Sequential Updating of Conditional Probabilities on Directed Graphical Structures

David J. Spiegelhalter
MRC Biostatistics Unit, Cambridge CB2 2BW, England
Steffen L. Lauritzen
Department of Mathematics and Computer Science, Institute of Electronic Systems, Aalborg University, DK-9000 Aalborg, Denmark

A directed acyclic graph or influence diagram is frequently used as a representation for qualitative knowledge in some domains in which expert system techniques have been applied, and conditional probability tables on appropriate sets of variables form the quantitative part of the accumulated experience. It is shown how one can introduce imprecision into such probabilities as a data base of cases accumulates. By exploiting the graphical structure, the updating can be performed locally, either approximately or exactly, and the setup makes it possible to take advantage of a range of well-established statistical techniques. As examples we discuss discrete models, models based on Dirichlet distributions and models of the logistic regression type.

1. INTRODUCTION

In a recent article (Lauritzen and Spiegelhalter [6], henceforth abbreviated as L–S), we described algorithms for fast computation of conditional and marginal probabilities in influence diagrams or "causal networks," in which independence properties of an initial joint distribution of a set of random variables X_v, $v \in V$, are related to an acyclic directed graph with each vertex (node) representing a random variable in V. The value of this graphical representation has been stressed by, for example, Pearl [9,10], Shachter [13], and Cooper [12].

A clear limitation of this approach is that it is assumed that the conditional probabilities necessary to initialize the system are specified precisely. This is liable to be unrealistic, whether the component probabilities are derived from subjective assessments or are based on specific data, and there are a number of reasons for wishing to be able to retain an explicit representation of this inevitable imprecision. First, this will allow probabilistic predictions provided by the system to be tempered with an allowance for possible error, traceable back to the current imprecision in the quantities held in the system. Second, it is only by allowing imprecision that the probabilities can be updated and im-

NETWORKS, Vol. 20 (1990) 579–605

CCC 0028-3045/90/050579-027$04.00

proved in the light of new data. Finally, the procedure for obtaining the conditional probability tables becomes more acceptable if doubt about subjective assessments can be incorporated or sampling error in data-based estimates recorded.

Essentially, we wish to process each individual case as best we can, but allow the quantitative aspects of the experience gained to be carried over to future cases. Figure 1 shows a general structure for this.

The *experience* is the quantitative memory, which may be based both on quantitative expert judgment and past cases and which is relevant to the processing of a future case. This provides the quantitative input to the *core* which expresses the qualitative relationships between the features of the case in hand, and the experience is *disseminated* into the core prior to the case-specific information being obtained. The case is then processed using whatever data have been gleaned, and the additional knowledge gained is then *retrieved* by the experience. The updated experience is then ready for passing on to a further case.

Within a Bayesian statistical paradigm, these general ideas become familiar operations. Specifically, the experience will consist of a distribution $p(\theta)$ over a parameter space Θ, and the core will specify the joint distribution $p(V|\theta)$ for a particular realization $\theta \in \Theta$. Here we adopt the abbreviated notation from L–S such that $p(V|\theta)$ is short for $p(X_v = x_v, v \in V|\theta)$ and so on. Dissemination of experience now involves integrating out over θ to produce a marginal distribution

$$p(V) = \int p(V)|\theta)p(\theta)d\theta \tag{1}$$

from which $p(V)$ may be used to process the case. Having observed some data E^*, the retrieval operation comprises the calculation and storage of the updated posterior probability distribution $p(\theta|E^*)$.

In most contexts it is reasonable that the qualitative core should also adapt

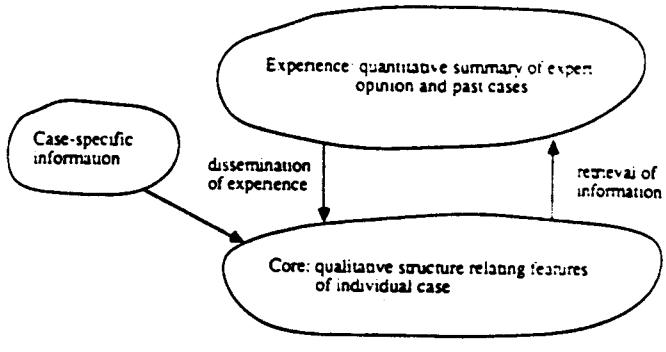

FIG. 1. General structure of case-to-case updating of experience.

its structure as a data base of cases accumulates, and this general framework allows cases to have different core structures and for the case specific information to include external elements such as control variables. However, in many situations there will be at least a period in which cases can be considered *exchangeable*, in that the core is common to each and conditional on a particular parametrization $\theta \in \Theta$, the feature vector V_i, $i = 1, \ldots, n$ for a set of n cases are independent each with distribution $p(V_i|\theta)$. Equivalently, the joint distribution $p(V_1, \ldots, V_n)$ is invariant to the order in which the cases are actually observed and is given by

$$p(V_1, \ldots, V_n) = \int \prod_{i=1}^{n} p(V_i|\theta)p(\theta)d\theta.$$

Although it will be common to assume exchangeability over a set of cases judged to be homogeneous, the following development will only deal with the operations for a single case, and, hence, exchangeability is not an essential prerequisite.

2. DIRECTED ACYCLIC GRAPHS AS CORE STRUCTURES

3.1. The General Model

The basis for the algorithms of L–S rests on assuming that the qualitative core structure may be represented as a directed acyclic graph, in which each variable $v \in V$ forms a node with a set of "parent" nodes $\text{pa}(v)$, where for each $w \in \text{pa}(v)$ there exists a directed link $w \to v$. The quantitative specification comprises a conditional probability distribution for v for each configuration of parents $\text{pa}(v)$. These conditional probability tables form the fundamental components from which case-specific inferences are eventually derived and are the natural parametrization for initialization from clinical opinion or available data. It therefore seems reasonable to break a general parametrization θ, into components θ_v corresponding to each node v. Hence, $\Theta = \times_{v \in V} \Theta_v$, and $\theta_v \in \Theta_v$ completely specifies the relationship between a node v and its parents $\text{pa}(v)$. The conditional probability table corresponding to a realization θ_v is denoted $p(v|\text{pa}(v), \theta_v)$, and, hence, our conditional distribution on V is

$$p(V|\theta) = \prod_{v \in V} p(v|\text{pa}(v), \theta_v) \tag{2}$$

using the conditional independence assumptions represented in the directed graph (L–S).

In order to explore Bayesian updating of experience we need to specify a form for the prior distribution $p(\theta)$. This must depend both on what is known about the context and on more pragmatic issues of computational and representational simplicity. In some applications of influence diagrams, particularly

genetics [16]. θ may have only a few elements related to unknown quantities in, say, the Mendelian segregation probabilities, genotype-phenotype penetrance function, and population frequency of alleles. Such a parametrization makes the likelihood $p(E^*|\theta)$, $E \subseteq V$ reasonably straightforward to calculate [18] and to be used as a basis for inference on θ. A posterior distribution $p(\theta|E^*)$ could also be carried over to future cases, which would take the form of analysis of similar traits on the same genealogy. However, such a parsimonious parametrization, in which a single parameter enters into many conditional probability tables, tends to give a complex structure for the joint distribution (1) and, hence, makes it difficult to calculate efficiently conditional probabilities on individual cases, say $p(D|E^*)$, where $D \subseteq V \setminus E$, which are not made conditional on $\theta \in \Theta$. It is precisely this kind of case-specific probability statement, *conditional* on known case-data, but *unconditional* on unknown parameters, that is crucial in making individualized predictions in applications such as expert systems.

In this paper we therefore explore in some detail the simplifying assumptions that will provide the marginal distribution (1) with the simplest possible structure. Our major assumption is that of *global independence*, i.e., the parameters $\{\theta_v, v \in V\}$ are *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

$$p(V,\theta) = \prod_v p(v|\mathrm{pa}(v),\theta_v)p(\theta_v)\,. \tag{3}$$

From (3) it is clear that θ_v may be considered, formally, as another parent of v in a general influence diagram, such as that shown in Figure 2 for the example introduced in L–S. This diagram expresses, for example, that θ_τ is a random quantity whose realization, possibly vector-valued, fully specifies the conditional probability table $p(\tau|\alpha)$. We note that even if conditional probabilities on V are known, such as in the logical connection $p(\epsilon|\tau,\lambda)$, it is convenient to retain an explicit representation of θ_ϵ as a random quantity with a degenerate distribution.

For readers familiar with the details of the L–S procedure, we note that, in theory, it would be possible to treat (V,θ) in a uniform fashion, not distinguishing between core and experience variables and, hence, carry out the operation of "marrying" and "filling-in" on the influence diagram corresponding to (3) as described in L–S and forming cliques as the basic representation for processing of evidence. In particular, we note that we may form a "moral graph" by joining parents $(\mathrm{pa}(v),\theta_v)$ of each node v and dropping directions. From (2), the joint distribution of (V,θ) can be seen to be expressed in terms of functions on joined nodes in this moral graph and, hence, forms a Markov random field. An example of such a moral graph is given in Figure 3.

