Propagation of Probabilities, Means, and Variances in Mixed Graphical Association Models

STEFFEN L. LAURITZEN*

A scheme is presented for modeling and local computation of exact probabilities, means, and variances for mixed qualitative and quantitative variables. The models assume that the conditional distribution of the quantitative variables, given the qualitative, is multivariate Gaussian. The computational architecture is set up by forming a tree of belief universes, and the calculations are then performed by local message passing between universes. The asymmetry between the quantitative and qualitative variables sets some additional limitations for the specification and propagation structure. Approximate methods when these are not appropriately fulfilled are sketched. It has earlier been shown how to exploit the local structure in the specification of a discrete probability model for fast and efficient computation, thereby paving the way for exploiting probability-based models as parts of realistic systems for planning and decision support. The purpose of this article is to extend this computational scheme to networks, where some vertices represent entities that are measured on a quantitative and some on a qualitative scale. An extension has the advantage of unifying several known techniques, but allows more flexible and faithful modeling and speeds computation as well. To handle this more general case, the properties of (CG) conditional Gaussian distributions are exploited. A fictitious but simple example is used for illustration throughout the paper, concerned with monitoring emissions from a waste incinerator. From optical measurements of the darkness of the smoke, the concentration of CO_2—which are both on a continuous scale—and possible knowledge about qualitative characteristics such as the type of waste burned, one wants to infer about the state of the incinerator and the current emission of heavy metals.

KEY WORDS: Bayesian methods; Causal network; CG distribution; Expert system; Recursive model; Strongly triangulated graph.

Recent developments have shown that graphical models provide a flexible framework for specification and computation in probabilistic expert systems. We abstain from a detailed survey of the literature in the area but refer to the bibliographies in Lauritzen and Spiegelhalter (1988); Jensen, Lauritzen, and Olesen (1990); and Pearl (1988), as well as the volumes of Oliver and Smith (1990) and Shafer and Pearl (1990).

For illustrative purposes we discuss a fictitious example throughout the article. This is taken from a problem connected with controlling the emission of heavy metals from a waste incinerator:

> The emissions from a waste incinerator differ because of compositional differences in incoming waste. Another important factor is the waste burning regimen, which can be monitored by measuring the concentration of CO_2 in the emissions. The filter efficiency depends on the technical state of the electrofilter and on the amount and composition of waste. The emission of heavy metals depends on both the concentration of metals in the incoming waste and the emission of dust particulates in general. The emission of dust is monitored through measuring the penetrability of light.

Here we have ignored the obvious time aspect of the monitoring problem and concentrated on a single point in time, for the sake of simplicity. The essence of the description is represented in the network of Figure 1.

The described network could in principle be used for several purposes. Typically, the emission of dust and heavy metals, the filter efficiency, and the actual concentration of heavy metals in the waste normally would not be directly available. The filter state might or might not be known, as is also the case for the type of waste.

From the measurements and knowledge available at any time, the emissions of heavy metals can be predicted—in particular, the mean and standard deviation of the predictive distribution for that emission is of interest. Diagnostic probabilities for stability of the burning regimen and/or the state of the filter could be required. Finally, the network can be used for management purposes, in that the influence of, for example filter efficiency and burning regimen on objective variables, such as the emission of heavy metals, can be computed.

The distributional theory of graphical models with both quantitative and qualitative variables is fundamental to the computational methods. A brief account of the basic elements of this theory is obtained in Section 1.

Another formal element of the computational structure is the decompositional theory of marked graphs, which are graphs with two types of vertices, here corresponding to the discrete and continuous variables. The necessary concepts are explained in Section 2.

The remaining sections describe in some detail the model specification and the elements of the computational architecture. The previously cited example is used to illustrate the various phases of the process of specification and implementation of a system based on the methods developed. We conclude with some remarks on computational complexity and approximative modifications when conditions for the exact results are not satisfied.

1. CG DISTRIBUTIONS AND POTENTIALS

The models behind the computations described in this article are based on the assumption that the conditional distribution of the continuous variables given the discrete is multivariate Gaussian. We briefly review some standard notation but refer the reader to Lauritzen and Wermuth (1989)

* Steffen L. Lauritzen is Professor of Mathematics and Statistics, Department of Mathematics and Computer Science, University of Aalborg, Fredrik Bajers Vej 7, DK-9220 Aalborg Ø, Denmark. The author is indebted to Kristian G. Olesen for computational assistance and helpful technical discussions on the way. The author is also grateful to Philip Dawid, David Spiegelhalter, a referee, and an associate editor for helpful comments concerning presentation of the results. This research has been supported in part by a grant from the SCIENCE program of the EEC.

Journal of the American Statistical Association
December 1992, Vol. 87, No. 420, Theory and Methods

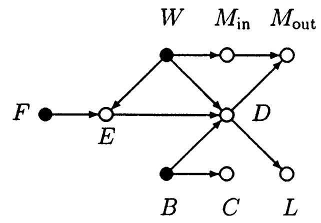

Figure 1. Graphical Representation of the Emission Problem. The variables Filter State (F), Waste Type (W), and Burning Regimen (B)—corresponding to filled circles—are conceived as qualitative variables with states {intact, defect}, {industrial, household}, and {stable, unstable}. The remaining variables are measured on a quantitative scale. These are: Metals in Waste (M_{in}), Metals Emission (M_{out}), Filter Efficiency (E), Dust Emission (D), CO_2 Concentration in Emission (C), and Light Penetrability (L).

or Whittaker (1990) for further details and derivations of formulas.

The set of variables V is partitioned as $V = \Delta \cup \Gamma$ into variables of *discrete* (Δ) and *continuous* (Γ) type. A typical element of the joint state space is denoted as in one of the following possibilities:

$$\mathbf{x} = (x_\alpha)_{\alpha \in V} = (\mathbf{i}, \mathbf{y}) = ((i_\delta)_{\delta \in \Delta}, (y_\gamma)_{\gamma \in \Gamma}),$$

where i_δ are qualitative and y_γ are real valued. The particular combination $\mathbf{i} = (i_\delta)_{\delta \in \Delta}$ is referred to as a *cell*, and the set of cells is denoted by $\mathcal{I}$. The joint distribution of the variables is supposed to have a density f with

$$f(\mathbf{x}) = f(\mathbf{i}, \mathbf{y}) = \chi(\mathbf{i})\exp\{g(\mathbf{i}) + \mathbf{h}(\mathbf{i})'\mathbf{y} - \mathbf{y}'\mathbf{K}(\mathbf{i})\mathbf{y}/2\},$$

where $\chi(\mathbf{i}) \in \{0, 1\}$ indicates whether f is positive at $\mathbf{i}$ and $\mathbf{A}'$ is the transpose of the matrix $\mathbf{A}$. We then say that X follows a *CG distribution,* which is equivalent to the statement

$$\pounds(\mathbf{X}_\Gamma \mid \mathbf{X}_\Delta = \mathbf{i}) = \mathcal{N}_{|\Gamma|}(\boldsymbol{\xi}(\mathbf{i}), \boldsymbol{\Sigma}(\mathbf{i})) \quad \text{whenever} \quad p(\mathbf{i}) = P\{\mathbf{X}_\Delta = \mathbf{i}\} > 0,$$

where $\mathbf{X}_A = (X_\alpha)_{\alpha \in A}$ and so on, and

$$\boldsymbol{\xi}(\mathbf{i}) = \mathbf{K}(\mathbf{i})^{-1}\mathbf{h}(i), \qquad \boldsymbol{\Sigma}(\mathbf{i}) = \mathbf{K}(\mathbf{i})^{-1}, \qquad (1)$$

the latter being positive definite. The triple $(\mathbf{g}, \mathbf{h}, \mathbf{K})$—defined only for $\chi(\mathbf{i}) > 0$—constitutes the *canonical characteristics* of the distribution, and $\{\mathbf{p}, \boldsymbol{\xi}, \boldsymbol{\Sigma}\}$ are the *moment characteristics.*

Note that there is a slight difference between the notation used here and in Lauritzen and Wermuth (1989), in that we allow $p(\mathbf{i})$ to be equal to 0 for some entries $\mathbf{i}$. Also, strictly speaking, χ belongs to the characteristics of the distribution, but we assume this to be implicitly represented through the domain where the other components are well defined. In the case where we have only one kind of variable, the undefined components are denoted by 0s; that is, $(\mathbf{g}, \mathbf{0}, \mathbf{0})$ or $(\mathbf{0}, \mathbf{h}, \mathbf{K})$.

A basic part of the computational task consists of updating the joint distribution in the light of evidence, corresponding to a conditioning process. A simple way of doing this is by computing with unnormalized density functions. It is also an important part of the computational process to recognize and exploit a product structure in the joint density, with the factors not necessarily being densities themselves.

