Bayesian Analysis in Expert Systems

David J. Spiegelhalter, A. Philip Dawid, Steffen L. Lauritzen, Robert G. Cowell

Statistical Science, Volume 8, Issue 3 (Aug., 1993), 219-247.

Your use of the JSTOR database indicates your acceptance of JSTOR's Terms and Conditions of Use. A copy of JSTOR's Terms and Conditions of Use is available at http://www.jstor.org/about/terms.html, by contacting JSTOR at jstor-info@umich.edu, or by calling JSTOR at (888)388-3574, (734)998-9101 or (FAX) (734)998-9113. No part of a JSTOR transmission may be copied, downloaded, stored, further transmitted, transferred, distributed, altered, or otherwise used, in any form or by any means, except: (1) one stored electronic and one paper copy of any article solely for your personal, non-commercial use, or (2) with prior written permission of JSTOR and the publisher of the article or other text.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

Statistical Science is published by Institute of Mathematical Statistics. Please contact the publisher for further permissions regarding the use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/ims.html.

Statistical Science
©1993 Institute of Mathematical Statistics

JSTOR and the JSTOR logo are trademarks of JSTOR, and are Registered in the U.S. Patent and Trademark Office. For more information on JSTOR contact jstor-info@umich.edu.

©1999 JSTOR

http://www.jstor.org/
Wed Feb 3 16:16:11 EST 1999

Statistical Science
1993, Vol. 8, No. 3, 219–283

Bayesian Analysis in Expert Systems

David J. Spiegelhalter, A. Philip Dawid, Steffen L. Lauritzen and Robert G. Cowell

Abstract. We review recent developments in applying Bayesian probabilistic and statistical ideas to expert systems. Using a real, moderately complex, medical example we illustrate how qualitative and quantitative knowledge can be represented within a directed graphical model, generally known as a *belief network* in this context. Exact probabilistic inference on individual cases is possible using a general propagation procedure. When data on a series of cases are available, Bayesian statistical techniques can be used for updating the original subjective quantitative inputs, and we present a set of diagnostics for identifying conflicts between the data and the prior specification. A model comparison procedure is explored, and a number of links made with mainstream statistical methods. Details are given on the use of Dirichlet prior distributions for learning about parameters and the process of transforming the original graphical model to a *junction tree* as the basis for efficient computation.

Key words and phrases: Graphical models, subjective probability, conditional independence, local computation, triangulation, junction tree, unsupervised learning, Dirichlet distribution, Bayes factors, prequential analysis, prediction, monitors.

1. INTRODUCTION

The work reviewed in this paper represents the synthesis of two important developments in the modelling of complex stochastic phenomena: first, the introduction of formal probabilistic and statistical methodology into the area of applied artificial intelligence known as *expert systems* and second, the use of a pictorial representation of conditional independence assumptions known as *graphical modelling*. To understand why these two strands have come together, it is useful to examine briefly the short but eventful history of research on uncertainty management in expert systems.

Although no agreed upon definition exists, the term *expert system* is generally applied to a computer program that is able to give some sort of reasoned guidance on a fairly tightly delineated problem. The boundaries between engineering, decision-science and artificial intelligence (AI) become somewhat blurred at this point. Traditionally expert systems have been seen as a branch of AI, for which a central tenet has been that of *symbolic reasoning*; logic was used as a tool for representing knowledge and solving problems, and such qualitative reasoning was seen as the antithesis of quantitative methods using techniques such as differential equations. When a fully deterministic representation was unreasonable, attention focussed on qualitative ways of handling uncertainty such as nonmonotonic logic (Reiter, 1987) or novel numerical schemes such as fuzzy logic (Zadeh, 1983), certainty factors (Shortliffe and Buchanan, 1975) or Shafer–Dempster belief functions (Gordon and Shortliffe, 1985). Probability theory was held to be epistemologically inadequate and computationally infeasible. The latter claim arose from two realisations. First, that complex applications would require the specification of huge joint distributions. Second, that what was known as "evidence propagation" within a logical framework would require the efficient computation of probabilities of certain events of interest, conditional on arbitrary configurations of other variables which constituted the observed evidence. See the papers collected by Shafer and Pearl (1990) for a wide-ranging discussion of these issues.

While research into the alternative representations of uncertainty continued to be vigorously pursued,

David J. Spiegelhalter is Senior Statistician, MRC Biostatistics Unit, Institute of Public Health, Robinson Way, Cambridge CB2 2SR, United Kingdom. A. Philip Dawid is Professor and Robert G. Cowell is Professor, Department of Statistical Science, University College London, Gower Street, London WC1E 6BT, United Kingdom. Steffen L. Lauritzen is Professor, Department of Mathematics and Computer Science, Institute for Electronic Systems, Fredrik Bajers Vej 7, DK-9220 Aalborg O, Denmark.

probability theory made a comeback. Naive methods for calculating marginal and conditional distributions were clearly impractical, but it was perceived that exploitation of conditional independence assumptions, implicit in the qualitative structure of the expert knowledge, might reduce the problem of specification and evidence propagation to a feasible level. It was noted that evidence propagation techniques within alternative formalisms clearly made use of some form of implicit conditional independence in their use of "local" propagation schemes.

One branch of statistical modelling had emphasised such qualitative structure over quantitative specification and further had provided a pictorial scheme that directly translated into conditional independence statements. Models using directed graphs had been introduced by Wright (1934) in the context of path analysis, while Darroch, Lauritzen and Speed (1980) had represented particular classes of log-linear models by undirected graphs; they also took the vital step of exploiting the representation to use graph-theoretic proofs of probabilistic properties of particular models. As Wermuth and Lauritzen (1983) were exploring the relationship between undirected and directed graphs, Pearl (1982) and Kim and Pearl (1983) were introducing simple computational schemes that showed probabilistic reasoning was quite feasible when the computations were governed by the conditional independencies expressed by particular simple directed graphical structures. Remarkably similar schemes had previously been developed by mathematical geneticists working in pedigree analysis (Cannings, Thompson and Skolnick, 1978), although the connections with this area were not to be made for another decade.

Contributions then came from two apparently unrelated areas of work. First, the issue of *triangulation* of undirected graphs had been carefully studied in the context of relational data bases (Tarjan and Yannakakis, 1984) and was found to be essential in reducing the problem faced with general graphical structures to one of *local computation* between communicating entities. Second, the axiomatic approach to conditional independence (Dawid, 1979a) was used to show that much of the proposed scheme extended far beyond probabilistic reasoning, and that the basic ideas, and hence even the same software, could be used for handling the forms of nonprobabilistic reasoning that had been developing in parallel. Thus the challenge of introducing rigorous probabilistic methodology has had the exciting consequence of providing a unifying framework for all forms of reasoning that exploit conditional independence.

In this paper we shall concentrate on new developments since Lauritzen and Spiegelhalter (1988), emphasizing statistical rather than expert system issues: for tutorial introductions to the basic ideas, see Pearl (1988), Neapolitan (1990) or Henrion, Breese and Horvitz (1991). In the next section the construction of directed graphical models is explored informally, with particular reference to an application in the diagnosis of congenital heart disease. Section 3 discusses how such a model might be used as a basis for an expert system, and the algorithm used for evidence propagation is displayed pictorially using a preliminary version of a network representing a set of congenital conditions leading to a "blue" baby. Aspects such as "explanation" and software are also covered. We then deal with the statistical, rather than probabilistic, aspects. Section 4 describes how we can use accumulating data both to revise the initial quantitative inputs and to provide diagnostic checks on the quality of any proposed model, and the choice between alternative qualitative structures is covered in Section 5.

Three more technical sections follow, intended for those interested in investigating these procedures in some detail. Section 6 discusses the conditional independence properties of directed graphical models, and how such models can be transformed into an undirected form more suitable for computation. The resulting general algorithm underlies the propagation procedure shown in Section 3. Section 7 covers the use of Dirichlet prior distributions when carrying out batch or sequential parameter estimation, and Section 8 deals with the issue of ensuring priors within alternative qualitative structures cohere in a reasonable way. Finally we discuss the trend of this work, emphasising the increasing links with mainstream statistical modelling.

2. MODELLING THE DOMAIN

We can divide the construction of a model into three distinct components. The first *qualitative* stage considers only general relationships between the variables of interest, in terms of the *relevance* of one variable to another under specified circumstances. This naturally leads to a graphical representation of conditional independence, but one that is not restricted to a probabilistic interpretation. The next *probabilistic* stage introduces the idea of a joint distribution defined on the variables in the model and relates the form of this joint distribution to the structure of the graph. The final *quantitative* step requires the numerical specification of the necessary conditional probability distributions.

2.1 Qualitative Modelling

In this paper we shall use a running example that forms part of a study with the Great Ormond Street Hospital for Sick Children in London. The hospital (here abbreviated to GOS) acts as a referral centre for

newborn babies with congenital heart disease, and since early appropriate treatment is essential, a preliminary diagnosis must be reached using information reported over the telephone; this data may concern clinical signs, blood gases, electrocardiogram (ECG) and x-ray. An algorithm to help the junior doctor at GOS has already been developed and evaluated on 400 cases (Franklin et al., 1991). The results suggested that use of a formal decision aid could be of substantial benefit, with the algorithm having a diagnostic accuracy of 76%, compared with 64% for the doctor in GOS and 45% for the referring paediatrician.

This algorithmic formation is attractive in its simplicity and transparency, but suffers from problems of observer variability and missing data. There is strong interest in developing a probabilistic system that will be more forgiving of limitations in the data but also can exploit the available accumulated data on nearly 600 babies. Such a probabilistic system needs to be based on considering the true disease and possible clinical findings as a set of random variables and requires the specification of a full joint distribution over these variables to represent clinical understanding of the disease process. It is tasks of this nature that have been considered challenging to constructors of expert systems.

The framework of a *graphical model* allows experts to concentrate on building up the qualitative structure of a problem, before even beginning to address issues of quantitative specification. As emphasised by Pearl (1988), such models are intended to encode natural judgements of relevance and irrelevance and can be formed prior to any probabilistic considerations. Nodes in the graph represent variables; missing links in the graph represent the irrelevance properties. Loosely, a directed edge is put between two variables to represent a direct influence. To avoid inconsistencies, we must not permit a sequence of directed edges which returns to its starting node: the graph is thus a *direct acyclic graph* or DAG. (Generalisations allow undirected edges to represent associations that cannot be explained by introducing a common "cause"; see Section 9.)

For any node, once the direct influences on it are known, all other potential influences become irrelevant. Such irrelevance judgements may be made intuitively, taking account of one's understanding of causal structure; they do not require probabilistic modelling (although, in its presence, they may be represented as assertions of probabilistic conditional independence between a node and its remaining potential influences, given its direct influences). Terms such as *influence diagrams, causal networks* and *relevance diagrams* are sometimes used to describe such graphical representations: we shall use directed *graphical model* and *belief network* interchangeably, and we will refer to a *probabilistic expert system* when such a structure is used for diagnosis.

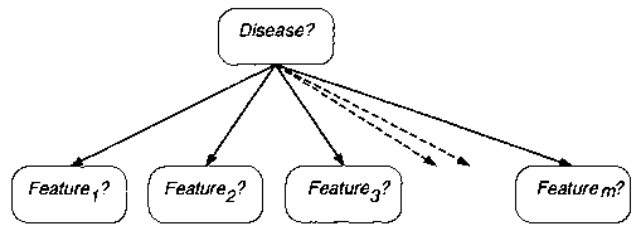

FIG. 1. *Directed graphical model representing conditional independence of feature variables within each disease class – the "idiot's Bayes" model.*

As a very simple example, Figure 1 may be regarded as expressing the view that, once the disease class is known, information about one set of feature variables is of no further relevance to predicting the values of some other disjoint set. Initial experiments with such a naive network were disappointing (Franklin et al., 1989), and a "deeper" model appears particularly appropriate in a domain such as congenital heart disease in which the basic physiological mechanisms are well understood.

Figure 2 shows a preliminary network for part of the spectrum of diseases, specifically the node *Disease?* includes six possible conditions, assumed mutually exclusive and exhaustive, that lead to particularly "blue" babies; its elicitation is described later. We shall call this particular model the CHILD network. The graph represents, for example, that the level of oxygen in the lower body (node 16) is directly related to the underlying level when breathing oxygen (node 11) and whether the hypoxia is equally distributed around the body (node 10). In turn, the level when breathing oxygen depends on the degree of mixing of the blood in the heart (node 6) and the state of the blood vessels (parenchyma) in the lungs (node 7). It is these intermediate variables that are directly influenced by the underlying disease (node 2). When we reach the next stage of assessing the detailed probabilistic structure associated with a graphical model, such qualitative judgements translate into algebraic properties of the overall joint distribution.

The aim of the CHILD network is to provide a mechanism so that both clinical expertise and available data can be properly exploited to produce a reasonably transparent diagnostic aid. In practice, however good the diagnostic ability of such a system, implementation would be severely limited by the lack of appropriate computing facilities currently available. Experience suggests that systems will only be used when they form part of an established clinical information system, such as planned for GOS hospital. In the meantime the CHILD system is intended to serve as a demonstra-

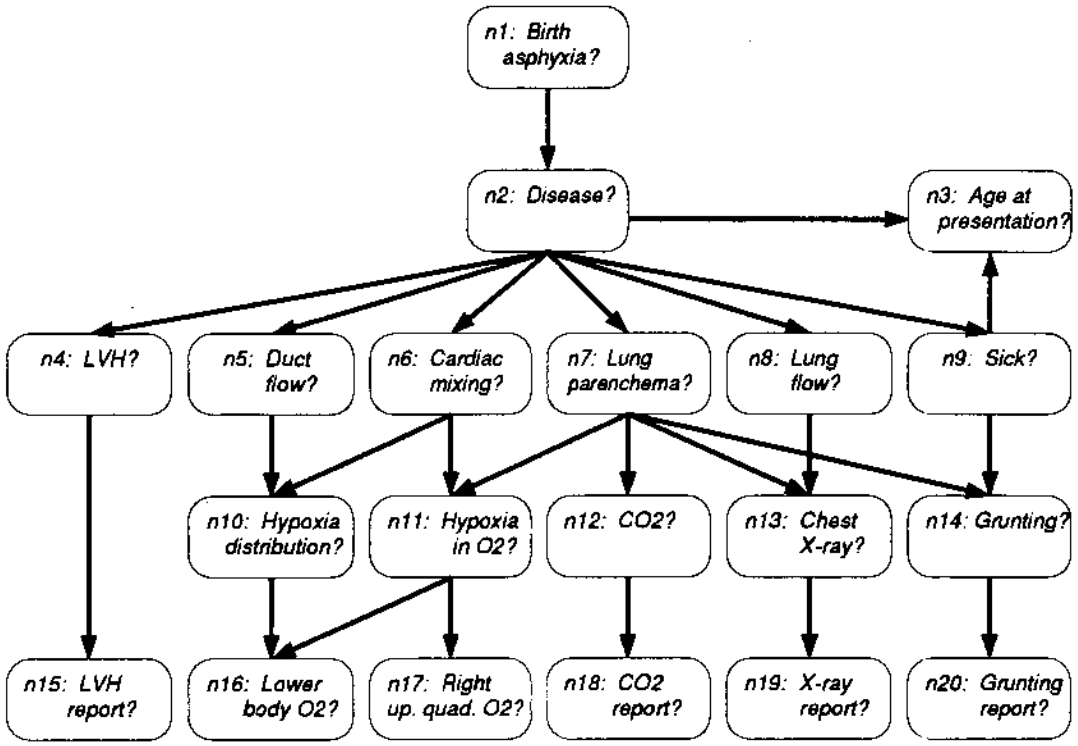

FIG. 2. *Directed acyclic graph representing the incidence and presentation of six possible diseases that would lead to a "blue" baby. LVH, left ventricular hypertrophy.*

tor of the technical possibilities and possibly as an educational tool.

2.2 Probabilistic Modelling

A probabilistic expert system functions by providing a representation of the joint distribution for all the variables, with underlying algorithms that allow fast calculation of the distribution for any node, conditional on any configuration of observed data. We therefore need to relate the qualitative structure described above to a formal expression for a joint distribution: the computational schemes for calculating conditional probabilities will be displayed in Section 3 and discussed in detail in Section 6.

Let the nodes in the graph represent a set of discrete random variables X_v, $v \in V$. The state space for each of the variables in the CHILD network can be seen in Figure 3. For $A \subseteq V$, we let X_A denote $\{X_v, v \in A\}$, although generally we rather loosely use v to stand for X_v in formulae. If we interpret the general idea of irrelevance in terms of probabilistic conditional independence, then the directed acyclic graph is a pictorial means of specifying the formal assumption that the joint distribution of X_V can be expressed as a product of the conditional distributions of each node given its direct influences (parents) in the graph. Hence, letting pa(v) denote the parents of node v, the graph implies that the joint distribution $p(V)$ has the form

$$p(V) = \prod_{v \in V} p(v|\text{pa}(v)) \tag{1}$$

This is also known as a *recursive* model with respect to a DAG $\mathfrak{D}$. Thus, for example, the model for Figure 1 is equivalent to

$$p(Disease, Feature_1, \ldots, Feature_m) = p(Disease) \prod_{i=1}^{m} p(Feature_i|Disease).$$

This joint distribution requires as numerical inputs only the prior distribution over the diseases and the distribution of each of the features in each of the disease categories. Calculating the posterior probability of each disease on the basis of observed findings is extremely straightforward: this simple model has been termed naive or even idiot's Bayes (Titterington et al., 1981).