By adding the dotted "fill-in" in Figure 3, we have a triangulated graph. We note that this operation requires no additional fill-in beyond those required for the triangulation of the core graph. This can be seen by the following argument.

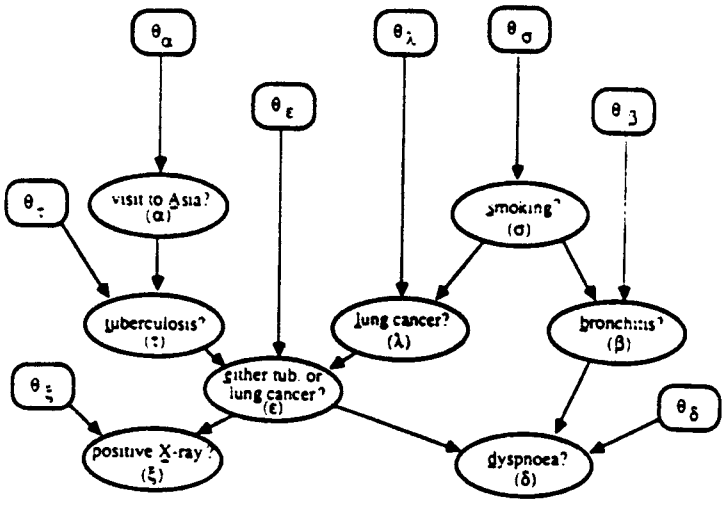

FIG. 2. Influence diagram representation of our assumption that "core" variables $V = \{\alpha, \tau, \xi, \epsilon, \delta, \lambda, \beta, \sigma\}$ have conditional independence properties described in L–S, with conditional probability tables generated by realization of marginally independent random quantities $\theta_v, v \in V$ forming a level of "experience."

A graph is triangulated if there exists a node ordering such that for each node all lower-ordered neighbors form complete sets (are mutual neighbors) (L–S). Consider such an ordering of the triangulated core, and add each node θ_v in turn, joined to v and to the parents pa(v); since parents are married, $v \cup \text{pa}(v)$

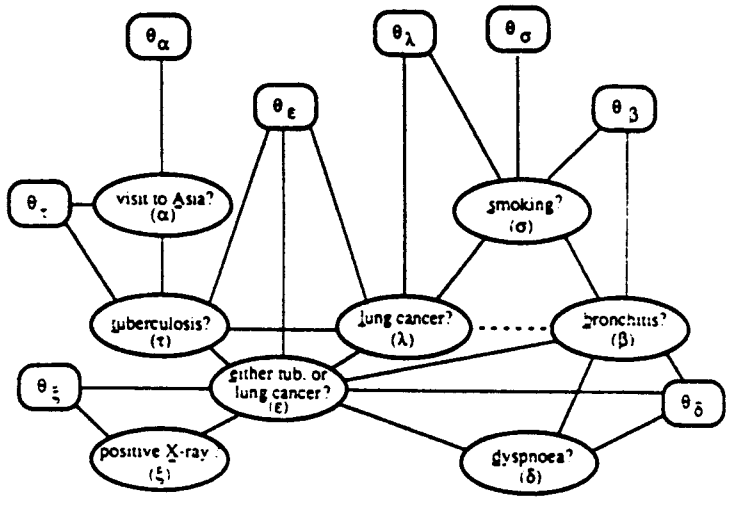

FIG. 3. Moral graph (solid lines) formed by joining parents and dropping directions from Figure 2. (Dashed line is fill-in needed for triangulation.) Joint distribution on (V, θ) is Markov on this moral graph.

are all joined and the boundary of θ_v is complete. It then follows that the overall graph is triangulated without additional fill-ins.

However, although this general structure may be useful in providing properties of the representation, it appears most useful to exploit the experience/core structure, and we now consider the processes of "dissemination" of experience and "retrieval" of information.

2.2. Dissemination and Retrieval of Information

Dissemination involves integrating out parameters in order to process a new case, and we obtain that

$$p(v) = \int p(v,\theta)\, d\theta = \int \prod_v p(v|\mathrm{pa}(v),\theta_v)p(\theta_v)\, d\theta_v = \prod_v p(v|\mathrm{pa}(v)), \qquad (4)$$

where

$$p(v|\mathrm{pa}(v)) = \int p(v|\mathrm{pa}(v),\theta_v)p(\theta_v)\, d\theta_v \qquad (5)$$

is simply the expectation of the local conditional probability table for v. Hence, our assumption implies that the global dissemination operation can be performed locally. (Note that in some parametrizations approximation may be necessary in carrying out this local operation.)

We emphasise that expression (4) shows that the conditional independence assumptions concerning the features V have been preserved under dissemination, and, hence, the standard evidence propagation described in L–S can take place, essentially using the "mean probability values" derived from the experience. After observing, say, evidence E^* on a current case, revised beliefs $p(v \cup \mathrm{pa}(v)|E^*)$ may be easily obtained as the set $v \cup \mathrm{pa}(v)$ will always be a subset of a clique whose revised marginal distribution will have been calculated in the L–S procedure.

We can now retrieve the new information about θ_v by the following operation. Since $v \cup \mathrm{pa}(v)$ separates θ_v from the remaining nodes in the moral graph, it follows that θ_v is conditionally independent of E^* given $v \cup \mathrm{pa}(v)$. Hence, a revised opinion concerning θ_v is given as

$$p(\theta_v|E^*) = \sum_{v \cup \mathrm{pa}(v)} p(\theta_v|v \cup \mathrm{pa}(v))p(v \cup \mathrm{pa}(v)|E^*). \qquad (6)$$

We refer to this calculation (6) as the *local retrieval* of information. We first note that a *local inversion* is required to obtain

$$p(\theta_v|v \cup \mathrm{pa}(v)) \propto p(v|\mathrm{pa}(v),\theta_v)p(\theta_v). \qquad (7)$$

Second, unless the configuration at nodes in $v \cup \mathrm{pa}(v)$ is observed, (6) forms a

mixture distribution. Thus, although this local retrieval formula is exact as it stands, we shall see that it may be convenient to approximate both these operations within particular parametrizations.

Although (6) provides the correct marginal posterior distribution for θ_v, it cannot be guaranteed without conditions explored in the next section that the θ_v's are still marginally independent. However, our basic philosophy is to approximate the process of *global retrieval* by a sequence of local retrievals, assume independence of the θ_v's, and leave the experience in the same localized form for the next case. The consequences of this approximation are to be explored.

2.3 Local Independence

A considerable simplification in the local retrieval operation is achieved if we assume what we shall term *local independence*. The distribution of v conditional on each configuration $x_{\mathrm{pa}(v)}$ of its parent set is then individually parametrized, and these parameters are assumed *a priori* independent. Thus,

$$\Theta_v = \times_{x_{\mathrm{pa}(v)} \in X_{\mathrm{pa}(v)}} \Theta_{v|x_{\mathrm{pa}(v)}}$$

and $\theta_{v|x_{\mathrm{pa}(v)}}$ are independent.

Consider then a particular configuration of parent nodes and denote this $\mathrm{pa}(v)^+$. Then θ_v^+ parametrizes the conditional probability table $p(v|\mathrm{pa}(v)^+, \theta_v^+)$, and conditional on $v \cup \mathrm{pa}(v)^+, \theta_v^+$ is independent of the remaining parameters $\theta_v \backslash \theta_v^+$. In addition, (6) is the sum of a set of terms for each parent configuration, with contribution from, say, $\mathrm{pa}(v)^+$ of

$$p(\theta_v \backslash \theta_v^+) \sum_v p(\theta_v^+ | v \cup \mathrm{pa}(v)^+) p(v \cup \mathrm{pa}(v)^+ | E^*) .$$

In general, this will create posterior dependence between the $\theta_{v|x_{\mathrm{pa}(v)}}$'s unless the parent configuration is observed to be, say $\mathrm{pa}(v)^+$, in which case posterior independence is retained and the only change is to revise the distribution of θ_v^+ to be

$$p(\theta_v^+ | E^*) = \sum_v p(\theta_v^+ | v \cup \mathrm{pa}(v)^+) p(v \cup \mathrm{pa}(v)^+ | E^*)$$

since $\mathrm{pa}(v)^+ \subseteq E^*$. This may itself need to be approximated unless v is also observed. Even if the true parent configuration is not observed, we may approximate by making the local independence assumption and just retain the true marginal distribution

$$p(\theta_v^+|E^*) = \sum_v p(\theta_v^+|v \cup \text{pa}(v)^+)p(v \cup \text{pa}(v)^+|E^*)$$

$$+ \sum_v \sum_{\text{pa}(v)^* \neq \text{pa}(v)^+} p(\theta_v^+|v \cup \text{pa}(v)^*)p(v \cup \text{pa}(v)^*|E^*)$$

$$= \sum_v p(\theta_v^+|v \cup \text{pa}(v)^+)p(v \cup \text{pa}(v)^+|E^*)$$

$$+ p(\theta_v^+)(1 - p(\text{pa}(v)^+|E^*)) . \tag{8}$$

We have here used that θ_v^+ is independent of $v \cup \text{pa}\,(v)^*$ for $\text{pa}(v)^* \neq \text{pa}(v)^+$, which is true since

$$p(v \cup \text{pa}(v)^*|\theta_v^+) = \sum_{\theta_v^+} p(v \cup \text{pa}(v)^*|\theta_v^+, \theta_v^*)p(\theta_v^*|\theta_v^+)$$

and neither term depends on θ_v^+.