For this reason we extend the notion of a CG distribution to that of a *CG potential,* which is any function ϕ of the form

$$\phi(\mathbf{x}) = \phi(\mathbf{i}, \mathbf{y}) = \chi(\mathbf{i})\exp\{g(\mathbf{i}) + \mathbf{h}(\mathbf{i})'\mathbf{y} - \mathbf{y}'\mathbf{K}(\mathbf{i})\mathbf{y}/2\},$$

where $\mathbf{K}(\mathbf{i})$ is only assumed to be a symmetric matrix. Thus ϕ is not necessarily a density. We still use the triple $(\mathbf{g}, \mathbf{h}, \mathbf{K})$ as canonical characteristics for the potential ϕ.

A basic difference is that the moment characteristics for a CG potential are only well defined when $\mathbf{K}$ is positive definite for all i with $\chi(\mathbf{i}) > 0$. Then $\boldsymbol{\Sigma}$ and $\boldsymbol{\xi}$ are given as in (1), whereas

$$p(\mathbf{i}) \propto \{\det \boldsymbol{\Sigma}(\mathbf{i})\}^{1/2}\exp\{g(\mathbf{i}) + \mathbf{h}(\mathbf{i})'\boldsymbol{\Sigma}(\mathbf{i})\mathbf{h}(\mathbf{i})/2\},$$

where $\propto$ means "proportional to." Conversely, if the moment characteristics $\{\mathbf{p}, \boldsymbol{\xi}, \boldsymbol{\Sigma}\}$ are given, then we can calculate the canonical characteristics as $\mathbf{K}(\mathbf{i}) = \boldsymbol{\Sigma}(\mathbf{i})^{-1}$, $\mathbf{h}(\mathbf{i}) = \mathbf{K}(\mathbf{i})\boldsymbol{\xi}(\mathbf{i})$, and

$$g(\mathbf{i}) = \log p(\mathbf{i}) + \{\log \det \mathbf{K}(\mathbf{i}) - |\Gamma|\log(2\pi) - \boldsymbol{\xi}(\mathbf{i})'\mathbf{K}(\mathbf{i})\boldsymbol{\xi}(\mathbf{i})\}/2.$$

2. MARKED GRAPHS AND JUNCTION TREES

In this section we give a brief exposition of the graphtheoretic notions used in the article. Many of the graphtheoretic terms have suggestive names that are really self-evident, and the reader might want to skip this section at the first reading of the article. When needing a more accurate understanding of the graphtheoretic details, the reader can return.

2.1 Notation and Terminology

First we need to establish the terminology, in particular to ensure accurate understanding of the details in future developments. A section of this type must necessarily be somewhat terse, so we ask the reader to be patient.

In this article a *network* or *graph* is formally a pair $\mathcal{G} = (V, E)$, where V is a finite set of *vertices* and the set of *edges* E is a subset of the set $V \times V$ of ordered pairs of distinct vertices. Thus our graphs have no multiple edges and no loops. Edges $(\alpha, \beta) \in E$ with both (α, β) and (β, α) in E are called *undirected,* whereas an edge (α, β) without its *opposite* (β, α) being contained in E is called *directed.*

In particular, we need to work with graphs where the vertices are *marked* in the sense that they are partitioned into two groups, Δ and Γ. We use the term *marked graph* for a graph of this type.

The vertices in set Δ are to represent qualitative variables, those in set Γ, quantitative variables. Therefore, we say that the vertices in Δ are *discrete* and those in Γ are *continuous.*

A marked graph is conveniently represented by a picture, where we use a dot for a discrete vertex and a circle for a continuous. Further, a line joining α to β represents an undirected edge, whereas an arrow from α pointing towards β is used for a directed edge (α, β).

If the graph has only undirected edges (lines), it is an *undirected* graph. If all edges are directed (arrows), the graph is said to be *directed.*

A subset C is *complete* if all vertices in C are joined by an arrow or a line. A complete subset that is maximal in the sense that no other vertex in the graph is connected to all its elements is called a *clique*. The cliques of the graph can be considered to be the fundamental blocks of the structure that the graph describes.

If there is an arrow from α pointing towards β, then α is said to be a *parent* of β and β a *child* of α. The set of parents of β is denoted as $\mathrm{pa}(\beta)$. If there is a line or an arrow between α and β, then α and β are said to be *neighbors*. A selection of graphtheoretic concepts are illustrated in Figure 2.

A *path* of length n from α to β is a sequence $\alpha = \alpha_0, \ldots, \alpha_n = \beta$ of distinct vertices such that $(\alpha_{i-1}, \alpha_i) \in E$ for all $i = 1, \ldots, n$. An undirected graph is a *tree* if there is a unique path between any two vertices.

A *n cycle* is a path of length n with the modification that $\alpha = \beta$; that is, it begins and ends in the same point. A graph is *acyclic* if it has no cycles. In particular, a *directed acyclic graph* will be of interest in that it is the basic structure used in the model specification.

For a directed graph $\mathcal{G}$, we define its *moral graph* $\mathcal{G}^m$ as the undirected graph with the same vertex set as $\mathcal{G}$ but with α and β adjacent in $\mathcal{G}^m$ if and only if either α is a parent of β or, conversely, or if α and β have a common child γ. The moral graph plays the same role in this article as it did in Lauritzen and Spiegelhalter (1988), in that it identifies groups of variables that enter simultaneously into the factors of the expression for the joint density. Figure 3 displays the moral graph of the network corresponding to the basic example studied.

2.2 Decomposition of Marked Graphs

The basic trick enabling the computational task to be performed locally is the decomposition of a suitably modified network into partly independent components formed by the cliques of that graph. The inherent asymmetry between discrete and continuous variables in the CG distributions implies that one also needs to take into proper account the behavior of the markings of the graph. We refer the interested reader to Leimer (1989) for a detailed graphtheoretic study of the problems as well as all proofs. Here we introduce the notion of a decomposition by stating the following formal definition.

Definition 1. A triple (A, B, C) of disjoint subsets of the vertex set V of an undirected, marked graph $\mathcal{G}$ is said to form a (strong) *decomposition* of $\mathcal{G}$ if $V = A \cup B \cup C$ and the following three conditions all hold:

1. C separates A from B.
2. C is a complete subset of V.
3. $C \subseteq \Delta \vee B \subseteq \Gamma$.

When this is the case, we say that (A, B, C) *decomposes* $\mathcal{G}$ into the components $\mathcal{G}_{A \cup C}$ and $\mathcal{G}_{B \cup C}$.

If only conditions 1 and 2 hold, we say that (A, B, C) form a *weak decomposition*. Thus weak decompositions ignore the markings of the graph.

In the pure cases condition 3 holds automatically and all weak decompositions are also decompositions. Note that what we have chosen to call a decomposition (without a qualifier) is what Leimer (1989) called a strong decomposition. Figure 4 illustrates the notions of (strong) and weak decompositions.

A decomposable graph is one that can be successively decomposed into its cliques. Again we choose to state this formally through the following recursive definition.

Definition 2. An undirected, marked graph is said to be *decomposable* if it is complete or if there exists a decomposition (A, B, C), with A and B both nonempty, into decomposable subgraphs $\mathcal{G}_{A \cup C}$ and $\mathcal{G}_{B \cup C}$.

Note that the definition makes sense, because both subgraphs $\mathcal{G}_{A \cup C}$ and $\mathcal{G}_{B \cup C}$ must have fewer vertices than the original graph $\mathcal{G}$.

Decomposable, unmarked graphs are characterized by being triangulated; that is, the graphs do not have chordless

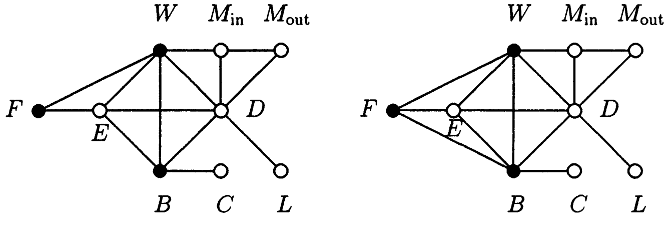

Figure 3. Modified Graphs for the Waste Example. The graph obtained after marrying of parents and dropping directions is shown in (a). When further a link between B and F is added, the strongly decomposable graph in (b) is obtained.

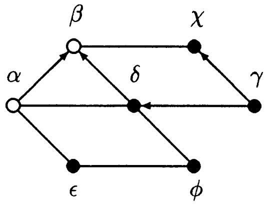

Figure 2. Illustration of Graph Theoretic Concepts. The vertices α and β are continuous, and the remaining are discrete. We have $\mathrm{pa}(\chi) = \{\gamma\}$. The vertices ε and ϕ are neighbors. The set $\{\alpha, \beta, \delta\}$ is a clique. A cycle is formed by $(\alpha, \delta, \phi, \varepsilon)$.