In discussing the precise conditional independence properties that such a factorisation of the joint density implies, it is first helpful to extend the use of "parent" to more distant relations such as "ancestor" and "descendant"—the natural use of this language reflects the fact that pedigree analysis and genetic counselling provide some of the best examples of the use of conditional independence graphs. The factorisation (1) then implies that, given the values of the parents of a node v, X_v is independent of all other nodes in the graph that are not descendants of v. In the CHILD network we have, for example, that given values of *n7: Lung*

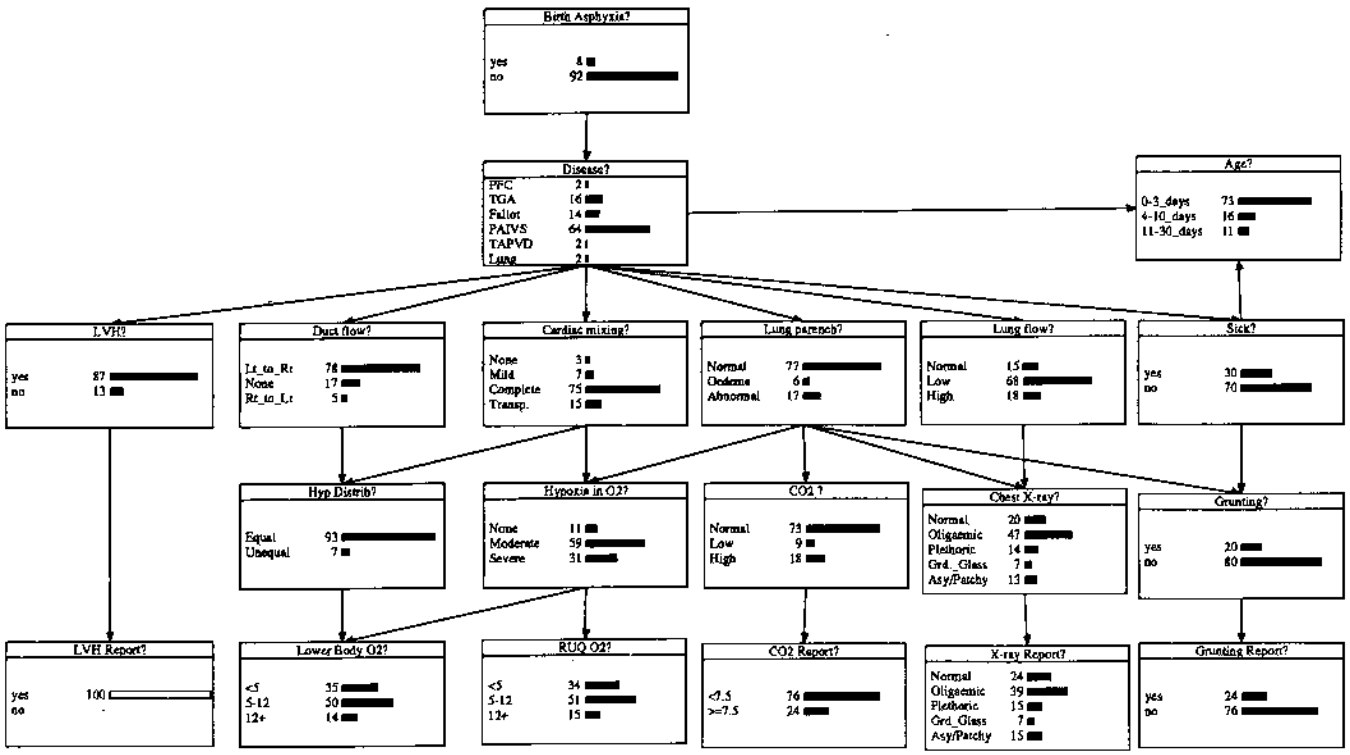

FIG. 3. *Conditional probability distributions on all nodes after propagation of evidence LVH-report = yes. The numbers and the length of the bars represent the current probability: for example, 64% belief that PAIVS is the true diagnosis, compared to a prior 22% belief. For observed evidence, that is, LVH-report = yes, the bar is hollow.*

parenchyma? and *n8: Lung flow?*, the node *n13: Chest x-ray?* is independent of all other nodes except *n19: X-ray report?* Pearl (1986) described an exhaustive set of conditional independence relations on a DAG $\mathcal{D}$ in terms of the concept of *d-separation.* However, in any particular instance this condition is often difficult to verify, and a simpler tool was derived by Lauritzen et al. (1990). To use this we need to consider undirected graphs, and this step is discussed in Section 6.

2.3 Quantitative Modelling

The preceding discussion implies that the expert must also provide sufficient conditional probabilities to specify fully the joint distribution. For example, Table 1 gives the assessments for the *Disease?* → *LVH?* link and the *LVH?* → *LVH-report?* link: the elicitation of these assessments is discussed later. The judgements show that LVH (left ventricular hypertrophy found on an ECG) is essentially a feature of PAIVS, although on fairly rare occasions it can appear with other conditions, and that there is an estimated 5% false positive and a 10% false negative rate in reporting. Criticism and revision of these quantitative assumptions are dealt with in Section 4. In total, for the 20 variables, 114 distributions were assessed for a mean of 3 states each, requiring the specification of 230 independent numerical assessments.

There has been research into incomplete specification of the full joint distribution, say using upper and lower probabilities (Walley, 1990), in which resulting probabilities may be computed only up to an interval using linear or nonlinear programming techniques (van der Gaag, 1991). Alternatively, if the number of assess-

TABLE 1
Subjective assessments of conditional probability tables assessed by expert for links n2 → n4 and n4 → n15

	n4: LVH?	
n2: Disease?	Yes	No
PFC	0.10	0.90
TGA	0.10	0.90
Fallot	0.10	0.90
PAIVS	0.90	0.10
TAPVD	0.05	0.95
Lung	0.10	0.90

	n15: LVH-report?	
n4: LVH?	Yes	No
Yes	0.90	0.10
No	0.05	0.95

Diseases are persistent foetal circulation (PFC), transposition of the great arteries (TGA), tetralogy of Fallot, pulmonary atresia with intact ventricular septum (PAIVS), obstructed total anomalous pulmonary venous connection (TAPVD) and lung disease.

ments made is insufficient to specify a joint distribution uniquely, it has been suggested that the distribution be completed by maximum entropy arguments (Nilsson, 1986).

2.4 Practical Issues in Model Specification

2.4.1 *The CHILD network*

The network in Figure 2 is adapted from one shown in Spiegelhalter and Cowell (1992), which was elicited using the technique of *similarity graphs* described by Heckerman (1990). This procedure is useful when there is a (possibly very large) set of diseases to discriminate. The expert first identifies sets of diseases that are *similar* and hence typically constitute a diagnostic problem – such a subset is known as a *differential diagnosis*. The directed graphical structure appropriate to each differential diagnosis is then elicited, by first asking the clinician about underlying physiological features that distinguish between the relevant diseases and then exploring the clinical consequences of those features. Links are informally interpreted as *direct causation*, which clinicians find a useful and intuitive idea, and absent links are confirmed by asking whether knowledge of the parent nodes does render other nondescendants irrelevant. By an iterative process a graph is built up for each differential diagnosis, and these graphs are finally superimposed. Software is available for this exercise (Heckerman, 1990), although the CHILD network was elicited on paper.

A series of changes has been made in response to the analysis of fictitious cases and scrutiny of the model. First, in the original network a link existed between $n10$ and $n17$: this not only is unrealistic, since hypoxia distribution should not influence upper body oxygen, but it was also found that the assessed conditional probabilities did not actually show a quantitative dependence. Second, *Age at presentation?* was originally a parent of *Disease?*, since it was felt fairly natural to think of different incidence rates for different ages at presentation. After discussion, this link was reversed and a dependence on *Sick?* introduced, as an attempt to model the referral behaviour of distant paediatricians. This issue of modelling the selection process deserves deeper analysis – ideally the network should reflect unselected cases, and a node *Case selected?* introduced as a child of variables that would influence the decision to refer. Conditioning on the case being selected would then give the appropriate joint distribution over the remaining nodes.

Finally, some conditional probability tables were adjusted to give the system reasonable behaviour when "archetypal" cases were put through the system. Of the 114 distributions, 13 (11%) were altered at this stage.

We note that, apart from the somewhat anomalous *Age at presentation?*, the graph naturally falls into a set of five "blocks," representing, respectively, risk factors, underlying physiological anomaly (disease), physiological disturbances caused by the disease, clinical signs as elicited by specialist staff in GOS and reported signs as obtained over the telephone from a paediatrician who is not a specialist in cardiology. Many other clinical problems appear to follow a similar structure, although sometimes the direction of an arrow is not clear. For example, birth asphyxia is a possible *cause* of PFC (persistent foetal circulation), but a possible *effect* of the other diseases. This suggests an undirected link (see Section 9).

Subjective probabilities of the type shown in Table 1 may be elicited using a range of standard techniques. If we view these as estimates of frequencies, then it is natural to place some imprecision on these assessments – this not only makes the expert feel less threatened but also provides a basis for empirical learning (see Section 4). It is helpful if the expert is familiar with thinking in terms of frequencies of events.

2.4.2 *Applications and software*

Historically, the idiot's Bayes model in Figure 1 was first used by Warner et al. (1961), coincidentally for the diagnosis of congenital heart disease. Later applications of this model are too numerous to list, but a notable example is the acute abdominal pain system dating from de Dombal et al. (1972), that has been implemented in a number of hospitals and remote sites such as submarines and is claimed to have a significant impact on care and resources (Adams et al., 1986).

It has long been argued that in most applications the assumptions underlying such a model are blatantly inappropriate, but only recently have computational techniques and computing environments allowed more realistic representations of substantive knowledge: Henrion, Breese and Horvitz (1991) give the background to a number of large applications. Although many of these implementations are at an early stage, some extremely challenging problems are being tackled: a reconstruction of the QMR/INTERNIST system (Miller, Pople and Myers, 1982) as a probabilistic model involves 4,500 nodes and over 40,000 links (Shwe et al., 1991), while the MUNIN network representing part of the muscles and nerves necessary for interpreting electromyographic data (Andreassen et al., 1987) already has over 1,000 nodes each with up to 27 states. The PATHFINDER system for the diagnosis of lymph node pathology concerns over 60 diseases and required the specification of over 75,000 subjective probabilities (Heckerman, Horvitz and Nathwani, 1992); it has been successfully converted to a commercial system, INTELLIPATH.

In such large systems there are so many numerical assessments required that it is unreasonable to expect

each to be individually specified. A variety of simplifying models for the conditional probabilities has been exploited: these include assuming an underlying continuous mathematical model (MUNIN), two-stage assessments in which large sets of conditional probabilities are first assessed to be equal and then a single common value elicited (PATHFINDER) and "noisy-or gates" (QMR). In QMR, up to 80 diseases may be parents of a symptom, but it is assumed that any single one of the diseases is sufficient to cause the symptom, and this causation occurs independently of other diseases present. The conditional distribution in this "competing-risk" model then only requires the specification of as many parameters as there are parents.

A number of distinct application areas share representational and computational issues. The first area is broadly *diagnostic*, which includes the medical examples discussed above, as well as causality assessment in drug safety (Spiegelhalter et al., 1991a), legal reasoning, forensic science and fault diagnosis in complex systems. The second area is essentially *spatial*, in which networks can be constructed for high-level vision problems (Jensen, Christensen and Nielson, 1992). Third, *dynamic* problems, involving prediction, monitoring and smoothing, seem ideally suited to this approach (Kjærulff, 1992b): examples that involve dynamic restructuring of the graph include plan recognition in text understanding (Charniak and Goldman, 1989) and monitoring in drug therapy (Berzuini et al., 1992).

Increasingly developments are being made either on commercial tools or programs available for research, all of which use a version of the algorithm described in Section 6. Commercial programs include HUGIN (Hugin Expert Ltd) and ERGO (Noetic Systems Inc), while freely available software includes BAIES (Cowell, 1992) and IDEAL (Srinivas and Breese, 1990).

3. BAYESIAN NETWORKS AS EXPERT SYSTEMS

3.1 Principle of Local Computation

Once the model has been established, the tasks become computational. Thus in the CHILD problem, we might obtain information on a new case that *Lower body* $O_2 < 5$ and *X-ray report* = plethoric and wish to know what we can deduce about the disease. This question could be answered, in principle, by calculating the conditional probabilities $p(Disease? \mid Lower\ body\ O_2 < 5, X\text{-}ray\ report = \text{plethoric})$ generated from the full multivariate distribution for all the variables, which is implicit in our graphical description and quantitative inputs. However, the state space of all possible configurations of values for all variables has dimensionality 20, there being more than a billion such configurations altogether. A naive approach to calculating the required conditional probabilities would require, first the explicit computation of the probability of each configuration, and then the construction, for each disease d, of a ratio in which the numerator summed these probabilities over the approximately 11 million configurations for which disease $= d$, lower body O_2 is < 5 and the X-ray report is plethoric, while the denominator further summed these answers over d. Calculations such as these can rapidly outrun the capabilities of large computers. A natural question is whether it might not be possible to exploit the qualitative structure expressed in the graphical model so as to simplify and streamline such computations.

A crucial feature of a graphical model is that it describes a joint distribution as built up out of local relationships within groups of variables – such as a node and its parents. This suggests a general strategy of "divide and conquer" whereby, instead of tackling the whole collection of variables simultaneously, we seek first to break it down into subgroups – which we may call *belief universes* (Jensen, Olesen and Andersen, 1990) – in such a way that the naive computations described above need only be performed within each belief universe. If we can do such a breakdown so as to obtain belief universes which are relatively small, the calculations then become manageable. Additionally, to ensure that we obtain the correct answers when considering all the variables together, we need to develop ways for the belief universes to communicate with each other, so that (for example) the effect of conditioning on a variable in one universe can be felt by those in another.

While – as we shall see – the above strategy is indeed implementable, this turns out to be somewhat less straightforward in general than might be expected, although Pearl (1986) describes some efficient techniques for tree-structured graphs. In particular, the appropriate belief universes are not always easily identified from a brief inspection of the graphical structure. Instead, a fairly complex chain of transformations of both the qualitative and the quantitative inputs is needed: a process which we may term "compilation." Although such compilation can be computationally demanding, it only needs to be performed once: after we have identified the appropriate belief universes and the ways in which they are to communicate with each other, all probabilistic calculations can then make use of this compiled structure to minimise the computational burden.

In Section 6 we describe in detail the process of identifying and organising the belief universes, initializing the quantities held on those universes and algorithms for propagating the effects of observed evidence. All these procedures are invisible to the user of a system, and so here we jump to an example of the use of a prototype system implemented on available software.

3.2 Propagation Example on CHILD

Figures 3–5 show a sequence of screen dumps, taken from the HUGIN system (Andersen et al., 1989), illustrating the propagation algorithm in use on CHILD. Figure 3 shows the status of the network after observing the evidence *LVH-report* = yes. After two more findings have been added, *X-ray report* = oligaemic and *Lower body* $O_2 < 5$, the posterior probability that the disease is PAIVS is 0.76 (Figure 4). The status of the intermediate nodes gives some explanation of the findings: the X-ray shows oligaemia because of the low lung flow, although the lung vessels appear normal. There is likely to be complete cardiac mixing, with left-to-right (aorta-to-pulmonary artery) flow in the arterial duct providing the only source of blood to the lungs. Such depiction of the effects of changing evidence are particularly valuable for rapid sensitivity analysis.

3.3 Expert System Aspects

Here we only give an indication of developments related to handling particular cases, which is of prime importance if systems are to have a practical role. For further details we refer to the series of proceedings for the workshops on Uncertainty in Artificial Intelligence which have been held annually since 1985, for a wide range of case studies and ideas.

First, *explanation* facilities have received attention both with regard to quantitative techniques, primarily based on weights of evidence [see, e.g., Heckerman, Horvitz and Nathwani (1992) for a description within PATHFINDER], and qualitative methods (Sember and Zuckerman, 1989). Pearl (1988) described how an adjustment to the propagation algorithm could lead to straightforward identification of the joint configuration of the variable with the highest probability. This has been found to be an attractive feature, since it can be thought of as providing a plausible explanation for the observed findings. Figure 5 shows the implementation of this idea within HUGIN, based on the algorithm described in subsection 6.6. The node bars display normalized conditional "profile probabilities," that is, the maximal obtainable probabilities, compatible with the given state and conditional on the evidence, normalized to have maximum 100. The configuration with highest probability can be read off by picking out the bars of length 100.

A related topic is the *sensitivity* of the probabilistic conclusion to both additional findings and imprecision in the parameter estimates. Spiegelhalter (1989) suggested a unified approach, which requires a decomposition of the total variance of the predictive distribution of the posterior probability of interest. Sensitivity leads naturally into *selection of questions,* which has again been discussed at length in the PATHFINDER project. Finally, *conflict* in evidence is an extremely useful finding in that it may lead to doubt about the

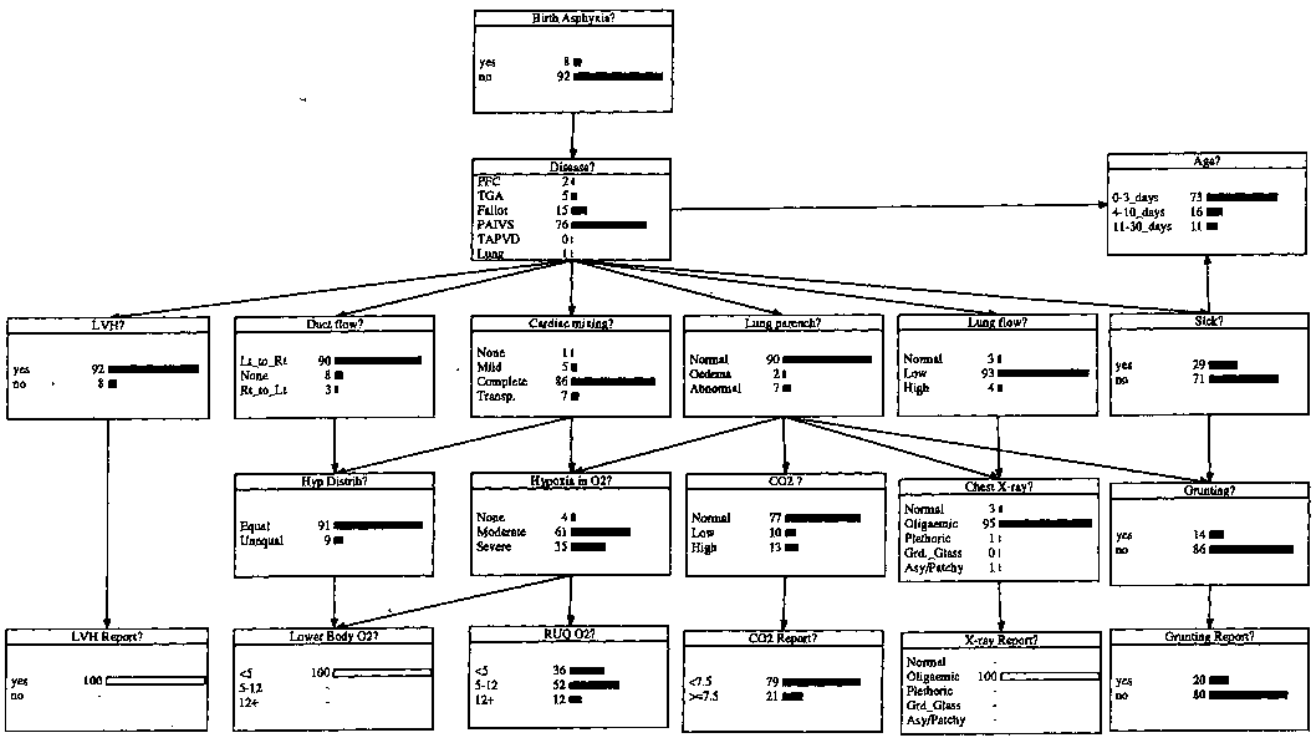

FIG. 4. *Status after propagation of additional evidence X-ray report = oligaemic and Lower body* $O_2 < 5$.

