GRAPHICAL BAYESIAN MODELS
IN
MULTIVARIATE EXPERT JUDGEMENTS
AND
CONDITIONAL EXTERNAL BAYESIANITY

ÁLVARO E. FARIA JR., BSc., MSc.

Thesis submitted for the degree of Doctor of Philosophy

DEPARTMENT OF STATISTICS
UNIVERSITY OF WARWICK

OCTOBER 1996

Typeset by $\mathcal{A}\mathcal{M}\mathcal{S}$-TEX

TABLE OF CONTENTS

SUMMARY v

1. INTRODUCTION AND HISTORICAL REMARKS 1

1.1 Historical Remarks on Combining Opinions 1

1.1.1 Impossibility results in group decision theory 2

1.1.2 The axiomatic and the modelling approaches 3

1.2 Historical Remarks on Statistical Graphical Models 4

1.2.1 Decomposable graphical models 5

1.2.2 Chain graphs 5

1.2.3 Influence diagrams 6

1.3 The Thesis Structure 6

2. A REVIEW ON THE AGGREGATION OF BELIEFS . . . 8

2.1 Some Fundamental Issues in Aggregating Beliefs 9

2.1.1 Quantifying degrees of beliefs 9

2.1.2 Calibration : measures and corrections 10

2.1.3 Coherence and honesty 12

2.1.4 Non-independence between experts 13

2.2 The Axiomatic Approach 14

2.2.1 Pooling operators, axioms and impossibilities 14

2.2.2 The marginalization property and the linear opinion pool . 17

2.2.3 The external Bayesianity property 19

2.2.4 The modified logarithmic opinion pool 20

2.2.5 A problem with external Bayesianity 22

2.3 Some Weighting Schemes 22

2.3.1 The optimal combination of Bates and Granger 23

2.3.2 Bunn's outperformance approach 24

2.3.3 The quasi-Bayes approach of Smith and Makov 25

2.4 The Bayesian Modelling Approach 26

2.4.1 The supra-Bayesian decision maker 27

2.4.2 The general supra-Bayesian analysis 27

2.4.3 Conditional independence and Bayes rule 28

2.4.4 Winkler's consensus model for dependent experts 29

2.4.5 Lindley's reconciliation of discrete probability distributions 31

3. A BASIC REVIEW ON GRAPHICAL MODELS 34

3.1 Notation and Terminology 34

3.2 Chain Graph Models 36

3.3 Conditional Independence 38

3.4 The d-separation Theorem 39

3.5 Influence Diagrams 40

3.6 Decomposability and Structure Preservation 42

4. GENERALISING EB LOGARITHMIC POOLS 46

4.1 A Generalisation of External Bayesianity and LogOps 46

4.2 Deterministic Weights in LogOps 49

4.3 The Bivariate Normal Model 50

A4.1 : The Conditional Normal Density Raised to a Power . . 54

A4.2 : The Product of Normal Densities 56

5. COMMON KNOWLEDGE CI STRUCTURES 57

5.1 Associations and Chain Graphs 58

5.2 A Useful Subclass of CG Models : the PCGs 59

5.3 The Cutting Likelihoods and External Bayesianity 60

5.4 PCG Preservation After Sampling 62

5.5 The Class of Decomposable PCGs 63

6. CONDITIONAL EXTERNAL BAYESIANITY 67

6.1 The Definition of Conditional External Bayesianity 67

6.2 A Characterisation of CEB Pooling Operators 67

6.3 Group Learning in Time Consistent Chains 70

6.4 How CEB Poolings Appear EB 71

6.5 CEB Linear Opinion Pools 75

A6 : Proof of Theorem 6.2 79

A6.1 Characterisation of the Conditional LogOp 79

A6.2 Equivalence Classes for CEB Pooling Operators 82

A6.3 The Proof of Theorem 6.2 89

7. GRAPHICAL MODELS FOR THE EXPERT PROBLEM . 91

7.1 Conditional Independence Structures for the Expert Problem . . 91

7.1.1 A particular subclass of expert problems 92

7.2 Common Knowledge, Value of Information and Sufficiency 95

7.3 Degenerate Situations in the Expert Problem 96

7.4 Non-overlapping Information Bases 97

7.5 Overlapping Information Bases 103

7.5.1 When the shared information is given 104

7.5.2 Accomodating the shared information 105

7.6 The Choice of Experts 106

8. EXTERNALLY BAYESIAN RECONCILIATION RULES . 109
8.1 The Experts' Miscalibration 109
8.2 Independence and Over-confidence 110
8.3 External Bayesianity and Reconciliation Rules 111
8.4 The Binomial Model 112
8.4.1 The case of exchangeable experts 112
8.4.2 Non-exchangeable experts 116
8.5 The Gaussian Case 118
8.6 The CEB Reconciliation Rules 119
9. CONCLUSIONS AND FURTHER RESEARCH 120
9.1 The CEB Pools 120
9.1.1 The flexibility on the weights 120
9.1.2 Complete graphs 121
9.1.3 The non-decomposable chain graphs 121
9.1.4 Using CEB pools not based on complete graphs 122
9.1.5 The CEB linear pools 123
9.2 A Link of the Modelling with the Axiomatic Approach 124
9.2.1 General expert problems 124
9.2.2 The SB analysis of uncalibrated experts 124
REFERENCES 125

ACKNOWLEDGEMENTS

I would like to thank:

- God, for the world we try to understand and describe,
- my parents, Álvaro and Neuza, my son Vitor and my fiancée Katie,
- my spiritual friends for all their support at all times,
- the Brazilian research council, CNPq, for the financial support,
- my friend (and supervisor in the masters) Reinaldo C. Souza for sending me to Coventry,
- my supervisor, Jim Smith for the guidance, encouragement, patience and willingness in all these years of hard work, and finally
- all the staff and research students at the Department of Statistics for their assistance and friendship.

SUMMARY

This thesis addresses the multivariate version of the *group decision problem* (French, 1985), where the opinions about the possible values of n random variables in a problem, expressed as subjective conditional probability density functions of the k members of a group of experts, are to be combined together into a single probability density.

A particular type of graphical chain model a bit more general than an *influence diagram* defined as a *partially complete chain graph* (PCG) is used to describe the multivariate causal (ordered) structure of associations between those n random variables. It is assumed that the group has a *commonly agreed* PCG but the members diverge about the actual conditional probability densities for the component variables in the common PCG. From this particular situation we investigate some suitable solutions.

The axiomatic approach to the group decision problem suggests that the group adopts a combination algorithm which demands, at least on learning information which is common to the members and which preserves the originally agreed PCG structure, that the pools of *conditional* densities associated with the PCG are *externally Bayesian* (Madansky, 1964). We propose a logarithmic characterisation for such *conditionally externally Bayesian* (CEB) poolings which is more flexible than the logarithmic characterisation proposed by Genest et al. (1986). It is illustrated why such a generalisation is practically quite useful allowing, for example, the weights attributed to the joint probability assessments of different individuals in the pool to differ across the distinct conditional probability densities which compound each joint density. A major advantage of this scheme is that it may allow the weights given to the group's members to vary according to the areas of prediction they can perform best. It is also shown that the group's commitment to being CEB on *chain elements* can be accomplished with the group appearing externally Bayesian on the whole PCG. Another feature of the CEB logarithmic pools is that with them the impossibility theorems related to the preservation of independence by opinion pools can be avoided. Yet, in the context of the axiomatic approach, we show the conditions under which the types of pools that satisfy McConway's (1981) *marginalization property*, i.e. the linear pools, can also be CEB.

Also, the *expert judgement problem* (French, 1985) is investigated through the Bayesian modelling approach where a *supra-Bayesian* decision maker treats the experts' opinions as data in the usual Bayesian framework. Graphical representations of standard combination models are discussed in the light of the issues of dependence among experts and sufficiency of experts' statements in certain cases. Most importantly, a supra-Bayesian analysis of uncalibrated experts allows the establishment of a link between the axiomatic and Bayesian modelling approaches. Reconciliation rules which are externally Bayesian are obtained. This result most naturally extends those rules to be CEB in the above mentioned multivariate structures.

CHAPTER 1
INTRODUCTION AND HISTORICAL REMARKS

In view of the vast field in which the research we do here stands and the methodologies employed in this thesis, we shall start by describing some historical remarks on both the aggregation of experts opinions and the graphical statistical models. The objective is also to put the results we obtain in a historical context. So, the Section 1.1 is dedicated to the combination of opinions for decision making with some of the difficulties and the main methodologies being mentioned. The Section 1.2 deals with the evolution of graphical models with emphasis on decomposable structures, chain graphs and influence diagrams. In Section 1.3 the structure of the thesis is presented.

1.1 Historical Remarks on Combining Opinions.

According to Bacharach (1979), one of the earliest works on combining probability distributions is due to Laplace, to whom the linear pool is attributed. Naturally, works on combining *opinions* (as expressed by subjective probability distributions) for decision making could only start to develop from the late 1920's after Ramsey (1926), and more effectively from the early 1950's after Savage (1954a) laid the foundations of the modern Bayesian decision theory. This is because before preferences (utilities) or opinions (probabilities) can be combined, they must first be elicited and expressed in some quantitative form. Basically, Savage (1954a) —encompassing the works of Von Neumann and Morgenstern (1944) on game theory, of Wald (1950) on statistical decision functions, of de Finetti (1937) on subjective probability, and of Ramsey (1926) in general— proposed that through an axiomatic definition of rationality, preferences and beliefs of an idealised decision maker (DM) who has to choose amongst actions to take under uncertainty, should be modelled by utility functions and subjective probability assessments. In effect, the extension of those concepts to questions related to group decision making or to the aggregation of opinions and preferences as an expression of group consensus started then being formulated in the light of Savage's rationale. The problem here is that the Bayesian theory is a coherent normative theory for the individual (rational DM); and groups of individuals are left in difficulties since no concept equivalent to the classical notion of objectivity is available to them.

1.1.1 Impossibility results in group decision theory.

In 1951, Arrow showed an impossibility theorem (reconfirmed later in 1963 in a revised second edition) related to group decision theory in which there was no unrestricted algorithm able to give fair and democratic pooled group *preferences*. That is, a group's ranking (based on the group's expected utility) R_g could not be formed from the members' individual rankings R_i $(i = 1, \ldots, k)$ of strategies (actions) in a group decision problem [see also Kelly (1978)] unless it was a dictatorship. However, despite its evident drawback for the consensus of preferences, this result was proved under the assumption that there are no restrictions on the possible forms of the R_i's.

Also, impossibility results related to the aggregation of *beliefs* then started being obtained. For example, Dalkey (1972, 1975) showed that the linear opinion pool could only be extended to conditional probabilities if it were a dictatorship. Later, Lehrer and Wagner (1983) proved a similar result for independent events. Other impossibility theorems associated with linear pools can be found in Bacharach (1975), Genest (1984b) and Wagner (1984). More closely related to our work, Genest and Wagner (1984) proved that only dictatorships can accommodate pooling operators with a certain functional form when they are also required to hold for a property Laddaga (1977) called *independence preservation property*. This property requires that the pooling on joint events must equal the product of the poolings marginally for each event (in the same spirit of the statistical independence for probability functions), whenever the individuals preserve the events independence in their subjective assessments (see Section 2.2.1).

In general, contrarily to Lehrer's (1975, 1976, 1983) claims, all those impossibility results mean that no pooling rule for preferences or beliefs works sensibly well in *all* situations. Combining rules general enough to satisfy certain 'desirable' axioms, on the other hand, can fail to be democratic. Since the arrival of these results there have been two lines of work trying to circumvent those problems. One line, attempted to weaken group behavioural assumptions (Pill, 1971) in the hope that the impossibility results would vanish but with little success (Winkler, 1968). The other line, to which this thesis is also a contribution, has looked at weakening the insistence on universal applicability. In fact, by assuming that the group consensus is obtained for agreed belief networks with a particular type of structure, the application of the related impossibility theorems to our case can be prevented as we shall see in Section 6.6.

Because we will be concentrating on the issue of combining beliefs, we shall assume throughout this thesis that although the members of the group may disagree on their beliefs, they share the same preferences for the possible outcomes of unknown quantities. This implies that the group members agree to share the same utilities for the problem under consideration, and thus prevents Arrow's (1951) impossibility theorem (Raiffa, 1968, and Bacharach, 1975).

1.1.2 The axiomatic and the modelling approaches.

Some authors (de Finetti, 1951, and Lindley, 1985a) refer to the problem of aggregating the opinions (beliefs) of the members of a group of individuals expressed by subjective probability functions as the *reconciliation of probability assessments* while others (Winkler et al., 1986) as the *expert resolution*, the *opinion pooling* (Stone, 1961) or the *combination of subjective probability distributions* (Genest and Zidek, 1986).

There are basically two distinct forms of dealing with this sort of problem (French, 1985 and Winkler et al., 1986). One way is to adopt what is called the *axiomatic approach*, where basically the group is assumed to choose certain 'desirable' and 'reasonable' conditions (axioms) that the aggregating formula should obey and which at the end will characterise its form. Another possibility would be to adopt the *Bayesian modelling approach* where expert opinion is to be treated as information (data) in the Bayesian paradigm. To avoid operational problems with the modelling approach such as who specifies prior densities and likelihood functions (French, 1985 or Winkler, 1968) we shall employ the figure of a *supra-Bayesian* (SB) decision maker who at the end will be the responsible for performing the aggregation method. In particular, the use of the figure of a SB gives us the abstraction to study the circumstances under which reconciliation formulas obtained via modelling approach can be identified with opinion pools in group decision situations (Chapter 8). See Section 2.4.1 for further comments on the SB. Particular aspects of both the axiomatic and the Bayesian modelling approaches are investigated in this thesis.

There has been a lot of discussion in the literature about the merits of adopting one approach or another (Winkler et al., 1986, Lindley, 1985a, French, 1985 and Genest and Zidek, 1986). We shall not enter this discussion. However, we claim that the methods we propose, and therefore the results we obtain in this thesis, are not general purpose but appropriate to the specific configurations (situations) of the problems we define. In fact, it is well known (Winkler's comments – quoted below – on Genest and Zidek, 1986, pp.

139) that there is no general prescription which is suitable for every kind of aggregation problem, but "different combining rules are suitable for different situations, and any search for a single, all purpose, 'objective' combining procedure is futile". In fact, there are several factors which can influence the form of the most suitable combining model to be used in a practical situation, as for example:

(i) the nature of the problem itself (group decision or expert judgement problem),

(ii) the characteristics of the underlying sample space (finite or infinite), and even

(iii) the form in which the members express their opinions (full density functions, point estimates, log odds or a collection of quantiles), together with

(iv) other issues like dependence, calibration, coherence and honesty of the group members (Section 2.1).

1.2 Historical Remarks on Statistical Graphical Models.

The use of graphical association models to represent probabilistic information was first proposed in statistical physics by Gibbs (1902), and in genetics by Wright (1921, 1923, 1934). The latter developed a method called *path analysis* for the analysis of linear equations of continuous random variables. This method makes use of directed graphs where the random variables represented by vertices are linked by directed arcs (arrows moving from parents to children) indicating correlation and causality in a one-to-one correspondence. Rules were developed which allowed measurements of influence along each path in the graph as well as the degree to which an effect was determined by a particular cause.

Later, path analysis ideas were further developed and applied in social sciences, economics and statistics (Wold, 1954 and 1960, Blalock, 1971, and Kiiveri and Speed, 1982, amongst others). The discrete cases were also investigated by Birch (1963), Goodman (1970, 1973), Haberman (1974) and others. Those authors together with Vorobev (1962) realized that some decomposition characteristics of statistical tables could be best expressed in graphical terms for log-linear models.

Good (1961) used *directed acyclic graphs* (DAGs) to represent causal hierarchies of qualitative variables with disjunctive causes. A DAG, when representing a conditional independence structure, allows the associated joint probability function to be uniquely factorized into conditional probability functions (Smith 1989, 1990). Lemmer (1983) suggested the use of trees for Bayesian updating, and Spiegelhalter (1986) proposed the fill-in algorithm to transform Bayesian networks into join trees.

The development of Markov fields presumed that the graphical model (network topology) was given and the problem was to characterise the probabilistic behaviour of a system complying with the dependencies prescribed by such a model (Darroch et al., 1980, Wermuth, 1980, Lauritzen et al., 1984, and Lauritzen and Wermuth, 1984). A survey of Markov fields can be found in Isham (1981). Also, Lauritzen (1982) applied the theory of Markov fields to the analysis of statistical tables and derived some theorems for independencies embedded in strictly positive probability distributions. The Markov properties on DAGs were systematically studied by Kiiveri et al. (1984), Pearl and Verma (1987), Verma and Pearl (1990), Smith (1989), Geiger and Pearl (1990a) and Lauritzen et al. (1990), to name but a few.

1.2.1 Decomposable graphical models.

An interesting class of models are ones called *decomposable*, introduced by Darroch et al. (1980) for undirected graphs and further developed by Lauritzen and Wermuth (1984) for graphs with mixed (both qualitative and quantitative) variables and Leimer (1985, 1989, 1993). They have some attractive properties such as that their statistical analysis can be broken down into small analyses of sub-models in an elegant way. Also, models in this class can, under certain restrictions on the form of input data, retain conditional independence structure in prior-to-posterior analyses (as we shall see in Chapter 5). This fact is used intensively to create quick algorithms for calculating posterior distributions in high dimensional systems (Dawid, 1992, Jensen and Jensen, 1994, Smith and Papamichail, 1996). Other algorithms for manipulating decomposable graphs in an efficient way can be found in Rose et al. (1976) and Tarjan and Yannakakis (1984) for verifying decomposability and finding their cliques; in Tarjan (1985) and Leimer (1993) for finding optimal decompositions; and in Kjærulff (1990, 1992) for finding decomposable subgraphs.

1.2.2 Chain graphs.

Chain graphs (CGs), a special class of mixed graphs characterised by having no directed cycles (see Section 3.3), were first introduced in the literature by Lauritzen and Wermuth (1989) and Wermuth and Lauritzen (1990). Special classes of CGs are the partially complete CGs we define in Section 5.2 and the influence diagrams (see Section 3.5). The Markov properties of CGs can be found in Frydenberg (1990) as well as in Andersson et al. (1996) where further results on the conditions of Markov equivalence of CGs are obtained.

1.2.3 Influence diagrams.

Howard and Matheson (1981), from their experience in decision analysis, proposed the representation of a decision problem by an influence diagram (ID), that is, a DAG representing conditional independence statements. They showed the connection between decision trees and IDs as well as established the notion of arc-reversal corresponding to the application of Bayes theorem in IDs. More recently, Shachter (1986) proposed simplifications of the IDs and the related decision problem by introducing barren nodes and general results for arc-reversal and other operations. In this context, Verma (1987) proposed an elegant way of applying Bayes theorem to graphical models in general.

The Markov property of IDs was studied amongst others by Pearl (1986a) which proposed the *d-separation* theorem (see section 3.4) for characterising relevances in IDs. Also Smith (1989) generalised earlier results on factorisation of joint probability densities associated to IDs to arbitrary distributions (not necessarily strictly positive).

The modelling aspects of IDs which are in fact a very natural representation of Bayesian models were studied by Oliver (1986), Barlow (1986) and Rege and Agogino (1986) to name some but a few.

Other important results concerning efficient computation in probabilistic expert systems as represented by IDs are summarised in Lauritzen and Spiegelhalter (1988).

1.3 The Thesis Structure.

After presenting in this chapter some historical remarks on the topics we shall be dealing with in this thesis, the theory and the concepts related to the most important approaches to the aggregation of beliefs are revised in Chapter 2, together with the main weighting schemes for the axiomatic approaches. In Chapter 3, the notation and terminology as well as the basic theory of CGs and IDs are introduced.

A motivation for the proposed generalisations of EB to CEB opinion pools is shown in Chapter 4. Some intermediate results for the examples in this chapter are found in the Appendices A4.1 and A4.2.

In Chapter 5, the partially complete chain graph models are defined and the conditions for their *a posteriori* conditional independence preservation are stated.

The definition of *conditional external Bayesianity* and its characterisation by CEB pooling operators are shown in Chapter 6. The proof of such characterisation is presented in the Appendix A6. In Section 6.5, it is shown how some particular linear opinion pools (LinOps) can allow the joint density of variables associated with a conditional independence structure, to be factorized into conditional densities on that structure, that is, how LinOPs can be CEB.

In Chapter 7, we make use of influence diagrams to introduce the Supra-Bayesian (SB) analysis of the expert judgement problem. The issue of dependence among experts' information sources is revised through the modelling of diverse situations characterised by the way those sources overlap. The role played by the diverse overlapping structures when obtaining the posterior distribution for the quantity of interest is investigated through considerations of sufficiency of experts' statements in a degenerate situation. Also, it works as an introduction to the following Chapter 8 where a SB analysis of uncalibrated experts is employed to characterise those reconciliation rules which can also be externally Bayesian. This establishes a link between the axiomatic and the Bayesian modelling approaches to the aggregation of beliefs.

Chapter 9 contains comments on the main results and indicates some directions for further research.

CHAPTER 2
A REVIEW ON THE AGGREGATION OF BELIEFS

In this chapter we not only present a review about the main axioms for group decision situations as the *external Bayesianity* and the *marginalization* properties (that characterise the logarithmic and the linear opinion pools respectively), but also describe the main Bayesian modelling approaches for the aggregation of beliefs such as those of Winkler (1981) and Lindley (1985a). The notation and terminology related to the aggregation of beliefs which will be used throughout the thesis are introduced.

The revisions we make here are not extensive but orientated towards the assumptions and generalisations we propose in the thesis. In addition to that, we also provide short descriptions of some important well-known properties (axioms), concepts and methods which are relevant in the area although not employed directly in our developments. In those cases, the reader should refer to the provided bibliographies for further details on the particular subject of interest. Alternatively, the paper by Genest and Zidek (1986), for example, provides a comprehensive review and annotated bibliography for axiomatic approaches, while French (1985) defines the classes of problems most commonly found in aggregating opinions and reviews the main concepts and methods on both the axiomatic and the modelling procedures (French et al., 1989). Clemen (1989) also provides a review and an extensive bibliography for combining forecasts. Naturally the above mentioned reviews are of results obtained up to the date they were written. Also, the papers by Winkler et al. (1986) discuss the main issues about expert resolution in the Bayesian modelling methodology. Other surveys of related topics include those of Winkler (1968) on consensus of dependent subjective probability functions, Pill (1971) on the Delphi method, Beach (1975) on expert judgements, Hogarth (1975) on cognitive psychology and Weerahandi and Zidek (1981) on Multi-Bayesian decision making.

In Section 2.1, some issues of importance on the aggregation of beliefs are described such as the assessment of subjective probability quantities, coherence, calibration, honesty and non-independence between group's members. In Section 2.2, the most important axiomatic approaches are revised with some impossibility results being presented in Section 2.2.1, while McConway's (1981) *marginalization property* (MP) that leads to the *generalised linear opinion pools* (generalised LinOPs) of Genest (1984b) is revised in Section 2.2.2.

Madansky's (1964, 1978) *external Bayesianity property* is described in Section 2.2.3 and in Section 2.2.4 we can see the only type of pooling operator, the *modified logarithmic opinion pool* (modified LogOp), which under certain restrictions satisfy the external Bayesianity property. In Section 2.3, some methods for obtaining the weights in opinion pools are presented. A description and a review on the some Bayesian approaches can be found in Section 2.4.

2.1 Some Fundamental Issues in Aggregating Beliefs.

There are some issues which are of fundamental importance for any combination model in the aggregation of beliefs such as how the beliefs or uncertainties can be quantified, how good those quantifications are in expressing beliefs, as well as the amount of dependence between them. In fact, those models generally expect as inputs expert opinions coded as probability functions of one form or the other which are informative, somehow comparable, obey the laws of probability, tend in the long run to the true probability function of the target population and truthfully express the individuals' opinion. In many practical cases, none or just some the above assumptions hold. However, there are a number of ways of dealing with this problem as we shall see below.

2.1.1 Quantifying degrees of belief.

The question of how probability functions (probabilities, densities or mass functions) can be chosen to represent the structure of an individual's beliefs is inherent to the Bayesian paradigm and as such has been investigated by Ramsey (1926), Savage (1954), de Finetti (1970), Fine (1973), French (1982) and Lindley (1982c), amongst others. Because repeated independent measurements cannot be obtained from the same individual who is likely to remember his previous reasonings, the law of large number cannot be applied to reduce the measurement errors. Therefore, pragmatic tests (comparisons with reality) usually employed for estimating objective probability functions are not applicable. However, semantic and syntactic tests have been suggested (Lindley et al., 1979) which help in improving the quality of the assessments as can be seen in the following Sections 2.1.2 and 2.1.3. Some other difficulties encountered in the experimental elicitation of opinions are reviewed in Savage (1971).

The issue of whether probability functions are adequate to summarise individuals' or groups' information has also been the object of discussion with some arguing against (e.g.

Baird, 1985) and others supporting its use (Savage, 1954, de Finetti, 1970, French, 1986, Gardenfors and Sahlin, 1988 and Cooke, 1991). See also Apostolakis et al. (1988) and van Steen and Oortman Gerlings (1989) for discussions on this.

Certainly, the nature of the underlying measure space associated with the problem of interest, can help in determining the most suitable form for the probability quantity to be adopted. For example, if Ω denotes a collection of disjoint events, exactly one of which is true at a certain time, then a probability function P which obeys the axioms of probability over Ω should be assessed to each possible subset E of Ω, according to the degree to which this subset is believed to contain the fixed, but unobserved realization, say, $\theta \in \Omega$. However, in general, more than just point estimates will need to be produced with at least some measure of the individual's confidence in his/her assessment being also required. Moreover, important correlations may be lost if point estimates are used when several events are considered. Other alternatives to single probabilities might include confidence intervals, quantiles, cumulative distribution functions, and density functions, although it seems that single probabilities, odds and log-odds are equivalent as are cumulative functions and mass or density functions.

Note that the form in which the group members express their beliefs can influence not only the type of aggregation rule most suitable to be adopted in a problem but also its form (Genest and Zidek, 1986). In most of the cases in the thesis (first 6 chapters) we assume that the belief or uncertainty assessments are provided in the form of probability density functions (either discrete or continuous) by the members of the group of experts. In Chapter 7, the experts provide summary statistics as their statements and in Chapter 8 their probability assessments.

As far as the Bayesian modelling approach for the aggregation of beliefs is concerned, there are two levels in which uncertainty must be encoded. One, is that the group's members must encode their beliefs, and the other, is that the SB analyst must encode her synthesis of these (see Section 2.4). Naturally, in principle, there is no reason why the same methodology should be employed for both elicitations.

In some axiomatic approaches, in addition to the group members' belief assessments, the weights attributed to those members in the combination rule, can be interpreted as probabilities and as such must also be elicited, sometimes subjectively by the group (Section 2.3).

2.1.2 Calibration : measures and corrections.

Empirical results have shown that without training, few individuals are good probability assessors and even with training there is no guarantee that any individual expert can assess probability well and thus communicate his beliefs accurately (Kahneman et al., 1982, Arkes and Hammond, 1986, and Wright and Ayton, 1987). However, it is worthwhile pointing out that papers highlighting the difficulties are much more frequently cited (citational bias) than those showing what people can do well as probability assessors (Beach and Christensen-Szalanski, 1984, Beach et al., 1987). In fact, there is evidence that, with training, many people can effectively make better assessments (Alpert and Raiffa, 1982, Lichtenstein and Fischhoff, 1980a, and Phillips, 1987), especially in areas in which they have expertise (Lichtenstein and Fischhoff, 1980b, Christensen-Szalanski and Bushyhead, 1981, Cooke et al., 1988, and Cooke, 1991).

The quality of subjective probability assessments can be evaluated in terms of a semantic criterion called *calibration*. Briefly, a person is said to be well-calibrated "if the proportion of correct statements, among those that were assigned the same probability, matches the stated probability, i.e. if his hit-rate matches his confidence" (Lindley et al., 1979). See also Dawid (1982).

Perhaps the main factor that contributes to people (even experts) being miscalibrated is *over-confidence* (Lichtenstein et al., 1982). Other factors are the difficulty of the task, the experience and expertise in the area of interest, and training in encoding uncertainty.

Several methods for measuring calibration have been proposed. The most known measure is a quadratic scoring rule called Brier score proposed by Brier (1950) for meteorological forecasts. Murphy (1973) generalised the Brier score to multi-alternative items and showed that it can be partitioned in additive parts which measure not only calibration but also the assessor's ability to sort the events into sub-categories. Similar measures have been proposed by Adams and Adams (1961), by Oskamp (1962) and more recently by DeGroot and Fienberg (1982, 1983) who adopt proper scoring rules for comparing forecasters. In DeGroot and Fienberg (1982) there is a definition of conditionally well-calibrated forecaster in multivariate forecasting. Discussion about proper scoring rules with references is given by Staël von Holstein (1970). However good they are those methods rely on past data being available such that the assessor's performance can be measured by successive comparisons between statements and occurrences.

The interest in empirically investigating the calibration of assessments in terms of probability density functions started in 1969 with a paper only published later by Alpert and Raiffa (1982). Some other papers here are those by Brown (1973), Hession and MacCarthy (1974), Selvidge (1975), Moskowitz and Bullers (1978), and Lichtenstein and Fischhoff (1980b). For log-odds assessments see Seaver et al. (1978) and for a review on all of these see Lichtenstein et al. (1982).

In the context of correcting miscalibrated assessments Morris (1977) suggested, within the expert resolution problem, that it should be done through the use of calibration function obtained by the decision maker. The calibration function is obtained from an auxiliary experiment where exchangeable experts are asked to provide probability assessments on many unrelated variables and a performance function is built. Lindley (1982a) defines calibration as the process of adjusting a decision maker's likelihood function for the experts' assessments and, as Morris (1977), suggests the use of an auxiliary experiment. Agnew (1985) uses a similar approach for uncalibrated experts whose assessment errors have a multivariate normal density. Also, Cooke et al. (1988) propose a classical approach which considers measures of relative information to form a calibration score.

In Section 7.7 we also make use of calibration functions to obtain reconciliation rules in the context of the Bayesian modelling approach described in Section 2.4 (see also Section 1.1.2), which can be identified with multiplicative pools originated from the axiomatic approach (Sections 1.1.2 and 2.2).

2.1.3 Coherence and honesty.

Basically, coherence has to do with the members' probability assessments following the axioms of the probability theory. For example, the sum of probabilities of disjoint sets of events over their underlying sample space must add up to unity. Although in some simple situations it is relatively straightforward to obtain coherent probability assessments for a person with just basic knowledge of probability and statistics, in many other cases it is a difficult task even for well trained and experienced assessors. Specially in multivariate situations as the ones we deal with in the thesis, the joint probability functions can be broken down into conditional probability functions and the task of assessing them can be very demanding since the sample space in those cases is a product space (Section 6.2). In the case of probability assessments, Lindley et al. (1979) provide the means for which a coherent "investigator" can correct for a subject's incoherence.

There may be situations where it is advantageous for an assessor to deliberately misstate his/her probability function. A dishonest statement of opinion certainly has its implications when used in any decision making situation. A solution to this problem has been proposed with the use of proper scoring rules which would encourage the assessor's honesty. In fact, it is only by giving the actual subjective probability function as the true opinion that the assessor can maximise his/her expected score. However, this has its problems, for example, if the assessor's utility function is not a linear function of his/her score (Staël von Holstein, 1970, and DeGroot and Fienberg, 1983). Also, French (1985) gives an example when the use of proper scoring rules are not sufficient to ensure honest statements.

2.1.4 Non-independence between experts.

As mentioned before, inherent to the aggregation of multiple probability statements is the issue of non-independence (or dependence) among the assessors.

Within the Bayesian modelling approach, the non-independence between expert statements must be assessed by the decision maker through the specification of a joint likelihood function for those statements. This is not an easy task and simplifications have been proposed by several authors as Winkler (1981) and Lindley (1985a) —see Section 2.4. Also, in Chapter 7 we make use of graphical models to represent diverse situations of dependence and conditional independence between experts.

In the context of the axiomatic approaches, the issue of dependence has been avoided and none of the pooling formulas known so far takes explicit account of it. It seems that the only way in which the pooling rules could deal with the dependence issue, would be through the weights allowing for some discrimination on that basis. However, this still would be a non-elegant, vague and ill defined way. On the other hand, one might argue that it is rather unlikely that a group of individuals with conflicting opinions can reach an agreement on the amount their opinions overlap. For example, Genest and Zidek (1986) claim that "the absence of a group leader or an external decision maker may be one practical reason for avoiding the issue of opinion dependence".

Concerning its origins, the dependence between experts can come from experts having (i) overlapping data or information, (ii) overlapping methodology, and (iii) direct interaction with exchange of viewpoints. Overlapping information results from many situations where the assessors have access to the same basic information sources, that is, they use the same data to make their assessments. Overlapping methodology may exist if assessors have similar academic and professional training, when even having non-overlapping information they are expected to use many of the same modelling methods or the same sort of reasoning. Certainly, the direct observation of other assessor opinions, the presentation of public reports and the open discussion of viewpoints will add to the overlap among the assessors judgements. In this thesis we shall not consider the problems where direct interaction between experts occur. Also, we shall assume the overlapping of information as the only cause of dependence.

Note that, when the issues of dependence and miscalibration are jointly considered, it may well be reasonable for a decision maker to treat uncalibrated experts statements as being independent, even if those experts have overlapping information.

2.2 The Axiomatic Approach.

As mentioned before in Section 1.1.2, the axiomatic approach focuses on certain axioms or properties that a fair and democratic combining rule should obey in order to look appealing. The usual starting point of this approach is to consider a general hyper-function T called *pooling operator* (see Definition 2.1 below) which maps the k members assessed probability measures on Ω, i.e. $f_1, \ldots, f_k$ where $f_i : \Omega \to [0, \infty)$ $(i = 1, \ldots, k)$, and possibly the underlying σ-algebra S of events defined over Ω, into the interval $[0, \infty)$. This hyper-function T has then its form shaped according to the axioms the group chooses to follow.

In fact, the axioms (properties) are restrictions applied to T. We describe in Section 2.2.1 the main axioms found in the related literature and show that certain properties when required to apply simultaneously can restrict the class of eligible pooling operators to the extreme degree of dictatorships being the only possible solutions.

2.2.1 Pooling operators, axioms and impossibilities.

Now, we define T as a class of pooling operators which are essential components of the usual characterisation of some of the properties we describe in the following sections.

Definition 2.1 (Pooling Operator). *Let (Ω, μ) be a measure space. Let Δ be the class of all μ-measurable functions $f : \Omega \to [0, \infty)$ such that $\int f d\mu = 1$ with μ almost everywhere (a.e.). A pooling operator $T : \Delta^k \to \Delta$, is that one which maps a vector of functions $(f_1, ..., f_k)$, where $f_i \in \Delta$ (for all $i = 1, ..., k$), into a single function also in Δ.*

Because a pooling operator can also be defined for when Δ is the class of functions $f : \Omega \to [0,1]$, we shall indicate when this is the case in the text below.

As examples of restrictions on T we can mention McConway's (1981) *Weak Setwise Function Property* (WSFP), viz. :

$$T(f_1,\ldots,f_k)(x) = F[x, f_1(x),\ldots,f_k(x)] \quad \forall x \in \mathcal{S} \ , \tag{2.1}$$

where $\Delta = \{f : \Omega \to [0,1]\}$, and $F : \Omega \times [0,1]^k \to [0,1]$, which is equivalent to the *Marginalization Property* (MP) we define in Section 2.2.2, when the underlying σ-field $\mathcal{S}$ is tertiary. The WSFP together with the *Zero Probability Property* (ZPP), viz :

$$f_1(x) = \cdots = f_k(x) = 0 \implies T(f_1,\ldots,f_k)(x) = 0 \quad \forall x \in \mathcal{S} \ , \tag{2.2}$$

is equivalent to the *Strong Setwise Function Property* (SSFP), viz :

$$T(f_1,\ldots,f_k)(x) = G[f_1(x),\ldots,f_k(x)] \quad \forall x \in \mathcal{S} \ , \tag{2.3}$$

where $G : [0,1]^k \to [0,1]$. The important result here proved by McConway (1981) is that the MP together with the ZPP is equivalent to the SSFP and restricts T to be a linear combination of $f_1,\ldots,f_k$ (Section 2.2.2).

On the other hand, Wagner (1982) showed that for tertiary σ-fields the WSFP, together with the *external Bayesianity* property (Definition 2.3 in Section 2.2.3), restricts T to a dictatorial form, i.e. $T(f_1,\ldots,f_k) = f_i$ for a fixed i. However, Genest (1984c) and Genest et al. (1986) circumvented this impossibility by introducing a normalised form of both WSFP and SSFP which together with external Bayesianity leads a form of T to a logarithmic (multiplicative) form, as we shall see in Section 2.2.3. Obviously, the logarithmic pools take the ZPP to the extreme, being sufficient that just one of the f_i's be zero. To avoid this undesirable fact, we shall invoke Cromwell's rule (Lindley, 1982b) to support our assumption that $f_i > 0$ for all $i = 1,\ldots,k$ in the logarithmic pools.

Another property called *Independence Preservation Property* (IPP) was proposed by Laddaga (1977). It requires that one should have

$$T(f_1,\ldots,f_k)(x \cap y) = T(f_1,\ldots,f_k)(x)T(f_1,\ldots,f_k)(y) \ , \tag{2.4}$$

whenever $f_i(x \cap y) = f_i(x)f_i(y)$ for some x and y in $\mathcal{S}$ $(i = 1,\ldots,k)$. Note that Laddaga's IPP demands the pooling rule to preserve the independence amongst distinct events and not the independence among the members' (experts') opinions about events. In fact, the works on axiomatic approaches avoid the issue of dependence between the members of the group which is better investigated within the modelling approaches – see Section 2.4. Also in this regard, Lehrer and Wagner (1983) proved that linear pools cannot satisfy the IPP unless they are *dictatorial*, i.e. $T(f_1, \ldots, f_k) = f_i$ for some fixed i. This impossibility result is indeed related to that of Dalkey (1972, 1975) who showed that only dictatorships could satisfy the SSFP when it is also required to hold for conditional probability functions such that

$$T(f_1, \ldots, f_k)(x|y) = G[f_1(x|y), \ldots, f_k(x|y)] \ . \tag{2.5}$$

Bordley and Wolf (1981) in a critical review of Dalkey's work, argue against the SSFP in which it is inconsistent with the way probabilities are aggregated in Bayesian models. They claim without proof that Bayes' theorem results in an equation of the form (2.1), the WSFP. On the other hand, McConway (1981) argues against Dalkey's impossibility theorem in that it requires the same G to apply to both conditional and unconditional probability functions.

A more general and stronger result is that of Genest and Wagner (1984), we mentioned in Section 1.1.1, in which no pooling operator of the form

$$T(f_1, \ldots, f_k)(x_j) = \frac{F_j[f_1(x_j), \ldots, f_k(x_j)]}{\sum_{j=1}^{n} F_j[f_1(x_j), \ldots, f_k(x_j)]} \ , \tag{2.6}$$

where $F_j : (0,1)^k \to (0,1)$ is also allowed to depend on the identity of x_j $(j = 1, \ldots, n)$, except dictatorships can also obey the IPP when Ω is at least quarternary (continuous case included). For Ω with exactly four points, they showed that the only class of pools which satisfy both (2.6) and IPP has the form

$$T(f_1, \ldots, f_k)(x_j) \propto \prod_{i=1}^{k} [f_i(x_j)]^{b_i} e^{\{a_i f_i(x_j)[1-f_i(x_j)]\}} \ , \tag{2.7}$$

for a_i and b_i arbitrary real constants and for at least one of the F_j's in (2.6) Lebesgue measurable $(j = 1, \ldots, n)$. Note that the logarithmic pool is a particular case of (2.7) when all the a_i's are set to zero.

Although the IPP can look appealing at a first glance, it is a very demanding criterion. In particular, it demands that if experts happen to assign probability distributions which agree that any two functions are independent —even if this is an accidental artifact of the probability assignments— then T must preserve this independence. Although this may well be, it is exceedingly fierce to demand independence preservation for all events in the case where the agreement of independence originates from commonly held logical arguments that the group would like T to preserve.

However, in this thesis —and also in Faria and Smith (1994, 1996, 1997)— we propose much weaker conditions for the prior-to-posterior preservation of conditional independence structures. The preservation is demanded *only* for logical associations encoded by particular types of chain graphs or influence diagrams (see Chapter 5). The pooling operators that are devised are conditional on a form which is a generalisation of the WSFP for conditional variables with the equality sign being replaced by a proportionality symbol in (2.1) —see formula (6.1) in Section 6.2. They also obey the external Bayesianity property on those conditional variables and are limited to be applied only to components of a given conditional independence preserving structure which is agreed to be logically supported by the group. The consensus on a variable conditioned in another, say $x_j|y_j$, may depend on occurred (fixed) values of the conditioning variable y_j which influence outcomes of x_j (Chapter 4). Thus, the type of impossibility results obtained by Dalkey (1972, 1975) and Genest and Wagner (1984) are only very loosely connected to our characterisation.

2.2.2 The marginalization property and the linear opinion pool.

Suppose that, in the context of the expert judgement problem or in the context of the group problem where a supra-Bayesian (SB) is assumed to perform the pooling, a combination of the group members opinions must be obtained from their subjective probability assessments over some space Ω of states of nature. The members agree not only on the space Ω but also on the σ-algebra $\mathcal{S}$ over Ω they use.

A σ-algebra of events $\mathcal{S}$ over Ω is characterised by an infinite set of events $X_1, \ldots, X_n, \ldots$ belonging to $\mathcal{S}$ such that $\cup_{i=1}^{n} X_i \in \mathcal{S}$. In this sense, the marginalization of a probability measure defined on a σ-algebra $\mathcal{S}$ to a probability measure on a sub-σ-algebra $\mathcal{V}$ of $\mathcal{S}$ consists of simply restricting its domain from $\mathcal{S}$ to $\mathcal{V}$. If the probability measure is a density function, the marginalization can be obtained by summing or integrating out the events or quantities in $\mathcal{S}$ but not in $\mathcal{V}$. The resulting density is then called a marginal density.

Now, we can define the marginalization property as follow:

Definition 2.2 (Marginalization property). *If the same combined probability density function for the group is obtained whether (a) the members opinions are first combined into a single density over $\mathcal{S}$ and then a marginal density taken, or (b) the group members each give their marginal densities over $\mathcal{V}$ and a combination is formed from these, then the combined density is said to obey the Marginalization Property (MP).*

As mentioned in the last section, McConway (1981) proved that the *linear opinion pool* (LinOp) of Stone (1961), to which Bacharach (1979) attributes to Laplace,

$$T(f_1,\ldots,f_k) = \sum_{i=1}^{k} w_i f_i \quad , \tag{2.8}$$

where $\sum_{i=1}^{k} w_i = 1$ with w_i non-negative, is the only pooling formula to satisfy both the WSFP in (2.1) and the MP when Ω contains at least three points. Generalising this result, Aczél et al. (1984) and Genest (1984b) proved that the *generalised linear opinion Pool* (generalised LinOp) below, is under the WSFP the only type of pool to satisfy the MP for a tertiary σ-field on Ω :

$$T(f_1,\ldots,f_k)(x) = \sum_{i=1}^{k} w_i f_i(x) + [1 - \sum_{i=1}^{k} w_i] g(x) \tag{2.9}$$

for all $f_1, ..., f_k \in \Delta$ and $x \in \mathcal{S}$, where $g \in \Delta$ is a fixed arbitrary probability measure and the weights $w_i \in [-1, 1]$ are such that $|\sum_{i=1}^{k} w_i| \leq 1$. Obviously, the LinOp is a particular case of the generalised LinOp when $\sum_{i=1}^{k} w_i = 1$. This corresponds to the ZPP being required for (2.9).

The possible negative weights in (2.9) are of difficult interpretation as pointed out by Genest (1984b) despite the efforts of Winkler (1981) and Genest and Schervish (1983) to give them some meaning in specific contexts. However, if the pooling is also required to follow a *dominance principle* (Aczél et al., 1984) in which

$$T(f_1,\ldots,f_k)(x) \leq T(q_1,\ldots,q_k)(x) \quad ,$$

whenever $f_i(x) \leq q_i(x)$ for all $i = 1,\ldots,k$, then the weights in (2.9) will be nonnegative. Genest (1984b) showed that negative weights may only occur if the σ-field associated with Ω is finite.

Also, regarding the weights in LinOps some authors (e.g. Laddaga, 1977 and Bordley and Wolf, 1981) have argued that because the experts opinions may change from one event to the next, that the weights should be allowed vary with it, that is $w_i(x)$ should replace w_i in (2.8) and (2.9) above. However, albeit intuitively appealing this is incompatible with the results of McConway (1981) and Genest (1984b) which imply (2.8) and (2.9) with weights not depending on x.

2.2.3 The external Bayesianity property.

In the context of the group decision problem, a hypothetical case is considered where a group of Bayesians jointly receive new information Y about a random variable X and agree a common likelihood function $l(X|Y)$ associated with the data y. Thus, the *external Bayesianity property* can be defined as follows :

Definition 2.3 (External Bayesianity property). *If whatever the common likelihood function $l(x|y)$, the group's combined probability density function, updated by Bayes rule using $l(x|y)$, produces the same posterior density as if the combination has been performed after each individual in the group has revised his opinion also using Bayes rule on $l(x|y)$, then that group is said to obey the external Bayesianity property and is called an externally Bayesian (EB) group.*

Briefly in other words, an EB pooling policy ensures that the combination rule will give the same result a posteriori, independently of being obtained before or after the members of the group update their beliefs when receiving a new data y for which they agree an associated likelihood $l(x|y)$ on the variable x.

Madansky (1964, 1978) characterised the EB pooling operators as those synthesising the diverging individual opinions $f_1, ..., f_k$ into a *group opinion* expressed by a single density $\bar{f} = T(f_1, ..., f_k)$ that must satisfy

$$T\left(\frac{l.f_1}{\int l.f_1 d\mu}, ..., \frac{l.f_k}{\int l.f_k d\mu}\right) = \frac{l.T(f_1, ..., f_k)}{\int l.T(f_1, ..., f_k) d\mu} \quad , \quad \mu \text{ a.e. }, \tag{2.10}$$

where $l : \Omega \to (0, \infty)$ is the group's common likelihood function for all the data that all the members observe, such that $0 < \int l.f_i d\mu < \infty$ $(i = 1, ..., k)$. Condition (2.10) has also been considered under other denominations like *data independence preservation* by McConway (1978) and *weak likelihood ratio axiom* by Bordley (1982) and Morris (1983). Notice that as the condition (2.10) allows the pooling of posterior densities to be achieved by the pooling of prior densities through the commonly agreed likelihood l, T can be said to be a *prior-to-posterior coherent* (PPC) pooling operator (Weerahandi and Zidek, 1978).

In fact, the PPC property is more general than the external Bayesianity property in that the requirement of a common likelihood for the group is relaxed. In this sense, a PPC pooling operator must satisfy

$$T\left(\frac{l_1 f_1}{\int l_1 f_1 d\mu}, \ldots, \frac{l_k f_k}{\int l_k f_k d\mu}\right) = \frac{T(l_1, \ldots, l_k) T(f_1, \ldots, f_k)}{\int T(l_1, \ldots, l_k) T(f_1, \ldots, f_k) d\mu} \quad ; \quad \mu \text{ a.e. }, \tag{2.11}$$

where $l_i : \Omega \to (0, \infty)$ is the i-th member likelihood for the new data y informative about x such that $0 < \int l_i f_i d\mu < \infty$ $(i = 1, \ldots, k)$. This means that the group's combined probability density function, updated by Bayes rule using the group's combined likelihood $T(\underline{l})$ where $\underline{l} = (l_1, \ldots, l_k)$, produces the same posterior density as if the combination had been performed after each member i in the group has revised his opinion also using Bayes rule on $l_i(x|y)$ $(i = 1, \ldots, k)$. Note that the external Bayesianity property is a particular case of PPC when $l_i = l$ for all $i = 1, \ldots, k$ and $T(\underline{l}) = l$.

Also, notice that while the external Bayesianity property can be invoked only when the likelihood l is *common knowledge* to the group (see Chapter 5), the PPC property demands all the individual likelihoods $l_1, \ldots, l_k$ to be common knowledge to the group.

2.2.4 The modified logarithmic opinion pool.

Genest et al. (1986) proved that the following *modified logarithmic opinion pool* is, under certain regularity conditions discussed later, the most general externally Bayesian logarithmic pool satisfying (2.10) :

$$T(f_1, ..., f_k) = \frac{g \prod_{i=1}^{k} f_i^{w_i}}{\int g \prod_{i=1}^{k} f_i^{w_i} d\mu} \quad , \quad \mu \text{ a.e. }, \tag{2.12}$$

where $g : \Omega \to (0, \infty)$, in general, is an arbitrary bounded function on Ω and w_i, $i = 1, ..., k$, are arbitrary weights (not necessarily nonnegative) adding up to one. The w_i's are experts' opinions weights in the combination and must be suitably chosen to reflect relative expertise. They should possibly be elicited by the experts based on their common knowledge about their own relative predictive capabilities.

One of the above mentioned regularity conditions on (Ω, μ) is that for an existing Lebesgue measurable function $P : \Omega \times (0, \infty)^k \to (0, \infty)$, the pooling operator $T : \Delta^k \to \Delta$ is such that

$$T(f_1, ..., f_k)(x) = \frac{P[x, f_1(x), ..., f_k(x)]}{\int P(\cdot, f_1, ..., f_k) d\mu} \quad , \quad \mu \text{ a.e.}. \tag{2.13}$$

This condition, a relaxation of the WSFP in (2.1), is called *likelihood principle*. This is because it restricts the likelihood of the combined density T at a particular point x in Ω to depend, except for a normalising constant, only on that x and on the individual densities f_i $(i = 1, ..., k)$ assigned to x, and not upon other points and densities of the points which might have occurred but did not. It is illustrated in Examples 4.1 and 4.4 in Chapter 4, that (2.13) is a strong and rather arbitrary requirement particularly in the case where T is a multivariate density modelling a causal structure of associations between variables. In the generalisation we make in Chapter 6 we require (2.13) to hold only for marginal and conditioned variables related to individual nodes in a graphical representation structure.

Another assumption which they make here is that the underlying space (Ω, μ) can be partitioned in at least four non-negligible sets. In this case such a measure space is called *quarternary*. This therefore includes the case where μ is Lebesgue and many (but not all) cases where μ is a counting measure.

McConway (1978) proved that in the case where Ω is countable and μ is a counting-type (purely atomic) measure then the formulas of the type (2.13) would be the only ones to qualify as EB if (2.13) holds. In the case where Ω is purely continuous (excluding thus the case where the sample space is countable), Genest (1984) showed that the only non-dictatorship externally Bayesian pooling operator satisfying (2.13), but when P is not indexed by x, i.e.

$$T(f_1, ..., f_k)(x) = \frac{P[f_1(x), ..., f_k(x)]}{\int P(f_1, ..., f_k)d\mu}\ , \mu \text{ a.e.}\ , \tag{2.14}$$

is the *logarithmic opinion pool* (LogOp) :

$$T(f_1, ..., f_k) = \frac{\prod_{i=1}^{k} f_i^{w_i}}{\int \prod_{i=1}^{k} f_i^{w_i} d\mu}\ , \mu \text{ a.e.}\ , \tag{2.15}$$

where $w_i \geq 0$ $(i = 1, ..., k)$ are *arbitrary constants* such that $\sum_{i=1}^{k} w_i = 1$.

The major intrinsic problem with the modified LogOp (2.12) in practice, is the difficulty of choosing the essentially bounded function g (which is even difficult to interpret). At a first glance one could be tempted to conclude that the function g could represent a decision maker's *prior* distribution for X. But according to Genest et al. (1986), even if the pooling rule (2.12) is adopted by a decision maker, there is no reason to assume in advance a bounded prior density. Nevertheless, in the context of our group decision problem, if *unanimity preservation* is required, i.e. $T(f, ..., f) = f$, $\forall f \in \Delta$, then g would be forced to unity and (2.12) would be reduced to an ordinary LogOp of the form (2.15). We shall employ combination rules which preserve unanimity of conditional independence structures in the examples related to the axiomatic approaches in this thesis.

2.2.5 A problem with external Bayesianity.

The EB group has several advantages. For example Raiffa (1968) illustrated how the relevance over the order in which the pooling and updating are done can lead to subjects trying to increase their influence on the consensus by insisting that their opinions be computed before the outcome of an experiment is known. Being EB all such argument would be pointless. Also, as Genest (1984a) points out, if such a pool can be agreed then it has great practical advantages. For example its members need not meet again after data has been observed.

Unfortunately as currently developed, EB pools have serious drawbacks. Perhaps the most obvious one is that, as Lindley (1985) points out, it appears perverse that the weights $\underline{w} = (w_1, ..., w_k)$ must a priori be common knowledge to the group (see Section 2.4.5). Surely as evidence appears which sheds light on the relative expertise of its members, the group should agree to adapt its weights to favour the better forecasters.

In Chapter 4, we use some examples to show the above mentioned difficulty with EB pools and to point out how an extension of external Bayesianity to conditional external Bayesianity in multivariate group decision situations can allow the weights to adapt reflecting the members' relative expertise. Other issues specific to multivariate problems are outlined in Chapter 6 where we generalise the external Bayesianity properties so that on the one hand they are suitable for the analysis of multivariate structures and on the other they allow the group to learn about the combination weights vector $\underline{w}$.