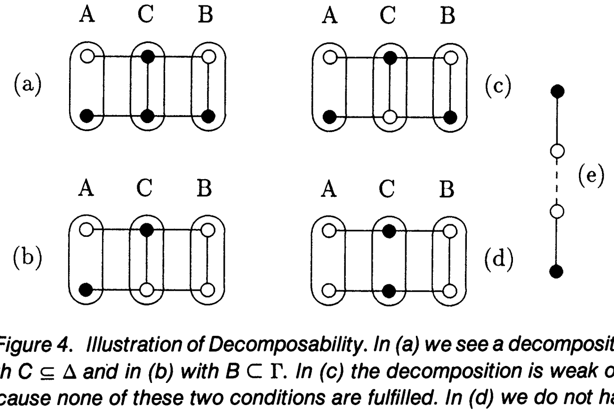

Figure 4. Illustration of Decomposability. In (a) we see a decomposition with $C \subseteq \Delta$ and in (b) with $B \subset \Gamma$. In (c) the decomposition is weak only because none of these two conditions are fulfilled. In (d) we do not have a decomposition because the separator C is not complete. The graph (e) displays a path that is forbidden in a decomposable graph.

cycles of length three or more. Decomposable marked graphs are characterized further by not having any path of a particular type.

Proposition 1. An undirected, marked graph is decomposable if and only if it is triangulated and does not contain any path ($\delta_1 = \alpha_0, \ldots, \alpha_n = \delta_2$) between two discrete vertices passing through only continuous vertices, with the discrete vertices not being neighbors.

We have illustrated the typical forbidden path in Figure 4.

2.3 Junction Trees With Strong Roots

The construction of the computational structure in Lauritzen and Spiegelhalter (1988) and Jensen et al. (1990) began with a directed acyclic graph, formed its moral graph, added links to make it triangulated, and formed a junction tree of the cliques of the triangulated graph. This procedure is to be generalized and modified.

The first part of the manipulations is unchanged in that we begin with a directed acyclic graph. We then form the moral graph by adding undirected edges between parents that are not already linked and drop directions to obtain an undirected graph. Finally, we add further links such that we obtain a decomposable, marked graph. In our example we have made this modification in Figure 3. Note that the link between B and F is necessary to remove the forbidden path (B, E, F) and make the graph decomposable, whereas it is (weakly) triangulated even without this. In this particular example the discrete variables end up forming a complete subset, but this is not always the case.

The next step is constructing the *junction tree*, an organization of a collection of subsets of the set of variables V into a tree that satisfies the condition that $A \cap B$ is a subset of all sets on the path in the tree between A and B. When a collection of subsets is organized in a junction tree, one can show that it must be a set of complete subsets of a triangulated graph containing the cliques. For any two sets C and D that are neighbors in the junction tree, their intersection $S = C \cap D$ is called their *separator* because it separates $C \backslash D$ from $D \backslash C$ in the graph. When we make a picture of the junction tree, the separators are drawn as rectangles (see Fig. 5).

The junction tree is the basic computational structure, but the asymmetry between continuous and discrete variables make a further condition necessary for the propagation scheme to work properly. Again we present a formal definition.

Definition 3. A subset R on a junction tree is a *strong root* if any pair A, B of neighbors on the tree with A closer to R than B satisfies

$$(B \backslash A) \subseteq \Gamma \vee (B \cap A) \subseteq \Delta. \tag{2}$$

The condition (2) is equivalent to the triple ($A \backslash B$, $B \backslash A$, $A \cap B$) forming a strong decomposition of $\mathcal{G}_{A \cup B}$. In words, it expresses that when a separator between two neighboring cliques is not purely discrete, the clique furthest away from the root has only continuous vertices beyond the separator.

Statement (iii)′ of Theorem 2′ of Leimer (1989) ensures that *the cliques of a decomposable marked graph can be organized in a junction tree with at least one strong root.* We assume henceforth that this has been done.

Figure 5 displays a junction tree for our example. The clique $\{W, E, B, F\}$ could be used as a strong root. For example, $\{W, M_{in}, D\}$ has only the continuous variable M_{in} beyond the separator $\{W, D\}$.

3. MODEL SPECIFICATION

As in the discrete case, the qualitative part of the model is initially specified by a directed acyclic graph, such as the one in Figure 1. The graph specifies the basic dependencies among the variables by assuming that the joint distribution of these has the directed Markov property with respect to the graph. In other words, we assume that the density is equal to the product of the conditional densities of the variables attached to each vertex in the graph, given the states at their parent vertices (Kiiveri, Speed, and Carlin 1984; Lauritzen, Dawid, Larsen and Leimer 1990).

To exploit the properties of CG potentials, we need to further assume that the graph satisfies the constraint that *no continuous vertices have discrete children.* If this assumption is not fulfilled, we have to use approximate methods in the specification phase (see Section 6). In our example the assumption is clearly satisfied.

Next we specify, for each discrete variable A, the conditional distribution at A given the states at its parent vertices (which are then all discrete). In the case where A is continuous, we assume the conditional distribution of the response variable Y associated with A to be of the type

$$\pounds(Y \mid \mathrm{pa}(A)) = \mathcal{N}(\alpha(\mathbf{i}) + \beta(\mathbf{i})'\mathbf{z}, \gamma(\mathbf{i})).$$

Here pa(A) is a short notation for the combination of discrete and continuous states ($\mathbf{i}$, $\mathbf{z}$) of the variables that are parents of A. In this formula $\gamma(\mathbf{i}) > 0$, $\alpha(\mathbf{i})$ is a real number, and $\beta(\mathbf{i})$ is a vector of the same dimension as the continuous part $\mathbf{z}$ of the parent variables.

Note that we assume that the mean depends linearly on continuous parent variables and that the variance does not depend on the continuous part of the parent variables. The linear function, like the variance, is allowed to depend on the discrete part of the parent variables.

The conditional density then corresponds to a CG potential ϕ_A with defined on the combination ($\mathbf{i}$, $\mathbf{z}$, y) of parent variables ($\mathbf{i}$, $\mathbf{z}$) and response variable y with canonical char-

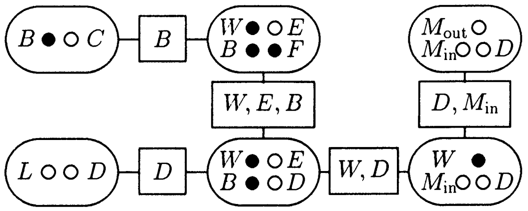

Figure 5. A Junction Tree of the Waste Example. The separators are drawn as rectangular boxes. Possible strong roots are {B, C} and {W, E, B, F}.

acteristics $(\mathbf{g}_A, \mathbf{h}_A, \mathbf{K}_A)$, where

$$g_A(\mathbf{i}) = -\frac{\alpha(\mathbf{i})^2}{2\gamma(\mathbf{i})} - [\log\{2\pi\gamma(\mathbf{i})\}]/2, \tag{3}$$

$$\mathbf{h}_A(\mathbf{i}) = \frac{\alpha(\mathbf{i})}{\gamma(\mathbf{i})}\begin{pmatrix} 1 \\ -\beta(\mathbf{i}) \end{pmatrix}, \tag{4}$$

and

$$\mathbf{K}_A(i) = \frac{1}{\gamma(\mathbf{i})}\begin{pmatrix} 1 & -\beta(\mathbf{i})' \\ -\beta(\mathbf{i}) & \beta(\mathbf{i})\beta(\mathbf{i})' \end{pmatrix}. \tag{5}$$

This follows from direct calculation using the expression for the normal density. We simply write

$$\phi(\mathbf{i}, \mathbf{z}, y) = \{2\pi\gamma(\mathbf{i})\}^{-1/2}\exp[-\{y - \alpha(\mathbf{i}) - \beta(\mathbf{i})'\mathbf{z}\}^2/\{2\gamma(\mathbf{i})\}],$$

resolve the parentheses, take logarithms, and identify terms. Note that $\mathbf{K}_A(\mathbf{i})$ has rank one and is, therefore, typically not positive definite.

In our basic example we specify the conditional distributions as follows.