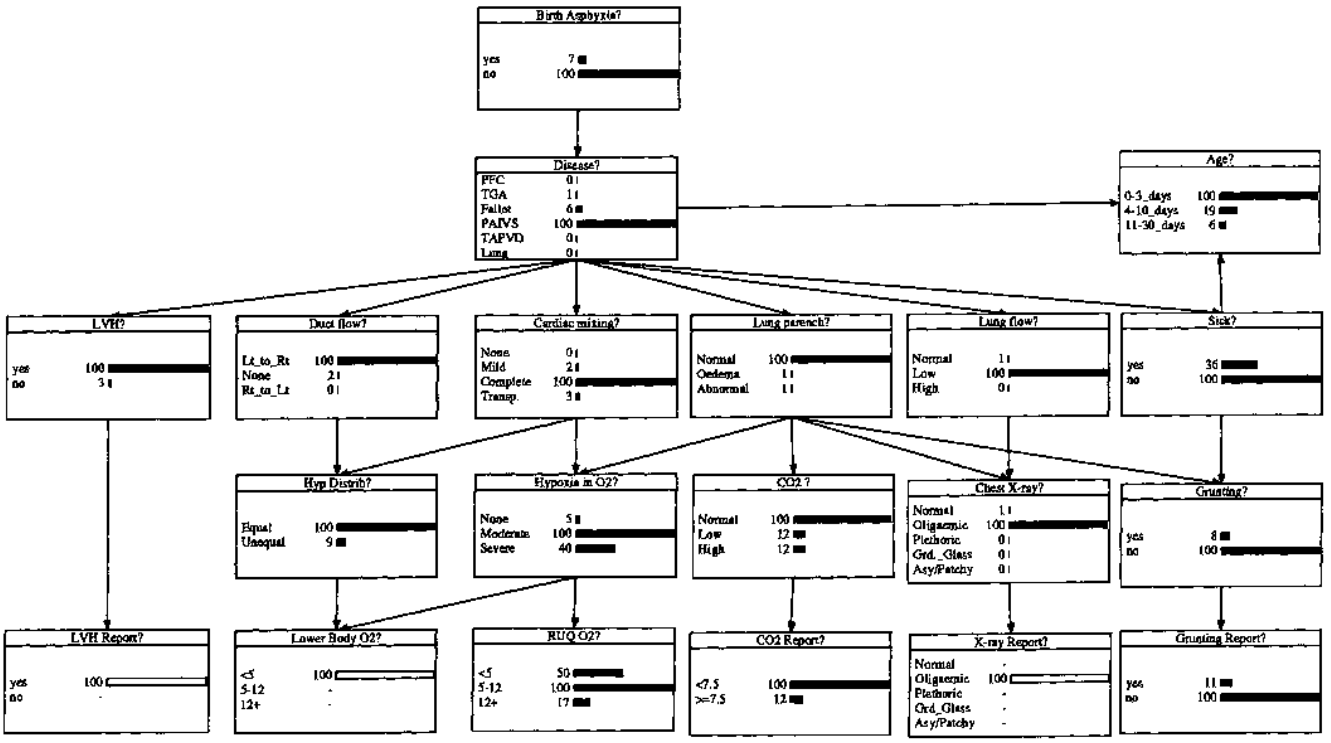

FIG. 5. *Profile probabilities of variables after entering evidence, identifying the most likely configuration. The probabilities have been normalised by the probability of this configuration which is* 0.0029.

conclusions of the system, as would be appropriate were a case with a bizarre combination of features to be observed. Jensen et al. (1991) show the value of identifying the source of conflict with relation to the junction tree; this being just one use of the normalisation factors that fall out of the propagation algorithm (Jensen, 1991).

4. USING DATA TO REFINE MODELS

4.1 Random Probabilities and Learning from Experience

Up to now we have made the strong assumption that all conditional probability distributions are precisely specified. This is clearly unrealistic, regardless of whether the distribution is derived from analysis of data or from subjective assessment, and in this section we relax this assumption and allow the conditional probabilities themselves to be unknown quantities. Initially specified distributions over these quantities can then be updated as data on patients accumulate. Hence we are extending the discussion from Bayesian probabilistic reasoning to Bayesian statistical reasoning. In this section we only deal with a simple situation in which complete data are observed. Section 7 deals with the general situation.

Spiegelhalter and Lauritzen (1990) introduce the natural extension of considering the conditional probabilities of the system as being generated by parameters θ_v, which are components of an overall parameterisation θ. Thus (1) becomes

$$p(V|\theta) = \prod_{v \in V} p(v|\mathrm{pa}(v), \theta_v) .$$

An attractive assumption is that of *global independence*, that is, the parameters $\{\theta_v, v \in V\}$ are assumed a priori independent random variables and so $p(\theta) = \Pi_v p(\theta_v)$. This assumption leads to the joint distribution of case-variables V and parameters θ being expressed as

$$p(V, \theta) = \prod_v p(v|\mathrm{pa}(v), \theta_v)\, p(\theta_v) . \tag{2}$$

From (2) it is clear that θ_v may be considered, formally, as another parent of v in an extended network, such as that shown in Figure 6. Thus, for example, $\theta_{LVH?}$ is a random quantity whose realisation would provide the conditional probability distribution $p(LVH?|Disease?)$, that is, the incidence of LVH in each disease category.

When processing a new case we require the joint distribution of the potentially observable quantities V, which from (2) is given by

$$\begin{aligned} p(V) &= \int p(V, \theta) d\theta \\ &= \int \prod_v p(v|\mathrm{pa}(v), \theta_v) p(\theta_v) d\theta_v = \prod_v p(v|\mathrm{pa}(v)) \end{aligned} \tag{3}$$

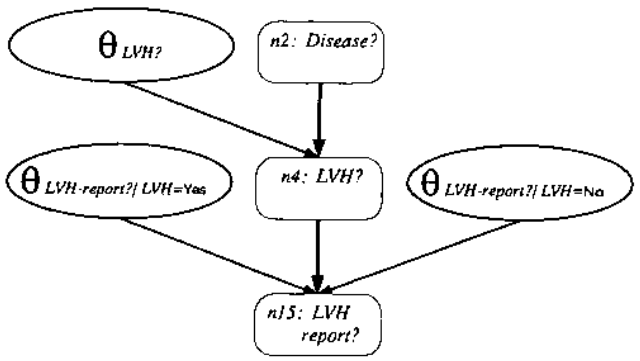

FIG. 6. *Part of CHILD network with supplementary "parameter" nodes, representing marginally independent random quantities* θ_v, $v \in V$ *whose realisations specify the conditional probability tables for the network. In addition, for LVH-report?, the parameter nodes are shown to be locally independent.*

where

$$p(v|\text{pa}(v)) = \int p(v|\text{pa}(v), \theta_v) p(\theta_v) d\theta_v$$

is the expectation of the conditional probability table for v. Hence we can simply use the current "mean probabilities" within the standard evidence propagation techniques assuming known parameters.

A further simplification is obtained if we are willing to assume *local independence,* by which we mean that θ_v breaks into components corresponding to the different configurations of pa(v), which are then assumed mutually independent random quantities. For example, Figure 6 shows $\theta_{LVH\text{-}report?}$ broken into $\theta_{LVH\text{-}report?|LVH=yes}$ and $\theta_{LVH\text{-}report?|LVH=no}$, assumed to be marginally independent variables specifying the respective conditional probability distributions.

A number of alternative parametrisations of the conditional probability tables are possible, but the most intuitive appears to be to assume a Dirichlet distribution, reducing to a beta distribution for binary variables. Let v have K states. Then for a particular parent configuration pa(v)*, we assume a parametrisation $p(v|\text{pa}(v)^*, \theta_{v|\text{pa}(v)^*}) = \theta_{v|\text{pa}(v)^*}$, with $\theta_{v|\text{pa}(v)^*}$ having a Dirichlet distribution $\mathfrak{D}[\alpha_1, \ldots, \alpha_K]$. We can think of the α_k as representing counts of past cases which are stored as a summary of our experience. For the next case, the conditional probability used for the kth category is from (3)

$$p(v_k|\text{pa}(v)^*) = E[\theta_{v|\text{pa}(v)^*}]_k = \alpha_k / \alpha$$

where $\alpha = \sum_{j=1}^{K} \alpha_j$ is the current "precision" underlying our beliefs concerning $\theta_{v|\text{pa}(v)^*}$.

If we observe v to be in the jth state, and pa(v) to take configuration pa(v)*, we have by standard conjugate Bayesian updating that

$$\theta_{v|\text{pa}(v)^*}|v_j, \text{pa}(v)^* \sim \mathfrak{D}[\alpha_1, \ldots, \alpha_j + 1, \ldots, \alpha_K],$$

a Dirichlet distribution denoted by $\mathfrak{D}_j$. Hence if both pa(v) and v are observed, we have a simple accumulation of cases gradually revising our point estimates of the conditional probabilities underlying the system. However, in general we may find that neither v nor its parents are observed with certainty, and this is discussed in Section 7.

4.2 A Numerical Example

When eliciting subjective judgements one can view the above discussion as allowing imprecise specification of the conditional probabilities. For example, Table 1 only displayed the point estimates obtained from the expert, who in fact provided additional ranges to reflect the perceived imprecision. These are shown in Table 2. The judgements shown in Tables 1 and 2 can be thought of as representing prior beliefs concerning unknown frequencies, and we need to transform these to parametric prior distributions. Our (somewhat simplistic) current procedure is as follows. The point values given in Table 1 are taken as the prior means. The range for each response is assumed to represent a one standard error interval [other values have been tried but good predictive performance has been found with this assumption (Spiegelhalter et al., 1991b)]. So, for example, we assume the range 0.05–0.20 given for $p(LVH = \text{yes}|\text{PFC}) = 0.10$ in Table 1 corresponds to a standard error of 0.075 for a mean of 0.10. This would be obtained with a beta distribution with parameters (1.50, 13.50). Similarly the range 0.70–0.99 for $p(LVH = \text{no}|\text{PFC}) = 0.90$ translates to a beta (0.33, 2.95) distribution. We take the minimum precision over the list of responses to decide the overall precision of the beta

TABLE 2

Subjective assessments of conditional probability tables, with expressed imprecision, and their translation into implicit samples underlying a beta(α_1, α_2) *distribution (parameters given to one decimal place)*

	LVH?			
	Yes		No	
Disease?	Range	α_1	Range	α_2
PFC	0.05–0.20	0.3	0.70–0.99	3.0
TGA	0.05–0.20	0.3	0.70–0.99	3.0
Fallot	0.05–0.20	0.3	0.70–0.99	3.0
PAIVS	0.70–0.99	3.0	0.05–0.20	0.3
TAPVD	0.02–0.08	1.2	0.90–0.99	21.3
Lung	0.05–0.20	0.3	0.70–0.99	3.0

	LVH-report?			
	Yes		No	
LVH?	Range	α_1	Range	α_2
Yes	0.80–0.99	18.3	0.02–0.15	2.0
No	0.01–0.10	1.2	0.90–0.99	21.3

or Dirichlet distribution, and so adopt the parameters shown in Table 2—these are thought of as implicit sample sizes. Our approach is clearly quite conservative—we want to use the expert's judgements to give our system a "headstart," but we want to ensure that accumulating data will be able to adapt those judgements reasonably rapidly if necessary.

The actual data available for learning comprise 168 cases of the six diseases of interest which were referred to GOS in 1988 and 1989. For each case information is generally available on nodes $n1$, $n2$, $n3$, $n9$, $n15$, $n16$, $n17$, $n18$, $n19$, $n20$, that is, the data that are reported over the telephone, plus the final diagnosis as established at GOS. It is, however, feasible that data on the "internal" nodes might become available if the hospital records of the patients were retrieved.

4.3 Model Criticism

4.3.1 *Batch and sequential monitors*

In systems that heavily exploit prior information it seems essential that not only should there be a capacity for learning, but the initial assumptions should also be critically examined in the light of data obtained [see, for example, Box (1980, 1983) for an exposition of the importance of this iterative process]. Here we concentrate on sequential techniques for such model diagnostics, but first make some comments on batch monitoring of assumptions.

The issue of testing the compatibility of a batch of data with an assumed model is general to the whole of statistical science, and we do not deal with this in detail here. If we observe complete data then the problem is fairly straightforward within a significance testing framework. The prior assessments for conditional probabilities can be directly compared to the observed counts, and the techniques suggested by Box used to derive significance tests (Spiegelhalter et al., 1991b). Structural assumptions of conditional independence can also be directly assessed using significance tests.

With incomplete data the classical approach runs into problems with deriving an appropriate sampling distribution. However, we shall see that a Bayesian approach is invariant to the order of the data, and hence there is no difference between batch and those sequential techniques discussed below that have a Bayesian derivation.

Three types of diagnostic monitors will be described; a *parent-child monitor* is a direct check on the adequacy of the prior beliefs in the conditional probability distribution of a node given its parents, a *node monitor* checks how well each node is predicted given all other available evidence on a case, and a *global monitor* assigns an overall degree of support for a particular directed graphical model. Explicit comparison of global monitors then forms the basis for the discussion in Section 5 on comparing different graphical structures. Many other types of monitor can be envisaged, but they all share a common foundation of a standardised check on how well the system is predicting the incoming data. Hence we are following the prequential approach of Dawid (1984) in basing our criticisms solely on the quality of the predictions made sequentially. See Cowell, Dawid and Spiegelhalter (1993) for an empirical investigation of the behaviour of these monitors using simulated data.

Each monitor can be thought of as a measure of the "surprise" felt when the data are observed. A formalisation of this concept is provided by *scoring rules*, which are a general procedure for evaluating the quality of probability statements concerning events that are then observed and have been widely used in areas such as probabilistic weather forecasting (Murphy and Winkler, 1984). The basic idea is that a penalty is incurred if a low probability is given to an event that actually occurs. When the scoring rule obeys certain properties the forecaster is encouraged to provide an honest assessment of their uncertainty: examples of such *proper* scoring rules include the Brier score, that is essentially a quadratic penalty, and the *logarithmic score* which we have used in our monitors.

In general, let Y denote a discrete random variable which can take on values y_k, $k = 1, \ldots, K$, and let $p_i(Y)$ denote the probability distribution for Y after $i-1$ cases have been analysed [$p_i(Y)$ may differ from $p_{i-1}(Y)$ if sequential updating of probabilities is occurring]. Suppose that $Y = y^*$ occurs in the ith case. Then we associate with the ith observation of Y a *logarithmic score*, S_i, given by

$$S_i = -\log p_i(y^*) .$$

By accumulating over a series of N cases we obtain a total penalty $S = \Sigma_{i=1}^{N} S_i$, which is the negative logarithm of the overall probability of all the data observed, and is thus invariant to the order in which the data have been observed.

For a Dirichlet prior $\mathfrak{D}[\alpha_1, \ldots, \alpha_K]$, $\Sigma_k \alpha_k = \alpha$ and observed data $(n_1, \ldots, n_K)$, $\Sigma_k n_k = n$ we have that the total logarithmic penalty is

$$S = -\log\left[\frac{\Gamma(\alpha)}{\prod_{k=1}^{K}\Gamma(\alpha_k)} \frac{\prod_{k=1}^{K}\Gamma(\alpha_k+n_k)}{\Gamma(\alpha+n)}\right] . \tag{4}$$

The problem is to decide whether the total observed penalty is a cause for alarm or not, and to do this we require some standardising technique. Two approaches are possible. The *relative*, Bayesian approach explicitly sets up an alternative predictive system, which we shall term a reference system, that gives rise to a total penalty S^{ref}. The sign of $S^{ref} - S$ determines the better predictor, and we shall show that the size of $S^{ref} - S$ has a direct probabilistic interpretation.

The second, *absolute* standardisation takes place without any consideration of an alternative predictive system and as such fits into the general significance testing framework. We form a standardised test statistic for the null hypothesis that the observed events are in fact occurring with the probability stated by the system. Under this hypothesis, before Y is observed the penalty to be incurred is a random quantity with expectation E_i and variance V_i given by

$$E_i = -\sum_{k=1}^{K} p_i(y_k) \log p_i(y_k),$$

$$V_i = \sum_{k=1}^{K} p_i(y_k) \log^2 p_i(y_k) - E_i^2.$$

From these we may calculate a standardised test statistic,

$$Z_N \equiv \frac{\sum_{i=1}^{N} S_i - \sum_{i=1}^{N} E_i}{\sqrt{\sum_{i=1}^{N} V_i}}$$

which will have approximate mean 0 and variance 1 under the null hypothesis that we are making appropriate predictions [see Cox (1958) for an early investigation of this form of test of predictive probabilities]. If Y is not observed in the ith case, then S_i, E_i and V_i are defined to be zero. It can be shown that under broad conditions Z_N is asymptotically standard normal when the assumed model holds (Seillier-Moiseiwitsch and Dawid, 1993).