2.3 Some Weighting Schemes.

One of the major difficulties associated with the opinion pools in the axiomatic approach, is the question of obtaining the group members' weights to be used on those rules, like formulas (2.8) and (2.15). This is because there is, up to date, no normative theory behind them to support their choice. However, there are several operational methods proposed, most of them specifically developed for linear pools.

Winkler (1968) suggested four possibilities :

(i) attribute equal weights to all the members,

(ii) set the weights proportional to a ranking based on the members' "goodness" as assessors,

(iii) allow the members to weight themselves, and

(iv) use of proper scoring rules.

Also, DeGroot (1974) developed a more mathematical than practical procedure for Winkler's suggestion (iii) in which each member provides weights for a LinOp and the pools then obtained are applied repeatedly to the members' judgements. Other methods on appropriate updating of weights can be found for example in Mendel and Sheridan (1987), Bayarri and DeGroot (1988), DeGroot and Mortera (1991) and Cooke (1991). Also, Cooke (1991) comments on the main problems in developing a theory of weights.

The objective of this section is to illustrate some weighting schemes that could be employed together with the axiomatic approaches we have described above as well as with the generalisation we will be proposing of external Bayesianity to conditional external Bayesianity. As we shall see, this generalisation is such that it allows the group to learn about the combination's vector of weights. This responds at least in part to some of the criticisms (see Section 2.4.4) about the inflexibility of the weights on the well discussed LogOps which usually are not allowed to vary with the members' individual expertise (see Section 4.1). Some examples on how the weights are allowed to be updated based on the individuals' relative expertise on causal variables are given in Chapter 4. An example with deterministic weights is given in Section 4.2.

It would be interesting in group decision situations that the group itself could choose the weights on the basis of its members commonly held information about their specialities and expertise. However, a group which cannot or is not prepared to agree about the member's relative expertise may prefer to adopt a class of combination rules which sequentially adapts the weights based on each individual relative predictive performances as done by methods such as the *optimal combination* of Bates and Granger (1969), the *quasi-Bayes* approach of Smith and Makov (1978), and by the *outperformance* of Bunn (1975, 1978), we describe in the following sections.

For the methods we describe below, we shall assume that for an independent time series X_t, the variables $X_1, X_2, X_3, ...$ are strictly comparable in the sense that success by member i in predicting X_t, after having seen $x_1, \ldots, x_{t-1}$ is strictly comparable with his success in predicting X_{t+1}, after having seen $x_1, \ldots, x_t$. Examples when this assumption is reasonable include the case when $X_1, X_2, \ldots$ is a homogeneous time series.

2.3.1 The optimal combination of Bates and Granger.

The first general analytical linear model for combining forecasts (point estimates) was proposed by Bates and Granger (1969). Because it minimises the variance of the combined forecast error for estimating the weights of each individual forecast in the linear pool, this model was called *optimal combination*. In this method the weight of each forecaster i in a particular time t, can be given by the convex combination

$$w_{i,t} = \lambda w_{i,t-1} + (1-\lambda)\frac{{S_i}^{-1}}{\sum_{i=1}^{k} {S_i}^{-1}} \tag{2.16}$$

where $S_i = \sum_{j=t_0}^{t-1} {e_{i,j}}^2$; $e_{i,j} = x_j - \mathbb{E}[f_i(x_j)]$; t_0 is an arbitrary initial time; $f_i(x_j)$ is the i-th forecaster predictive density for the random variable X at the time j and λ is a factor used to give relative weight between the past w_i and the function of the squares of $(t-1)-t_0$ past errors.

The optimal combination has been shown to be a robust and efficient method (Newbold and Granger, 1974, and Winkler and Makridakis, 1983). It is also expected to present good performance over medium and large samples (Menezes and Bunn, 1990, 1991) which motivates its use for comparison with other methods (Faria and Souza, 1995).

2.3.2 Bunn's outperformance approach.

The need for a not very complex Bayesian approach that would allow the incorporation of subjective opinion into the synthesis of forecasts has justified, at least in part, the appearance of the *outperformace* method proposed by Bunn (1975), also for linear combination of forecasts but based on probabilities with a simple non-parametric interpretation.

Outperformance interprets a component $w_{i,t}$ $(i = 1, ..., k)$ of the weights vector $\underline{w}_t$ as the probability that the member i will produce the most appropriate forecasting model (the closest to a good model) of X_t. The sufficient statistics with which to learn about $\underline{w}_t$ is assumed to be both $\underline{w}_{t-1}$ and the identity of the member whose model produced the closest forecast of X_t. This identity, viewed as a random variable, is assumed to follow a Multinomial $(1, \underline{\theta}_t)$ distribution while the $\underline{w}_t$ is interpreted as the prior mean of $\underline{\theta}_t$. The parameter vector $\underline{w}_t$ is then successively updated in the usual Bayesian framework in the light of forecasts. Using this method with the assumption that the i-th member relative expertise about X_t is independent of every other member expertise about X_t, it is easily checked that for $t \geq 1$

$$w_{i,t} = (1-\rho_{t-1})w_{i,t-1} + \rho_{t-1}(t-1)^{-1} r_{i,t-1} \tag{2.17}$$

for $i = 1, ..., k$, where $\rho_{t-1} = (t-1)/[\bar{\alpha}_{t-1} + (t-1)]$, with $\bar{\alpha}_{t-1} = \sum_{i=1}^{k} \alpha_{i,t-1}$, where

$\alpha_{i,t-1}$ are the parameters of the conjugate Dirichlet prior distribution of $\underline{\theta}_t$, and $r_{i,t-1}$ is the number of successes of forecasting model i up to time $t-1$.

A more attractive formulation of this approach (Bunn, 1978) allows the probability $\theta_{il,t}$ that the member i will produce a more appropriate model than member l at time t, to be revised in the light of all the members relative performances. Pairwise comparisons between models are set up and a relative performance ranking is obtained. The weight $w_{il,t}$ is assumed to be the posterior mean of the $\theta_{il,t-1}$ whose density function is now assumed to be a beta with parameters $(\alpha_{il}, \alpha_{li})$. These parameters are updated in the usual Bayesian way. Also assuming outperformance independence among estimators, the estimate of the probability of model i outperform all other models, $w_{i,t}$, can be obtained for $l \neq i$ as

$$w_{i,t} \propto \prod_{j=1}^{k} w_{ji,t} \quad . \tag{2.18}$$

Such a method of updating weights transfers directly onto both the LinOp (2.8) and the modified LogOp (2.12) discussed in Sections 2.2.2 and 2.2.4 above.

Notice that the rule is fair to members in the group in that it is symmetric in their success if a priori we set $w_{i,0} = w_{l,0}$ for $i, l = 1, ..., k$. It also has the desirable property that a member whose models consistently produce closest forecasts has those models ever more closely followed by the group. However, there has been some criticism of this approach on the basis that while other methods such as the quasi-Bayes we describe in the next section are intentional approximations, the outperformance is an out ranking method built upon unintentional mistakes (French and Bunn, 1980).

2.3.3 The quasi-Bayes approach of Smith and Makov.

Although originally developed for signal detection and pattern recognition, where unsupervised learning and sequential classification play an important role, the *quasi-Bayes* approach of Smith and Makov (1978) is suitable for any situation where it is possible to make an extensive study of the distributions of observations belonging to individual populations, so that it can be completely specified but the mixture of these populations is unknown.

The quasi-Bayes consists of a simplification of the sequential Bayesian estimation procedure, which presents a serious problem in the calculation of the posterior probability distribution for the vector of weights $\underline{w}_t$ in linear pools. In fact, Bayes rule applied for estimating those weights gives at a time t the posterior distribution

$$f(\underline{w}_t|\underline{x}^t) \propto \mathcal{L}(x_t|\underline{w}_t)f(\underline{w}_t|\underline{x}^{t-1}) \tag{2.19}$$

where $\underline{x}^r = (\underline{x}_1, ..., \underline{x}_r)$, $f(\underline{w}_t|\underline{x}^{t-1})$ is the weights predictive distribution at time $t-1$ and $\mathcal{L}(\cdot)$ the likelihood function. However, successive computations of (2.19) cause a combinatorial explosion of terms in the likelihood function when this likelihood is identified with the linear pool (2.8) applied to $x_t|\underline{w}_t$. The quasi-Bayes approach simplifies (2.19) by attributing a Dirichlet with parameters $(\alpha_1, \ldots, \alpha_k)$ as a prior distribution for the weights which are then sequentially estimated by a posterior update on those parameters. This gives for $i = 1, \ldots, k$

$$w_{i,t} = \frac{\alpha_{i,t}}{t + \sum_{j=1}^{k} \alpha_{j,t}} \tag{2.20}$$

where

$$\alpha_{i,t} = \alpha_{i,t-1} + \frac{f_i(x_t)\alpha_{i,t-1}}{\sum_{j=1}^{k} f_j(x_t)\alpha_{j,t-1}} \quad . \tag{2.21}$$

Therefore, the weights in this approach are obtained by considering the likelihoods of the individual predictive densities as measures of distance between observations and population means.

2.4 The Bayesian Modelling Approach.

The Bayesian modelling approach for combining the opinions of the members of a group treats the individuals probability assessments as *data* within the usual Bayesian framework. For this, because a likelihood function for those individuals assessments must be elicited, we shall make use of the figure of a *Supra-Bayesian* (SB) decision maker who is supposed to be *herself* (at least) an expert on the members' assessing abilities (see Section 2.4.1).

The first conceptual and methodological Bayesian framework for "the use of experts in decision situations" was introduced by Morris (1974, 1977) —see also Roberts (1965)— and built upon by Lindley et al. (1979), French (1980, 1981), Winkler(1981), Bordley (1983), Lindley (1982c, 1983, 1985a, 1987), Agnew (1985), Clemen and Winkler (1985), Mendel and Sheridan (1987), West (1989), Wiper and French (1995) and Wiper and Pettit (1996), among others.

After introducing the general Bayesian modelling approach (Section 2.4.2) we refer to the question of how the SB's likelihood function for the members can be simplified when conditional independence assumptions can be made (Section 2.4.3) and then describe in more detail some modelling methods we refer to in the thesis (Sections 2.4.4 and 2.4.5).

2.4.1 The supra-Bayesian decision maker.

The term *supra-Bayesian* (SB) was first used by Keeney and Raiffa (1976) referring to a fictitious decision maker who would represent the "synthetic personality" of a group of individuals (Hogarth, 1975, p. 282).

We use that term here as a general designation for the (feminine) figure of a Bayesian decision maker (fictitious or not) who is responsible for somehow collating the group members' beliefs.

In group decision situations where there is a group leader then this leader is the natural SB decision maker. In other cases, such as the "jury problem", there may be enough cohesion (or affinity) in the group albeit some differences of opinions so that a fictitious SB would be justifiable. The problem with the virtual SB within the Bayesian modelling approach is that a likelihood function for the members' statements must be elicited as well as a prior distribution for the variable (or vector) of interest and this task would fall on the group. French (1985) proposes a benevolent SB who would start with a vague prior over the σ-field $\mathcal{S}$ of interest and then would learn about $\mathcal{S}$ only through the members' beliefs. The absence of a group leader or an external decision maker would imply in the group itself having to construct the SB by agreement about how she should behave.

2.4.2 The general supra-Bayesian analysis.

A quite general way of viewing the SB analysis of the aggregating problem is to see it as the problem of a SB decision maker seeking for expert advice to make use of the most relevant information available to effectively make better inferences about the unknown parameters $\Theta = (\underline{\theta}_1, \ldots, \underline{\theta}_n)'$ of statistical models she attributes to uncertain (future) events or quantities of interest $\underline{X} = (X_1, ..., X_n)'$. The parameter matrix Θ is composed of vectors $\underline{\theta}_j = (\theta_{j1}, ..., \theta_{j,d(j)})'$ of dimension $d(j)$ where each element θ_{jl} $(l = 1, ..., d(j))$ represents the unknown parameter of the model attributed to X_j $(j = 1, ..., n)$.

Typically, the relevant information associated with the expert problem consists of :

(a) other events or quantities $\underline{Z} = (Z_1, ..., Z_m)'$, not in $\underline{X}$, which are also informative about Θ (these could include past observations of $\underline{X}$), and whose parameters associated to their underlying statistical models are $\Phi = (\underline{\varphi}_1, ..., \underline{\varphi}_m)'$, where $\underline{\varphi}_j = (\varphi_{j1}, \ldots, \varphi_{jd'(j)})$ is of dimension $d'(j)$ for $j = 1, \ldots, m$;

(b) experts' statements, $\mathbf{Q} = (\mathbf{Q}_1, ..., \mathbf{Q}_k)$, which can be objective reports or subjective opinions they have related to Θ, and which could possibly be obtained from $\underline{Z}$;

and

(c) how the SB has access to what information (about variables, parameters and the experts) and how for the SB that information relates to each other in the particular configuration of the problem being considered.

On the other hand, an expert statement $\mathbf{Q}_i = (\underline{q}_{i1}, ..., \underline{q}_{in})$, where each component vector $\underline{q}_{ij} = (q_{ij1}, ..., q_{i,j,d(j)})'$ with q_{ijl} being the expert's E_i statement about the parameter θ_{jl} of Θ ($i = 1, ..., k$; $j = 1, ..., n$ and $l = 1, ..., d(j)$), is usually based on all information he has about quantities associated to $\underline{X}$ and which characterises his information base $\underline{I}_i$. Those statements can be characterised by :

(i) summary statistics obtained from random samples of $\underline{Z}$ (Clemen, 1987),

(ii) subjective discrete probabilities when $\underline{X}$ is a vector of events (Lindley, 1985a),

(iii) quantiles partially specifying density functions (Mendel and Sheridan, 1987),

(iv) the values of parameters fully specifying subjective probability density functions (Winkler, 1981), or

(v) measures of location and scale only, where specifications of density functions are not required (Lindley, 1983).

We shall assume throughout the paper that observations of the quantities $\underline{X}$ and $\underline{Z}$ have the same underlying dominating measures associated to the product space Ω of spaces related to components of $\underline{X}$ and $\underline{Z}$.

2.4.3 Conditional independence and Bayes rule.

Usually, before consulting a group of experts and beginning to receive their statements $\mathbf{Q}$, the SB has already her prior opinion about the unknown parameter matrix Θ, as expressed by a density function $f(\Theta)$. As a Bayesian, if she has other information $(\underline{Z}, \Phi)$ somehow informative about Θ then she would incorporate this information into her model, by updating her prior $f(\Theta)$ to the form $f(\Theta|\underline{z}, \Phi)$ through her assessment of a conditional joint density function $f(\underline{z}, \Phi|\Theta)$. Moreover, if she consults experts and still wants to be Bayesian, then she must treat their reported statements $\mathbf{Q}$ as data and proceed the updating of $f(\Theta|\underline{z}, \Phi)$ to the posterior form $f(\Theta|\mathbf{Q}, \underline{z}, \Phi)$, now through her likelihood function $f(\mathbf{Q}|\Theta, \underline{z}, \Phi)$ for $\mathbf{Q}$ given Θ, $\underline{Z} = \underline{z}$ and Φ, by using Bayes theorem again. Nevertheless, this sequential approach assumes implicitly that $\mathbf{Q}$ and $(\underline{Z}, \Phi)$ are all independent given Θ. Thus, if there is a structure of conditional independence between $\mathbf{Q}$, $\underline{Z}$ and Φ, then an appropriate approach would be to update her prior $f(\Theta|\Phi)$ to a posterior form $f(\Theta|\mathbf{Q}, \underline{z}, \Phi)$ through her conditional joint density $f(\mathbf{Q}|\underline{z}, \Phi, \Theta)$, taking that conditional independence structure into consideration. When there are no conditional independence restrictions, the posterior density would have the general form

$$f(\Theta|\mathbf{Q}, \underline{z}, \Phi) = \frac{f(\mathbf{Q}|\underline{z}, \Phi, \Theta) f(\Theta|\underline{z}, \Phi)}{f(\mathbf{Q}|\underline{z}, \Phi)} \ . \tag{2.22}$$

Henceforth, every density function $f(\cdot)$ refers to the SB's probability density.

Certainly, the most difficult task related to obtaining the posterior density for Θ conditionally on $\mathbf{Q}, \underline{Z}, \Phi$ in (2.22), is the SB's assessment of $f(\mathbf{Q}|\underline{z}, \Phi, \Theta)$. However, simplifications of that function can be achieved if the SB is prepared to make additional conditional independence statements about those variables. For instance, if a particular class of problems is modelled such that $\underline{Z}$ is independent of Θ given $\mathbf{Q}$ and/or Φ, then the SB's reconciliation rule (2.22) could be simplified to

$$f(\Theta|\mathbf{Q}, \underline{z}, \Phi) = \frac{f(\mathbf{Q}|\Phi, \Theta) f(\Theta|\Phi)}{f(\mathbf{Q}|\Phi)} \ .$$

Note that even a simpler density function of the form $f(\Theta|\mathbf{Q}) = f(\mathbf{Q}|\Theta)f(\Theta)/f(\mathbf{Q})$ is in general difficult to obtain. Perhaps the main difficulty occurs when $\mathbf{Q}$ is a vector (or matrix) of density functions for which the SB must specify a likelihood function $f(\mathbf{Q}|\Theta)$. In this case, $f(\mathbf{Q}|\Theta)$ must represent a joint probability distribution for density functions, that is, an hyper likelihood function.

In Chapter 7, we characterise diverse univariate situations in the modelling approach context where (2.22) can be further simplified and the SB's task of assessing the likelihood function for the experts is facilitated.

2.4.4 Winkler's consensus model for dependent experts.

Winkler's (1981) consensus model with dependence was developed for the univariate setting where the variable of interest $\theta \in \mathbb{R}$ could be either a parameter of a statistical model or a future observation. The model's general formulation considers that the experts are calibrated and provide the Bayesian decision maker their continuous probability densities $\underline{q} = (q_1, \ldots, q_k)$ for θ. The decision maker then assesses her joint likelihood function $\mathcal{L}(\underline{e})$ for the experts' errors $\underline{e} = (e_1, \ldots, e_k)'$ where $e_i = \mu_i - \theta$ and $\mu_i = \mathbb{E}_i[\theta] = \int_{-\infty}^{+\infty} \theta q_i(\theta) d\theta$ for $i = 1, \ldots, k$ is taken as the experts' point estimates for θ. Winkler assumes that $\underline{e}$ is location invariant so that $\mathcal{L}(\underline{e})$, the decision maker's likelihood function for the experts' errors, can be assessed. If $f_0(\theta)$ represents the decision maker prior density for θ then her "consensus distribution" is given by

$$f(\theta|\underline{q},\mathcal{L},f_0) \propto \mathcal{L}(\underline{e})f_0(\theta) \quad . \tag{2.23}$$

Although this formulation seems to be a good solution to the problem of assessing an hyper joint likelihood function for the experts' densities $\underline{q}$, information provided by the experts (their confidence on their assessments) are being lost by the use of just point estimates obtained from those densities. However, Winkler proposes that this also could be solved if $\mathcal{L}(\underline{e})$ is chosen such that

$$\int_{-\infty}^{+\infty} \cdots \int_{-\infty}^{+\infty} \mathcal{L}(\underline{e})de_1 \dots de_{i-1}de_{i+1} \dots de_k = q_i(\mu_i - e_i) \quad .$$

Another alternative proposition that Winkler makes and which would enable the modelling of the dependence among the experts and the decision maker herself is to treat her as the $k+1$th expert with $q_{k+1}(\theta) = f_0(\theta)$, in which case the consensus distribution would be $f(\theta|q_1,\dots,q_{k+1},\mathcal{L}) \propto \mathcal{L}(\mu_1 - \theta,\dots,\mu_{k+1} - \theta)$ and that then would reflect her prior information.

Restricting $\mathcal{L}$ in (2.23) to the family of k-variate normal densities with a mean vector of k zeroes and a positive definite covariance matrix Σ with variances σ_i^2 for $i = 1,\dots,k$ and covariances σ_{ij} for $i,j = 1,\dots,k$, $i \neq j$, Winkler obtains, for when Σ is known and the prior density for θ is an improper diffuse density, a standard normal density $\phi[(\theta-\mu^*)/\sigma^*]$ for the consensus where

$$\mu^* = \frac{\underline{1}'\Sigma^{-1}\underline{\mu}}{\underline{1}'\Sigma^{-1}\underline{1}} \tag{2.24}$$

and

$$\sigma^{*2} = \frac{1}{\underline{1}'\Sigma^{-1}\underline{1}} \quad . \tag{2.25}$$

The vector $\underline{1} = (1,\dots,1)'$.

Note that, μ^* is a linear combination $\sum_{i=1}^{k} w_i\mu_i$ of the experts' means, where

$$w_i = \frac{\sum_{j=1}^{k} \tau_{ij}}{\sum_{l=1}^{k}\sum_{j=1}^{k} \tau_{lj}} \tag{2.26}$$

with $\Sigma^{-1} = (\tau_{ij})$.

For Σ not known, applying the natural-conjugate analysis with the assumption that Σ and θ are independent, we have for an inverted Wishart density for Σ with parameters

(δ_0, Σ_0), a multivariate normal density for $\underline{\mu}|\Sigma, \theta$ as above, and an improper diffuse prior density for θ, that the posterior consensus is a t density with $\delta_0 + k - 1$ degrees of freedom, mean m^* and variance $(\delta_0 + k - 1)s^{*2}/(\delta_0 + k - 3)$, that is

$$f(\theta|\underline{\mu}) \propto \left[\frac{1 + (\theta - m^*)^2}{(\delta_0 + k - 1)s^{*2}}\right]^{-(\delta_0+k)/2} \tag{2.27}$$

where

$$m^* = \frac{\underline{1}'\Sigma_0^{-1}\underline{\mu}}{\underline{1}'\Sigma_0^{-1}\underline{1}} \tag{2.28}$$

and

$$s^{*2} = \frac{\delta_0 + (m^*\underline{1} - \underline{\mu})'\Sigma_0^{-1}\underline{\mu}}{(\delta_0 + k - 1)\underline{1}'\Sigma_0^{-1}\underline{1}} \tag{2.29}$$

If the decision maker's prior density for θ is normal or t then the posterior density is poly-t (Dréze, 1976) and can be evaluated numerically. The subjective assessment of covariance coefficients is studied by Gokhale and Press (1982). The above results overlap with those of Lindley (1983).

Also, Clemen (1987) shows by using the above formulation of Winkler in a degenerate situation (see Chapter 7), where the experts report to the decision maker the summary statistics they obtain from samples they have observed, that the Bernoulli model for non-independent experts corresponds to a mixture of beta distributions. He then studies the effect of different overlapping patterns to conclude that commonly observed data "introduce dependence among the experts and produce a confounding effect that more than offsets the additional information, resulting in the mixed distribution being more spread out ...".

In Chapter 7, we also illustrate several possibilities of overlapping of information between experts by using influence diagrams. In particular, the above mentioned degenerate situation of the expert problem is investigated under considerations of sufficiency.

2.4.5 Lindley's reconciliation of discrete probability distributions.

Lindley (1985a) analyses the case of finite $\Theta = (\theta_1, \ldots, \theta_n)$ by assuming a joint multivariate normal density for the experts' log-probabilities $q_{ij} = \log[pr_i(\theta_j)]$ $(i = 1, \ldots, k; j = 1, \ldots, n)$. This gives, with further assumptions about expectations (and conditional expectations) and by applying Bayes formula, a posterior log-probability for θ_j which is linear in the elements q_{ij}. That is,

$$\log p(\theta_s|\mathrm{Q}, H) = c + \sum_{i,j} \beta_{ijs} q_{ij} + \alpha_s + \gamma_s \quad , \tag{2.31}$$

where $p(\cdot|\cdot)$ denotes the SB's conditional probability of its arguments; $\mathbf{Q} = (q_{ij})$ is a $k \times n$ matrix; H represents the SB's knowledge of the situation before consulting the experts; c is a normalising constant;

$$\beta_{ijs} = \sum_{m,l} \tau_{ijml}\mu_{mls} \ ; \tag{2.32}$$

τ_{ijml} are the elements of the SB's precision matrix for $\mathbf{Q}$, i.e. the inverse of the SB's covariance matrix $\Sigma = (\sigma_{ijml})$ for $\mathbf{Q}$; $\mu_{mls} = \mathbb{E}[q_{ml}|\theta_s, H]$ is the SB's expectation for q_{ml} given θ_s and H;

$$\alpha_s = -\frac{1}{2} \sum_{i,j,m,l} \mu_{ijs}\tau_{ijml}\mu_{mls} \ , \tag{2.33}$$

and $\gamma_s = \log p(\theta_s|H)$ is the SB's prior log-probability for θ_s. The covariances σ_{ijml} from which the precision terms τ_{ijml} were obtained, are not supposed to depend on the true event θ_s, that is, $\sigma_{ijml} = \text{cov}[q_{ij}, q_{ml}|\theta_s, H]$. In all the above, $i, m = 1, \ldots, k$ and $j, l = 1, \ldots, n$.

However, because the q_{ij} are negative, Lindley proposes to use the results for contrasts (i.e. linear forms $\sum_j c_j q_{ij}$ with $\sum_j c_j = 0$) instead. With that the normality assumption can hold if further restriction are imposed. Thus, by defining the contrast $\mathbf{Q}^* = ({q^*}_{ij})$ with ${q^*}_{ij} = q_{ij} - q_{in}$ for $i = 1, \ldots, k$ and $j = 1, \ldots, n-1$ so that $\mathbf{Q} = (\mathbf{Q}^*, \underline{q}_n)$ for $\underline{q}_n = (q_{1n}, \ldots, q_{kn})'$, and assuming that $\mathbb{E}[\underline{q}_n|\mathbf{Q}^*, \theta_l, H]$ does not depend on l, the reconciliation rule (2.31) applies if some "coherence" requirements hold. The above results subsume those of French (1980, 1981) who developed analogous formulation but for log-odds in univariate settings.

Although (2.31) allows, via μ_{ijm} and σ_{ijml}, for the modelling of dependencies among both the experts themselves and the experts and events (experts' calibration), it does not consider a possible conditional independence structure among the future events or parameters θ_j.

Note, as pointed out by Genest and Zidek (1986), that (2.31) corresponds to the LogOp (2.12) if we take anti-logarithms on both sides of (2.31) and identify $p(\theta_s|\mathbf{Q}, H)$ with T, $p(\theta_s|H)$ with g and β_{ijs} with the weights considered proportional to the independent 'information content" of each assessment (Freeling, 1981).

However, despite the fact that (2.31) expresses the SB's opinions about the experts' beliefs (she is not just doing the pool of their opinions), Lindley (1985a) verifies his results against the MP and the EB property to conclude that both criteria are faulty. It can be seen readily that (2.31) does not obey the MP, and it is only EB if $\sum_{i,j} \beta_{ijs} = 1$. Also, he claims with reason that "the fallacy in external Bayesianity is not to recognise that T changes with new information". T is the pooling operator in (2.12).

In Chapter 8 we characterise particular types of SB models in which reconciliation rules can be identified with logarithmic pools for which external Bayesianity applies.

CHAPTER 3

A BASIC REVIEW ON GRAPHICAL MODELS

In this chapter the main concepts in graph theory are described and the definitions of graph related terms that are used in the thesis are then introduced. The emphasis here is on the theory associated with chain graphs (CGs) and influence diagrams (IDs). The notation and terminology employed are based on those described by Lauritzen (1989, 1996), Frydenberg (1990) and Wermuth and Lauritzen (1990). A particularly important result for directed acyclic graphs or IDs, the *d-separation* theorem of Pearl (1986), which allows for statements of conditional independence as well as the issue of preservation of the conditional independence structures in prior-to-posterior analyses of IDs are also presented. However, because the type of CG for which conditional external Bayesianity applies, the partially complete CG (PCG) defined in Chapter 5, can be treated as an ID, the Markov properties for CGs are not reviewed here. The reader should refer to Frydenberg (1990) or Lauritzen (1996) for a complete description of those properties. In fact, the subgraphs induced by the chain elements of a PCG are complete and the possible conditional independence restrictions can only originate from the PCG's underlying ID (Section 5.2).

It is assumed that the set of random variables whose associations of conditional independence are represented by a graphical statistical model, is homogeneous in the sense that all variables are of the same type, i.e. they are all qualitative or all quantitative variables. Mixed graphical association models are studied by Lauritzen and Wermuth (1984), Lauritzen (1989) and Wermuth and Lauritzen (1990).

3.1 Notation and Terminology.

A *graph* $\mathcal{G}$ is characterised by a pair (V, E) of a finite set of *vertices* V and *edges* $E = \{(\alpha, \beta) : \alpha \in V, \beta \in V \text{ and } \alpha \neq \beta\}$, i.e. a subset of the set $V \times V$ of ordered pairs of distinct vertices.

An edge $(\alpha, \beta) \in E$ is called *undirected* if both $(\alpha, \beta) \in E$ and $(\beta, \alpha) \in E$; while an edge $(\alpha, \beta) \in E$ with $(\beta, \alpha) \notin E$ is *directed.* The notation $\alpha \rightarrow \beta$ and $\alpha \sim \beta$ is used to represent directed and undirected associations between vertices respectively.

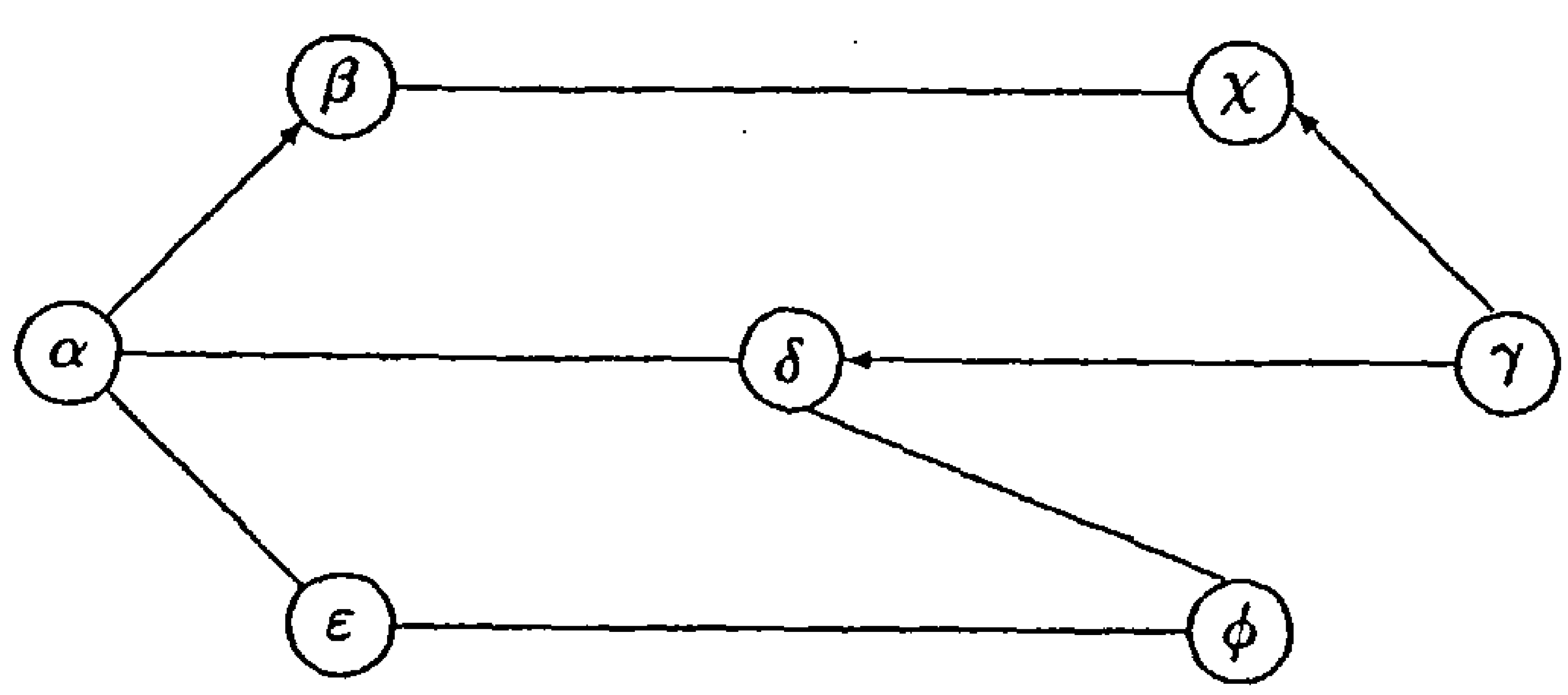
When drawing a graph, we shall use circles to represent vertices and arcs (lines with arrows) or just lines to represent directed or undirected edges respectively. Also, double-circles are used to represent vertices of a particular type (Section 3.2).

A graph $\mathcal{G}$ with only undirected edges is an *undirected graph* $\mathcal{G}^u$, whereas a graph with only directed edges is a *directed graph* $\mathcal{G}^{dir}$.

A *complete graph* is the graph which has all its vertices connected between themselves either by directed or undirected edges. Let $A \in V$ be a subset of vertices of a graph $\mathcal{G}$. The *induced subgraph* $\mathcal{G}_A = (A, E_A)$ is such that $E_A = E \cap A \times A$. A subset of vertices of a graph is a *complete subset* if it induces a complete subgraph and a maximal (with respect to $\subseteq$) complete subset defines a *clique*.

A vertex α is a *parent* of β, if there is an arc from α to β ($\alpha \rightarrow \beta$). In this case, $\alpha \in pa(\beta)$ the set of parents of β. The vertex β is then a *child* of α. Two parents of β are said to be *married* if they are connected by an edge (directed or not). The vertices α and β are *neighbours* if they are connected by an undirected edge ($\alpha \sim \beta$). The *boundary* of a subset $A \in V$ is the set of all vertices V not in A ($V \backslash A$) that are parents or neighbours of vertices in A.

As an illustration, in Figure 3.1 we have a graph where for example, $\alpha \rightarrow \beta$ but $\beta \not\rightarrow \alpha$, that is, $\alpha \in \pi(\beta)$. Also, $\beta \sim \chi$ and thus β and χ are neighbours and the boundary set of β is (α, χ).



$\mathcal{G}'$

FIGURE 3.1. A graph $\mathcal{G}'$ where $\alpha = pa(\beta)$, χ is the neighbour of β, whose boundary set is (α, χ).

It is said that in a graph $\mathcal{G} = (V, E)$ there is a *path* of length n from α to β if there is a sequence $\alpha = \alpha_0, \ldots, \alpha_n = \beta$ of distinct vertices in V such that $(\alpha_{j-1}, \alpha_j) \in E$ for all $j = 1, \ldots, n$, i.e. $\exists\ \alpha_{j-1} \rightarrow \alpha_j\ \forall j = 1, \ldots, n$.

The *ancestors* $an(\beta)$ of a vertex β are the vertices α such that there is a path from α to β and the *descendants* of α are the set of all vertices such that there is a path from α to β.

A set A is an *ancestral set* if the boundary of all $\alpha \in A$ is a subset of $(\subseteq)$ A. Thus, in a directed graph the set A is ancestral if and only if $an(\alpha) \in A$ for all $\alpha \in A$.

A *chain* of length n from α to β is a sequence of distinct vertices $\alpha = \alpha_0, \ldots, \alpha_n = \beta$ such that there is an edge (a line or an arc) between α_{j-1} and α_j for all $j = 1, \ldots, n$, i.e. either $(\alpha_{j-1}, \alpha_j) \in E$ or $(\alpha_j, \alpha_{j-1}) \in E$. In $\mathcal{G}'$ of Figure 3.1, there is a chain of length 2 from α to ϕ.

If all chains from $\alpha \in A$ to $\beta \in B$ intersect $S \in V$, where A, B and S are subsets of V, then S is said to *separate* A from B. Thus in $\mathcal{G}'$ of Figure 3.1, we can say that (δ, ε) separates α from ϕ whereas δ or ε alone does not.

A *cycle* is a path such that $\alpha = \beta$, that is, it begins and ends in the same vertex.

3.2 Chain Graph Models.

A graph when used to represent a set of conditional independence relationships between random variables in a statistical model is called a graphical model. Wermuth and Lauritzen (1990) defined a *graphical chain model* as a statistical model characterised by a graph representing a conditional independence structure. The vertices (circles) of that graph represent random variables (or parameters of statistical models related to random variables) and the edges (arcs or lines) associations between variables. A missing edge in this graphical model is interpreted as a conditional independence. Thus, complete graphs do not imply any conditional independence restrictions. Deterministic variables are represented by double-circles.

Chain graphs are graphical chain models in which the vector $\underline{X}$ of variables can be partitioned into ordered sub-vectors $\underline{X}_1, \ldots, \underline{X}_n$ called *chain elements* (or components) such that

(i) within subsets there are only undirected edges connecting the variables, and

(ii) the subsets themselves are connected by directed edges (arcs pointing) from a set with lower index to one with higher index number.

A chain element $\underline{X}_j$ for which directed edges can be pointed to is called a *response set.* Similarly, an *influence set* is a chain element which has arcs pointing out to a response set.

The direct associations of chain elements according to (ii) above constitute a *dependence chain*. Also, because of the ordering imposed by (ii) above, the CGs do not contain any directed cycles (Frydenberg, 1990).

As an example, in Figure 3.2 below we can see a CG $\mathcal{G}$ with $n = 3$ chain elements $\underline{X}_1 = \{X_{11}, X_{12}, X_{13}\}$, $\underline{X}_2 = \{\underline{X}_{21}, \underline{X}_{22}, \underline{X}_{23}\}$ and $\underline{X}_3 = \{X_3\}$ forming a dependence chain $\underline{X}_1 \to \underline{X}_2 \to \underline{X}_3$. While $\underline{X}_1$ is an influence set and $\underline{X}_3$ is a response set, $\underline{X}_2$ is both a response set to $\underline{X}_1$ and an influence set to $\underline{X}_3$.

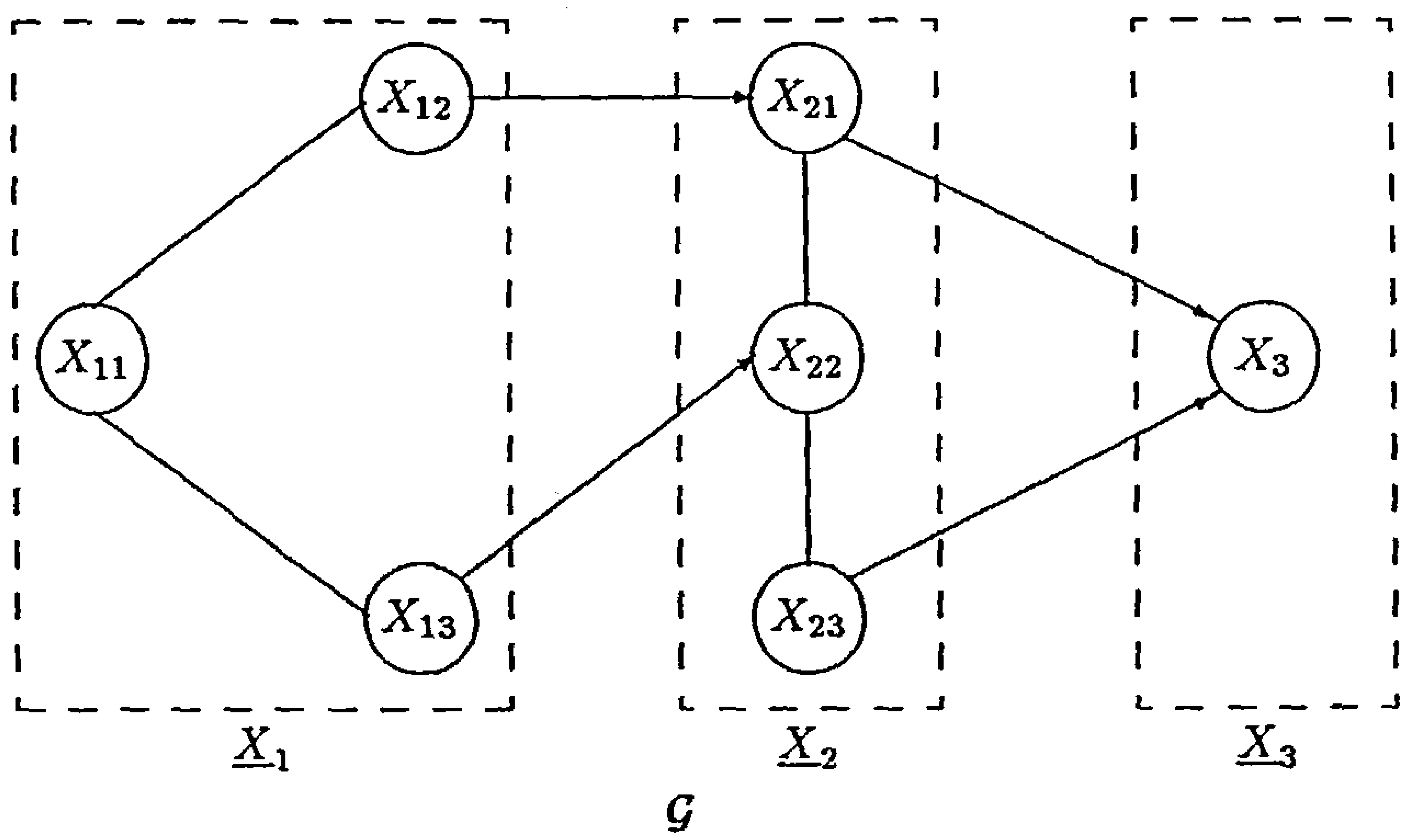


FIGURE 3.2. A chain graph $\mathcal{G}$ with 3 chain elements $\underline{X}_1, \underline{X}_2$ and $\underline{X}_3$ forming a dependence chain.

One of the features of CG models is that they can capture the association structure between variables being described by an individual in a way that can be elicited directly from a verbal description of the problem by the client. They require no numerical inputs early in the modelling process. Several authors (e.g. Smith, 1990) have argued that this property makes them more primitive than probability for specifying perceived relationships (an assertion which is exploited later in this thesis).

An important concept for the characterisation of conditional independence relationships between variables in IDs (as we shall see in Section 3.4) is that of a moral graph. The *moral graph* obtained from a CG $\mathcal{G} = (V, E)$ is the undirected graph $\mathcal{G}^m = (V, E^m)$ with the same vertex set V but with α and β made neighbours in $\mathcal{G}^m$ if and only if

(i) they were neighbours in $\mathcal{G}$, or

(ii) they both were parents of vertices in the same chain element.

Note that both the undirected and the directed acyclic graphs are particular characterisations of CGs. A CG with a single chain element and no dependence chain forms an undirected graph, while a CG with uni-dimensional chain elements constitutes a directed acyclic graph. The moral graph of a directed acyclic graph is built up by joining with a line all vertices that have a common child and replacing all directed edges by undirected arcs (i.e. arcs by lines).

3.3 Conditional Independence.

As mentioned in the last section, missing edges in the underlying graph of a graphical model can be used to represent probabilistic conditional independence (CI) between the random variables in that graph.

A notation which is commonly used to represent probabilistic CI is that of Dawid (1979) where for the random variables X, Y and Z,

$$X \amalg Y|Z \tag{3.1}$$

reads : given Z, X is conditionally independent of Y.

When (X, Y, Z) is defined on a probability space $(\Omega, \mathcal{F})$ with joint distribution P on $(\Omega, \mathcal{F})$, the usual definition of (3.1) is expressed in terms of the factorisation of the conditional joint distribution of (X, Y) given Z : for A, B,

$$P(X \in A, Y \in B|Z) = P(X \in A|Z)P(Y \in B|Z) \quad . \tag{3.2}$$

In his systematic study of CI, Dawid (1979, 1980 and 1996) has shown that the following properties hold for (X, Y, Z, W) defined on $(\Omega, \mathcal{F})$ where $h(\cdot)$ is a function of its arguments only :

(P1) $X \amalg Y|X$,

(P2) $X \amalg Y|Z \iff Y \amalg X|Z$,

(P3) $X \amalg Y|Z$ and $U = h(Y) \implies X \amalg U|Z$,

(P4) $X \amalg Y|Z$ and $U = h(Y) \implies X \amalg Y|(U, Z)$, and

(P5) $X \amalg Y|Z$ and $X \amalg W|(Y, Z) \implies X \amalg (Y, W)|Z$.

It is possible to obtain other properties of CI by regarding the above properties as axioms in a logical system. One of the most appealing reasons for doing so is that complicated manipulations of conditional probability distributions can be avoided in favour of clarity. For example, Dawid (1979) showed among others that the "nearest neighbour" property of a Markov chain can be expressed as :

$$X_3 \amalg X_1|X_2,\ X_4 \amalg (X_1,X_2)|X_3, \text{ and } X_5 \amalg (X_1,X_2,X_3)|X_4 \implies X_3 \amalg (X_1,X_5)|(X_2,X_4).$$

Many other results obtained through the use of (P1)–(P5) alone are given in Dawid (1979, 1996) where no other properties of Probability than those expressed in terms of conditional independence are employed. In effect, the above properties have been adopted as the standard set of CI axioms in most of the works in this area. In the case where X, Y, Z are disjoint subsets of a finite set and $U = h(Y)$ is replaced by $U \subseteq Y$, (P1)–(P5) define an algebraic structure called *semi-graphoid* (Pearl, 1988). If in addition to (P1)–(P5) the following property (P6) is considered,

(P6) $X \amalg Y|Z$ and $X \amalg Z|Y \implies X \amalg (Y,Z)|Y \cap Z$,

then the set (P1)–(P6) is called *graphoid* (Pearl, 1988). However, property (P6) does not hold universally, except under the additional condition that Ω is discrete with $P(\{\omega\}) > 0$ for all $\omega \in \Omega$ (Dawid, 1996).

Although it appeared for a certain time that all properties of probabilistic CI could be derived from the above set (P1)–(P5) alone (Pearl and Paz, 1987), Studený (1992) showed that no finite set of axioms can completely characterise probabilistic CI. However, (P1)–(P5) appear to be adequate for many purposes and can be considered as a reasonable set of axioms applicable in most situations where the concept of CI is required. See also Geiger and Pearl (1993) for a systematic study of the logical implications of CI.

In fact, depending on the interpretation given to the tertiary operator $\cdot \amalg \cdot|\cdot$, it is possible to broaden the range of applications of CI to other domains other than the purely probabilistic one. For example, if Ω is a linear space then $X \amalg_L Y|Z$ can be written to represent that the coefficient of Y in the linear regression of X on (Y, Z) may be taken as 0. Smith (1990) calls this interpretation *linear CI* while Dawid (1996) names it as *zero partial correlation.* Some other terms employed in other applications are : meta CI, for families of distributions on $(\Omega, \mathcal{F})$ and hyper CI for prior distributions in Bayesian analysis (Dawid, 1996). See Lauritzen (1996) for an example of properties of geometric orthogonality being represented by the above formulation.

3.4 The d-separation Theorem.

We have already seen that graphical models can be used to represent probabilistic CI structures. Now, the important issue of how CI relations can be read from given graphs will be presented. For CGs there are Markov properties which allow for this. However, for the particular case of directed acyclic graphs, which is the object of our interest here, there is a criterion called *d-separation* stated as a theorem by Pearl (1986a, 1986b) and fully formalised in Verma and Pearl (1990a, 1990b). See also, for example, Kiiveri et al. (1984), Pearl and Verma (1987), Smith (1989), Lauritzen et al. (1990) and Lauritzen (1996). Geiger and Pearl (1990) showed that the d-separation is the sharpest criterion for CI and therefore cannot be improved.

The version of the d-separation theorem presented below corresponds to a Proposition in Lauritzen et al. (1990) where the proof can be found :

Theorem 3.1 (d-separation). *Let $\underline{X}$, $\underline{Y}$ and $\underline{Z}$ be disjoint sub-vectors of a directed acyclic graph $\mathcal{G}^{dir}$. Then $\underline{Z}$ d-separates $\underline{X}$ from $\underline{Y}$ if and only if $\underline{Z}$ separates $\underline{X}$ from $\underline{Y}$ in $(\mathcal{G}^{dir}_{An(\underline{X}\cup\underline{Y}\cup\underline{Z})})^m$, the moral graph of the smallest ancestral set containing $\underline{X}\cup\underline{Y}\cup\underline{Z}$.*

For an illustration, refer to the directed acyclic graph $\mathcal{G}^{dir}$ of Figure 3.3(a). To investigate whether the statement $(X_4, X_5) \perp\!\!\!\perp X_1|X_2$ is valid, we build up the moral graph of the smallest ancestral set containing (X_1, X_2, X_4, X_5), see Figure 3.3(b), to verify that X_2 alone does not separate (X_4, X_5) from X_1 because of the path X_1, X_2, X_5 in $(\mathcal{G}^{dir}_{An(X_1,X_2,X_4,X_5)})^m$. Therefore the above statement is not valid. However, note that if we include X_3 in the 'separator set' together with X_2, then $(X_4, X_5) \perp\!\!\!\perp X_1|(X_2, X_3)$ is a valid statement. Is $X_2 \perp\!\!\!\perp X_3$ in $\mathcal{G}^{dir}$ valid ? The answer is no, because in $(\mathcal{G}^{dir}_{An(X_2,X_3)})^m$ there is the unblocked path X_2, X_1, X_3. But $X_2 \perp\!\!\!\perp X_3|X_1$ is valid.

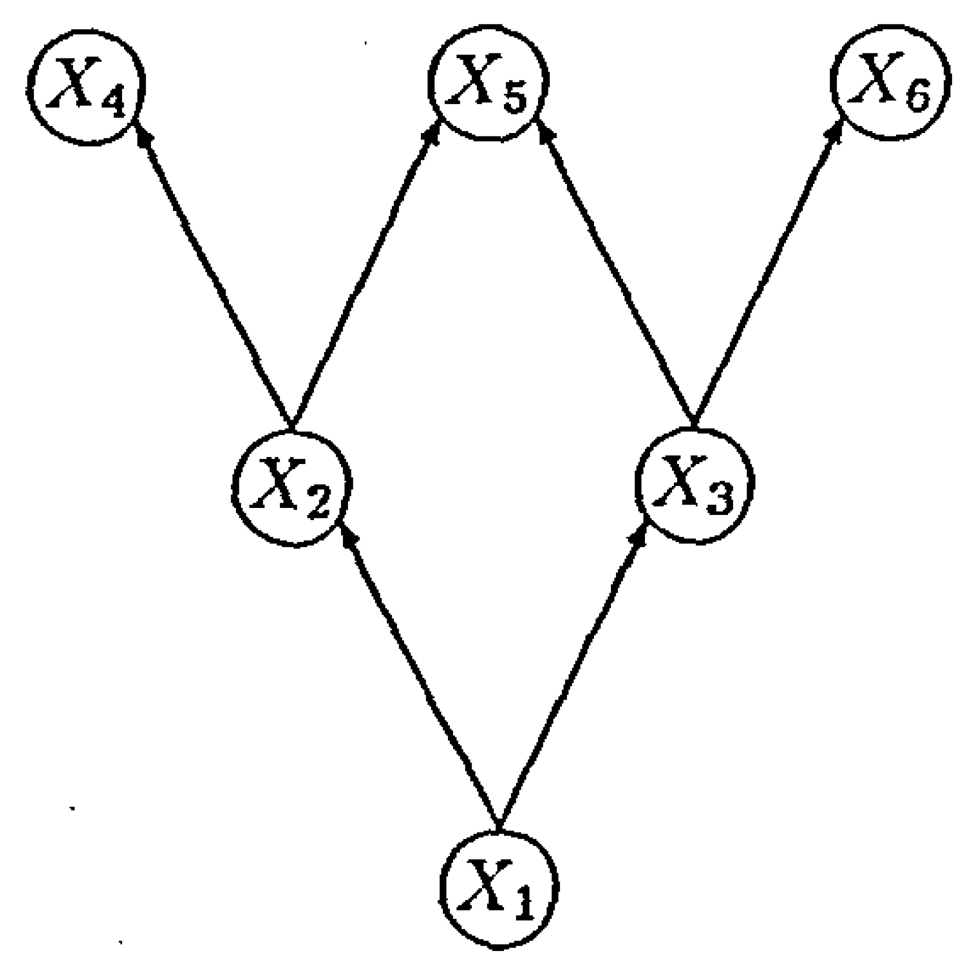
3.5 Influence Diagrams.