1. *Burning regimen.* This variable is discrete and denoted by B. We let

$$P(B = \textit{stable}) = .85 = 1 - P(B = \textit{unstable}).$$

2. *Filter state.* This is discrete and denoted by F. We let

$$P(F = \textit{intact}) = .95 = 1 - P(F = \textit{defect}).$$

3. *Waste type.* This is discrete and denoted by W. We let

$$P(W = \textit{industrial}) = 2/7 = 1 - P(W = \textit{household}).$$

4. *Filter efficiency.* This is represented on a logarithmic scale and denoted by E. We assume the relation $\text{dust}_{\text{out}} = \text{dust}_{\text{in}} \times \rho$ and get in the logarithmic scale that $\log \text{dust}_{\text{out}} = \log \text{dust}_{\text{in}} + \log \rho$. We let $E = \log \rho$ and admit that filter inefficiency might be a better word for the variable E. We then specify

$$\mathcal{L}(E \mid \textit{intact, household}) = \mathcal{N}(-3.2, .00002),$$

$$\mathcal{L}(E \mid \textit{defect, household}) = \mathcal{N}(-.5, .0001),$$

$$\mathcal{L}(E \mid \textit{intact, industrial}) = \mathcal{N}(-3.9, .00002),$$

and

$$\mathcal{L}(E \mid \textit{defect, industrial}) = \mathcal{N}(-.4, .0001).$$

This corresponds to filter efficiencies $1 - \rho$ on about 96%, 39%, 98%, and 33%. For example, when the filter is defect and household waste is burnt, the filter removes a fraction of $1 - \exp(-.5) = .39$ of the dust.

5. *Dust emission.* This is represented on a logarithmic scale as a variable D. We let

$$\mathcal{L}(D \mid \textit{stable, industrial}, e) = \mathcal{N}(6.5 + e, .03),$$

$$\mathcal{L}(D \mid \textit{stable, household}, e) = \mathcal{N}(6.0 + e, .04),$$

$$\mathcal{L}(D \mid \textit{unstable, industrial}, e) = \mathcal{N}(7.5 + e, .1),$$

and

$$\mathcal{L}(D \mid \textit{unstable, household}, e) = \mathcal{N}(7.0 + e, .1).$$

Thus on a day when household waste is burned under a stable regimen and the filter works perfectly, the typical concentration will be $\exp(6.0 - 3.2) = 16.4$ mg/Nm^3. Similarly, if the filter is defective on a day with industrial waste and the burning regimen is unstable, we typically will see an output concentration of dust on $\exp(7.5 - .4) = 1{,}212$ mg/Nm^3.

6. CO_2 *concentration.* This is represented on a logarithmic scale as a variable C. We let

$$\mathcal{L}(C \mid \textit{stable}) = \mathcal{N}(-2, .1), \quad \mathcal{L}(C \mid \textit{unstable}) = \mathcal{N}(-1, .3).$$

Thus the concentration of CO_2 is typically around 14% under a stable regimen and 37% when the burning process is unstable.

7. *Light penetrability.* This is represented on a logarithmic scale as a variable L. We let

$$\mathcal{L}(L \mid D = d) = \mathcal{N}(3 - d/2, .25).$$

This corresponds to the penetrability being roughly inversely proportional to the square root of the dust concentration.

8. *Metal in waste.* The concentration of heavy metals in the waste is represented as a continuous variable M_i on a logarithmic scale. We let

$$\mathcal{L}(M_{in} \mid \textit{industrial}) = \mathcal{N}(.5, .01),$$

$$\mathcal{L}(M_{in} \mid \textit{household}) = \mathcal{N}(-.5, .005).$$

The precise interpretation is unit dependent, but the main point is that industrial waste tends to contain heavy metals in concentrations about three times as high as in household waste. Also the variability of the metals concentrations is higher in industrial waste.

9. *Metal emission.* This is a continuous variable M_{out} on a logarithmic scale. We let

$$\mathcal{L}(M_{out} \mid d, m_{in}) = \mathcal{N}(d + m_{in}, .002).$$

Thus we simply assume that the concentration of emitted metals is about the same in the dust emitted as in the original waste.

The numbers have been constructed with a view to information from Hansen and Dalager (1985), but are otherwise purely fictitious.

4. BASIC OPERATIONS ON CG POTENTIALS

The basis for the computational scheme consists partly of a set of fundamental operations on the CG potentials, partly on a message-passing scheme in the junction tree. The latter is described in Section 5. Here we describe the elements of the local computations.

Recall from Section 1 that a CG potential is any function ϕ of the form

$$\phi(\mathbf{x}) = \phi(\mathbf{i}, \mathbf{y}) = \chi(\mathbf{i})\exp\{g(\mathbf{i}) + \mathbf{h}(\mathbf{i})'\mathbf{y} - \mathbf{y}'\mathbf{K}(\mathbf{i})\mathbf{y}/2\},$$

where $\mathbf{K}(\mathbf{i})$ is a symmetric matrix. The triple $(\mathbf{g}, \mathbf{h}, \mathbf{K})$ represents the canonical characteristics for the potential ϕ.

4.1 Extension

If $(\mathbf{g}, \mathbf{h}, \mathbf{K})$ are the characteristics of a CG potential ϕ defined on the variables $(\mathbf{i}, \mathbf{y})$, then we need sometimes to operate on this as if it was defined on a larger set $(\mathbf{i}, \mathbf{j}, \mathbf{y}, \mathbf{z})$ of variables. This is formally done by extending it to $\bar{\phi}$, letting $\bar{\phi}(\mathbf{i}, \mathbf{j}, \mathbf{y}, \mathbf{z}) = \phi(\mathbf{i}, \mathbf{y})$. Clearly, the corresponding characteristics are

$$\bar{g}(\mathbf{i}, \mathbf{j}) = g(\mathbf{i}), \qquad \bar{\mathbf{h}}(\mathbf{i}, \mathbf{j}) = \begin{pmatrix} \mathbf{h}(\mathbf{i}) \\ 0 \end{pmatrix},$$

$$\bar{\mathbf{K}}(\mathbf{i}, \mathbf{j}) = \begin{pmatrix} \mathbf{K}(\mathbf{i}) & 0 \\ 0 & 0 \end{pmatrix}.$$

Hence the extension essentially amounts to adjoining 0s to the characteristics such as to give them the desired dimensions.

4.2 Multiplication and Division

Multiplication of two functions is defined the natural way, after the functions have been extended to be defined on the same space of variables. Expressed in terms of the canonical characteristics, multiplication becomes simple addition:

$$(\mathbf{g}_1, \mathbf{h}_1, \mathbf{K}_1) * (\mathbf{g}_2, \mathbf{h}_2, \mathbf{K}_2) = (\mathbf{g}_1 + \mathbf{g}_2, \mathbf{h}_1 + \mathbf{h}_2, \mathbf{K}_1 + \mathbf{K}_2).$$

Division likewise is defined in the obvious way, but special care must be taken when dividing by 0. Thus for $\mathbf{x} = (\mathbf{i}, \mathbf{y})$ we let

$$\begin{aligned} (\phi/\psi)(\mathbf{x}) &= 0 && \text{if } \phi(\mathbf{x}) = 0 \\ &= (\phi(\mathbf{x})/\psi(\mathbf{x})) && \text{if } \psi(\mathbf{x}) \neq 0 \\ &= \text{undefined} && \text{otherwise.} \end{aligned}$$

4.3 Marginalization

An essential difference between the pure discrete case and the situation described in this article is due to fact that adding two CG potentials generally will result in a function of a different structure. Hence there will be some complications.

We distinguish several cases. First, we discuss *marginals over continuous variables.* In this case we simply integrate. Let

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix}, \qquad \mathbf{h} = \begin{pmatrix} \mathbf{h}_1 \\ \mathbf{h}_2 \end{pmatrix}, \qquad \mathbf{K} = \begin{pmatrix} \mathbf{K}_{11} & \mathbf{K}_{12} \\ \mathbf{K}_{21} & \mathbf{K}_{22} \end{pmatrix},$$

with $\mathbf{y}_1$ having dimension p and $\mathbf{y}_2$ having dimension q. We then have the following.