We illustrate the above discussion of local independence with a simple example. Consider the example of L–S shown in Figure 2, and let $v = \beta$ (Bronchitis?) and so $\text{pa}(v) = \{\sigma\}$ (Smoking?). Thus, the experience node θ_β needs to specify a joint distribution over $p(b|s,\theta_\beta)$ and $p(b|\bar{s},\theta_\beta)$, denoted q_b^+ and q_b^-; i.e., to express the uncertainty concerning the frequencies with which bronchitis occurs in smokers and nonsmokers, respectively.

Assuming local independence entails q_b^+ and q_b^- being independent random quantities, and, hence, Figure 4 represents the initial independencies.

However, Figure 4 does not reveal all the conditional independencies that exist. Suppose we observe both σ and β to be, say, s and b (i.e., a smoker with bronchitis). Then the parameter revision leaves opinion concerning q_b^- unchanged and q_b^+ has a revised distribution

$$p(q_b^+|s,b) \propto p(s,b|q_b^+)p(q_b^+) \propto p(b|s,q_b^+)p(q_b^+) = q_b^+ p(q_b^+) .$$

Local independence is retained, and, hence, q_b^+ and q_b^- are conditionally inde-

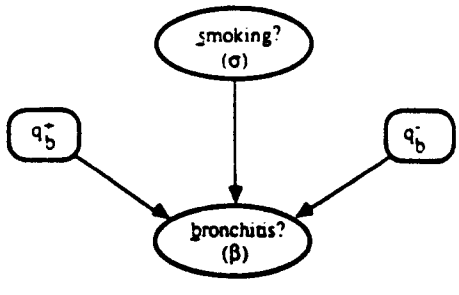

FIG. 4. Influence diagram assuming local independence of conditional probabilities relevant to β.

pendent given σ and β. This is not a direct consequence of the influence diagram, since β would normally introduce a dependency between q_b^+ and q_b^-, but in a sense this is cancelled by also observing σ. Essentially, our minimal Markov field is obtained without full moralization, since a link does not need to be introduced between q_b^+ and q_b^-. However, even more is concealed by Figure 4, since if we observe σ, one of q_b^+ or q_b^- is rendered independent of β. The answer lies in seeing that, conditional on σ, only one of q_b^+ and q_b^- is 'relevant'; essentially, observing σ revises the structure of the graph so that nonrelevant parts are pruned and, hence, made independent of the remainder. A systematic approach to the manipulation of such relevance links would be an important development.

The dependency between q_b^+ and q_b^- introduced by observing general data E^* may be examined by studying $p(q_b^+|E^*.q_b^-)$. We have

$$p(q_b^+|E^*.q_b^-) = \sum_{\sigma.\beta} p(q_b^+|\sigma.\beta.E^*.q_b^-)p(\sigma.\beta|E^*.q_b^-)$$

$$= \sum_{\beta} p(q_b^+|s.\beta)p(s.\beta|E^*) + \sum_{\beta} p(q_b^+)p(\bar{s}.\beta|E^*.q_{\bar{s}b}^-) .$$

since q_b^- is conditionally independent of E^* and q_b^- given s and β, and q_b^- is conditionally independent of E^*, q_b^-, and β given $\bar{s}$. Thus, we obtain

$$p(q_b^+|E^*.q_b^-) = \sum_{\beta} p(q_b^+|s.\beta)p(s.\beta|E^*) + p(q_b^+)p(\bar{s}|E^*.q_b^-) .$$

Comparison with (8) shows that the assumption of local independence is equivalent to ignoring the dependence on q_b^- of the current belief in the parent node. We note that if E^* includes σ, the dependence disappears.

2.4. Recollection

To summarize so far: An assumption of global independence of θ_i's allows global dissemination to be carried out locally. Local inversion procedures need to be specified, and then global retrieval can take place locally if global independence is either correct or to be assumed. Assumed local independence allows each conditional probability distribution to be individually updated. Each of these *a priori* assumptions only remains valid under certain sampling schemes, and these are explored in the next section. The fundamental local operations to be carried out at each node are thus dissemination (5), inversion (7), and retrieval (6). In Section 4 we consider a number of different parametrizations and examine the approximation procedures that may be appropriate to make these operations computationally feasible.

3. SAMPLING SCHEMES AND APPROXIMATIONS

We now consider the need for approximations under three sampling situations: complete data, a special pattern of observed data known as "ancestral sets," and arbitrary patterns of missing data.

3.1. Complete Data

In some situations it may be possible to develop a system using complete training data. If this is the case, the expression (6) degenerates to

$$p(\theta_v|E^*) = p(\theta_v|v^* \cup \text{pa}(v)^*)$$

and so no mixing is necessary in the local retrieval operation, while expression (8) also degenerates and local independence, as well as global independence, is retained. The only possible need for approximation is in storing the local inversion expression $p(\theta_v|v \cup \text{pa}(v))$ required for local retrieval (see Subsection 4.3)

3.2. Sampling from an Ancestral Set

We first define an *ancestral set* of a directed acyclic graph as a set of nodes W in which for each node $w \in W$, the parents of w are also in W. For example, in Figure 2, nodes $\{\theta_\alpha, \alpha, \theta_\tau, \tau\}$ form an ancestral set, but $\{\theta_\alpha, \tau\}$ do not. It follows that $p(W)$ is directed Markov on the influence diagram comprising the ancestral set and, hence, that $p(W)$ is Markov on the moral graph formed from this. For any three sets of nodes A, B, and C in a influence diagram, we therefore can state that A is conditionally independent of B given C, if C separates A from B in the moral graph formed from an ancestral set containing $A \cup B \cup C$; the minimal ancestral set to consider is that containing $A \cup B \cup C$ and all their ancestors [5].

Our objective is to see under what conditions $\theta_v, v \in V$ are conditionally independent given a core node set E. The minimal ancestral set containing these nodes comprises $\theta \cup E \cup \text{an}(E)$, where $\text{an}(E)$ are the ancestors of E in V. The moral graph for this segment comprises cliques $\{v \cup \text{pa}(v) \cup \theta_v\}$ for $v \in E \cup \text{an}(E)$. The nodes $\theta_v, v \in V$ are all separated by E in this moral graph if and only if $\text{an}(E) = \emptyset$, i.e., E itself forms an ancestral set of the core influence diagram. Hence, for example, marginal independence of $\{\theta_v\}$ in Figure 2 is retained if we were to observe α and τ, but not if we were to just observe τ.

Our assumption of marginal independence of θ_v is thus retained if we observe data on a case with the property that if node v is observed, then its core parents $\text{pa}(v)$ are observed, too.

We note that local independence is also preserved under sampling from ancestral sets. For a specific node v with parents $\text{pa}(v)$, two patterns are possible: We may observe either $v \cup \text{pa}(v)$ or neither v nor any descendants of v. In the first case, local independence is protected by the argument of the previous section, and in the second case, no information on $p(v|\text{pa}(v))$ is obtained.

3.3 General Patterns of Incomplete Data

After the system has been initialized and trained with, optimistically, complete data, it is still possible to carry out parameter updating with a general

configuration of missing data. A number of approximations may be necessary, as will be shown in the following sections. First, as with complete data, it may be necessary in the local inversion form (7) for $p(\theta_v | v \cup \text{pa}(v))$. Second, in the local retrieval operation, the mixture expression (6) may require approximation. Third, local dependencies may be introduced (and possibly ignored) between conditional probability distributions for a specific node v. Finally, global dependencies may arise.

We observe that global independence may, fortuitously, remain although this will not generally hold. Secondly, subsets of θ may still retain independence within themselves. For example, if in our basic example we observe a nonsmoking patient with dyspnoea that has been to Asia and has a negative chest X-ray, the experience-nodes break into three independent groups: θ_α, θ_σ, and the remaining experience-nodes . This is seen in Figure 5 where the observed set of nodes separate these groups from each other.