As in Smith (1989), an *influence diagram* (ID) $\mathcal{I}$ on a set of random vectors $\mathrm{X}_n = \{\underline{X}_1, \ldots, \underline{X}_n\}$ can be defined as follows:

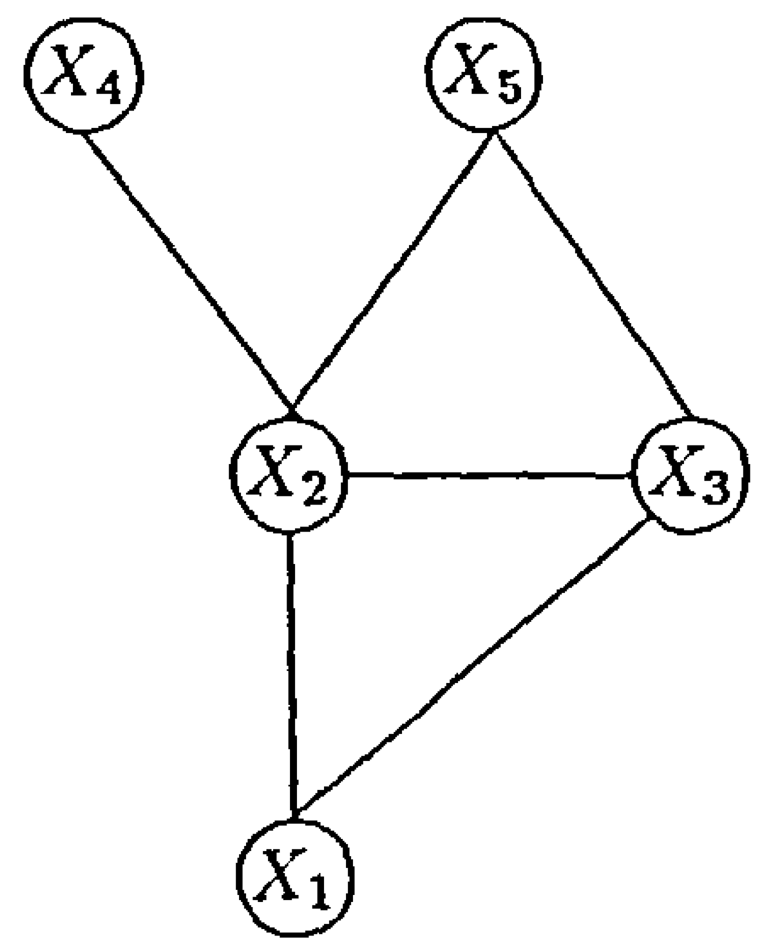
Definition 3.2 (Influence diagram). *Let $\mathcal{G}^{dir}$ be a directed acyclic graph characterised by the pair (X_n, E) and let $\alpha : \mathrm{X}_n \to \{1, \ldots, n\}$ be a numbering of the vertices in $\mathcal{G}^{dir}$ such that $\alpha(\underline{X}_i) < \alpha(\underline{X}_j)$ for a directed edge $(\underline{X}_i, \underline{X}_j) \in E$. An ID $\mathcal{I}$ on X_n is a pair $(\mathcal{G}^{dir}, \alpha)$ together with the following $n-1$ CI statements :*

$$\underline{X}_r \perp\!\!\!\perp \{\underline{X}_j : \alpha(\underline{X}_j) < \alpha(\underline{X}_r)\}|pa(\underline{X}_r); \quad \alpha(\underline{X}_r) = 2, \ldots, n. \tag{3.3}$$

Because of the ordering imposed by the directed edges between the variables in an ID, it can also be called a *causal network* (Jensen, 1996). Also, according to Definition 3.2,



(a) $\mathcal{G}^{dir}$



(b) $(\mathcal{G}^{dir}{}_{An(X_1,X_2,X_4,X_5)})^m$

FIGURE 3.3. (a) A directed acyclic graph, and (b) the moral graph of the smallest ancestral set containing (X_1, X_2, X_4, X_5) to verify the statement $(X_4, X_5) \amalg X_1 | X_2$.

an ID can be viewed as a particular chain graph model where the dependence chain is formed by single elements (i.e. each chain element has just one vertex) in some compatible ordering.

The joint probability density function $f(\mathrm{X}_n)$ of the random variables in an ID $\mathcal{I}$, can be factorized according to (3.3) such that

$$f(\underline{X}_1, \ldots, \underline{X}_n) = \prod_{\alpha=1}^{n} f[\underline{x}_\alpha | pa(\underline{x}_\alpha)] \;, \tag{3.4}$$

where each conditional density $f(\underline{x}_\alpha | pa(\underline{x}_\alpha)$, as well as $f(\underline{x}_\alpha | \emptyset) = f(\underline{x}_\alpha)$, is strictly positive. According to this, it is not difficult to see that for a fully specified ID, there exists a unique joint distribution corresponding to it, whereas the inverse does not hold. According to the ID in Figure 3.4, the joint density function $f(\underline{X})$, where $\underline{X} = (X_1, X_2, X_3, X_4, X_5)$, can be factorized as

$$f(x_1, x_2, x_3, x_4, x_5) = f(x_1) f(x_2) f(x_3 | x_1, x_2) f(x_4 | x_3, x_5) f(x_5) \;.$$

When the Bayesian framework is applied in calculating the conditional densities in the RHS of (3.4), the ID which characterise that breakdown in $f(\mathrm{X}_n)$ can be called a *Bayesian network* or a *belief network*.

One important issue related to the calculations of conditional densities associated to nodes of IDs is how new evidence entering the model can be efficiently propagated through

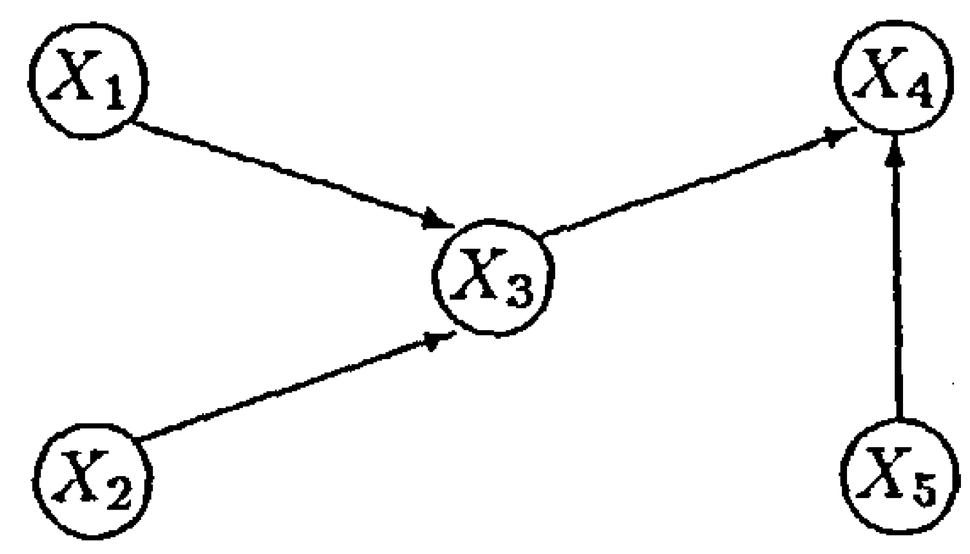


FIGURE 3.4. An ID on $\underline{X} = (X_1, X_2, X_3, X_4, X_5)$ for which the CI statement $(X_1, X_2) \amalg (X_4, X_5) | X_3$ holds and the joint density function factorizes as $f(\underline{x}) = f(x_1)f(x_2)f(x_3|x_1, x_2)f(x_4|x_3, x_5)f(x_5)$.

the network. In particular, we will be interested on the question of preservation of ID structures in terms of the original set of CI statements being kept valid in prior-to-posterior analyses on those structures. For that, a review on the class of decomposable graphs is essential as we shall see in the next section.

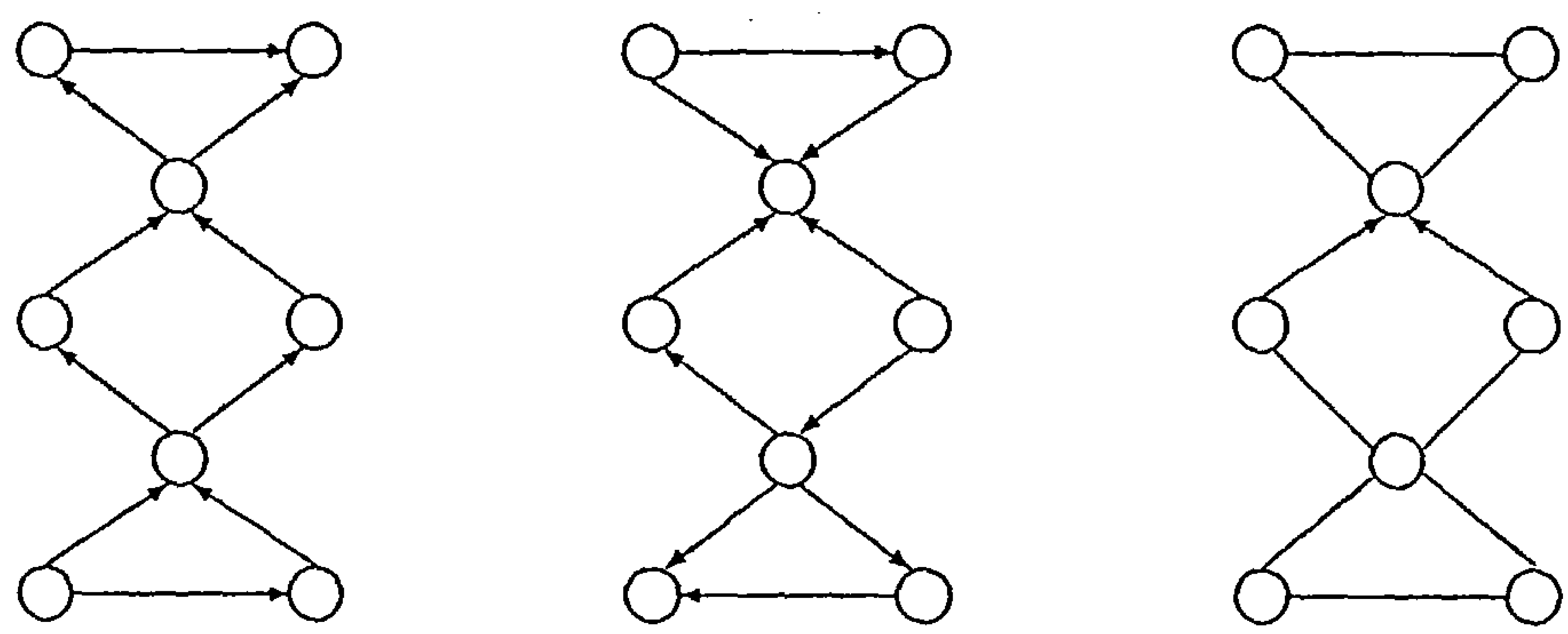
Bayes theorem applied to an ID may correspond to an *arc reversal.* The arc reversal operation in IDs was first addressed by Howard and Matheson (1981) and more formally by Shachter (1986a). The problem in reversing conditioning orderings in IDs is that information concerning CI can be lost. For example, Shachter (1986a) showed that for a directed arc (X_i, X_j) in an ID to be replaced by (X_j, X_i), both nodes must inherit each other's parents. This can clearly introduce loss of CI statements in the original ID. However, a result due to Verma (1987) makes the application of Bayes theorem in graphical models a lot easier as we shall see in the following section.

3.6 Decomposability and Structure Preservation.

The class of decomposable graphical models introduced by Lauritzen et al. (1984), is the one which under certain restrictions on the form of input data, retains the structure (coded in terms of CI statements) in a prior-to-posterior analysis. This fact is used extensively together with the construction of junction trees, to create quick algorithms for calculating posterior distributions in high dimensional problems (Dawid, 1992, Jensen et al. 1994, and Smith and Papamichail, 1996). However in this thesis we shall use this same property for IDs (which are used to represent the informative side —in terms of CI— of more general structures defined in Section 5.2) to define classes of combination rules which are, in a partial sense, externally Bayesian (Chapter 6).

An ID $\mathcal{I}$ is said to be *decomposable* if the parents of any node in the graph of $\mathcal{I}$ are married (i.e. they are all pairwise connected by an edge) in that graph.

Verma (1987) defined a *Verma graph* as one in which all the directed edges are substituted by edges except the ones leaving nodes which are unmarried parents of a common child, and proved that if two graphs have the same Verma graph then they have exactly the same set of CI statements. This result corresponds to the application of Bayes theorem to graphs. See also Pearl and Verma (1987). An illustration can be seen in Figure 3.5 where the graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ have the same Verma graph $\mathcal{G}$ and thus carry the same CI information, despite having some edges pointing in different directions. Note that both graphs $\mathcal{G}_1$ and $\mathcal{G}_2$ have the directed edges in their corresponding Verma graph pointing in the same direction.

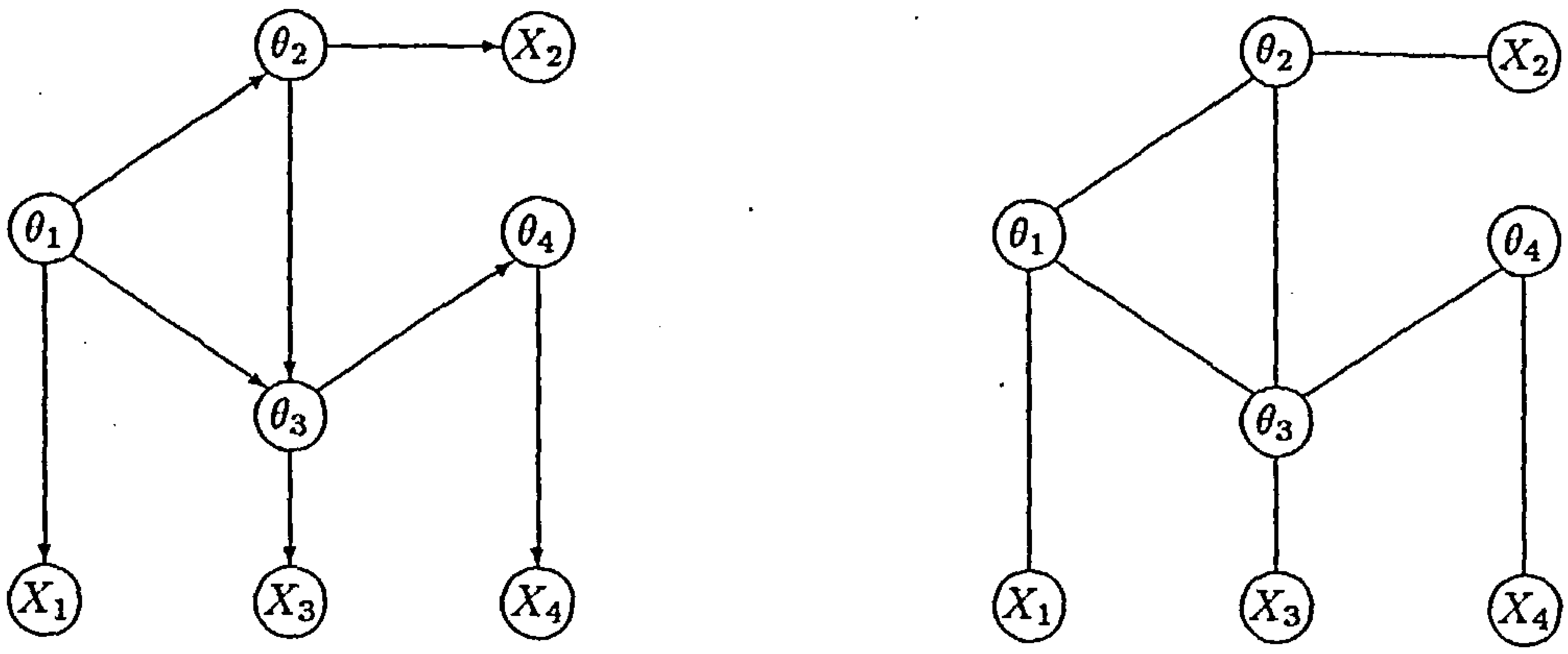


(a) $\mathcal{G}_1$ (b) $\mathcal{G}_2$ (c) $\mathcal{G}$

FIGURE 3.5. Two non-decomposable graphs (a) $\mathcal{G}_1$ and (b) $\mathcal{G}_2$, with a common Verma graph (c) $\mathcal{G}$.

Certainly, if two decomposable graphs have the same Verma graph then they imply the same CI statements even if they have all their arcs pointing to different directions when compared.

Now, Smith (1989) showed a particular case of Verma's result for *similar* IDs, that is, the ones which have the same underlying undirected graph when all their arcs are substituted by lines. He proved that if two similar IDs $\mathcal{I}_1$ and $\mathcal{I}_2$ are decomposable then all CI statements contained in one can be deduced from the CI statements in the other. In practical terms, this result means that a decomposable ID can have its structure preserved in a prior-to-posterior analysis with no information being lost when 'conditioning out' random variables if a certain conditioning order is obeyed.



(a) $\mathcal{I}^d$ (b) $\mathcal{G}^m(\mathcal{I}^d)$

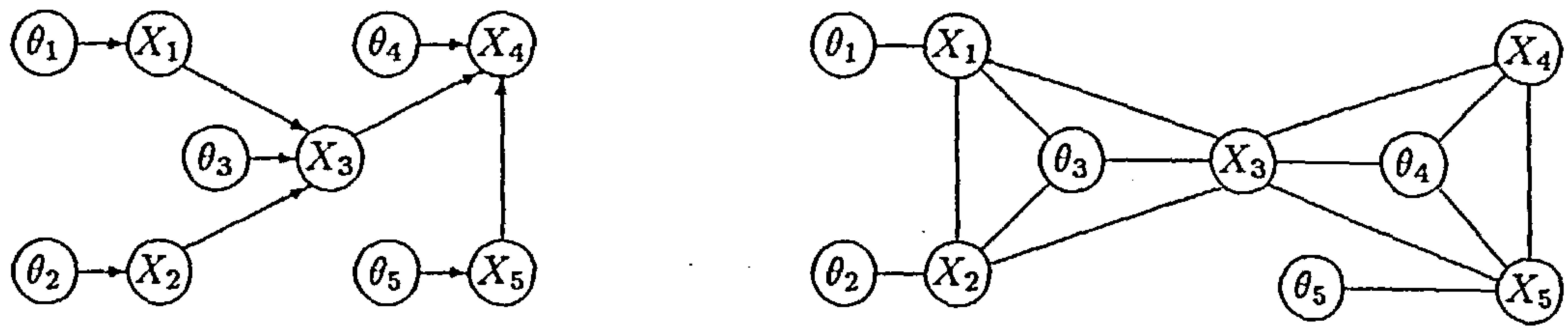
FIGURE 3.6 (a) A decomposable ID $\mathcal{I}^d$ on $\underline{\theta}$ and $\underline{X}$, and (b) its moral graph $\mathcal{G}^m(\mathcal{I}^d)$ showing that CI on $\underline{\theta}$ is preserved under any sampling of $\underline{X}$.

As an example, consider the decomposable ID $\mathcal{I}^d$ of Figure 3.6(a) where the CI between the parameters $\theta_1, \theta_2, \theta_3$ and θ_4 is considered under a form of likelihood called "separable" (Smith, 1990) for $\underline{X} = (X_1, X_2, X_3, X_4)$. The corresponding moral graph $\mathcal{G}^m(\mathcal{I}^d)$ is shown in Figure 3.6(b). According to this, the joint density of $\underline{X}, \underline{\theta}$, where $\underline{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4)$, $f(\underline{x}, \underline{\theta}) = \prod_{i=1}^{4} f[\theta_i|pa(\theta_i)]f(x_i|\theta_i)$, is such that the CI properties on $\underline{\theta}$ are not affected by sampling on $\underline{X}$.

On the other hand, Figure 3.7(a) shows the non-decomposable ID of Figure 3.4 with additional parameter nodes θ_i $(i = 1, \ldots, 5)$ considered as independent random quantities (as in Smith, 1990). For example $\theta_i = P[X_i|pa(X_i)]$. In this case the full ID corresponds to a joint density factorized accordingly

$$p(\underline{x}, \underline{\theta}) = \prod_{i=1}^{5} p[x_i|pa(x_i), \theta_i]p(\theta_i) \ .$$

Figure 3.7(b) shows the moral graph of $\mathcal{I}$. Note that, the independence between the parameters θ_i is preserved only under (i) complete sampling, or (ii) sampling over ancestral graph, of the variables X_i $(i = 1, \ldots, 5)$. For instance, if X_3 is observed first then the only valid CI statement between the parameters would be that $(\theta_1, \theta_2, \theta_3) \amalg (\theta_4, \theta_5)|X_3$



(a) $\mathcal{I}$ (b) $\mathcal{G}^m(\mathcal{I})$

FIGURE 3.7. (a) a non-decomposable ID with parameters $\underline{\theta}$ considered as independent random quantities, and (b) its moral graph showing that CI on $\underline{\theta}$ is preserved under complete or ordered (taken over ancestral set) sampling of $\underline{X}$.

but, for example, $\theta_1 \not\perp\!\!\!\perp \theta_2 \not\perp\!\!\!\perp \theta_3 | X_3$. However, if X_1 is observed first, followed by X_2, and then X_3 and so on, then the independence would be preserved as when all the variables are given at once.

CHAPTER 4

GENERALISING EB LOGARITHMIC POOLS

The reasons we initially chose the external Bayesianity criterion on which to base group pooling rules for members who agree on a conditional independence (CI) structure for a set of variables were twofold. The first reason was a pragmatic one. The types of multiplicative pools which result from external Bayesianity tend to preserve the CI structures related to the associations of variables in a problem, whereas other pools, like the linear opinion pools described in Section 2.2.2 do not (see Section 6.5). Second, by using a loosened form of the external Bayesianity criterion with a CI structure allows a generalisation of the modified LogOp that answers some of the criticisms of its use (see Sections 2.2.5 and 2.4.5). In particular we will show that with our generalisation for a multivariate modified LogOp (the conditional modified LogOp) it is possible to include an algorithm that learns on the weights given to different group members and allows different weights to be attributed to group members depending on their particular expertise in different components of the problem. Since external Bayesianity forms an axiomatic basis for group consensus, our approach can be viewed as essentially normative.

4.1 A Generalisation of External Bayesianity and LogOps.

We shall start our generalisation by showing the need for such an extension using some hypothetical examples for the bivariate case in a situation where experts have very distinct levels of expertise. The first example considers the problem of forming consensus about two independent random variables. The second considers the case when these variables are related in a dependence chain representing a consensus as to the causal relationship linking the two observations. In both these examples we assume that the pooling rule preserves unanimity.

Example 4.1 (independent random variables) : Suppose that a group of 3 experts are to reach consensus about the probability distribution of the occurrence of undetected accidents in a nuclear plant. Any leakage at the nuclear reactor is to be detected by a measurement instrument. Let the random variable X_1 represent the quantity of leakage that might occur in a given scenario. Let X_2 be the error of an instrument measuring that leakage. The expert 1 is a physicist with particular knowledge of the nuclear plant, the expert 2 is a manager of the plant and the expert 3 is the maker of the measuring instruments. All the experts agree that X_1 and X_2 can be assumed independent of each other and each one expresses his own beliefs about the densities of X_1 and X_2, i.e. $f_i(x_1)$ and $f_i(x_2)$ respectively, for $i = 1, 2, 3$. A natural way of deriving the consensus distribution is to note that the joint density $f(x_1, x_2)$ can be expressed in the form $f(x_1, x_2) = f(x_1)f(x_2)$ by the agreed independence assumption. If we require EB to hold individually on both X_1 and X_2 and that the combination rule preserves unanimity, the result (2.15) demands we set

$$f(x_1) = a_1 \prod_{i=1}^{3} [f_i(x_1)]^{w_{i1}}$$
$$f(x_2) = a_2 \prod_{i=1}^{3} [f_i(x_2)]^{w_{i2}}$$

where a_1 and a_2 are proportionality constants for $f(x_1)$ and $f(x_2)$ respectively, which allow these distributions to integrate to one. The weights w_{ij} ($i = 1, 2, 3, j = 1, 2$) would then be arbitrary positive constants such that $\sum_{i=1}^{3} w_{ij} = 1$ for $j = 1, 2$. Note that in the setting of this example, given the different skills of the three experts, it would be unreasonable to demand that $w_{i1} = w_{i2}$ ($i = 1, 2, 3$). Instead we would like to set $w_{31} < w_{32}$, for example.

The combination rule thus sets

$$f(x_1, x_2) = a \prod_{i=1}^{3} [f_i(x_1)]^{w_{i1}} [f_i(x_2)]^{w_{i2}} \tag{4.1}$$

where a is also a normalising constant as a_1 and a_2 above, and the weights w_{i1} and w_{i2} ($i = 1, 2, 3$) are allowed to take arbitrary values on the simplex.

Now under the condition (2.13) imposed by Genest et al. (1986), such a rule cannot be EB on the joint density $f(x_1, x_2)$ unless $w_{i1} = w_{i2}$, $i = 1, 2, 3$. One reason for this is that unless $w_{i1} = w_{i2}$, $i = 1, 2, 3$, we cannot guarantee that two points $\underline{x} = (x_1, x_2)$ and $\underline{x}^* = (x_1^*, x_2^*)$ giving the same value of f will be combined in the same way. However this is a rather dubious and restrictive regularity condition to force the result (2.13) and certainly in this context appears unreasonable. A second reason why the demand of EB forces this is that a likelihood associated with an observation Y whose distribution depends on both X_1 and X_2, may introduce a dependence between X_1 and X_2 conditional on $Y = y$. The combination of densities after the individual accommodation by each member of information y is then not even defined by (4.1). The demand that EB holds for likelihoods of the type described above seems to us too strong in contexts like the one in this example. Note however that the rule (4.1) is clearly EB with respect to sampling distributions which *separate* in X_1 and X_2, i.e. for which $f(y|x_1, x_2) = l_1(x_1|y)l_2(x_2|y)f(y)$ where l_1 and l_2 are functions of their arguments only. So provided we restrict the PPC (prior-to-posterior coherence) to apply *only* to data which preserves, in the mind of each group's member, the independence across the variables they are considering, new pooling rules, reflecting diversity of ability across experts, can be devised. This is one motivation for the rules developed in Chapter 6.

Example 4.2 (dependence chain) : Now suppose that instead of agreeing to the independence of X_1 and X_2, the group makes a weaker agreement that the release X_1 causes the measurement error X_2, i.e. that there is a directed association from X_1 to X_2. Each member has his own particular expertise and individual beliefs about the two variables. With such a causal structure, it is natural for each expert to give his beliefs about X_2 conditional on each value of the variable X_1 causing it. So assume each expert is prepared to provide a density for X_1, and a density for X_2 given a value of X_1. It is natural for the group to construct its combined density $f(x_1, x_2)$ on (X_1, X_2) from a combined density $f(x_1)$ on X_1 and a combined density $f(x_2|x_1)$ on $X_2|X_1$ using the formula

$$f(\underline{x}) = f(x_1)f(x_2|x_1) \tag{4.2}$$

Thus we can write that for PPC and unanimity on likelihoods involving X_1 and not X_2 we obtain a pooling formula

$$f(x_1) = a_1 \prod_{i=1}^{3} f_i{}^{w_{i1}}(x_1) \quad , \tag{4.3}$$

and for PPC on a likelihood where X_2 is a controlled and conditioned variable where unanimity is required we obtain

$$f(x_2|x_1) = a_2(x_1) \prod_{i=1}^{3} f_i{}^{w_{i2}(x_1)}(x_2|x_1) \ . \tag{4.4}$$

The proportionality terms a_1 and a_2 are such as in Example 4.1. Note now however that a_2 is possibly a function of the given value x_1. The weights w_{i1} and w_{i2} $(i = 1, 2, 3)$ must reflect the experts' relative expertise in predicting the variables X_1 and their abilities to predict X_2 once given fixed values x_1 of X_1. Due to the ordered association between the variables, it is natural to allow for the possibility that any i-th's expertise on X_2, coded by the weight w_{i2}, be allowed to be a function of an observed value of X_1 : for example, an expert may forecast X_2 well when x_1 is small but not so well when x_1 is large. Obviously however the restrictions on these weights summing to one, $\sum_{i=1}^{3} w_{ij} = 1$ $(j = 1, 2)$ still apply.

The former relations (4.2) to (4.4) imply that

$$f(\underline{x}) = a(x_1) \prod_{i=1}^{3} f_i^{w_{i1}}(x_1) f_i^{w_{i2}(x_1)}(x_2|x_1) \tag{4.5}$$

where $a(x_1)$ allows $f(\underline{x})$ to integrate to one.

It is interesting now to compare equation (4.5) with (2.15). Clearly (4.5) is more general than (2.15) which can be obtained from (4.5) by setting $w_{i2}(x_1)$ functionally independent of x_1 for $i = 1, 2, 3$. Notice that, unless we have this lack of dependence, equation (4.5) violates Genest's requirement in (2.15) that the LogOp, f, defined as $f = T(f_1, f_2, f_3)$, would only be a function of the range of distributions (f_1, f_2, f_3). We have already stated in Section 2.2.4, and the same arguments of the last paragraph of Example 4.1 can be used here to show how unnatural this regularity condition is when applied to multivariate densities. Here, our chosen transform is a composite of two transforms, one on the margin of X_1, and the other on the conditional density of X_2 given x_1. In this case, a weaker restriction such as

$$T_2(f_1, f_2, f_3)(x_2|x_1) = \frac{P[x_1, f_1(x_2|x_1), f_2(x_2|x_1), f_3(x_2|x_1)]}{\int P[x_1, f_1, f_2, f_3] dx_2}$$

would be more pertinent in that it allows the consensus on $x_2|x_1$ to also depend through $w_{i2}(x_1)$ on occurred values of x_1 which may cause or directly influence outcomes of X_2. In Chapter 6 a n-variate formula for a more general causal structure will be presented.

4.2 Deterministic Weights in LogOps.

We argued above that the weights associated with the combination on $\underline{X}_2$ should be allowed to depend on the observed value of the causal variable X_1, i.e. $w_{i2}(x_1)$ for $i = 1, 2, 3$. Here is another example to reinforce this argument.

Example 4.3 (deterministic weighting) : Let X_1 be the level of blood sugar in diabetes patients in a certain clinic. Let X_2 be the precision of a treatment imposed to reduce that level of blood sugar to its proper value administered to those patients. Assume it is well known that X_2 depends on X_1. The clinic has two doctors, D_1 and D_2. D_1 only ever sees patients for which $X_1 \geq 100$, while D_2 only sees patients for which $X_1 < 100$. Then a sensible rule for a consensus between the two doctors about their joint distribution for $\underline{X} = (X_1, X_2)$ is, according to the consensus probability breakdown for the common ID (Section 3.5), set to be

$$f(x_1, x_2) = f(x_1) f(x_2|x_1) \tag{4.6}$$

where $f(x_1)$ is set under some criterion, but

$$f(x_2|x_1) = \begin{cases} f_1(x_2|x_1), & \text{for } x_1 \geq 100 \\ f_2(x_2|x_1), & \text{for } x_1 < 100 \end{cases} \tag{4.7}$$

or

$$f(x_2|x_1) = a(x_1)[f_1(x_2|x_1)]^{w_{12}(x_1)}[f_2(x_2|x_1)]^{w_{22}(x_1)}$$

where f_1 and f_2 are the doctors D_1 and D_2 respective individual conditioning beliefs about $X_2|x_1$. The $a(x_1)$ is a proportionality term allowing for the $f(x_2|x_1)$ to integrate to one.

The individual levels of expertise of both D_1 and D_2 on X_1 and consequently on X_2 are very distinct and, thus, a natural rule for choosing the weights would be to set

$$w_{12}(x_1) = \begin{cases} 1, & \text{for } x_1 \geq 100 \\ 0, & \text{for } x_1 < 100 \end{cases} \tag{4.8}$$

and $w_{22}(x_1) = 1 - w_{12}(x_1)$.

More generally, if the two doctors had records which gave different expertises depending on X_1 (e.g. they had some overlapping experience about the effect of treatments for patients with the same level of blood sugar), then a more elaborate function of x_1 might be made.

In fact, these weights should be chosen by the doctors on the basis of their commonly held information I about their specialities and expertises. Nevertheless, as we have seen, it is natural given this I to allow the weights regarding the relative expertise on the variable X_2, to depend on x_1. For more complex graph structures, it is natural for the weights to be allowed to depend on the parents of variables as we shall see in Chapters 5 and 6.

4.3 The Bivariate Normal Model.

To obtain some insight into how such conditional weighting methods combine, it is helpful to consider a standard case. So assume that the members of the group assess normal probability density functions as their expert opinions for the involved variables. We shall now see that the group consensus density is not multivariate normal as it was under the LogOp. We shall denote a normal probability density function with mean μ and variance σ^2 by $\mathbf{n}(\mu, \sigma^2)$.

Example 4.4 (bivariate normal): Suppose that in Example 4.3 the doctors D_1 and D_2 assume the process of treating patients with different blood sugar levels as being normally distributed. Thus each one, D_1 and D_2, specifies a marginal density for X_1, $f_i(x_1)$, as $n(\mu_{i1}, {\sigma_{i1}}^2)$, $i = 1, 2$, and a conditional density for $X_2|x_1$, $f_i(x_2|x_1)$, as $n(\mu'_{i2}(x_1), {\sigma'_{i2}}^2)$, $i = 1, 2$. Notice that alternatively the doctors could have assessed joint normal densities for $\underline{X} = (X_1, X_2)$, $f_i(x_1, x_2)$, as $n(\underline{\mu}_i, \Sigma_i)$ where $\underline{\mu}_i = (\mu_{i1}, \mu_{i2})$ is the mean vector and

$$\Sigma_i = \begin{pmatrix} {\sigma_{i1}}^2 & \rho_i \\ \rho_i & {\sigma_{i2}}^2 \end{pmatrix}$$

is the covariance matrix of $\underline{X}$, $i = 1, 2$. The ρ_i's are the assessed correlation coefficients for the variables. In this case, the parameters of the conditional densities $n(\mu'_{i2}(x_1), {\sigma'_{i2}}^2)$, $i = 1, 2$, would be :

$$\begin{aligned} \mu'_{i2}(x_1) &= \mu_{i2} + \frac{\sigma_{i2}}{\sigma_{i1}} \rho_i [x_1 - \mu_{i1}] \\ {\sigma'_{i2}}^2 &= {\sigma_{i2}}^2 (1 - {\rho_i}^2) \end{aligned} \tag{4.9}$$

where the μ'_{i2} are the regression functions in x_1.

We know from (4.6) that the doctors consensus about the joint distribution of the variables is $f(x_1, x_2) = f(x_1) f(x_2|x_1)$ where

$$f(x_1) = a_1 [f_1(x_1)]^{w_{11}} [f_2(x_1)]^{w_{21}} \tag{4.10}$$

and

$$f(x_2|x_1) = a_2(x_1) [f_1(x_2|x_1)]^{w_{12}(x_1)} [f_2(x_2|x_1)]^{w_{22}(x_1)} \tag{4.11}$$

with a_1 and $a_2(x_1)$ being proportionality constants which allow the functions above to integrate to one, and w_{ij} being the weights where w_{12} is possibly a function of x_1, i.e. for $i, j = 1, 2$, $w_{i2} = w_{i2}(x_1)$.

The functions of the right-hand side of the above equations are normal densities and by the results of the Appendix A4.1 we have that $f(x_1)$ is also normal. In fact, by the results of the Appendix A4.1 applied to the marginal density of X_1, we have that

$$[f_i(x_1)]^{w_{i1}} = b_i h_i(x_1)$$

where $b_i = (\sqrt{2\pi}\sigma_{i1})^{1-w_{i1}} / \sqrt{w_{i1}}$ and

$$h_i(x_1) \sim n(\mu_{i1}, {\sigma_{i1}}^2 / w_{i1}).$$

And so, $f(x_1) = c_1 h_1(x_1) h_2(x_1)$ where c_1 is a proportionality constant such that $f(x_1)$ integrates to one. By Theorem A4.2.1,

$$f(x_1) \sim \mathrm{n}(m_1, v_1{}^2)$$

with $c_1 = 1$ and, where, according to equations (A4.2.1) and (A4.2.2) with n set to 2,

$$m_1 = \frac{w_{11}\mu_{11}\sigma_{21}{}^2 + w_{21}\mu_{21}\sigma_{11}{}^2}{w_{11}\sigma_{21}{}^2 + w_{21}\sigma_{11}{}^2}$$

and

$$v_1{}^2 = \frac{1}{r_1} = \frac{\sigma_{11}{}^2\sigma_{21}{}^2}{w_{11}\sigma_{21}{}^2 + w_{21}\sigma_{11}{}^2}.$$

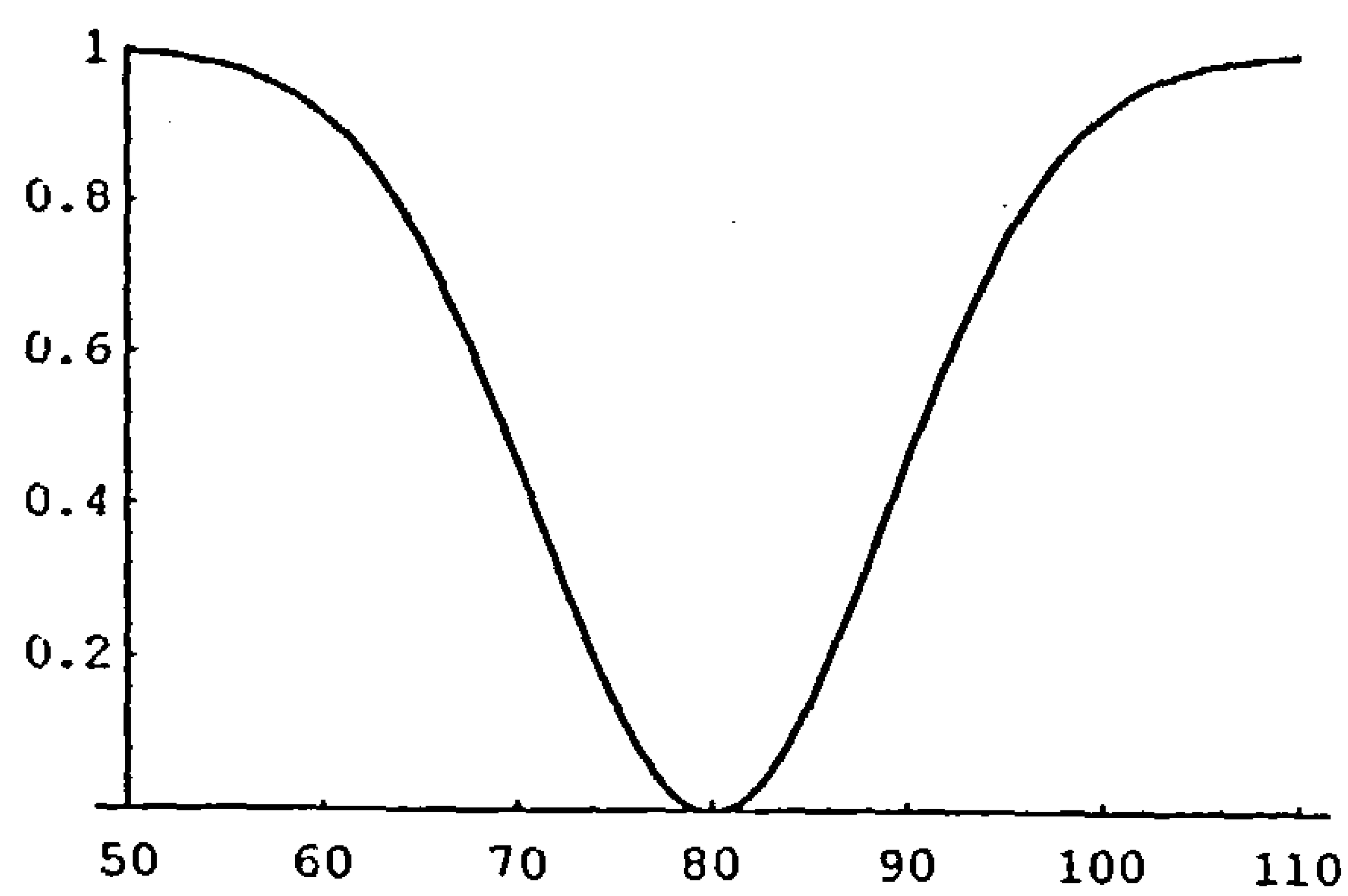


FIGURE 4.1. $w_{12}(x_1)$

Therefore, using the same approach for the conditional of $X_2|x_1$, we have that

$$f(x_2|x_1) \sim \mathrm{n}[m_2(x_1), v_2{}^2(x_1)]$$

where

$$m_2(x_1) = \frac{w_{12}(x_1)\mu'_{12}(x_1)\sigma'_{22}{}^2 + w_{22}(x_1)\mu'_{22}(x_1)\sigma'_{12}{}^2}{w_{12}(x_1)\sigma'_{22}{}^2 + w_{22}(x_1)\sigma'_{12}{}^2}$$

and

$$v_2{}^2(x_1) = \frac{1}{r_2} = \frac{\sigma'_{12}{}^2\sigma'_{22}{}^2}{w_{22}(x_1)\sigma'_{12}{}^2 + w_{12}(x_1)\sigma'_{22}{}^2}$$

with the $\mu'_{i2}(x_1)$'s and $\sigma'_{i2}{}^2$'s, $i = 1, 2$, obtained from (4.9). But notice that the parameters of the conditional mean m_2 as well as the conditional variance $v_2{}^2$ are functions of x_1, the values of X_1, such that $m_2(x_1)$ may be quadratic in x_1 through the weights $w_{i2}(x_1)$'s and the means $\mu'_{i2}(x_1)$'s. Because of this, the joint density of the consensus $f(x_1, x_2) =$

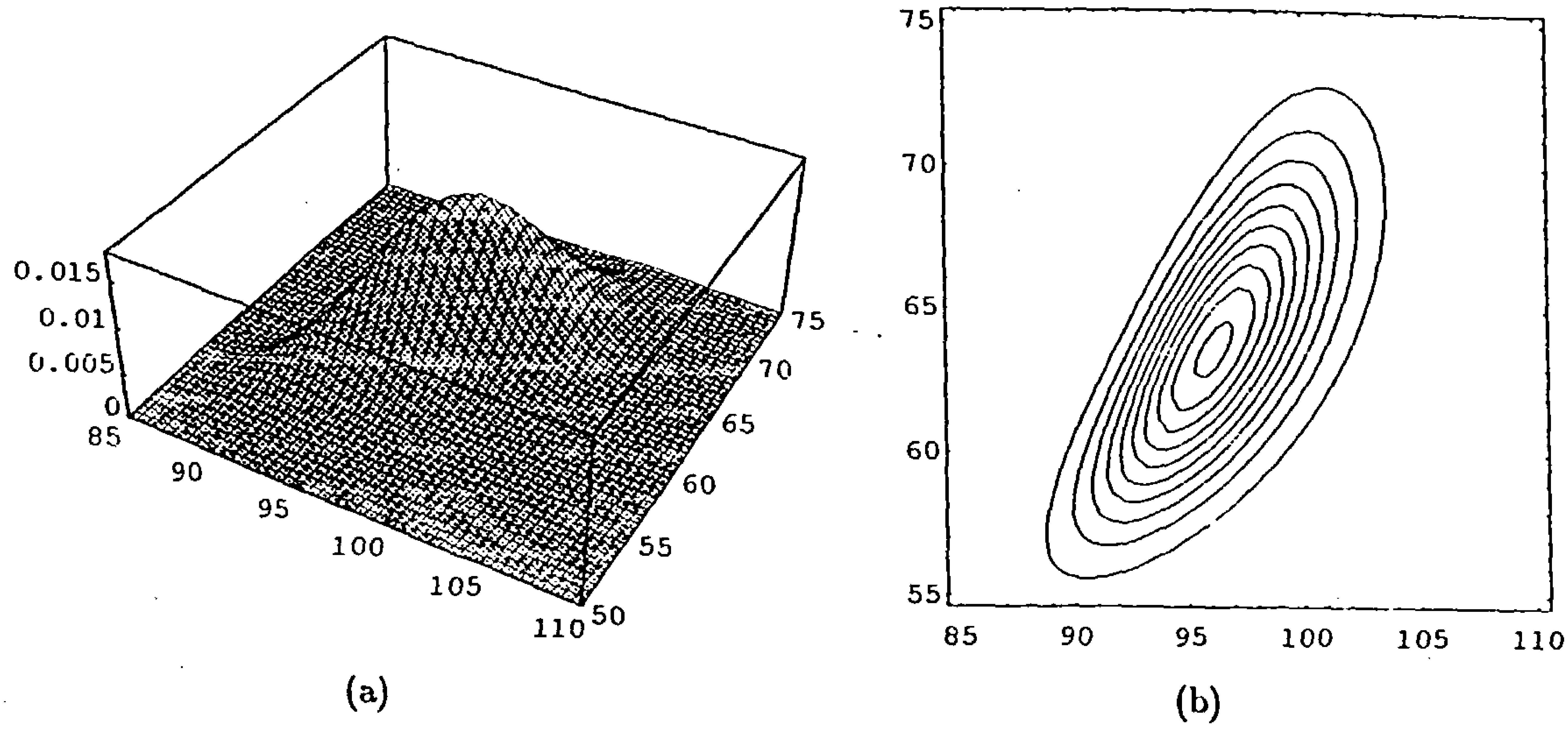


FIGURE 4.2. (a) the joint consensus density and (b) its contour curves.

$f(x_1)f(x_2|x_1)$ is not a multivariate normal despite the fact that $f(x_1)$ and $f(x_2|x_1)$ are normal densities. Sometimes these contours are very non-elliptical in approximate non-regular ellipsoid forms, in other cases the distributions are close to normality.

As an illustration, suppose that at a certain stage D_1 assesses $\mu_{11} = 105$, $\sigma_{11} = 5$, $\mu_{12} = 70$, $\sigma_{12} = 4$ and $\rho_1 = 0.4$; while D_2 assesses $\mu_{21} = 95$, $\sigma_{21} = 3$, $\mu_{22} = 60$, $\sigma_{22} = 2.5$ and $\rho_2 = 0.75$. Also assume that the doctors have distinct specialities such that D_1 only treats extreme cases where $X_1 < 70$ or $X_1 > 90$, while D_2 treats patients whose blood sugar level falls in the range $70 \leq X_1 \leq 90$. Suppose this distinction is expressed by setting $w_{12}(x_1) = 1 - e^{-.5[\frac{(x_1-80)}{9}]^2}$ (see Figure 4.1) and that at this time $w_{11} = .35$. In this case the joint consensus density $f(x_1, x_2) = f(x_1)f(x_2|x_1)$ and its contour curves take the form shown in Figures 4.2(a) and 4.2(b) respectively.

APPENDIX A4.1

The Conditional Normal Density Raised to a Power

In this Appendix we find that a conditional normal density raised to the power w, $w \in \Re^+$, is proportional to a normal density. The conditional densities themselves are of a particularly simple nature since the mean depend only linearly on the variates held fixed and the variance-covariance do not depend at all on the values of the fixed variate. In fact, for a normally distributed random vector $\underline{X}$ which can be partitioned into two component sub-vectors $\underline{X}_1$ and $\underline{X}_2$ of sizes $(q-p)$ and p respectively, it is well known (Anderson, 1971) that

$$f(\underline{x}_2|\underline{x}_1) = \frac{1}{(2\pi)^{\frac{1}{2}p}\sqrt{|\Sigma_{22.1}|}} \exp\left\{-\frac{1}{2}[(\underline{x}_2-\underline{\mu}_2)-\Sigma_{21}\Sigma_{11}^{-1}(\underline{x}_1-\underline{\mu}_1)]' \times \Sigma_{22.1}^{-1}[(\underline{x}_2-\underline{\mu}_2)-\Sigma_{21}\Sigma_{11}^{-1}(\underline{x}_1-\underline{\mu}_1)]\right\}. \quad (A4.1.1)$$

The conditional density $f(\underline{x}_2|\underline{x}_1)$ is a p-variate normal density with mean

$$\underline{\mu}'_2(\underline{x}_1) = \underline{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(\underline{x}_1-\underline{\mu}_1) \quad (A4.1.2)$$

and covariance matrix

$$\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \quad (A4.1.3)$$

where, $\underline{\mu}_1$ and $\underline{\mu}_2$ are the component mean sub-vectors of $\underline{\mu}$ (the $\underline{X}$ mean vector) relative to $\underline{X}_1$ and $\underline{X}_2$ respectively. Similarly, the $\underline{X}$ covariance matrix is

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}.$$

Let $w \in \Re^+$, be a non-negative arbitrary constant, possibly a function of the fixed variates $\underline{x}_1$, $w(\underline{x}_1)$. Thus, (A4.1.1) raised to $w(\underline{x}_1)$ give

$$\begin{aligned} g(\underline{x}_2|\underline{x}_1) &= [f(\underline{x}_2|\underline{x}_1)]^{w(\underline{x}_1)} \\ &= \left[\frac{1}{(2\pi)^{\frac{1}{2}p}\sqrt{|\Sigma_{22.1}|}}\right]^{w(\underline{x}_1)} \\ &\quad \times \exp\left\{-\frac{1}{2}w(\underline{x}_1)[(\underline{x}_2-\underline{\mu}'_2(\underline{x}_1))]'\Sigma_{22.1}^{-1}[(\underline{x}_2-\underline{\mu}'_2(\underline{x}_1))]\right\}. \end{aligned}$$

Defining

$$h(\underline{x}_2|\underline{x}_1) = \frac{g(\underline{x}_2|\underline{x}_1)}{\int g(\underline{x}_2|\underline{x}_1)d\underline{x}_2}$$

we have that $h(\underline{x}_2|\underline{x}_1)$ is a *p-variate normal density* with mean vector $\underline{\mu}'_2(\underline{x}_1)$ and covariance matrix $w^{-1}(\underline{x}_1)\Sigma_{22.1}$. That is,

$$h(\underline{x}_2|\underline{x}_1) = \frac{\sqrt{w(\underline{x}_1)}}{(2\pi)^{\frac{1}{2}p}\sqrt{|\Sigma_{22.1}|}} \exp\{-\frac{1}{2}w(\underline{x}_1)[\underline{x}_2 - \underline{\mu}_2'(\underline{x}_1)]'\Sigma_{22.1}^{-1}[\underline{x}_2 - \underline{\mu}'(\underline{x}_1)]\} . \qquad \text{(A4.1.4)}$$

This is because,

$$\begin{aligned} b(\underline{x}_1) &= \int g(\underline{x}_2|\underline{x}_1) d\underline{x}_2 \\ &= \{1/[(2\pi)^{\frac{1}{2}p}\sqrt{|\Sigma_{22.1}|}]\}^{w(\underline{x}_1)} \int \exp\{-\frac{1}{2}w(\underline{x}_1)[\underline{x}_2 - \underline{\mu}_2'(\underline{x}_1)]'\Sigma_{22.1}^{-1}[\underline{x}_2 - \underline{\mu}_2'(\underline{x}_1)]\} d\underline{x}_2 \\ &= \left\{\frac{1}{[(2\pi)^{\frac{1}{2}p}\sqrt{|\Sigma_{22.1}|}]}\right\}^{w(\underline{x}_1)} \left\{\frac{(2\pi)^{\frac{1}{2}p}\sqrt{|\Sigma_{22.1}|}}{\sqrt{w(\underline{x}_1)}}\right\} \end{aligned}$$

Notice that (i) $b(\underline{x}_1)$ depends on $\underline{x}_1$ only through $w(\underline{x}_1)$ and (ii) $g(\underline{x}_2|\underline{x}_1)$ is not a normal density but $h(\underline{x}_2|\underline{x}_1)$ is.

If $\underline{X}_1$ and $\underline{X}_2$ are independent random vectors then $\Sigma_{12} = \Sigma_{21} = \underline{0}$ and $f(\underline{x}_2|\underline{x}_1) = f(\underline{x}_2)$. In this case, we have that $h(\underline{x}_2) = \frac{1}{b}[f(\underline{x}_2)]^w$ is a normal density with mean $\underline{\mu}_2$ and covariance matrix $w^{-1}\Sigma_{22}$, where $b = \int [f(\underline{x}_2)]^w d\underline{x}_2$ does not depend neither on $\underline{x}_1$ or $\underline{x}_2$.

APPENDIX A4.2

The Product of Normal Densities

Theorem A4.2.1. *Let* $h_i(x)$, $i = 1, ..., n$, *be normal probability density functions with both means* μ_i *and precisions* $\tau_i = 1/\sigma_i{}^2$ *held fixed. Then* $q(x) = \prod_{i=1}^{n} h_i(x)$ *is a normal probability density function with mean* m *and precision* r, $\mathrm{n}(m, 1/r)$, *where*

$$m = \frac{\sum_{i=1}^{n} \tau_i \mu_i}{\sum_{i=1}^{n} \tau_i} \tag{A4.2.1}$$

and

$$r = \sum_{i=1}^{n} \tau_i \quad . \tag{A4.2.2}$$

proof. *Follows by induction over n based on the development by DeGroot (1970, pp.167) for the case when* $n = 2$, *where* $m = (\tau_1\mu_1 + \tau_2\mu_2)/(\tau_1 + \tau_2)$ *and* $r = \tau_1 + \tau_2$. □

CHAPTER 5

COMMON KNOWLEDGE CI STRUCTURES

In this Chapter we define a particular class of graphical models, the *partially complete chain graphs* (PCGs), which we use to represent the conditional independence (CI) structure for the random variables in our problem. Conditions are imposed over the sampling in the problem in order to preserve that structure a posteriori. The notation, terminology, definitions and important results related to *chain graphs* (CGs) and *influence diagrams* (IDs) were introduced in Chapter 3.

Chain graphs are a useful tool for the Bayesian statistical modelling as well as for the decision analysis. They are drawn to follow ordered directional associations (dependence chains of response and influence variables) among the variables in a problem (Section 3.2). This is done by a graph whose vertices represent random variables and whose directed (undirected) edges between vertices represent the directional (undirectional) associations. It has been argued (Smith, 1990 and Pearl, 1993) that the relationships of associations represented in a graph describe beliefs at a coarse enough level to expect that informed individuals may well agree about the structure of associations represented in a CG, at least in many straightforward situations, even if they disagree about distributions of variables within that graph.

