Stable Local Computation with Conditional Gaussian Distributions[*]

Steffen L. Lauritzen
Aalborg University

Frank Jensen
HUGIN Expert Ltd.

September 1999

Abstract: This article describes a propagation scheme for Bayesian networks with conditional Gaussian distributions that does not have the numerical weaknesses of the scheme derived in Lauritzen (1992). The propagation architecture is that of Lauritzen and Spiegelhalter (1988).

In addition to the means and variances provided by the previous algorithm, the new propagation scheme yields full local marginal distributions. The new scheme also handles linear deterministic relationships between continuous variables in the network specification.

The new propagation scheme is in many ways faster and simpler than previous schemes and the method has been implemented in the most recent version of the HUGIN software.

Key words: Artificial intelligence, Bayesian networks, CG distributions, Gaussian mixtures, probabilistic expert systems, propagation of evidence.

1 Introduction

Bayesian networks have developed into an important tool for building systems for decision support in environments characterized by uncertainty (Pearl 1988; Jensen 1996; Cowell *et al.* 1999).

The exact computational algorithms that are most developed are concerned with networks involving discrete variables only.

[*] This is Research Report R-99-2014, Department of Mathematical Sciences, Aalborg University.

Lauritzen (1992) developed a computational scheme for exact local computation of means and variances in networks with conditional Gaussian distributions. Unfortunately the scheme turned out to have fatal numerical difficulties, basically due to a computationally unstable transformation between two different representations of these distributions.

The motivation for the present work is to remedy this numerical instability. The fundamental idea behind the developments below is at all times to keep the interesting quantities represented in units that have a direct meaning such as probabilities, means, regression coefficients, and variances. These must necessarily be of a reasonable order of magnitude.

The computational scheme to be developed is rather remote from the computational architecture used to deal with the discrete variables in the HUGIN software and similar schemes as represented, for example, in abstract form in Shenoy and Shafer (1990) and Lauritzen and Jensen (1997). The difference is partly related to the fundamental operations of combination and marginalization being only partially defined, but also the handling of evidence is quite different. The scheme is closest to the original scheme developed in Lauritzen and Spiegelhalter (1988), but abstract considerations such as those in Shafer (1991) seem necessary to embed the scheme in a unifying framework.

Additional benefits of the present scheme includes that deterministic linear relationships between the continuous variables can be represented without difficulty, and we show how to calculate full local marginals of continuous variables without much computational effort. Both of these represent major improvements over the original scheme of Lauritzen (1992).

2 CG distributions and regressions

The Bayesian networks to be considered have distributions that are conditionally Gaussian, a family of distributions introduced by Lauritzen and Wermuth (1984, 1989). We shall briefly review some standard notation but otherwise refer the reader to Lauritzen (1996) for further details.

The set of variables V is partitioned as $V = \Delta \cup \Gamma$ into variables of *discrete* (Δ) and *continuous* (Γ) type and the joint distribution of the continuous variables given the discrete is assumed to be multivariate Gaussian, i.e.

$$\mathcal{L}(Y \,|\, I = i) = \mathcal{N}_{|\Gamma|}(\xi(i), \Sigma(i)) \quad \text{whenever} \quad p(i) = P\{I = i\} > 0,$$

where Y denotes the continuous variables, I the discrete, $|\Gamma|$ denotes the cardinality of Γ, and $\Sigma(i)$ is positive semidefinite. We then say that $X = I \cup Y$ follows a *CG distribution.*

The symbol $\mathcal{N}_{|\Gamma|}(\xi, \Sigma)$ denotes the multivariate Gaussian distribution with mean ξ and covariance matrix Σ. In the case where Σ is positive definite, this distribution has density

$$f(y \,|\, \xi, \Sigma) = \left\{(2\pi)^{|\Gamma|} \det \Sigma\right\}^{-1/2} \exp\left\{-\tfrac{1}{2}(y-\xi)^\top \Sigma^{-1}(y-\xi)\right\}.$$

If Σ is singular, the multivariate Gaussian distribution has no density but is implicitly determined through the property that for any vector v, the linear combination $v^\top Y$ has a univariate Gaussian distribution:

$$\mathcal{L}(v^\top Y) = \mathcal{N}_1(v^\top \xi, v^\top \Sigma v),$$

where $\mathcal{N}_1(\mu, 0)$ is to be interpreted as the distribution degenerate at μ. See for example Rao (1973), Chapter 8, for a description of the Gaussian distribution at this level of generality.