Also if, as is the case here, there is almost sure evidence that none of the diseases tuberculosis and lung cancer are present, it is to be expected that θ_τ, θ_ξ, θ_λ, and $(\theta_\beta, \theta_\delta)$ are very close to being posterior independent such that the only essential error in the approximation is due to ignoring the posterior dependence between θ_β and θ_δ.

A standard measure of the goodness of the global approximation to the true joint distribution is the Kullback–Leibler distance

$$\mathbf{I}(p,q) = -\mathbf{E}_p[\log(p/q)]$$

and a decomposition of this measure may be of help in monitoring the goodness of the approximation. When approximating the true posterior density $p(\theta|E^*)$

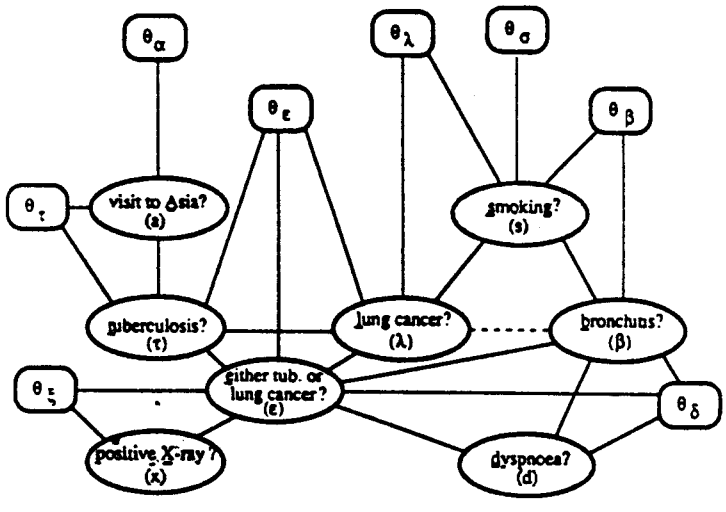

FIG. 5. Observed nodes (shaded) separate experience nodes into three groups ($\theta_\alpha, \theta_\sigma$, and the remaining) of mutually posterior independent variables.

by the product of the marginal densities $p_v = p(\theta_v|E^*)$ and these, in turn, get approximated by $\tilde{p}_v = \tilde{p}(\theta_v|E^*)$, we obtain

$$\mathbf{I}\left(p, \prod_v \tilde{p}_v\right) = \mathbf{I}\left(p, \prod_v p_v\right) + \sum_v \mathbf{I}(p_v, \tilde{p}_v). \tag{9}$$

This equation decomposes the Kullback–Leibler distance into a global and local term. More precisely, the first term on the right-hand side of (9) is the error due to assuming posterior global independence. This can easily be seen to be additive over groups of nodes that are independent as in the example. The second term is the contribution to the total distance from the errors in the local approximations, and this is additive over the set of nodes in the system.

4. SPECIFIC MODELS FOR CONDITIONAL PROBABILITIES

We next consider a number of alternative representations for a conditional probability table $p(v|\text{pa}(v), \theta_v)$. These range from discrete to strongly parametrized models, and we emphasize the means of specifying the distribution $p(\theta_v)$ and carrying out the local dissemination (4), inversion (7), and retrieval (6) operations. In particular, we consider approximations that may be necessary to allow $p(\theta_v|E^*)$ to remain in a suitable form for future use, having observed incomplete data E^* on a specific case.

4.1. Models with Discrete Parameter Space

Consider a single node v and a particular configuration $\text{pa}(v)^-$ of its coparents $\text{pa}(v)$. The conditional probability distribution $p(v|\text{pa}(v)^-, \theta_v)$, denoted q_v^-, may be considered as a random quantity taking on values in a discrete domain Q, with $p(\theta_v)$ taking the form of a fully specified distribution $p(q_v^-)$. This approach was suggested by Spiegelhalter [15] and involves full storage of a discrete distribution on q_v^-. The dissemination is trivially carried out by calculating the mean of q_v^-, while the local inversion expression $p(q_v^-|v \cup \text{pa}(v)^-)$ may be calculated and stored for each possible realization of v, thus allowing straightforward retrieval to obtain a revised discrete distribution $p(q_v^-|E^*)$. No approximation problem occurs in the local retrieval operation since the distribution of q_v^- has no assumed parametric structure that it would be desirable to retain. We note local independence has been adopted. We consider again the example of the last section.

In L–S we specified a single value $q_b^- = p(b|s, \theta_B) = .60$, but, in practice, this quantity would not be exactly known, and so a distribution over q_b^- is necessary. For illustration, we specify our prior distribution as

$$p(q_b^- = .40) = .2 \qquad p(q_b^- = .60) = .6 \qquad p(q_b^- = .80) = .2$$

reflecting a moderate confidence in our assessment (standard deviation = .13) but recognizing the proportion of smokers who present with bronchitis could

be considerably lower or higher than 60%. Before processing a case, we can first carry out the dissemination (4) to give

$$p(b|s) = \sum_{\theta_\beta} p(b|s,\theta_\beta)p(\theta_\beta) = \sum q_b^* \, p(q_b^*) = .60$$

the mean of the distribution of q_b^*, and, henceforth, the analysis of a case will proceed exactly as in L–S. Further, we may calculate in advance the inverse relation $p(\theta_\beta|\beta,\sigma)$ required for retrieval. First, we note that conditional on $\bar{s}$(not smoker), our opinion concerning q_b^* is unchanged due to local independence. We therefore only have to consider revisions conditional on s. This is best considered by calculating a joint distribution, conditional on s, of β and $q_b^* = p(b|s,\theta_\beta)$. From Table I the prior distribution on q_b^* appears as the lower margin, the predictive distribution over β, calculated in the dissemination operation, appears as the right-hand margin, and the calculations for retrieval are easily obtained. Specifically, we see that the general distribution $p(\theta_i|v \cup \mathrm{pa}(v))$ is given in this case by

$$p(q_b^*|b,s) = p(b|s,q_b^*)p(q_b^*|s)/p(b|s) = q_b^* p(q_b^*)/p(b|s),$$

while

$$p(q_b^*|\bar{b},s) = (1 - q_b^*)p(q_b^*)/p(\bar{b}|s).$$

The solutions are shown in Table II and can be seen to be simply obtained by normalizing the rows of Table I to add to 1. We note that a single observation on whether a smoker has bronchitis is surprisingly influential on our beliefs concerning the proportion of smokers who have bronchitis: if the patient has bronchitis the chance that the condition is common ($q_b^* = .80$) goes up from 20% to 27%, whereas if the patient does not have bronchitis, the chance that the condition is not so common ($q_b^* = .40$) goes up from 20% to 30%. If we do observe b, the value of $p(b|s)$ to be disseminated to the next case is $\mathbf{E}[q_b^*|b] = .628$.

In fact, suppose we observe neither whether the patient is a smoker nor has bronchitis, but only the indirect evidence E^*, described in L–S, that the patient has dyspnoea and has been to Asia. The retrieval operation (6) then consists

TABLE I. Joint distribution conditional on s (smoker), of β (bronchitis), and q_b^* (conditional probability of bronchitis given smoker).

	q_b^*			
β	0.40	0.60	0.80	$p(\beta)$
b	0.08	0.36	0.16	0.60
$\bar{b}$	0.12	0.24	0.04	0.40
$p(q_b^*)$	0.20	0.60	0.20	

TABLE II. Conditional distribution of q_b^* (proportion of smokers who have bronchitis) having observed a smoker with bronchitis (row b) or without bronchitis (row $\bar{b}$).

	q_b^*		
β	0.40	0.60	0.80
b	0.13	0.60	0.27
$\bar{b}$	0.30	0.60	0.10

of calculating

$$p(q_b^*|E^*) = \sum_{\beta,\sigma} p(q_b^*|\beta,\sigma)p(\beta,\sigma|E^*) .$$

This calculation is shown in Table III, where values for $p(q_b^*|\beta,\sigma)$ are obtained from Table II, and $p(\beta,\sigma|E^*)$ is derived from the calculation in L–S. We note that, because of local independence, conditioning on s produces no revision of opinion concerning q_b^*, and so it is not really necessary to explicitly represent rows 2 and 4 of Table III. The evidence E^* raised the probability that the patient were a smoker from .5 to .616 (L–S) and the probability that they had bronchitis from .45 to .812. This pattern supports the idea that bronchitis is very common among smokers, and, hence, by indirect observation the belief in conditional probability $q_b^* = p(b|s,\theta_B)$ is revised upwards.

For the next case, the estimated value $p(b|s)$ to be disseminated to the core will be

$$p(b|s) = \sum q_b^* \, p^*(q_b^*) = .612 ,$$

where $p^*(q_b^*)$ is the revised distribution on q_b^*. This shows a small increase over .6 due to observation of the previous case.