In the following Sections 5.1 to 5.4, we define special subclasses of both CGs, the PCGs, and likelihood functions related to sampling on the elements of the PCGs. These definitions allow us to state conditions that a group might obey in order to obtain combination rules more appropriate than the EB ones in multivariate problems like those illustrated in Examples 4.1 and 4.2. In particular we demand that external Bayesianity only holds when the agreed likelihood does not destroy the conditional independence (CI) structure on variables represented by the group's agreed PCG.

The conditions we shall state here are assumed to be *common knowledge* (CK) to all members of the group. According to Aumann (1976), "two people, 1 and 2, are said to have common knowledge of an event E if both know it, 1 knows that 2 knows it , 2 knows that 1 knows it, 1 knows that 2 knows that 1 knows it, and so on". See also Geanokoplos (1992).

5.1 Associations and Chain Graphs.

We saw in Chapter 4 that a more useful set of EB combination rules than the LogOp ones characterised by (2.15) could be obtained for multivariate problems. Nevertheless, two basic conditions on the group's members beliefs about the set of uncertain measurements $\underline{X} = (\underline{X}_1, ..., \underline{X}_n)$ must be met. The first is that all members agree on the association structure on $\underline{X}$ such that :

Condition 5.1.

(i) *The vector $\underline{X}$ can be represented as an ordered list of sub-vectors $\underline{X}_1, ..., \underline{X}_n$, where it is CK to all members of a group G that for $j = 1, ..., n$, the random vector $\underline{X}_j$ receives a directed association from (or more loosely is caused by) $pa(\underline{X}_j)$ where $pa(\underline{X}_j)$ is a sub-vector of $(\underline{X}_1, \underline{X}_2, ..., \underline{X}_{j-1})$, henceforth called the group parent set of the chain element $\underline{X}_j$;*

(ii) *it is CK to the group G that $\underline{X}_j$ is conditionally independent of the elements in $(\underline{X}_1, ..., \underline{X}_{j-1})$ which are not in $pa(\underline{X}_j)$, given its parent set $pa(\underline{X}_j)$, i.e. in the usual notation (Section 3.3)*

$$\underline{X}_j \amalg \{(\underline{X}_1, ..., \underline{X}_{j-1}) \backslash pa(\underline{X}_j)\} | pa(\underline{X}_j) \quad , \; j = 1, ..., n \; .$$

In Example 4.1 we saw the simplest association structure of this kind where it was agreed that $X_1 \amalg X_2$. So $pa(X_2) = \emptyset$ where $\emptyset$ is the empty set. In the second example, $pa(X_2) = X_1$ is that the statement in (ii) were vacuous, and therefore automatically agreeable by the group but substantive because the dependence between X_1 and X_2 was causally ordered in the sense given in (i).

If it is assumed that all members have a common dominating measure then we can assert that the i-th member of the group can write his joint density over $\underline{X}$, $f_i(\underline{x})$, in the form

$$f_i(\underline{x}) = \prod_{j=1}^{n} f_{ij}(\underline{x}_j | pa(\underline{x}_j)) \quad ; \quad i = 1, ..., k \quad ,$$

where $\underline{x}_j$ and $pa(\underline{x}_j)$ are defined above for $j = 2, ..., n$ and $pa(\underline{x}_1) = \emptyset$. Henceforth we shall impose the usual positivity condition that $f_{ij}(\underline{x}_j | pa(\underline{x}_j)) > 0$ for each value of $pa(\underline{x}_j)$ at all values of $\underline{x}_j$ in this space, which is common to all members of the group.

The association structure described by (i) and (ii) above define a class of multivariate structures which are a subclass of CG models (Section 3.2).

Now in our context, it is not implausible to assume that in many structures it will be possible for the members in a group of experts, who possibly interact between themselves, to agree this association structure but differ on their quantification of the probabilities within this structure. This will be our starting point.

5.2 A Useful Subclass of CG Models : the PCGs.

All CGs can be represented in an evocative way using mixed graphs – with both directed and undirected associations. In our context we will consider only CGs which represent associations and conditional independencies a bit more general than those described in the last section, in that undirected associations represent unconditional dependencies within chain elements with the subgraph formed by the components (of a chain element) being *complete*, i.e. all the components are associated between themselves. These will be called PCGs (partially complete chain graphs). Note that they are ones which, by judicious changes of definition, can be represented by IDs on vectors of variables (Smith, 1989, and Queen and Smith, 1993). The PCGs are defined as follows (see Figure 5.1 for an example):

Definition 5.2 (PCG) :. *A chain graph $\mathcal{G}^p(\underline{X})$ with nodes labelled by its chain elements $\underline{X}_1, ..., \underline{X}_n$ is called a partially complete chain graph if, for each $j = 1, ..., n$, all the components of the sub-vector (chain element) $\underline{X}_j$ of $\mathcal{G}^p(\underline{X})$ are connected together to form a complete undirected subgraph. A directed edge connects $\underline{X} \in \underline{X}_i$ to $\underline{Y} \in \underline{X}_j$ if and only if $\underline{X}_i$ is a sub-vector of $pa(\underline{X}_j)$, where $pa(\underline{X}_j)$ is the parent set of $\underline{X}_j$.*

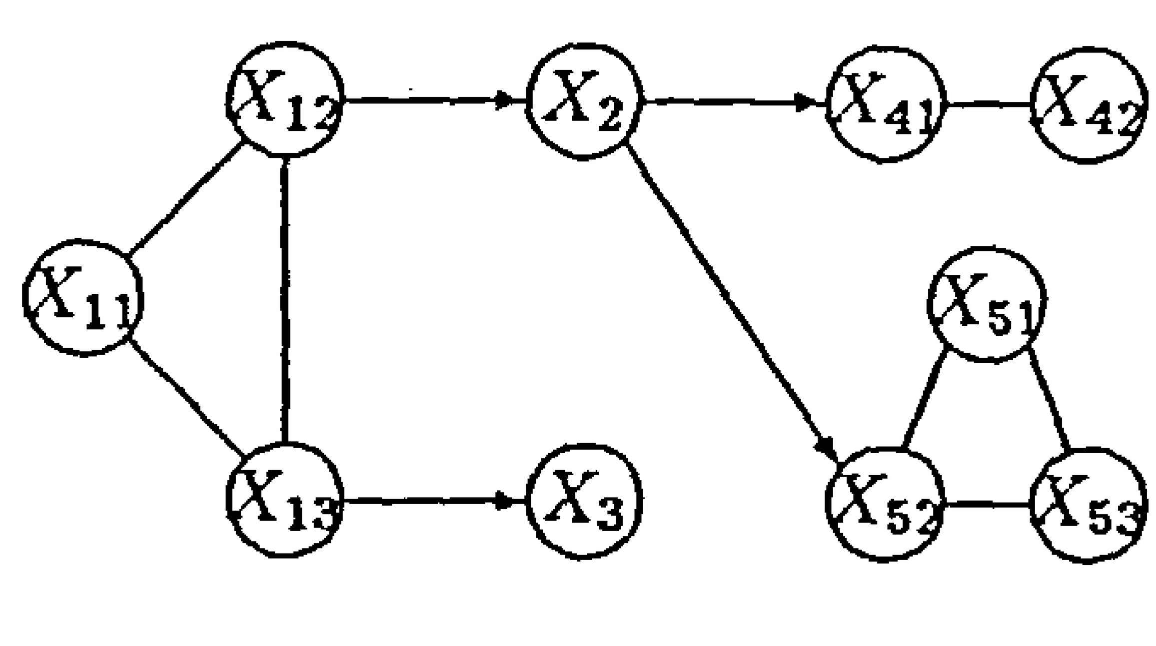
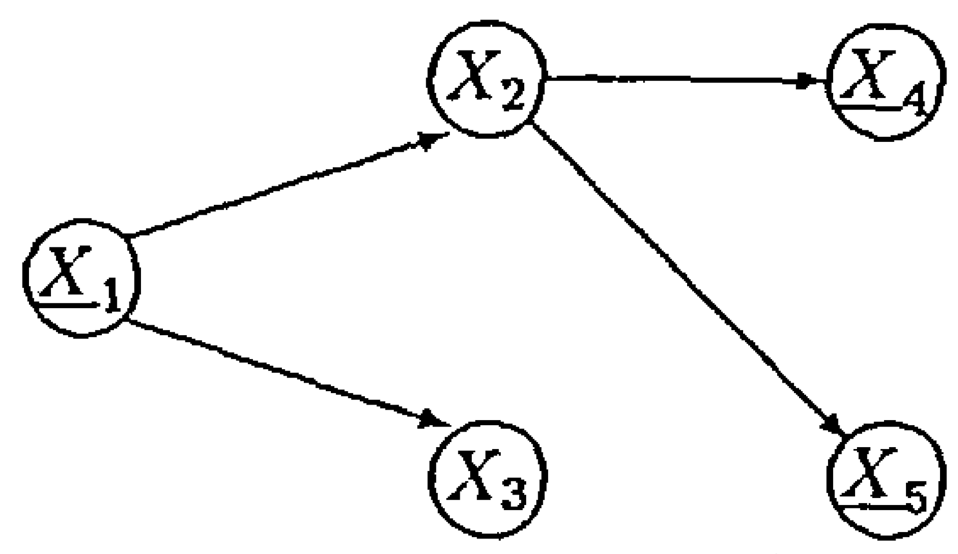


FIGURE 5.1. A partially complete CG $\mathcal{G}^p$.

The ID *induced* by the PCG defined above consists of the statements of Condition 5.1 in Section 5.1, together with a directed graph whose n nodes are labelled $\underline{X}_1, ..., \underline{X}_n$ and where $\underline{X}_i$ is connected by an edge to $\underline{X}_j$ if and only if $\underline{X}_i$ is a sub-vector of $pa(\underline{X}_j)$. Figure 5.2 shows the ID induced by the PCG $\mathcal{G}^p$ of Figure 5.1.



$\mathcal{I}$

FIGURE 5.2. The ID $\mathcal{I}$ induced by $\mathcal{G}^p$ of Figure 5.1.

We shall call a PCG *decomposable*, $\mathcal{G}^{pd}(\underline{X})$, if the ID induced by $\mathcal{G}^p(\underline{X})$ is decomposable, i.e. when for some k $(k = 1, ..., n)$, $\underline{X}_i, \underline{X}_j$ are both sub-vectors of $pa(\underline{X}_k)$, $i \neq j$, then either $\underline{X}_i$ is a sub-vector of $pa(\underline{X}_j)$ or $\underline{X}_j$ is a sub-vector of $pa(\underline{X}_i)$.

Note that because components of the chain elements $\underline{X}_j$ of a PCG $\mathcal{G}^p(\underline{X})$ form a complete undirected graph, any new information about a set of such components is informative about all the other components of that chain element and no CI assumption is destroyed within that chain element.

A practical statistical analysis which uses PCG models is given in Queen and Smith (1994) where dynamic graphical models are derived for multivariate time series and applied to brand sales forecasting of products in supermarkets.

5.3 The Cutting Likelihoods and External Bayesianity.

As mentioned before in Section 5.1, to generalise external Bayesianity, it is also necessary to introduce a second condition, as illustrated in Example 4.1, which will act on the class of likelihoods for models defined on a PCG $\mathcal{G}^p(\underline{X})$. In a general setting, these likelihoods, which we shall call *cutting*, are those which can be informative about one chain element $\underline{X}_j$ and/or its parents $pa(\underline{X}_j)$ only, of the group's common PCG.

Definition 5.3 (Cutting likelihood). *Say that $l(\underline{x}|\underline{z})$ is in the class of cutting likelihoods related to a PCG $\mathcal{G}^p(\underline{X})$, henceforth denoted by $\mathcal{L}(\mathcal{G}^p)$, if it is a likelihood function which could have resulted from a sample $\underline{Z}$ whose density $g(\underline{z}|\underline{x})$ can be written in the following form :*

$$g(\underline{z}|\underline{x}) = g_1(\underline{z}_1|\underline{x}_1)g_2(\underline{z}_2|\underline{x}_2, pa(\underline{x}_2), \underline{z}_1)...g_n(\underline{z}_n|\underline{x}_n, pa(\underline{x}_n), \underline{z}^{n-1}) \tag{5.1}$$

where $\underline{x} = (\underline{x}_1, ..., \underline{x}_n)$ are values of the components of $\underline{X}$ in $\mathcal{G}^p$, $pa(\underline{x}_j)$ are fixed values of the parents of $\underline{X}_j$ in $\mathcal{G}^p(\underline{X})$ and $\underline{z}^l = (\underline{z}_1, \underline{z}_2, ..., \underline{z}_l)$, for $k > 1$, are the observed values of $\underline{Z}^l = (\underline{Z}_1, ..., \underline{Z}_l)$.

What does this class of cutting likelihoods $\mathcal{L}(\mathcal{G}^p)$ look like ? Well, first note that if $\mathcal{G}^p$ is complete then $\mathcal{L}(\mathcal{G}^p)$ is the class of all likelihoods. So we only constrain our class of likelihoods when there is some substantive agreement between the members about some lack of association between certain sets of variables given another.

In the case when $\mathcal{G}^p(\underline{X})$ is not complete there is an ordering of $\underline{X}$ to $(\underline{X}_1, ..., \underline{X}_n)$ such that $\underline{X}_i$ can be thought (loosely) of as being caused by $pa(\underline{X}_i)$, a sub-vector of $(\underline{X}_1, \underline{X}_2, ..., \underline{X}_{i-1})$, $i = 2, ..., n$. When $l(\underline{x}|\underline{z}) \in \mathcal{L}(\mathcal{G}^p(\underline{X}))$ we assume that we have taken an observation $\underline{Z}$ which, possibly after some transformation (see Example 5.1 below), can be represented as $(\underline{Z}_1, ..., \underline{Z}_n)$ where $\underline{Z}_i$ is dependent on $\underline{X}$ only through $\underline{X}_i$ and the values of its direct causes $pa(\underline{X}_i)$. Observe that in this case, the ID induced from $\mathcal{G}^p(\underline{X})$ must be decomposable otherwise original relevances may not be valid anymore after sampling.

Example 5.1 : Gaussian linear models of the form $\underline{Z} = A\underline{X} + \underline{\epsilon}$, where $\underline{Z}$ is a vector of observable random variables, A is a matrix, $\underline{X}$ is a vector of independent random variables and $\underline{\epsilon}$ is the error vector whose components are independent and normally distributed, can be represented by IDs which may not be decomposable if A is not diagonal. Therefore, by the d-separation theorem, edges between components of $\underline{X}$ can be induced in the moral graph when $\underline{Z}$ is observed, thus appearing to destroy the original independence assumptions on $\underline{X}$. Nevertheless, in the Gaussian case, it is always possible to find an orthogonal transformation L which diagonalizes the matrix A (Anderson, 1971). The transformed model, $\underline{Z}^* = L'AL\underline{X} + \underline{\eta}$, where $\underline{\eta} = L\underline{\epsilon}$ and for which $\underline{Z}^*$ is also normally distributed, can now be represented by an ID with components of $\underline{Z}^*$ being separable observables of the respective components of $\underline{X}$ and the independence assumptions on $\underline{X}$ are preserved a posteriori.

The class of cutting likelihoods may look contrived. However such types of likelihood functions arise naturally in a number of situations as we shall see in Section 5.5. They can also be thought as having occurred from designed experiments where $pa(\underline{X}_j)$ have been fixed.

When $\underline{X}_1, ..., \underline{X}_n$ are mutually independent, then $pa(\underline{x}_j) = \emptyset$ (the empty set) for $j = 1, ..., n$. So $\mathcal{L}(\mathcal{G}^p)$ just contains likelihoods arising from an independent observation on a single chain element $\underline{X}_j$ for some j (see Example 4.1). In Example 4.2, the condition on the likelihood is automatically satisfied. So all we require is that external Bayesianity on

X_2 is demanded for all observations y_2 about X_2, which are observed *after* the cause x_1, is known.

Now, the mentioned second condition the group is required to obey regarding the likelihoods obtained from sampling over a decomposable PCG is that :

Condition 5.4. *External Bayesianity is required to hold only with respect to incoming information $\underline{Z}_j$, about the chain elements $\underline{X}_j$ of a decomposable PCG $\mathcal{G}^{pd}(\underline{X})$, for which the value of the ancestral set $An(\underline{z}_j) = (\underline{z}_1, \ldots, \underline{z}_{j-1})$ is already known, and whose likelihood $l_j(\underline{x}_j|\underline{z}^j, pa(\underline{x}_j))$, $j = 1, ..., n$, is a component of a cutting likelihood $l(\underline{x}|\underline{z})$ related to $\mathcal{G}^{pd}(\underline{X})$.*

Note that in particular, the form (5.1) prevents a variable $\underline{X}_k$ in a decomposable PCG, $\mathcal{G}^{pd}$, not belonging to the parent set of $\underline{X}_j$ – which is associated with the index of the product component g_j in (5.1) – $pa(\underline{X}_j)$, to condition the observation $\underline{z}_j$ in g_j, or in other words, to be directly associated with $\underline{z}_j$ in $\mathcal{G}^{pd}(\underline{X}|\underline{Z})$. This in its turn, avoids the introduction of any new association in the moral graph of $\mathcal{G}^{pd}(\underline{X}|\underline{Z})$ not originally present in $\mathcal{G}^{pd}(\underline{X})$.

5.4 - PCG Preservation After Sampling.

The class of cutting likelihoods is a very natural one to consider in the context of PCGs, for CGs $\mathcal{G}^p$ whose induced ID is decomposable, $\mathcal{G}^{pd}$. This is because, provided condition 5.4 is satisfied by the group, the class is determined by those data sets which, for each member, are guaranteed to preserve the CI structure implicit in $\mathcal{G}^{pd}$ after data assimilation. Thus it is simply information which does not destroy the association structure agreed by members of the group (see Example 5.1). Now we present the formal statement and proof of this result.

Theorem 5.5.

(a) *If condition 5.4 is satisfied by the group G for likelihood functions $l(\underline{x}|\underline{z})$ related to the variables of a PCG whose induced ID is decomposable, $\mathcal{G}^{pd}(\underline{X})$, then for each member of G, $\mathcal{G}^{pd}(\underline{X}|\underline{Z})$ is a CG of his joint density of $\underline{X}|\underline{Z}$;*

(b) *if condition 5.4 is not satisfied then, for some member of G, $\mathcal{G}^{pd}(\underline{X}|\underline{Z})$ may not be the CG of $\underline{X}|\underline{Z}$.*

proof.

(a) First we draw the ID $\mathcal{J}(\underline{X},\underline{Z})$ *whose nodes are* $(\underline{Z}_1,...,\underline{Z}_n,\underline{X}_1,...,\underline{X}_n)$. *The ID of* $\underline{X}_1,...,\underline{X}_n$, $\mathcal{J}(\underline{X})$, *is the one induced by* $\mathcal{G}^{pd}(\underline{X})$. *A node* $\underline{Y}$ *is a parent of* $\underline{Z}_j$ $(j = 1,...,n)$ *if either (i)* $\underline{Y} = \underline{X}_j$ *or* $\underline{Y} \in pa(\underline{X}_j)$ *or (ii)* $\underline{Y} = \underline{Z}_l$ $(l = 1,...,j-1)$. *See Fig. 5.3 below for an example.*

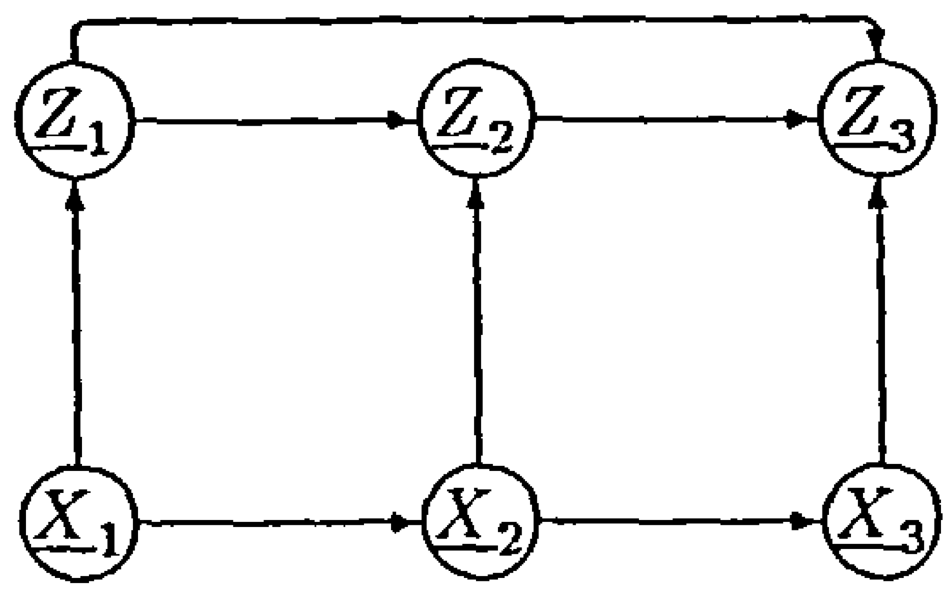


FIGURE 5.3. ID induced by a PCG $\mathcal{G}^{pd}(\underline{X})$ with added nodes $\underline{Z}$.

Now use the d-separation theorem (Section 3.4). New edges inducing the marriage of parents of $\underline{Z}_j$ $(j = 1,...,n)$ *can only occur between a Z node and an X node, since because* $\mathcal{G}^{pd}(\underline{X},\underline{Z})$ *has an induced decomposable ID within* $\mathcal{J}(\underline{X},\underline{Z})$, *the subgraph on sub-vectors of* $(pa(\underline{X}_j),\underline{X}_j)$ *is complete. It follows that all new paths in* $\mathcal{J}(\underline{X},\underline{Z})$ *between nodes on the subgraph* $\mathcal{J}(\underline{X})$ *of* $\mathcal{J}(\underline{X},\underline{Z})$ *induced by marrying of parents are blocked by* $\{\underline{Z}_1,..,\underline{Z}_n\}$. *Also because the value of the ancestral set* $An(\underline{Z}_j)$ *is assumed to be already known when* $\underline{Z}_j$ *is observed (condition 5.4), there will be no unblocked new path linking X nodes which were not linked in the original association structure. This in turn implies that if a CI statement is implied in* $\mathcal{J}(\underline{X})$ *on* $\underline{X}$ *it is also implied on* $\mathcal{J}(\underline{X},\underline{Z})$.

(b) If $l(\underline{x}|\underline{z}) \notin \mathcal{L}(\mathcal{G}^{pd})$, *then for some index j* $(j = 1,...,n)$ *there exists a* $\underline{Z}_j$ *such that the parents of* $\underline{Z}_j$ *will have an edge induced to* $\underline{X}_j$ *after moralization. Marrying parents on conditioning on* $\underline{Z}_j$ *will now produce an unblocked path between another node* $\underline{X}_k$ $(k = 1,...,j)$ *not connected to* $\underline{X}_j$ *in* $\mathcal{J}(\underline{X})$, *and* $\underline{X}_j$. *Thus whereas a priori all members agreed that*

$$\underline{X}_j \amalg \underline{X}_l | pa(\underline{X}_j)$$

now, after observing $\underline{Z}_j$ *we can no longer deduce this.* □

5.5 The Class of Decomposable PCGs.

In this section we shall illustrate some cases in which, whenever the group needs to combine its beliefs about the chain element $\underline{X}_j$, $j = 1,...,n$, in a PCG $\mathcal{G}^p(\underline{X})$, it will have already observed the value of the parents $pa(\underline{X}_j)$ of $\underline{X}_j$. When this is so it is very natural for the group to combine beliefs *conditional* on $pa(\underline{x}_j)$. By demanding external Bayesianity on these conditional densities, we shall see in the next chapter, gives the conditionally externally Bayesian combination rules. So in this context the direct use of those rules is both straightforward and natural.

In fact although the restrictions we impose to our CI structures impose limitations in their applicability there are many modelling situations when combination rules may act under these conditions (Smith et al., 1993) making them a fairly rich class of structures. Here are some cases and examples:

Case 5.1 : Assume $\underline{X}_1, ..., \underline{X}_n, ...$ is a multivariate ARIMA(n,p,0) time series with known coefficients. Here $pa(\underline{X}_1) = \emptyset$ and for $j \geq 2$,

$$pa(\underline{X}_j) = pa(\underline{X}_t, ..., \underline{X}_{j-1})$$

where $t = max\{1, j-1-n-p\}$. One step ahead forecasting by the group now fulfils this requirement. These models are in fact very special examples of the class given below.

Case 5.2 (The Bayes prequential models) : Prequential specification of a model (Dawid, 1992) can be used in a wide class of problems. In such a specification it is assumed that each probabilistic forecaster provides a sequence of probabilistic forecasts, in the form of densities, viz. :

$$f_1(\underline{x}_1), f_2(\underline{x}_2|\underline{x}_1), ..., f_r(\underline{x}_r|\underline{x}_1, ..., \underline{x}_{r-1}), ... \quad ; \quad r > 1 ,$$

i.e. each forecaster gives a one step ahead predictive density for the next observation in a series given the whole past history. Any Bayesian time series model, for example the dynamic linear model (DLM) of West and Harrison (1989), is designed to be able to provide these outputs. But it is worth noting that such outputs are often available *implicitly* from classical models as well. For example current maximum likelihood estimates can substitute parameters in the sample density of successive observations. This is called the "plug-in rule" by Dawid (1992).

For such models we can use our established notation to write

$$f_r(\underline{x}_r|\underline{x}_1, ..., \underline{x}_{r-1}) = f_r(\underline{x}_r|pa(\underline{x}_r)) \quad ; \quad r > 1 ,$$

where $pa(\underline{x}_r) = \underline{t}_r(\underline{X}^{r-1})$ is a vector of statistics based on $\underline{X}_1, ..., \underline{X}_{r-1}$ which is *predictively sufficient* for $\underline{X}_r$, i.e. those arguments (not $\underline{x}_r$) which appear explicitly in the functional form of the density f_r defined above.

Now in many homogeneous time series we can write

$$\underline{t}_r(\underline{X}^{r-1}) = \underline{t}_{r-1}(\underline{X}^{r-2}) + \underline{T}_{r-1}(\underline{X}_{r-1}) \quad ; \quad r > 1 \ ,$$

where $\underline{t}_r, \underline{T}_r$ are k-vectors which depend on their arguments only, $\underline{T}_r$ is a *smoothed* error vector term. Initially, $\underline{t}_0$ is a vector of constants and $\underline{T}_0 = \underline{0}$ is a vector of zeros. When this property holds, the relationships can be usefully embodied in the CG of Figure 5.4. We shall call such a time series *state regular.*

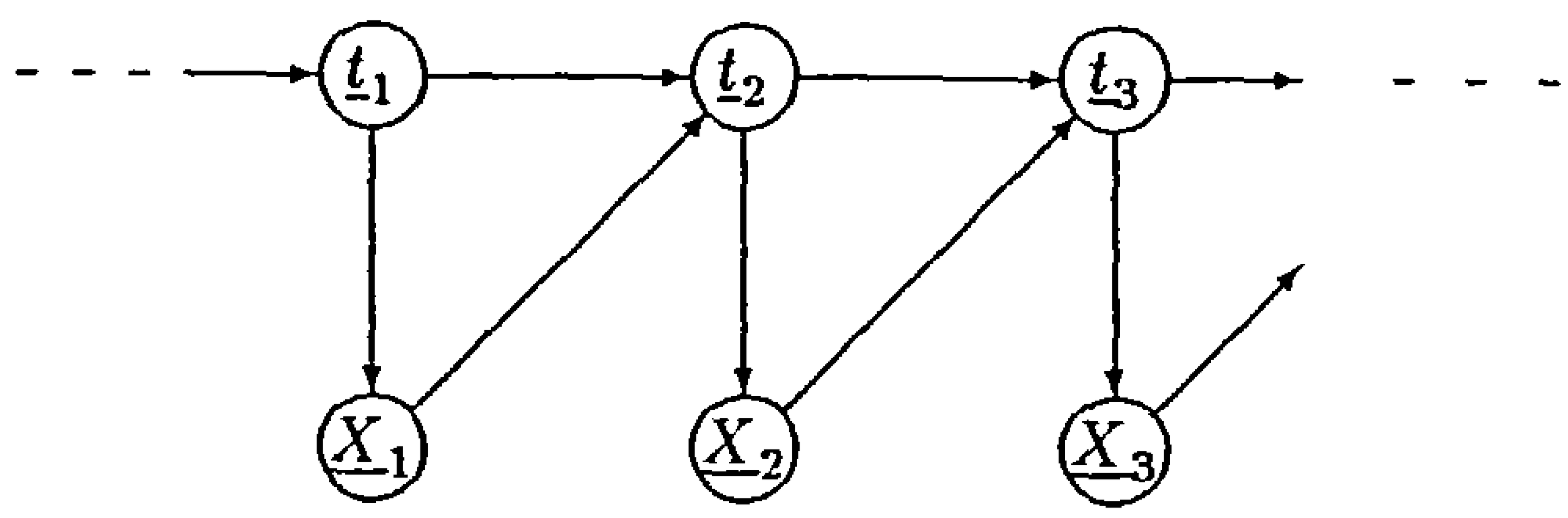


FIGURE 5.4. The CG of a state regular time series.

Here are some examples of state regular time series.

Example 5.2 : The ARIMA(n,p,0) model of case 5.1 is state regular with

$$\underline{t}_r = (\underline{X}_t, ..., \underline{X}_{r-1})$$

where for $r \geq 2$, $t = max\{1, r-1-n-p\}$.

Example 5.3 (The constant DLM) : In the notation of West and Harrison (1989), when variances are known

$$\underline{t}_r = \underline{a}_r$$

where $\underline{a}_r$ is the one step ahead predictive mean vector of the state vector $\underline{\theta}_r$ at time $r-1$. For the DLM with unknown observational variance updates,

$$\underline{t}_r = (\underline{a}_r, \underline{R}_r)$$

where $\underline{R}_r = \underline{S}_{r-1}\underline{R}_r^*$, $\underline{R}_r^* = \underline{G}_r\underline{C}_{r-1}^*\underline{G}_r' + \underline{W}_r^*$, with $\underline{G}_r$ being the state transition matrix, $\underline{C}_{r-1}^*$ and $\underline{W}_r^*$ being the covariance matrixes of the prior density of $\underline{a}_r$ and of the state error density respectively; $\underline{S}_{r-1} = d_{r-1}/n_{r-1}$ is a prior location estimate of the observational variance, with d_{r-1} and n_{r-1} being respectively the mean and variance of the prior density for the observational precision matrix. The same argument shows that Dynamic Generalised Linear Models (West and Harrison, 1989) are also state regular.

Example 5.4 (The multiprocess DLM class II) : In its most common form (Harrison and Stevens, 1976) these act on families of growth models. Here set

$$\underline{t}_r(i) = \underline{a}_r(i)$$

where $\underline{a}_r$ is the $\underline{\theta}_r$ predictive mean vector as given above for the i-th model M_i where models (M_1, M_2, M_3, M_4) represent respectively, no change, outliers, change in level and change in growth, last time step, and $\underline{t}_r(i)$ is a 2 vector of state means. Then

$$\underline{t}_r = (t_r(1), t_r(2), t_r(3), t_r(4), \underline{S}_r)$$

where $\underline{S}_r$ is the 4 vector of probabilities assigned to M_1, M_2, M_3, M_4 on the basis of the first $r-1$ observations.

In all the models described above if used for one-step-ahead predictions of $\underline{X}_r$ all members of a group can be assumed to know $\underline{t}_r$ at time r. Hence the group prediction can be reasonably assumed to used this fact.

Example 5.5 (Dynamic junction trees) : In Gargoum and Smith (1994), the *Adjust Operator*, $Adj(C(i), S(i))$, applied on cliques $C(i), i = 1, ..., n_t$, of a dynamic decomposable Gaussian junction tree, updates the mean vectors $\underline{\mu}_t(i)$ and the covariance matrixes $\Sigma_t(i)$ of the marginal distribution on $C(i)$, as new observations $\underline{y}(i)$ –informative about $C(i)$ only– arrives. Thus,

$$\underline{t}_r(i) = (\underline{\mu}_1^*(i), \Sigma_{11}^*(i))$$

where $\underline{\mu}_1^*(i)$ and Σ_{11}^* are the parameters (mean vector and covariance matrix respectively) of the distribution of the set of variables in the separator set $S(i) \subseteq C(i)$, updated by the usual Bayesian approach.

CHAPTER 6

CONDITIONAL EXTERNAL BAYESIANITY

6.1 The Definition of Conditional External Bayesianity.

Suppose that the k members of a group have agreed in the structure of a decomposable PCG $\mathcal{G}^{pd}(\underline{X})$ relating n random vectors, $\underline{X} = (\underline{X}_1, ..., \underline{X}_n)$ in a certain problem. Let A_j be the event that the parent nodes of $\underline{X}_j$, $pa(\underline{X}_j)$ have fixed values $\underline{X}_j$, i.e. $\mathrm{A}_j = \{pa(\underline{X}_j) = pa(\underline{x}_j)\}$ for $j = 1, ..., n$. Despite believing the common PCG, $\mathcal{G}^{pd}$, each member has its own particular opinion about the parameters of his conditional densities, $f_{ij}(\underline{X}_j|\mathrm{A}_j)$, associated with the graph structure. For technical reasons we shall assume that $f_{ij} > 0$ for $i = 1, ..., k$ and $j = 1, ..., n$ (see Section 2.2.1). The members agree to follow the external Bayesianity axiom, satisfying thus formula (2.10) but for sets of conditional densities. Hence, according to the probability breakdown in CGs, external Bayesianity on $\underline{X}_j|\mathrm{A}_j$ is required for all possible values of the $\underline{X}_j$ $(j = 1, ..., n)$. Unlike in our Definition 2.3 of external Bayesianity in Section 2.2.3, the likelihood over which we demand external Bayesianity to hold for these conditional densities, is restricted to the family $\mathcal{L}(\mathcal{G}^{pd})$ of cutting likelihoods defined in the Section 5.3. Thus we demand condition 5.4 to be satisfied. Particularly, this means that external Bayesianity is required *only* for new information that might come from a designed experiment whose design points are the parents or causes of that variable.

We can now define the conditional external Bayesianity property which will characterise conditionally externally Bayesian (CEB) pooling operators representing combined probability density functions associated with a PCG as follows :

Definition 6.1 (Conditional External Bayesianity). *Say a group G obeys the conditional external Bayesianity property if the joint density $f(\underline{x})$ of the variables in its common decomposable PCG $\mathcal{G}^{pd}(\underline{X})$ is combined in the following way. For each component $\underline{X}_j$ of $\underline{X}$ and each set A_j of possible values of the parents $pa(\underline{X}_j)$ of $\underline{X}_j$, each of the conditional densities $f_j(\underline{x}_j|\mathrm{A}_j)$ is pooled to preserve the external Bayesianity property $(j = 1, ..., n)$ with respect to the component $l_j(\underline{x}_j|\mathrm{A}_j, \underline{z}^j)$ of the common cutting likelihood $l(\underline{x}|\underline{z})$ associated with $\mathcal{G}^{pd}(\underline{X})$, where $\underline{z}^j = (z_1, \ldots, z_j)$.*

6.2 A Characterisation of CEB Pooling Operators.

In order to obtain a characterisation of conditional external Bayesianity through a class of pooling operators, we propose, in line with Madansky (1964, 1978) and Genest et al. (1986), that a CEB pooling operator, T_j, associated with a PCG $\mathcal{G}^p(\underline{X})$, is one which is EB but only for data that respect an ordering of conditioning, $\underline{X}_j|\mathrm{A}_j$ $(j = 1, ..., n)$, implicit in that chain when Condition 5.4 is satisfied.

Define a measure space $[\Omega_j(\mathrm{A}_j), \mu_j^*(\mathrm{A}_j)]$, thereafter denoted (Ω_j, μ_j^*), with Ω_j being the product space of spaces related to components of $\underline{X}_j$ and μ_j^* being the product reference measure associated with the $r(j)$ dimensional vector $\underline{X}_j$ in Ω_j (Rudin, 1986). Let $T_j : \Delta_j^k \to \Delta_j$ be a CEB pooling operator, where Δ_j is the class of all μ_j^*- measurable functions $f_{ij} : \Omega_j \to (0, \infty)$ with $f_{ij} > 0$ (μ_j^* a.e.) such that $\int \cdots \int f_{ij} d\mu_{1j} \ldots d\mu_{r(j)j} = 1$ for all i, j and $r(j)$. The μ_{qj}'s $(q = 1, \ldots, r(j)$ for $j = 1, \ldots, n)$ are measures associated with components of $\underline{X}_j$. If such a pooling operator satisfies the following condition :

$$T_j[f_{1j}, ..., f_{kj}](\underline{x}_j|\mathrm{A}_j) = \frac{P_j[\underline{x}_j|\mathrm{A}_j, f_{1j}(\underline{x}_j|\mathrm{A}_j), ..., f_{kj}(\underline{x}_j|\mathrm{A}_j)]}{\int \cdots \int P_j[\cdot, f_{1j}, ..., f_{kj}] d\mu_{1j} ... d\mu_{r(j),j}} \ ; \mu_j^* \text{ a.e.}, \tag{6.1}$$

for each $j = 1, ..., n$, where $P_j : \Omega_j \times (0, \infty)^k \to (0, \infty)$ is some arbitrary Lebesgue measurable function, then $\underline{T} = (T_1, ..., T_n)$ is said to satisfy the PCG $\mathcal{G}^{pd}$ *likelihood principle.* This condition means that, except for a normalisation factor which does not depend on $\underline{X}_j$, the density of the consensus at $\underline{X}_j|\mathrm{A}_j$ is required to depend only on $\underline{X}_j$ and its fixed parents $pa(\underline{X}_j)$ as well as on the individual densities at the actual value of the unseen quantities given its parents, but not upon the densities of the values which might have obtained but did not.

Similar to Madansky's condition (2.10) for EB pooling operators, the CEB pooling operator T_j is required to satisfy the condition :

$$T_j\left(\frac{l_j f_{1j}}{\int \cdots \int l_j f_{1j} d\mu_{1j} ... d\mu_{r(j),j}}, ..., \frac{l_j f_{kj}}{\int \cdots \int l_j f_{kj} d\mu_{1j} ... d\mu_{r(j),j}}\right) = \frac{l_j T_j(f_{1j}, ..., f_{kj})}{\int \cdots \int l_j T_j(f_{1j}, ..., f_{kj}) d\mu_{1j} ... d\mu_{r(j),j}} \tag{6.2}$$

with μ_j^* a.e. for $j = 1, ..., n$ and for likelihoods l_j component of $l \in \mathcal{L}(\mathcal{G}^{pd})$.

Assuming the underlying measure space (Ω_j, μ_j^*) of each vector $\underline{X}_j$ in $\mathcal{G}^{pd}(\underline{X})$ can be partitioned in at least four non-negligible sets (that includes the continuous and most of the countable cases), it is straightforward to extend the characterisation theorem of Genest et al. (1986) for such an operator in the following way :

Theorem 6.2 (Conditional modified LogOp). *Let (Ω_j, μ_j^*) be a quarternary measure space. Let $\bar{f}_j(\underline{x}_j|\mathrm{A}_j) : \Delta_j^k \to \Delta_j$ be a CEB pooling operator representing the k subjects combined conditional density for the $r(j)$ dimensional random vector $\underline{X}_j$, $j = 1, ..., n$, given its parents in a PCG $\mathcal{G}^{pd}(\underline{X})$. If*

$$\bar{f}_j(\underline{x}_j|\mathrm{A}_j) = T_j[f_{1j}, ..., f_{kj}](\underline{x}_j|\mathrm{A}_j)$$

$j = 1, ..., n$, where for all f_{ij} $(i = 1, ..., k)$ in Δ_j^k and for an existing $\mu_j^* \times$ Lebesgue measurable function $P_j : \Omega_j \times (0, \infty)^k \to (0, \infty)$, $T_j : \Delta_j^k \to \Delta_j$ satisfies (6.1), then $\bar{f}_j$ takes the form

$$\bar{f}_j(\underline{x}_j|\mathrm{A}_j) = \frac{p_j \prod_{i=1}^{k} [f_{ij}(\underline{x}_j|\mathrm{A}_j)]^{w_{ij}(\mathrm{A}_j)}}{\int \cdots \int p_j \prod_{i=1}^{k} [f_{ij}(\underline{x}_j|\mathrm{A}_j)]^{w_{ij}(\mathrm{A}_j)} d\mu_{1j} ... d\mu_{r(j),j}} \quad ; \mu_j^* \text{ a.e. }, \tag{6.3}$$

where $p_j : \Omega_j \to (0, \infty)$ are essentially bounded functions and $w_{ij}(\mathrm{A}_j)$ are weights such that $\sum_{i=1}^{k} w_{ij}(\mathrm{A}_j) = 1$ holds for each index $j = 1, ..., n$, and A_j are the variables whose values are commonly known by the group when the combination rule is applied. Furthermore, the weights are nonnegative unless Ω_j is finite or there does not exist a countably infinite partition of (Ω_j, μ_j^*) into non-negligible sets.

proof. *The proof is straightforward by a slight adaptation of Theorem 4.4 in Genest et al. (1986) individually for each node in $\mathcal{G}^{pd}$ associated with the underlying measure space (Ω_j, μ_j^*), but conditioned on $\underline{X}_j|\mathrm{A}_j$ $(j = 1, ..., n)$. For completeness a full proof is given in Appendix A6.* □

The $w_{ij}(\mathrm{A}_j)$ are the weights that should be a measure of the member i expertise associated with the vector $\underline{X}_j$. They can be possibly a function of other components in $\underline{X}_j$. We will see in Section 6.3 that in time series problems we can often set A_j so that $\mathrm{A}_j \subset \underline{X}_1, ...\underline{X}_{j-1}$ for $j = 2, ..., n$.

As mentioned in Section 2.2.4, it is rather difficult to give an interpretation to p_j in the context of our group decision problem. However, it is reasonable in the majority of problems to require that T_j preserves the group's unanimity. This leads us to setting p_j equal one. Note that in this case, condition (6.1) can be restricted to

$$T_j[f_{1j}, ..., f_{kj}](\underline{x}_j|\mathrm{A}_j) = \frac{P_j[\mathrm{A}_j, f_{1j}(\underline{x}_j|\mathrm{A}_j), ..., f_{kj}(\underline{x}_j|\mathrm{A}_j)]}{\int \cdots \int P_j[\cdot, f_{1j}, ..., f_{kj}] d\mu_{1j} ... d\mu_{r(j),j}} \quad ; \mu_j^* \text{ a.e. }, \tag{6.4}$$

thus not allowing T_j to depend on $\underline{x}_j$ directly.

The following corollary can be stated, with the proof being easily obtained from Theorem 6.2 :

Corollary 6.3 (Conditional LogOp). *Let (Ω_j, μ_j^*) be a quaternary measure space and let $T_j : \Delta_j^k \to \Delta_j$ be a CEB pooling operator which preserves unanimity. If there exists a $\mu_j^* \times$ Lebesgue measurable function $P_j : \Omega_j \times (0, \infty)^k \to (0, \infty)$ such that T_j satisfies (6.1) for all vectors of conditional opinions $(f_{1j}, ..., f_{kj}) \in \Delta_j^k$, then T_j is a logarithmic opinion pool, i.e.*

$$T_j(f_{1j}, ..., f_{kj})(\underline{x}_j|\mathrm{A}_j) = \frac{\prod_{i=1}^k [f_{ij}(\underline{x}_j|\mathrm{A}_j)]^{w_{ij}(\mathrm{A}_j)}}{\int \cdots \int \prod_{i=1}^k [f_{ij}(\underline{x}_j|\mathrm{A}_j)]^{w_{ij}(\mathrm{A}_j)} d\mu_{1j} ... d\mu_{r(j),j}} \quad ; \mu_j^* \text{ a.e. }, \quad (6.5)$$

for some arbitrary weights w_{ij}, $i = 1, ..., k$, $j = 1, ..., n$, possibly functions of A_j (the commonly known past when the densities of $\underline{X}_j$ are combined) adding up to one. Moreover, the weights $w_{ij}(\mathrm{A}_j)$ are nonnegative unless Ω_j is finite or there does not exist a countable partition of (Ω_j, μ_j^*) into non-negligible sets.

6.3 Group Learning in Time Consistent Chains.

Now coming back to the time series application domain of Section 5.5, in all those types of setting, CEB is requiring that if information about $\underline{X}_r$ arrives between time point $r-1$ and r (whose sample distribution is agreed by all members), then it should not matter whether the group combines *one-step-ahead predictive densities* of observations after each member has assimilated this information or to combine densities first and then let the group assimilate it.

A valuable point here is that from the characterisation given in the last section, the group can agree beforehand about the algorithm they require to make the weights given to each expert depend upon the available past, A_r. In this context A_r is just $\underline{X}_1, ..., \underline{X}_{r-1}$. As we have already mentioned, such flexibility answers one of the criticisms of the usual LogOp pool. On the other hand, this extra flexibility demands that we find other criteria that might help the group to choose an appropriate algorithm.

In the past various classical and Bayesian methods for updating weights in pooling rules have been suggested and tested on data (see e.g. Bunn, 1985, and Faria and Souza, 1995). Although most of these have just been used in conjunction with point forecasting methods in linear combining models (Section 2.3), it is possible that many of them could be relatively simply adjusted to define methods of combining full probabilistic forecasts.

Perhaps the simplest such method is the *outperformance* we described in Section 2.3.2, which can be generalised for the n-variate setup described above by just introducing an index j in formula (2.17), that is, for $t \geq 1$:

$$w_{ij,t} = (1 - \rho_{j,t-1})w_{ij,t-1} + \rho_{j,t-1}(t-1)^{-1}r_{ij,t-1} ,$$

where again a component $w_{ij,t}$ $(i = 1, ..., k)$ of the weights vector $\underline{w}_{j,t}$ $(j = 1, ..., n)$ is interpreted as the probability that the member i will produce the most appropriate forecasting model of $X_{j,t}$; $r_{ij,t-1}$ is the number of successes obtained by the member's i forecasting model for the variable X_j up to time $t-1$; $\rho_{ij,u} = u/[\bar{\alpha}_{j,u} + u]$, with $\bar{\alpha}_{j,u} = \sum_{i=1}^{k} \alpha_{ij,u}$ and $\alpha_{ij,u}$ are the parameters of the Dirichlet prior distribution associated to the variable X_j.

Such a method of updating weights transfers directly onto the conditional LogOp discussed above as does other more sophisticated methods such the quasi-Bayes (Section 2.3.3).

Although developed for linear combinations, there would be in principle no reason why such methods should not be adapted to be employed in CEB pools such as the LogOps which are log-linear functions. It seems that at least in the case of Gaussian models, when logarithmic pools give linear combinations on the means (see e.g. Example 4.4 in Section 4.3), this is case. Nevertheless, this subject requires further investigation. Among the methods described in Section 2.3, the outperformance is the least dependent on the linearity of the combination rule.

6.4 How CEB Poolings Appear EB.

We have already shown in Chapter 5 that in order to ensure that the CEB rules are well-defined, the common PCG $\mathcal{G}^p(\underline{X})$ must be decomposable. Note that the CEB rules are *not* based on pooling operators since their arguments are not necessarily just the values of the joint densities in those pools.

The question now is, when $l(\underline{x}|\underline{z}) \in \mathcal{L}(\mathcal{G}^{pd})$, in what sense, if any, are the CEB combinations on chain elements, EB on the whole PCG $\mathcal{G}^{pd}$?

Certainly, when data $\underline{Z}$ about $\underline{X}$ in $\mathcal{G}^{pd}(\underline{X})$ is observed, that evidence must be propagated through the PCG. Therefore, all the conditional pools $T_j(f_{1j}, \ldots, f_{mj})[\underline{x}_j|\underline{x}^{(j-1)}]$ on the chain elements $\underline{X}_j$ and its predecessors on $\mathcal{G}^{pd}(\underline{X})$, that is $\underline{X}_1, \ldots, \underline{X}_{j-1}$, must be updated to (omitting the members' densities) $T_j[\underline{x}_j|\underline{x}^{(j-1)}, \underline{z}]$ for $j = 1, \ldots, n$.

Suppose we demand that the group agrees to update those conditional densities in a backwards sequence. Thus assume that the group agrees to update $\underline{X}_n|\underline{X}^{(n-1)}$ first,

$\underline{X}_{n-1}|\underline{X}^{(n-2)}$ second, and so on to $\underline{X}_1$.

Let $\tilde{f}$ be the group's combined joint density which takes the individual posterior densities on $\underline{X}_n|[\underline{X}^{(n-1)}, \underline{Z}]$, pools them, uses the derived (agreed) density of $\underline{Z}|\underline{X}^{(n-1)}$ to obtain individual densities of $\underline{X}_{n-1}|[\underline{X}^{(n-2)}, \underline{Z}]$, pool these, and so on down to the density of $\underline{X}_1$. Note that to do that the members must adopt the densities (associated to the marginal likelihoods) for $\underline{Z}|\underline{X}_j, pa(\underline{X}_j)$ $(j = n, n-1, ..., 1)$ as common updating factors (see the proof of Theorem 6.4 below).

Also, let $\bar{f}$ be the group's combined joint density which pools the individual prior densities of $\underline{X}_n|\underline{X}^{(n-1)}$, forms a group's posterior density of $\underline{X}_n|[\underline{X}^{(n-1)}, \underline{Z}]$ and a density of $\underline{Z}|\underline{X}^{(n-1)}$, uses this agreed density of $\underline{Z}|\underline{X}^{(n-1)}$ and the pool of the prior densities on $\underline{X}_{n-1}|\underline{X}^{(n-2)}$ to obtain the posterior density $\underline{X}_{n-1}|[\underline{X}^{(n-2)}, \underline{Z}]$ and so on down to the pooling of $\underline{X}_1$. Thus, the question is when does $\tilde{f}(\underline{x}|\underline{z}) = \bar{f}(\underline{x}|\underline{z})$? The answer is provided by the following theorem.

Theorem 6.4. *Suppose that Property 5.4 is satisfied by the group G for a PCG $\mathcal{G}^{pd}(\underline{X})$. Also assume that the vector of weights $\underline{w}_j$ of the conditional LogOps used to combine the beliefs of the members of G on chain elements $\underline{X}_j$ of $\mathcal{G}^{pd}(\underline{X})$, is a function only of variables in $pa(\underline{x}_j)$ for all $j = 1, ..., n$. Then for the whole graph $\mathcal{G}^{pd}(\underline{X}|\underline{Z})$,*

$$\tilde{f}(\underline{x}|\underline{z}) = \bar{f}(\underline{x}|\underline{z}) \ ,$$

where the conditional LogOps components of $\tilde{f}$ or $\bar{f}$ are backwards sequentially updated.

Note that the above result guarantees that the original PCG structure is preserved after new information is incorporated to the model. However, it is important to point out that neither $\tilde{f}$ or $\bar{f}$ are strictly EB in general. When there is agreement on how the graph is updated then they are in particular *sequentially* PPC. The updating is only strictly EB when the corresponding graph is completely disconnected.

proof. *Here we use the convention that a function will explicitly depend only on the values of its arguments. Also, for simplicity, the argument of w_{il}, that is $pa(\underline{x}_l)$, will be omitted in this proof. First note that there is a proportionality constant for each $l = 1, ..., n$, $h_l(pa(\underline{x}_l), \underline{z}^l) = 1/\int ... \int \prod_{i=1}^k f_{il}{}^{w_{il}}(\underline{x}_l|pa(\underline{x}_l), \underline{z}^l)dx_{1l}...dx_{r(l)l}$, $r(l)$ being the dimension of the vector element $\underline{X}_l$, such that*

$$\tilde{f}_l(\underline{x}_l|pa(\underline{x}_l), \underline{z}^l) = h_l(pa(\underline{x}_l), \underline{z}) \prod_{i=1}^k f_{il}^{w_{il}}(\underline{x}_l|pa(\underline{x}_l), \underline{z}^l) \ .$$

Also note that if information $\underline{Z}^l$ about $\underline{X}^l$ is observed then the density of $(\underline{X}_{l+1}, ... \underline{X}_n)$ given $\underline{X}^l$ remains unchanged.

We begin the proof by showing that, under the conditions of the theorem, the density of $\underline{X}_n | (pa(\underline{X}_n), \underline{Z}^n)$ does not depend on when $\underline{Z}^n$ is incorporated. Notice that $\underline{Z}^n = \underline{Z}$.