Note: there is a slight difference between the terminology used here and in Lauritzen (1996) in that we allow $p(i)$ to be equal to 0 for some entries i. We also avoid using the so-called canonical characteristics of the CG distribution as the numerical instability of the scheme in Lauritzen (1992) is associated with switching between these and the moment characteristics (p, ξ, Σ). As an additional benefit, we can then allow singular covariance matrices Σ.

Occasionally it is of interest to describe how a CG distribution depends on additional variables. If the dependence on a set of discrete variables j and a vector of continuous variables z is determined as

$$\begin{aligned} p(i \,|\, J = j, Z = z) &= p(i \,|\, j), \\ \mathcal{L}(Y \,|\, I = i, J = j, Z = z) &= \mathcal{N}(A(i \,|\, j) + B(i \,|\, j)z, C(i \,|\, j)), \end{aligned}$$

we refer to this dependence as a (simple) *CG regression.* Note that neither the covariance matrix nor the discrete part depends on the continuous variables z and the conditional expectation of the continuous variables depends linearly on the continuous variables for fixed values of the discrete variables (i, j). In a general CG regression, p is also permitted to depend on z in a specific way (Lauritzen 1996), but this is not relevant here.

3 Mixed Bayesian networks

We consider probabilistic networks over a directed acyclic graph (DAG), known as Bayesian networks (Pearl 1986). A mixed Bayesian network with conditional Gaussian distributions is specified over a set of nodes or variables V, partitioned as $V = \Delta \cup \Gamma$ into discrete and continuous variables as above. The DAG associated with the network must satisfy the restriction that discrete nodes have no continuous parents. The conditional distributions of discrete variables given their (discrete) parent variables are specified as usual, whereas the conditional distribution of continuous variables are given by CG regressions

$$\mathcal{L}(Y \,|\, I = i, Z = z) = \mathcal{N}(\alpha(i) + \beta(i)^\top z, \gamma(i)).$$

Note that as Y is one-dimensional, $\gamma(i)$ is just a nonnegative real number. If $\gamma(i) = 0$, this conditional distribution specifies a linear and deterministic dependence of Y on Z.

The assumptions above imply that the joint distribution of all variables in the Bayesian network is a CG distribution.

The computational task to be addressed is that of computing the joint distribution of interesting subsets of these variables — in particular of a single variable — possibly given specific evidence, i.e. given known values of arbitrary subsets of other variables in the network. This distribution will in general be a mixture of conditional Gaussian distributions.

The propagation scheme to be described involves the usual steps: Construction of a junction tree with strong root, initialization of the junction tree, incorporation of evidence, and local computation of marginals to cliques.

4 Potentials and their operations

4.1 CG potentials

The basic computational object is that of a *CG potential.* A CG potential is represented as $\phi = [p, A, B, C](H \,|\, T)$. Here $(H \,|\, T)$ denotes a partitioning of the continuous variables in the domain D of ϕ into *head* and *tail*: $D \cap \Gamma = H \cup T$. We denote the variables in the head by Y and those in the tail by Z and assume these to be r and s-dimensional. An arbitrary configuration of the discrete variables in the domain is denoted by i. Thus, every potential has a domain with discrete nodes, head nodes and tail nodes, some of which could be absent. In the expression above

- $p = \{p(i)\}$ is a table of nonnegative numbers, i.e. a 'usual' potential as in the discrete case;
- $A = \{A(i)\}$ is a table of $r \times 1$ vectors;
- $B = \{B(i)\}$ is a table of $r \times s$ matrices;
- $C = \{C(i)\}$ is a table of $r \times r$ positive semidefinite symmetric matrices.

The potential represented by $[p, A, B, C](H \mid T)$ specifies the CG regression

$$P(I = i) \propto p(i), \quad \mathcal{L}(Y \mid I = i, Z = z) = \mathcal{N}_r(A(i) + B(i)z, C(i)).$$

The abstract notion of potentials with head and tail is due to Shafer (1991). In many ways it would be more natural also to partition the discrete variables into head and tail variables, then reflecting that the potentials always represent a conditional distribution of head variables given their tail. But as the partitioning of discrete variables is not exploited in our propagation scheme, we have chosen not to do so. A propagation scheme of the 'lazy' type (Madsen and Jensen 1998) could exploit such a partitioning.

The initial conditional distribution for a continuous variable v with parent nodes $\mathrm{pa}(v)$ in a mixed Bayesian network corresponds to the CG potential $[1, \alpha, \beta^\top, \gamma](H \mid T)$ with $H = \{v\}$, $T = \mathrm{pa}(v) \cap \Gamma$, and discrete part of the domain equal to $\mathrm{pa}(v) \cap \Delta$. Similarly, the specification of the conditional distribution of a discrete variable given its parents corresponds to the CG potential $[p, -, -, -](- \mid -)$, where p is determined by the conditional probability tables. The discrete part of the domain is equal to the family $\mathrm{fa}(v) = v \cup \mathrm{pa}(v)$, and hyphens indicate that the corresponding parts of the potential are void.