We make a number of observations concerning the use of discrete distributions. First, no approximation is involved in the retrieval operation and, hence,

TABLE III. Derivation of the retrieved distribution $p(q_b^*|E^*)$, having observed a patient with dyspnoea who has visited Asia, showing small increase in belief that bronchitis in smokers is common (q_b^*=0.80).

		$p(q_b^*\|\beta,\sigma)$			
β	σ	q_b^*=0.40	q_b^*=0.60	q_b^*=0.80	$p(\beta,\sigma\|E^*)$
b	s	0.13	0.60	0.27	0.54
b	$\bar{s}$	0.20	0.60	0.20	0.27
$\bar{b}$	s	0.30	0.60	0.10	0.08
$\bar{b}$	$\bar{s}$	0.20	0.60	0.20	0.11
$p(q_b^*\|E^*)$		0.17	0.60	0.23	

$p(q_v^+|E^*)$ is the correct revised distribution, although approximation still takes place in the assumption that the components of θ are both locally and globally marginally independent having observed data such as E^* that does not form an ancestral set. Second, the dissemination and retrieval operations are attractively straightforward.

However, the unparametrized means of representing $p(\theta_v)$, $v \in V$ is extremely expensive in storage, particularly if, as could be desirable, the domains of the conditional probability tables are fairly extensive. Nevertheless, the simplicity of the operations suggest that discrete distributions might be a reasonable option, if one only allows, say, three values that a probability could take on (low, medium, high), which could be changed once it was clear what direction the data were taking.

4.2. Conditional Probabilities as Dirichlet Random Variables

In multinomial sampling, the conjugate prior for the parameters is a Dirichlet distribution, which specializes to the beta distribution as a prior for the parameter of a binomial experiment. Such a parsimonious modeling of a conditional probability table is an attractive alternative to full discrete distributions, particularly for variables with more than two states, although we shall find problems in approximations when complete case-data is not obtained.

Consider a specific conditional probability distribution $p(v|\text{pa}(v)^+, \theta_v^+)$ for a fixed configuration $\text{pa}(v)^+$ and assume local independence. We shall denote this distribution over the states $x_{v1}, \ldots, x_{vk}$ by $q_v^+ = (q_{v1}^+, \ldots, q_{vk}^+)$, and assume that *a priori* it has a Dirichlet distribution $\mathcal{D}$ with parameters $\alpha^+ = (\alpha_1^+, \ldots, \alpha_k^+)$ so

$$p(q_v^+|\alpha^+) \propto \prod_i (q_{vi}^+)^{\alpha_i^+ - 1}.$$

We note that the number of parameters necessary to specify a distribution over the possible conditional probabilities is only one more (k) than required for specifying the distinct probabilities ($k - 1$), and so this is an extremely efficient means of representing $p(\theta_v)$. (Essentially, instead of a table of probabilities, a table of "counts" for each state is stored as a memory of past experience.)

The dissemination is straightforward, since by (4) the table to be used for processing the next case is

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

$$= \int q_v^+ p(q_v^+|\alpha^+)\, dq_v^+ = \alpha^+ / \sum_i \alpha_i^+ \qquad (10)$$

so that, for example, the conditional probability $p(x_{vj}|\text{pa}(v)^+)$ is taken to be $\alpha_j^+ / \Sigma_i \alpha_i^+$. The retrieval operation can, however, be more complex. From (6)

and (8), we require a specification of $p(\theta_v^* | v \cup \text{pa}(v)^*)$ for each state of v. If we observe v to be in state x_{vj} and parent configuration $\text{pa}(v)^*$, we have by standard conjugate Bayesian updating that

$$q_v^* | v \sim \mathcal{D}[\alpha_1^*, \ldots, \alpha_j^* + 1, \ldots, \alpha_k^*].$$

i.e., one is added to the relevant parameter, corresponding to an additional case in the memory. Thus, if both $\text{pa}(v)$ and v are observed, retrieval is trivial. However, if v is not observed, this mixture is over k Dirichlet distributions, each with unity added to the appropriate α term, whereas if $\text{pa}(v)$ is not definitely known to be $\text{pa}(v)^*$, an additional Dirichlet distribution is specified with the original parametrization. Thus, in general, (8) takes the form

$$p(q_v^* | E^*) = \sum_j \mathcal{D}[\alpha_1^*, \ldots, \alpha_j^* + 1, \ldots, \alpha_k^*] p(x_{vj}, \text{pa}(v)^* | E^*)$$

$$+ \mathcal{D}[\alpha_1^*, \ldots, \alpha_k^*](1 - p(\text{pa}(v)^* | E^*). \quad (11)$$

Note that if $\text{pa}(v)^*$ is observed, and neither v nor any descendants of v are observed, then (11) is formally a mixture of k Dirichlet distributions but is identical to a single Dirichlet distribution with parameters α^*, i.e., no updating has taken place. Before discussing how we may deal with this potential explosion of terms, parallel to the issues faced in unsupervised learning [4, 20], we consider our specific example examined previously. Our initial opinion concerning the proportion of smokers who had bronchitis had mean 0.6 and standard deviation 0.13, which we would also obtain if we adopted a beta distribution for q_b^* with parameters $\alpha_1^* = 9$, $\alpha_2^* = 6$; so this opinion can be roughly thought of as equivalent to having observed $\alpha_1^* + \alpha_2^* = 15$ smokers, of whom nine had bronchitis. (These parameters are most easily obtained by noting that if we were to observe n smokers with $0.6n$ having bronchitis our estimate of q_b^* would have mean 0.6 and standard deviation $(0.6 \times 0.4/n)^{1/2}$; equating the latter to 0.13 gives $n \approx 15$.)

It can be seen from Table IV that if, for example, we were to observe a smoker with bronchitis (b,s) our "experience" would change to 16 smokers, 10 of whom had bronchitis, giving a revised mean of 0.625 to be disseminated to the next case. The true revised marginal distribution on q_b^*

$$p(q_b^* | E^*) = \sum_{\beta,\sigma} p(q_b^* | \beta, \sigma) p(\beta, \sigma | E^*)$$

is, however, a mixture of three distinct beta distributions. It has mean 0.611 and standard deviation 0.1210. The mean is almost precisely that obtained by assuming the crude three point distribution as the previous subsection.

A number of possibilities exist for handling such a mixture. First, we could explicitly store a number of tables of counts and mix them when disseminating. This number will multiply with each incomplete case observed, but a maximum, say, five tables, could be stored, and above this, similar tables could be amalga-

TABLE IV. Retrieved distribution $p(q_k^{\rightarrow}|\beta,\sigma)$ were we to observe each possible configuration of β and σ together with the posterior distribution of β,σ after observing the patient with dyspnoea who had been to Asia.

Observed configuration		$q_k^{\rightarrow}*\|\beta,\sigma \sim \mathcal{B}[\alpha_1^{\rightarrow}*,\alpha_2^{\rightarrow}*]$				
β	σ	$\alpha_1^{\rightarrow}*$	$\alpha_2^{\rightarrow}*$	Mean	s.d.	$p(\beta,\sigma\|E^*)$
b	s	10	6	0.625	0.1174	0.54
$\underline{b}$	s	9	6	0.60	0.1225	0.27
$\overline{b}$	s	9	7	0.5625	0.1203	0.08
$\bar{b}$	s	9	6	0.60	0.1225	0.11

mated using one of the methods discussed below. Second, we could approximate the mixture with a single Dirichlet distribution after each observation, providing a conjugate prior for the next case. A number of combination rules have been explored in the literature. One possibility is the "fractional updating" procedure [14, 19], in which the parameters of the approximate Dirichlet distribution are the appropriate mixtures of the parameters of the individual Dirichlet distributions. Thus, in the general formulation, we assume that (11) is approximated by a Dirichlet distribution with revised parameters

$$(\alpha_i^{\rightarrow})^* = \alpha_i^{\rightarrow} + p(x_{v,i}, \mathrm{pa}(v)^{\rightarrow}|E^*), \quad i = 1, \ldots, k. \tag{12}$$

This simple updating scheme has been claimed to have a number of good properties [14]. For the specific example, the fractional updating method will approximate the mixture by a beta distribtuion with parameters (9.54, 6.08), since $p(b,s|E^*) = 0.54$ and $p(\bar{b},s|E^*) = 0.08$. This distribution has the correct mean, 0.611, but a somewhat smaller standard deviation, 0.1196.