Since $g(\underline{z}|\underline{x})$ is a function of $\underline{x}_n$ only through $g_n(\underline{z}_n|\underline{x}_n, pa(\underline{x}_n), \underline{z}^{n-1})$, we can write that

$$\tilde{f}_n(\underline{x}_n|pa(\underline{x}_n), \underline{z}^n) = h_n(pa(\underline{x}_n), \underline{z}^n) \prod_{i=1}^{k} [\frac{f_{in}(\underline{x}_n|pa(\underline{x}_n))g_n(\underline{z}_n|pa(\underline{x}_n), \underline{x}_n, \underline{z}^{n-1})}{h_{in}(pa(\underline{x}_n), \underline{z}^n)}]^{w_{in}}$$

where throughout we let

$$h_{in}(pa(\underline{x}_l), \underline{z}^n) = \int ... \int f_{in}(\underline{x}_n|pa(\underline{x}_n))g_n(\underline{z}_n|\underline{x}_n, pa(\underline{x}_n), \underline{z}^{n-1})dx_{1n}...dx_{r(n),n}$$

for $i = 1, ..., k$. Noting that $\sum_{i=1}^{k} w_{in} = 1$ this can be arranged as

$$\begin{aligned}\tilde{f}_n(\underline{x}_n|pa(\underline{x}_n), \underline{z}^n) &= \frac{h_n(pa(\underline{x}_n), \underline{z}^n)}{\prod_{i=1}^{k} h_{in}^{w_{in}}(pa(\underline{x}_n), \underline{z}^n)} \prod_{i=1}^{k} f_{in}^{w_{in}}(\underline{x}_n|pa(\underline{x}_n))g_n(\underline{z}_n|\underline{x}_n, pa(\underline{x}_n), \underline{z}^{n-1}) \\ &= V_n(pa(\underline{x}_n), \underline{z}^n)\bar{f}_n(\underline{x}_n|pa(\underline{x}_n), \underline{z}^n)\end{aligned}$$

where

$$V_n(pa(\underline{x}_n), \underline{z}^n) = \frac{h_n(pa(\underline{x}_n), \underline{z}^n)v_n(pa(\underline{x}_n), \underline{z}^n)}{\prod_{i=1}^{k} h_{in}^{w_{in}}(pa(\underline{x}_n), \underline{z}^n)}$$

and where

$$v_n(pa(\underline{x}_n), \underline{z}^n) = 1/\int ... \int \prod_{i=1}^{k} f_{in}^{w_{in}}(\underline{x}_n|pa(\underline{x}_n))g_n(\underline{z}_n|\underline{x}_n, pa(\underline{x}_n), \underline{z}^{n-1})dx_{1n}...dx_{r(n),n} \quad .$$

Now we know that $\tilde{f}_n$ and $\bar{f}_n$ must integrate to 1 over $\underline{x}_n$ for all values of $pa(\underline{x}_n)$ and $\underline{z}^n$. It follows therefore that $V_n(pa(\underline{x}_n), \underline{z}^n)$ is identically one and so

$$\tilde{f}_n(\underline{x}_n|pa(\underline{x}_n), \underline{z}^n) = \bar{f}_n(\underline{x}_n|pa(\underline{x}_n), \underline{z}^n) \tag{6.6}$$

as required for each node of the ID induced by $\mathcal{G}^{pd}$.

After having shown that whether the combination is done before or after observing $\underline{Z}$ does not affect the conditional density of $\underline{X}_n|\{pa(\underline{X}_n), \underline{Z}^n\}$, we next consider the updating of $\underline{X}_{n-1}|pa(\underline{X}_{n-1})$ given $\underline{Z}^n$. First note that to update the distribution of $\underline{X}^{n-1} = (\underline{X}_1, \underline{X}_2, ..., \underline{X}_{n-1})$ [and hence of $\underline{X}_{n-1}|pa(\underline{X}_{n-1})$] in the light of $\underline{Z}^n$, we need to calculate the density $g^{(1)}$ of $\underline{z}^n = (\underline{z}_1, ..., \underline{z}_n)$ given $\underline{x}^{n-1} = (\underline{x}_1, ..., \underline{x}_{n-1})$. We can then simply use

Bayes rule as above. Now since the likelihood of $\underline{Z}$ is cutting, from the usual probability calculus we can write that,

$$g^{(1)}(\underline{z}^n|\underline{x}^{n-1})$$
$$= g_1(\underline{z}_1|\underline{x}_1)g_2(\underline{z}_2|\underline{x}_2, pa(\underline{x}_2), \underline{z}_1)...g_{n-1}(\underline{z}_{n-1}|\underline{x}_{n-1}, pa(\underline{x}_{n-1}), \underline{z}^{n-2})\tilde{g}_n(\underline{z}_n|\underline{x}^{n-1}, \underline{z}^{n-1})$$

where

$$\tilde{g}_n(\underline{z}_n|\underline{x}^{n-1}, \underline{z}^{n-1}) = \int ... \int g_n(\underline{z}_n|\underline{x}_n, pa(\underline{x}_n), \underline{z}^{n-1}) f_n(\underline{x}_n|pa(\underline{x}_n)) dx_{1n}...dx_{r(n),n} \ .$$

Note that f_n above is unambiguously defined since $\bar{f}_n = \tilde{f}_n$ by equation (6.6) and

$$f_n(\underline{x}_n|pa(\underline{x}_n)) = \bar{h}(pa(\underline{x}_n)) \prod_{i=1}^{k} f_{in}^{w_{in}}(\underline{x}_n|pa(\underline{x}_n)) \ .$$

Now, provided that w_{in} is a function of $\underline{x}$ only of terms in $pa(\underline{x}_n)$, it is clear that $\tilde{g}_n$ is a function of $\underline{z}^n$ and $pa(\underline{x}_n)$ only. So we can write that

$$\tilde{g}_n(\underline{z}_n|\underline{x}^{n-1}, \underline{z}^{n-1}) = \tilde{g}_n(\underline{z}_n|pa(\underline{x}_n), \underline{z}^{n-1})$$

where $\tilde{g}_n$ is a function of its arguments only. Also, since $\mathcal{G}^{pd}(\underline{X})$ has a decomposable induced ID, it will exhibit the running intersection property (see Lauritzen and Speigelhalter, 1988, p. 169). This states that there will exist an index $j(n)$ (say) such that $pa(\underline{x}_n) \subseteq \{\underline{x}_{j(n)}, pa(\underline{x}_{j(n)})\}$ with $j(n) = 1, ..., n-1$. So $\tilde{g}$ can be further simplified into

$$\tilde{g}_n(\underline{z}_n|\underline{x}^{n-1}, \underline{z}^{n-1}) = \tilde{g}_n(\underline{z}_n|\underline{x}_{j(n)}, pa(\underline{x}_{j(n)}), \underline{z}^{n-1}) \ .$$

Hence

$$g^{(1)}(\underline{z}^n|\underline{x}^{n-1}) = g_1^{(1)}(\underline{z}_1^{(1)}|\underline{x}_1)g_2^{(1)}(\underline{z}_2^{(1)}|\underline{x}_2, \underline{z}_1^{(1)})...g_{n-1}^{(1)}(\underline{z}_{n-1}^{(1)}|\underline{x}_{n-1}, pa(\underline{x}_{n-1}), \underline{z}^{(1),n-2})$$

where

$$g_j^{(1)}(\underline{z}_j^{(1)}|\underline{x}_j, pa(\underline{x}_j), \underline{z}^{(1),j-1}) = g_j(\underline{z}_j^{(1)}|\underline{x}_j, pa(\underline{x}_j), \underline{z}^{(1),j-1}) \ ,$$

and $\underline{Z}_l^{(1)} = \underline{Z}_l$ for $l \neq j(n)$, $l = 1, ..., n-1$ and

$$g_{j(n)}^{(1)}(\underline{z}_{j(n)}^{(1)}|\underline{x}_{j(n)}, pa(\underline{x}_{j(n)}), \underline{z}^{j(n)-1})$$
$$= g_{j(n)}(\underline{z}_{j(n)}|\underline{x}_{j(n)}, pa(\underline{x}_{j(n)}), \underline{z}^{j(n)-1})\tilde{g}_n(\underline{z}_n|\underline{x}_{j(n)}, pa(\underline{x}_{j(n)}), \underline{z}^{n-1}) \ .$$

It follows that $g^{(1)}(\underline{z}|\underline{x}^{n-1})$ *is a likelihood function which is cutting on* $\underline{x}^{(1)} = \underline{x}^{n-1}$ *since the only changed term is the likelihood arising as if from two conditionally independent observations* $\underline{Z}_j^{(1)} = (\underline{Z}_{j(n)}, \underline{Z}_n)$.

Now consider the density of $\underline{X}_{n-1}|\{pa(\underline{X}_{n-1}), \underline{Z}\}$. *The argument leading to equation (6.6) can now be directly applied but with* $n-1$ *replacing* n, g *replaced by* $g^{(1)}$ *and* $\underline{x}$ *by* $\underline{x}^{(1)}$ *to give us that*

$$\bar{f}_{n-1}(\underline{x}_{n-1}|pa(\underline{x}_{n-1}), \underline{z}^{n-1}) = \tilde{f}(\underline{x}_{n-1}|pa(\underline{x}_{n-1}), \underline{z}^{n-1}) \quad .$$

Since the induced ID is decomposable, we can find an index $j(n-1)$ *such that* $pa(\underline{x}_{n-1}) \subseteq \{\underline{x}_{j(n-1)}, pa(\underline{x}_{j(n-1)})\}$. *We can therefore use an argument exactly analogous to the one above, replacing* $\underline{x}^{(1)}$ *by* $\underline{x}^{(2)} = \underline{x}^{n-2} = (\underline{x}_1, ..., \underline{x}_{n-2})$, $(n-1)$ *by* $(n-2)$, g *by* $g^{(1)}$ *and* $g^{(1)}$ *by* $g^{(2)}$, *where*

$$g^{(2)}(\underline{z}^{n-1}|\underline{x}^{n-2}) = g_1^{(2)}(\underline{z}_1^{(2)}|\underline{x}_1)...g_{n-2}^{(2)}(\underline{z}_{n-2}^{(2)}|\underline{x}_{n-2}, pa(\underline{x}_{n-2}), \underline{z}^{(2),n-3})$$

and where

$$g_j^{(2)}(\underline{z}_j^{(2)}|\underline{x}_j, pa(\underline{x}_j), \underline{z}^{(2),j-1}) = g_j(\underline{z}_j^{(2)}|\underline{x}_j, pa(\underline{x}_j), \underline{z}^{(2),j-1})$$

and $\underline{z}_j^{(2)} = \underline{z}_j^{(1)}$ *for* $j \neq j(n-1)$, $j = 1, ..., n-2$, *and*

$$\begin{aligned}
&g_{j(n-1)}^{(2)}(\underline{z}_{j(n-1)}^{(2)}|\underline{x}_{j(n-1)}, pa(\underline{x}_{j(n-1)}), \underline{z}^{(2),j(n-1)-1}) \\
&= g_{j(n-1)}^{(1)}(\underline{z}_{j(n-1)}|\underline{x}_{j(n-1)}, pa(\underline{x}_{j(n-1)}), \underline{z}^{j(n-1)-1}) g_{n-1}^{(1)}(\underline{z}_{n-1}^{(1)}|pa(\underline{x}_{j(n-1)}), \underline{z}^{(2),j(n-1)})
\end{aligned}$$

we have that

$$\bar{f}_{n-2}(\underline{x}_{n-2}|pa(\underline{x}_{n-2}), \underline{z}^{n-1}) = \tilde{f}_{n-2}(\underline{x}_{n-2}|pa(\underline{x}_{n-2}), \underline{z}^{n-1}) \quad .$$

It should now be clear that we can proceed inductively backwards through the indices l, *starting from* $l = n$ *to prove that*

$$\bar{f}_l(\underline{x}_l|pa(\underline{x}_l), \underline{z}^l) = \tilde{f}_l(\underline{x}_l|pa(\underline{x}_l), \underline{z}^l) \quad ; \quad l = 1, ..., n \ .$$

But since the density of $\bar{f}$ *is uniquely determined by the product of* $\bar{f}_l$, $l = 1, ..., n$, *and* $\tilde{f}$ *by the product of conditional densities* $\tilde{f}_l$, $l = 1, ..., n$, *we have then proved that*

$$\bar{f}(\underline{x}|\underline{z}) = \tilde{f}(\underline{x}|\underline{z}) \quad ,$$

i.e. that the required external Bayesianity holds for $\mathcal{G}^{pd}(\underline{X}|\underline{Z})$. □

The above results guarantee that provided (i) the common PCG is decomposable and (ii) the common likelihood for sampling over ancestral sets associated with the PCG is cutting, the original PCG structure is preserved after new information is incorporated to the model. Moreover, under these circumstances, if the group is CEB on every chain element and agrees to be sequentially PPC then it *appears* EB on the PCG.

6.5 CEB Linear Opinion Pools.

If we impose linear opinion pools (LinOps), that is, pooling operators T which obey both MP and ZPP (see Sections 2.2.1 and 2.2.2), to the CI preserving PCGs $\mathcal{G}^{pd}(\underline{X})$, then we should have

$$T(f_1,\ldots,f_k)(\underline{x}) = \sum_{i=1}^{k} w_i f_i(\underline{x}) \ , \tag{6.7}$$

where $f_i(\underline{x})$ is the member E_i assessed joint density for $\underline{X}$ and w_i his weight in the pool, such that all the weights add up to one. Since the PCG is common to the group, $f_i(\underline{x})$ can be factorized accordingly for each $i = 1,\ldots,k$ and (6.7) would give

$$T(f_1,\ldots,f_k)(\underline{x}) = \sum_{i=1}^{k} \{w_i \prod_{j=1}^{n} f_{ij}[\underline{x}_j|pa(\underline{x}_j)]\} \ , \tag{6.8}$$

where $f_{ij}[\underline{x}_j|pa(\underline{x}_j)]$ is the member E_i assessed conditional density for the chain element $\underline{X}_j$.

Although this formulation preserves the MP, the weights are restricted and the possibility of their use to reflect relative expertise on components of $\mathcal{G}^{pd}(\underline{X})$ is not allowed. In fact, the LinOp (6.8) is not really a combination rule on densities associated to chain elements of $\mathcal{G}^{pd}(\underline{X})$. However, this could be circumvented by imposing the LinOps on the PCG's components $\underline{X}_j$ conditionally on their parent sets $pa(\underline{X}_j)$ instead (as done for the LogOps). In this case we would have that

$$T_j(f_{1j},\ldots,f_{kj})[\underline{x}_j|pa(\underline{x}_j)] = \sum_{i=1}^{k} w_{ij}[pa(\underline{x}_j)] f_{ij}[\underline{x}_j|pa(\underline{x}_j)] \ , \tag{6.9}$$

where $w_{ij}[pa(\underline{x}_j)]$ is E_i's weight for his assessment for $\underline{X}_j$ given $pa(\underline{x}_j)$ in the pool. The restriction $\sum_{i=1}^{k} w_{ij}[pa(\underline{x}_j)] = 1$ applies. Note that with this formulation the weights are not only allowed to reflect individual relative expertises on components of $\mathcal{G}^{pd}(\underline{X})$ but also to be set according to occurred values of the variables in $pa(\underline{x}_j)$, provided the restriction of adding up to unity remains. This possibility certainly does not violate the MP as happens when the weights are allowed to vary with $\underline{x}_j$ (e.g. Genest, 1984b). Also, for the same reasons described in Section 2.2.1 for the LogOps, the impossibility results of Dalkey (1972, 1975) and of Genest and Wagner (1984) do not apply here. See also McConway (1981).

Now, following the probability breakdown of the joint consensus density for all the components according to the associations in $\mathcal{G}^{pd}(\underline{X})$ gives

$$T(f_{11},\ldots,f_{1n},\ldots,f_{k1},\ldots,f_{kn})(\underline{x}) = \prod_{j=1}^{n}\Big\{\sum_{i=1}^{k} w_{ij}[pa(\underline{x}_j)]f_{ij}[\underline{x}_j|pa(\underline{x}_j)]\Big\}\ . \qquad (6.10)$$

Clearly (6.10) does not preserve both the MP and the external Bayesianity property. It also does not preserve the product form of the factorization of the joint density into conditional densities according to the PCG structure. Also the factors in (6.10) due to the cross products of different weighted densities over different components is of difficult interpretation. In fact, the products of densities within those factors could possibly be seen as components of dependence measurements among different members assessements for different variables, but the product of weights are rather arbitrary as weightings for those measurements. Consider, for example, the simple case where $k = n = 2$, $\dim(\underline{X}_i) = r(i) = 1$ for $i = 1,2$, and the common ID has a directed edge from X_1 to X_2. Thus, according to this

$$\begin{aligned} T[f_{11},f_{12},f_{21},f_{22}](x_1,x_2) = {} & w_1w_2(x_1)f_{11}(x_1)f_{12}(x_2|x_1) \\ & + w_1[1-w_2(x_1)]f_{11}(x_1)f_{22}(x_2|x_1) \\ & + (1-w_1)w_2(x_1)f_{21}(x_1)f_{12}(x_2|x_1) \\ & + (1-w_1)[1-w_2(x_1)]f_{21}(x_1)f_{22}(x_2|x_1)\ . \end{aligned}$$

Observe that the parcels formed by the products and cross products of the original weights in the sum above also add up to one. Now, taking the second factor of the above sum, for example, the weights cross-product parcel $w_1[1-w_2(x_1)]$ is itself weighting the product of $f_{11}(x_1)$ with $f_{22}(x_2|x_1)$. The issue here is whether the cross-product of densities can be interpreted as representing the dependence between E_1's assessment for X_1 and E_2's assessment for $X_2|x_1$. If one chooses to interpret this way then how reasonable would it be to take the parcel $w_1[1-w_2(x_1)]$, where w_1 measures E_1's relative expertise on X_1 and $1-w_2(x_1)$ measures E_2's relative expertise on $X_2|x_1$, as a weight (or discount) for the relative strength of that dependence. Also, it is not always that the product of densities is itself a probability density function.

In fact, it is well known (Wagner, 1982 and Genest, 1984c) that the linear opinion pools (LinOps) are not EB for tertiary σ-fields (see Section 2.2.1). This corresponds in the univariate setting to (2.8) or (2.9) not satisfying (2.10) unless they are dictatorships. Naturally dictatorships are undesirable in that context. However, if in our multivariate setting, one member of the group has so much more expertise about certain chain elements than the other members then it is natural for the group to choose that member's opinion as the consensus (the group's conditional density) on those elements. If this is the case for every chain element in $\mathcal{G}^{pd}$, then Theorem 6.4 applies and the joint consensus T on the whole PCG $\mathcal{G}^{pd}(\underline{X})$ is EB for cutting likelihoods. Note that, the selected model of member E_i for the element $\underline{X}_j$ synthesizes his opinion about $\underline{X}_j|pa(\underline{X}_j)$ and could have been obtained from a linear combination of other statistical models.

Alternatively consider the situation where for each chain element $\underline{X}_j$, only one of the members' models $f_{ij}[\underline{x}_j|pa(\underline{x}_j)]$ for $\underline{X}_j|pa(\underline{X}_j)$ is known to be the 'right' model but that member identity is unknown. Let $\alpha_j = i$ if the member E_i model is the 'best' (or right) model for $\underline{X}_j|pa(\underline{X}_j)$. Also, let $\underline{w}_j = (w_{1j}, \ldots, w_{kj})$ where the weight $w_{ij} = Pr\{\alpha_j = i\}$, that is, the probability that E_i's model is the best model for $\underline{X}_j|pa(\underline{X}_j)$. Thus, the consensus on $\underline{X}_j|pa(\underline{X}_j)$ also conditioned on $\alpha_j = i$,

$$f_j[\underline{x}_j|pa(\underline{x}_j), \alpha_j = i] = f_{ij}[\underline{x}_j|pa(\underline{x}_j)] \tag{6.11}$$

is CEB regarding the likelihood for $\underline{w}_j$ if all members agree on a distribution for $\underline{w}_j$. On the other hand,

$$f_j[\underline{x}_j|pa(\underline{x}_j)] = \sum_{i=1}^{k} w_{ij} f_j[\underline{x}_j|pa(\underline{x}_j), \alpha = i] \tag{6.12}$$

is a LinOp.

It seems that the above mentioned cases characterize the only instances where a LinOp can be CEB.

APPENDIX A6

Proof of Theorem 6.2

The proof of Theorem 6.2 needs intermediate results that we state here as lemmas and other theorems. Basically, the proof itself is split up to include all possible configurations of the underlying measure spaces, that is, the cases in which (Ω_j, μ_j^*) does not contain any atoms, or is purely atomic, or contains atoms but is not purely atomic. Some of these intermediate theorems and lemmas themselves need other results and concepts.

We begin Subsection A6.1 with a characterisation theorem of the conditional LogOp. This theorem is used to prove Theorem A6.4, a characterisation of the *conditional modified* LogOp for the case in which (Ω_j, μ_j^*) is not purely atomic and condition (6.1) holds for the functions P_j $(j = 1, \ldots, n)$. In Subsection A6.2 we introduce the concept of equivalence classes for CEB pooling operators which is useful for the statement of a theorem that characterises such operators without imposing any restrictions on the measure space and without requiring (6.1) to hold, namingly Theorem A6.6. For instance, this theorem is used to demonstrate Lemmas A6.7 and A6.8 which characterise the conditional modified LogOp for the case in which (Ω_j, μ_j^*) contains atoms but is not purely atomic. Finally, Subsection A6.3 proves the Theorem 6.2 .

A6.1 Characterisation of the Conditional LogOp.

Theorem A6.1. *Suppose that μ_j^* is not purely atomic in the measure space (Ω_j, μ_j^*). Let $\bar{f}_j : \Delta^k \to \Delta$, $j = 1, \ldots, n$, be pooling operators for which there exists Lebesgue measurable functions $P_j : (0, \infty)^k \to (0, \infty)$ such that*

$$\bar{f}_j(f_{1j}, \ldots, f_{kj})(\underline{x}_j|\mathrm{A}_j) = \frac{P_j(f_{1j}(\underline{x}_j|\mathrm{A}_j), \ldots, f_{kj}(\underline{x}_j|\mathrm{A}_j))}{\int \cdots \int P_j(f_{1j}, \ldots, f_{kj}) d\mu_{1j} \ldots d\mu_{r(j),j}} \; ; \mu_j^* \text{ a.e. }, \tag{A6.1}$$

holds for each $j = 1, \ldots, n$, where $\underline{x}_j|\mathrm{A}_j$ denotes the j-th set of variables given their parents in a chain graph $\mathcal{G}$ and where f_{ij} denotes the group member's conditional density function for this set of variables $\underline{x}_j|\mathrm{A}_j$. Then $\bar{f}_j$ is EB if and only if there exist weights $w_{ij}(\mathrm{A}_j) \geq 0$, $\sum_{i=1}^k w_{ij}(\mathrm{A}_j) = 1$, for $j = 1, \ldots, n$, such that

$$\bar{f}_j(f_{1j}, \ldots, f_{kj})(\underline{x}_j|\mathrm{A}_j) = \frac{\prod_{i=1}^k [f_{ij}(\underline{x}_j|\mathrm{A}_j)]^{w_{ij}(\mathrm{A}_j)}}{\int \cdots \int \prod_{i=1}^k [f_{ij}(\underline{x}_j|\mathrm{A}_j)]^{w_{ij}(\mathrm{A}_j)} d\mu_{1j} \ldots d\mu_{r(j),j}} \; ; \mu_j^* \text{ a.e. }, \tag{A6.2}$$

holds.

As in Genest (1984a), the proof of Theorem A6.1 is broken up into two lemmas. But first note that the function P_j in (A6.1) is the same as in (6.1) except that it is not indexed by $\underline{x}_j|\mathrm{A}_j$. This condition restricts the value of the likelihood of the combined conditional density at a particular point $\underline{y}_j$ in Ω_j to depend only on a function $f_{ij}(\underline{y}_j)$, $i = 1, ..., k$, and a proportionality constant which ensures $\bar{f}_j$ integrates to one. Also notice that since μ_j^* is not purely atomic in (Ω_j, μ_j^*) we can choose arbitrarily small disjoint subsets B_q $(q = 1, ..., s)$ of Ω_j such that $0 < \mu_j^*(B_q) < \epsilon$, $\epsilon > 0$.

The following Lemmas A6.2 and A6.3 are respectively the Lemmas 2.2 and 2.3 in Genest (1984) where P_j replaces P, c_j replaces c, z_{ij} replaces z_i, $l_j(\underline{x}_j|\mathrm{A}_j)$ replaces l, and so far. Their proofs are ommited here since apart from the mentioned replacements, they are completely analogous to those in Genest (1984a).

Lemma A6.2. *The functions $P_j : (0,\infty)^k \to (0,\infty)$ in (A6.1) are homogeneous, i.e. $P_j(c_jz_{1j}, ..., c_jz_{kj}) = c_jP_j(z_{1j}, ..., z_{kj})$ for all $c_j > 0$ and $z_{ij} > 0$, where $i = 1, \ldots, k$ and $j = 1, \ldots, n$.*

Lemma A6.3. *The function P_j in (A6.1) satisfies the functional equation*

$$P_j(s_{1j}z_{1j}, ..., s_{kj}z_{kj})P_j(1, ..., 1) = P_j(s_{1j}, ..., s_{kj})P_j(z_{1j}, ..., z_{kj})$$

for all s_{ij}, z_{ij} in $(0,\infty)$, where $i = 1, \ldots, k$ and $j = 1, ..., n$.

proof of Theorem A6.1. *For each $j = 1, \ldots, n$, let*

$$H_j(z_{1j}, ..., z_{kj}) = \frac{P_j(z_{1j}, ..., z_{kj})}{P_j(1, ..., 1)}.$$

Observe that H_j is Lebesgue measurable and satisfies the functional equation of Lemma A6.3 above, $H_j(s_{1j}z_{1j}, ..., s_{kj}z_{kj}) = H_j(s_{1j}, ..., s_{kj})H_j(z_{1j}, ..., z_{kj})$ on $(0,\infty)^k$ for all $j = 1, \ldots, n$. Therefore for each $j = 1, \ldots, n$, $H_j(z_{1j}, ..., z_{kj}) = \prod_{i=1}^k z_{ij}^{w_{ij}(\mathrm{A}_j)}$ for some $w_{ij}(\mathrm{A}_j) \in \Re^+$ $(i = 1, \ldots, k)$, and since $H_j(c_jz_{1j}, ..., c_jz_{kj}) = c_jH_j(z_{1j}, ..., z_{kj})$ for all $c_j > 0$ and $z_{ij} > 0$ ($i = 1, \ldots, k$ and $j = 1, \ldots, n$) then by Lemma A6.2, $\sum_{i=1}^k w_{ij}(\mathrm{A}_j) = 1$. The non-negativity of the w_{ij}'s come from the fact that $\int \cdots \int \prod_{i=1}^k f_{ij}^{w_{ij}(\mathrm{A}_j)} d\mu_{1j}...d\mu_{r(j),j} < \infty$ for all possible $f_{ij} \in \Delta$ ($i = 1, \ldots, k$ and $j = 1, \ldots, n$). □

Theorem A6.4. *Assume that (Ω_j, μ_j^*) is not purely atomic and that N is the complement set of the atoms. Let $\bar{f}_j : \Delta^k \to \Delta$, $j = 1, \ldots, n$, be externally Bayesian. Assume that (6.1) holds for $\mu_j^* \times$ Lebesgue measurable functions $P_j : \Omega_j \times (0,\infty)^k \to (0,\infty)$, $j = 1, \ldots, n$. Then,*

$$\bar{f}_j(f_{1j}, ..., f_{kj})(\underline{x}_j|\mathrm{A}_j) = \frac{p_j(\underline{x}_j|\mathrm{A}_j)\prod_{i=1}^k[f_{ij}(\underline{x}_j|\mathrm{A}_j)]^{w_{ij}(\mathrm{A}_j)}}{\int \cdots \int P_j(\cdot, f_{1j}, ..., f_{kj})d\mu_{1j}...d\mu_{r(j),j}}\ ;\ \mu_j^* \text{ - a.e. on } N, \quad \text{(A6.3)}$$

where $w_{ij}(\mathrm{A}_j) \geq 0$ and $\sum_{i=1}^k w_{ij}(\mathrm{A}_j) = 1$, $j = 1, \ldots, n$.

proof. *Define a new function* $NP_j : \Omega_j \times (0,\infty)^k \to (0,\infty)$ *by*

$$NP_j(\underline{x}_j|\mathbf{A}_j, z_{1j}, ..., z_{kj}) = P_j(\underline{x}_j|\mathbf{A}_j, z_{1j}, ..., z_{kj})/P(\underline{x}_j|\mathbf{A}_j, 1, ..., 1)$$

for all $z_{ij} \in (0,\infty)$, $i = 1, \ldots, k$ *and* $j = 1, \ldots, n$.

Notice that $NP_j(\underline{x}_j|\mathbf{A}_j, z_{1j}, ..., z_{kj})$, $j = 1, \ldots, n$, *is a function of the* z_{ij}*'s only,* $i = 1, \ldots, k$, *i.e.* $NP_j(\underline{x}_j|\mathbf{A}_j, z_{1j}, ..., z_{kj}) = NP_j(z_{1j}, ..., z_{kj})$, $j = 1, \ldots, n$. *In fact, given* $z_{ij} > 0$ *(*$i = 1, \ldots, k$ *and* $j = 1, \ldots, n$*) we can arbitrarily choose* $0 < \varepsilon_j < min\{1/2, 1/2z_{ij} : i = 1, \ldots, k\}$ *and let* B_{qj}, $q = 1, ..., 4$, *be a partition of* N *such that* $0 < \mu_j^*(B_{qj}) < \varepsilon_j$ *for* $k = 1, 2$; $0 < \mu_j^*(B_{3j}) < \infty$ *and* $\mu_j^*(B_{4j}) > 0$ *(*$j = 1, \ldots, n$*). If we define for each* $j = 1, \ldots, n$,

$$\begin{aligned}\bar{z}_{ij} = \mathcal{I}(B_{1j}) + \mathcal{I}(B_{2j}) &+ \frac{t_{ij} - \gamma_{ij} - \lambda_j(1-\gamma_{ij})}{\mu_j^*(B_{3j})(\xi_j - \lambda_j)}\mathcal{I}(B_{3j}) \\ &+ \frac{\xi_j(1-\gamma_{ij}) - (t_{ij} - \gamma_{ij})}{R_j(\xi_j - \lambda_j)} h_j \mathcal{I}(B_{4j})\end{aligned}$$

where $R_j = \int \cdots \int \mathcal{I}(B_{4j}) h_j d\mu_{1j} ... d\mu_{r(j),j} > 0$; $\mathcal{I}(B)$ *is the indicator function of the set* B; $h_j \in \Delta$ *is a fixed arbitrary density;* $t_{ij} = 1/z_{ij}$; $\gamma_j = \mu_j^*(B_{1j}) + \mu_j^*(B_{2j})$ *(note that* $0 < \gamma_j < min\{t_{ij}, 1 : i = 1, \ldots, k\}$ *and,* $\lambda_j > 0$ *and* $\xi_j < \infty$ *are chosen such that*

$$\begin{aligned}&\lambda_j < min\{[1-\gamma_{ij}]^{-1}(t_{ij} - \gamma_{ij}) : i = 1, \ldots, k\} \\ &\qquad\qquad \le max_{i,j}\{[1-\gamma_{ij}]^{-1}(t_{ij} - \gamma_{ij}) : i = 1, \ldots, k\} < \xi_j .\end{aligned}$$

Considering the likelihood

$$l_j = \mathcal{I}(B_{1j}) + \mathcal{I}(B_{2j}) + \xi_j \mathcal{I}(B_{3j}) + \lambda_j \mathcal{I}(B_{4j})$$

for each j, *we have that* $\int \cdots \int l_j \bar{z}_{ij} d\mu_{1j} ... d\mu_{r(j),j} = t_{ij}$ $(i = 1, \ldots, k)$.

Letting $\bar{h}_{ij} = l_j \bar{z}_{ij} / \int \cdots \int l_j \bar{z}_{ij} d\mu_{1j} ... d\mu_{r(j),j}$ *we have that for* $(\underline{x}_j|\mathbf{A}_j) \in B_1 \cup B_2$, $\bar{z}_{ij}(\underline{x}_j|\mathbf{A}_j) = 1$ *and* $\bar{h}_{ij} = z_{ij}$ *(*$i = 1, \ldots, k$ *and* $j = 1, \ldots, n$*). Since* $\bar{f}_j$ *is externally Bayesian then*

$$\frac{\bar{f}_j(\bar{h}_{1j}, ..., \bar{h}_{kj})(\underline{x}_j|\mathbf{A}_j)}{l_j(\underline{x}_j|bA_j)\bar{f}_j(\bar{z}_{1j}, ..., \bar{z}_{kj})(\underline{x}_j|\mathbf{A}_j)}$$

is constant (μ_j^* *- a.e.); and since* $l_j(\underline{x}_j|bA_j) = 1$ *on* $B_1 \cup B_2$, *and since (6.1) holds for* μ_j^* *- a.e. on* N, *it follows that*

$$NP_j(\underline{x}_j|\mathbf{A}_j, z_{1j}, ..., z_{kj}) = \frac{\bar{f}_j(\bar{h}_{1j}, ..., \bar{h}_{kj})(\underline{x}_j|\mathbf{A}_j)}{l_j(\underline{x}_j|bA_j)\bar{f}_j(\bar{z}_{1j}, ..., \bar{z}_{kj})(\underline{x}_j|\mathbf{A}_j)}$$

on $B_{1j} \cup B_{2j}$ $(j = 1, \ldots, n)$, *i.e.* NP_j *is essentially constant as a function of* $\underline{x}_j|\mathrm{A}_j$ *on* $B_1 \cup B_2$. *To see that this is valid on* N, *lets assume to the contrary that there exists a subset* C *of* N *with* $\mu_j^*(C) > 0$ *such that* $NP_j(\underline{x}_j|\mathrm{A}_j, z_{1j}, ..., z_{kj}) > (<) NP_j(z_{1j}, ..., z_{kj})$, $j = 1, \ldots, n$, *for almost all* $\underline{x}_j|\mathrm{A}_j$ *in* C. *Choosing a subset of* C *with positive measure at most* ϵ *and naming it* B_{2j}, *we repeat the above development with* B_{1j} *being kept the same as before.* $NP_j(\underline{x}_j|\mathrm{A}_j, z_{1j}, ..., z_{kj})$ *for* $j = 1, \ldots, n$, *is still essentially constant on* $B_{1j} \cup B_{2j}$ *in contradiction to the assumption that* $NP_j(\underline{x}_j|\mathrm{A}_j, z_{1j}, ..., z_{kj}) > (<) NP_j(z_{1j}, ..., z_{kj})$ *on* B_{2j}, $j = 1, \ldots, n$.

Now that we have shown that $NP_j(\underline{x}_j|\mathrm{A}_j, z_{1j}, ..., z_{kj})$, $j = 1, \ldots, n$, *is a function of the* z_{ij}*'s only* $(i = 1, \ldots, k$ *and* $j = 1, \ldots, n)$, *the proof is straightforward by defining a new pooling operator* $f_j^* : \Delta^k \to \Delta$ *such that*

$$f_j^*(f_{1j}, ..., f_{kj})(\underline{x}_j|\mathrm{A}_j) = \frac{NP_j(f_{1j}(\underline{x}_j|\mathrm{A}_j), ... f_{kj}(\underline{x}_j|\mathrm{A}_j))}{\int \cdots \int NP_j(f_{1j}, ..., f_{kj}) d\mu_{1j} ... d\mu_{r(j),j}} \; .$$

It is clear that f_j^* *is externally Bayesian and of the form*

$$f_j^*(f_{1j}, ..., f_{kj})(\underline{x}_j|\mathrm{A}_j) \propto P_j(f_{1j}(\underline{x}_j|\mathrm{A}_j), ..., f_{kj}(\underline{x}_j|\mathrm{A}_j)) \quad ; \mu_j^*\text{-a.e.} \; ,$$

where the proportionality constant is independent of $\underline{x}_j|\mathrm{A}_j$ *for each* $j = 1, \ldots, n$, *with* NP_j *replaced by* P_j. *Applying Theorem A6.1 we conclude that*

$$NP_j(z_{1j}, ..., z_{kj}) = \prod_{i=1}^{k} z_{ij}^{w_{ij}(\mathrm{A}_j)}$$

for some $w_{ij}(\mathrm{A}_j) \geq 0$ *such that* $\sum_{i=1}^{k} w_{ij}(\mathrm{A}_j) = 1$ *for each* $j = 1, \ldots, n$. *The relation (A6.3) is then achieved if we set* $p_j(\underline{x}_j|\mathrm{A}_j) = P_j(\underline{x}_j|\mathrm{A}_j, 1, ..., 1)$ *for all* $\underline{x}_j|\mathrm{A}_j$ *in* N $(j = 1, \ldots, n)$. □

A6.2 Equivalence Classes for CEB Pooling Operators.

The concept of *equivalence classes* allows the characterisation of a theoretically more general class of CEB pooling operators than that of the conditional LogOp. In fact, the domain of pooling operators is divided into particular sets of density functions (equivalence classes) in such a way that, given the value of an operator at one member of an equivalence class, the CEB property defines the value of the operator at all other members of that class.

Definition A6.5 (equivalence class). *Two vectors* $(f_1, ..., f_n)$ *and* $(f_1^*, ..., f_n^*)$ *in* Δ^k *belongs to the same equivalence class* α *if and only if there exists a likelihood function* $l : \Omega_j \to (0, \infty)$ *such that*

$$f_i^* = \frac{l f_i}{\int \cdots \int l f_i d\mu_{1j} ... d\mu_{r(j),j}} \; ; \mu_j^* \text{ - a.e. },$$

for all $i = 1, \ldots, k$.

Observe that if $(f_1, ..., f_n)$ and $(f_1^*, ..., f_n^*)$ belongs to the same equivalence class α then for all k, r there exists $c_{kr} > 0$ such that

$$\frac{f_k}{f_r} = c_{kr} \frac{f_k^*}{f_r^*} \; ; \mu_j^* \text{ - a.e. },$$

and, in this case $(f_1, ..., f_n)$ is said to equivalent to $(f_1^*, ..., f_n^*)$, i.e. $(f_1, ..., f_n) \sim (f_1^*, ..., f_n^*)$.

We can now state the following theorem.

Theorem A6.6. *Let* $\mathcal{A}$ *be the space of equivalence classes and let* $\bar{f}_j : \Delta^k \to \Delta$ *be an arbitrary pooling operator for each* $j = 1, \ldots, n$ *associated with the partially complete chain graph* $\mathcal{G}^p$*. Then* $\bar{f}_j$ $(j = 1, ..., n)$ *is CEB if and only if, for each* $j = 1., , , .n$,

$$\bar{f}_j(f_{1j}, ..., f_{kj}) \propto b_{\alpha_j} v_{\alpha_j} \frac{f_{1j}}{f_{1j}^{\alpha_j}} \; ; \mu_j^* \text{ - a.e. }, \tag{A6.4}$$

where α_j *is the equivalence class of* $(f_{1j}, ..., f_{kj})$*, and for each* $\alpha_j in \mathcal{A}$*,* b_{α_j} *is some essentially bounded function and* v_{α_j} *is some function such that* $v_{\alpha_j} \geq max\{f_{1j}^{\alpha_j}, ..., f_{kj}^{\alpha_j}\}$*,* μ_j^* *almost everywhere.*

Although formula (A6.4) appears to be f_{1j}'s dictatorship, the opinions of all other members of the group are considered in the consensus through the knowledge of the equivalence class corresponding to the vector $(f_{1j}, ..., f_{kj})$. Note that each equivalence class is characterised by an arbitrary vector $(f_{1j}^{\alpha_j}, ..., f_{kj}^{\alpha_j})$ in the class or, equivalently, by a component $f_{kj}^{\alpha_j}$ and the ratios $f_{ij}^{\alpha_j} / f_{kj}^{\alpha_j}$, $i = 1, \ldots, k$, $i \neq k$, which are invariant within a given class $(j = 1, \ldots, n)$.

proof. *First, for any fixed* b_{α_j} *and* v_{α_j}*,* $\bar{f}_j(f_{1j}, ..., f_{kj})$ *in (A6.4) is well defined and is CEB for all* $j = 1, \ldots, n$ *in the chain graph* $\mathcal{G}$*. In fact, for the common likelihood functions* l_j *in* $\mathcal{G}^p$*, since* $f_1/f_1^\alpha = c_{1i} f_i / f_i^\alpha$*,* μ_j^* *- a.e. , we have that*

$$\begin{aligned} \bar{f}_j(l_j f_{1j}, ..., l_j f_{kj}) &\propto b_{\alpha_j} v_{\alpha_j} \frac{l_j f_{1j}}{f_{1j}^{\alpha_j}} \propto b_{\alpha_j} v_{\alpha_j} l_j \frac{f_{ij}}{f_{ij}^{\alpha_j}} \\ &\propto l_j \bar{f}(f_{1j}, ..., f_{kj}) \end{aligned}$$

and $\bar{f}_j$ is CEB for all j.

To prove that if $\bar{f}_j$ is CEB then it is of the form (A6.4), lets for each $\alpha_j \in \mathcal{A}$, denote $h_{\alpha_j} = \bar{f}_j(f_{1j}^{\alpha_j}, ..., f_{kj}^{\alpha_j})$ for all j. For arbitrary sets of density functions $(f_{1j}, ..., f_{kj})$ equivalent to $(f_{1j}^{\alpha_j}, ..., f_{kj}^{\alpha_j})$, consider the likelihoods $l_j = f_{1j}/f_{1j}^{\alpha_j}$. Since for each j, $f_{ij}^{\alpha_j}/f_{1j}^{\alpha_j} = c_{i1j} f_{ij}/f_{1j}$, μ_j^* - a.e. , we have that $\int \cdots \int l_j f_{ij}^{\alpha_j} d\mu_{1j}...d\mu_{r(j),j} = c_{i1j} < \infty$ and

$$l_j f_{ij}^{\alpha_j} / \int \cdots \int l_j f_{ij}^{\alpha_j} d\mu_{1j}...d\mu_{r(j),j} = f_{ij}$$

for all i and j. From the fact that $\bar{f}_j$ is CEB, we have that

$$\begin{aligned} \bar{f}(f_{1j}, ..., f_{kj}) &= \bar{f}\Big(\frac{l_j f_{1j}^{\alpha_j}}{\int \cdots \int l_j f_{1j}^{\alpha_j} d\mu_{1j}...d\mu_{r(j),j}}, ..., \frac{l_j f_{kj}^{\alpha_j}}{\int \cdots \int l_j f_{kj}^{\alpha_j} d\mu_{1j}...d\mu_{r(j),j}}\Big) \\ &\propto l_j \bar{f}_j(f_{1j}^{\alpha_j}, ..., f_{kj}^{\alpha_j}) = \frac{f_{1j} h_{\alpha_j}}{f_{1j}^{\alpha_j}}, \ \mu_j^* \text{ - a.e. }, \end{aligned}$$

is valid as long as $(f_{1j}, ..., f_{kj}) \in \alpha_j$.

Now, to see that b_{α_j} is essentially bounded, assume that for each j,

$$v_{\alpha_j} \geq max\{f_{1j}^{\alpha_j}, ..., f_{kj}^{\alpha_j}\} .$$

Let $b_{\alpha_j} = h_{\alpha_j}/v_{\alpha_j}$, pick an arbitrary function p_j in Δ and define $f_{1j} = p_j f_{1j}^{\alpha_j}/v_{\alpha_j}$, μ_j^* - a.e. $(j = 1, \ldots, n)$. Since $\int \cdots \int f_{1j} f_{ij}^{\alpha_j}/f_{1j}^{\alpha_j} d\mu_{1j}...d\mu_{r(j),j} < \infty$ for i and j, we can also define $f_{2j}, ..., f_{kj} \in \Delta$ such that $(f_{1j}, ..., f_{kj}) \sim (f_{1j}^{\alpha_j}, ..., f_{kj}^{\alpha_j})$ by $f_{ij} \propto f_{1j} f_{ij}^{\alpha_j}/f_{1j}^{\alpha_j}$ for $i \geq 2$ and $j = 1, \ldots, n$. Then $\bar{f}_j(f_{1j}, ..., f_{kj}) \propto p_j h_{\alpha_j}/v_{\alpha_j}$, and thus $\int \cdots \int p_j h_{\alpha_j}/v_{\alpha_j} d\mu_{1j}...d\mu_{r(j),j} < \infty$ for all $p_j \in \Delta$, $j = 1, \ldots, n$.

Now, the conclusion follows from Theorem 20.15 in Hewitt & Stromberg (1965). In fact, according to that theorem, if $\int \cdots \int p_j b_{\alpha_j} d\mu_{1j}...d\mu_{r(j),j} < \infty$ for all p_j such that $\int \cdots \int p_j d\mu_{1j}...d\mu_{r(j),j} < \infty$ then b_{α_j} is bounded. □

Lemma A6.7. *Let (Ω_j, μ_j^*) be a quaternary measure space that contains at least two atoms, and let $\bar{f}_j : \Delta^k \to \Delta$ be a CEB pooling operator. Suppose there exist $\mu_j^* \times$ Lebesgue measurable functions $P_j : \Omega_j \times (0, \infty)^k \to (0, \infty)$ such that (6.1) holds for all conditional densities $f_{ij} \in \Delta$, $i = 1, \ldots, k$ and $j = 1, \ldots, n$, associated to a partially complete chain graph $\mathcal{G}^p$. Then for every pair of atoms $(\underline{x}_j|\mathrm{A}_j, \underline{y}_j|\mathrm{D}_j)$ in Ω_j^2, the identity*

$$\frac{\bar{f}_j(f_{1j}, ..., f_{kj})(\underline{x}_j|\mathrm{A}_j)}{\bar{f}_j(f_{1j}, ..., f_{kj})(\underline{y}_j|\mathrm{D}_j)} = \frac{\bar{f}_j(h_{1j}, ..., h_{kj})(\underline{x}_j|\mathrm{A}_j)}{\bar{f}_j(h_{1j}, ..., h_{kj})(\underline{y}_j|\mathrm{D}_j)} \tag{A6.5}$$

where $\mathbf{D}_j = \{\pi(\underline{Y}_j) = \pi(\underline{y}_j)\}$ *is the set of fixed values for the parent set of* $\underline{Y}_j$ *for each* j, *holds for all* f_{ij} *and* h_{ij} $(i = 1, \ldots, k$ *and* $j = 1, \ldots, n)$ *in* Δ *for which*

$$\frac{f_{ij}(\underline{x}_j|bA_j)}{f_{ij}(\underline{y}_j|\mathbf{D}_j)} = \frac{h_{ij}(\underline{x}_j|bA_j)}{h_{ij}(\underline{y}_j|\mathbf{D}_j)} \tag{A6.6}$$

for all i *and* j.

proof. *If for each* j *the vectors of* μ_j^**-densities* $(f_{1j}, ..., f_{kj})$ *and* $(h_{1j}, ..., h_{kj})$ *satisfy condition (A6.6) and belong to the same equivalence class* α_j, *then substituting (A6.4) of Theorem A6.6 in (A6.6), (A6.5) is obtained.*

Now we have to show that if there exist f_{ij}, h_{ij} *in* Δ *for all* i *and* j *such that (A6.6) holds, then such densities belong to the same equivalence class. To do so, set* $s_{ij} = f_{ij}(\underline{x}_j|\mathbf{A}_j)$, $u_{ij} = f_{ij}(\underline{y}_j|\mathbf{D}_j)$, $s_{ij}^* = h_{ij}(\underline{x}_j|\mathbf{A}_j)$ *and* $u_{ij}^* = h_{ij}(\underline{y}_j|\mathbf{D}_j)$ *so that*

$$\frac{s_{ij}}{s_{ij}^*} = \frac{u_{ij}}{u_{ij}^*} = t_{ij}$$

for all i *and* j. *Choose four sets in* Ω_j *for each* j, B_{qj}, *with* $0 < \mu_j^*(B_{qj}) < \infty$ *for* $k = 1, 2, 3$ *and* $\mu_j^*(B_{4j}) > 0$, *such that* $s_{ij}\mu_j^*(B_{1j}) + u_{ij}\mu_j^*(B_{2j}) < 1$ *and* $s_{ij}^*\mu_j^*(B_{1j}) + u_{ij}^*\mu_j^*(B_{2j}) < 1$ *for all* i. *Make* $B_{1j} = \{\underline{x}_j|\mathbf{A}_j\}$ *and* $B_{2j} = \{\underline{y}_j|\mathbf{D}_j\}$ *and construct densities* $(f_{1j}', ..., f_{kj}')$ *and likelihoods* l_j *for these densities such that* $f_{ij}'(\underline{x}_j|\mathbf{A}_j) = s_{ij}$, $f_{ij}'(\underline{y}_j|\mathbf{D}_j) = u_{ij}$, $h_{ij}'(\underline{x}_j|\mathbf{A}_j) = s_{ij}^*$ *and* $h_{ij}'(\underline{y}_j|\mathbf{D}_j) = u_{ij}^*$ *where* $h_{ij}' = l_j f_{ij}' / \int \cdots \int l_j f_{ij}' d\mu_{1j}...d\mu_{r(j),j}$ *(for all* i *and* j*), in a similar way of the preceding proofs.*

Denote $\gamma_{ij} = s_{ij}\mu_j^*(B_{1j}) + u_{ij}\mu_j^*(B_{2j})$. *Observe that* $s_{ij}\mu_j^*(B_{1j}) + u_{ij}\mu_j^*(B_{2j}) = t_{ij}[s_{ij}^*\mu_j^*(B_{1j}) + u_{ij}^*\mu_j^*(B_{2j})] < t_{ij}$ *for all* i *and* j, *and thus* $0 < \gamma_{ij} < min\{t_{ij}, 1\}$ *for each* i *and* j. *Also choose* $\lambda_j > 0$ *and* $\xi_j < \infty$ *such that*

$$\begin{aligned}\lambda_j &< min\{[1 - \gamma_{ij}]^{-1}(t_{ij} - \gamma_{ij}) : i = 1, \ldots, k\} \\ &\geq max\{[1 - \gamma_{ij}]^{-1}(t_{ij} - \gamma_{ij}) : i = 1, \ldots, k\} < \xi_j \ .\end{aligned}$$

Fixing an arbitrary density $g_j \in \Delta$ *we can define for all* i *and* j

$$\begin{aligned}f_{ij}' = s_{ij}\mathcal{I}(B_{1j}) + u_{ij}\mathcal{I}(B_{2j}) &+ \frac{t_{ij} - \gamma_{ij} - \lambda_j(1 - \gamma_{ij})}{\mu_j^*(B_{3j})(\xi_j - \lambda_j)}\mathcal{I}(B_{4j}) \\ &+ \frac{\xi_j(1 - \gamma_{ij}) - (t_{ij} - \gamma_{ij})}{R_j(\xi_j - \lambda_j)} g_j \mathcal{I}(B_{4j})\end{aligned}$$

where $R_j = \int \cdots \int \mathcal{I}(B_{4j}) g_j d\mu_{1j} ... d\mu_{r(j),j} > 0$ *and* $\mathcal{I}(B)$ *is the indicator of the set* B. *Note that* $\int \cdots \int f'_{ij} d\mu_{1j} ... d\mu_{r(j),j} = 1$ *and* $f'_{ij}(\underline{x}_j|\mathbf{A}_j) = s_{ij}$, $f'_{ij}(\underline{y}_j|\mathbf{D}_j) = u_{ij}$ *for all* i *and* j. *Finally, consider the likelihoods*

$$l_j = \mathcal{I}(B_{1j}) + \mathcal{I}(B_{2j}) + \xi_j(B_{3j}) + \lambda_j \mathcal{I}(B_{4j}).$$

It can be seen that $\int \cdots \int l_j f'_{ij} d\mu_{1j} ... d\mu_{r(j),j} = t_{ij}$ *so that for all* i *and* j, $h'_{ij}(\underline{x}_j|\mathbf{A}_j) = s^*_{ij}$ *and* $h'_{ij}(\underline{y}_j|\mathbf{D}_j) = u^*_{ij}$. $\square$

The preceding Lemma A6.7 shows that if $\bar{f}_j : \Delta^k \to \Delta$ is CEB and satisfies (6.1) for some functions $P_j : \Omega_j \times (0,\infty) \to (0,\infty)$ then for all pairs of atoms $(\underline{x}_j|\mathbf{A}_j, \underline{y}_j|\mathbf{D}_j)$ in Ω_j^2, there must exist Lebesgue measurable functions $Q_j(\underline{x}_j|\mathbf{A}_j, \underline{y}_j|\mathbf{D}_j) : (0,\infty)^k \to (0,\infty)$ for all j such that for all conditional densities $f_{1j}, ..., f_{kj} \in \Delta$,

$$\frac{\bar{f}_j(f_{1j}, ..., f_{kj})(\underline{x}_j|\mathbf{A}_j)}{\bar{f}_j(f_{1j}, ..., f_{kj})(\underline{y}_j|\mathbf{D}_j)} = Q_j(\underline{x}_j|\mathbf{A}_j, \underline{y}_j|\mathbf{D}_j)\left[\frac{f_{1j}(\underline{x}_j|\mathbf{A}_j)}{f_{1j}(\underline{y}_j|\mathbf{D}_j)}, ..., \frac{f_{kj}(\underline{x}_j|\mathbf{A}_j)}{f_{kj}(\underline{y}_j|\mathbf{D}_j)}\right]. \tag{A6.7}$$

In fact, according to (6.1) we have for each j that

$$\frac{\bar{f}_j(f_{1j}, ..., f_{kj})(\underline{x}_j|\mathbf{A}_j)}{\bar{f}_j(f_{1j}, ..., f_{kj})(\underline{y}_j|\mathbf{D}_j)} = \frac{P_j(\underline{x}_j|\mathbf{A}_j, f_{1j}(\underline{x}_j|\mathbf{A}_j), ..., f_{kj}(\underline{x}_j|\mathbf{A}_j))}{P_j(\underline{y}_j|\mathbf{D}_j, f_{1j}(\underline{y}_j|\mathbf{D}_j), ..., f_{kj}(\underline{y}_j|\mathbf{D}_j))}.$$

If we replace, for each j, the right-hand side of the above identity for a function of the type $Q_j(\underline{x}_j|\mathbf{A}_j, \underline{y}_j|\mathbf{D}_j)$, then Lemma A6.7 allows us to obtain (A6.7). Also notice that $d_j = dim\{\mathbf{A}_j\} = dim\{\mathbf{D}_j\}$ for all j, since we have the same chain graph $\mathcal{G}^p$.