Lemma 1. The integral $\int \phi(\mathbf{i}, \mathbf{y}_1, \mathbf{y}_2)\, d\mathbf{y}_1$ is finite if and only if $\mathbf{K}_{11}$ is positive definite. It is then equal to a CG potential ϕ with canonical characteristics given as

$$\begin{aligned} \tilde{g}(\mathbf{i}) &= g(\mathbf{i}) \\ &\quad + \{p \log(2\pi) - \log\det \mathbf{K}_{11}(\mathbf{i}) + \mathbf{h}_1(\mathbf{i})'\mathbf{K}_{11}(\mathbf{i})^{-1}\mathbf{h}_1(\mathbf{i})\}/2, \end{aligned}$$

$$\tilde{\mathbf{h}}(\mathbf{i}) = \mathbf{h}_2(\mathbf{i}) - \mathbf{K}_{21}(\mathbf{i})\mathbf{K}_{11}(i)^{-1}\mathbf{h}_1(\mathbf{i}),$$

and

$$\tilde{\mathbf{K}}(\mathbf{i}) = \mathbf{K}_{22}(\mathbf{i}) - \mathbf{K}_{21}(\mathbf{i})\mathbf{K}_{11}(\mathbf{i})^{-1}\mathbf{K}_{12}(\mathbf{i}).$$

Proof. Let

$$\mu(\mathbf{i}) = -\mathbf{K}_{11}(\mathbf{i})^{-1}\mathbf{K}_{12}(\mathbf{i})\mathbf{y}_2 + \mathbf{K}_{11}(\mathbf{i})^{-1}\mathbf{h}_1(\mathbf{i}).$$

Then we find by direct calculation that

$$\begin{aligned} \phi(\mathbf{i}, \mathbf{y}) &= \exp\{-(\mathbf{y}_1 - \mu(\mathbf{i}))'\mathbf{K}_{11}(\mathbf{i})(\mathbf{y}_1 - \mu(\mathbf{i}))/2\} \\ &\quad \times \exp\{\mathbf{y}_2'(\mathbf{h}_2(\mathbf{i}) - \mathbf{K}_{21}(\mathbf{i})\mathbf{K}_{11}(\mathbf{i})^{-1}\mathbf{h}_1(\mathbf{i}))\} \\ &\quad \times \exp\{-\mathbf{y}_2'(\mathbf{K}_{22}(\mathbf{i}) - \mathbf{K}_{21}(\mathbf{i})\mathbf{K}_{11}(\mathbf{i})^{-1}\mathbf{K}_{12}(\mathbf{i}))\mathbf{y}_2/2\} \\ &\quad \times \exp\{g(\mathbf{i}) + \mathbf{h}_1(\mathbf{i})'\mathbf{K}_{11}(\mathbf{i})^{-1}\mathbf{h}_1(\mathbf{i})/2\}. \end{aligned}$$

Now $\mathbf{y}_1$ appears only in the first factor. This can be integrated by letting $\mathbf{z} = \mathbf{y}_1 - \mu(\mathbf{i})$ and recalling that if $\mathbf{z} \in \mathcal{R}^p$ and $\mathbf{K}$ is positive definite, then

$$\int e^{-\mathbf{z}'\mathbf{K}\mathbf{z}/2}\, d\mathbf{z} = (2\pi)^{p/2}(\det \mathbf{K})^{-1/2}.$$

The result follows.

When calculating *marginals over discrete variables,* we distinguish two cases. First, if $\mathbf{h}$ and $\mathbf{K}$ do not depend on $\mathbf{j}$—that is, $\mathbf{h}(\mathbf{i}, \mathbf{j}) \equiv \mathbf{h}(\mathbf{i})$ and $\mathbf{K}(\mathbf{i}, \mathbf{j}) \equiv \mathbf{K}(\mathbf{i})$—we define the marginal $\tilde{\phi}$ of ϕ over $\mathbf{j}$ the direct way:

$$\begin{aligned} \tilde{\phi}(\mathbf{i}, \mathbf{y}) &= \sum_{\mathbf{j}} \phi(\mathbf{i}, \mathbf{j}, \mathbf{y}) \\ &= \sum_{\mathbf{j}} \chi(\mathbf{i}, \mathbf{j})\exp\{g(\mathbf{i}, \mathbf{j}) + \mathbf{h}(\mathbf{i})'\mathbf{y} - \mathbf{y}'\mathbf{K}(\mathbf{i})\mathbf{y}/2\} \\ &= \exp\{\mathbf{h}(\mathbf{i})'\mathbf{y} - \mathbf{y}'\mathbf{K}(\mathbf{i})\mathbf{y}/2\} \sum_{j} \chi(\mathbf{i}, \mathbf{j})\exp g(\mathbf{i}, \mathbf{j}), \end{aligned}$$

which leads to the following canonical characteristics for the marginal

$$\tilde{g}(\mathbf{i}) = \log \sum_{\mathbf{j}:\chi(\mathbf{i},\mathbf{j})=1} \exp g(\mathbf{i}, \mathbf{j}),$$

$$\tilde{\mathbf{h}}(\mathbf{i}) = \mathbf{h}(\mathbf{i}), \qquad \tilde{\mathbf{K}}(\mathbf{i}) = \mathbf{K}(\mathbf{i}).$$

Second, if either $\mathbf{h}$ or $\mathbf{K}$ depends on $\mathbf{j}$, then the marginalization process is more subtle, because simple addition of CG potentials will not result in a CG potential. The procedure we shall then use is only well defined for $\mathbf{K}(\mathbf{i}, \mathbf{j})$ positive definite and is best described in terms of the moment characteristics $\{\mathbf{p}, \xi, \Sigma\}$. The marginal $\tilde{\phi}$ is defined as the potential with moment characteristics $\{\tilde{\mathbf{p}}, \tilde{\xi}, \tilde{\Sigma}\}$, where

$$\tilde{p}(\mathbf{i}) = \sum_{\mathbf{j}} p(\mathbf{i}, \mathbf{j}), \qquad \tilde{\xi}(\mathbf{i}) = \sum_{\mathbf{j}} \xi(\mathbf{i}, \mathbf{j})p(\mathbf{i}, \mathbf{j})/\tilde{p}(\mathbf{i})$$

and

$$\begin{aligned} \tilde{\Sigma}(\mathbf{i}) &= \sum_{\mathbf{j}} \Sigma(\mathbf{i}, \mathbf{j})p(\mathbf{i}, \mathbf{j})/\tilde{p}(\mathbf{i}) \\ &\quad + \sum_{\mathbf{j}} (\xi(\mathbf{i}, \mathbf{j}) - \tilde{\xi}(\mathbf{i}))'(\xi(\mathbf{i}, \mathbf{j}) - \tilde{\xi}(\mathbf{i}))p(\mathbf{i}, \mathbf{j})/\tilde{p}(\mathbf{i}). \end{aligned}$$

The "marginalized" density will then have the correct moments; that is,

$$P(\mathbf{I} = \mathbf{i}) = \tilde{p}(\mathbf{i}), \qquad E(\mathbf{Y} \mid \mathbf{I} = \mathbf{i}) = \tilde{\xi}(\mathbf{i}),$$

$$V(\mathbf{Y} \mid \mathbf{I} = \mathbf{i}) = \tilde{\Sigma}(\mathbf{i}),$$

where expectations are taken with respect to the CG distribution determined by ϕ. This is a direct consequence of the familiar relations

$$E(\mathbf{Y} \mid \mathbf{I} = \mathbf{i}) = E\{E(\mathbf{Y} \mid (\mathbf{I}, \mathbf{J})) \mid \mathbf{I} = \mathbf{i}\} \qquad (6)$$

and

$$V(\mathbf{Y}\,|\,\mathbf{I}=\mathbf{i}) = E\{V(\mathbf{Y}\,|\,(\mathbf{I},\mathbf{J}))\,|\,\mathbf{I}=\mathbf{i}\} + V\{E(\mathbf{Y}\,|\,(\mathbf{I},\mathbf{J}))\,|\,\mathbf{I}=\mathbf{i}\}. \quad (7)$$

When *marginalizing over both continuous and discrete variables,* we first marginalize over the continuous variables and then over the discrete. If in the second of these stages we have $(\mathbf{h}, \mathbf{K})$ independent of j, we say that we have a *strong marginalization.* In the other case we must use the marginalization process just described, and we speak of a *weak marginalization.* In both cases we use the symbol $\sum_{W\setminus V}\phi_W$ for the marginalized potential, where V denotes the set of variables marginalized to and $W\setminus V$ denotes the set of variables marginalized over.

We leave to the reader to verify that the weak marginalization satisfies the standard composition rule such that when $U \subset V \subset W$,

$$\sum_{V\setminus U}\left(\sum_{W\setminus V}\phi_W\right) = \sum_{W\setminus U}\phi_W. \quad (8)$$

However, only the strong marginalizations behave well when products are involved. In general we have

$$\sum_{W\setminus V}(\phi_W\phi_V) \neq \phi_V\left(\sum_{W\setminus V}\phi_W\right). \quad (9)$$

A consequence is that the axioms of Shenoy and Shafer (1990) are not fulfilled. Hence we must establish correctness of our propagation scheme directly without exploiting their general computational theory.

In the special case of *strong* marginalizations, equality holds in (9). This follows by elementary calculations, because strong marginalizations are just ordinary integrations.