However, a number of objections can be raised against the fractional updating procedure. Suppose, for simplicity, that the parent nodes $\mathrm{pa}(v)^{\rightarrow}$ have been observed and so the final term in (11) is absent. This situation is now equivalent to standard unsupervised learning in which the child node takes the role of the unobserved true class of the observation, the Dirichlet distributions express the uncertainty about the prior distribution over the classes, and the class-conditional feature probabilities are known and so provide the posterior probability that the observation comes from each class. Expression (12) is one method of revising the Dirichlet parameters, but Bernardo and Girón [1] have suggested two plausible desiderata that any such procedure should satisfy. We now consider these using the expressions relevant to our context. The first states that there should be no updating if no information concerning the node v has been obtained, i.e., $p(x_{v,i}|E^*) = p(x_{v,i}|\mathrm{pa}(v)^{\rightarrow})$. This is not obeyed in fractional updating since (12) will perform an update of the α's even if there is no relevant evidence. The second desideratum states that the implicit sample size $\Sigma\,\alpha_i^{\rightarrow}$ underlying the Dirichlet distribution should increase by 1 if and only if the true state of v is observed. This is also not obeyed by (12), since $\Sigma(\alpha_i^{\rightarrow})^* = \Sigma\,\alpha_i^{\rightarrow} + 1$ whatever the evidence E^*. Fractional updating therefore

seems capable of unwarranted inflation of the implicit sample size and, hence, the precision. A final criticism is that if the parent nodes pa(v) are not observed with certainty so that the final term of (11) is nonzero then the updating procedure (12) does not even necessarily provide the correct mean for the conditional probabilities.

An alternative method is to use the method of moments, i.e., to choose the parameters of the beta distribution such as to match the mean and standard deviation; see, for example, Titterington et al. [20] for a short discussion of this approach. In the example, the approximating beta distribution would have parameters (9.32, 5.94). The difference between the approximating distribution and the true mixture is almost invisible by eye on a drawing, see Table V.

The moment approach can be generalized to nonbinary response variables in a number of ways. One possibility is to equate the "average variance" of the approximating Dirichlet destribution to that of the mixture as follows. Let m_{ij} and v_{ij} denote the mean and variance of q_i^- in the jth term of (11), where the 0th term is the last. Then, the true posterior mean and variance of q_i^- are equal to

$$m_i = \sum_{j=1}^{k} m_{ij} p(x_{v,j}, \text{pa}(v)^- | E^*) + m_{i0}(1 - p(\text{pa}(v)^- | E^*))$$

and

$$v_i = \sum_{j=1}^{k} \left(v_{ij} + (m_{ij} - m_i)^2 \right) p(x_{v,j}, \text{pa}(v)^- | E^*)$$

$$+ \left(v_{i0} + (m_{i0} - m_i)^2 \right) (1 - p(\text{pa}(v)^- | E^*)) .$$

The *average variance* of the mixture distribution is

$$\bar{v} = \sum_i m_i v_i .$$

TABLE V. The density of the exact mixture distribution and the approximating beta distribution with the same mean and standard deviation in the example.

	0.20	0.30	0.40	0.50	0.55	0.60
Mixture	0.011	0.157	0.790	2.046	2.689	3.099
$\mathcal{B}$[9.32,5.94]	0.010	0.154	0.787	2.048	2.684	3.100
	0.65	0.70	0.75	0.80	0.85	0.90
MIxture	3.121	2.701	1.947	1.103	0.438	0.094
$\mathcal{B}$[9.32,5.94]	3.119	2.699	1.947	1.106	0.442	0.096

and if we approximate the mixture with a Dirichlet distribution with parameters $(sm_1, \ldots, sm_k)$ for some $s > 0$, it will have the correct mean but average variance equal to

$$\bar{v} = \sum_i m_i^2(1 - m_i)/(s + 1).$$

Equating $\bar{v}$ to $\tilde{v}$ leads to using a Dirichlet distribution with

$$s = \frac{\sum_i m_i^2(1 - m_i)}{\sum_i m_i v_i} - 1$$

as the approximating distribution.

The Dirichlet assumption therefore requires approximation in the mixture operation in local retrieval, in the assumptions of both local and global independence. However, the virtue is in its simplicity of operation, and its ready interpretation, in terms of 'counts.'

4.3. Modeling Log-odds as Gaussian Variables

The previous two subsections have discussed representations of conditional probability tables as random quantities that required specification of more parameters than did the conditional probabilities of interest. We now consider a more parsimonious assumption in which the conditional probabilities for all possible parent configurations are simultaneously modeled. Hence, 'local independence' does not hold. We hope thereby to illustrate how to exploit some more sophisticated statistical modeling.

Consider $(r + 1)$ functions $t_0, t_1, \ldots, t_r$ of the configuration at parent nodes of a node v with $q + 1$ states, numbered as $x_{v0}, \ldots, x_{vq}$. For notational convenience, t_0 is the constant function assumed to take on the value 1. Choose the state x_{v0} as a particular reference state, and let for $i = 1, \ldots, q$

$$\theta_i(x_{\mathrm{pa}(v)}) = \log[p(x_{vi}|x_{\mathrm{pa}(v)})/p(x_{v0}|x_{\mathrm{pa}(v)})].$$

i.e., the log-odds on x_{vi} against x_{v0} having observed a particular parent configuration. Our basic assumption is that $\theta_i(x_{\mathrm{pa}(v)})$ are linear combinations of the functions $t_j(x_{\mathrm{pa}(v)})$. We also create a random vector Z with $Z_i = 1$ if $X_v = x_{vi}$ and $Z_i = 0$ otherwise, for $i = 1, \ldots, q$, and, hence, we may state our assumption as

$$\theta_i(x_{\mathrm{pa}(v)}) = \theta_i(x_{\mathrm{pa}(v)}, \alpha_i) = \theta_i(t, \alpha_i) = \alpha_i' t.$$

where $\alpha_i = (\alpha_{i0}, \ldots, \alpha_{ir})$ are unknown coefficients, $t = (t_0(x_{\mathrm{pa}(v)}), \ldots, t_r(x_{\mathrm{pa}(v)}))$ are the given values of the functions and $'$ denotes transpose. We shall first consider the binary case where $q = 1$, corresponding to standard logistic regression.

4.3.1. *The Binary Case*

In this case, there is only one state apart from the reference state and we may ignore the subscript i. The prior assumption is that the coefficients α have a multivariate normal distribution with mean μ and covariance matrix Σ and, hence, $\theta(x_{\mathrm{pa}(v)},\alpha) \sim \mathcal{N}(\mu' t, t'\Sigma t)$. This assumption of prior normality for log-odds has been explored by, for example, Lindley [8] and Leonard [7].

The dissemination operation to obtain the conditional probability $p(Z = 1|x_{\mathrm{pa}(v)})$ requires the evaluation of

$$\mathbf{E}_\alpha[p(Z = 1|x_{\mathrm{pa}(v)},\alpha)] = \mathbf{E}_\theta[e^\theta/(1 + e^\theta)],$$

where $\theta \sim \mathcal{N}(\nu,\sigma^2)$; $\nu = \mu' t$, $\sigma^2 = t'\Sigma t$. In Appendix A we show that this may be well approximated by

$$\hat{p} = \hat{p}(Z = 1|x_{\mathrm{pa}(v)}) = e^{cv}/(1 + e^{cv}),$$

where $c = (1 + 0.368\sigma^2)^{-1/2}$ is a shrinking factor to what we term the direct estimate

$$\check{p} = \check{p}(Z = 1|x_{\mathrm{pa}(v)}) = e^v/(1 + e^v).$$

Hence, dissemination is a simple operation.

The retrieval operation (6), after having observed evidence E^*, is given by

$$p(\alpha|E^*) = \sum_{x_{\mathrm{pa}(v)},z} p(\alpha|x_{\mathrm{pa}(v)},z)p(x_{\mathrm{pa}(v)},z|E^*). \tag{13}$$

where the second term is obtained by the L–S procedure. The local inversion $p(\alpha|x_{\mathrm{pa}(v)},z)$ could be performed exactly in the previous parametrizations, but here we need to consider approximations. Specifically, we would like a multivariate normal approximation to the posterior density, $p(\alpha|x_{\mathrm{pa}(v)},z)$ for any specified value of $x_{\mathrm{pa}(v)}$, where

$$p(\alpha|x_{\mathrm{pa}(v)},z) \propto p(\alpha)p(z|x_{\mathrm{pa}(v)},\alpha) \propto \exp\{-(\alpha - \mu)'\Sigma^{-1}(\alpha - \mu)/2\}\frac{e^{z\alpha' t}}{(1 + e^{\alpha' t})}.$$

In Appendix B.1 we show that a Gaussian approximation to this posterior has mean $\tilde{\mu}$ and covariance matrix $\tilde{\Sigma}$ where

$$\tilde{\Sigma} = \Sigma - \left(\frac{\check{p}(1 - \check{p})}{1 + \check{p}(1 - \check{p})\sigma^2}\right)(\Sigma t)(\Sigma t)', \tag{14}$$

where $\sigma^2 = t'\Sigma t$ was the prior variance of $\theta(t,\alpha) = \alpha' t$ and $\check{p}$ was the direct estimate mentioned above, while

$$\tilde{\mu} = \mu + (z - \check{p})\tilde{\Sigma} t.$$

Hence, the mean is changed by the residual predictive error times a function that relates the parents to the coefficients μ. Thus, $\theta(x_{\mathrm{pa}(v)},\alpha)$ has its variance and mean adjusted to

$$\tilde{\sigma}^2 = \sigma^2 - \frac{\check{p}(1-\check{p})\sigma^4}{1+\check{p}(1-\check{p})\sigma^2} = \frac{\sigma^2}{(1+\check{p}(1-\check{p})\sigma^2)}$$

and

$$\tilde{\nu} = \nu + (z - \check{p})\tilde{\sigma}^2.$$

This local inversion operation is straightforward to perform. If pa(v) and v are not observed, then (13) may be approximated by a single multivariate normal distribution with the correct mean and covariance.