The following lemma derive a more specific form for the right-hand side of (A6.7) and allows us to derive a formula similar to (6.2) but for atoms.

Lemma A6.8. *In addition to the hypotheses of Lemma A6.7, assume that* (Ω_j, μ^*_j) *contains at least three atoms. Then, there exist, for each vector* $j = 1, \ldots, n$ *of a partially complete chain graph* $\mathcal{G}^p$, *constant terms* $v_{1j}(\mathbf{A}_j, \mathbf{D}_j), ..., v_{kj}(\mathbf{A}_j, \mathbf{D}_j)$ *such that for all functions* $g_{1j}, ..., g_{kj} > 0$ *and every pair of atoms* $(\underline{x}_j|\mathbf{A}_j, \underline{y}_j|\mathbf{D}_j)$ *in* Ω_j^2, *we have for each* j *that*

$$Q_j(\underline{x}_j|\mathbf{A}_j, \underline{y}_j|\mathbf{D}_j)(g_{1j}, ..., g_{kj}) = Q_j(\underline{x}_j|\mathbf{A}_j, \underline{y}_j|\mathbf{D}_j)(\underline{1}) \prod_{i=1}^{k} g_{ij}^{v_{ij}(\mathbf{A}_j, \mathbf{D}_j)} \tag{A6.8}$$

where $\underline{1} = (1, ...1)$ *is* n*-dimensional.*

proof. *Define for each j the new functions $NQ_j(\underline{x}_j|\mathrm{A}_j, \underline{y}_j|\mathrm{D}_j) : (0,\infty)^k \to (0,\infty)$ by setting for all atoms $\underline{x}_j|\mathrm{A}_j \neq \underline{y}_j|\mathrm{D}_j$,*

$$NQ_j(\underline{x}_j|\mathrm{A}_j, \underline{y}_j|\mathrm{D}_j)(g_{1j}, ..., g_{kj}) = \frac{Q_j(\underline{x}_j|\mathrm{A}_j, \underline{y}_j|\mathrm{D}_j)(g_{1j}, ..., g_{kj})}{Q_j(\underline{x}_j|\mathrm{A}_j, \underline{y}_j|\mathrm{D}_j)(1, ..., 1)} .$$

Let $\underline{x}_j|\mathrm{A}_j$, $\underline{y}_j|\mathrm{D}_j$ and $\underline{z}_j|\mathrm{E}_j$ be three distinct atoms in Ω_j where $E_j = \{\pi(\underline{Z}_j) = \pi(\underline{z}_j)\}$ represent fixed values for the parents set of the $\underline{Z}_j$'s. Pick $\epsilon_j > 0$ small enough that there exist conditional densities in Δ which assume any of the values ϵ_j, $\epsilon_j g_{ij}$, or ϵ_j/h_{ij} at any of these three atoms. Writing $\underline{\epsilon}_j = (\epsilon_j, ..., \epsilon_j)$, $\underline{g}_j = (g_{1j}, ..., g_{kj})$ and $\underline{h}_j = (h_{1j}, ..., h_{kj})$, and assuming that all operations on vectors are carried out component-wise, we have for all j that

$$\begin{aligned} NQ_j(\underline{x}_j|\mathrm{A}_j, \underline{y}_j|\mathrm{D}_j)(\underline{g}_j\underline{h}_j) &= \frac{P_j(\underline{x}_j|\mathrm{A}_j, \epsilon_j\underline{g}_j)/P_j(\underline{y}_j|\mathrm{D}_j, \epsilon_j/\underline{h}_j)}{P_j(\underline{x}_j|\mathrm{A}_j, \underline{\epsilon}_j)/P_j(\underline{y}_j|\mathrm{D}_j, \underline{\epsilon}_j)} \\ &= \frac{P_j(\underline{z}_j|\mathrm{E}_j, \underline{\epsilon}_j)/P_j(\underline{y}_j|\mathrm{D}_j, \epsilon_j/\underline{h}_j)}{P_j(\underline{z}_j|\mathrm{E}_j, \underline{\epsilon}_j)/P_j(\underline{y}_j|\mathrm{D}_j, \underline{\epsilon}_j)} \frac{P_j(\underline{x}_j|\mathrm{A}_j, \epsilon_j\underline{g}_j)/P_j(\underline{z}_j|\mathrm{E}_j, \underline{\epsilon}_j)}{P_j(\underline{x}_j|\mathrm{A}_j, \underline{\epsilon}_j)/P_j(\underline{z}_j|\mathrm{E}_j, \underline{\epsilon}_j)} \\ &= NQ_j(\underline{x}_j|\mathrm{A}_j, \underline{y}_j|\mathrm{D}_j)(\underline{g}_j) NQ_j(\underline{x}_j|\mathrm{A}_j, \underline{y}_j|\mathrm{D}_j)(\underline{h}_j) \qquad \text{(A6.9)} \end{aligned}$$

for all $\underline{g}_j$ and $\underline{h}_j$ in $(0,\infty)^k$ and $j = 1, \ldots, n$. Assuming for a moment (and we will prove it later) that for all j, $NQ_j(\underline{x}_j|\mathrm{A}_j, \underline{y}_j|\mathrm{D}_j)(.) = NQ_j(.)$ does not depend on $\underline{x}_j|\mathrm{A}_j$ or $\underline{y}_j|\mathrm{D}_j$, the above equation reduces to Cauchy's functional equation

$$NQ_j(\underline{g}_j\underline{h}_j) = NQ_j(\underline{g}_j)NQ_j(\underline{h}_j) \qquad \text{(A6.10)}$$

and we can use a multivariate extension of Aczél's (1966) Theorem 3 to conclude that

$$NQ_j(\underline{g}_j) = \prod_{i=1}^{k} g_{ij}^{v_{ij}(A_j, D_j)}$$

for some arbitrary constant terms $v_{ij} \in \Re$, possibly functions of the fixed A_j and D_j, $v_{ij}(\mathrm{A}_j, \mathrm{D}_j)$ for all i and j. Notice that for each i and j, $v_{ij}(\mathrm{A}_j, \mathrm{D}_j)$ is neither a function of $\underline{x}_j|\mathrm{A}_j$ nor $\underline{y}_j|\mathrm{D}_j$. The non-measurable solutions of (A6.10) are ruled out here because Q_j, P_j, and hence NQ_j were assumed to be Lebesgue measurable.

Now to see that $NQ_j(\underline{x}_{1j}|\mathrm{A}_{1j}, \underline{x}_{2j}|\mathrm{A}_{2j})(\underline{g}_j) = NQ_j(\underline{g}_j)$ is a function of $\underline{g}_j$ only, consider $NQ_j(\underline{y}_{1j}|\mathrm{D}_{1j}, \underline{y}_{2j}|\mathrm{D}_{2j})$ for some $\underline{y}_{1j}|\mathrm{D}_{1j}, \underline{y}_{2j}|\mathrm{D}_{2j} \in \Omega_j$. Assume without loss of generality that $\underline{g}_j \neq 1$ so that $\underline{x}_{1j}|\mathrm{A}_{1j} \neq \underline{x}_{2j}|\mathrm{A}_{2j}$ and $\underline{y}_{1j}|\mathrm{D}_{1j} \neq \underline{y}_{2j}|\mathrm{D}_{2j}$. If $\{\underline{x}_{1j}|\mathrm{A}_{1j}, \underline{x}_{2j}|\mathrm{A}_{2j}\} = \{\underline{y}_{1j}|\mathrm{D}_{1j}, \underline{y}_{2j}|\mathrm{D}_{2j}\}$, then either $\underline{x}_{kj}|\mathrm{A}_{kj} = \underline{y}_{kj}|\mathrm{D}_{kj}$ or $\underline{x}_{kj}|\mathrm{A}_{kj} = \underline{y}_{3-k,j}|\mathrm{D}_{3-k,j}$. The *former case is trivial, and in the latter case we can choose* $\underline{x}_{3j}|A_{3j} \notin \{\underline{x}_{1j}|A_{1j}, \underline{x}_{2j}|A_{2j}\}$ *and apply (A6.9) to obtain successively for each* j,

$$\begin{aligned} NQ_j(\underline{x}_{1j}|A_{1j}, \underline{x}_{2j}|A_{2j})(\underline{g}_j) &= NQ_j(\underline{x}_{1j}|A_{1j}, \underline{x}_{3j}|A_{3j})(\underline{1})NQ_j(\underline{x}_{3j}|A_{3j}, \underline{x}_{2j}|A_{2j})(\underline{g}_j) \\ &= NQ_j(\underline{x}_{3j}|A_{3j}, \underline{x}_{1j}|A_{1j})(\underline{g}_j)NQ_j(\underline{x}_{1j}|A_{1j}, \underline{x}_{2j}|A_{2j})(\underline{1}) \\ &= NQ_j(\underline{x}_{2j}|A_{2j}, \underline{x}_{3j}|A_{3j})(\underline{1})NQ_j(\underline{x}_{2j}|A_{2j}, \underline{x}_{1j}|A_{1j})(\underline{g}_j) \\ &= NQ_j(\underline{x}_{2j}|A_{2j}, \underline{x}_{1j}|A_{1j})(\underline{g}_j) \ . \end{aligned}$$

On the other hand, if $\{\underline{x}_{1j}|A_{1j}, \underline{x}_{2j}|A_{2j}\} \neq \{\underline{y}_{1j}|D_{1j}, \underline{y}_{2j}|D_{2j}\}$, *then at least one of* $\underline{y}_{1j}|D_{1j}$, $\underline{y}_{2j}|D_{2j}$ *must be different from both* $\underline{x}_{1j}|A_{1j}, \underline{x}_{2j}|A_{2j}$. *Choose, for example, it to be* $\underline{y}_{2j}|D_{2j}$. *Using (A6.9) again, we have for each* j *that*

$$\begin{aligned} NQ_j(\underline{x}_{1j}|A_{1j}, \underline{x}_{2j}|A_{2j})(\underline{g}_j) &= NQ_j(\underline{x}_{1j}|A_{1j}, \underline{y}_{2j}|D_{2j})(\underline{g}_j)NQ_j(\underline{y}_{2j}|D_{2j}, \underline{x}_{2j}|A_{2j})(\underline{1}) \\ &= NQ_j(\underline{y}_{1j}|D_{1j}, \underline{x}_{1j}|A_{1j})(\underline{1})NQ_j(\underline{x}_{1j}|A_{1j}, \underline{y}_{2j}|D_{2j})(\underline{g}_j) \\ &= NQ_j(\underline{y}_{1j}|D_{1j}, \underline{y}_{2j}|D_{2j})(\underline{g}_j) \ , \end{aligned}$$

and hence $NQ_j(\underline{x}_{1j}|A_{1j}, \underline{x}_{2j}|A_{2j})(\underline{g}_j) = NQ_j(\underline{g}_j)$. *Now, by definition we have for each* j *that*

$$\begin{aligned} Q_j(\underline{x}_{1j}|A_{1j}, \underline{x}_{2j}|A_{2j})(\underline{g}_j) &= Q_j(\underline{x}_{1j}|A_{1j}, \underline{x}_{2j}|A_{2j})(\underline{1})NQ_j(\underline{x}_{1j}|A_{1j}, \underline{x}_{2j}|A_{2j})(\underline{g}_j) \\ &= Q_j(\underline{x}_{1j}|A_{1j}, \underline{x}_{2j}|A_{2j})(\underline{1})\prod_{i=1}^{k} g_{ij}^{v_{ij}(A_{1j},A_{2j})} \end{aligned}$$

for all $\underline{x}_{1j}|A_{1j}, \underline{x}_{2j}|A_{2j} \in \Omega_j$, *and the proof is complete.* □

Moreover, if for all j we fix an atom $\underline{y}_j|D_j \in \Omega_j$ and choose $\epsilon_j > 0$ such that $\epsilon_j < 1/\mu_j^*(\underline{y}_j|D_j)$, we can define for all atoms $\underline{x}_j|A_j \in \Omega_j$ the functions

$$P_j(\underline{x}_j|A_j) = Q(\underline{x}_j|A_j, \underline{y}_j|D_j)(\underline{1})P_j[\underline{y}_j|D_j, \underline{\epsilon}_j]\epsilon^{-v_j(A_j,D_j)}$$

where $v_j(A_j, D_j) = \sum_{i=1}^{k} v_{ij}(A_j, D_j)$. Then for atoms $\underline{x}_j|A_j$, we have that

$$P_j(\underline{x}_j|A_j, g_{1j}, ..., g_{kj}) = p_j(\underline{x}_j|A_j)\prod_{i=1}^{k} g_{ij}^{v_{ij}(A_j)}$$

for all $0 < g_{1j}, ..., g_{kj} < 1/\mu_j^*(\underline{x}_j|A_j)$. This implies that for all j,

$$\bar{f}_j(f_{1j}, .., f_{kj})(\underline{x}_j|A_j) = \frac{p_j(\underline{x}_j|A_j)\prod_{i=1}^{k}[f_{ij}(\underline{x}_j|A_j)]^{v_{ij}(A_j)}}{\int \cdots \int p_j \prod_{i=1}^{k} f_{ij}^{v_{ij}(A_j)} d\mu_{1j}...d\mu_{r(j),j}} \tag{A6.11}$$

for all vector of atoms $\underline{x}_j|A_j$ in Ω_j.

We are now in position to prove Theorem 6.2 .

A6.3 The Proof of Theorem 6.2.

If (Ω_j, μ_j^*) contains no atoms at all then the proof is immediate from Theorem A6.4.

If (Ω_j, μ_j^*) is purely atomic, then (6.3) is easily obtained from (A6.11) with $w_{ij}(\mathrm{A}_j) = v_{ij}(\mathrm{A}_j)$ for all $i = 1, \ldots, k$ and $j = 1, \ldots, n$. Moreover, from the fact that $\bar{f}_j$ is CEB, it is easy to verify that the weights must sum to one for each j, i.e. $\sum_{i=1}^{k} w_{ij}(\mathrm{A}_j) = 1$.

If (Ω_j, μ_j^*) has atoms but is not purely atomic, we use Theorem A6.4 to obtain the result on the set N, the complement of the atoms set of (Ω_j, μ_j^*). Consider the atoms $\underline{x}_{1j}|\mathrm{A}_{1j}, \underline{x}_{2j}|\mathrm{A}_{2j}, \ldots$ and let $P_{ij}(g_{1j}, \ldots, g_{kj})$ denote $P_{ij}(\underline{x}_{ij}|\mathrm{A}_{ij}, g_{1j}, \ldots, g_{kj})$ for all g_{ij} such that $0 < g_{ij} < 1/\mu_j^*(\underline{x}_{ij}|\mathrm{A}_{ij})$ with $i = 1, \ldots, k$ and $j = 1, \ldots, n$. From the definition of CEB for a chain graph $\mathcal{G}$ we have for each j that

$$\frac{l_j(\underline{x}_j|\mathrm{A}_j) P_j[\underline{x}_j|\mathrm{A}_j, f_{1j}(\underline{x}_j|\mathrm{A}_j), \ldots, f_{kj}(\underline{x}_j|\mathrm{A}_j)]}{P_j[\underline{x}_j|\mathrm{A}_j, h_{1j}(\underline{x}_j|\mathrm{A}_j), \ldots, h_{kj}(\underline{x}_j|\mathrm{A}_j)]} = \text{constant}\ ,\ \mu_j^* \text{ - a.e.}\ , \tag{A6.12}$$

whenever $h_{ij} \propto l_j f_{ij}$ for all i and j. From (A6.3) and for $\underline{x}_j|\mathrm{A}_j$ on N, the left-hand side of the above equation equals $\prod_{i=1}^{k} t_{ij}^{w_{ij}(\mathrm{A}_j)}$, where $t_{ij} = \int \cdots \int l_j f_{ij} d\mu_{1j} \ldots d\mu_{r(j),j}$ for all i and j.

Now, fix $t_{1j}, \ldots, t_{kj}$, pick a single atom $\underline{x}_{kj}|\mathrm{A}_{kj}$ in Ω_j, and let ϵ_j be small enough such that $0 < \epsilon_j/t_{ij} < 1/\mu_j^*(\underline{x}_{kj}|\mathrm{A}_{kj})$ for each i and j. Then, let $B_{1j} = \{\underline{x}_{kj}|\mathrm{A}_{kj}\}$, and setting $s_{ij} = \epsilon_j$ and $s_{ij}^* = \epsilon_j/t_{ij}$ for each i and j, construct the same densities and likelihoods as in the proof of Lemma A6.7 . Using the fact that (A6.12) also holds on all of Ω_j, we have

$$\frac{P_{kj}(\epsilon_j, \ldots, \epsilon_j)}{P_{kj}(\epsilon_j/t_{1j}, \ldots, \epsilon_j/t_{kj})} = \prod_{i=1}^{k} t_{ij}^{w_{ij}(\mathrm{A}_{kj})}$$

for each j and an arbitrary k. This means that for all $g_{1j}, .., g_{kj}$ between 0 and $1/\mu_j^*(\underline{x}_{kj}|\mathrm{A}_{kj})$, $P_{kj}(g_{1j}, \ldots, g_{kj}) = P_{kj}(\epsilon_j, \ldots, \epsilon_j)\epsilon_j^{-1} \prod_{i=1}^{k} g_{ij}^{w_{ij}(\mathrm{A}_{kj})}$. Let $P_j(\underline{x}_{kj}|\mathrm{A}_{kj}) = P_{kj}(\epsilon_j, \ldots, \epsilon_j)\epsilon_j^{-1}$ and (6.3) is proved.

About the weights we can say that for all j, they are constant relative to $\underline{x}_j|\mathrm{A}_j$, but can possibly vary with the fixed values A_j. Also observe that they must be nonnegative if μ_j^* is not purely atomic. This same statement can be done if (Ω_j, μ_j^*) is purely atomic but includes a countably infinite number of atoms, for which we can easily construct densities $f_{1j}, \ldots, f_{kj}$ which will make the integrals

$$\int \cdots \int p_j \prod_{i=1}^{k} f_{ij}^{w_{ij}(\mathrm{A}_j)} d\mu_{1j} \ldots d\mu_{r(j),j}$$

infinite unless all the weights are nonnegative. However, these integrals are always finite when Ω_j is finite and μ_j^* is a counting type measure and the weights can take negative values in this case. Furthermore, p_j must be essentially bounded, or else there exist f_j such that the above integrals are infinite when all the f_{ij} are equal to f_j for each j, according to Hewitt & Stromberg's (1965) Theorem 20.15 . □

CHAPTER 7

GRAPHICAL MODELS FOR THE EXPERT PROBLEM

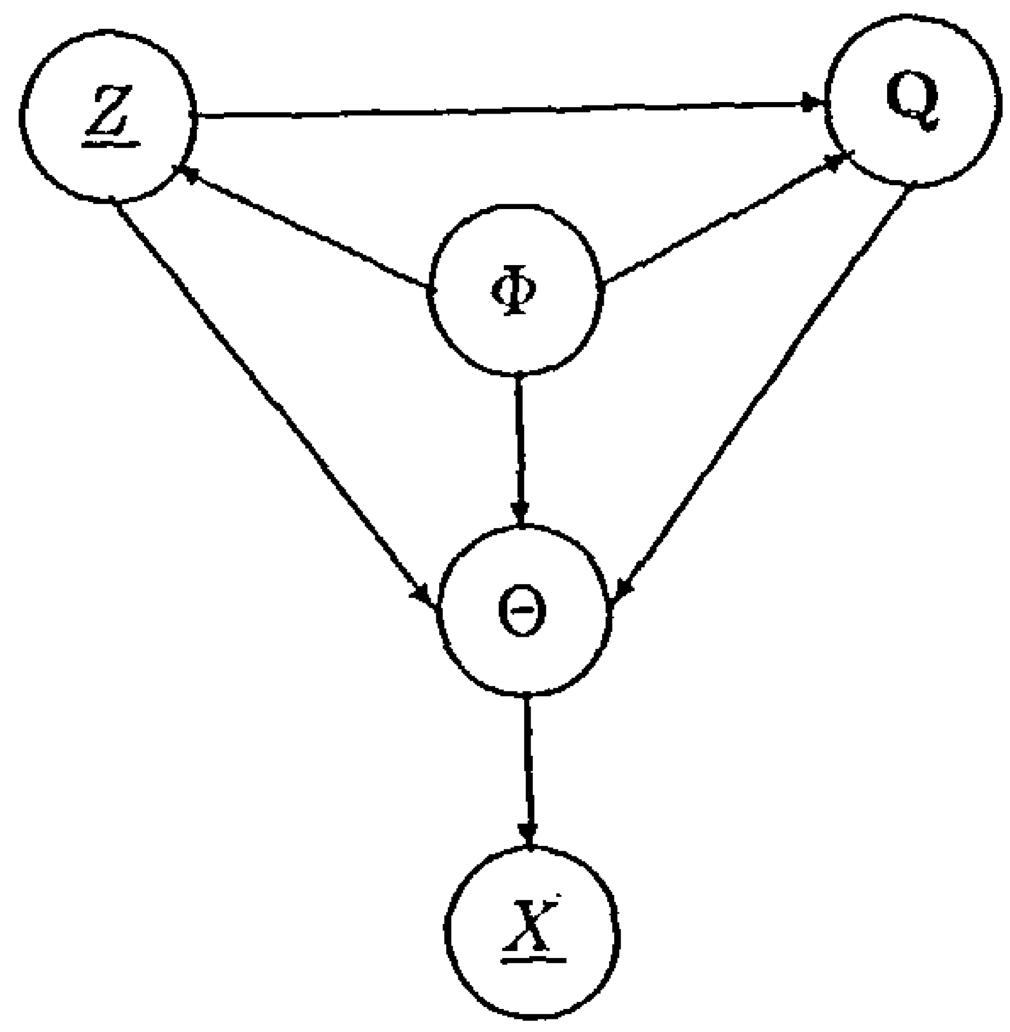
In this chapter we show a graphical representation for the expert judgement problem in general and make use of influence diagrams (IDs) to represent some special situations. In particular, we deal with degenerate situations where the experts report to the supra-Bayesian (SB) only their summaries (and not opinions) of the outcomes of experiments which are informative about the underlying parameters of interest and which are components of their information bases. The SB represents each situation according to the way the experts' information bases overlap, as well as through considerations about the sufficiency of the experts' reports for the estimation of the underlying parameters. In certain cases, if the SB cannot obtain statistics of the experts' shared information, then the reconciliation rule must account for all possible outcomes of those.

Although degenerate, the situations we describe are useful enough to give some insight on how the dependencies induced by the overlapping of information can influence on the form of the SB's reconciliation rules. Also, the graphical modelling approach is quite useful to show how rich the class of expert judgement problems is, pointing out the complexity of such problems and the potential for future research in this area.

7.1 Conditional Independence Structures for the Expert Problem.

The SB by making use of IDs to represent her view of the associations between variables is able to represent relevance statements which are usually obscured in the generally complex modelling approach of the expert judgement problem (see Smith, 1995). In fact, the SB's aim is to find relatively simple descriptions which are easier to manipulate and yet contain all the information relevant to the problem. In our context, the SB's ideal ID representation of the problem would be one which is capable of simultaneously capturing both (a) the way various sources of expert's information are related to each other, as well as (b) how the information contained in those sources become available to her (according to the problem's configuration) such that she could estimate the unknown parameters with the resources she has in the best possible manner.

The description of the expert problem made in Section 2.4.2, stands for a general multivariate set-up. In that setting, a representation of the expert problem is that shown



$\mathcal{I}$

FIGURE 7.1. The general ID $\mathcal{I}$ of a subclass of expert problems.

by the ID $\mathcal{I}$ in Figure 7.1.

Now, referring to Section 2.4.3, associated to the above ID $\mathcal{I}$, the posterior density for Θ, given $\mathbf{Q}, \underline{z}$ and Φ, has the form (2.22). However, as mentioned before, to obtain (2.22) the SB needs to engage in the difficult task of specifying a conditional joint likelihood function for $(\mathbf{Q}, \underline{z}, \Phi)$ given Θ. Although this is difficult even in one-dimensional problems, that task could be facilitated if the SB is prepared to make additional CI statements about those variables as we shall see.

The notation we use here for CI statements obtained from IDs is that introduced in Section 3.3. Within considerations of sufficiency that notation could have an alternative interpretation. For example, *if B is a parameter and both A and C are random variables then $A \amalg B|C$ could read: C is sufficient for B relative to A.*

7.1.1 A particular subclass of expert problems.

In the following sections, we shall consider a particular subclass of the general expert problem for which, using the notation of Section 2.4.2, $\underline{X} = X$ with $n = 1$, $\Theta = \Phi = \theta$ with $d(1) = d'(1) = 1$, $\underline{Z} = (Z_1, \ldots, Z_m)$ and $\mathbf{Q} = \underline{q} = (q_1, \ldots, q_k)$. In this setting, the formula corresponding to (2.22) would be

$$f(\theta|\underline{q}, \underline{z}) \propto f(\underline{q}|\underline{z}, \theta) f(\theta|\underline{z}) \ , \tag{7.1}$$

where $f(\underline{q}|\underline{z}, \theta)$ is the SB's joint likelihood function for $\underline{q}$ for all possible values of $\underline{z}, \theta$ and $f(\theta|\underline{z})$ is her density for θ posterior to observing $\underline{z}$ and prior to receiving $\underline{q}$ (and observing x).

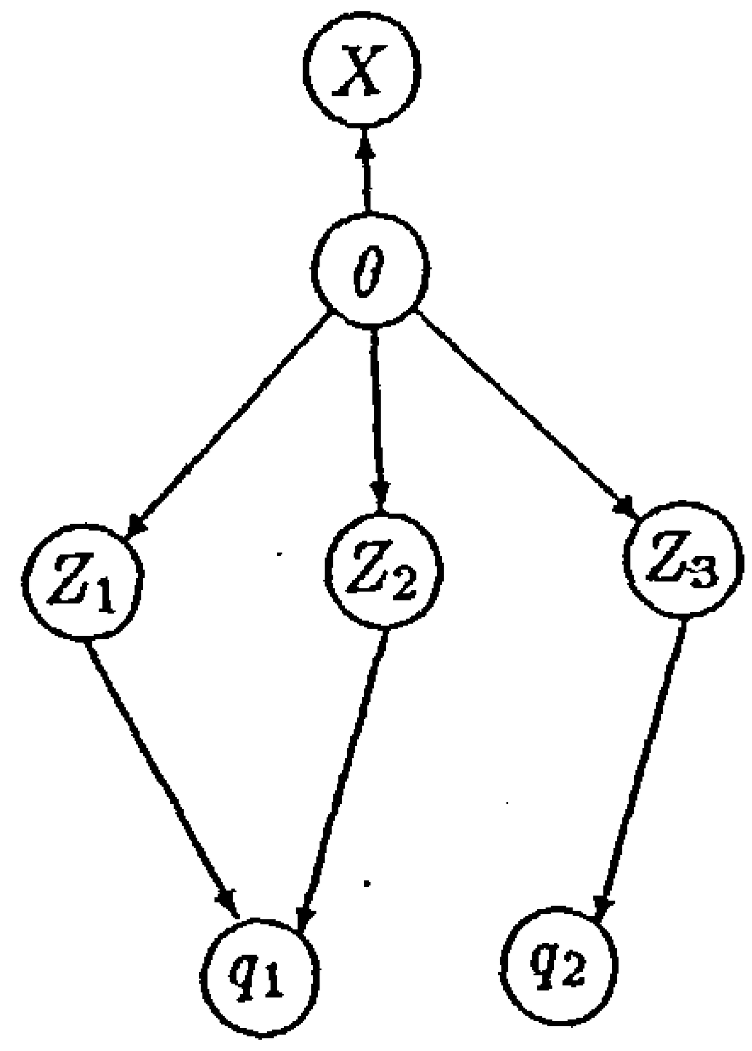
In this context, we define the subclass of problems we shall be dealing with by imposing restrictions on the form of $f(\underline{q}|\underline{z},\theta)$. We start with the definition of expert information base and introduce some of the notation as well as an example that shall be used throughout this chapter.

Definition 7.1 (Information base). *An expert's E_i information base $\underline{I}_i$ consists of the set of experiments (usually a subset of $\underline{Z}$) related to θ that E_i performs or necessarily observes, and in which he bases to make his statement q_i about θ $(i = 1, ..., k)$.*

We can now define our problem. Formally, each expert E_i is supposed to report to the SB his statement q_i about θ which is based on his information base $\underline{I}_i$, where $\underline{I}_i \subseteq \underline{Z}$ and $\underline{Z} = (Z_1, \ldots, Z_m)$ $(i = 1, ..., k)$. The statements $\underline{q} = (q_1, \ldots, q_k)$ are assessed by the experts based on the outcomes of random samples of components of $\underline{Z}$ which the SB does not observe. However, the SB knows what the information base I_i of each expert E_i is. Therefore she knows together with $\underline{I} = (\underline{I}_1, ..., \underline{I}_k)$, the set $\underline{S} = (\underline{S}_{12}, ..., \underline{S}_{k-1,k}, ..., \underline{S}_{1...k})$ of all information shared by experts, where $\underline{S}_{i...l}$ represents the subset of the information shared by experts $E_i \cdots E_l$, i.e. $\underline{S}_{i...l} = \underline{I}_i \cap \cdots \cap \underline{I}_l$. To make it clearer, consider the following example describing the problem above in the context of a coin tossing experiment:

Example 7.1 (coin tossing experiment) Suppose that the SB is interested in determining the probability θ that the outcome will be a head in a coin tossing experiment X she performs or observes. She does not know whether the coin is fair. However, she knows that the same coin has already been used in other independent experiments $\underline{Z} = (Z_1, Z_2, Z_3)$ whose outcomes are were partially observed by two experts E_1 and E_2 but not observed by the SB. Although the SB did not observe any outcome z_j of Z_j $(j = 1, 2, 3)$, she knows what the experts' information bases are, i.e. which experiment Z_j each expert has observed (and, thus, all the experiments that E_1 and E_2 share). The SB also knows that each expert's statement q_i about θ is based on the number of heads occurred in the outcomes he has observed. For instance, if E_1 has observed $\underline{I}_1 = (Z_1, Z_2)$ and E_2 has observed $\underline{I}_2 = (Z_2, Z_3)$ then $\underline{S} = (\underline{S}_{12}) = Z_2$ is the information they share, and their statements $\underline{q} = (q_1, q_2)$ are based on the number of heads occurred in their respective information bases.

A graphical representation of a coin tossing experiment similar to that of Example 7.1 but for when $\underline{S} = \emptyset$, can be seen in the ID $\mathcal{I}_1$ of Figure 7.2. From $\mathcal{I}_1$ the SB can state, by



$\mathcal{I}_1$

FIGURE 7.2. The ID $\mathcal{I}_1$ for 2 experts observing (Z_1, Z_2) and Z_3 respectively, which they use to make their statements q_1 and q_2 about θ

using the d-separation theorem (see Section 3.4), the following useful relevances :

(i) $Z_1 \perp\!\!\!\perp Z_2 \perp\!\!\!\perp Z_3 \perp\!\!\!\perp X|\theta$, representing the conditional independence of the experiments. Also, note that given q_1, Z_1 and Z_2 are no longer conditionally independent, that is $Z_1 \not\perp\!\!\!\perp Z_2|\theta, q_1$. In general for k experts and m experiments, $\perp\!\!\!\perp_{j=1}^{m} Z_j|\theta$ and $Z_j \not\perp\!\!\!\perp Z_l|q_i, \theta$ for $j \neq l$ and $\{Z_j, Z_l\} \in \underline{I}_i$, where $\underline{I}_i$ is the expert E_i information base compounded of only those experiments Z_j he observes $(j, l = 1, \ldots, m; i = 1, \ldots, k)$;

(ii) $\underline{q} \perp\!\!\!\perp \theta|\underline{z}$, meaning that if given, the experiments $\underline{z}$ would be sufficient for the SB to estimate θ relative to $\underline{q} = (q_1, q_2)$, that is, knowing $\underline{z}$, $\underline{q}$ would bring no further information for the SB regarding θ. Also note that $q_1 \perp\!\!\!\perp \theta|\underline{I}_1$ and $q_2 \perp\!\!\!\perp \theta|\underline{I}_2$, where $\underline{I}_1 = (Z_1, Z_2)$ and $\underline{I}_2 = Z_3$;

(iii) $q_1 \perp\!\!\!\perp q_2|\theta$, i.e. the experts statements are independent given θ. For k experts, $\perp\!\!\!\perp_{i=1}^{k} q_i|\theta, \varphi$; and

(iv) $\underline{q} \perp\!\!\!\perp X|\theta$.

The consideration of sufficiency described in item (ii) above is a bit awkward in the context of expert opinion assessment, since usually the experts consider factors, such as intuition and experience, other than the observation of objective experiments alone in order to assess their opinions. Certainly, in this setting the experts opinions would bring further information to the SB other than that just contained in the experiments they observe. This is also supposedly valid when considering the choice of experts. However, a tightened concept of sufficiency can be used in this regard (see Section 7.2). Also, in the degenerate situation we shall examine, it is not unreasonable at all to consider the sufficiency of experts' statements.

Taking the above CI statements for general k and m, and from the fact that the SB does not observe $\underline{Z}$, her posterior density for θ given $\underline{q}$ when the experts do not share any experiment would be the simplified version of (7.1) :

$$f(\theta|\underline{q}) \propto f(\underline{q}|\theta)f(\theta) = f(\theta)\prod_{i=1}^{k} f_i(q_i|\theta) \ , \tag{7.2}$$

where $f_i(q_i|\theta)$ is the SB's likelihood function for the expert E_i $(i = 1, 2)$.

Note that even in the above simpler case, restrictions on the form of $f_i(q_i|\theta)$ are required if $\underline{q}$ corresponds to a vector of probability densities as commented in Section 2.4.3. Obviously, this includes the commonly considered case when the only extraneous information available to the SB are the experts' probability statements.

7.2 Common Knowledge, Value of Information and Sufficiency.

An issue worth pointing out here is the concept of *common knowledge* (CK) (Aumann, 1976) which was originally developed in the context of game theory and the economics of information, for the study of the interactions between people concerning their exchange of information and opinion. In the group decision problem, where the whole group of experts is responsible for taking decisions, it is appropriate to consider the events of interest as being CK to all members of the group. In fact, EB groups must have a common likelihood function for their agreed CI structure for the events they observe (Chapter 6). On the other hand, in the expert judgement problem where interactions between experts are in general coarser than in the group problem, CK would generally be a rather strong assumption. Usually, the SB consults experts who may share some information but who do not necessarily interact and exchange opinions. Nevertheless, there are certain situations in which the expert problem can be seen as a Bayesian cooperative game where CK can occur at a certain level. In cases where the SB knows that some experts have the same priors and that their posteriors for certain events are CK, then according to Aumman's (1976) theorem, those posteriors must be equal and the SB may well consider discarding experts opinions from the reconciliation process.

As far as the choice of experts is concerned, the criterion of *value of information* (Raiffa and Schlaifer, 1961), where the increase in utility which would result from learning an expert's opinion is computed, will not be applied here. According to Clemen and Winkler (1985), when costs of consulting experts are not considered, "consulting an extra expert can never be detrimental in a value-of-information sense, regardless of whether the experts are dependent or independent."

When the criteria for choosing (or discarding) experts is solely based upon sufficiency, DeGroot and Fienberg (1983) found out that information could be lost. That is, even if an expert A is sufficient for another, B, in the sense that "from A's prediction together with a simple auxiliary randomisation, one can simulate a prediction with the same stochastic properties as B's prediction", it is possible that by learning the prediction of B in addition to the prediction of A, more information might be gained than from the prediction of A alone. This could be the case if for example A is known to be uncalibrated while B is well-calibrated. We shall see in Section 7.6 a situation where this is the case even for 'perfectly calibrated' experts.

Also, as mentioned before in Section 7.1.1, another type of sufficiency is that of the experts' opinions relative to their information bases for estimating θ. We shall consider here that for the SB, an expert's statement $q_i(\theta|\underline{I}_i)$ is sufficient for θ relative to his information base $\underline{I}_i$ $(i = 1, \ldots, k)$ if no additional information would be given were $\underline{I}_i$ observed by her directly. However, the fact that each q_i is sufficient for θ relative to $\underline{I}_i$ $(i = 1, \ldots, k)$, does not guarantee that $\underline{q}$ is sufficient for θ relative to $\underline{Z}$ in all cases as we shall see. This is closely related to the issue of non-independence among experts.

Certainly it is easier to process information if the SB knows that expert judgements are independent. A question that arises here and which will also be dealt with in this chapter, is when it is possible to discard those experts who introduce dependence and whose statements do not bring any extra information for the SB in obtaining statistics which are sufficient regarding all the experiments the experts have observed.

7.3 Degenerate Situations in the Expert Problem.

We shall assume here, in the context of the problem modelled by Winkler (1981) —see Section 2.4.4— and as also described in Clemen (1987), that the SB is supposed to only know

(i) what the experts information bases are, and thus all the information they share, but she does *not* actually observe any information on those bases; and

(ii) that the experts statements $\underline{q}$ are summary statistics obtained from their information bases and not their subjective opinions.

Note that because the experts' statements $\underline{q}$ are objective summaries rather than probability statements per se, Winkler's problem is degenerate in the context of the Bayesian paradigm. Despite that, it is very useful to the understanding of the role that dependence of expert information bases and sufficiency of expert statements play in the expert problem to start at this point.

Formally, each expert E_i $(i = 1, ..., k)$ is supposed to report to the SB his statistic q_i which is a summary of his information base $\underline{I}_i$. Each statistic $q_i = h_i(\underline{I}_i)$, where h_i is a function of its arguments only $(i = 1, ..., k)$, is obtained by E_i from $\underline{I}_i$.

The dependence between experts information bases can be modelled for the above described problem, in three basic distinct situations which are characterised by the way the experts share information. Those are when the experts have :

(1) non-overlapping information bases, with each expert observing his own independent set of experiments (e.g. coin tosses) when there is no experiment being shared by the experts, i.e. $\underline{S}_{i...l} = \emptyset$, for all $i \neq l$ ($i, l = 1, ..., k$) ;

(2) overlapping structure with experts sharing experiments although no expert's information base is subsumed by any other, i.e. $\underline{S}$ is non-empty and for any two experts E_i and E_l, neither $\underline{I}_i \not\subseteq \underline{I}_l$ nor $\underline{I}_l \not\subseteq \underline{I}_i$ for $i \neq l$ $(i, l = 1, ..., k)$, and finally ;

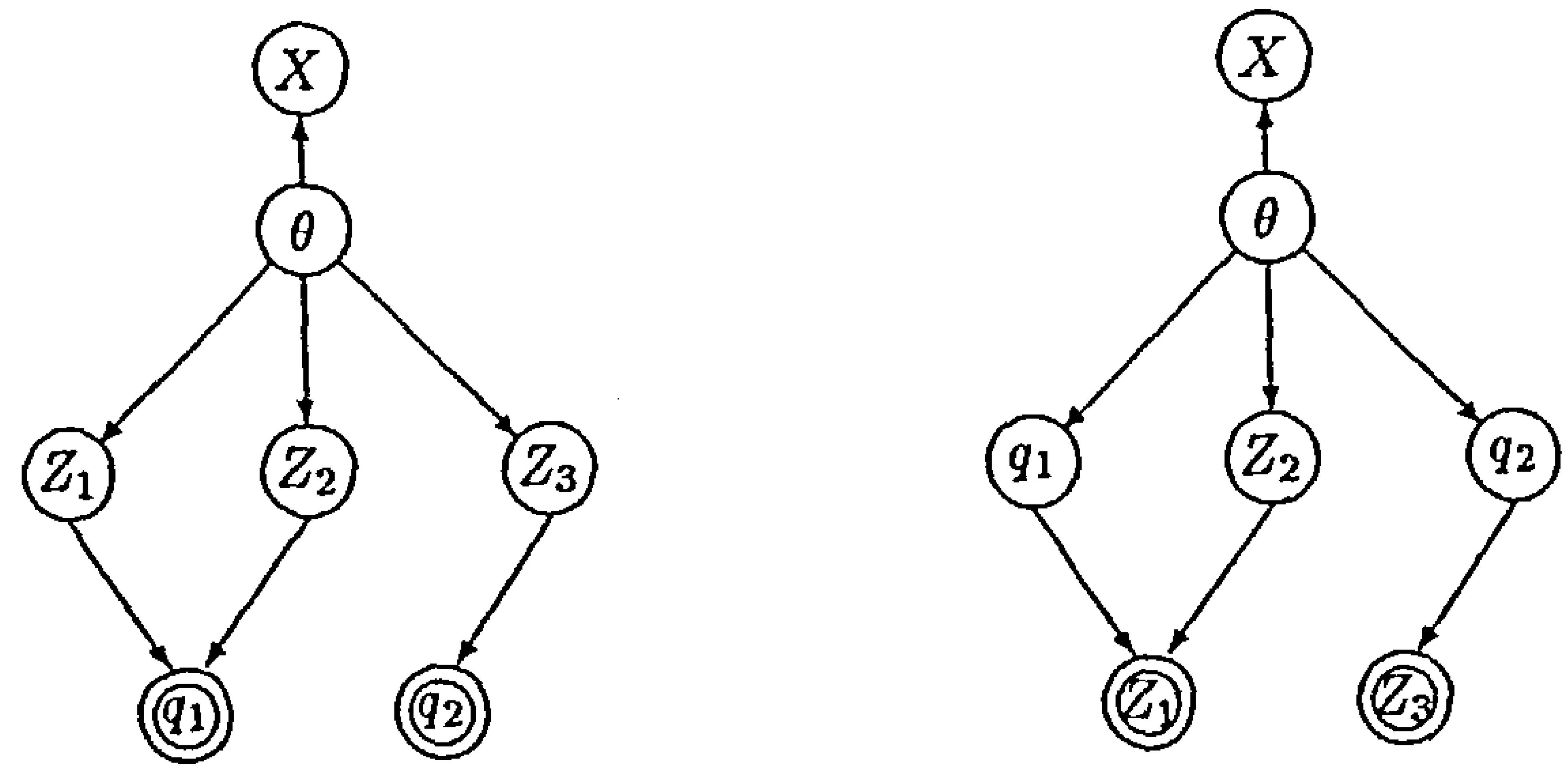
(3) overlapping with dependent information bases, where an expert's base is subsumed by other, i.e. there is at least one expert E_l whose experiments are also observed by another expert E_i such that $\underline{S}_{il} = \underline{I}_i \cap \underline{I}_l = \underline{I}_l$ implying that $\underline{I}_l \subseteq \underline{I}_i$ for $i \neq l$.

In the following Sections 7.4, 7.5 and 7.6, we make use of Example 7.1 to illustrate the modelling of the above cases (1), (2) and (3) respectively. Particular types of reconciliation rules are obtained in each case.

7.4 Non-overlapping Information Bases.

Suppose that in Example 7.1, the experts E_1 and E_2 do not perform or observe any experiment together, that is, they share no information (i.e. $\underline{S}_{12} = \emptyset$). If their information bases are $\underline{I}_1 = (Z_1, Z_2)$ and $\underline{I}_2 = Z_3$, respectively, then the SB's ID representing this problem could be $\mathcal{I}_2$ in Figure 7.3(a). Note that apart from the double-circled nodes used to represent deterministic variables, the ID $\mathcal{I}_2$ in Figure 7.3(a) is the same as $\mathcal{I}_1$ in Figure 7.2. Because of that, the CI statements (i)–(iv) in Section 7.1.1 are also valid here. The ID

$\mathcal{I}_3$ in Figure 7.3(b) is equivalent to $\mathcal{I}_2$ since we can rewrite $q_1 = Z_1 + Z_2$ as $Z_1 = q_1 - Z_2$. Thus, we can also say that once q_1, q_2 and Z_2 are given, Z_1 and Z_3 bring no further information about θ.



(a) $\mathcal{I}_2$ (b) $\mathcal{I}_3$

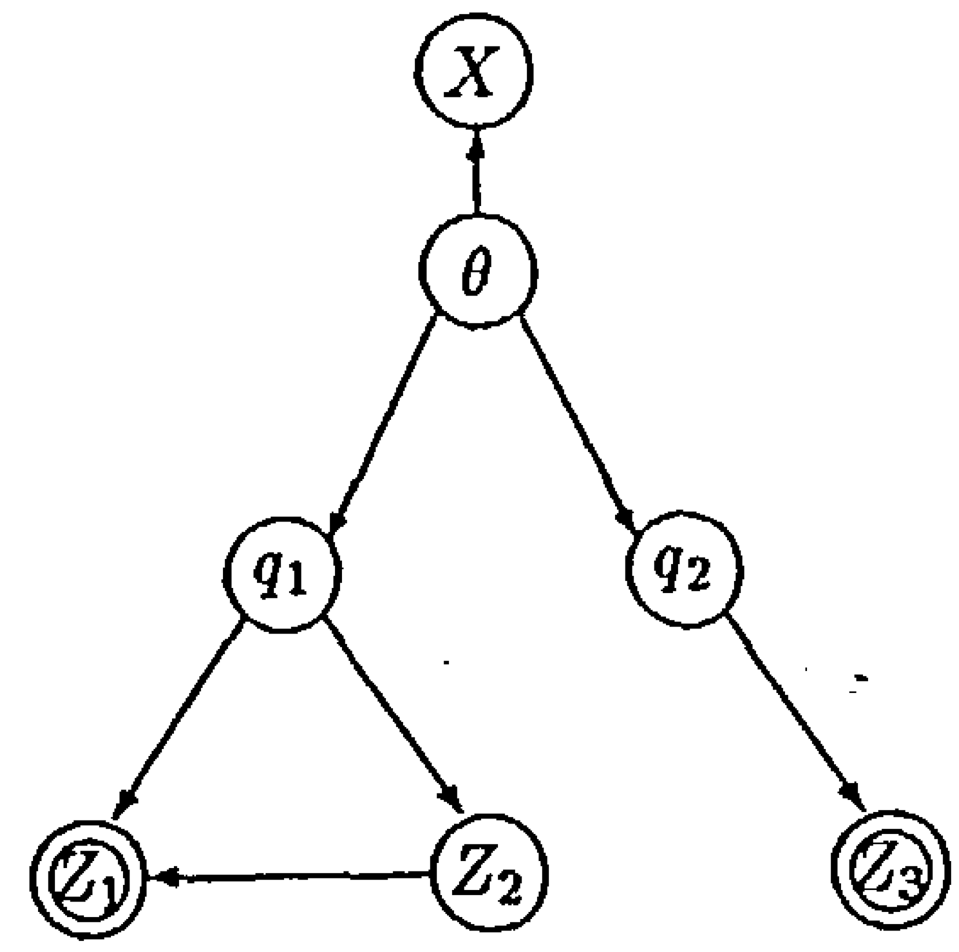
FIGURE 7.3. Two IDs representing the same problem for 2 experts' non-overlapping information bases when $q_1 = Z_1 + Z_2$ and $q_2 = Z_3$

Observe that from the SB's point of view, $Z_j \not\perp\!\!\!\perp Z_l | q_i, \theta$ for $\{Z_j, Z_l\} \in \underline{I}_i$, $j \neq l$ $(j, l = 1, \ldots, m; i = 1, \ldots, k)$, i.e. the fact that an expert E_i observed two or more experiments and is reporting a summary of the outcomes, introduces a conditioning factor amongst the components of his information base $\underline{I}_i$. For example, if in a coin tossing experiment, it is given that $q_1 = (Z_1 + Z_2) = 1$ then $Z_1 = 1 - Z_2$.

Therefore, the SB's posterior distribution for θ given $\underline{q}$, $f(\theta|\underline{q})$, is obtained by formula (7.2) in which the SB's task of assessing $f(\underline{q}|\theta)$ is rather simplified by that factorisation.

Also, regarding the example above, since $\underline{S}_{12} = \emptyset$, if $T(\underline{Z}) = Z_1 + Z_2 + Z_3$ (the total number of heads occurred in $\underline{Z}$) is a sufficient statistic for θ with respect to $\underline{Z} = \underline{I}_1 \cup \underline{I}_2$, then $t(\underline{q}) = q_1 + q_2$ is also sufficient for θ. This is because $t(\underline{q}) = U(\underline{Z})$ where $\underline{q}$ is a one-to-one function. In terms of an ID representing this sufficiency we would have $\mathcal{I}_4$ of Figure 7.4 below, where $\underline{q}$ separates θ from $\underline{Z}$.

It is worthwhile pointing out here that *the fact that for some expert E_i, q_i is sufficient for his information base $\underline{I}_i \subseteq \underline{Z}$ $(i = 1, \ldots, k)$, where $\cup_{i=1}^{k} \underline{I}_i = \underline{Z}$ does not necessarily imply, in general, that a function of the vector $\underline{q}$ of all experts statements is for the SB, sufficient for $\underline{Z}$ concerning θ* as we shall see. In the next section, conditions are derived in order to allow a function t (not necessarily a function of $\underline{q}$ only) to be such a sufficient



$\mathcal{I}_4$

FIGURE 7.4. ID $\mathcal{I}_4$ representing the sufficiency of the experts' statements q_1, q_2 for θ relative to (Z_1, Z_2, Z_3) when there is no overlap of information.

statistic.

Now, examples in which the SB's posterior density function is obtained for the coin tossing Example 7.1 and for the normal case are shown.

Example 7.2 (Bernoulli model). Consider that in Example 7.1, $\underline{I}_1 = (Z_1, Z_2)$ and $\underline{I}_2 = Z_3$ as the case in this section. Also, each of the binary variables X, Z_1, Z_2 and Z_3, is defined as 1 if a head occur and 0 otherwise, in its respective coin toss, have each a Bernoulli density with parameter $\theta = Pr[head]$. Thus, the SB's joint distribution for the experts' statements $q_1 = (z_1+z_2)$ and $q_2 = z_3$ conditioned on θ is binomial with parameters $(3, \theta)$. Note that an usual sufficient statistic with respect to $\underline{Z}$ is $U(\underline{Z}) = (Z_1 + Z_2 + Z_3)$ and that after receiving q_1 and q_2, the SB can obtain $t(\underline{q}) = (q_1 + q_2) = U(\underline{z})$. In fact, from the relation $q_1 \amalg q_2 | \theta$,

$$f(\underline{q}|\theta) = f(q_1|\theta) f(q_2|\theta) \propto \theta^{(q_1+q_2)}(1-\theta)^{[3-(q_1+q_2)]} \ .$$

If the SB chooses a beta (α, β) as a prior for θ, then, according to the usual conjugate analysis, the posterior density is also a beta but with parameters (α^*, β^*) where $\alpha^* = \alpha + (q_1 + q_2)$ and $\beta^* = \beta + 3 - (q_1 + q_2)$.

In a general setting with k experts, where q_i given θ would have a binomial density with parameters (n_i, θ) with n_i being the number of experiments (coin tosses) E_i observed, the posterior density for θ given $\underline{q}$ would also be a beta density with parameters

$$\alpha^* = \alpha + \sum_{i=1}^{k} q_i \tag{7.3}$$

and

$$\beta^* = \beta + \sum_{i=1}^{k} n_i - \sum_{i=1}^{k} q_i \ . \tag{7.4}$$

If the SB had taken the LogOp pool of her posterior densities $f_1(\theta|q_1)$ and $f_2(\theta|q_2)$ for θ given the statements of E_1 and E_2, which are beta densities with parameters ($\alpha_1 = \alpha + q_1, \beta_1 = \beta + n_1 - q_1$) and ($\alpha_2 = \alpha + q_2, \beta_2 = \beta + n_2 - q_2$) respectively (where α and β are the parameters of the SB's prior for θ), the combination would give

$$T(f_1, f_2)(\theta|\underline{q}) \propto f_1(\theta|q_1)^{w_1} f_2(\theta|q_2)^{w_2} \quad ,$$

where T is a pooling operator and w_1 and w_2 are weights such that $w_1 + w_2 = 1$, would be a beta density with parameters (α^*, β^*) where

$$\alpha^* = \alpha + w_1 q_1 + w_2 q_2$$

and

$$\beta^* = \beta + (w_1 n_1 + w_2 n_2) - (w_1 q_1 + w_2 q_2) \ .$$

For k experts,

$$\alpha^* = \alpha + \sum_{i=1}^{k} w_i q_i \tag{7.5}$$

and

$$\beta^* = \beta + \sum_{i=1}^{k} w_i n_i - \sum_{i=1}^{k} w_i q_i \ . \tag{7.6}$$

Note that (7.5) and (7.6) basically differ of (7.3) and (7.4) respectively in that implicitly in the later the weights are all unity and sum to k instead of one as in the latter.

Although, rigorously the LogOp should be employed in the context of the group decision problem when a group of experts agrees to be EB and there is no individual decision maker, the use of (7.5) and (7.6) by the SB would for example allow the inclusion of her confidence on the experts' statements and expertises in her aggregation rule.