4.2 Extension and reduction

A CG potential can be extended by adding discrete variables to its domain or continuous variables to its tail. When adding discrete variables to its domain, the parts of ϕ are extended as $p^*(i, j) = p(i)$ etc. When adding continuous variables to its tail, the B matrices are extended by adding zero columns for each of the new tail variables:

$$B^*(i) = \{B(i) : 0\}.$$

Similarly, if B has columns that are identically zero for all values of i, the corresponding variables can be removed from the tail of the potential, and we say that the tail is *reduced*. If no columns of B are identically zero, the tail of the potential is said to be *minimal*.

4.3 Marginals

As in the propagation scheme of Lauritzen (1992), marginals of a CG potential are only defined under certain conditions and when marginals over groups of discrete and continuous variables are calculated, the marginals over continuous variables are calculated first.

Marginals over continuous variables can only be calculated over head variables. If $[p, A, B, C](H \,|\, T)$ is decomposed as

$$H = (H_1, H_2), \quad A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}, \quad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}, \quad C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix},$$

corresponding to a partitioning of the head variables as $Y = (Y_1, Y_2)$, the marginal of ϕ to $D' = D \setminus H_2$ is given as

$$\phi^{\downarrow D'} = [p, A_1, B_1, C_{11}](H_1 \,|\, T).$$

We say that these marginals are *strong* as they correspond to calculating ordinary marginals of the relevant conditional Gaussian distributions.

When all head variables have been removed by marginalization, the tail can be reduced to become empty so that a discrete potential emerges. This leads indirectly to marginalization of tail variables.

Marginals over discrete variables are defined only when the tail of the potential is empty, i.e. when there are no continuous conditioning variables and therefore no B matrix. Then the marginal of the CG potential $\phi = [p, A, -, C](H \,|\, -)$ with discrete domain partitioned as $U \cup W$ over W is

$$\phi^{\downarrow U \cup H} = [\tilde{p}, \tilde{A}, -, \tilde{C}](H \,|\, -),$$

where

$$\begin{aligned}
\tilde{p}(i_U) &= \sum_{i_W} p(i) \\
\tilde{A}(i_U) &= \frac{1}{\tilde{p}(i_U)} \sum_{i_W} A(i) p(i) \\
\tilde{C}(i_U) &= \frac{1}{\tilde{p}(i_U)} \sum_{i_W} \left\{ C(i) + \left[A(i) - \tilde{A}(i_U)\right] \left[A(i) - \tilde{A}(i_U)\right]^{\top} \right\} p(i),
\end{aligned}$$

where $i = (i_U, i_W)$. This marginalization is said to be *weak* when it does not correspond to calculating the full marginal distribution.

In general the full marginal distribution will be a discrete mixture of CG distributions, and the distribution represented by the weakly marginalized potential will be the CG distribution closest in Kullback–Leibler distance to the true marginal, see Lauritzen (1996), Lemma 6.4.

4.4 Direct combination

The combination operation for CG potentials will not be defined for an arbitrary pair of potentials and as such the scheme is quite different from most other propagation schemes.

The direct combination of two CG potentials $\phi = [p, A, B, C](H_1 \,|\, T_1)$ and $\psi = [q, E, F, G](H_2 \,|\, T_2)$ is defined only if the head of ψ is disjoint from the domain of ϕ, i.e. satisfies that

$$H_2 \cap D_1 = \emptyset. \tag{1}$$

Here we always assume that the potentials have first been reduced so that the tails are minimal.

If (1) is fulfilled for the reduced potentials, these are subsequently extended such that the extensions have $T_2 = H_1 \cup T_1$. This is done by extending T_1 to $T_1 \cup (T_2 \setminus H_1)$ and T_2 to $T_2 \cup H_1 \cup T_1$.