5. OPERATING IN THE JUNCTION TREE

When the model has been specified, the handling of incoming evidence and calculation of specific probabilities is done in the junction tree representation using the elementary operations described in the previous section. Essentially, the junction tree representation of the cliques in the strongly triangulated, moralized graph captures the computationally interesting aspects of the product structure in the joint density of the variables involved. Then the computations can be performed locally within the cliques and between the cliques that are neighbors in the junction tree. Hence we assume that a junction tree with strong root has been established on the basis of the original graphs, such as discussed in Section 2.

Each subset of variables in the junction tree is referred to as a *belief universe.* The set of all variables is termed the *total universe.* The collection of belief universes in the junction tree is denoted by $\mathcal{C}$, to indicate that it is the set of cliques in a strongly decomposable graph.

Recall from Section 2 that the intersections of neighbors in the junction tree are called separators. The collection of separators is denoted by $\mathcal{S}$, where this may involve multiple copies of the same separator set. Both the belief universes and the separators can have belief potentials ϕ_W attached to them, and these are all assumed to be CG potentials defined on the corresponding spaces of variables. The *joint system belief* ϕ_U associated with the given attachment of potentials is defined as

$$\phi_U = \frac{\prod_{V\in\mathcal{C}}\phi_V}{\prod_{S\in\mathcal{S}}\phi_S} \quad (10)$$

and is assumed to be proportional to the joint density of all the variables. Because all potentials involved are CG potentials, the joint density will be a CG density itself.

We always assume that the tree is *supportive,* meaning that for any universe V with neighboring separator S we have $\phi_S(\mathbf{x}) = 0 \Rightarrow \phi_V(\mathbf{x}) = 0$. This enables us to deal correctly with cases in which some states are ruled out as impossible by having potentials equal to 0.

5.1 Initializing the Junction Tree

As a first step, the junction tree with strong root must be initialized according to the model specification, to make sure that the tree is supportive and the joint system belief given by (10) is the joint density specified by the model, as in Section 3. This is done as follows.

First, we assign each vertex A in the original graph to a universe V in the tree. This must be done in such a way that $(A \cup \mathrm{pa}(A)) \subseteq V$ but is otherwise arbitrary. This ensures that the universe is so large that the CG potential ϕ_A obtained from the conditional density of A given $\mathrm{pa}(A)$ can be extended to V.

Second, for each universe V we let ϕ_V be the product of all the (extensions of) potentials ϕ_A for vertices assigned to it. On the separators we let $\phi_S \equiv 1$; that is, the potential with canonical characteristic $(\mathbf{0}, \mathbf{0}, \mathbf{0})$. This is also the potential on universes with no vertices assigned to them.

In our basic example are several possibilities for initializing the junction tree in Figure 5. An assignment of vertices to universes could be B and C to $\{B, C\}$; F, W, and E to $\{B, F, W, E\}$; D to $\{B, W, E, D\}$; L to $\{L, D\}$; M_{in} to $\{W, D, M_{in}\}$; and M_{out} to $\{D, M_{in}, M_{out}\}$.

The potentials in the belief universes are then obtained as follows: The assignment of B to the universe gives for the value *stable* a potential with characteristics (log .85, 0, 0) = (−.16252, 0, 0). The assignment of C gives for the same value a potential where from (3)

$$g_{\{C\}}(stable) = -\frac{(-2)^2}{2\times .1} - \{\log(2\pi\times .1)\}/2$$
$$= -20 + .23235 = -19.76765,$$

and from (4) and (5) we get $h_{\{C\}}(stable) = -2/.1 = 20$ and $k_{\{C\}}(stable) = (1)/.1 = 10$. Adding these numbers and rounding off leads to the potentials

$$g_{\{B,C\}}(stable) = -19.930,$$
$$h_{\{B,C\}}(stable) = -20,\ k_{\{B,C\}}(stable) = 10.$$

Analogously, for the unstable case we find

$$g_{\{B,C\}}(unstable) = -3.881,$$
$$h_{\{B,C\}}(unstable) = -3.333,\ k_{\{B,C\}}(unstable) = 3.333.$$

Only the variable L was assigned to the universe $\{L, D\}$, and this has no discrete parents. Hence only $\mathbf{h}_{\{L,D\}}$ and $\mathbf{K}_{\{L,D\}}$ are needed. From (4) we find that

$$\mathbf{h}_{\{L,D\}} = \frac{3}{.25}\begin{pmatrix} 1 \\ -.5 \end{pmatrix} = \begin{pmatrix} 12 \\ 6 \end{pmatrix},$$

and from (5) we find that

$$\mathbf{K}_{\{L,D\}} = \frac{1}{.25}\begin{pmatrix} 1 & .5 \\ .5 & .25 \end{pmatrix} = \begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix}.$$

Similar calculations must be performed for the remaining belief universes.

The basic computational structure is now established. The belief universes are objects carrying information in the form of potentials, and the separators are communication channels through which information can flow. The computational scheme described in the next section is a combination of entering evidence to relevant universes and a message-passing algorithm.

5.2 Entering Evidence

Incoming evidence is envisaged to be of the type such that certain states are impossible for particular discrete variables or combinations of these, and that certain continuous variables are in specific states.

To be able to handle evidence in our local computational scheme, each item of evidence must be concerned with groups of variables that are members of the same universe in the junction tree. Thus an *item of evidence* is one of the following:

1. a function $\chi_W(\mathbf{i}_W) \in \{0, 1\}$, where W is a set of discrete variables that is a subset of some universe V in the junction tree
2. a statement that $Y_A = y_A^*$ for a particular continuous variable A.

The first type of evidence—which we shall term *discrete* evidence—is entered simply by multiplying χ_W onto the potential ϕ_V. Then it holds that the joint system belief will be proportional to the conditional density, given that the states for which χ_W is equal to 0, are impossible; that is, it represents the conditional belief, given the evidence.

If the second type of evidence—*continuous* evidence—is entered, then the potentials must be modified in all universes V containing A. We must modify the potentials to become those where y_A becomes fixed at the value y_A^*. If the potential ϕ has canonical characteristics $(\mathbf{g}, \mathbf{h}, \mathbf{K})$ with

$$\mathbf{h}(\mathbf{i}) = \begin{pmatrix} \mathbf{h}_1(\mathbf{i}) \\ h_A(\mathbf{i}) \end{pmatrix}, \qquad \mathbf{K}(\mathbf{i}) = \begin{pmatrix} \mathbf{K}_{11}(\mathbf{i}) & \mathbf{K}_{1A}(\mathbf{i}) \\ \mathbf{K}_{A1}(\mathbf{i}) & K_{AA}(\mathbf{i}) \end{pmatrix},$$

then the transformed potentials ϕ^* will have canonical characteristics $(\mathbf{g}^*, \mathbf{h}^*, \mathbf{K}^*)$ given as

$$\mathbf{K}^*(\mathbf{i}) = \mathbf{K}_{11}(\mathbf{i}),$$

$$\mathbf{h}^*(\mathbf{i}) = \mathbf{h}_1(\mathbf{i}) - y_A^* \mathbf{K}_{A1}(\mathbf{i}),$$

and

$$g^*(\mathbf{i}) = g(\mathbf{i}) + h_A(\mathbf{i}) y_A^* - K_{AA}(\mathbf{i})(y_A^*)^2/2.$$

Note that a continuous item of evidence must be entered to all universes and separators of which A is a member.

In the example, we could know that the waste burned was of industrial type and enter this information as the function χ_W with χ_W(industrial) = 1 and χ_W(household) = 0. Similarly, we might have measured the light penetration to be 1.1 and the concentration of CO_2 to be $-.9$, both on the logarithmic scale applied when specifying the conditional distributions. The latter translates to a CO_2 concentration in the emission of 41%. Then the potentials from the initialization are modified to become, for example,

$$g^*_{\{B\}}(\mathit{stable}) = -19.930 + 18 - 4.050 = -5.980$$

and

$$g^*_{\{B\}}(\mathit{unstable}) = -3.881 + 3 - 1.350 = -2.231,$$

as well as

$$h^*_{\{D\}} = 6 - 1.1 \times 2 = 3.8, \qquad k^*_{\{D\}} = 1.$$

5.3 Absorption

The fundamental process in the message-passing algorithm is that of a universe absorbing information from its neighbors in the junction tree. So consider a tree of belief universes with collection $\mathcal{C}$ and separators $\mathcal{S}$ assumed to be supportive. Let $V \in \mathcal{C}$ and let $W_1, \ldots, W_m$ be neighbors of V with separators $S_1, \ldots, S_m$. The universe V is said to *absorb* from $W_1, \ldots, W_m$ if the following operations are performed on the belief potentials,

$$\phi'_{S_i} = \sum_{W_i \setminus V} \phi_{W_i}$$

and

$$\phi'_V = \phi_V * (\phi'_{S_1}/\phi_{S_1}) * \cdots * (\phi'_{S_m}/\phi_{S_m}).$$

In words, the potentials of all the neighbors are marginalized to the separators, and the ratio between the new and old separator potential is passed on as a "likelihood ratio" and multiplied onto the potential at V.

