, C., and Scheines, R. (1999). *Causality, Prediction and Search.*, 2nd ed. New York, N. Y. MIT Press.

[87] Studeny, M. (2002). Characterization of essential graphs by means of an operation of legal component merging. In *Proceedings of the First European Workshop on Probabilistic Graphical Models*, J. A. Gamez, A. Salmeron (Eds.). University Castilla la Mancha, Spain, 161-172.

[88] Verma, T. S., Pearl, J. (1990). Equivalence and synthesis of causal models. In *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence,*(July, Cambridge, MA), 146-155. Reprinted in Bonissone, P., Henrion, M., Kanal, L. N. and Lemmer, J. F.(Eds.), *Uncertainty in Artificial Intelligence,* vol. **6**, 255-68. Amsterdam: Elsevier.

[89] Verma, T. S., and Pearl, J. (1992). An Algorithm for Deciding if a Set of Observed Independencies has a Causal Explanation. In *Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Mateo, CA, 323-330.

[90] Wasserman, L. (1992). Recent methodological advances in robust Bayesian inference. In *Bayesian Statistics 4*, J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.). Oxford: University Press, 447-465.

[91] Whiley, M., and Titterington, D. M. (2002). Model Identifiability in Naive Bayesian Networks. The paper is available from WWW URL: `http://www.stats.gla.ac.uk/Research/TechRep2002/02-1.pdf`.

[92] Zolotarev, V. M. (1983). Probability metrics. *Theory. Probab. Appl.*, **28**, 278-302.