Next, let $F = [F_1 : F_2]$ be partitioned into $r_2 \times r_1$ and $r_2 \times s_1$ matrices corresponding to (H_1, T_1). We then define the *direct combination* as the (apparently non-commutative) product

$$[\rho, U, V, W](H \,|\, T) = [p, A, B, C](H_1 \,|\, T_1) \,\dot{\otimes}\, [q, E, F, G](H_2 \,|\, T_2),$$

where

$$H = H_1 \cup H_2, \quad T = (T_1 \cup T_2) \setminus H, \quad D = D_1 \cup D_2,$$

and

$$\begin{aligned} \rho &= pq \\ U &= \begin{pmatrix} A \\ E + F_1 A \end{pmatrix} \\ V &= \begin{pmatrix} B \\ F_2 + F_1 B \end{pmatrix} \\ W &= \begin{pmatrix} C & C F_1^\top \\ F_1 C & G + F_1 C F_1^\top \end{pmatrix}. \end{aligned}$$

This combination operation corresponds to ordinary composition of conditional distributions. Note that if both of $\phi \mathbin{\dot{\otimes}} \psi$ and $\psi \mathbin{\dot{\otimes}} \phi$ exist, they are equal. The direct combination also satisfies

$$(\phi \mathbin{\dot{\otimes}} \psi) \mathbin{\dot{\otimes}} \eta = \phi \mathbin{\dot{\otimes}} (\psi \mathbin{\dot{\otimes}} \eta)$$

in the sense that if the combinations on one side are well defined, so are those on the other side and the resulting potentials are the same. Shafer (1991) has called this type of algebraic structure a *partial commutative semigroup*.

The notation above reflects that the operation of direct combination in some sense is similar to that of forming disjoint union of sets.

Unfortunately, direct combination of CG potentials is not sufficient for our propagation scheme to work for an arbitrary mixed Bayesian network. But before we can define a more general combination, we need to introduce the notion of complement.

4.5 Complements

If the head of a CG potential $\phi = [p, A, B, C](H \,|\, T)$ is partitioned as

$$H = (H_1, H_2), \quad A = \begin{pmatrix} A_1 \\ A_2 \end{pmatrix}, \quad B = \begin{pmatrix} B_1 \\ B_2 \end{pmatrix}, \quad C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix},$$

and $[p^*, A_1, B_1, C_{11}](H_1 \,|\, T)$ is the strong marginal of ϕ, then we define its *complement* $\phi^{|H_1 \cup T}$ as the CG potential $[q, E, F, G](H_2 \,|\, H_1 \cup T)$, where

$$\begin{aligned} q &= p/p^* \\ E &= A_2 - C_{21} C_{11}^{-} A_1 \\ F &= [\, C_{21} C_{11}^{-} : B_2 - C_{21} C_{11}^{-} B_1 \,] \\ G &= C_{22} - C_{21} C_{11}^{-} C_{12}. \end{aligned}$$

Here M^{-} denotes an arbitrary generalized inverse of the matrix M (Penrose 1955), i.e. an arbitrary matrix M^{-} satisfying

$$M M^{-} M = M, \tag{2}$$

see also Rao (1973), pp. 24–27, and Rao and Mitra (1971). Then

$$[p, A, B, C](H \,|\, T) = [p^*, A_1, B_1, C_{11}](H_1 \,|\, T) \mathbin{\dot{\otimes}} [q, E, F, G](H_2 \,|\, H_1 \cup T),$$

which is easily checked by using the formulae for combination together with (2) and the fact that for any generalized inverse C_{11}^{-} of C_{11} it also holds that

$$C_{21} C_{11}^{-} C_{11} = C_{21},$$

see e.g. Rao (1973), formula (8a.2.12).

Note that in the above expressions we either have $p^* = p$ or $H_1 = \emptyset$. The decomposition of a potential into its strong marginal and its complement corresponds exactly to the decomposition of a probability distribution into its marginal and conditional.

4.6 Recursive combination

We next define a more general combination of CG potentials. This is required for the initialization process described in the section below. Consider again two potentials $\phi = [p, A, B, C](H_1 \,|\, T_1)$ and $\psi = [q, E, F, G](H_2 \,|\, T_2)$ with minimal tails. If $H_1 \cap H_2 \neq \emptyset$ the combination will remain undefined. If the heads of the potentials are disjoint, we let

$$\phi \otimes \psi = \phi \,\dot{\otimes}\, \psi \quad \text{or} \quad \phi \otimes \psi = \psi \,\dot{\otimes}\, \phi$$

if at least one of the right-hand-side expressions are defined. As we have $\phi \,\dot{\otimes}\, \psi = \psi \,\dot{\otimes}\, \phi$ if both are defined, there is no ambiguity in this definition.