We note that after an absorption, the belief potential for S_i is the marginal of W_i with respect to S_i, and the tree remains supportive. We also have that $\phi_V/(\phi_{S_1} * \cdots * \phi_{S_m}) = \phi'_V/(\phi'_{S_1} * \cdots * \phi'_{S_m})$, whence the joint system belief is unchanged by the absorption process.

In the particular case where $m = 1$, the universes V and W will (under certain circumstances) after absorption "contain the same information" on common variables. More precisely, we say that V and W are *consistent* if $\sum_{V \setminus S} \phi_V \propto \phi_S \propto \sum_{W \setminus S} \phi_W$, and a tree of belief universes is *locally consistent* if all mutual neighbors in the tree are consistent. We then have the following lemma.

Lemma 2. If V absorbs from W and ϕ_S is the strong marginal of ϕ_V, then V and W are consistent after absorption. In fact, $\sum_{V \setminus W} \phi'_V = \phi'_S = \sum_{W \setminus V} \phi'_W$.

Proof. Because the marginalization over $V \setminus W$ is strong, it is composed of integrations and summations only. Hence we find that

$$\sum_{V\setminus W} \phi'_V = \sum_{V\setminus W} \phi_V * (\phi'_S/\phi_S) = (\phi'_S/\phi_S) * \sum_{V\setminus W} \phi_V = \phi'_S.$$

The other equality is trivial.

We emphasize that the corresponding result is false when ϕ_S is only the weak marginal of ϕ_V; see (9). The necessity of using junction trees with strong roots to obtain exact propagation is a consequence of this fact. In the situation described in the lemma, we say that V has *calibrated* to W.

5.4 Collecting Evidence

Based on the notion of absorption, the propagation scheme can now be constructed exactly as in the discrete case. Each $V \in \mathcal{C}$ is given the action COLLECTEVIDENCE. When COLLECTEVIDENCE in V is called from a neighbor W, then V calls COLLECTEVIDENCE in all its other neighbors. When these neighbors have finished their COLLECTEVIDENCE, V absorbs from them (see Fig. 6). We note that because COLLECTEVIDENCE is composed of absorptions only, after COLLECTEVIDENCE the joint system belief is unchanged and the tree remains supportive.

The idea is now to evoke COLLECTEVIDENCE from a strong root R in the junction tree. A flow of activation of neighbors will move through the tree, and a flow of absorptions towards the root will take place. When the flow terminates, the root R will have absorbed the information available from all parts of the tree.

If COLLECTEVIDENCE is evoked from a strong root R and W and W^* are neighbors with separator S such that W is closer in the tree to R than W^*, then the COLLECTEVIDENCE from R has caused W to absorb from W^*. Thus, after COLLECTEVIDENCE, the belief potential for S is the marginal of W^* with respect to S. Because the root is strong, the marginal will be strong. This can be exploited for a second flow through the tree, to be described subsequently.

5.5 Distributing Evidence

After COLLECTEVIDENCE, the root R has absorbed all information available. Next, it must pass this information on to the remaining universes in the tree, formalized as the operation DISTRIBUTEEVIDENCE.

Each $V \in \mathcal{C}$ is given the action DISTRIBUTEEVIDENCE. When DISTRIBUTEEVIDENCE is called in V from a neighbor W, then V absorbs from W and calls DISTRIBUTEEVIDENCE in all its other neighbors.

The activation of DISTRIBUTEEVIDENCE from the root R will create an outward flow of absorptions that will stop when it reaches the leaves of the tree. Again, the joint system belief and supportiveness remain unchanged under DISTRIBUTEEVIDENCE. But when DISTRIBUTEEVIDENCE has terminated, the resulting tree of belief universes will be locally consistent.

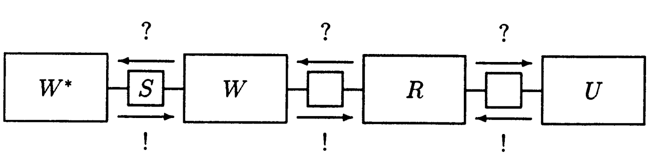

Figure 6. The Calls and Message Passing in COLLECTEVIDENCE.

This follows because after COLLECTEVIDENCE, all separator potentials will be strong marginals of potentials further away from the strong root. When DISTRIBUTEEVIDENCE is subsequently performed, Lemma 2 ensures that all absorptions are calibrations.

We shall now argue that this locally consistent tree is the representation we are aiming for. Ideally, we would want a tree of belief universes such that the probability distributions can be directly inferred from the local belief potentials without having to calculate the joint system belief. This is clearly too much to demand, because the true marginal distribution at any universe would be a mixture of CG distributions and not a CG distribution itself. What we can hope to get is an equality of moments; that is, for each local potential to be the correct (weak) marginal of the joint system belief. That this is in fact true is the main result of this article.

Theorem 1. Let T be a locally consistent junction tree of belief universes with a strong root R and collection $\mathcal{C}$. Let ϕ_U be the joint system belief for T, and let $V \in \mathcal{C}$. Then

$$\sum_{U\setminus V} \phi_U \propto \phi_V. \tag{11}$$

Proof. Let n denote the number of universes in the collection $\mathcal{C}$. We first realize that it is enough to consider the case $n = 2$: If $n = 1$, then the statement obviously holds. If $n > 2$, we can find a leaf L in the tree and use the case $n = 2$ on the junction tree with strong root $R' = \cup_{V\in\mathcal{C}\setminus\{L\}} V$ and one leaf L. By induction, the case gets reduced to $n = 2$.

So assume that $U = R \cup L$, where R is a strong root, and let $S = R \cap L$ be the separator. The marginal to R is a strong marginal, and we find by Lemma 2 that

$$\sum_{L\setminus R} \phi_U = \sum_{L\setminus R} \frac{\phi_L \phi_R}{\phi_S} = \frac{\phi_R}{\phi_S} \sum_{L\setminus R} \phi_L = \phi_R.$$

If S contains only discrete variables, then the marginal to L is also strong and the same calculation applies.

If S contains continuous variables, then $L\setminus R$ contains only continuous variables. Then $L\setminus S \subseteq \Gamma$; that is, only continuous vertices are in the external part of the leaf. Denote the states of the variables in S by $(\mathbf{i}, \mathbf{y})$ and those in $L\setminus S$ by $\mathbf{z}$. Because ϕ_S in the weak marginal of ϕ_U, the moments $p(\mathbf{i})$, $E(\mathbf{Y} \mid \mathbf{I} = \mathbf{i})$, and $V(\mathbf{Y} \mid \mathbf{I} = \mathbf{i})$ are correct when calculated according to ϕ_S or, because these are identical, also according to ϕ_V. That the remaining moments now are correct follows because

$$\begin{aligned} E(\mathbf{Z} \mid \mathbf{I} = \mathbf{i}) &= E\{E(\mathbf{Z} \mid \mathbf{Y}, \mathbf{I} = \mathbf{i})\} \\ &= E\{\mathbf{A}(\mathbf{i}) + \mathbf{B}(\mathbf{i})\mathbf{Y} \mid \mathbf{I} = \mathbf{i}\} \\ &= \mathbf{A}(\mathbf{i}) + \mathbf{B}(\mathbf{i})E(\mathbf{Y} \mid \mathbf{I} = \mathbf{i}), \end{aligned}$$

where $\mathbf{A}(\mathbf{i})$ and $\mathbf{B}(\mathbf{i})$ are determined from ϕ_L/ϕ_S alone. Similarly,

$$\begin{aligned} E(\mathbf{Z}'\mathbf{Y}\,|\,\mathbf{I}=\mathbf{i}) &= E\{E(\mathbf{Z}'\mathbf{Y}\,|\,\mathbf{Y},\mathbf{I}=\mathbf{i})\} \\ &= E\{(\mathbf{A}(\mathbf{i})+\mathbf{B}(\mathbf{i})\mathbf{Y})'\mathbf{Y}\,|\,\mathbf{I}=\mathbf{i}\} \\ &= \mathbf{A}(\mathbf{i})'E(\mathbf{Y}\,|\,\mathbf{I}=\mathbf{i}) + E(\mathbf{Y}'\mathbf{B}(\mathbf{i})\mathbf{Y}\,|\,\mathbf{I}=\mathbf{i}) \\ &= \mathbf{A}(\mathbf{i})'E(\mathbf{Y}\,|\,\mathbf{I}=\mathbf{i}) \\ &\quad + E(\mathbf{Y}\,|\,\mathbf{I}=\mathbf{i})'\mathbf{B}(\mathbf{i})E(\mathbf{Y}\,|\,\mathbf{I}=\mathbf{i}) \\ &\quad + \operatorname{tr}\{\mathbf{B}(\mathbf{i})V(\mathbf{Y}\,|\,\mathbf{I}=\mathbf{i})\}. \end{aligned}$$

Finally,

$$\begin{aligned} \mathbf{V}(\mathbf{Z}\,|\,\mathbf{I}=\mathbf{i}) &= E\{V(\mathbf{Z}\,|\,\mathbf{Y},\mathbf{I}=\mathbf{i})\} + V\{E(\mathbf{Z}\,|\,\mathbf{Y},\mathbf{I}=\mathbf{i})\} \\ &= \mathbf{C}(\mathbf{i}) + \mathbf{B}(\mathbf{i})'\mathbf{C}(\mathbf{i})\mathbf{B}(\mathbf{i}), \end{aligned}$$

where the conditional covariance $\mathbf{C}(\mathbf{i})$ also is determined from ϕ_L/ϕ_S alone, whence the moments are correct.