Now consider the case, similar to the one mentioned in Section 6.5, in which the SB believes that one of the experts' statements is the true probability of heads but she does not know the identity of that expert. Thus, The SB will use a combination rule which is a LinOP. In this case, the 'weight w_i' assigned to the expert E_i ($i = 1, \ldots, k$) would represent the SB's probability that E_i's statement is the one based on the largest data set. Clearly if the SB knows that the expert E_i^* observes a set of experiments which contains all the other experts' experiments (see Section 7.6) then the SB need only employ this expert's statement q_i^* which would be sufficient for the SB by the usual statistical arguments.

The LinOp for k experts,

$$T(f_1,...,f_k) \propto \sum_{i=1}^{k} w_i f_\beta(\theta|q_i) \ , \tag{7.7}$$

where $\sum_{i=1}^{k} w_i = 1$, would be a mixture of beta densities since $f_\beta(\theta|q_i)$ is a beta density function with parameters $(\alpha + q_i, \beta + n_i - q_i)$. Note that the updating rule for $\underline{w}$ is now automatic if the SB has her own data on which to check the different experts' predictions and $\underline{w}$ is invariant under different experiments.

The other example for the normal case is introduced next :

Example 7.3 (Normal model). Suppose that the SB and $k = 2$ experts know that the variable of interest X is normally distributed with unknown mean θ but with known precision τ_x, i.e. $X \sim N[\theta, \tau_x]$, and each of the variables Z_j, $j = 1,2,3$, informative about θ, are all normally distributed with means $\mu_j(\theta)$, which are known linear functions of the unknown θ, i.e. $\mu_j(\theta) = a_j\theta + b_j$, where a_j and b_j are fixed, and known precisions τ_j $(j = 1,2,3)$, i.e. $Z_j \sim \mathrm{N}[\mu_j(\theta), \tau_j]$ for each $j = 1,2,3$. Assume two experts E_1 and E_2 observe $\underline{I}_1 = \{Z_1, Z_2\}$ and $\underline{I}_2 = Z_3$ reporting $q_1 = \frac{1}{2}(z_1 + z_2)$ and $q_2 = z_3$, respectively to the SB (who does not observe $\underline{Z}$).

Assume that before hearing the experts' reports, the SB has a normal prior density for θ with parameters (m_0, τ_0). According to Winkler's (1981) model described in Section 2.4.4, the likelihood function in (7.2) $f(\underline{q}|\theta)$ is Gaussian with mean $\underline{\mu}_q(\theta) = (\mu_{q_1}(\theta), \mu_{q_2}(\theta))'$, where $\mu_{q_1}(\theta) = \frac{1}{2}[\mu_1(\theta) + \mu_2(\theta)]$ and $\mu_{q_2}(\theta) = \mu_3(\theta)$, and precision matrix

$$\mathrm{T} = \begin{pmatrix} \frac{4\tau_1\tau_2}{\tau_1+\tau_2} & 0 \\ 0 & \tau_3 \end{pmatrix} .$$

Notice that T is known just because the SB knows (i) $\underline{\tau}_z = (\tau_1, \tau_2, \tau_3)'$, and (ii) which variables Z each expert observed to report his summary of the form above (i.e. as the arithmetic average of the observed data).

It is easy to see that $f(\theta|\underline{q})$ is normally distributed with parameters (μ^*, T^*) where according to (2.25) and (2.26),

$$\mu^* = \sigma^{*2}\underline{1}'\Sigma^{-1}\underline{q} = \frac{1}{\mathrm{T}^*}\underline{1}'\mathrm{T}\underline{q} \tag{7.8}$$

and

$$\mathrm{T}^* = \underline{1}'\mathrm{T}\underline{1} \tag{7.9}$$

with, in the k-variate setup, $\underline{1} = (1, ..., 1)'$ and $\underline{q} = (q_1, ..., q_k)'$ being k-dimensional.

This leads to

$$\mu^* = \left[\frac{4\tau_1\tau_2}{(4\tau_1\tau_2 + \tau_1\tau_3 + \tau_2\tau_3)}\right] q_1 + \left[\frac{\tau_3(\tau_1 + \tau_2)}{(4\tau_1\tau_2 + \tau_1\tau_3 + \tau_2\tau_3)}\right] q_2 \tag{7.10}$$

and

$$\tau^* = \frac{4\tau_1\tau_2 + \tau_1\tau_3 + \tau_2\tau_3}{\tau_1 + \tau_2} \tag{7.11}$$

in our example. If $\tau_j = \tau$ for $j = 1, 2, 3$ then $\mu^* = t = U = \frac{1}{3}(2q_1 + q_2)$ and $\tau^* = 3\tau$.

If the SB does not know the I_i's and the S_{ij}'s ($i \neq j$; $i, j = 1, ..., k$) and consequently is not able to assess T, then she might be able to make use of subjective judgement and past data to determine a prior density for T. For instance, assuming a diffuse prior distribution for θ and that $\theta \amalg \mathrm{T}$ (note that this seems reasonable since T does not relate to the process generating θ but to the expected deviations and partial correlations of the experts' opinions) we have that,

$$f(\theta, \mathrm{T}|\underline{q}) \propto f(\underline{q}|\mathrm{T}, \theta) f(\mathrm{T}) \; , \tag{7.12}$$

where, in the natural conjugate analysis, $f(\mathrm{T})$ is the SB's Wishart prior density for T with $\delta_0 + k$ degrees of freedom and with symmetric positive definite precision matrix T_0. The density $f(\underline{q}|\mathrm{T}, \theta)$ is the SB's multivariate normal prior density for $\underline{q}|\mathrm{T}, \theta$ with mean $\underline{\mu}_{\underline{q}}(\theta)$ and precision matrix T. Thus, according to Winkler's (1981) results described in Section 2.4.4, the posterior density for $\theta|\underline{q}$ is a Student-t as in (2.27), with $\delta^* = \delta_0 + k - 1$ degrees of freedom. Replacing Σ_0^{-1} by T_0 and $\underline{\mu}$ by $\underline{q}$ in formulas (2.28) and (2.29), we have the mean and s^{*2} in the precision $(\delta^* - 2)/\delta^* s^{*2}$, respectively Also, as in Section 2.4.4, if $f(\theta)$ is a normal or a t distribution, then $f(\theta|\underline{q})$ is a Poly-t distribution.

Note that the above approach for unknown T differs from the approach that should be employed in the case of considering a second unknown parameter, i.e. $\underline{\theta}_1 = (\theta, \tau_x)$ when $d(1) = 2$.

The configuration with non-overlapping information bases described above can be seen as an *orthogonal design* experiment (see e.g. Geramita and Seberry, 1979, and Montgomery, 1991) with the experts' statements $\underline{q}$, treated as data by the SB, forming a separable likelihood function (Smith, 1990). Winkler (1981), Clemen (1987) and Clemen and Winkler (1985) among others, have shown this design to provide the smallest posterior variance for θ, being, thus, optimal for incorporating relevant information into the SB's prior distribution for θ. No available information is lost in the process due to overlap.

7.5 Overlapping Information Bases.

Now, assume that the experts are allowed to share information without however one completely encompassing any other expert's information base, i.e. there is at least one non-empty set $\underline{S}_{ij}$, $i \neq j$ $(i, j = 1, ..., k)$, such that $\underline{I}_i \not\subseteq \underline{I}_j$, for all $i \neq j$ $(i, j = 1, ..., k)$.

In this setting, the problem described in Example 7.1 can be represented by the ID $\mathcal{I}_5$ of Figure 7.5(a). Also, because $Z_1 = q_1 - Z_2$ and $Z_3 = q_2 - Z_2$, we can draw an equivalent ID $\mathcal{I}_6$ of Figure 7.5(b).

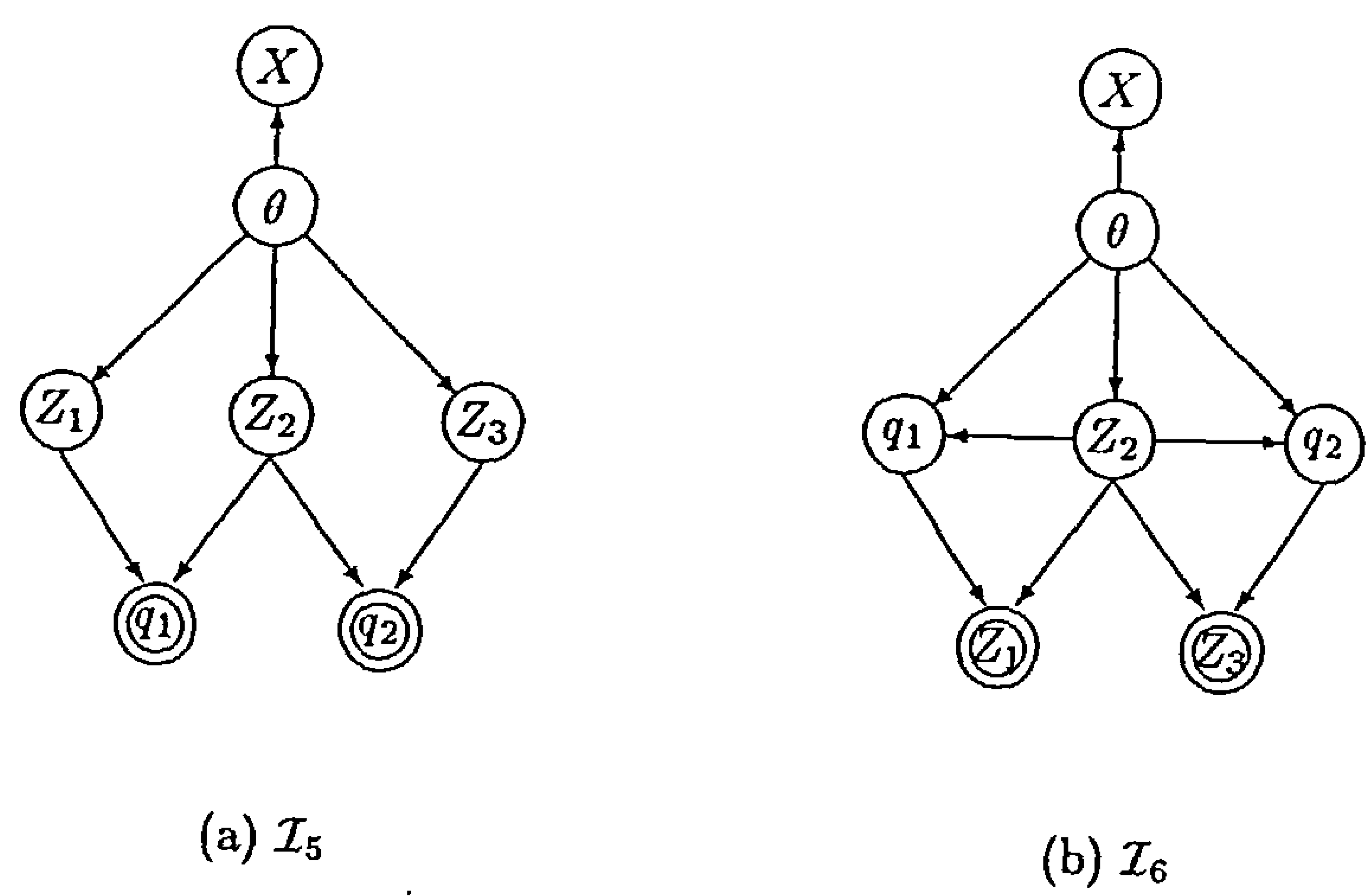


FIGURE 7.5. Two IDs representing the same problem for 2 experts' overlapping information bases when $q_1 = Z_1 + Z_2$ and $q_2 = Z_2 + Z_3$

From $\mathcal{I}_5$, we can make the same CI statements as in (i)–(iv) in Section 7.1.1 except (iii). This is because now, $q_1 \not\perp\!\!\!\perp q_2|\theta$, since there is a path from q_1 to q_2 $(q_1—Z_1—Z_2—q_2)$ in the moral graph of the smallest ancestral set of q_1, q_2 and θ obtained from $\mathcal{I}_5$, which is not being separated by θ. In fact, this statement is in accordance with the fact that both E_1 and E_2 are now sharing a common experiment Z_2 and therefore their statements are not independent anymore conditionally on θ. However, note that if Z_2 is also given together with θ, then $q_1 \perp\!\!\!\perp q_2|\theta, Z_2$. This suggests that if the SB could somehow be able to obtain a sufficient statistic for Z_2, then this would correspond to the introduction of a new variable in the system which could then separate the experts opinions as we shall see below (Figure 7.6).

Note that for general k and m, since $\underline{S}$ is non-empty, $U(\underline{Z})$, the SB's desired sufficient statistics for θ, cannot be written as a function of $\underline{q}$ only ($\underline{q}$ is not one-to-one anymore). In many cases this may be implicit, since different experts may have access to certain information bases unavailable to and unknown by the SB. It is easy to see that although each q_i is sufficient for θ with respect to $\underline{I}_i$, $t(\underline{q})$, a function of $\underline{q}$ only, is *not* sufficient for θ with respect to $\underline{Z}$, unless $k = n$, i.e. the number of experts assessments q_i equals the number of variables $Z_j \subset \underline{Z}$, when the SB could deduce every Z_j $(j = 1, ..., n)$ from $\underline{q}$. In fact, summary statistics will be insufficient unless they enable the identification of the observations that overlap. Therefore, in the situations where the SB can obtain summary statistics $\underline{Y}(\underline{S})$ of the shared data, then this would enable her to obtain a statistic $u(\underline{q}, \underline{y})$ which is sufficient for θ relative to $\underline{Z}$.

7.5.1 When the shared information is given.

For instance, assume that for $m = 5$ and $k = 3$ such that $\underline{I}_1 = (Z_1, Z_2)$, $\underline{I}_2 = (Z_2, Z_3)$ and $\underline{I}_3 = (Z_3, Z_4, Z_5)$, thus with the experts sharing the experiments $\underline{S} = (\underline{S}_{12}, \underline{S}_{23})$ where $\underline{S}_{12} = Z_2$ and $\underline{S}_{23} = (Z_3, Z_4)$. Notice that $\underline{I}_i \not\subseteq \underline{I}_j$, $i \neq j$ $(i, j = 1, 2, 3)$. Now, let $\underline{Y}(Y_{12}, Y_{23}) = (Z_2, Z_3 + Z_4)$ be the SB's summary statistic of $\underline{S}$ obtained from some information source, then the SB can construct the statistic $u(\underline{q}, \underline{y}) = q_1 + q_2 + q_3 - y_{12} - y_{23}$, where $q_1 = Z_1 + Z_2$, $q_2 = Z_2 + Z_3 + Z_4$ and $q_3 = Z_3 + Z_4 + Z_5$, which is sufficient for $\underline{Z}$.

In this context, assuming that the SB is able to obtain $\underline{Y}$, the situation above can be represented by the ID $\mathcal{I}_7$ of Figure 7.6, from which the following interesting relevances can be stated by using the d-separation theorem:

(i) $q_1 \amalg q_2 \amalg q_3 | \theta, \underline{Y}$, i.e. the experts' summaries are independent given θ and $\underline{Y}$. In general, $\amalg_{i=1}^{k} q_i | \theta, \underline{Y}$;

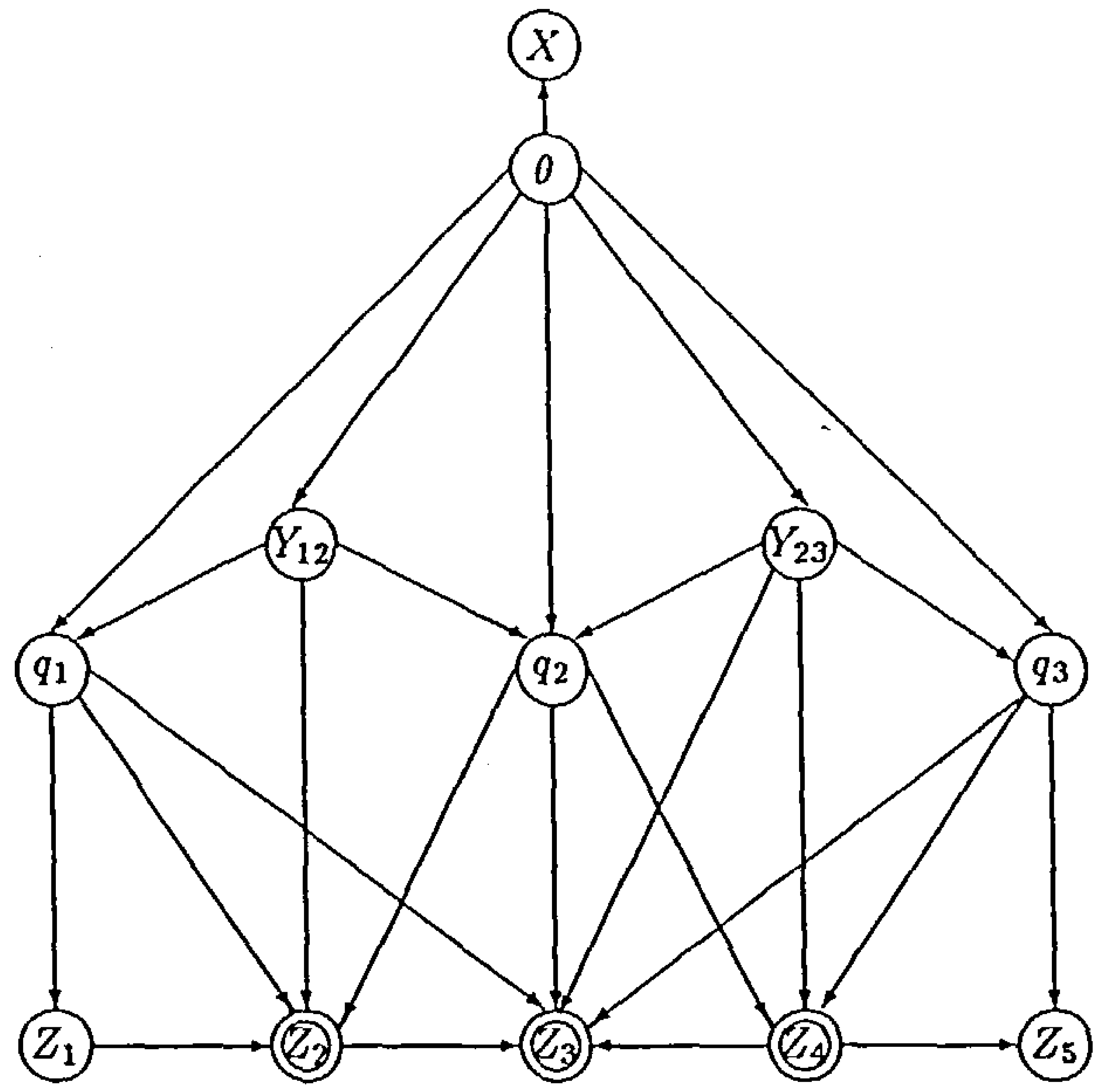
(ii) $Y_{12} \amalg Y_{23} | \theta$ for the conditional independence of the statistics of shared information. If we let $L = \{12, \ldots, 1k, \ldots, 123, \ldots, 1 \cdots k\}$ be the set of indexes of components of $\underline{Y}$, then in general $\amalg_{j \in L} Y_j | \theta$; and

(iii) $\underline{Z} \amalg \theta | \underline{q}, \underline{Y}$, meaning that $\underline{q}$ together with $\underline{Y}$ are sufficient for θ relative to $\underline{Z}$.

Now, making use of the above relations (i) and (ii) and applying Bayes theorem, we have that the SB's posterior density for θ is

$$f(\theta | \underline{q}, \underline{y}) = c(\underline{q}, \underline{y}) f(\theta) \prod_{j \in J} f(y_j | \theta) \prod_{i=1}^{k} f_i(q_i | \theta, \underline{y}) \ , \tag{7.15}$$

where $c(\underline{q}, \underline{y}) = [f(\underline{q}, \underline{y})]^{-1}$ is constant relative to θ.



$\mathcal{I}_7$

FIGURE 7.6. ID $\mathcal{I}_7$ of overlapping experts' information bases representing their statements q_1, q_2 and q_3 which together with the SB's access to Y_{12} and Y_{23}, the summaries of the shared information, are sufficient for θ relative to $(Z_1, Z_2, Z_3, Z_4, Z_5)$

Formula (7.15) most naturally compares with (7.2) when each y_j $(j \in J)$ is taken as an extra expert's honest statement. Also, $f_i(q_i|\theta, \underline{y})$ has a deterministic part for certain values of $\underline{y}$. For instance, if we consider the Example 7.1 with $\underline{Y} = Y = Z_2$ then for $y = 1$, the SB would expect either $q_i = 1$ or $q_i = 2$ with higher probability if she consider the experts to be honest. On the other hand, if the SB believes some experts to be inaccurate in their statements then she could make use of $f_i(q_i|\theta, \underline{y})$ to assess this.

7.5.2 Accommodating the shared information.

Nevertheless, if the SB cannot obtain the outcomes of shared information by redesigning the experiment or consulting other sources of information, then one possible approach is to combine all the densities (or probabilities) for all possible outcomes of the shared information. This can be done through the introduction of the auxiliary variables $\underline{S}^* = (S_1^*, \ldots, S_k^*, S_{12}^*, \ldots, S_{1k}^*, S_{1kl}^*, \ldots, S_{1\ldots k}^*)$, where $S_{i\ldots l}^*$ is the set of experiments observed *exactly* by the experts $E_{i\ldots l} = E_i \cap \cdots \cap E_l$ $(i \neq l)$. Also, call $J = \{1, \ldots, k, 12, \ldots, 1k, \ldots, 1 \cdots k\}$ the set of indexes of components of $\underline{S}^*$. Those variables which separate $\underline{I} = \cup_{i=1}^k \underline{I}_i$ into disjoint sets and are conditionally independent given θ, can be integrated (or summed) out of $f(\underline{q},\theta,\underline{s}^*)$ to give the posterior

$$f(\underline{q}|\theta) = \int f(\underline{q}|\theta,\underline{s}^*) \prod_{j\in J} f(s_J^*|\theta) d\underline{s}^* \ , \tag{7.16}$$

where the integration sign stands for multiple integration over the domain of $\underline{S}^*$. Note that integration methods as Gibbs sampling can be used here to obtain $f(\theta|\underline{q})$.

This is the approach adopted by Clemen (1987) who shows that in the Bernoulli case, the discrete version of formula (7.16) with the summation replacing the integration, gives

$$f(\underline{q}|\theta) \propto \sum_{s_1^*=0}^{n_1^*} \cdots \sum_{s_{1\ldots k}^*=0}^{n_{1\ldots k}^*} f(\underline{q}|\theta,\underline{s}^*) f(\underline{s}^*|\theta) \ , \tag{7.17}$$

where $n_{i\ldots l}^*$ is the number of observations seen *only* by experts $E_{i\ldots l}$. After observing that $f(\underline{q}|\theta,\underline{s}^*)$ equals one if the $s_{j\ldots l}^*$'s add up correctly to each q_i, and zero otherwise, and that $f(\underline{s}^*|\theta)$ is a joint probability of non-overlapping independent binomial samples, he obtains a mixture of beta distributions when the SB has an improper diffuse prior for θ, viz.

$$f(\theta|\underline{q}) = \sum_{\alpha=0}^{m} v_\alpha f_\beta(\theta|\alpha, m) \ , \tag{7.18}$$

where v_α obtained form (7.17) can be interpreted as the posterior probability, conditional on $\underline{q}$, that the total number of Heads in the m trials is equal to α, and $f_\beta(\theta|\alpha,m)$ is a beta distribution with parameters (α, m).

The normal case for a situation as in Example 7.3 but with overlapping information is straightforward by the application of Winkler's (1981) model (see Section 2.4.4) which was obtained in a way similar to the above formulation which led to (7.16). In this case, the covariance matrix Σ in formulas (2.25) and (2.26) is such that its ith diagonal element is $1/n_i$ and the (i,j)th off-diagonal element is n_{ij}/n_in_j for $i,j = 1,\ldots,k$ as in Clemen (1987). Therefore, according to them, the posterior density $f(\theta|\underline{q})$ is also Gaussian.

7.6 The Choice of Experts.

The particular case of the overlapping information bases of the previous section, when there are at least two experts E_i and E_j such that one's information base is completely contained into the other's, i.e. $\underline{S}_{ij} = \underline{I}_i \cap \underline{I}_j = \underline{I}_j$ ($\underline{I}_j \subseteq \underline{I}_i$) for $i \neq j$, can sometimes help the SB with the selection of experts and some other times with determining the shared information.

This is because if $\underline{Y}(\underline{S})$, the summary statistics of shared information, could be provided by some experts, then they should be consulted by the SB. On the other hand, if $\underline{I}_j$ is subsumed by $\underline{I}_i$ and E_i does not share information with nobody else, then q_j would bring to the SB's aggregation rule no information which were not already provided by other experts. In the case of our Example 7.1 when $\underline{I}_2 \subset \underline{I}_1$ this mean that the expert E_2 statement plays no part whatsoever in SB's posterior density function for θ given $\underline{q}$. However, if $\underline{I}_j$ and/or $\underline{I}_i$ contain information which is shared with other experts, then q_j must be considered as we shall see.

We shall make use of another example to illustrate some possibilities in this context:

Example 7.4 (not choosing an expert). Consider the situation where $k = 3$, $m = 4$ and $\underline{I}_2 \subset \underline{I}_1$ ($\underline{S}_{12} = I_2$) such that $q_1 = Z_1 + Z_2 + Z_3$, $q_2 = Z_2 + Z_3$ and $q_3 = Z_4$. Note that E_2 does not share information with E_3 ($\underline{S}_{23} = \underline{S}_{13} = \emptyset$). An ID for this case could be $\mathcal{I}_8$ of Figure 7.7(a).

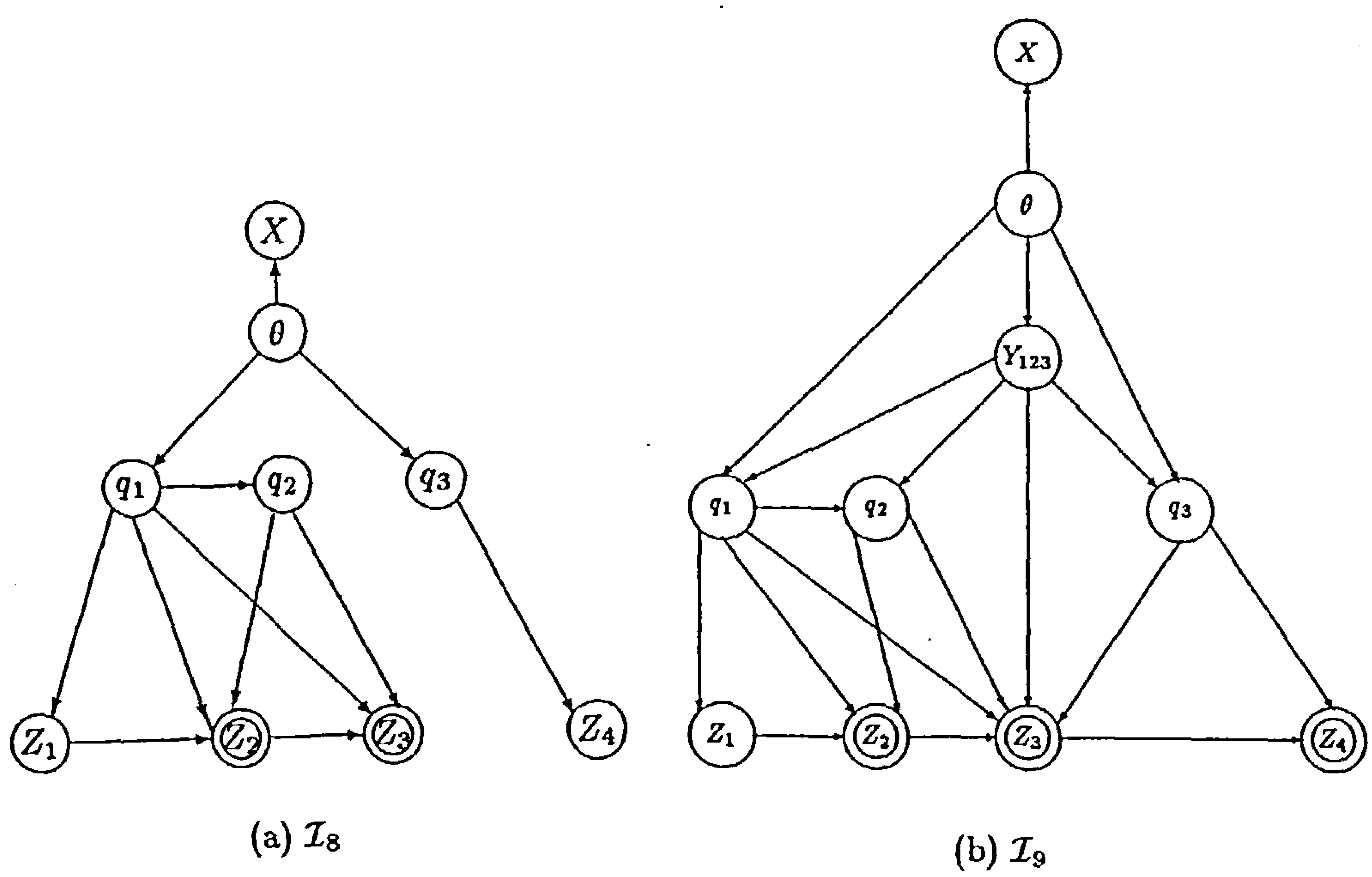
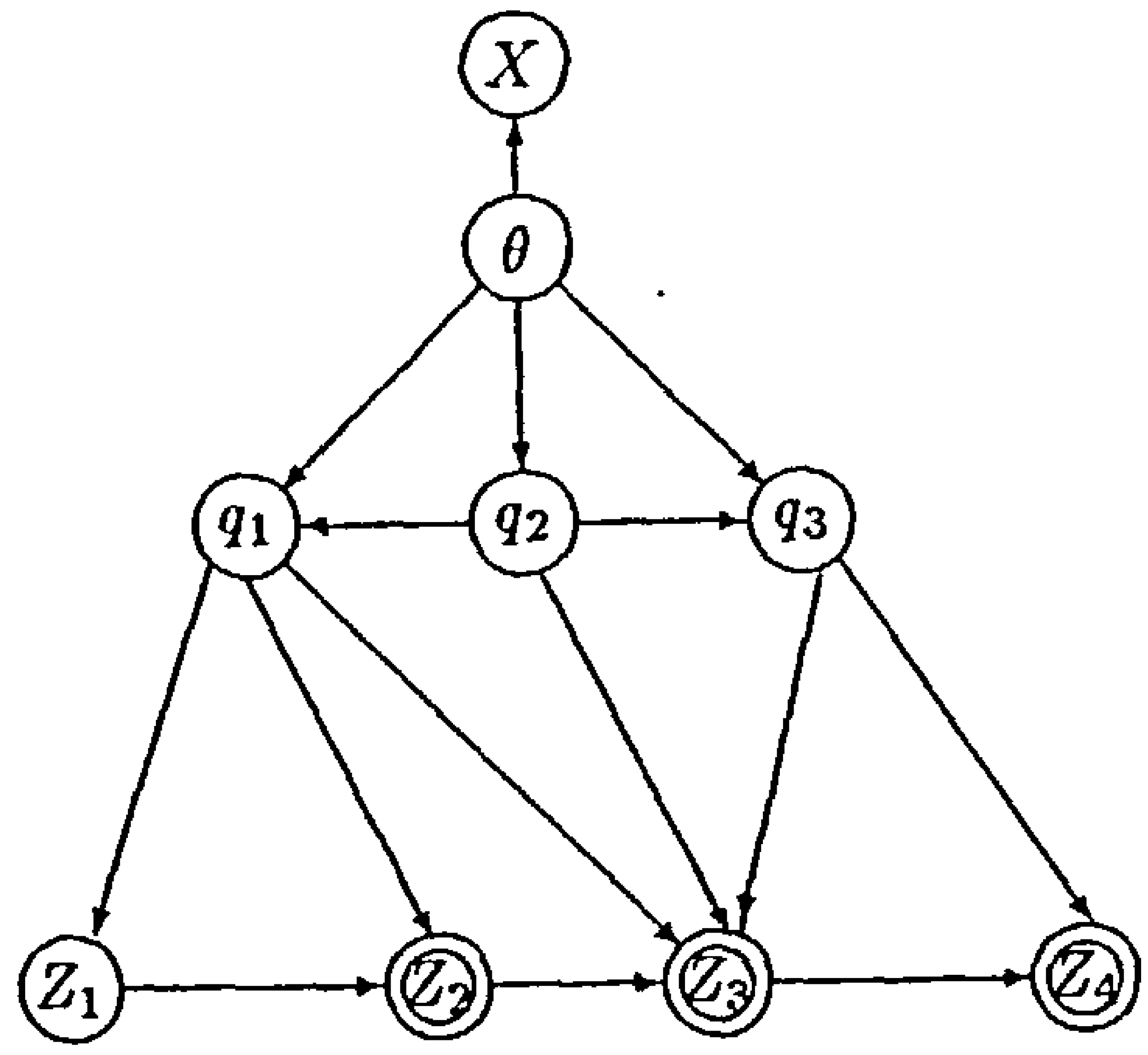


FIGURE 7.7. (a) $\mathcal{I}_8$ for $\underline{I}_2 \subset \underline{I}_1$ but $\underline{S}_{23} = \emptyset$, and (b) $\mathcal{I}_9$ for $\underline{I}_2 \subset (\underline{I}_1, \underline{I}_2)$ but $\underline{S}_{123} = Z_3$ in Example 7.4.

Observe that $u(\underline{q}) = q_1 + q_3$ is sufficient for θ relative to $\underline{Z}$, and from $\mathcal{I}_8$ the SB can state that (i) $q_2 \amalg \theta | q_1$, that is given q_1, q_2 brings no further information about θ, and (ii) $q_3 \amalg q_2 | \theta$.

However, if for example $q_3 = Z_3 + Z_4$ instead, then $\mathcal{I}_9$ in Figure 7.7(b) would represent the case, and although $\underline{I}_2 \subset \underline{I}_1$, (i) and (ii) above would not hold since $q_2 \not\perp\!\!\!\perp \theta | q_1$ neither $q_3 \not\perp\!\!\!\perp q_2 | \theta, q_1$. On the other hand, $q_2 \perp\!\!\!\perp \theta | q_1, Y_{123}$. Therefore, q_2 should be considered. In particular, it brings partial information about y_{123} to the model.



$\mathcal{I}_{10}$

FIGURE 7.8. An ID $\mathcal{I}_{10}$ representing the case where $\underline{I}_2 \subset (\underline{I}_1, \underline{I}_3)$ but $S_{123} = q_2$.

Certainly in the situation where one expert's information base $\underline{I}_i$ is encompassed by another expert's $\underline{I}_j$ but that base corresponds to the information shared between E_j and another expert E_l, q_i would bring valuable information to the SB in order to separate q_j of q_l. As an example of this situation refer to $\mathcal{I}_{10}$ in Figure 7.8 where $\underline{I}_2 \subset (\underline{I}_1, \underline{I}_2)$ but $q_2 = Z_3 = S_{123}$ is crucial for the SB to have $\underline{q}$ sufficient for θ regarding $\underline{Z}$. Note that, (i) $q_1 \perp\!\!\!\perp q_3 | \theta, q_2$ but $q_1 \not\perp\!\!\!\perp q_3 | \theta$ and, (ii) $\underline{Z} \perp\!\!\!\perp \theta | \underline{q}$. This corresponds to the case of Section 7.5.1 when the shared information is given to the SB by one of the experts' statements,

CHAPTER 8

EXTERNALLY BAYESIAN RECONCILIATION RULES

In this Chapter we make a link between the axiomatic and the modelling approaches to the problem of aggregating opinions. It is done by showing that it is possible (via the SB analysis of uncalibrated experts in non-degenerate situations) to obtain reconciliation rules for the experts' opinions which can be identified with EB LogOps as if performed by those experts as a group.

Let's start by assuming that the members of a group of k experts have non-overlapping information bases. Each member E_i considering his information base $\underline{I}_i$ $(i = 1, \ldots, k)$, reports to the SB his subjective opinion about the unknown parameter θ of his statistical model for X. It is assumed that the SB and the experts adopt the same dominating measure for their statistical models in a problem.

The opinion of the expert member E_i is expressed by the report of either

(i) his assessment of the hyper-parameters of his posterior density for θ given $\underline{I}_i$, or

(ii) the outcomes of experiments in his information base $\underline{I}_i$ $(i = 1, \ldots, k)$ and which characterise his likelihood function for the components of $\underline{I}_i$, together with his prior density $q_i(\theta)$ for θ.

From this later type of reports, the SB can build up her likelihood function for E_i, $\mathcal{L}(\underline{I}_i|\theta)$, and apply Bayes theorem to obtain E_i's posterior density function for θ given $\underline{I}_i$:

$$q_i(\theta|\underline{I}_i) \propto \mathcal{L}(\underline{I}_i|\theta)q_i(\theta) \ . \tag{8.1}$$

Note that even if, unlike E_i, the SB does not observe $\underline{I}_i$ directly, she is able to assess a likelihood function for $\underline{I}_i$ given θ from E_i's statements $(i = 1, \ldots, k)$ as in (ii) above.

8.1 The Experts' Miscalibration.

Although each member E_i reports what he thinks is his opinion about the *true* parameter θ, the SB knows that it is, in fact, his opinion about $\phi_i = h_i(\theta)$, where h_i is an increasing function of θ $(i = 1, \ldots, k)$. Therefore, the SB treats each $q_i(\theta|\underline{I}_i)$ in (8.1) above as $q_i(\phi_i|\underline{I}_i)$.

It is well known that in many practical situations "people are overconfident with general-knowledge items of moderate or extreme difficulty" (Lichtenstein et al., 1982 and references there in). Although experts tend to be better calibrated than non-experts, the above quote still apply (not as extremely) for expert judgements.

With this in mind, the SB really believes that the experts will choose either extreme values for their probability assessments or rather small variances for their probability densities for θ, or too, small interquartile ranges for their quantile assessments, depending on the problem.

Therefore, following Savage's (1971) suggestion that "you might discover with experience that your expert is optimistic or pessimistic in some respect and therefore temper his judgement", the SB adopts in line with Morris (1977) an approach in which she 'corrects' for the experts' *over-confidence* (that causes inaccuracies and miscalibration in their assessments) by obtaining the function $h_i(\theta)$ for each member E_i of the group $(i = 1, \ldots, k)$. The SB is assumed to be an accurate, well calibrated and coherent assessor. Another situation which reinforces the SB's belief about the experts' being miscalibrated occurs when those experts condition their assessments on the fact that they all see the same information even with this not being the case (and the SB knows it).

8.2 Independence and Over-confidence.

Since the experts have non-overlapping information bases we can make the following assumption:

Assumption 8.1. *The SB treats all the experts' uncalibrated statements $q_i(\phi_i|\underline{I}_i)$ $(i = 1, \ldots, k)$ as being independent conditionally on θ.*

Note that even if the experts had overlapping information bases it would not be completely unreasonable to the SB to adopt Assumption 8.1. This is because the experts are inaccurate and miscalibrated in their statements and also because their assessments are subjective opinions and not objective reports, such that even if they all have the same information bases they could still produce completely diverging statements (reflecting their diverging beliefs). The simplest model by the SB in this case would be one which chooses to treat their statements as being independent conditionally on θ. Observe that this choice is a maximum entropy choice for the SB on the densities associated to such assessments (see e.g. Newbold and Granger, 1974, Winkler, 1981, and Clemen, 1987).

Another assumption which supports the experts' miscalibration and therefore is useful for the analysis in this section is :

Assumption 8.2. *The SB, when assessing $q_i(\phi_i|\underline{I}_i)$ $(i = 1, \ldots, k)$, believes that the expert E_i believes that all other experts' observed information are equivalent to an independent replication of the information he observed.*

Assumptions (8.1) and (8.2) together mean that although the SB *knows* that E_i observes $\underline{I}_i$ to make his statement q_i about what he thinks is θ but she knows is $\phi_i = h_i(\theta)$, she *believes* that:

(i) the experts have disjoint information bases, i.e., $\underline{I}_i \cap \underline{I}_j = \emptyset$ for $i \neq j$ $(i, j = 1, \ldots, k)$, and

(ii) E_i thinks that the other experts' experiments gave the same result as his did.

For instance, this corresponds to the SB believing that each expert is substituting his maximum likelihood estimator of the result of the other experts and using this inputted data to calculate his posterior distribution.

Note that the above Assumption 8.2 also supports the SB's belief about the experts over-confidence. Another possibility for Assumption 8.2 would be that the SB believes that each expert E_i conditions his statement on the belief that $\underline{I}_i = \underline{I}$ is common knowledge to everyone. See also comments on Aumann (1976).

In the above context, if the SB adopts the Bayesian modelling approach, thus treating the experts' statements $\underline{q} = [q_1(\phi_1|\underline{I}_1), \ldots, q_k(\phi_k|\underline{I}_k)]$ as data, then her posterior density for θ is given by (7.2) which applied here gives

$$f(\theta|\underline{q}) \propto \prod_{i=1}^{k} f_i[q_i(\phi_i|\underline{I}_i)|\theta] f(\theta) \quad , \tag{8.2}$$

where $f_i[q_i(\phi_i|\underline{I}_i)|\theta]$ is the SB's likelihood function for the experts' statements about $\phi_i = h_i(\theta)$ given their information bases, and $f(\theta)$ is the SB's prior density for θ.

8.3 External Bayesianity and Reconciliation Rules.

We shall see in the following Sections 8.4 and 8.5 that with the above considerations about the members' miscalibration together with Assumptions 8.1 and 8.2, the SB can obtain at least in certain specific cases (such as for the binomial and normal models), reconciliation rules which are EB LogOp pools (see Section 2.2.4). Those LogOp pools also obey the unanimity condition.

Naturally, for the EB property to hold, the likelihood function must be common knowledge between the SB and the experts in all future data acquisition. Note that, the demand on the reconciliation rule being EB implies in θ and ϕ being identified to one another after new data is commonly observed. Therefore the modified LogOp for the group of k experts would be

$$\bar{f}(\theta|\underline{D}) = T[q_1(\theta|\underline{I}_1), ..., q_k(\theta|\underline{I}_k)](\theta|\underline{D}) \\ \propto g(\theta)\prod_{i=1}^{k}[q_i(\theta|\underline{I}_i)]^{w_i} \quad , \tag{8.3}$$

where $\underline{D}$ is the data set $\{\underline{I}, \underline{w}\}$, with $\underline{I} = (\underline{I}_1, \ldots, \underline{I}_k)$ and weights $\underline{w} = (w_1, \ldots, w_k)$ such that $\sum_{i=1}^{k} w_i = 1$. T is a pooling operator, and $g(\theta)$ is a bounded function of θ only.

8.4 The Binomial Model.

To illustrate the binomial case we make use of a coin toss experiment in which we assume that each expert E_i after observing n_i' tosses of a coin whose unknown *true* probability is θ, reported to the SB that he had observed r_i *heads* (H) and $n_i - r_i$ *tails* (T) $(i = 1, \ldots, k)$. Notice that those reports are not necessarily the outcomes of the coin tosses E_i observed but could be his opinion formed after that observation (i.e. including the parameters of his prior density for θ). Each expert's information base $\underline{I}_i$ is compounded of the outcomes of all coin tosses E_i has observed. The SB does not observe $\underline{I}_i$ but only the experts' statements.

8.4.1 The case of exchangeable experts.

If the SB considers the experts $E_1, \ldots, E_k$ to be exchangeable in terms of their lack of calibration (i.e. they are equally miscalibrated), then $\phi_i = \phi$ for all $i = 1, \ldots, k$. Thus, under Assumptions 8.1 and 8.2, the SB's probability table for the experts possible statements about the possible outcomes of a coin toss Z, would be like that in Table 8.1. There, the first row corresponds to the number of Heads (NoH) seen by $0, 1, 2, \ldots, k$ experts simultaneously. For instance, if we define $\theta = Pr\{H|Z\}$, then the second row shows the SB's probability : $c(\theta)\theta$ that all will report H; $c(\theta)t_1(\theta)$ that just one expert will report H; $c(\theta)t_2(\theta)$ that two experts will report H; and so on to $c(\theta)(1-\theta)$ that all will report T. The terms $c(\theta)$ and $t_i(\theta)$ thereafter denoted by just c and t_i $(i = 1 \ldots, k)$, are functions of θ only. The third row shows the total number of combinations (TNC) and the fourth and fifth rows show the number of combinations with H and T (respectively) in a given margin (NCH and NCT), all associated with the first row.

The terms c and t_i $(i = 1, \ldots, k)$ can be fixed by using Assumption 8.1 and $\phi = h(\theta)$ obtained as in the following theorem :

NoH	0	1	2	$\cdots$	r	$\cdots$	$k-1$	k
Pr.	$c(1-\theta)$	ct_1	ct_2	$\cdots$	ct_r	$\cdots$	ct_{k-1}	$c\theta$
TNC	$\binom{k}{0}$	$\binom{k}{1}$	$\binom{k}{2}$	$\cdots$	$\binom{k}{r}$	$\cdots$	$\binom{k}{k-1}$	$\binom{k}{k}$
NCT	$\binom{k-1}{0}$	$\binom{k-1}{1}$	$\binom{k-1}{2}$	$\cdots$	$\binom{k-1}{r}$	$\cdots$	$\binom{k-1}{k-1}$	0
NCH	0	$\binom{k-1}{0}$	$\binom{k-1}{1}$	$\cdots$	$\binom{k-1}{r-1}$	$\cdots$	$\binom{k-1}{k-2}$	$\binom{k-1}{k-1}$

TABLE 8.1. The SB's probabilities for k exchangeable uncalibrated experts in a binomial experiment.

Theorem 8.3. *Suppose that the SB considers the members of a group of k experts to be exchangeable in their lack of calibration on their statements about an experiment whose outcomes are binomially distributed with true probability θ. If the Assumptions 8.1 and 8.2 hold then, the probability of success for which each member produces a statement is*

$$\phi = \frac{\theta^{\frac{1}{k}}}{\theta^{\frac{1}{k}} + (1-\theta)^{\frac{1}{k}}} \quad . \tag{8.4}$$

Proof.

From rows 4 (NCT) and 5 (NCH) of the Table 8.1 we have that

$$P(H) = c\Big[\theta + \sum_{r=0}^{k-1} \binom{k-1}{r} t_r\Big] \ , \tag{8.5}$$

and

$$P(T) = c\Big[(1-\theta) + \sum_{r=0}^{k-2} \binom{k-1}{r} t_{r+1}\Big] \ , \tag{8.6}$$

where $c = [1 + \sum_{r=1}^{k-1} \binom{k}{r} t_r]^{-1}$.

Let $\lambda = \frac{P(T)}{P(H)}$, *then for the independence hypothesis of Assumption 8.1 we must have that*

$$c\theta = [P(H)]^k \tag{8.7}$$

and

$$c(1-\theta) = [P(T)]^k \ . \tag{8.8}$$

From the above equations (8.7) and (8.8) we can obtain $\lambda = (\frac{(1-\theta)}{\theta})^{\frac{1}{k}}$. *Also, we must have for the diverging opinions that*

$$ct_r = [P(T)]^r[P(H)]^{k-r} \tag{8.9}$$

for $r = 1, \ldots, k-1$. *Equations (8.9) and (8.7) imply that* $t_r = \theta\lambda^r = \theta^{\frac{k-r}{k}}(1-\theta)^{\frac{r}{k}}$. *Consequently,*

$$c = [\theta^{\frac{1}{k}}(1-\theta)^{\frac{1}{k}}]^{-k} \; . \tag{8.10}$$

Finally, substituting (8.10) in (8.5) and (8.6) we have that

$$P(H) = \frac{\theta^{\frac{1}{k}}}{\theta^{\frac{1}{k}} + (1-\theta)^{\frac{1}{k}}} \; ,$$

and

$$P(T) = \frac{(1-\theta)^{\frac{1}{k}}}{\theta^{\frac{1}{k}} + (1-\theta)^{\frac{1}{k}}} \; ,$$

respectively. □

Notice that in particular this sets the re-calibration function of each expert so that

$$\log\left[\frac{\phi}{(1-\phi)}\right] = \frac{1}{k}\log\left[\frac{\theta}{(1-\theta)}\right] \; .$$

In particular, the LogOp pool arising from this set of assumptions, has the uncomfortable implication that the SB assures the more experts the more extreme are their misspecification, i.e. by the SB adding another expert makes the other more optimistic. However, adjustments can be made to the SB's assumptions so that a LogOp is obtained with weights adding up to some value between 1 and k and thus lie somewhere between the case above and Winkler's (1981) model with independent sources of variation. This will incidentally sacrifice the EB property for the weaker prior-to-posterior coherence (PPC) —see Section 2.2.3. Alternatively one could consider this result as questioning the LogOp when the number of experts was large, if the assumptions above were believed reasonable. A detailed analysis of such models will be the topic of further research.

This above result leads, in our case, to the argument that just enough experts should be employed such that their mutual independence confirmation convinces the SB that an event has occurred.

The Theorem 8.3 can now be used to characterise the SB's EB reconciliation rule :

Theorem 8.4. *Under the hypotheses and result of Theorem 8.3, also assume that*

(1) *the likelihood function for the outcomes of the binomial experiment is common to the SB and to the experts in all new data acquisition ;*

(2) *the weights associated to the modified LogOp (8.3) obtained by the group of experts is such that* $w_1 = \cdots = w_k = \frac{1}{k}$, *and*

(3) $g(\theta)$ *in (8.3) equals* $g'(\theta)f(\theta)$ *where* $g'(\theta) = [\theta^{\frac{1}{k}} + (1-\theta)^{\frac{1}{k}}]^{-2\bar{n}}$, $\bar{n} = \frac{1}{k}\sum_{i=1}^{k} n_i$ *and* $f(\theta)$ *is the SB's prior density for* θ *as in (8.2).*

Then the SB's posterior density $f(\theta|\underline{q})$ *is a reconciliation rule which is also an EB modified LogOp pool.*

proof.

First, under Assumption 8.1, the SB's likelihood function for the experts' statements in (8.2) can be obtained from her knowledge of $\underline{n} = (n_1, \ldots, n_k)$ *and* $\underline{r} = (r_1, \ldots, r_k)$, *giving*

$$f(\underline{n}, \underline{r}|\theta) \propto \prod_{i=1}^{k} \phi^{r_i}(1-\phi)^{n_i - r_i} \quad . \tag{8.11}$$

Substituting (8.4) in (8.11) gives that

$$f(\underline{n}, \underline{r}|\theta) \propto g'(\theta)\theta^{\bar{r}}(1-\theta)^{\bar{n}-\bar{r}}$$

where $\bar{r} = \frac{1}{k}\sum_{i=1}^{k} r_i$, $\bar{n} = \frac{1}{k}\sum_{i=1}^{k} n_i$ *and* $g'(\theta) = [\theta^{\frac{1}{k}} + (1-\theta)^{\frac{1}{k}}]^{-2\bar{n}}$. *Therefore, the SB's posterior density in (8.2) becomes*

$$f(\theta|\underline{n}, \underline{r}) \propto \theta^{\bar{r}}(1-\theta)^{\bar{n}-\bar{r}} g'(\theta)f(\theta) \quad . \tag{8.12}$$

On the other hand, the modified LogOp pool (8.3) for the experts binomial densities has the form

$$\bar{f}(\theta|\underline{D}) \propto g(\theta)\theta^{\sum_{i=1}^{k} w_i r_i}(1-\theta)^{\sum_{i=1}^{k} w_i(n_i - r_i)} \quad . \tag{8.13}$$

Now, for the modified LogOp (8.13) to be identified with the reconciliation rule (8.12) the requirements (2) and (3) of Theorem 8.4 must hold. To be EB this pool also requires that (1) holds. □

Notice that because the pooling should be performed by the group itself, the densities in (8.13) are not 'corrected' by $h(\theta)$. They consider themselves as being well calibrated experts.

If the SB has a diffuse prior density for θ then $g(\theta) = g'(\theta)$ and (8.13) equals (8.12) if just $w_i = \frac{1}{k}$ for $i = 1\ldots, k$. This choice $w_i = \frac{1}{k}$ is coherent with with the experts' being considered exchangeable. Also, the result above gives a precise interpretation for $g(\theta)$ in the modified LogOp pool as corresponding to a weighted prior density function of a SB. The only restriction here is that this prior must be such that

$$\int g(\theta)\theta^{\sum_{i=1}^{k} w_i r_i}(1-\theta)^{\sum_{i=1}^{k} w_i(n_i - r_i)}d\theta$$

is finite (see Genest et al., 1986).

For k=2,

$$\phi = h(\theta) = \frac{\theta^{1/2}}{\theta^{1/2} + (1-\theta)^{1/2}} \cdot$$

Note that $h(\theta)$ really 'smoothes' the values of θ specially on the boundaries of the extremes 0 and 1, 'correcting' the overconfident experts as can be seen in Figure 8.2.