4.3.2 *Parent-child monitors*

These monitors are intended to detect discrepancies between prior beliefs in $p(v|\text{pa}(v)^*)$, for a particular parent configuration $\text{pa}(v)^*$, and the observed distribution of v when $\text{pa}(v)^*$ obtains. Hence the monitor is only applicable when a node and its parents are observed, and so $\text{pa}(v)^*$ will no longer be explicitly mentioned. Spiegelhalter et al. (1991b) explore a variety of sequential and nonsequential approaches to this problem.

Consider the node *Disease?* for the parent configuration *Birth asphyxia* = yes. Table 3 shows the initial prior estimates and ranges and their transformation to a Dirichlet distribution. These were used to process the 31 cases (of 168) who had birth asphyxia. The relative approach is to contrast the total penalty S with the penalty S^{ref} that would have been obtained had the expressed prior opinion been irrelevant to v, and instead a "reference" prior assumed: $S^{ref} - S$ is the log(Bayes factor) for testing the null hypothesis that the expert's prior is appropriate, since

$$\exp(S^{ref} - S) = \frac{p(\text{all data}|\text{expert's prior})}{p(\text{all data}|\text{reference prior})}.$$

Using a reference prior $\mathfrak{D}[1/K, \ldots, 1/K]$ results in a

TABLE 3
Raw and transformed prior assessments for p(Disease? | Birth asphyxia = yes)

Disease	Prior estimate	Prior range	Dirichlet parameter α_k
PFC	0.20	0.05–0.30	0.85
TGA	0.30	0.10–0.50	1.28
Fallot	0.25	0.15–0.35	1.06
PAIVS	0.15	0.10–0.30	0.64
TAPVD	0.05	0.02–0.10	0.21
Lung	0.05	0.02–0.10	0.21
	1.00		4.25

statistic which was examined in detail in Spiegelhalter et al. (1991b). Figure 7 shows the accumulating reference penalty, in which the reference predictions initially have smaller penalty than those based on the expert prior. However, allowing the priors to be adapted by the data finally gives $S = 42.8$, $S^{ref} = 43.8$, providing a Bayes factor of 2.7 in favour of the prior provided by the expert, over a reference prior. Figure 7 also shows the substantially higher penalty incurred by using the initial prior estimates without learning from the observed data.

Alternatively, the frequentist absolute standardisation gives the following results. By the time the first case with birth asphyxia arrived, indirect evidence had slightly revised the initial distribution to (0.196, 0.295, 0.263, 0.147, 0.049, 0.049), from which we can calculate that the penalty has expectation $E_1 = 1.609$ and variance $V_1 = 0.272$. In fact the true disease of this first case was PFC, which leads to a penalty $S_1 = -\log 0.196 = 1.628$. Thus $Z_1 = 0.035$. The second case with birth asphyxia had disease TAPVD, which received a penalty of 3.281, giving $Z_2 = 2.25$. Figure 7 shows the accumulating standardised parent-child penalty with and without learning; if no learning is allowed the prior assessments remain constant and the standardised penalty increases, while with learning the predictions adapt to the data after a while and the standardised penalty eventually stabilises around 0. In fact, as may be seen from Table 4, the observed data showed PFC and lung disease were both considerably more common in cases with birth asphyxia than expected.

In general the Bayesian relative approach gives similar results to the significance tests: in 3 of 21 parent-child monitors the reference prior would have given better predictions, and these had Z statistics of 2.44, 1.84 and 2.30. The highest Z was for the assessment of the proportion of those with lung disease that would be sick: a prior assessment of 0.70 (range 0.50–0.90) was made, translating to a beta(3.0, 1.25) distribution, while in fact only 6/16 cases were reported as sick. This gave

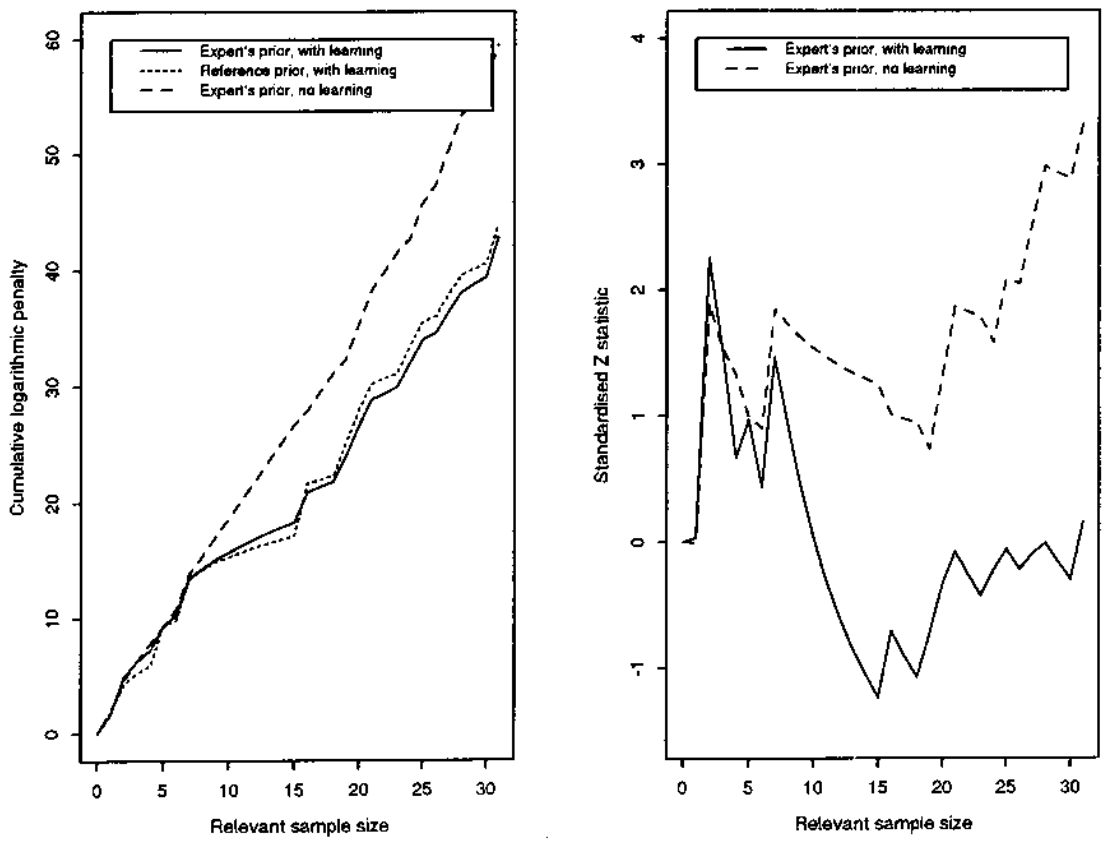

FIG. 7. *Parent-child monitor for the distribution of Disease? for parent configuration Birth asphyxia = yes. On the left are the cumulative logarithmic penalties for three sets of predictions: those based on the unchanged point prior estimates shown in Table 3 (expert's prior, no learning), those based on the Dirichlet distribution in Table 3 (expert's prior, with learning), and those based on a reference Dirichlet distribution* $\mathcal{D}[0.16, \ldots, 0.16]$. *On the right are standardised monitors with and without learning, showing the early indications of poor predictions which are eventually overcome by accumulating data.*

$Z = 2.44$ and $S^{ref} - S^{*} = -0.02$, showing only marginal preference for the reference beta(0.5, 0.5) prior.

Although we have concentrated on sequential monitoring, it is also possible to consider *batch* monitors which are calculated simply on the basis of the observed counts such as shown in Table 4.

First, we note that since the statistic $S^{ref} - S$ is invariant to the order of the data, the sequential Bayesian monitor also acts as a batch monitor. In contrast, from a frequentist perspective we need to calculate the chance of getting such an extreme value of S, given the null hypothesis that the initial prior assessment is appropriate. Spiegelhalter et al. (1991b) show how this is essentially the proposal of Box (1980) for comparing prior with likelihood, and that the test can be approximated by an adjusted Pearson's χ^2 statistic. Specifically, we may calculate the expected counts under the point prior estimates, as shown in Table 4. We then calculate a χ^2 statistic based on the observed and expected counts, which in our example gives $\chi^2 = 54.14$ on 5 degrees of freedom. To allow for the prior imprecision, this must be discounted by a factor $(\alpha + 1)/(\alpha + n) = (4.25 + 1)/(4.25 + 31) = 0.149$ in our case. The final test statistic is then $54.14 \times 0.149 = 8.07$, which on 5 degrees of freedom gives $P = 0.17$. There is therefore slight evidence against the expert's prior: the contrast between this and the small Bayesian support for the null hypothesis is an example of Lindley's paradox (Lindley, 1957).

TABLE 4
Observed frequency distribution for Disease? when Birth asphyxia = yes, with counts expected under the initial prior estimates

Disease	Observed count	Observed proportion	Expected counts
PFC	19	0.61	6.20
TGA	3	0.10	9.30
Fallot	1	0.03	7.75
PAIVS	2	0.06	4.65
TAPVD	0	0.00	1.55
Lung	6	0.20	1.55
	31	1.00	31

4.3.3 *Node monitors*

Two types of monitors exist for each node in the graph and are intended to identify parts of the network in which the modelling is poor. *Unconditional* monitors simply assess the marginal distribution given to each node, in that if node v is observed to have value v^* for case i, then the unconditional penalty for node v is increased by $-\log p_{i-1}(v^*)$. The cumulative penalties may be absolutely standardised by calculating the mean and variance as before—a relative standardisation against a reference prior may also be carried out but is not reported here.

Conditional monitors are concerned with the quality of the predictions on each node observed, conditional on all other current evidence on that case. Formally, we calculate total and standardised scores based on $p(v^*|\mathcal{E}_i \setminus v^*)$, where $\mathcal{E}_i$ denotes the total evidence on case i. The idea is to rapidly identify nodes of the graph that are not well predicted by what else is known on that case. This is also valuable for identifying poor prior assessments when incomplete data are observed and the parent–child monitors are ineffective.

At first sight it may seem that if a case has evidence about M nodes, then up to M propagations would be necessary to evaluate the conditional node monitors, a process which could be quite time consuming. However, as mentioned in Section 6.6 the standard evidence propagation algorithm can be modified to provide for the necessary "fast retraction."

Table 5 shows unconditional and conditional standardised monitors for all observed nodes in the graph. Examining the unconditional nodes it is clear that the overall incidence of *Birth asphyxia?* and *Disease?* are poorly modelled, even after learning. Conditional monitors for *Lower body O_2?* and *RUQ O_2?* suggest the assessments in this part of the graph should be carefully examined. The remaining conditional monitors appear reasonable, although the monitors without learning suggest that this is due to the considerable prior imprecision and hence the rapid adaptation of the conditional probabilities. The considerable number of negative Z statistics suggests that if anything the probability assignments have been rather conservative, in that the observed penalty has been much less than that expected, indicating the lower scoring (most likely) events are occurring more often than predicted.

The diagnostics shown in Table 5 are only the first step in the improvement of the system. Since errors at prior specification at the top of the graph may filter through to affect all the unconditional monitors, it is appropriate to adjust the priors in sequence to identify additional poor assessments. When reasonable unconditional monitors are obtained then aberrant conditional monitors should better reflect poor structure. We do not report this further iterative development here.

4.3.4 *Global monitors*

We define the contribution of the ith case to the global monitor of a model to be the logarithmic score of the probability of the evidence observed, that is, $-\log p_i(\mathcal{E}_i)$, when $i-1$ cases have been processed, and $p_i(\mathcal{E}_i)$ is the probability of the ith evidence $\mathcal{E}_i$. As pointed out in subsection 6.4, this probability is simply the normalisation of any clique table in the junction tree after the evidence $\mathcal{E}_i$ has been propagated.

If we let the total global monitor be $G = \Sigma_i - \log p_i(\mathcal{E}_i)$, we have that

$$G = -\log \prod_i p_i(\mathcal{E}_i) = -\log \prod_i p(\mathcal{E}_i|\mathcal{E}_1, \ldots, \mathcal{E}_{i-1})$$
$$= -\log p(\mathcal{E}_1, \ldots, \mathcal{E}_i) = -\log p(\mathcal{E}),$$

the marginal or integrated likelihood of all the evidence $\mathcal{E}$ that has been observed.

With complete data observed on case i we obtain

$$p_i(\mathcal{E}_i) = p_i(x_V) = -\sum_{v \in V} \log p_i(v|\mathrm{pa}(v))$$

by the factorisation (1), and hence the contribution to the global monitor is simply the sum of the contribution to the individual parent–child penalties for each node. Hence the total G is the sum over all nodes and

TABLE 5

Final conditional and unconditional standardised monitors for observed nodes, both with and without parameter learning

Node	N	With learning		Without learning	
		Conditional	Uncond.	Conditional	Uncond.
n1: Birth asphyxia?	120	0.38	1.96	1.55	5.78
n2: Disease?	168	0.64	2.61	1.39	5.93
n3: Age at presentation?	165	1.41	−0.59	0.47	−3.39
n9: Sick?	168	1.47	0.02	−0.18	0.48
n15: LVH-report?	141	−1.17	0.06	−3.01	−4.11
n16: Lower body O_2?	45	−2.13	−0.98	−2.26	−1.39
n17: RUQ O_2?	120	−1.91	0.36	−1.22	1.96
n18: CO_2?	146	−1.25	−1.55	−5.10	−5.57
n19: X-ray report?	168	−0.98	0.99	−2.74	−1.89
n20: Grunting report?	165	−0.52	−0.62	−3.99	−3.79

all parent configurations of the parent-child penalty formula derived from (4).

The calculation of E_i, V_i and the reliability statistic Z_i of the global monitor is in general quite laborious, except for the two cases of complete data, for which a local propagation scheme exists or for models having small total state space. However, a relative approach to global monitors may be used to compare the overall predictive quality of competing models, and this is considered in the next section.

5. MODEL COMPARISON

Although it may be reasonable to start off with the structure provided by the expert we should use the data to monitor the conditional independence assumptions the graph expresses. In particular, we need to be able to compare two or more candidate structures. As an example, in this section we shall compare the CHILD structure given in Figure 2 with the naive structure of Figure 1; the latter model only incorporates those observed nodes listed in Table 5.

5.1 Bayes Factors

The Bayesian view is straightforward, being based on Bayes factors which are simply contrasts of global monitors. If for two possible models we obtain global monitors G^1 and G^2, then their difference is

$$\Delta^{12} = G^2 - G^1 = \log p^1(\mathcal{E}) - \log p^2(\mathcal{E}) = \log \frac{p^1(\mathcal{E})}{p^2(\mathcal{E})}.$$

The Bayes factor to compare these two models is thus simply $\exp(\Delta^{12})$. If we are willing to assign a prior log-odds δ^{12} on model 1 versus model 2, then the posterior log-odds on model 1 is just $\Delta^{12} + \delta^{12}$.

It is not in general feasible, nor indeed appropriate, to require separate prior elicitation for each possible model. In particular we would like the priors in nested models to be *compatible*, in the sense of being obtainable by appropriate marginalisations. In Section 8 we describe a procedure for obtaining such compatible priors by a process we term *expansion and contraction*.

5.2 Global Monitors for CHILD Network

We now take as our baseline model the CHILD network, with the expert's priors and parameter learning, so that $p^1(\mathcal{E})$ is the marginal probability of the observed evidence under this assumed model. We contrast it with the naive network, using the priors derived from the process of expansion and contraction described in Section 8. Additional comparisons are made with the naive network with reference priors (no expert input) and the CHILD network with no learning (keeping to the expert's point estimates). Figure 8 shows the log(Bayes factors) for the three alternative models versus the baseline model.

It is apparent from Figure 8 that the ability to learn about the parameters starts showing benefit after about 25 cases, and thereafter the no-learning model is clearly inferior. The naive model is initially poor, but with learning it becomes better than CHILD without learning after about 52 cases, although it becomes increasingly inferior to the more structured baseline model. The naive model that starts from reference priors almost catches up the model with the expert "headstart" after about 120 cases. Overall, it is apparent that the data strongly support the more structured model, particularly in view of Jeffreys' rule-of-thumb (Jeffreys, 1961, p. 432) that a log(Bayes factor) more than 4.6 (i.e., $\log_{10}$ Bayes factor = 2) constitutes "decisive" evidence, which occurs after only 11 cases. Of course, further comparison with more local structural adjustments should now be investigated using these techniques.

5.3 Diagnostic Comparisons

The above model comparison has used the ability of each model to predict the totality of data that is observed, whereas it may be more appropriate to place more weight on a model's success in predicting the disease node. Table 6 summarises the diagnostic accuracy of the four models being examined. The conditional monitor S is the total logarithmic penalty given to the node *Disease?* conditional on all other evidence, and hence $\exp(S/N)$ is the geometric-mean posterior probability of true disease. The diagnostic accuracy is simply the number of cases for which the disease receiving the highest posterior probability was the true disease.

It is apparent that from the perspective of a proper scoring rule the baseline model is superior. However, this benefit of additional quantitative and structural input is not fully manifested into increased diagnostic accuracy. Further analysis reveals this is primarily due to the evidence being somewhat attenuated by the conservative probability assignments in CHILD and hence becoming insufficient to push the rarer diseases into first place. It is also important to note that all the models in Table 6 do considerably worse, in terms of simple accuracy, than a simple algorithmic approach: the algorithm in Franklin et al. (1991) obtains an accuracy of 132. The clinicians at GOS had an accuracy of 114, with the referring paediatricians 76. Thus a probabilistic system in the process of learning matches the quality of the middle-level paediatric cardiologists at GOS, but clearly more work is required into exploiting the interactions that the algorithmic approach incorporates. For an evaluation of a fully trained network, see subsection 5.4 below. It is also important to consider the relative importance of diagnostic errors, and to this end a loss function has already been elicited (Franklin et al., 1991).