We illustrate the technique with a trivial example. In subsection 4.1 we considered the single conditional probability of observing bronchitis in a smoker whose prior distribution had mean 0.6 and standard deviation 0.13. This is comparable to taking $\theta \sim \mathcal{N}(\mu,\sigma^2)$, where

$$\theta = \log[p(b|s)]/p(\bar{b}|s)]$$

and $\mu = 0.427$, $\sigma^2 = 0.293$ since then the approximate mean of $p(b|s)$ is $\hat{p} = 0.6$ and the approximate standard deviation, calculated by the delta method, is $\hat{p}(1-\hat{p})\sigma = 0.13$. The direct estimate becomes $\check{p} = 0.605$. Having observed a smoker with bronchitis, the variance is modified to $\tilde{\sigma}^2 = 0.284$ and the new mean of θ is updated to $\tilde{\mu} = 0.539$, implying that the value of $p(b|s)$ to be disseminated next has increased to 0.625 (compared to 0.628 for the discrete distribution and 0.625 for the beta distribution).

4.3.2. *The Case with Multiple States*

In the case of multiple states, things get slightly more complex although we shall see that there is much analogy with the binary case. First, we organize the coefficients α_i as columns in an $(r+1)\times q$ matrix denoted α. The prior assumption is that the coefficients α have an $(r+1)\times q$-variate normal distribution with mean μ and covariance matrix $\Sigma \otimes A$, where $\otimes$ denotes Kronecker product of matrices and A is to be defined below. Hence, $\theta(x_{\mathrm{pa}(v)},\alpha) \sim \mathcal{N}_q(\mu' t, t'\Sigma t A)$ such that A, apart from a scalar depending on t, is the covariance matrix of the log-odds. To describe A, we first introduce the vector $\pi = \pi(\mu) =$ as

$$\pi_i = p(Z_i = 1|\mu, x_{\mathrm{pa}(v)}) = \exp\{\mu_i' t - \psi(\mu, t)\},$$

where

$$\psi(\mu, t) = \log\left[1 + \sum_{i=1}^{q} e^{\mu_i' t}\right].$$

and let the matrix $\Pi = \Pi(\mu)$ be a $q \times q$ diagonal matrix with the entries of π in the diagonal. Further, let E be the matrix with all entries equal to one. We then let

$$A = A(\mu) = \Pi^{-1} + \pi_0^{-1}E.$$

where $\pi_0 = \exp(-\psi(\alpha,t))$, is the probability of observing the state $x_{(0)}$. Thus, A has $\pi_i^{-1} + \pi_0^{-1}$ in the diagonal and π_0^{-1} in all off-diagonal elements. Then, the variance of $\alpha_i't$ becomes $\sigma_i^2 = \sigma^2(\pi_0^{-1} + \pi_0^{-1})$, where $s^2 = t'\Sigma t$. Note that σ^2 has a slightly different meaning here as in the previous subsection.

The dissemination operation to obtain the conditional probabilities $p(Z_i = 1|x_{\mathrm{pa}(v)})$ is made approximately in analogy with the binary case as

$$\hat{p}_i = \hat{p}(Z_i = 1|x_{\mathrm{pa}(v)}) = \pi_i(\hat{\mu}),$$

where $\hat{\mu}_i = \mu_i't(1 + 0.368\sigma_i^2)^{-1/2}$ is obtained by shrinking the direct estimate as before. Hence, dissemination creates no difficulties.

The local inversion $p(\alpha|x_{\mathrm{pa}(v)},z)$ is done by a multivariate normal approximation to the posterior density,

$$p(\alpha|x_{\mathrm{pa}(v)},z) \propto \exp\{-\mathrm{tr}[A^{-1}(\alpha - \mu)'Z^{-1}(\alpha - \mu)]/2 + t'\alpha z - \psi(\alpha,t)\}.$$

In Appendix B.2, we show that a Gaussian approximation to this posterior has mean $\tilde{\mu}$ and covariance matrix $\tilde{\Sigma} \otimes A$, where

$$\tilde{\Sigma} = \Sigma - (\Sigma t)(\Sigma t)'/(1 + \sigma^2), \tag{15}$$

while

$$\tilde{\mu} = \mu + \tilde{\Sigma}t(z - \pi)'A.$$

To make the approximation work for the next case, we need to change the A to $\tilde{A} = A(\tilde{\mu})$. This latter change gives a slightly different approximation than the one used in the binary case. Thus, the approximate local inversion operation is carried out by using $\mathcal{N}_{(r+1)\times q}(\tilde{\mu},\tilde{\Sigma} \otimes \tilde{A})$ as posterior (and prior for the next case). If $\mathrm{pa}(v)$ and v are not observed, then we again have to approximate the mixture in the retrieval by a single multivariate normal distribution.

5. FURTHER ISSUES

The preceding development provides a computationally straightforward basis for local revision of conditional probabilities as data accumulates. However, there are a number of stages at which assumptions have been made, and we can currently point to possible alternatives.

First, our assumption of exchangeability over a sequence of cases will not generally hold over a substantial period, and we would expect more recent

cases to be most important in our updating procedure. Such an adaptive system, in which past cases are gradually eliminated from the memory summarized by the experience, could be established, say, by downweighting past experience by a small amount at each updating step. Specifically, all measures of precision, which will usually be represented by a sample size, are multiplied by $(1 - \epsilon)$ when being combined with the newest observation.

Second, it is clear that our global independence assumption, while making the updating scheme attractively simple, will often be inappropriate when similar fragments of knowledge are represented in many parts of a network. Pedigree analysis is an extreme example, in which only a few unknown parameters may exist. One possibility is to do the L–S evidence propagation with algebraic expressions and, hence, provide both a final likelihood for the unknown parameters obtained from the case in hand such as indicated in our reply to the discussion of (L–S) as well as an expression for $p(C|E^*,\theta)$ for each clique C. Dissemination could then take place after propagation by combining with the prior in force before the case was analyzed, and the resulting posterior distribution calculated by combining the prior with the likelihood. Alternatives are to carry out within-case analysis for a range of parameter values that would be equivalent to a numerical version of the procedure just sketched. When more parameters are unknown, a full simulation procedure using the IP algorithm [17] may be appropriate.

Third, the robustness of the proposed schemes in the face of extreme amounts of missing data has yet to be explored. For example, in the face of totally unsupervised learning, when, say, we wish to learn the sensitivity and specificity of a diagnostic test and yet the true disease is never known, we might expect the discrete procedure to be rather unstable.

Finally, we have not addressed the crucial area of criticism of the qualitative structure of the model. If we consider a number of possible network structures denoted by $H_1, \ldots, H_k$, say, then we can monitor the predictive probability of the data obtained on each case

$$p(E_1^*, \ldots, E_n^*|H_i) = \prod_{j=1}^{n} p(E_j^*|H_i),$$

which makes the global comparison simply the product of the Bayes factors obtained while treating each case.

We note that it should be possible to decompose such a Bayes factor into components relative to cliques of the filled-in network. Hence, local model comparisons, concerned with whether a link could be dropped, or whether two adjacent cliques should be merged, could be monitored locally. Provided a full database on past cases had been retained, it would be reasonable to systematically explore the possible benefits, in terms of better prediction, that would be obtained by introducing a single additional link. However, we believe that major reconfigurations of the network should not be contrived automatically, and a domain-expert is essential for structural "learning."