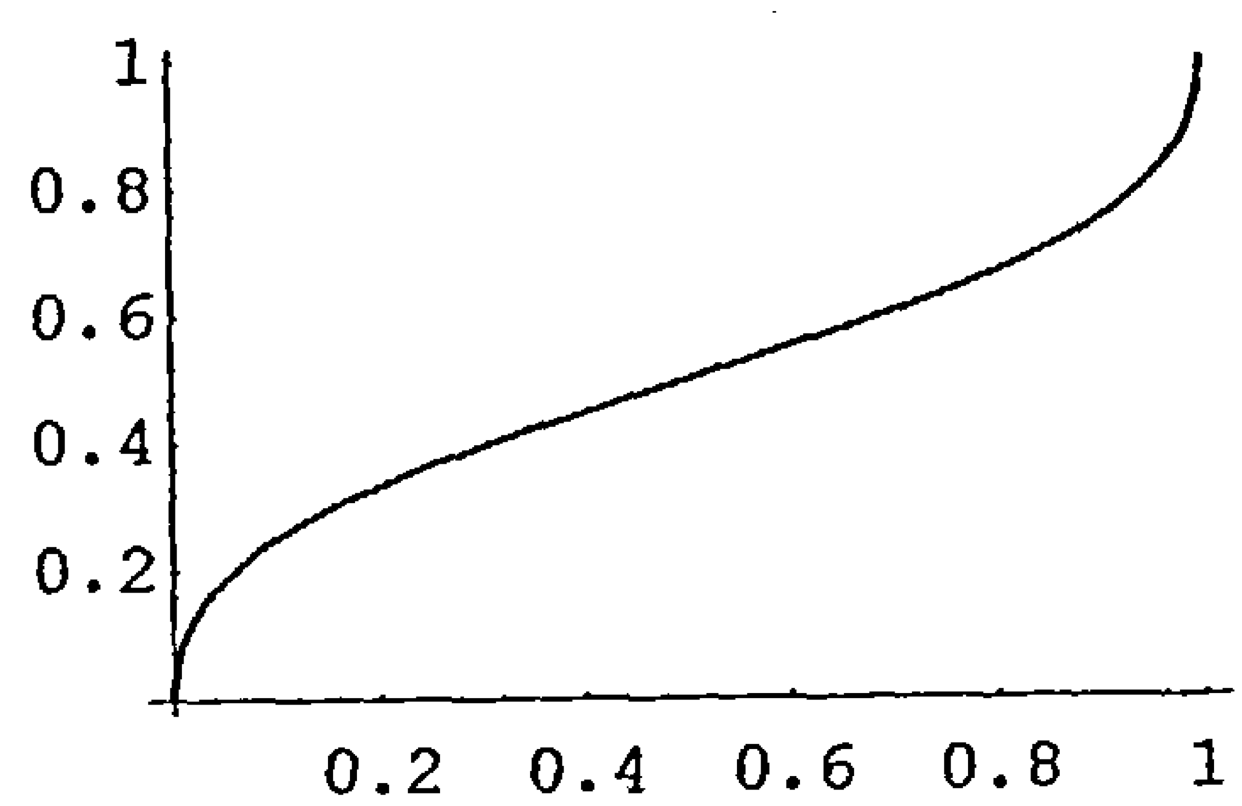


FIGURE 8.2. The function $\phi = h(\theta)$ of two exchangeable uncalibrated experts in the binomial case.

8.4.2 Non-exchangeable experts.

If the SB thinks the experts have different degrees of miscalibration then they cannot be considered exchangeable anymore. Under the Assumptions 8.1 and 8.2, the joint probability table for their possible statements would be for example for $k = 2$ that of Table 8.3 in which

(i) $c\theta$ is the probability that both experts will report H,

(ii) $c\theta^{w_1}(1-\theta)^{w_2}$ is the probability that E_1 will report H and E_2 will report T,

(iii) $c\theta^{w_2}(1-\theta)^{w_1}$ is the probability that E_2 will report H and E_1 will report T, and

(iv) $c(1-\theta)$ is the probability that both will report T.

Again here $c = c(\theta)$. The terms w_1 and w_2 are weights such that $w_1 + w_2 = 1$. It is through those weights that the SB express how relatively miscalibrated the experts are.

E_1 \ E_2	H	T
H	$c\theta$	$ct(\theta)$
T	$ct(\theta)$	$c(1-\theta)$

TABLE 8.3. Joint probability table for two non-exchangeable experts in a binomial experiment.

Using the axiom of probability which says that $P[\Omega] = 1$ we can obtain $c(\theta) = [\prod_{i=1}^{2} \theta^{w_i}(1-\theta)^{w_i}]^{-1}$. The SB could then assess her calibration function for each expert regarding the true probability as

$$\phi_i = \frac{\theta^{w_i}}{\theta^{w_i} + (1-\theta)^{w_i}} , \tag{8.14}$$

where $\sum_{i=1}^{2} w_i = 1$ for $i = 1, 2$.

Now, using the same notation as in Section 8.4.1, analogous to (8.13), the modified LogOp pool for the group of experts obtained from their stated densities is

$$\bar{f}(\theta|\underline{D}) \propto g(\theta)\theta^{\bar{r}}(1-\theta)^{\bar{n}-\bar{r}} , \tag{8.15}$$

with $\bar{r} = \sum_{i=1}^{2} w_i r_i$, $\bar{n} = \sum_{i=1}^{2} w_i n_i$ and $g(\theta)$ being a bounded function of θ only.

On the other hand, the SB's reconciliation rule analogous to (8.12) is

$$f(\theta|n_1, r_1, n_2, r_2) \propto g'(\theta)\theta^{\bar{r}}(1-\theta)^{\bar{n}-\bar{r}} f(\theta) , \tag{8.16}$$

where $g'(\theta) = 1/c(\theta)$ and $\bar{r}$ and $\bar{n}$ are defined as above. Thus, setting $g(\theta) = g'(\theta)f(\theta)$ gives our identity again.

For k experts, set ϕ_i as (8.14) above but with $i = 1, \ldots, k$ and $\sum_{i=1}^{k} w_i = 1$. Then again using independence and resolution of margins gives the identity between (8.15) and (8.16) but with $\bar{r} = \sum_{i=1}^{k} w_i r_i$, $\bar{n} = \sum_{i=1}^{k} w_i n_i$ and $c(\theta) = \prod_{i=1}^{k}[\theta^{w_i} + (1-\theta)^{w_i}]$. Also, for $i \neq r$, $t_r(\theta) = \theta^{w_r} \prod[\theta^{w_i} + (1-\theta)^{w_i}]$ $(r = 2, ..., k-1)$.

In fact, we have just proved the following theorem :

Theorem 8.5. *Suppose that Assumptions 8.1 and 8.2 hold for k non-exchangeable uncalibrated experts. Also, assume that the SB calibrates each expert's likelihood function in* *her reconciliation rule (8.2) by using*

$$\phi_i = \frac{\theta^{w_i}}{\theta^{w_i} + (1-\theta)^{w_i}} \quad , \tag{8.17}$$

where $\sum_{i=1}^{k} w_i = 1$ *for* $i = 1, \ldots, k$, *Then the SB's reconciliation rule (8.2) identifies with the experts' EB modified LogOp pool (8.3), if and only if*

(1) $g(\theta) = g'(\theta) f(\theta)$ *where* $g'(\theta) = \{\prod_{i=1}^{k} [\theta^{w_i} + (1-\theta)^{w_i}]\}^{-1}$ *and* $f(\theta)$ *is a prior density function for* θ *;*

(2) $\sum_{i=1}^{k} w_i = 1$ *; and*

(3) *the likelihood of all new data acquisition in the binomial experiment is common to the SB and to all the experts.*

Notice the interesting point here that, under this model, the weights $\underline{w}$ reflect not the quantity of information available to each expert but his lack of calibration in communicating his information.

8.5 The Gaussian Case.

Assume it is CK that X is Gaussian with unknown mean θ but with known precision τ. Thus, each expert E_i, based on his information base $\underline{I}_i$, assesses his mean μ_i and precision τ_i for a normal density for θ ($i = 1, \ldots, k$). The Assumptions 8.1 and 8.2 lead to a posterior distribution for SB of the form (8.3).

Assuming initially that the experts are exchangeable in their lack of calibration, they all report $\tau_i = \tau$ ($i = 1, \ldots, k$).

Because of the Assumption 8.1 , the SB's likelihood function for the experts can be factorized as $f(\underline{\mu}, \tau|\phi) = \prod_{i=1}^{k} f_i(\mu_i, \tau|\phi)$, where $\underline{\mu} = (\mu_1, \ldots, \mu_k)'$ and ϕ is the parameter that the SB knows the experts' assessments are about. Therefore, this function has the form of a normal density with mean $\bar{\mu}$ and precision τ, where $\bar{\mu} = \sum_{i=1}^{k} \mu_i$. However, the SB is interested in $f(\underline{\mu}, \tau'|\theta)$ where $\tau' = \tau/k$ is her 'correction' for the experts' stated precisions when θ and not ϕ is considered. This likelihood also has the form of a normal but with parameters $(\bar{\mu}, \tau')$.

Notice that unanimity is preserved here and the correction τ' on the experts precisions corresponds to the experts having in fact assessed their opinions for

$$\phi = \frac{\theta}{\sqrt{k}}$$

instead.

The modified LogOp for the uncalibrated experts with the weights $w_i = 1/k$ for all $i = 1, \ldots, k$ gives

$$\bar{f}(\theta|\underline{\mu}, \tau) \propto \exp\{-\frac{\tau}{2}(\bar{\mu} - \theta)^2\}g(\theta) \quad ,$$

which identifies with the Bayesian reconciliation rule when $g(\theta) = f(\theta)$, i.e. the SB's prior for θ.

Again here, this gives an interpretation to $g(\theta)$ in the modified LogOp. Although Genest et al. (1986) point out that $g(\theta)$ must be essentially bounded and the there is no reason why a prior density should be bounded, in many practical situations this is the case.

When the experts are not exchangeable, the SB can choose

$$\phi_i = \theta\sqrt{w_i} \quad ,$$

where w_i corresponds to the SB's weight or 'correction' factor for the expert E_i. In this case, the identity between the Bayesian and the pooling approaches occurs when $g(\theta) = f(\theta)$ and the weights in the LogOp pool are w_i $(i = 1, \ldots, k)$.

8.6 The CEB Reconciliation Rules.

The previous results for $\Omega = \theta$ most naturally extend for when $\Omega = \{\theta_1, ..., \theta_n\}$ and the conditional independence structure of associations between the variables in $\underline{X} = (\underline{X}_1, ..., \underline{X}_n)$ is a decomposable PCG of the form defined in Section 5.2, with the group of experts obeying Condition 5.4 (see Section 5.3) for cutting likelihoods which are common to the group.

CHAPTER 9

CONCLUSIONS AND FURTHER RESEARCH

In this concluding chapter the results obtained in the thesis are discussed. The extend of their applicability is commented and directions for further research are indicated.

We begin in Section 9.1 by discussing the conditionally externally Bayesian (CEB) pools defined in Chapter 6. We comment on the flexibility of their weights as well as on their application to complete chain graphs, to non-decomposable influence diagrams induced from partially complete chain graph and to decomposable non-complete induced influence diagrams. The issue of likelihoods in decomposable structures is also discussed. The discussions on the CEB linear pools end the section. In Section 9.2 the results which allow a link between the axiomatic and the Bayesian modelling approaches are discussed.

9.1 The CEB Pools.

9.1.1 The flexibility on the weights.

One of the main features of a CEB pool when applied to a variable X of the group's graphical model, is that its weights can vary according to the occurred values of variables in the parent set of X. This suggests the development of methods for obtaining weights which would also consider the group's members expertise on those variables related to X. Alternatively, an interpretation of the weights in terms of the members' calibration would be plausible, with the difference that in this case the auxiliary experiments usually employed for calibrating an expert need not be purely independent. Another question here is whether this flexibility would also allow an interpretation which would include measures of dependence among the group's members.

Although the extension of the externally Bayesian (EB) to the CEB pools is more general in terms of the flexibility on the weights, it imposes some restrictions on their domain of application. However, those restrictions are useful in order to avoid the impossibility results of non-preservation of independence for general pooling operators.

Immediate natural extensions of this thesis would be to consider the implications of other cases such as when:

(i) the commonly agreed graph is a chain graph and undirected edges are allowed within incomplete chain elements of the partially complete chain graphs,

(ii) the use of mixed graphs (Lauritzen, 1996) with both discrete and continuous variables being represented,

(iii) the weights, included as variables in the common partially complete chain graph of a problem, are themselves object of group consensus, and

(iv) the members do not agree on the same conditional independence structure for the problem.

9.1.2 Complete graphs.

Complete chain graphs are always decomposable and make no statements about conditional independence being probabilistically valid for all situations. So, for complete chain graphs, Theorem 6.4 implies that all CEB rules are EB to general likelihoods. This is not in contradiction to the original theorems of Genest et al. (1986) because they make an additional requirement that the combination rule must be achieved through a pooling operator. In particular, the value of a pooled density at a point $\underline{x}$ depends only on $\underline{x}$ and on the values of the joint densities $f_1(\underline{x})$ and $f_2(\underline{x})$ at that point $\underline{x}$ for the first and second expert.

The CEB pools are not formed as pooling operators on joint densities on all the variables of a system. They act as components of pooling operators on conditional densities. The argument that a pooling should be a pooling operator on a joint density appears to be weak.

Each ordering of variables in a complete induced influence diagram gives a different class of EB pool, so with n variables there are $n!$ different classes of EB pools defined by different chain graphs. Therefore, the class of EB pools is extremely rich, a fact obscured by the insistence that a pooling should be an operator on a joint density.

This multiplicity in the complete case is, in one sense, a problem, since we need to choose which CEB pool to use. But this will largely be determined by the time order in which the random variables are observed. We need this to fix the weights $\underline{w}(j)$ associated with the j^{th} variable $X(j)$ since $\underline{w}(j)$ is allowed only to depend on the parents of $X(j)$. In the complete case this is $X(1), ..., X(j-1)$. And for all possible pools to operate in the class of CEB pools associated with this graph, $X(1), ..., X(j-1)$ will not need to be known before the pooling takes place.

9.1.3 The non-decomposable chain graphs.

An influence diagram is just a set of conditional independence statements. When the influence diagram induced by the underlying partially complete chain graph is not decomposable, assimilation of data tends to destroy that agreed structure. Within the context of external Bayesianity, which addresses the group's behaviour when assimilating information, it is natural therefore to work only with information in the influence diagram which is not destroyed by such assimilation.

An agreed non-decomposable influence diagram $\mathcal{I}$ can always be made decomposable by marrying parents and adding directed edges until it is decomposable (see e.g. Lauritzen and Spiegelhalter, 1888). By using this derived influence diagram $\mathcal{J}$ as a basis for the CEB pools instead of $\mathcal{I}$, we do not deny that the conditional independence statements supplied by $\mathcal{I}$ but not $\mathcal{J}$ do not exist. Rather we say that it is only sensible to *explicitly formulate the pooling* on those agreed conditional independencies which can reasonably be assumed to be preserved after simple types of sampling.

9.1.4 Using CEB pools not based on complete graphs.

One legitimate question that might be asked is : why use chain graphs whose induced influence diagrams are not complete ?

Because an influence diagram is always acyclic there will always be a complete dimensional graph which has it as a subgraph and so is a valid influence diagram description of the problem. Furthermore, the CEB pools which are associated with complete graphs have the advantage that they are EB with respect to all likelihoods. On the other hand, CEB pools related to incomplete graphs are only EB to data which gives rise to a certain structure of likelihood (the cutting likelihood). There are three answers to this question :

(i) *Simplicity.* If the type of information you expect to receive will automatically preserve conditional independence structures it seems perverse to demand methods of combination of densities which exhibit individual's dependence structures they will never believe.

(ii) *Preservation of symmetry.* Suppose two random variables X_i, X_j in the random vector $\underline{X}$ are agreed to exhibit conditional independence and are symmetric in the sense that there is no clear order of causality or association between them. So, to introduce such an association into the pooling algorithms seems to be artificial and undesirable.

(iii) *Fixing a frame.* In a given problem, experts will agree a set of random vectors $\underline{X}_1, ..., \underline{X}_n$ on which they will pool their opinions. However, in most circumstances, they will also each have beliefs about other variables $\underline{X}^*$, agreed as independent of $\underline{X}$. Implicitly in any pooling, they will ignore the disparity between their beliefs about $\underline{X}^*$. Similarly at a future time if asked to combine their beliefs about $\underline{X}^*$ they will choose to ignore the disparity in their beliefs about $\underline{X}$. But to do this implies the use of a CEB rule which explicitly demands in its associated chain graph that $\underline{X} \amalg \underline{X}^*$.

So if we do not allow incomplete cases, in different and independent forecasting problems about $\underline{X}$ and $\underline{X}^*$ then we would need to prioritise $\underline{X}$ and $\underline{X}^*$.

Despite the restrictions imposed to our problem and thus to the pooling operators discussed here, they are quite general to be useful in practical situations as we have seen. For instance, the assumption of a common partially complete chain graph structure is a rather mild restriction for one can fairly easily devise situations where this is the case. Moreover, as we have already mentioned, non-decomposable partially complete chain graph can always be made decomposable.

The class of decomposable graphical models introduced by Lauritzen et al. (1984), is the one which under certain restrictions on the form of input data, retains the structure (coded in terms of conditional independence statements) in a prior-to-posterior analysis. This fact is used extensively to create quick algorithms for calculating posterior distributions in high dimensional problems (see e.g. Dawid, 1992, Jensen et al., 1994, and Smith and Papamichail, 1996). Here we have used this same property to define classes of combination rules which are, in a partial sense, externally Bayesian.

The demands on certain forms of input data (cutting likelihoods) together with decomposability for the graphical model are necessary and sufficient conditions for the a posteriori preservation of the group's conditional independence structure. Decomposability alone is not a sufficient condition. However, depending on the problem, the demand on sampling over ancestral sets can be relaxed. This is the case when the likelihoods in a problem are separable.

9.1.5 The CEB linear pools.

We have seen that even allowing the marginalization property to hold for set of variables, the application of linear pools to elements of conditional independence structures brings other difficulties associated to the odd form that the resulting joint pool in general takes.

It seems that the only instances when the linear pools can be CEB are when they are dictatorships on elements of the agreed structure. In this case the joint pool breaks down nicely and is EB with the difference that the weights can reflect relative expertises on different components of the structure.

Further research could possibly lead to interpretations of the cross-product of weights and densities in the joint pool, which would allow the incorporation to the pooling of various correlations that might occur between the members' assessments, their expertises and the diverse variables of the problem. It may not be an elegant approach for treating the dependence issue but could answer the problem of modelling dependencies in axiomatic approaches.

9.2 A Link of the Modelling with the Axiomatic Approach.

9.2.1 General expert problems.

Because expert judgement problems are in general quite complex with diverse possibilities of associations among their various underlying components, the graphical modelling approach shows to be extremely useful in helping both the characterisation and the understanding of the situation being modelled.

We have investigated the class of expert problems for which a variable X and the components of the vector $\underline{Z}$ of variables informative about X are conditionally independent given the parameter θ of the models associated to X and $\underline{Z}$. The results obtained can be extended for more general structures of association between those variables. In fact, the graph of Figure 7.1, that represents a general subclass of expert problems, suggests several other possibilities for further investigation in this area. For example, the issue of $\underline{Z}$ having associated models with parameters Φ distinct but informative about θ implies that the supra-Bayesian (SB) decision maker also has to assess $f(\theta|\phi)$.

Also, considerations of sufficiency of expert opinions regarding their individual information is a quite interesting issue to be developed, as there are not many works in this area.

9.2.2 The SB analysis of uncalibrated experts.

We have shown that a supra-Bayesian (SB) analysis of uncalibrated experts with considerations resembling game theory allows a connection of the results obtained for CEB pools with combination rules obtained via the Bayesain paradigm applied to the expert problem. In fact, a SB calibrating experts via calibration functions seems to be one of the few, if not the only, way in which such a connection can be achieved. This subject is actually under investigation and all the related concepts and the material presented here (in Chapters 7 and 8) are actually being further developed.

REFERENCES

Aczél, J. (1966), *Lectures on functional equations and their applications*, Academic Press, New York.

Aczél, J., Ng, C. T. and Wagner, C.G. (1984), *Aggregation theorems for allocation problems*, SIAM J. Alg. Disc. Meth. **5**, 1–8.

Adams, J. K. and Adams, P. A. (1961), *Realism of confidence judgements*, Psychological Review **68**, 33–45.

Agnew, C. E. (1985), *Multiple probability assessments by dependent experts*, J. Amer. Statist. Assoc. **80**, 343–347.

Alpert, M. and Raiffa, H. (1982), *A progress report on the training of probability assessors*, Judgement under uncertainty : Heuristics and biases (D. Kahnemann, P. Slovik and A. Tversky, eds.), Cambridge University, U. K., pp. 294–306.

Anderson, T. W. (1971), *An introduction to the multivariate statistical analysis*, John Wiley and Sons, New York.

Andersson, S. A., Maidigan, D. and Perlman, M. D. (1996), *On the Markov equivalence of chain graphs, undirected graphs, and acyclic digraphs*, Scand. J. Statist. **26** (to appear).

Apostolakis, G. E., Farmer, R. R. and van Otterloo, R. W. (eds.) (1988), Special Issue of Reliability, Engineering and System Safety, vol. 23, pp. 247–320.

Arkes, H. R. and Hammond, K. R. (1986), *Judgement and Decision Making*, Cambridge University Press.

Arrow, K. J. (1951), *Social choice and individual values*, Wiley, New York.

Aumann, R. J. (1976), *Agreeing to disagree*, Ann. Statist. **4**, **No.6**, 1236–1239.

Bacharach, M. (1972), *Scientific disagreement*, Unpublished manuscript, Christchurch, Oxford.

Bacharach, M. (1975), *Group consensus in the face of differences of opinions*, Manag. Sci. **22**, 182–191.

Bacharach, M. (1979), *Normal Bayesian dialogues*, J. Amer. Statist. assoc. **74**, 837–846.

Baird, D. (1985), *Lehrer-Wagner consensual probabilities do not adequately summarize the available information*, Synthese **62**, 47–62.

Barlow, R. (1986), *Using influence diagrams to solve a calibration problem*, Proceedings of International Symposium on Probability and Bayesian Statistics, Austria, pp. 23–36.

Bates, J. M. and Granger, C. W. J. (1969), *The combination of forecasts*, Opl. Res. Q. **20**, 451–468.

Bayarri, M. and DeGroot, M. H. (1989), *Gaining weight : a Bayesian approach*, Bayesian Statistics (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. M. F. Smith, eds.), vol. 3, North Holland, Amsterdam.

Beach, B. H. (1975), *Expert judgement about uncertainty : Bayesian decision making in realistic settings*, Org. Behav. Hum. Perform. **14**, 10–59.

Beach, L. R. and Christensen-Szalanski (1984), *Citation bias: fad and fashion in the judgement and decision literature*, Amer. Psych. **39**, 75–78.

Beach, L. R., Christensen-Szalanski and Barnes, V. (1987), *Assessing human judgement: has it been done, can it be done, should it be done ?*, Judgemental Forecasting (G. Wright and P. Ayton, eds.), John Wiley, Chichester, pp. 49–62.

Bernardo, J. M. (1979), *Expected information as expected utility*, Ann. Statist. **7**, 686–690.

Birch, M. N. (1963), *Maximum likelihood in three way contingency tables*, J. R. Statist. Soc. Series B **25**, 220–233.

Blalock, H. M., Jr. (ed.) (1971), *Causal Models in the Social Sciences*, Aldine-Atheston, Chicago.

Bordley, R. F. and Wolf, R. W. (1981), *On the aggregation of individual probability estimates*, Man. Sci. **27**, 959–964.

Bordley, R. F. (1982), *A multiplicative formula for aggregating probability assessments*, Man. Sci. **28**, 1137–1148.

Bordley, R. F. (1983), *Bayesian group decision theory*, Conference on Information Pooling and Group Decision Making, Irvine, CA, March 1983.

Brier, G. (1950), *Verification of forecasts expressed in terms of probability*, Mon. Weath. Rev. **75**, 1–3.

Brown, T. A. (1973), *An experiment in probabilistic forecasting*, The RAND corp. Report **R-944-ARPA**, Santa Monica, CA.

Bunn, D. W. (1975), *A Bayesian approach for linear combining models*, Op. Res. Q. **40**, 322–327.

Bunn, D. W. (1978), *A simplification of the matrix beta distribution for combining estimators*, J. Opl. Res. Soc. **29**, 1013–1016.

Bunn, D. W. (1985), *Statistical efficiency in the linear combination of forecasts*, Int. J. Forecasting **1**, 151–163.

Christensen-Szalanski, J. and Bushyhead, J. B. (1981), *Physicians' use of probabilistic information in a real clinical setting*, J. Experim. Psych. : Human Perception and Performance **7**, 928–935.

Clemen, R. T. (1987), *Combining overlapping information*, Manag. Sci. **33**, 373–380.

Clemen, R. T. (1989), *Combining forecasts : a review and annotated bibliography*, Int. J. Forecasting **5**, 559–583.

Clemen, R. T. and Winkler, R. L. (1985), *Limits for the precision and value of information from dependent sources*, Op. Res. **33**, 427–442.

Cooke, R. M. (1991), *Experts in uncertainty : expert opinion and subjective probability in science*, Oxford University Press, New York.

Cooke, R. M., Mendel, M. and Thys, W. (1988), *Calibration and information in expert resolution: a classical approach*, Automatica **24**, 87–94.

Dalkey, N. C. (1972), *An impossibility theorem for group probability functions*, Rand Corporation Memo **P-4862**.

Dalkey, N. C. (1975), *The Delphi Method : Techniques and Applications*, Eds. Linstone, H.A. and Turoff, M., Addison-Wesley, Reading, MA.

Darroch, J. N., Lauritzen, S.L. and Speed, T.P. (1980), *Markov fields and log-linear models for contingency tables*, Ann. Statist. **8**, 522–539.

Dawid, A. P. (1979), *Conditional independence in statistical theory (with discussion)*, J. R. Statist. Soc. B **41**, 1-31.

Dawid, A.P. (1979b), *Some misleading arguments involving conditional independence*, J. R. Statist. Soc. B **41**, 249-252.

Dawid, A. P. (1980), *Conditional independence for statistical operations*, Ann. Statist. **8**, 598–617.

Dawid, A. P. (1982), *The well-calibrated Bayesian*, J. Amer. Statist. Assoc. **77**, 605–610.

Dawid, A. P. (1992), *Prequential analysis, stochastic complexity and Bayesian inference (with discussion)*, Bayesian Statistics (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. M. F. Smith, eds.), vol. 4, Clarendon Press, Oxford, pp. 109–125.

Dawid, A. P. (1996), *Conditional independence*, Encyclopedia of Statistical Sciences (Update), Prepared as an entry, Wiley (to appear).

de Finetti, B. (1937), *Foresight : its logical laws, its subjective sources*, Studies in Subjective Probability (H. E. Kyburg and H. E. Smokler., eds.) (Italian); translated and reprinted (1964), Wiley, New York, pp. 217–225.

de Finetti, B. (1951), *Proceedings 2^{nd} Berkeley Symposium on Math. Statistics and Probability*, Univ. of California Press, pp. 217–225.

de Finetti, B. (1970), *Teoria delle probabilità*, (translated to English as) Theory of Probability. 2 volumes. Wiley, N.Y. (1974, 1975), Giulio Einaudi editore, Turino.

DeGroot, M. H. (1970), *Optimal statistical decisions*, McGraw-Hill.

DeGroot, M. H. (1974), *Reaching a consensus*, J. Amer. Statist. Assoc. **69**, 118–121.

DeGroot, M. H. and Fienberg, S. E. (1982), *Assessing probability assessors: calibration and refinement*, Statistical Decision Theory and Related Topics III (S. S. Gupta and J. O. Berger, eds.), vol. 1, Academic Press, N. Y., pp. 291–314.

DeGroot, M. H. and Fienberg, S. E. (1983), *The comparison and evaluation of forecasters*, The Statistician **32**, 12–22.

DeGroot, M. H. and Mortera, J. (1991), *Optimal linear opinion pools*, Man. Sci. **37**, 546–558.

Dréze, J. H. (1976), *Bayesian regression analysis using Poly-t densities*, CORE Discussion Paper **7613**, Univ. Catholique de Louvain, Heverleé, Belgium.

Faria, A. E. and Smith, J. Q. (1994), *Conditional external Bayesianity in chain graphs*, Tech. Rep. **271**, Dept. of Statistics, Univ. of Warwick.

Faria, A .E. and Souza, R. C. (1995), *A re-evaluation of the quasi-Bayes approach to the linear combination of forecasts*, J. Forecasting **14**, 533–542.

Faria, A. E. and Smith, J. Q. (1996), *Conditional external Bayesianity in decomposable influence diagrams*, Bayesian Statistics (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.), vol. 5, Clarendon Press, Oxford, pp. 551–560.

Faria, A. E. and Smith, J. Q. (1997), *Conditionally externally Bayesian pooling operators in chain graphs*, Ann. Statist. (accepted subject to minor corrections).

Fine, T. L. (1973), *Theories of Probability*, Academic Press, New York.

Freeling, A. N. S. (1981), *Reconciliation of multiple probability assessments*, Org. Behav. Hum. Perform. **28**, 395–414.

French, S. (1980), *Updating of belief in the light of someone else's opinion*, J. R. Statist. Soc. A **143**, 43–48.

French, S. (1981), *Consensus of opinions*, Eur. J. Op. Res. **7**, 332–340.

French, S. (1982), *On the axiomatisation of subjective probabilities*, Theory and Decision **14**, 19–33.

French, S. (1985), *Group consensus probability distributions : a critical survey*, Bayesian Statistics (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. M. F. Smith, eds.), vol. 2, Elsevier, North-Holland, pp. 183–202.

French, S. (1986), *Decision Theory : An Introduction to the Mathematics of Rationality*, Ellis Horwood, Chichester.

French, S. and Bunn, D. W. (1980), *Outranking probabilities and the synthesis of forecasts*, J. Opl. Res. Soc. **31**, 545–551.

French, S., Cooke, R. M. and van Steen, J. F. J. (1989), *The use of expert judgement in risk assessment*, Report for the European Space Agency ESTEC 8051/88/NL/RE(SC).

Frydenberg, M. (1990), *The chain graph Markov property*, Scan. J. Statist. **17**, 333–354.

Gardenfors, P. and Sahlin, N-E. (eds.) (1988), *Decision, Probability and Utility*, Cambridge University Press.

Gargoum, A. S. and Smith, J. Q. (1994), *Approximation schemes for efficient probability propagation in evolving high dimensional Gaussian processes*, Tech. Rep. **266**, Dept. of Statistics, Univ. of Warwick.

Geanokoplos, J. (1992), *Common Knowledge*, J. Economic Perpectives **6**, 53–82.

Geiger, D. and Pearl, J. (1990), *On the logic of causal models*, Uncertainty in artificial intelligence IV (R.D. Schachter, T.S. Levitt, L.N. Kanal and J.F. Lemmer, eds.), North-Holland, Amsterdam, pp. 136–147.

Geiger, D. and Pearl, J. (1993), *Logical and algorithmic properties of conditional independence and graphical models*, Ann. Statist. **21**, 2001–2021.

Genest, C. (1984a), *A characterization theorem for externally Bayesian groups*, Ann. Statist. **12**, 1110–1105.

Genest, C. (1984b), *Pooling operators with the marginalization property*, Can. J. of Statist. **12**, 153–163.

Genest, C. (1984c), *A conflict between two axioms for combining subjective distributions*, J. R. Statist. Soc. B **46**, 403–405.

Genest, C., Weerahandi, S. and Zidek, J.V. (1984), *Aggregating opinions through logarithmic pooling*, Theory and Decision **17**, 61–70.

Genest, C. McConway, K. J. and Schervish, M. J. (1986), *Characterization of externally Bayesian pooling operators*, Ann. Statist. **14**, **No.2**, 487–501.

Genest, C. and Schervish, M. J. (1985), *Modeling expert judgements for Bayesian updating*, Ann. Stat. **13**, 1198–1212.

Genest, C. and Wagner, C. G. (1984), *Further evidence against independence preservation in expert judgement synthesis*, Tech. Rep. **84–10**, Dept. of Statistics, Univ. of Waterloo.

Genest, C. and Zidek, J. V. (1986), *Combining probability distributions: a critique and an annotated bibliography (with discussion)*, Statist. Science **1**, 114–148.

Geramita, A. V. and Seberry, J. (1979), *Orthogonal Designs, Quadratic Forms and Hadamard Matrices*, Marcel Dekker, Inc., New York.

Gibbs, W. (1902), *Elementary Principles of Statistical Mechanics*, Yale University Press, NewHaven, Connecticut.

Gokhale, D. V. and Press, S. J. (1982), *Assessment of a prior distribution for the correlation coefficient in a bivariate normal distribution*, J. R. Statist. Soc. A **145**, 237–249.

Goldstein, M. (1990), *Influence and belief adjustment (with discussion)*, Influence Diagrams, Belief Nets and Decision Analysis, Eds. R.M. Oliver and J.Q. Smith, Wiley, pp. 143–174.

Goldstein, M. and O'Hagan, A. (1996), *Bayes linear sufficiency and systems of expert posterior assessments*, J. R. statist. Soc. B (to appear).

Good, I. J. (1961), *A causal calculus*, Philosophy of Science **11**, 305–318.

Goodman, L. A. (1970), *The multivariate analysis of qualitative data : interaction among multiple classifications*, J. Amer. Statist. Assoc. **65**, 226–256.

Goodman, L. A. (1973), *The analysis of multidimensional contingency tables when some variables are posterior to others. A modified path analysis approach*, Biometrika **60**, 179–192.

Haberman, S. J. (1974), *The Analysis of frequency Data*, University of Chicago Press, Chicago.

Harrison, P. J. and Stevens, C. F. (1976), *Bayesian Forecasting (with discussion)*, J. R. Statist. Soc. B **38**, 205–239.

Hession, E. and McCarthy, E. (1974), *Human performance in assessing subjective probability distributions*, Unpublished Manuscript, Dept. of Business Administration, University College Dublin.

Hewitt, E. and Stromberg, K. (1965 publ Springer), *Real and Abstract Analysis*, New York.

Hogarth, R. M. (1975), *Cognitive processes and the assessment of subjective probability distributions (with discussion)*, J. Amer. Statist. Assoc. **70**, 271–294.

Howard, R. A. and Matheson, J. E. (1981), *Influence diagram*, Readings on the Principles and Applications of decision Analysis (R. A. Howard and J. E. Matheson, eds.), Strategic Decisions Group, Menlo Park, California, vol. II, pp. 719–762.

Isham, V. (1981), *An introduction to spatial point process and Markov random fields*, Intl. Statist. Review **49**, 21–43.

Jensen, F. V. and Jensen, F. (1994), *Optimal junction trees*, Proceedings of the 10th conference on uncertainty in artificial intelligence (R.L. de Mantaras and D. Poole, eds.), Morgan Kaufmann, San Mateo, pp. 360–366.

Jensen, F., Jensen, F. V. and Dittmer, S. L. (1994), *From influence diagrams to junction trees*, Report **R-94-2013**, Dept. of Mathematics and Computer Sciences, Aalborg University.

Jensen, F. V. (1996), *An Introduction to Bayesian Networks*, UCL Press, London.

Kahnneman, D. , Slovik, P. and Tversky, A. (eds.) (1982), *Judgement Under Uncertainty: Heuristics and Biases*, Cambridge University Press.

Kelly, J. S. (1978), *Arrow Impossibility Theorems*, Academic Press, New York.

Kiiveri, H. and Speed, T. P. (1982), *Structural analysis of multivariate data : a review*, Sociological methodology (S. Leinhardt, ed.), Jossey-Bass, San francisco.

Kiiveri, H., Speed, T. P. and Carlin, J. B. (1984), *Recursive causal models*, J. Austral. Statist. Soc. A **36**, 30–52.

Kjærulff, U. (1990), *Graph triangulation – algorithms giving small total state space*, Technical Report **R-90-09**, Aalborg University.

Kjærulff, U. (1992), *Optimal decomposition of probabilistic networks by simulated annealing*, Statistics and Computing **2**, 19–24.

Laddaga, R. (1977), *Lehrer and the consensus proposal*, Synthese **36**, 473–477.

Lauritzen, S.L. (1974), *Sufficiency, prediction and extreme models*, Scand. J. Statist. **1**, 128–134.

Lauritzen, S. L. (1982), *Lectures on contingency tables*, 2nd. ed., Univ. of Aalborg Press, Aalborg, Denmark.

Lauritzen, S. L. (1989), *Mixed graphical Association models*, Scand. J. Statist. **16**, 273–306.

Lauritzen, S. L. (1996), *Graphical Models*, Oxford University Press, Oxford.

Lauritzen, S. L., Speed, T. P. and Vijayan, K. (1984), *Decomposable graphs and hypergraphs*, J. Austral. Math. Soc. A **36**, 12–29.

Lauritzen, S. L. and Wermuth, N. (1984), *Mixed interactions models*, Research Report **R-84-8**, Aalborg University.

Lauritzen, S. L. and Wermuth, N. (1989), *Graphical models for associations between variables, some of which are qualitative and some quantitative*, Ann. Statist. **17**, 31–54.

Lauritzen, S. L. and Spiegelhalter, D. J. (1988), *Computations with Probabilities on Graphical Structures and their Application to Expert Systems*, J. R. Statist. Soc. B **50**, 157–224.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N. and Leimer, H.-G. (1990), *Independence properties of directed Markov fields*, Networks **20**, 491–505.

Lehrer, K. (1975), *Social consensus and rational agnoiology*, Synthese **31**, 141–160.

Lehrer, K. (1976), *When rational disagreement is impossible*, Noûs **10**, 327–332.

Lehrer, K. (1983), *Rationality as weighted averaging*, Synthese **57**, 283–295.

Lehrer, K. and Wagner,C. G. (1983), *Probability amalgamation and the independence issue : a reply to Laddaga*, Synthese **62**, 3–11.

Leimer, H.-G. (1985), *Strongly decomposable graphs and hypergraphs*, PhD Thesis. Berichte zur Stochastik und verwandte Gebiete 85-1, University of Mainz.

Leimer, H.-G. (1989), *Traingulated graphs with marked vertices*, Graph theory in memory of G. A. Dirac (L. D. Andersen, C. Thomassen, B. Toft and P. D. Vestergaard, eds.), Annals of Discrete Mathematics 41, Elsevier Science Publishers B. V. (North-Holland), Amsterdam, pp. 311–324.

Leimer, H.-G. (1993), *Optimal decomposition by clique separators*, Discrete Math. **113**, 99–123.

Lemmer, J. F. (1983), *Generalized Bayesian updating of incompletely specified distributions*, Large Scale Systems **5**, 51–68.

Lichtenstein, S. and Fischhoff, B. (1980a), *Training for calibration*, Org. Behav. Hum. Perform. **28**, 149–171.

Lichtenstein, S. and Fischhoff, B. (1980b), *How well do probability experts assess probability ?*, Decision Research Report **80-5**, Eugene, Oreg. : Decision Research.

Lichtenstein, S., Fischhoff, B. and Phillips, L. D. (1982), *Calibration of probabilities: the state of the art to 1980*, Judgement Under Uncertainty: Heuristic and Biases (D. Kahnneman, P. Slovic and A. Tversky, eds.), Cambridge University Press, pp. 306–334.

Lindley, D. V., Tversky, A. and Brown, R. V. (1979), *On the reconciliation of probability assessments*, J. R. Statist. Soc. A **142**, 146–180.

Lindley, D. V. (1982a), *Scoring rules and the inevitability of probability (with discussion)*, Int. Statist. Rev. **50**, 1–26.

Lindley, D. V. (1982b), *The Bayesian approach to statistics*, Some recent advances in statistics (J. T. de Oliviera and B. Epstein, eds.), Academic Press, London, pp. 65–87.

Lindley, D. V. (1982c), *The improvement of probability judgements*, J. R. Statist. Soc. A **145**, 117–126.

Lindley, D. V. (1983), *Reconciliation of probability distributions*, Oper. Res. **31**, 866–880.

Lindley, D. V. (1985a), *Reconcilition of discrete probability distributions*, Bayesian Statistics (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. M. F. Smith, eds.), vol. 2, Elsevier, North-Holland, pp. 375–390.

Lindley, D.V. (1985b), *Making Decisions*, Jonh Wiley & Sons, New York.

Lindley, D. V. (1987), *Using expert advice on a skew judgemental distribution*, Op. Res. **35**, 716–721.

Madansky, A. (1964), *Externally Bayesian groups*, Rand Corporation Memo RM-4141-PR.

Madansky, A. (1978), *Externally Bayesian groups*, Unpublished manuscript, University of Chicago.

McConway, K. J. (1978), *The combination of experts' opinions in probability assessment : some theoretical considerations*, PhD thesis, University College, London.

McConway, K. J. (1981), *Marginalization and linear opinion pools*, J. Amer. Statist. Assoc. **76**, 410-414.

Mendel, M. B. and Sheridan, T. B. (1987), *Optimal estimation using human experts*, Man-machine Systems Lab., Dept. of Mechanical Engineering, M.I.T..

Menezes, L. and Bunn, D. W. (1990), *Specification of the predictive distribution from a combination of forecasts*, Proc. 16th Symposium on Operational Research, Springer-Verlag, Berlin.

Menezes, L. and Bunn, D. W. (1991), *Diagnostic tracking and model-switching in the use of combined forecasts*, TIMS/SOBRAPO International Meeting, Rio de Janeiro, Brasil.

Montgomery, D. C. (1991), *Design and analysis of experiments*, 3rd edition, John Wiley and Sons.

Morris, P. A. (1974), *Decision analysis expert use*, manag. Sci. **20**, 1233–1241.

Morris, P. A. (1977), *Combining expert judgements : a Bayesian approach*, manag. Sci. **23**, 679–693.

Morris, P. A. (1983), *An axiomatic approach to expert resolution*, manag. Sci. **32**, 303–306.

Moskowitz, H. and Bullers, W. I. (1978), *Modified PERT versus fractile assessment of subjective probability distributions*, Paper **675**, Purdue University.

Murphy, A. H. (1973), *A new vector partition of the probability score*, J. of Applied Meteorology **12**, 595–600.

Newbold, P. and Granger, C. W. J. (1974), *Experience with forecasting univariate time And the combination of forecasts (with discussion)*, J. R. Statist. Soc. A **137**, 131–149.

Oliver, R. M. (1986), *Influence diagrams in forecasting*, Res. Report, Dept. of Industrial Engineering and Operations Research, University of California at Berkeley.

Olmsted, S. M. (1984), *On representing and solving decicion problems*, Ph.D. thesis, Dept. of engineering-Economic Systems, Stanford University, Stanford, California.

Oskamp, S. (1962), *The relationship of clinical experience and training methods to several criteria of clinical prediction*, Psychological Monographs 76 **547**.

Pearl, J. (1986a), *Fusion, propagation and structuring in belief networks*, Artificial Intelligence **29**, 241-288.

Pearl, J. (1986b), *A constraint propagation approach to probabilistic reasoning*, Uncertainty in Artificial Intelligence (L. M. Kanal and J. Lemmer, eds.), North-Holland, Amsterdam, pp. 357–370.

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems*, Morgan Kaufmann, San Mateo.

Pearl, J. (1993), *From Bayeasian networks to causal networks*, Adaptive Computing and Information Processing, vol. I, Unicom Seminars, pp. 165-194.

Pearl, J. and Paz, A. (1987), *Graphoids: a graph based logic for reasoning about relevance relations*, Advances in Artificial Intelligence II (B. D. Boulay, D. Hogg and L. Steel, eds.), North-Holland, Amsterdam, pp. 357–363.

Pearl, J. and Verma, T. (1987), *The logic of representing dependencies by directed graphs*, Proc. of the 6th conf. of Amer. Assoc. of Artif. Intelligence, American Association of Artificial Intelligence, pp. 374–379.

Phillips, L. D. (1987), *On the adequacy of judgmental forecasts*, Judgemental Forecasting (G. Wright and P. Ayton, eds.), John Wiley, Chichester, pp. 11–30.

Pill, J. (1971), *The Delphi method : substance, context, a critique and annotated bibliography*, Socio-Econ. Plan. Sci. **5**, 57–71.

Queen, C. M. and Smith, J. Q. (1993), *Multiregression dynamic models*, J. R. Statist. Soc. B **55**, **No. 4**, 849-870.

Queen, C. M. and Smith, J. Q. (1994), *Bayesian forecasts in markets with overlapping structures*, Int. J. of Forecasting **10**, 209–233.

Raiffa, H. and Schlaifer, R. (1961), *Applied Statistical Decision Theory*, Graduate School of Business Administration, Harvard University, Boston.

Raiffa, H. (1968), *Decision Analisys : Introductory Lectures on Choices under Uncertainty*, Random House, New York, pp. 211-226.

Ramsey, F. P. (1926), *Truth and Probability*, The Foundations of Mathematics and other Logical Essays, 1931, Kegan Paul, London.

Rege, A. and Agogino, A. M. (1986), *Sensor integrated expert system for manufacturing and process diagnostics*, ORC Report **86-15**, University of California, Berkeley.

Roberts, H. V. (1965), *Probabilistic prediction*, J. Amer. Statist. Assoc. **60**, 50–62.

Rose, D. J., Tarjan, R. E. and Lueker, G. S. (1976), *Algorithmic aspects of vertex elimination on graphs*, SIAM Journal of Computing **5**, 266–283.

Rudin, W. (1986), *Real and Complex Analysis*, McGraw-Hill, Singapore.

Savage, L. J. (1954a), *The Foundations of Statistics*, Wiley, New York.

Savage, L. J. (1954), *Probabilistic prediction*, J. Amer. Statist. Assoc. **60**, 50–62.

Savage, L. J. (1971), *Elicitation of personal probabilities and expectations*, J. Amer. Statist. Assoc. **66**, 783–801.

Schmitt, F. F. (1985), *Consensus, respect, and weighted averaging*, Synthese **62**, 25–46.

Seaver, D. A., von Winterfield, D. and Edwards, W. (1978), *Eliciting subjective probability distributions on continuous variables*, Org. Behav. Hum. Perf. **21**, 379–391.

Selvidge, J. (1975), *Experimental comparison of different methods for assessing the extreme probability distributions by the fractile method*, Man. Sci. Rep. Series, Report 75-13, Graduate School of Business Administration, University of Colorado.

Shachter, R. D. (1986a), *Evaluating influence diagrams*, Reliability and Quality Control (A.P. Basu, ed.), Elsevier, North-Holland, pp. 321–344.

Shachter, R. D. (1986b), *Intelligent probabilistic inference*, Uncertainty in artificial intelligence (L.N. Kanal and J. Lemmer, eds.), North-Holland, Amsterdam, pp. 371–382.

Shachter, R. D. (1988), *Probabilistic inference and influence diagrams*, Operations Research **36**, 589–604.

Slovic, P. , Fischhoff, B. and Lichtenstein, S. (1977), *Behavioural decision theory*, Ann. Rev. Psychol. **28**, 1–39.

Smith, A. F. M. and Makov, U. E. (1978), *A quasi-Bayes sequential procedure for mixtures*, J.R. Statist. Soc. B **40**, **No. 1**, 106-112.

Smith, J. Q. (1988), *Belief nets and influence diagrams: what they are and how they work*, Tech. Rep. **155**, Dept. of Statistics, Univ. of Warwick.

Smith, J. Q. (1989), *Influence diagrams for statistical modelling*, Ann. Statist. **17**, 654–672.

Smith, J. Q. (1990), *Statistical principles on graphs*, Influence Diagrams, Belief Nets and Decision Analysis (R.M. Oliver and J.Q. Smith, eds.), Wiley, pp. 89–120.

Smith, J.Q. (1995), *Handling multiple sources of variation using influence diagrams*, Eur. J. of Op. Res. **86**, 189–200.

Smith, J. Q. , French, S. and Ranyard, D. (1993), *An efficient graphical algorithm for updating the estimates of the dispersal of gaseous waste after an accidental release*, Tech. Rep. **254**, Dept. of Statistics, Univ. of Warwick.

Smith, J. Q. and Papamichail, K. N. (1996), *Fast Bayes and the dynamic junction forest*, RODOS(WG1)-TN(96)-02.

Spetzler, C.S. and Staël von Holstein, C.A.S. (1975), *Probability encoding in decision analysis*, Manag. Sci. **22**, 340–358.

Staël von Holstein, C. S. (1970), *Assessment and evaluation of subjective probability distributions*, The Economic Research Institute, Stockholm.

Stone, M. (1961), *The opinion pool*, Ann. Math. Statist. **32**, 1339–1342.

Studený, M. (1989), *Multi-information and the problem of characterization of conditional independence relations*, Problems of Control and Information Theory **18**, 3–16.

Studený, M. (1992), *Conditional independence relations have no finite complete characterization*, Transactions of the 11th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes B, Prague: Academia, pp. 377–396.

Tarjan, R. E. and Yannakakis, M. (1984), *Simple linear-time algorithms to test chordality of graphs, test acyclicity of hypergraphs and selectively reduce acyclic hypergraphs*, SIAM Journal of Computing **13**, 566–579.

Tarjan, R. E. (1985), *Decomposition by clique separators*, Discrete Mathematics **55**, 221–232.

Tversky, A. and Kahneman, D. (1974), *Judgement under uncertainty : heuristic and biases*, Science **135**, 1124–1131.

van Steen, J. F. J. and Oortman Gerlings, P. D. (1989), *Expert opinion in safety studies: volume 2*, Report Min. Housing, Physic. Plan. and Environm., Delft University of Technology and TNO, Netherlands.

Verma, T. (1987), *Causal networks : semantics and expressiveness*, Proc. of the third workshop on uncertainty in artificial intelligence, 352–359.

Verma, T. and Pearl, J. (1990a), *Causal networks : semantics and expressiveness*, Uncertainty in artificial intelligence IV (R.D. Schachter, T.S. Levitt, L.N. Kanal and J.F. Lemmer, eds.), North-Holland, Amsterdam, pp. 69–76.

Verma, T. and Pearl, J. (1990b), *Equivalence and synthesis of causal models*, Proc.of the 6th Conference on Uncertainty in Artificial Intelligence (P. Bonissone, M. Henrion, L.N. Kanal and J.F. Lemmer, eds.), North-Holland, Amsterdam, pp. 255–270.

Von Neumann, J. and Morgenstern, O. (1944), *Theory of Games and Economic Behaviour*, Princeton University Press.

Vorobev, N. N. (1962), *Consistent families of measures and their extensions*, Theory of Probability and its Applications **7**, 147–163.

Wagner, C. G. (1982), *Allocation, Lehrer models and the consensus of probabilities*, Theory and Decision **14**, 207–220.

Wagner, C. G. (1984), *Aggregating Subjective Probabilities: Some Limitative Theorems*, Notre Dame Journal of Formal Logic **25**, 233–240.

Wald, A. (1950), *Statistical Decision Functions*, Wiley, New York.

Weerahandi, S. and Zidek, J. V. (1978), *Pooling prior distributions*, Report **78-34**, Inst. Appl. Math. and Statist., Univ. of British Columbia.

Weerahandi, S. and Zidek, J. V. (1981), *Multi-Bayesian statistical decision theory*, J. R. Statist. Soc. Ser. A **144**, 85–93.

Wermuth, N. (1980), *Linear recursive equations, covariance selection and path analysis*, J. Amer. Statist. Assoc. **75**, 963–972.

Wermuth, N. and Lauritzen, S. L. (1990), *On substantive research hypotheses, conditional independence graphs and graphical chain models*, J. Roy. Statist. Soc. B **52**, 21–50.

West, M. (1989), *Modelling expert opinions*, Bayesian Statistics (J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds.), vol. 3, Oxford University Press, pp. 493–508.

West, M. and Harrison, P. J. (1989), *Bayesian Forecasting and Dynamic Models*, Springer, New York.

Winkler, R. L. (1968), *The consensus of subjective probability distributions*, Manag. Sci. **15**, B61–B75.

Winkler, R. L. (1977), *Rewarding expertise in probability assessment*, In Decision Making and Change in Human Affairs (H. Jugermann and G. de Zeeuw, eds.), D. Reidel, Dordrecht, Netherlands, pp. 127–140.

Winkler, R. L. (1981), *Combining probability distributions from dependent information sources*, Manag. Sci. **27**, 479–488.

Winkler, R. L. and Makridakis, S. (1983), *The combination of forecasts*, J. R. Statist. Soc. A **146**, 150–157.

Winkler, R. L., Clemen, R. T., French, S., Lindley, D. V., Morris, P. A. and Schervish, M. J. (1986), *Expert resolution*, Manag. Sci. **32**, 298–303.

Wiper, M. P. and French, S. (1995), *Combining experts' opinions using a normal-Wishart model*, J. Forecast. **14**, 25–34.

Wiper, M. P. and Pettit, L. I. (1996), *On improving a model for combining experts' forecasts'*, Bayesian Statistics (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.), vol. 5, Clarendon Press, Oxford, pp. 809–813.

Wright, S. (1921), *Correlation and causation*, Journal of Agricultural Research **20**, 557–585.

Wright, S. (1923), *The theory of path coefficients : a reply to Niles' criticism*, Genetics **8**, 239–255.

Wright, S. (1934), *The method of path coefficients*, Annals of Mathematical Statistics **5**, 161–215.

Wright, G. and Ayton, P. (eds.) (1987), *Judgemental Forecasting*, John Wiley, Chichester.

Wold, H. D. A. (1954), *Causality and econometrics*, Econometrica **22**, 162–177.

Wold, H. D. A. (1960), *A generalization of causal chain models*, Econometrica **28**, 443–463.