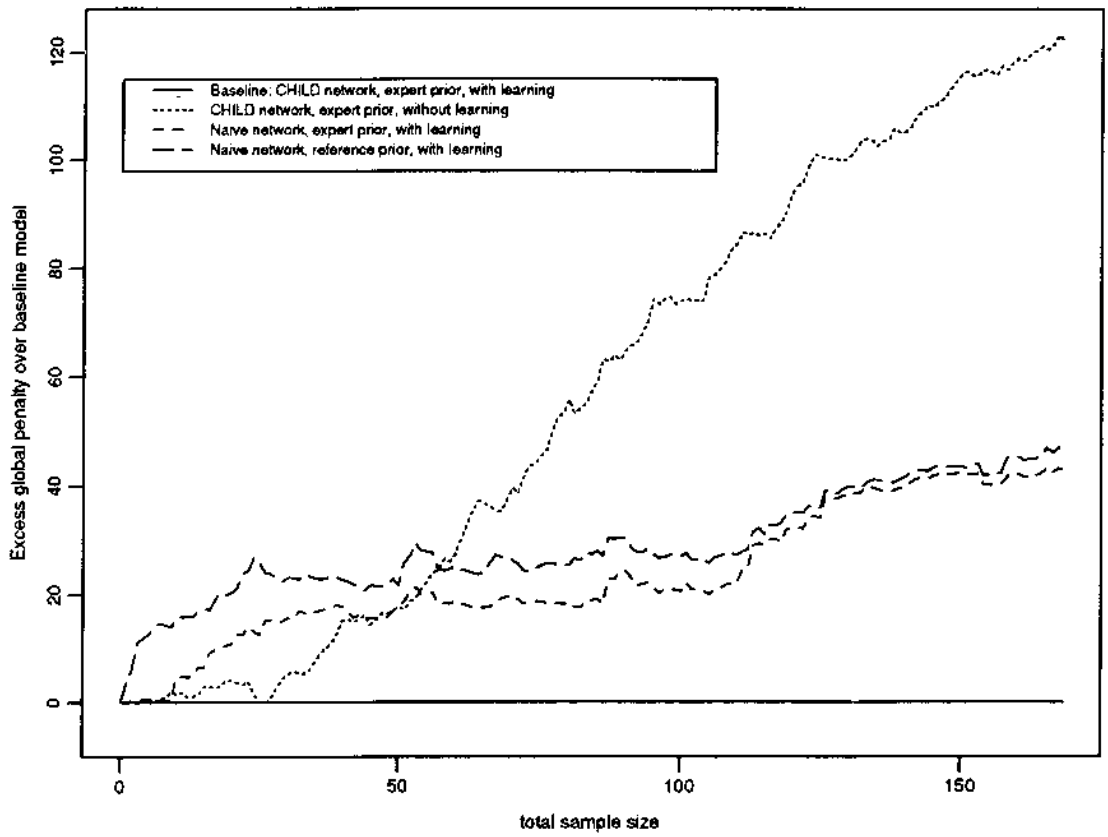

FIG. 8. *Global monitors for three alternative models compared to a baseline assumption. The ordinate also expresses the* log(*Bayes factor*) *in favour of the baseline model (CHILD network with learning).*

5.4 Statistical Model Selection

An approach that takes model comparisons to its full consequence is to induce the network directly from data using statistical model selection methods, ignoring the prior structural and quantitative information available. There is a number of different areas in which these ideas have been investigated. The *statistical* literature in general exploits strategies based on significance testing to construct log-linear models which have an undirected graphical representation: see, for example, Wermuth (1976), Edwards and Havránek (1985, 1987) and Andersen, Krebs and Andersen (1991). The program BIFROST (Højsgaard, Skjøth and Thiesson, 1992) exploits a contingency table program CoCo (Badsberg, 1991) to build networks in a systematic fashion using some prior structural information. Currently the methods are limited to being applied to complete data. Hence, in our example, the structure of the model must be lost to some extent, since it can be based only upon a limited number of nodes.

The performance of a system constructed by these methods has been investigated by Lauritzen, Thiesson and Spiegelhalter (1992). In a prospective test, the best automatically generated networks had an accuracy of 63 of 87 correctly diagnosed cases, compared to 64 of 87 for a fully trained version of the CHILD network. These accuracies (73–74%) are comparable to the simple algorithmic approach.

A second literature is founded more in the social sciences and seeks to investigate causality by building

TABLE 6
Measures of the diagnostic accuracy of four models

Model	Conditional penalty S	Mean posterior prob. ($e^{S/N}$)	Accuracy in 168 cases
Baseline: CHILD, expert priors, learning	156.3	0.40	110
CHILD, expert priors, no learning	197.4	0.31	99
Naive network, expert priors, learning	172.4	0.36	113
Naive network, reference priors, learning	176.4	0.35	111

directed graphical models from data: see, for example, Spirtes, Glymour and Scheines (1993). Finally, the AI or machine learning community is concentrating on construction of belief networks from data, both as a theoretical (Pearl and Verma, 1991) and a practical task (Cooper and Herskovits, 1992). The latter approach relies on complete data and uses the global monitors (Bayes factor) approach described below, but assuming uniform priors ($\alpha_i = 1$) throughout. We note that the "machine learning" literature is primarily concerned with classification, rather than the construction of a model for the underlying process.

6. GRAPHICAL ALGORITHMS AND EVIDENCE PROPAGATION

In this more technical section we give details of the graphical and algebraic procedures that underlie the evidence propagation displayed in Section 3. We begin by making a transformation from a directed to an undirected graphical model: this process is useful both in checking conditional independence statements made on a DAG and as a first step in the eventual re-representation of the graph in terms of locally communicating belief universes.

6.1 From a Directed to an Undirected Graph

Undirected graphs have been used increasingly in the specification of statistical models for analysis of multivariate data: see, for example, Lauritzen (1989) and Whittaker (1990). In such a graphical model $\mathcal{G}$ nodes are again random variables but are linked by undirected edges. The factorisation property analogous to (1) is that the joint distribution is expressed as a product of terms defined on the *cliques* of the graph, which are the maximal sets of nodes that are all joined to each other. This implies that the joint distribution is *Markov* with respect to $\mathcal{G}$, in the following sense. Let A, B and C be sets of variables such that any path in $\mathcal{G}$ from a node in A to one in B must pass through C. Then $A \perp\!\!\!\perp B|C$, that is, the sets A, B are conditionally independent of each other given the variables in C.

To establish the connection between a DAG $\mathcal{D}$ and an undirected graphical model, we first note that $p(v|\text{pa}(v))$ can trivially be considered as a function, denoted by b, defined on family(v), where *family* represents a node and its parents. Hence (1) can be written in the form

$$p(V) = \prod_{v \in V} b(\text{family}(v)). \tag{5}$$

If $\mathcal{G}$ is chosen to have cliques which are the families of $\mathcal{D}$, we have fulfilled the condition for an undirected graphical model $\mathcal{G}$ since the joint density p is expressed as a product of terms (5) defined on the cliques of $\mathcal{G}$. Such a graph $\mathcal{G}$ is obtained by placing a link between co-parents in $\mathcal{D}$ that are not currently connected and dropping directions on the links. For CHILD, the resulting "moral" graph (originally named as it married unmarried parents) is shown in Figure 9 and is in general denoted by $\mathcal{D}^m$.

We have therefore shown that a distribution $p(V)$ that is recursive with respect to a DAG $\mathcal{D}$ must be Markov with respect to $\mathcal{D}^m$. In particular, if A, B and C are subsets of V such that C separates A from B in $\mathcal{D}^m$, then $A \perp\!\!\!\perp B|C$. Still further conditional independencies will often be obtained on observing that, if $W \subset V$ is an ancestral set, that is, it contains its own ancestors, then the restriction of p to W is recursive

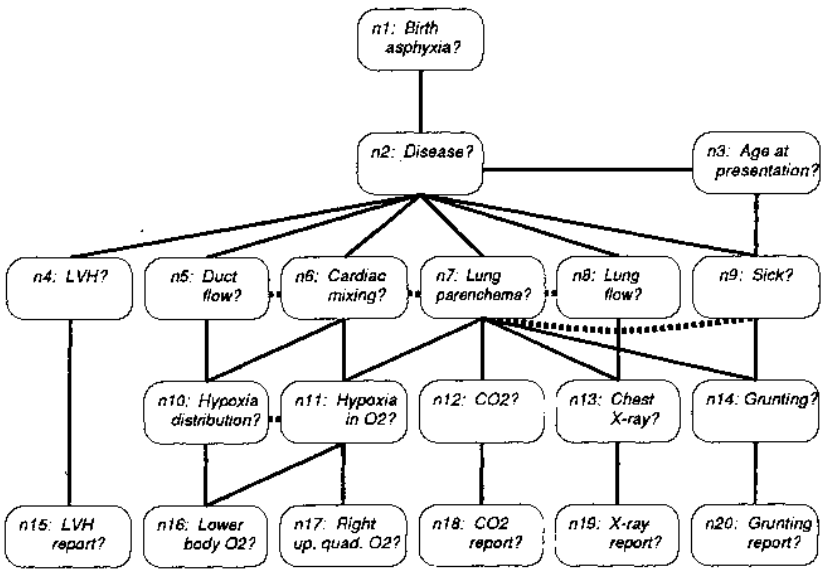

FIG. 9. *Moral graph formed from CHILD network by joining unconnected parents and dropping directions. The joint distribution of the variables is Markov with respect to this graph.*

with respect to $\mathfrak{D}_W$, and hence Markov with respect to $\mathfrak{D}_{\mathfrak{W}}$.

For (a somewhat contrived) example, we might ask whether, if we were to know the true disease ($n2$) and have measured the hypoxia when breathing oxygen ($n11$), would knowing the CO_2 report ($n18$) tell us anything additional about the distribution of hypoxia ($n10$)? Figure 10 shows the moral graph of the ancestral set of the variables of interest. It is clear that a path exists between $n18$ and $n10$ that bypasses $n2$ and $n11$. Hence $n18$ would be informative, essentially since it would tell us more about $n7$: *lung parenchyma?*, and hence whether the observed hypoxia level ($n11$) is explained by the state of the lungs. If not, this changes our belief in $n6$: *Cardiac mixing?*, which finally feeds down to $n10$. It is a crucial feature of these directed structures that they allow reasoning about joint causes through effects being "explained away," as is $n11$ by $n7$.

It may be shown (Lauritzen et al., 1990) that this technique of forming the moral graph of ancestral sets will reveal *all* the conditional independence properties logically implied by $p(V)$ being recursive with respect to $\mathfrak{D}$. However, we shall here be content with the re-representation of $p(V)$ as graphical over $\mathfrak{D}^m$. Under this transformation, some of the conditional independence properties displayed in the original DAG $\mathfrak{D}$, such as that between $n5$ and $n6$ given $n2$ in Figure 2, may lose their representation in graphical form. They still hold, but are effectively buried in the quantitative component of the model. Only those conditional independences that retain a graphical representation in $\mathfrak{D}^m$ can be utilised for further simplification of the analysis.

Having carried out this initial specification and the construction of the moral graph, the tasks involved in deriving the evidence propagation procedure illustrated in Section 3 may be listed as follows:

- Identification and organisation of belief universes
- Initialization
- Propagation

These are described in successive subsections using CHILD as an example. We then describe how slight adaption of the basic propagation algorithm provides additional tools, and finally we discuss a basic axiomatic approach that extends the algorithm to nonprobabilistic conditional independence.

6.2 Identification and Organisation of Belief Universes

We now identify the belief universes introduced above as the cliques $\mathcal{C}$ of a suitably chosen undirected graph $\mathcal{G}$. For computational reasons we will need to organise the cliques $\mathcal{C}$ of $\mathcal{G}$ into a tree $\mathfrak{I}$, the *junction tree,* with the property that, for any $v \in V$, the collection of all $C \in \mathcal{C}$ containing v forms a (connected) sub-tree of $\mathfrak{I}$; this ensures that communication from any node to another is via a unique path. It may be shown that this is possible if and only if $\mathcal{G}$ is *triangulated* (Jensen, Olesen and Andersen 1990); this means that we cannot find a cycle of length 4 or more without a *chord* (if there is more than one such junction tree, any one will serve for further manipulations).

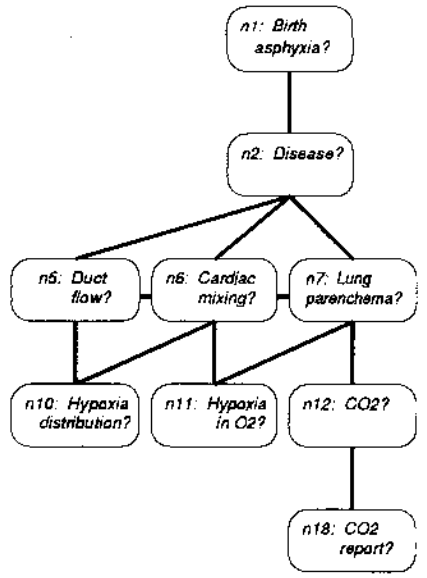

FIG. 10. *Moral graph formed from ancestral set of nodes* {$n2$, $n10$, $n11$, $n18$}. *From the Markov property,* $n18$ *and* $n10$ *are not conditionally independent given* {$n2$, $n11$}, *since there is a path in this graph between* $n18$ *and* $n10$ *that is not blocked by* {$n2$, $n11$}.

We have already shown how to construct an undirected (moral) graph from our original DAG, but this may not be triangulated. In the case of CHILD, the moral graph in Figure 9 is *not* triangulated although at first sight it may appear so: the cycle ($n2, n7, n11, n10, n5, n2$) is such that there is no edge between two nonconsecutive nodes. It may be rendered triangulated by adding suitable further edges, that is, between $n5$ and $n7$ and between $n5$ and $n11$, as shown in Figure 11. This will alter the cliques of the graph: thus ($n2, n5, n6, n7$), ($n5, n6, n7, n11$) and ($n5, n6, n10, n11$) are cliques of the new graph, but not of the old. Equation (5) represents the joint density as Markov on $\mathfrak{D}^m$. It will likewise be Markov on any graph $\mathcal{G}$ over V which is obtained by adding further edges to $\mathfrak{D}^m$, and hence our initial joint distribution will have the independence properties appropriate to this triangulated graph.

In general we would like the triangulated graph $\mathcal{G}$ obtained by adding edges to $\mathfrak{D}^m$ to have cliques that are *small*, say in the sense of their state space. However, any formalisation of this task leads to a difficult optimization problem [in fact one that is $\mathcal{NP}$-complete (Cooper, 1990)]. Various heuristic algorithms have been proposed (Kjærulff, 1992a), and research into good tri-

angulation methods continues – this issue also arises in the context of relational databases (Tarjan and Yannakakis, 1984). There exist however good algorithms for *checking* whether a given undirected graph is triangulated. Having chosen a suitable triangulated graph $\mathcal{G}$ as an extension of $\mathcal{D}^m$, all further computations are based on $\mathcal{G}$ – we note that, as in going from $\mathcal{D}$ to $\mathcal{D}^m$, some conditional independencies have become implicit in the numerical assignments rather than being explicit in the graph $\mathcal{G}$.

The simple *maximum cardinality search* (MCS) algorithm will simultaneously check whether $\mathcal{G}$ is triangulated and, if it is, construct a junction tree. It consists of a sequence of stages as follows.

At stage 1, select any node of V and label it 1. At the start of a later stage i, $i-1$ nodes will already have been labelled. The node to receive label i is that which has the most labelled neighbours. The stage is *successful* if those labelled neighbours of i are all neighbours of each other, that is, form a complete subgraph. It can be shown that all stages will be successful if and only if $\mathcal{G}$ is triangulated, thus providing a check for this condition.

Now suppose that all n stages have been successful. The graph $\mathcal{G}$ is then triangulated, and the labelled nodes form what is known as a *perfect* numbering; that is, for any node, the neighbours with lower number are all connected. Figure 11 shows the additional edges added to make the graph triangulated and the result of a maximum cardinality search starting at $n18$: *CO_2 report?*. For example, after $n7$ and $n2$ have been labelled 3 and 4 respectively, any of nodes $n5$, $n6$, $n8$ or $n9$ could be chosen, as each has two already labelled neighbours. However, once $n5$ is arbitrarily chosen to have label 5, the next node to be labelled must be $n6$, since it is the only node with three currently labelled neighbours. Thus each clique is completed before proceeding to the next.

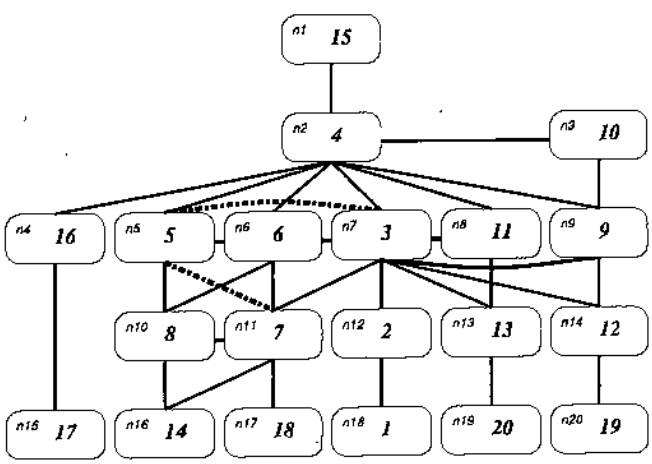

FIG. 11. *A perfect ordering of the nodes in CHILD arising from maximum cardinality search.*

Starting from the above perfect numbering, we can identify the cliques of $\mathcal{G}$. The highest labelled node within each clique is noted, and this produces an ordering of the cliques. Figure 12 shows each clique as a node in a tree – the number in the top right hand corner of each clique is its highest constituent label, and the corresponding clique ordering C_1 to C_{17} is shown.