APPENDIX

A. The Mean of the Logistic Transform

We wish to find an approximation to $\mathbf{E}[e^{\theta}/(1+e^{\theta})]$ when $\theta \sim \mathcal{N}(\nu,\sigma^2)$. This is achieved by twice making use of the approximation $e^{\theta}/(1+e^{\theta}) \approx \Phi(\theta\xi)$, with ξ chosen appropriately. If we, for example, choose $\xi = 0.607$, the approximation is excellent and differences are very small in absolute value, see table 2.1 of Cox [3]. Thus,

$$\mathbf{E}[e^{\theta}/(1+e^{\theta})] \approx \int_{-\infty}^{\infty} \Phi(\theta\xi)\phi(\theta;\nu,\sigma^2)\,d\theta,$$

where $\phi(\theta;\nu,\sigma^2)$ denotes a normal density function with mean ν and variance σ^2. The right-hand side can be written

$$\int_{-\infty}^{\infty} P(U<0|\theta)\phi(\theta;\nu,\sigma^2)\,d\theta,$$

where $U|\theta \sim \mathcal{N}(-\theta,\xi^{-2})$, implying that the integral can be found as $P(U<0)$. Since $\theta \sim \mathcal{N}(\nu,\sigma^2)$, we get that $U \sim \mathcal{N}(-\nu,\sigma^2+\xi^{-2})$ and therefore

$$\mathbf{E}[e^{\theta}/(1+e^{\theta})] \approx P(U<0) = \Phi(\nu/(\sigma^2+\xi^{-2})^{1/2}),$$

which, using the approximation in the reverse direction, is approximately equal to $e^{c\nu}/(1+e^{c\nu})$, where $c = (1+\xi^2\sigma^2)^{-1/2}$. For $\xi = 0.607$, this gives $c = (1+0.368\sigma^2)^{-1/2}$.

B. Local Inversion of Normal Log-odds

B.1. The Binary Case

The log-posterior density $l(\alpha)$ has, apart from an additive constant, the form

$$l(\alpha) = -(\alpha-\mu)'\Sigma^{-1}(\alpha-\mu)/2 + (\alpha' t)z - \log(1+e^{\alpha' t}).$$

Our estimate of a good multivariate normal approximation to this distribution are based on a single step in a Fisher–Newton approximation, i.e., approximating $l(\alpha)$ locally by a quadratic around μ. Hence, we find the gradient ∇ and hessian H of $l(\alpha)$, both evaluated at $\alpha = \mu$ and let $\tilde{\mu} = \mu - H^{-1}\nabla$. The new covariance matrix $\tilde{\Sigma}$ is set to $-H^{-1}$. We get

$$\nabla_i = \frac{\partial l}{\partial \alpha_i} = -w_{ii}(\alpha_i-\mu_i) - \sum_j w_{ij}(\alpha_j-\mu_j) + zt_i - t_i e^{\alpha' t}/(1+e^{\alpha' t}),$$

where $W = \Sigma^{-1}$. Hence, evaluating at $\alpha = \mu$, we find

$$\nabla = (z - \check{p})t.$$

the residual error times the observed value of the functions of the parent configuration. Further,

$$H_{ii} = \frac{\partial^2 l}{\partial \alpha_i^2} = -w_{ii} - t_i^2 e^{\alpha' t}/(1 + e^{\alpha' t})^2$$

and

$$H_{ij} = \frac{\partial^2 l}{\partial \alpha_i \partial \alpha_j} = -w_{ij} - t_i t_j e^{\alpha' t}/(1 + e^{\alpha' t})^2,$$

and, hence, H evaluated at $\alpha = \mu$ can be expressed as

$$H = -\Sigma^{-1} - \check{p}(1 - \check{p})tt'.$$

Fortunately, this means that its inverse can be simply evaluated by a standard identity, see, e.g., Rao [12], giving expression (14) for $-H^{-1} = \tilde{\Sigma}$. Hence,

$$\tilde{\mu} = \mu - H^{-1}\nabla = \mu + (z - \check{p})\tilde{\Sigma}t.$$

These recursive updating formulae are analogous to expressions in Walker and Duncan [21] and Pregibon [11].

B.2. The Case with Multiple States

In the case of multiple states, we proceed as before. The log-posterior density is, apart from a constant,

$$l(\alpha) = -\mathrm{tr}[A^{-1}(\alpha - \mu)'\Sigma^{-1}(\alpha - \mu)]/2 + t'\alpha z - \psi(\alpha, t).$$

Before differentiating, note that the function ψ has gradient and hessian equal to

$$\nabla\psi(\alpha, t) = \mathbf{E}[Zt'|\alpha] = \pi(\alpha)t'$$

$$H\psi(\alpha, t) = \mathbf{V}[Zt'|\alpha] = tt' \otimes (\Pi(\alpha) - \pi(\alpha)\pi(\alpha)').$$

Thus, we obtain by differentiation

$$\nabla l(\alpha) = -\Sigma^{-1}(\alpha - \mu)A(\mu)^{-1} + zt' - \pi(\alpha)t'$$

which, evaluated at $\alpha = \mu$, gives

$$\nabla = (z - \pi)t'.$$

Differentiating a second time yields

$$H = -\Sigma^{-1} \otimes A^{-1} - tt' \otimes (\Pi(\alpha) - \pi(\alpha)\pi(\alpha)').$$

But, we have, in fact, that

$$A^{-1} = (\Pi(\mu) - \pi(\mu)\pi(\mu)').$$

which can be seen by straightforward multiplication, such that when we evaluate at $\alpha = \mu$ we get

$$-H^{-1} = (\Sigma^{-1} + tt')^{-1} \otimes A.$$

Using the identity from before gives expression (15), and then $\tilde{\mu}$ can be calculated directly.

ACKNOWLEDGEMENTS

We are grateful to Phil Dawid for comments and discussions and to the British Science and Engineering Research Council as well as the Danish Natural Science Research Council for financial support.

References

[1] J. M. Bernardo and J. Girón, A Bayesian analysis of simple mixture models. *Bayesian Statistics*, vol. 3 (Bernardo, deGroot, Lindley, and Smith, Eds.). Clarendon Press, Oxford (1988) 67–78.
[2] G. F. Cooper, Current research directions in the development of expert systems based on belief networks. *Appl. Stochastic Models Data Anal.* **5** (1989) 39–52.
[3] D. R. Cox, *Analysis of Binary Data*, Methuen, London (1970).
[4] R. D. Duda and P. E. Hart, *Pattern Recognition and Scene Analysis*, John Wiley, New York (1973).
[5] S. L. Lauritzen, A. P. Dawid, B. N. Larsen, and H.-G. Leimer, Independence properties of directed Markov fields. *Networks* **20** (1990) 491–505.
[6] S. L. Lauritzen and D. J. Spiegelhalter, Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *J. R. Statist. Soc. Ser. B* **50** (1988) 157–224.
[7] T. Leonard, Bayesian methods for binomial data. *Biometrika* **59** (1972) 581–589.
[8] D. V. Lindley, The Bayesian analysis of contingency tables. *Ann. Math. Statist.* **35** (1961) 1622–1643.
[9] J. Pearl, Fusion, propagation and structuring in belief networks. *Artificial Intell.* **29** (1986) 241–288.
[10] J. Pearl, *Probabilistic Inference in Intelligent Systems*, Morgan Kaufmann, San Mateo, CA (1988).
[11] D. Pregibon, Logistic regression diagnostics. *Ann. Statist.* **9** (1981) 705–724.
[12] C. R. Rao, *Linear Statistical Inference and Its Applications*, 2nd. ed., John Wiley, New York.
[13] R. D. Shachter, Evaluating influence diagrams. *Operations Res.* **34** (1986) 871–882.
[14] A. F. M. Smith and U. E. Makov, A quasi-Bayes sequential procedure for mixtures. *J. R. Statist. Soc. Ser. B* **40** (1978) 106–111.

[15] D. J. Spiegelhalter. Probabilistic reasoning in predictive expert systems. *Uncertainty in Artificial Intelligence* (L. M. Kanal and J. Lemmer, Eds.), North-Holland, Amsterdam (1986) 357–370.
[16] D. J. Spiegelhalter. Fast algorithms for probabilistic reasoning in influence diagrams, with applications in genetics and expert systems. *Influence Diagrams, Belief Nets and Decision Analysis* (R. M. Oliver and J. Q. Smith, Eds.), John Wiley, Chichester (1990) pp. 361–384.
[17] M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *J. Am. Statist. Assoc.* **82** (1987) 528–540.
[18] E. A. Thompson. *Pedigree Analysis in Human Genetics*. John Hopkins University Press, Baltimore (1985).
[19] D. M. Titterington. Updating a diagnostic system using unconfirmed cases. *Appl. Stat.* **25** (1976) 238–247.
[20] D. M. Titterington, A. F. M. Smith, and U. E. Makov. *Statistical Analysis of Finite Mixture Distributions*. John Wiley, Chichester (1985).
[21] S. H. Walker and D. B. Duncan. Estimation of the probability of an event as a function of several independent variables. *Biometrika* **54** (1967) 167–179.

Received February 1989
Accepted March 1990