In summary, after entering evidence the junction tree can be made consistent by evoking CollectEvidence and then DistributeEvidence from a strong root. The weak marginal of the belief at a vertex A subsequently can be obtained from any universe (or even separator) containing A by further weak marginalization. In particular, this gives the correct updated probabilities of the states of any discrete variable and the correct updated mean and variance of any continuous variable.

If the full marginal density is required for a continuous variable, then further computations are needed. These typically involve all variables on the path between a strong root and a universe containing the variable in question. In general, both the density itself and the problem of its computation can be forbiddingly complex.

We want to point out that although the marginal density of variables cannot be obtained explicitly in practice, the tree still contains a fully correct representation of the joint system belief, given the evidence. No information is lost during the flow of evidence. Hence the system remains ready for a correct, exact updating of beliefs when more evidence is obtained.

In the example we have displayed the initial and updated marginal probabilities, the means, and the standard deviations in Table 1.

As was to be expected from measuring a CO_2 emission of 41%, there is strong evidence for an unstable burning regimen, whereas the filter must be intact to explain the penetrability. This, combined with the fact that industrial waste is burned, means that the expected emission of heavy metals has been increased with a factor of exp $1.3 \approx 3.7$.

6. FURTHER TOPICS

We shall first briefly touch on the issues involving feasibility of the computations. The most complex operation is the weak marginalization over a given clique. If the clique contains discrete variables $\delta \in d$ with state spaces of cardinality n_δ and q continuous variables, then the computational complexity of marginalization is of the order of magnitude $q^3 \prod_{\delta\in d} n_\delta$, whereas the storage requirements are about $q^2 \prod_{\delta\in d} n_\delta$. This is because matrix inversion of a $q \times q$ matrix takes about q^3 operations and about q^2 space, and this must be performed for every cell in the table of discrete configurations. These quantities should be compared with $2^q \prod_{\delta\in d} n_\delta$, which is the complexity (of both computation and storage) when the q variables are discretized as much as to a binary variable. Thus when q is large, dramatic savings are possible.

It should be remembered, however, that extra links may have to be added to make the graph strongly triangulated instead of just triangulated. In some cases these extra links may increase clique size so much that the savings are thereby lost.

Another possibility is to ignore the constraint that the junction tree needs a strong root and use an ordinary triangulation for constructing the tree. Then CollectEvidence also would involve weak marginalization and, after propagation, the tree would be only approximately consistent. The quality of such an approximation is to be explored. In particular the approximative methods could give rise to pathologies such as nonpositive definite covariance matrices.

In the case where the original directed graph has continuous variables with discrete children, the initialization will have to be done approximately rather than exactly to take advantage of the CG distributions in the previously described computational scheme. So let $\mathbf{i}$ denote a typical state for a discrete variable with discrete parent states $\mathbf{j}$ and continuous parent states $\mathbf{z}$, where $\mathbf{z} \in \mathcal{R}^q$. We then must approximate $\log p(\mathbf{i}\,|\,\mathbf{j},\mathbf{z})$ with a second degree polynomial in $\mathbf{z}$ for such pairs $(\mathbf{i}, \mathbf{j})$, where $p(\mathbf{i}\,|\,\mathbf{j},\mathbf{z})$ is strictly positive. In particular the positivity is not allowed to depend on $\mathbf{z}$.

An obvious suggestion is to use a CG regression model for the conditional probabilities and let

$$\log p(\mathbf{i}\,|\,\mathbf{j},\mathbf{z}) = a(\mathbf{i}\,|\,\mathbf{j}) + \mathbf{b}(\mathbf{i}\,|\,\mathbf{j})'\mathbf{z} + \mathbf{z}'\mathbf{C}(\mathbf{i}\,|\,\mathbf{j})\mathbf{z} - \kappa(\mathbf{j},\mathbf{z}),$$

Table 1. Probabilities, Means, and Standard Deviations of Single Variables in Our Example Before and After Evidence Has Been Entered

	F	W	B	Means and standard deviations					
Status	p(i)	p(h)	p(s)	M_{in}	M_{out}	E	D	C	L
Initial values	.95	.71	.85	−.21	2.83	−3.25	3.04	−1.85	1.48
				.46	.86	.71	.77	.51	.63
Updated values	.9995	0	.01	.50	4.11	−3.90	3.61	−0.90	1.10
				.10	.35	.07	.33	0	0

NOTE: The information strongly suggests that the filter is intact but the burning regimen is unstable. Consequently, there is increased emission of dust and metals.

where

$$\kappa(\mathbf{j}, \mathbf{z}) = \log \sum_{\mathbf{i}} \exp\{a(\mathbf{i} \mid \mathbf{j}) + \mathbf{b}(\mathbf{i} \mid \mathbf{j})'\mathbf{z} + \mathbf{z}'\mathbf{C}(\mathbf{i} \mid \mathbf{j})\mathbf{z}\}.$$

This seems natural, because CG regressions occur as conditional distributions in CG distributions (see Lauritzen and Wermuth 1989).

Because κ is not a quadratic in $\mathbf{z}$, $\log p$ is to be approximated by its second-order Taylor expansion around its maximal value. Various optimization methods can be used to find this maximal value; we abstain from discussing this point in detail here.

When the approximate initialization has been performed and one recalculates the conditional distribution of $\mathbf{i}$ given the parent states, this will be of the same type as initially specified but the coefficients may have changed slightly. This can be used indirectly to indicate the quality of the approximation. We believe that the error of approximation is negligible compared to the general uncertainty involved in the model building itself.

REFERENCES

Hansen, J. A., and Dalager, S. (eds.) (1985), *Emission fra affaldsforbrændingsanlæg* (in Danish), Copenhagen: DAKOFA.

Jensen, F. V., Lauritzen, S. L., and Olesen, K. G. (1990), "Bayesian Updating in Causal Probabilistic Networks by Local Computations," *Computational Statistics Quarterly,* 4, 269–282.

Kiiveri, H., Speed, T. P., and Carlin, J. B. (1984), "Recursive Causal Models," *Journal of the Australian Mathematical Society,* Ser. A, 36, 30–52.

Lauritzen, S. L. (1989), "Mixed Graphical Association Models" (with discussion), *Scandinavian Journal of Statistics,* 16, 273–306.

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), *Journal of the Royal Statistical Society,* Ser. B, 50, 157–224.

Lauritzen, S. L., and Wermuth, N. (1989), "Graphical Models for Associations Between Variables, Some of Which Are Qualitative and Some Quantitative," *The Annals of Statistics,* 17, 31–57.

Leimer, H.-G. (1989), "Triangulated Graphs With Marked Vertices," in Graph Theory in Memory of G. A. Dirac, eds. L. D. Andersen, C. Thomassen, B. Toft, and P. D. Vestergaard. *Annals of Discrete Mathematics,* 41, 311–324.

Oliver, R., and Smith, J. Q. (eds.) (1990), *Influence Diagrams, Belief Nets and Decision Analysis,* Chichester, U.K.: John Wiley.

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

Shafer, G. R., and Pearl, J. (eds.) (1990), *Readings in Uncertain Reasoning,* San Mateo, CA: Morgan Kaufmann.

Shenoy, P. P., and Shafer, G. R. (1990), "Axioms for Probability and Belief-Function Propagation," in *Uncertainty in Artificial Intelligence,* 4, eds. R. D. Shachter, T. S. Levitt, L. N. Kanal, and J. F. Lemmer. Amsterdam: North-Holland, pp. 169–198.

Whittaker, J. (1990), *Graphical Models in Applied Multivariate Statistics,* Chichester, U.K.: John Wiley.