The links in the tree are obtained as follows. The perfect ordering generated by maximum cardinality search leads to a clique ordering with the following *running intersection property*: let S_j be the nodes in the intersection of clique C_j and lower numbered cliques $C_1, \ldots, C_{j-1}$. Then there exists (at least) one clique in $C_1, \ldots, C_{j-1}$ that contains S_j. Placing a link between that clique and C_j leads to a tree such as Figure 12. For example, clique C_8 has $S_8 = \{n2, n7\}$, which is contained in either C_3 or C_6; the former has been (arbitrarily) chosen to link C_8 with the preceding cliques.

Note that, if a graph possesses a junction tree, this need not be unique. Any junction tree will serve our purposes. The tree has the property that if a node v is contained in any two cliques C_i and C_j, then it is contained in all the cliques in the unique path between C_i and C_j. For example, *n2: Disease?* is contained in C_3, C_6, C_7, C_8, C_{12} and C_{13}, which form a connected sub-tree. As we shall see, the general idea is that evidence on any node can then be passed to the rest of the network by a unique path.

6.3 Initialisation

For any v, family(v) is complete in $\mathcal{G}$, and so it is contained in at least one clique. Assign v to just one such C, and for each $C \in \mathcal{C}$ define $a_C(C)$ to be the product of $p(v|\text{pa}(v))$ for all v assigned to C (or 1 if none are assigned). Then (5) becomes

$$p(V) = \prod_{C \in \mathcal{C}} a_C(C) \tag{6}$$

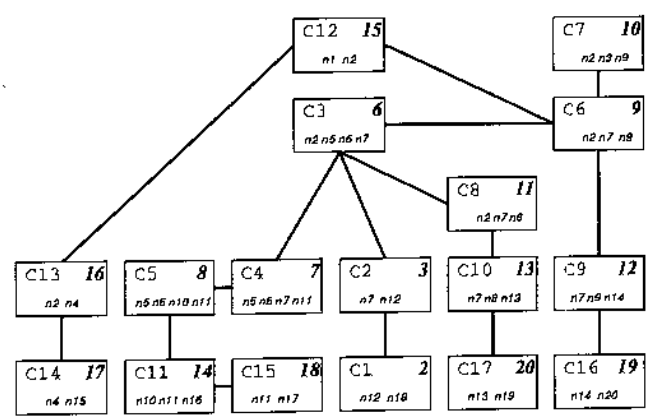

FIG. 12. *Junction tree of cliques derived from perfect ordering of the CHILD nodes. The members of each clique are shown, the highest label among the members is shown in the top right-hand corner, while the corresponding ordering of the cliques is shown in the top left-hand corner.*

which defines the numerical specification of $p(V)$ in turns of the functions $\{a_C\}$ on the cliques of $\mathcal{G}$.

There are many ways, other than the above, of choosing functions $\{a_C\}$ to satisfy (5), and indeed this very freedom forms the basis of the computational algorithms to be described. In fact, it turns out to be useful to generalise (5) to allow still more freedom, as follows. Let C_i and C_j be adjacent cliques of the junction tree. Then we associate with the edge joining them the set $S = C_i \cap C_j$ of nodes of $\mathcal{G}$. This is called a *separator*, and we shall denote the family of all separators by $\mathcal{S}$. (Distinct edges may yield the same separator—in this case we shall have repetitions in $\mathcal{S}$.) As well as having a function a_C for each clique C, we suppose that we have a function b_S for each separator S, such that we can express

$$p(V) = \frac{\prod_{C \in \mathcal{C}} a_C(C)}{\prod_{S \in \mathcal{S}} b_S(S)}. \tag{7}$$

(The right-hand-side of (7) is interpreted as 0 whenever its denominator is 0.) The individual a and b functions are called *potential functions*, and equation (7) is a *potential representation* for p. Initially we take $b_S \equiv 1$ and the a's as described above.

The computational algorithms proceed by modifying the individual potential functions in a sequence of steps, but in such a way that (7) holds at all times. After all steps have been completed, the final potential functions will have a special form containing the desired information. Thus the propagation algorithm discussed below finishes with every potential function being the *marginal density* for the relevant set of variables. With this choice, (7) defines the *marginal representation* of p, which may be written

$$p(V) = \frac{\prod_{C \in \mathcal{C}} p(C)}{\prod_{S \in \mathcal{S}} p(S)}. \tag{8}$$

From the clique marginals it is then trivial to marginalise to the marginal distribution on each node.

6.4 Incorporation of New Evidence

Suppose that we observe "evidence" $\mathcal{E}$: $X_A = x_A^*$. Define a new function p^* by

$$p^*(x) = \begin{cases} p(x), & \text{if } x_A = x_A^*, \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

Then $p^*(x) = p(x, \mathcal{E}) = p(\mathcal{E})p(x|\mathcal{E})$, where $p(.|\mathcal{E})$ is the density of the conditional distribution given $\mathcal{E}$. We can rewrite (9) as

$$p^* = p \prod_{v \in A} l(v),$$

where $l(v)$ is 1 if $x_v = x_v^*$, 0 otherwise. Thus $l(v)$ is the *likelihood function* based on the partial evidence $X_v = x_v^*$.

If we have any representation of the form (6) or (7) for p, we immediately obtain such a representation for p^*, by associating each $v \in A$ with any one clique containing v, and replacing each $a(C)$ by

$$a(C) = a(C) \prod \{l(v) : v \text{ is assigned to } C\} \tag{10}$$

(taking an empty product as unity).

The fact that p^* is proportional to, rather than equal to, a probability density function is of no consequence and can in fact be turned to advantage. In particular, if we apply a routine for finding the marginal representation directly to p^*, it will give us $p^*(U) = p(\mathcal{E})p(U|\mathcal{E})$ for any clique or separator U. Then the *normalising constant* for any such U [i.e., the sum of all the values of $p^*(U)$] will be just $p(\mathcal{E})$. This provides a means of calculating the joint density at specified values of any collection of variables, even when the corresponding set of nodes is not complete in $\mathcal{G}$. Finally, on performing the normalisation, we obtain $p(U|\mathcal{E})$, and so we shall have transmitted the effect of the evidence to every clique.

6.5 A Propagation Algorithm

We now describe an algorithm for calculating the marginal representation, starting from any potential representation. This proceeds by the propagation of simple "flows" through the junction tree. Each such flow involves only two adjacent cliques and the associated separator, and the crucial feature is that after each flow the representation (7) continues to hold.

6.5.1 *Flows between adjacent cliques*

Consider two adjacent cliques from Figure 12: $C_{14} = \{n4, n15\}$ and $C_{13} = \{n2, n4\}$ and their separator $S = \{n4\}$. Initially the potential tables for these are as in Figure 13a; as discussed at the start of subsection 6.3, these are given by $p(n15\,|\,n4)$ for C_{14}, and $p(n4\,|\,n2)$ for C_{13}. If we incorporate the evidence $\mathcal{E}$: *LVH-report* = yes, then the potential for C_{14} changes to that shown in Figure 13b. Starting from these modified potentials, we now describe the passage of a flow from C_{14}, the sender, to C_{13} the receiver, across S. This has two phases:

(i) Within C_{14}, we sum out over variables not in S ($n15$: *LVH-report?*) to obtain a new potential function over S. This gives the potential shown in Figure 13b.

(ii) The potential over C_{13} is now modified by multiplying each term by the associated *update ratio*, which is the ratio of the relevant value of the new potential over S to that of the old one. In this case, the update ratios are just the new potential values since the previous potentials were unity, and we obtain the potentials on C_{13} shown in Figure 13b.

(a)

Disease	LVH? Yes	LVH? No
PFC	0.100	0.900
TGA	0.100	0.900
Fallot	0.100	0.900
PAIVS	0.900	0.100
TAPVD	0.050	0.950
Lung	0.100	0.900

(b)

LVH? Yes	LVH? No
0.090	0.045
0.090	0.045
0.090	0.045
0.810	0.005
0.045	0.048
0.090	0.045

(c)

LVH? Yes	LVH? No
0.004	0.002
0.031	0.016
0.026	0.013
0.180	0.001
0.002	0.002
0.004	0.002

(d)

LVH? Yes	LVH? No
0.004	0.002
0.031	0.016
0.026	0.013
0.180	0.001
0.002	0.002
0.004	0.002

(e) C_{13}

LVH? Yes	LVH? No
0.015	0.007
0.109	0.055
0.091	0.045
0.634	0.004
0.008	0.008
0.016	0.008

(a)

LVH? Yes	LVH? No
1.000	1.000

(b)

LVH? Yes	LVH? No
0.900	0.050

(c)

LVH? Yes	LVH? No
0.900	0.050

(d)

LVH? Yes	LVH? No
0.248	0.036

(e) S

LVH? Yes	LVH? No
0.872	0.128

(a)

LVH Report?	LVH? Yes	LVH? No
Yes	0.900	0.050
No	0.100	0.950

(b)

LVH? Yes	LVH? No
0.900	0.050
0	0

(c)

LVH? Yes	LVH? No
0.900	0.050
0	0

(d)

LVH? Yes	LVH? No
0.248	0.036
0	0

(e) C_{14}

LVH? Yes	LVH? No
0.872	0.128
0	0

FIG. 13. *Propagation of evidence through cliques C_{13} and C_{14} of junction tree*: (a) *initial potentials*, (b) *after incorporation of evidence LVH-report = yes*, (c) *after propagation through rest of network and back to C_{13}*, (d) *final potentials*, (e) *marginal tables after normalisation.*

The above routine applies when any flow is passed between adjacent cliques. The potentials on the separator and on the remaining clique are modified, and a representation as (7) continues to hold but with equality replaced by proportionality.

To achieve our purpose we need to schedule the flows appropriately. Call a flow *active* if, before it is passed, the sender has already received active flows from all its neighbouring cliques, with the possible exception of the receiver (thus initially only peripheral cliques can pass active flows). We can schedule a sequence of flows so that eventually an active flow has been passed in both directions between each pair of neighbouring cliques. In this "equilibrium" state, which is unaffected by any further flows, the potential representation obtained will be the desired marginal representation (Dawid, 1992). In CHILD, a suitable sequence of active flows would be through clique numbers 14-13-12-6, 7-6, 16-9-6, 6-3, 17-10-8-3, 1-2-3, 15-11-5-4-3 and then the same in reverse order.

Figure 13c shows the potentials on C_{13}, C_{14} and S, having incorporated the evidence *LVH-report* = yes just prior to the final flow from C_{13} to C_{14}. Passage of the flow now produces the new potential function on S shown in Figure 13d and corresponding update ratios (0.248/0.900 = 0.276, 0.036/0.050 = 0.720). These ratios are then incorporated, in this case trivially, into the potential for C_{14}, yielding Figure 13d as the final display after propagation. At this point, for example, the potential a_{13} for C_{13} satisfies $a_{13}(d, l) = p(Disease = d, LVH = l, \mathcal{E})$ and so by normalisation of this or any other potential table we can find $p(\mathcal{E}) = 0.284$. Dividing each term by this marginal probability of the evidence observed gives us the final marginal representation (8) shown in Figure 13e. We can now see, for example, that $p(LVH = \text{yes}|\mathcal{E}) = 0.872$.

6.5.2 *Controlling information flow*

There are various ways of constructing a suitable schedule of active flows, the differences between them lying not in the results but in the control mechanisms which are employed. A schedule of the type constructed in 6.5.1 is *palindromic*: it starts and ends with flows out of and into some peripheral clique C, these sandwiching a palindromic schedule for the tree with C omitted – allowing recursive construction. Jensen, Olesen and Andersen (1990) proceed by selecting an arbitrary "root-clique" C_0, which requests incoming flows from its neighbours which in turn pass on similar requests to their neighbours, until they can be satisfied. After this initial "collection" phase is completed, the potential on C_0 will be its equilibrium value. Finally flows are passed out from C_0 towards the periphery.

At the end of this "distribution" phase, full equilibrium is reached.

In effect, any schedule of flows can be divided into a collection and distribution phase, with root the first clique to receive active flows from all its neighbours. Where appropriate, flows can be passed simultaneously in parallel.

Once a suitable scheduling sequence has been constructed it can be used repeatedly for various purposes, such as recomputation taking new evidence into account. However, this may be inefficient: for example, starting from equilibrium new evidence entered in a single clique C only requires distribution from C as root. More efficient is dynamic flow scheduling, handled by a production rule associated with each flow, which fires when the remaining incoming flows are active and at least one has just changed.

6.6 Generalisations

Minor variations on the propagation algorithm described above allow us to solve other problems (Dawid, 1992). For example, suppose that, in phase (i) of the passage of a flow, the new potential over S is constructed by maximisation, rather than summing, over the remaining variables. Then at equilibrium the potential on any clique or separator will be the joint density maximised over all other variables. Calculating node marginals by further maximisation within cliques will yield the "profile probabilities," that is, the probabilities (given the evidence) of a state and the remaining variables, maximised over the latter. By picking out and stringing together the arguments maximising the value at each node we can identify the overall configuration with highest probability, conditional on any evidence that is incorporated. Applying this method to CHILD yields Figure 5, in which the configuration with the highest probability can be thought of as providing the best possible explanation for the configuration of findings observed.

Another task which can be handled in a similar fashion is "fast retraction" of evidence, simultaneously for each node (Cowell and Dawid, 1992), which is relevant for the calculation of conditional node monitors (see subsection 4.3.3).

The essential common structure of all such problems could be captured in a set of axioms for abstract propagation. This line has been followed, although with a slightly different structure and emphasis, by Shenoy (1989) and Shenoy and Shafer (1990), who also apply these ideas to constraint satisfaction and to systems where the uncertainty is expressed using other formalisms than standard probability theory – for example, belief functions. The expert system shell *Pulcinella* (Saffiotti and Umkehrer, 1991), allows the user to select between several ways of representing uncertainty, all handled within a common propagation structure.

The essential strategy of the algorithms discussed has been to break down the problem into its cliques and restrict calculations to take place within one clique at a time. We shall now briefly discuss how this general strategy can also apply to more complex problems of statistical estimation and model testing, both classical and Bayesian.

6.7 Using the Junction Tree for Inference on Decomposable Models

The Bayesian learning methods described in subsection 4.1 are tailored to directed graphical structures. There is also a general theory of Bayesian learning for undirected graphs (Dawid and Lauritzen, 1993). This theory is based on the junction tree representation $\mathcal{T}$ of a triangulated graph $\mathcal{G}$, introduced to support the theory above, and it bears strong formal similarities with that theory.

If C is any clique of $\mathcal{G}$, there is an unknown joint distribution θ_C governing the observables $X_C = \{X_v : v \in C\}$. Assuming the full distribution θ to be Markov over $\mathcal{G}$, one can show that it is determined by $\{\theta_C : C \in \mathcal{C}\}$, and thus our prior distribution is expressible as a joint distribution for these quantities. Now consider any separator S, and let A and B denote the node-sets contained in the two parts of $\mathcal{T}$ remaining when the link through S is removed. The Markov property of X implies $X_A \perp\!\!\!\perp X_B | X_S$ (given θ). There is an analogous condition one might impose on the prior, namely $\theta_A \perp\!\!\!\perp \theta_B | \theta_S$. If this holds for all $S \in \mathcal{S}$, the prior distribution is termed *hyper-Markov*. This often reasonable condition turns out to streamline the Bayesian analysis. In particular, if we specify the marginal prior distributions for each θ_C then (subject to certain obvious compatibility conditions) there will be exactly one hyper-Markov prior for θ having these marginals – clearly this greatly simplifies prior specification. The posterior distribution (based on complete data), will again be hyper-Markov, so that we only need to find the marginal posterior for each θ_C.

In general the posterior for θ_C will depend on data about variables outside, as well as inside, C. However, we can strengthen the hyper-Markov property to require that $\theta_A \perp\!\!\!\perp \theta_{B|S}$. Under this *strong hyper-Markov* structure, the posterior for θ_C will only involve data on variables in C, and thus Bayesian analysis is entirely localised, proceeding quite independently within each clique.

Bayesian inference in the more general (weak) hyper-Markov case can in principle be performed using a propagation algorithm similar to those discussed in Section 3, which allows information about θ_C from data outside C to filter via the passage of suitable flows. However, this generally involves integrals of complicated functions and is thus not easy to implement. The problem is simplified if the prior can be expressed as

a mixture of strong hyper-Markov priors. This property is preserved on observing possibly incomplete data, thus allowing Bayesian inference from such data. However, the number of terms in the mixture can grow exponentially with the sample size, and approximating methods are needed. We are currently exploring a stochastic approximation approach to this problem.

6.8 Limitations and Alternatives

The crucial limitations on the schemes described above concern the size of the state spaces of the cliques obtained after triangulation. With good triangulation algorithms remarkably large and dense networks can be handled, but there comes a point when computational limits are exceeded. This is a particular problem in two common contexts. First, when a node represents an unknown quantity that influences many nodes in a graph, say when the node is a parameter in a model, then the triangulation can become infeasible if there are many such parameters. If there are only a limited number of such nodes, then they can be set at a grid of values and the likelihood of the observed data calculated for each combination – this relates to the "cut-set" proposal of Pearl (1986). This is typically the approach within pedigree analysis (Thompson, 1986).

The second common context is when there is a form of regularity in the graphical structure, such as nodes forming a lattice (e.g., in image analysis) or repeated multiply connected blocks (e.g., complex temporal models). Various analytic approximations may be possible, but current attention is focussing on simulation schemes which are derivatives of those explored in image processing (Geman and Geman, 1984). By allowing nodes to represent parameters, and repeated structures to represent individual cases, a link is provided to statistical analysis of data using hierarchical conditional independence models. In this context we note the recent attention to Gibbs sampling as a general statistical computational technique (Gelfand and Smith, 1990), which in fact was predated by the suggestion of Pearl (1987) for its use in expert systems.

7. LEARNING FROM INCOMPLETE DATA WITH DIRICHLET DISTRIBUTIONS

In this section we shall expand the general discussion of Section 5 to consider the more technical aspects of both batch and sequential learning about parameters (i.e., the conditional probabilities) of a directed graphical model, when the prior distributions are assumed to have a Dirichlet form.

7.1 Batch Learning

We first consider learning from a batch of data, where the usual (non-Bayesian) method for determining conditional probabilities from a database is some version of maximum likelihood. As in our running example, one must be able to handle data with a massive quantity of missing observations. In Lauritzen (1991) it is shown how to exploit the EM algorithm for computation of the maximum likelihood estimates of the unknown probabilities, assuming that N cases have been observed, and that for case i we have observed evidence $\mathcal{E}_i$. Specifically, it involves iteratively letting at the M step

$$(11) \qquad p(v_j|\mathrm{pa}(v)^*) = n(v_j, \mathrm{pa}(v)^*) \,/\, n(\mathrm{pa}(v)^*)\,,$$

where n are the expected number of observations in the marginal table: these expectations are calculated in the E step by adding results of a probability propagation, using the current estimated conditional probabilities, for each case in the database; for example,

$$n(\mathrm{pa}(v)^*) = \sum_{i=1}^{N} p(\mathrm{pa}(v)^*|\mathcal{E}_i)\,.$$

Experience (Thiesson, 1991) indicates that when there are as many missing data as in the CHILD example, the likelihood function has a number of local maxima and straight maximum likelihood gives results with unsuitably extreme probabilities at the unobserved intermediate nodes in the network. Hence it seems appropriate to exploit prior information and penalize the likelihood for deviating from the prior assessments, and the EM-algorithm applies almost as easily for maximizing various penalized likelihoods (Green, 1990). If the likelihood function is multiplied by the Dirichlet prior density described in subsection 4.1, the posterior mode can be found iteratively by replacing (11) with

$$p(v_j|\mathrm{pa}(v)^*) = \frac{n(v_j, \mathrm{pa}(v)^*) + \alpha_j - 1}{n(\mathrm{pa}(v)^*) + \alpha - k}\,,$$

provided this remains positive. If any α_j are less than 1, the posterior distribution may not have a mode in the interior and the above expression can turn negative. To avoid this we increased the precision at places where $\alpha_j < 1$ in order to ensure $\alpha_j \geq 1$ everywhere.

A more suitable approach may be to penalize the likelihood directly by interpreting the α values as counts in a likelihood, leading to the iteration

$$p(v_j|\mathrm{pa}(v)^*) = \frac{n(v_j, \mathrm{pa}(v)^*) + \alpha_j}{n(\mathrm{pa}(v)^*) + \alpha}\,.$$

The resulting estimates are posterior modes if densities are calculated with respect to a suitably chosen (improper) prior or, alternatively, they are approximately equal to posterior means.

We now come back to the part of the network discussed in subsection 4.2, in which the imprecise conditional probabilities for the links *Disease?* → *LVH?* and *LVH?* → *LVH-report?* were translated to Dirichlet

priors. Available data on this part of the network comprise 141 cases in which an LVH-report and the true disease are available: the true LVH status is not currently available. The observed frequency of *LVH-report?* = yes in the six diseases was 2/21 (0.095), 1/49 (0.020), 1/34 (0.029), 12/13 (0.923), 0/11 (0.000) and 2/13 (0.154). Table 7 shows the results for the batch learning procedure using posterior modes and posterior means. We note the tendency for the posterior modes to be at extreme values, and convergence for these estimates was much slower.

Two extensions are currently being investigated. First, from the second derivative of the (penalized) likelihood it is possible to obtain approximate expressions for the precision of the estimated values, which will allow standard error estimates for conditional probabilities. Second, the EM algorithm is known to converge relatively slowly. A closer look reveals that, in suitable parametrisations, the gradient of the likelihood function can be calculated with essentially the same amount of work as is involved in the E-step of the EM algorithm. It is therefore conceivable that algorithms exploiting gradient information could be preferable.

It seems reasonable to use these batch learning techniques to initialize a system with conditional probabilities obtained from a combination of prior information and currently available data. Then as more data accumulate one can proceed with the sequential procedure described below.

7.2 Sequential Learning

We now consider the situation in which data arrive one case at a time, and after observing evidence $\mathcal{E}$ on a current case, we wish to revise our beliefs concerning θ_v. Spiegelhalter and Lauritzen (1990) and Spiegelhalter and Cowell (1992) discuss this process in detail, and so here we provide only a brief summary.

In subsection 4.1 we noted that if we observe $v = v_j$ and $\mathrm{pa}(v) = \mathrm{pa}(v)^*$, then the updated distribution of $\theta_{v|\mathrm{pa}(v)^*}$ is a Dirichlet denoted $\mathfrak{D}_j$. Suppose now that neither v nor its parents are observed with certainty. It is then straightforward to show that, assuming local and global independence, the correct posterior distribution for $\theta_{v|\mathrm{pa}(v)^*}$ is

$$p(\theta_{v|\mathrm{pa}(v)^*}|\mathcal{E}) = \sum_j \mathfrak{D}_j p(v_j|\mathrm{pa}(v)^*, \mathcal{E})\, p(\mathrm{pa}(v)^*|\mathcal{E}) + \mathfrak{D}_0 (1 - p(\mathrm{pa}(v)^*|\mathcal{E})), \tag{12}$$

where $\mathfrak{D}_0$ is the initial distribution $\mathfrak{D}[\alpha_1, \ldots, \alpha_K]$. Expression (12) is a mixture of the posterior distribution had v, $\mathrm{pa}(v)^*$ been observed, plus a term that is the unchanged prior weighted by the chance that the relevant parent configuration had not occurred. This is the correct marginal posterior distribution under the expressed local and global independence assumptions, but with general patterns of missing data (12) will need approximation to prevent an explosion of terms. This is strongly related to unsupervised learning, and our approach follows that of the "probabilistic editor" (Titterington, Smith and Makov 1985), in which the approximating distribution attempts to match the moments of the correct mixture (12). There is a degree of arbitrariness in what aspect of the Dirichlet distribution is matched—we equate the "average" variance of the two distributions (see references for the details). Other techniques for approximating a mixture distribution could also be used.

As a fairly simple example, consider the distribution $\theta_{LVH|\mathrm{disease=PAIVS}}$. The first line of Table 8 shows the parameters of the prior distribution $\mathfrak{D}_0$ as derived in Table 2, together with its total precision 3.28, its mean 0.9 and variance 0.021. If we next observe a case with PAIVS and *LVH* = yes, then the posterior distribution ($\mathfrak{D}_1$) has precision increased by 1 and a mean of 0.923, which will be the value used for $p(LVH =$

TABLE 7

Estimates of conditional probabilities obtained from batch learning (using posterior mode and mean) and sequential learning using matching of moments approximation

	Posterior mode	Posterior mean	Sequential	Seq. precision
		LVH? = yes		
Disease?				
PFC	0.056	0.088	0.062	16.4
TGA	0.000	0.016	0.011	40.1
Fallot	0.003	0.026	0.020	30.8
PAIVS	1.000	0.959	0.965	13.1
TAPVD	0.004	0.035	0.035	32.1
Lung	0.079	0.139	0.106	11.4
		LVH-report? = yes		
LVH?				
Yes	0.937	0.916	0.913	29.6
No	0.020	0.018	0.021	112.9

TABLE 8
Exact and approximate distribution of $\theta_{LVH\,|\,disease\,=\,PAIVS}$ having observed different evidence on LVH? and Disease?

Distribution	α_1	α_2	α	Mean	Variance
Prior ($\mathfrak{D}_0$)	2.95	0.33	3.28	0.900	0.0210
Posterior having observed					
(a) *LVH* = yes and PAIVS ($\mathfrak{D}_1$)	3.95	0.33	4.28	0.923	0.0135
(b) *LVH* = no and PAIVS ($\mathfrak{D}_2$)	2.95	1.33	4.28	0.690	0.0405
(c) *LVH-report* = yes and PAIVS	3.82	0.32	4.14	0.922	0.0140
(d) *LVH-report* = yes and p(PAIVS|$\mathcal{E}$) = 0.765	3.59	0.32	3.91	0.917	0.0155

The mean of $\theta_{LVH|disease\,=\,PAIVS}$ is simply the current estimate of p(LVH|PAIVS).

yes|PAIVS) when the next case is processed. If PAIVS and *LVH* = no is observed, then the mean of the posterior ($\mathfrak{D}_2$) drops to 0.690. If we only observe PAIVS and *LVH-report* = yes, then we have that $p(LVH = \text{yes}|\mathcal{E}) = 0.994$, and our true posterior distribution is a mixture of ($\mathfrak{D}_1$) and ($\mathfrak{D}_2$), with slightly adjusted mean 0.922 and variance 0.0140. A single beta distribution with these moments has parameters shown in row (c) of Table 8—we note that we have added evidence equivalent to 0.86 (4.14–3.28) of a full case.

Finally, if we only observe the evidence shown in Figure 4, then we have additionally that p(PAIVS|$\mathcal{E}$) = 0.765, and we have directly observed neither the node of interest nor its parent. Our correct posterior distribution is now a mixture of three terms, and its mean and variance can be calculated to be 0.917 and 0.0155. Row d of Table 8 shows the parameters of a single beta distribution with these moments, suggesting that our evidence may be considered roughly equivalent to having observed 0.64 of a case of PAIVS with LVH.

Table 7 shows the consequences of this learning algorithm applied sequentially to the whole dataset. We note the similarity of the sequential point estimates and the batch posterior means. From Table 2 we may calculate total initial precisions $\alpha_{Disease?} = 39.0$ and $\alpha_{LVH?} = 42.5$, representing the number of implicit cases underlying the prior assessments of conditional probabilities for *LVH?* and *LVH-report?*, respectively. The final total precisions are seen from Table 7 to be 143.9 and 142.5. Since 141 cases have been observed, the sequential procedure has essentially obtained around 80% of the precision that would have been obtained had full data on *LVH?* been available.

Spiegelhalter and Cowell (1992) explore this sequential learning procedure algebraically and through simulation, and show a number of attractive and not-so-attractive properties. For example, it is quite feasible for the total precision α to *decrease* on receipt of certain incomplete evidence: this seems quite reasonable. However, they also show that with systematic missing data on intermediate nodes such as in the CHILD network, the estimation procedure may be inconsistent and strongly reliant on the prior distribution. Therefore considerable care is required when specifying priors for nodes that are not observed, and it may be preferable to marginalise over nodes that are not to be observed and learn on this collapsed graph. Olesen, Lauritzen and Jensen (1992) report similar experiments, adding the feature that precision is gradually decreased so as to put less weight on early cases.

8. COMPATIBLE PRIORS IN ALTERNATIVE MODELS

In Section 6 we introduced the idea of comparing alternative models through their predictive ability assessed by logarithmic penalty (the global monitor), which is equivalent to the Bayes factor procedure for Bayesian model comparison. Here we address the technical problem of assigning Dirichlet prior distributions within competing structures.

It is not generally feasible to reassess subjective probabilities for each possible model, and in any case it seems desirable for the comparison to be made between alternative qualitative structures, uninfluenced by the quantitative inputs. We would wish the models to have consistent prior beliefs where possible. We therefore need a procedure for adapting the initial prior assessments made on a baseline model to other structures that we may wish to entertain. Using evidence propagation it is in principle possible to calculate for any random variable its conditional probability distribution given *any* set of random variables in the model (not necessarily its parents). Hence for any other model structure involving the same set of random variables (or possibly a subset), it is possible to calculate a complete set of conditional probability distributions which both specifies the variant model numerically and which are compatible with the original model.

This process becomes more difficult if the baseline model is specified numerically by a set of imprecise conditional probability tables. This is because the precisions within each model should in some sense match, but how this should be achieved may not be clear. Here we make some basic suggestions.

Consider a node v which has parent sets $\text{pa}^1(v)$ and $\text{pa}^2(v)$ in two models, where imprecise quantitative assessments have been obtained for model 1. First, for

each configuration $\text{pa}^2(v)^*$, we can obtain the conditional probability distribution $p(v|\text{pa}^2(v)^*)$ appropriate for model 2, by simply instantiating evidence $\text{pa}^2(v)^*$ and propagating in model 1 to give $p^1(v|\text{pa}^2(v)^*)$. Thus it is straightforward to derive the point prior probability assessments necessary for any competitor network.

We now consider the precisions of these assessments and essentially obtain them by redistributing the implicit cases underlying the precisions in model 1. Let $\text{pa}^1(v) \cap \text{pa}^2(v) = I$, $\text{pa}^1(v) \setminus I = O$, $\text{pa}^2(v) \setminus I = N$, indicating the *I*ntersecting, the *O*ld and the *N*ew parents of v. For any parent configuration $\text{pa}^1(v)^*$ let $\alpha_{O^*I^*}$ be the precision of the Dirichlet distribution for $p(v|\text{pa}^1(v)^*)$. We first *expand* these "cases" to include the new parent nodes N, to give for a particular configuration of O, I, N a precision

$$\alpha_{O^*I^*N^*} = \alpha_{O^*I^*} p^1(N^*|O^*I^*)\,,$$

where $p^1(N^*|O^*I^*)$ is again obtained using the propagation scheme in model 1.

We now *contract* out the old parent nodes to give

$$\alpha_{I^*N^*} = \sum_{O^*} \alpha_{O^*I^*N^*} = \sum_{O^*} \alpha_{O^*I^*} p^1(N^*|O^*I^*)\,.$$

Table 9 shows this process of expansion and contraction for the section of the network examined in Sections 4 and 7, assuming we wish to collapse out the unobserved node *LVH?*. For node *LVH-report?* there are no intersecting parents, $O = \{LVH?\}$, $N = \{Disease?\}$. From Table 2 we have that the relevant precisions are $\alpha_{LVH?} = (20.3, 22.5)$. These are redistributed according to $p(Disease?|LVH = \text{yes}) = (0.04, 0.13, 0.10, 0.70, 0.01, 0.02))$, and $p(Disease?|LVH = \text{no}) = (0.13, 0.42, 0.30, 0.03, 0.06, 0.06)\}$, and then finally summed over *LVH?*. The total precision has remained constant at 42.9.

This procedure has some attractive properties. First, the total precision remains constant for each node. Second, if the estimates are all based on implicit data equivalent to a complete sample of size n, then in the original network we have the identity $\alpha_{O^*I^*} = np(O^*I^*)$. It is easily seen that this property is retained in model 2, since

$$\alpha_{I^*N^*} = \sum_{O^*} np(O^*I^*) p^1(N^*|O^*I^*) = np(I^*N^*)\,.$$

However, the procedure does ignore the imprecision associated with the $p^1(N^*|O^*I^*)$ and hence could give rather precise assessments. At an extreme, if we assume in model 1 that $p(v|\text{pa}^1(v)^*)$ are specified precisely, then $p(v|\text{pa}^2(v)^*)$ will also be specified precisely.

9. OUTLOOK

The work described can be seen as just one aspect in a general development in which complex stochastic models are constructed by modular combination of components with simple local structure and in which analysis is performed between communicating local elements. More traditional statistical application areas for this approach include image processing, dynamic models and general multivariate analysis. In each context, graphical models are an attractive medium for communicating the essentials of a problem between a subject matter specialist and a constructor of a formal model. It can happen that the graph will be too complex to allow the use of the exact techniques described here, but we feel these methods for structured probabilistic reasoning will prove useful beyond the motivating field of expert systems.

An important development that will be seen in the coming years is the extension of the basic methods described in the present paper to deal with a variety of model types, so that the technique is not limited to recursive graphical models for variables with a discrete state space.

Such developments include the ability to deal with a more sophisticated dependence structure using the Markov property for so-called chain graphs (Frydenberg, 1990). Chain graph models admit undirected links and therefore enable the researcher to accommodate associations that are of "correlation" rather than "causation" type, and to deal with the kind of bidirectional links discussed in Section 2. The program BIFROST mentioned in subsection 5.4 is based upon exploiting a subclass of these models.

Another line of development is concerned with admitting real-valued random variables. When the CG (Conditional Gaussian) distributions of Lauritzen and Wermuth (1989) are used, the propagation technique can be generalised to give fast computation of correct means and standard deviations of variables (Lauritzen, 1992); for a survey of chain graph models based upon

TABLE 9

Redistribution of precision associated with p(LVH-report? | LVH?) to new table p(LVH-report? | Disease?) upon removal of node LVH?

	Disease?						
	PFC	TGA	Fallot	PAIVS	TAPVD	Lung	$\alpha_{LVH?}$
$\alpha_{LVH?\,=\,\text{yes},\,Disease?}$	0.8	2.6	2.0	14.2	0.2	0.4	20.3
$\alpha_{LVH?\,=\,\text{no},\,Disease?}$	2.9	9.5	6.8	0.7	1.4	1.4	22.5
$\alpha_{Disease?}$	3.7	12.1	8.8	14.9	1.6	1.8	

CG distributions, see Lauritzen (1989) or Whittaker (1990). Naturally, a variety of other distributions in graphical models can be introduced if one is willing to substitute exact propagation with Monte-Carlo methods (Thomas, Spiegelhalter and Gilks, 1992), with both approaches playing a role in models that accommodate strong spatial and temporal dependencies (see subsection 6.8).

ACKNOWLEDGMENTS

We are indebted to Dr. Kate Bull, consultant paediatric cardiologist at Great Ormond Street Hospital, for the time and expertise given to this research. We are grateful to Bo Thiesson for performing the batch learning experiments and HUGIN Expert Ltd. for the use of software. Comments from Guido Consonni, the referees and the Editor were extremely valuable. Grant support was provided by the Science and Engineering Research Council Complex Stochastic Systems Initiative, and under the EEC SCIENCE Laboratory Twinning scheme.

REFERENCES

ADAMS, I. D., CHAN, M., CLIFFORD, P. C., COOKE, W. M., DALLOS, V., DE DOMBAL, F. T., EDWARDS, M. H., HANCOCK, D. M., HEWETT, D. J., MCINTYRE, N., SOMERVILLE, P. G., SPIEGELHALTER, D. J., WELLWOOD, J. and WILSON, D. H. (1986). Computer-aided diagnosis of acute abdominal pain: a multicentre study. *British Medical Journal* 292 800–804.

ANDERSEN, K. and HOOKER, J. (1992). Bayesian logic. *Decision Support Systems.*

ANDERSEN, L. R., KREBS, J. H. and ANDERSEN, J. D. (1991). STENO: An expert system for medical diagnosis based on graphical models and model search. *Journal of Applied Statistics* 18 139–153.

ANDERSEN, S. K., OLESEN, K. G., JENSEN, F. V. and JENSEN, F. (1989). HUGIN—A shell for building Bayesian belief universes for expert systems. In *Proceedings of the 11th International Joint Conference on Artificial Intelligence* 1080–1085. Morgan Kaufmann, San Mateo, CA. [Also reprinted in Shafer and Pearl (1990).]

ANDREASSEN, S., WOLDBYE, M., FALCK, B. and ANDERSEN, S. (1987). MUNIN—a causal probabilistic network for the interpretation of electromyographic findings. In *Proceedings of the 10th National Conference on AI* 121–123. AAAI: Menlo Park, CA.

BADSBERG, J. H. (1991). A guide to CoCo. Technical Report R-91-43, Dept. Mathematics and Computer Science, Aalborg Univ.

BERZUINI, C., BELLAZZI, R., QUAGLINI, S. and SPIEGELHALTER, D. J. (1992). Bayesian networks for patient monitoring. *Artificial Intelligence in Medicine* 4 243–260.

BOX, G. E. P. (1980). Sampling and Bayes inference in scientific modelling and robustness (with discussion). *J. Roy. Statist. Soc. Ser. A* 143 383–430.

BOX, G. E. P. (1983). An apology for ecumenism in statistics. In *Scientific Inference, Data Analysis and Robustness* (G. E. P. BOX, T. Leonard and C. F. Wu, eds.) 51–84. Academic, New York.

CANNINGS, C., THOMPSON, E. and SKOLNICK, M. (1978). Probability functions on complex pedigrees. *Adv. in Appl. Probab.* 10 26–61.

CHARNIAK, E. and GOLDMAN, R. (1989). Plan recognition in stories and in life. In *Uncertainty in Artificial Intelligence* (M. Henrion, R. D. Shachter, L. Kanal and J. Lemmer, eds.) 5 54–60. North-Holland, Amsterdam.

COOPER, G. F. (1990). The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence* 42 393–405.

COOPER, G. and HERSKOVITS, E. (1992). A Bayesian method for the induction of probabilistic networks from data. *Machine Learning* 9 309–347.

COWELL, R. G. (1992). BAIES—a probabilistic expert system shell with qualitative and quantitative learning. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 595–600. Clarendon Press, Oxford.

COWELL, R. G. and DAWID, A. P. (1992). Fast retraction of evidence in a probabilistic expert system. *Statistics and Computing* 2 37–40.

COWELL, R. G., DAWID, A. P. and SPIEGELHALTER, D. J. (1993). Sequential model criticism in probabilistic expert systems. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 15 209–219.

COX, D. R. (1958). Two further applications of a model for binary regression. *Biometrika* 45 562–565.

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. (1979a). Conditional independence in statistical theory (with discussion). *J. Roy. Statist. Soc. Ser. B* 41 1–31.

DAWID, A. P. (1984). Statistical theory—the prequential approach. *J. Roy. Statist. Soc. Ser. A* 147 277–305.

DAWID, A. P. (1992). Applications of a general propagation algorithm for probabilistic expert systems. *Statistics and Computing* 2 25–36.

DAWID, A. P. and LAURITZEN, S. L. (1993). Hyper Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* 21.

DE DOMBAL, F., LEAPER, D., STANILAND, J., MCCANN, A. and HORROCKS, J. (1972). Computer-aided diagnosis of acute abdominal pain. *British Medical Journal* 2 9–13.

EDWARDS, D. and HAVRÁNEK, T. (1985). A fast procedure for model search in multidimensional contingency tables. *Biometrika* 72 339–351.

EDWARDS, D. and HAVRÁNEK, T. (1987). A fast model selection procedure for large families of models. *J. Amer. Statist. Assoc.* 82 205–211.

FRANKLIN, R. C. G., SPIEGELHALTER, D. J., MACARTNEY, F. and BULL, K. (1989). Combining clinical judgements and statistical data in expert systems: Over the telephone management decisions for critical congenital heart disease in the first month of life. *International Journal of Clinical Monitoring and Computing* 6 157–166.

FRANKLIN, R. C. G., SPIEGELHALTER, D. J., MACARTNEY, F. and BULL, K. (1991). Evaluation of a diagnostic algorithm for heart disease in neonates. *British Medical Journal* 302 935–939.

FRYDENBERG, M. (1990). The chain graph Markov property. *Scand. J. Statist.* 17 333–353.

GELFAND, A. E. and SMITH, A. F. M. (1990). Sampling-based approaches to calculating marginal densities. *J. Amer. Statist. Assoc.* 85 398–409.

GEMAN, S. and GEMAN, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 6 721–741.

GLYMOUR, C., SCHEINES, R., SPIRTES, P. and KELLEY, K. (1987). *Discovering Causal Structure.* Academic, New York.

GORDON, J. and SHORTLIFFE, E. H. (1985). A method for managing evidential reasoning in hierarchical hypothesis spaces. *Artificial Intelligence* **26** 323–358.

GREEN, P. J. (1990). On use of the EM algorithm for penalized likelihood estimation. *J. Roy. Statist. Soc. Ser. B* **52** 443–452.

HECKERMAN, D. (1990). Probabilistic similarity networks. *Networks* **20** 607–636.

HECKERMAN, D., HORVITZ, E. and NATHWANI, B. (1992). Toward normative expert systems I: The PATHFINDER project. *Methods of Information in Medicine* **31** 90–105.

HENRION, M., BREESE, J. S. and HORVITZ, E. J. (1991). Decision analysis and expert systems. *AI Magazine* 12 64–91.

HØJSGAARD, S., SKJØTH, F. and THIESSON, B. (1992). User's guide to BIFROST. Technical Report R-92-2001, Dept. Mathematics and Computer Science, Aalborg Univ.

JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Oxford Univ. Press.

JENSEN, F. V. (1991). Calculation in HUGIN of probabilities for specific configurations—a trick with many applications. In *Scandinavian Conference of Artificial Intelligence* 176–186. IOS Press, Amsterdam.

JENSEN, F. V., CHAMBERLAIN, B., NORDAHL, T. and JENSEN, F. (1991). Analysis in HUGIN of data conflict. In *Uncertainty in Artificial Intelligence VI* (P. P. Bonissone, M. Henrion, L. N. Kanal and J. F. Lemmer, eds.) 519–528. North-Holland, Amsterdam.

JENSEN, F. V., CHRISTENSEN, H. I. and NIELSEN, J. (1992). Bayesian methods for interpretation and control in multi-agent vision systems. In *Applications of Artificial Intelligence X* **1708** 536–548. SPIE, Orlando, Florida.

JENSEN, F. V., OLESEN, K. G. and ANDERSEN, S. K. (1990). An algebra of Bayesian belief universes for knowledge-based systems. *Networks* **20** 637–659.

KIM, J. and PEARL, J. (1983). A computational model for causal and diagnostic reasoning in reference systems. In *Proceedings of the 8th International Joint Conference on Artificial Intelligence* 190–193. AAAI, Menlo Park, CA.

KJÆRULFF, U. (1992a). Optimal decomposition of probabilistic networks by simulated annealing. *Statistics and Computing* **2** 19–24.

KJÆRULFF, U. (1992). A computational scheme for reasoning in dynamic probabilistic networks. In *Uncertainty in Artificial Intelligence* (D. Dubois, M. P. Wellman, B. D'Ambrosio and P. Smets, eds.) **8** 121–129. Morgan Kaufmann, San Mateo, CA.

LAURITZEN, S. L. (1989). Mixed graphical association models (with discussion). *Scand. J. Statist.* **16** 273–306.

LAURITZEN, S. L. (1991). The EM algorithm for graphical association models with missing data. Technical Report R 91-05, Institute for Electronic Systems, Aalborg Univ.

LAURITZEN, S. L. (1992). Propagation of probabilities, means and variances in mixed graphical association models. *J. Amer. Statist. Assoc.* **87** 1098–1108.

LAURITZEN, S. L., DAWID, A. P., LARSEN, B. N. and LEIMER, H.-G. (1990). Independence properties of directed Markov fields. *Networks* **20** 491–505.

LAURITZEN, S. L. and SPIEGELHALTER, D. J. (1988). Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J. Roy. Statist. Soc. Ser. B* **50** 157–224.

LAURITZEN, S. L., THIESSON, B. and SPIEGELHALTER, D. J. (1992). Diagnostic systems created by model selection methods—a case study. Technical Report R-92-2018, Institute for Electronic Systems, Aalborg Univ.

LAURITZEN, S. L. and WERMUTH, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Ann. Statist.* **17** 31–57.

LINDLEY, D. V. (1957). A statistical paradox. *Biometrika* **44** 187–192.

MILLER, R. A., POPLE, H. E. and MYERS, J. (1982). Internist-1, an experimental computer-based diagnostic consultant for general internal medicine. *New England Journal of Medicine* **307** 468–476.

MURPHY, A. H. and WINKLER, R. L. (1984). Probability forecasting in meteorology. *J. Amer. Statist. Assoc.* **79** 489–500.

NEAPOLITAN, E. (1990). *Probabilistic Reasoning in Expert Systems*. Wiley, New York.

NILSSON, N. (1986). Probabilistic logic. *Artificial Intelligence* **28** 71–87.

OLESEN, K. G., LAURITZEN, S. L. and JENSEN, F. V. (1992). HUGIN: A system creating adaptive causal probabilistic networks. In *Uncertainty in Artificial Intelligence* (D. Dubois, M. P. Wellman, B. D'Ambrosio and P. Smets, eds.) **8** 223–229. Morgan Kaufmann, San Mateo, CA.

PEARL, J. (1982). Reverend Bayes on inference engines: a distributed hierarchical approach. In *Proceedings of American Association for Artificial Intelligence National Conference on AI, Pittsburgh*, 133–136. AAAI, Menlo Park, CA.

PEARL, J. (1986). Fusion, propagation and structuring in belief networks. *Artificial Intelligence* **29** 241–288.

PEARL, J. (1987). Evidential reasoning using stochastic simulation. *Artificial Intelligence* **32** 245–257.

PEARL, J. (1988). *Probabilistic Inference in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA.

PEARL, J. and VERMA, T. (1991). A theory of inferred causality. In *Proceedings of the Second International Conference on the Principles of Knowledge Representation and Reasoning* (J. A. Allen, R. Fikes and E. Sandewall, eds.) 441–452. Morgan Kaufmann, San Mateo, CA.

REITER, R. (1987). Nonmonotonic reasoning. *Annual Review of Computer Science* **2** 147–186.

SAFFIOTTI, A. and UMKEHRER, E. (1991). Pulcinella: a general tool for propagating uncertainty in valuation networks. In *Uncertainty in Artificial Intelligence* (B. D'Ambrosio, P. Smets and P. P. Bonissone, eds.) **7**. Morgan Kaufmann, San Mateo, CA.

SEILLIER-MOISEIWITSCH, F. and DAWID, A. P. (1993). On testing the validity of probability forecasts. *J. Amer. Statist. Assoc.* **88** 355–359.

SEMBER, P. and ZUCKERMAN, I. (1989). Strategies for generating micro explanations for Bayesian belief networks. In *Uncertainty in Artificial Intelligence* (M. Henrion, R. D. Shachter, L. N. Kanal and J. F. Lemmer, eds.) **5** 295–302. North-Holland, Amsterdam.

SHAFER, G. R. and PEARL, J. (ed.) (1990). *Readings in Uncertain Reasoning*. Morgan Kaufmann, San Mateo, CA.

SHENOY, P. P. (1989). A valuation-based language for expert systems. *Internat. J. Approx. Reason.* **3** 383–411.

SHENOY, P. P. and SHAFER, G. R. (1990). Axioms for probability and belief-function propagation. In *Uncertainty in Artificial Intelligence* (R. D. Shachter, T. S. Levitt, L. N. Kanal and J. F. Lemmer, eds.) **4** 169–198. North-Holland, Amsterdam.

SHORTLIFFE, E. H. and BUCHANAN, B. G. (1975). A model for inexact reasoning in medicine. *Math. Biosci.* **23** 351–379.

SHWE, M., MIDDLETON, B., HECKERMAN, D., HENRION, M., HORVITZ, E. and LEHMANN, H. (1991). Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base I: The probabilistic model and inference algorithms. *Methods of Information in Medicine* **30** 241–255.

SPIEGELHALTER, D. J. (1989). A unified approach to imprecision and sensitivity of beliefs in expert systems. In *Uncertainty*

in Artificial Intelligence (L. N. Kanal, J. Lemmer and T. S. Levitt, eds.) 3 199–208. North-Holland, Amsterdam.

SPIEGELHALTER, D. J. and COWELL, R. G. (1992). Learning in probabilistic expert systems. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 447–466. Clarendon Press, Oxford.

SPIEGELHALTER, D. J., DAWID, A. P., HUTCHINSON, T. A. and COWELL, R. G. (1991a). Probabilistic causality assessment after a suspected adverse drug reaction: a case study in Bayesian network modelling. *Philos. Trans. Roy. Soc. London Ser. A* 337 387–405.

SPIEGELHALTER, D. J., HARRIS, N. L., BULL, K. and FRANKLIN, R. C. G. (1991b). Empirical evaluation of prior beliefs about frequencies: methodology and a case study in congenital heart disease. Technical Report 91-4. MRC Biostatistics Unit, Cambridge.

SPIEGELHALTER, D. J. and LAURITZEN, S. L. (1990). Sequential updating of conditional probabilities on directed graphical structures. *Networks* 20 579–605.

SPIRTES, P., GLYMOUR, C. and SCHEINES, R. (1993). *Causation, Prediction, and Search.* Springer, New York.

SRINIVAS, S. and BREESE, J. (1990). IDEAL: a software package for the analysis of influence diagrams. In *Uncertainty in Artificial Intelligence* (L. N. Kanal, J. Lemmer and T. S. Levitt, eds.) 6 212–219. North-Holland, Amsterdam.

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 J. Comput.* 13 566–79.

THIESSON, B. (1991). (G)EM algorithms for maximum likelihood in recursive graphical association models. Master's thesis, Dept. Mathematics and Computer Science, Aalborg Univ.

THOMAS, A., SPIEGELHALTER, D. J. and GILKS, W. R. (1992). BUGS: A program to perform Bayesian inference using Gibbs sampling. In *Bayesian Statistics 4* (J. M. Bernardo, J. O. Berger, A. P. Dawid and A. F. M. Smith, eds.) 837–842. Clarendon Press, Oxford.

THOMPSON, E. A. (1986). Genetic epidemiology: a review of the statistical basis. *Statistics in Medicine* 5 291–302.

TITTERINGTON, D. M., MURRAY, G. D., MURRAY, L. S., SPIEGELHALTER, D. J., SKENE, A. M., HABBEMA, J. D. F. and GELPKE, G. J. (1981). Comparison of discrimination techniques applied to a complex data-set of head-injured patients (with discussion). *J. Roy. Statist. Soc. Ser. A* 144 145–175.

TITTERINGTON, D. M., SMITH, A. F. M. and MAKOV, U. E. (1985). *Statistical Analysis of Finite Mixture Distributions.* Wiley, Chichester.

VAN DER GAAG, L. (1991). Computing probability intervals under independency constraints. In *Uncertainty in Artificial Intelligence* (P. P. Bonissone, M. Henrion, L. N. Kanal and J. F. Lemmer, eds.) 6 457–466. North-Holland, Amsterdam.

WALLEY, P. (1990). *Statistical Reasoning with Imprecise Probabilities.* Chapman and Hall, London.

WARNER, H. R., TORONTO, A. F., VEASEY, L. G. and STEPHENSON, R. (1961). A mathematical approach to medical diagnosis—application to congenital heart disease. *Journal of the American Medical Association* 177 177–184.

WERMUTH, N. (1976). Model search among multiplicative models. *Biometrics* 32 253–263.

WERMUTH, N. and LAURITZEN, S. L. (1983). Graphical and recursive models for contingency tables. *Biometrika* 70 537–552.

WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Analysis.* Wiley, Chichester.

WRIGHT, S. (1934). The method of path coefficients. *Ann. Math. Statist.* 5 161–215.

ZADEH, L. A. (1983). The role of fuzzy logic in the management of uncertainty in expert systems. *Fuzzy Sets and Systems* 11 199–228.

Comment: Assessing the Science Behind Graphical Modelling Techniques

A. P. Dempster

These papers, labelled here CW (Cox and Wermuth) and SDLC (Spiegelhalter, Dawid, Lauritzen and Cowell), are welcome reviews of extensive collaborations. CW are the more limited of the pair in their aims, making a few points convincingly, most notably (1) that covariance-based regression models are conceptually distinct from the simultaneous causal models of econometrics, even when both varieties are expressed through identical linear equations, and (2) that models with covariance matrices corresponding to restricted graphical structures often give good fits to empirical matrices. The SDLC paper by contrast is a tour de force that aims to leave no relevant topic unmentioned.

Both sets of authors intend their formal models and computations to speak to issues of scientific knowledge and science-based decision making, and in particular both are concerned about the informal scientific understanding that motivates their formal models. CW are reluctant to use the term "causal," viewing it as too ambiguous, but the authors substitute nonspecific language such as "appropriate subject matter considerations." SDLC, in contrast, discuss "influence" and "relevance" that take "account of one's understanding of causal structure." The difference appears to be that CW wish to hold to the idea that informal prior knowl-

A. P. Dempster is Professor of Statistics, Harvard University, Statistics Department, Science Center, 1 Oxford Street, Cambridge, Massachusetts 02138.