If neither of the direct combinations are defined, we must have that

$$H_1 \cap D_2 \neq \emptyset \quad \text{and} \quad H_2 \cap D_1 \neq \emptyset. \tag{3}$$

Let $D_{12} = H_1 \setminus D_2$ and $D_{21} = H_2 \setminus D_1$. If both of these are empty, the combination will not be defined. Else we decompose one of the factors, say ϕ (assuming $D_{12} \neq \emptyset$), as

$$\phi = \phi^{\downarrow(D_1 \setminus D_{12})} \,\dot{\otimes}\, \phi^{|(D_1 \setminus D_{12})} = \phi' \,\dot{\otimes}\, \phi''$$

and attempt to combine ϕ and ψ as

$$\phi \otimes \psi = (\phi' \otimes \psi) \,\dot{\otimes}\, \phi''.$$

This equation is to be understood recursively in the sense that the procedure described is to be repeated for the product $\phi' \otimes \psi$, whereas the direct combination in the expression is well defined by construction.

The recursion terminates unsuccessfully if two potentials with minimal tails satisfy (3) and also

$$H_1 \setminus D_2 = H_2 \setminus D_1 = \emptyset. \tag{4}$$

Then the combination of ϕ and ψ remains undefined.

5 Initialization

Setting up the computational structure involves several steps: forming a strong junction tree with strong root, assigning potentials to cliques, transforming these to potentials of a specific form by sending messages first towards the root, then away from the root.

A junction tree with strong root is constructed in the usual way, see for example Cowell *et al.* (1999), Chapter 7. Thus, we assume to begin our computational scheme at the point where we have specified a mixed Bayesian network and an associated junction tree with cliques $\mathcal{C}$ and a root $R \in \mathcal{C}$ such that for all neighbouring cliques C and D with C closer to the root than D, we have that

$$S = C \cap D \subseteq \Delta \quad \text{or} \quad D \setminus C \subseteq \Gamma, \tag{5}$$

i.e. if the 'residual' $D \setminus C$ contains a discrete variable, then the separator S consists of discrete variables only. Also, it holds for all variables v that $\mathrm{fa}(v)$ is contained in some clique of the junction tree.

5.1 Assignment of potentials to cliques

Every CG potential corresponding to a specification of the conditional distribution of a node given its parents is assigned to an arbitrary clique of the junction tree that contains its family. The potentials assigned to a given clique are subsequently combined in some order. This can always be done using direct combination as the DAG is acyclic and each continuous node is head of exactly one potential.

5.2 Collecting messages at the root

The next step in the initialization process involves sending messages from the leaves of the junction tree towards the root in a way similar to the process known as CollectEvidence in the standard HUGIN architecture (Jensen *et al.* 1990), although the messages sent are slightly different. Thus, a clique is allowed to send a message if it is a leaf of the junction tree, or if it has received messages from all of its neighbours further away from the root. The process stops when the root has received messages from all of its neighbours. We use the term Collect for this operation.

When a Collect-message is sent from a clique C to its neighbour D towards the root with separator $S = C \cap D$, the potentials ϕ_C on C and ϕ_D

on D are modified to become ϕ_C^* and ϕ_D^*, where

$$\phi_C^* = \phi_C^{|S}, \quad \phi_D^* = \phi_D \otimes \phi_C^{\downarrow S}, \tag{6}$$

i.e. ϕ_C^* is the complement of ϕ_C after marginalization to the separator and ϕ_D^* is obtained by combining the original potential with the marginal of ϕ_C. It remains to be argued that the combination in (6) is indeed well defined.

To see this we first realize that the heads of any two potentials to be combined must necessarily be disjoint as a variable occurs only once as head.

Further, for any of the potentials involved in (6), it holds that tail variables have no parents in the DAG induced by the conditional specifications that have been combined and possibly marginalized to form the potential. Thus, if the potential is reduced to have minimal tail, there must be a directed path from every variable present in the tail of the potential to some variable in the head of the potential. Because then it holds for any tail variable u that it is not conditionally independent of the head given the remaining tail variables. Thus there must be a trail which d-connects u to some variable in the head. As tail variables have no parents, this trail must initially be directed away from u and leave the tail immediately. As only tail variables are in the conditioning set, there can be no head-to-head nodes on this active trail, which then must form a directed path from u to the head.

Assume that (4) is satisfied and H_1 and H_2 are both nonempty. This implies $H_1 \subseteq T_2$ and $H_2 \subseteq T_1$. From this we deduce that there must be a directed path from every variable $u \in H_1$ (implying $u \in T_2$) to some variable $v \in H_2$ (implying $v \in T_1$), and from v there must be a directed path to some variable $w \in H_1$. Thus, from every $u \in H_1$ there is a directed path to some $w \in H_1$, and since H_1 is nonempty and finite, this would contradict the acyclicity of the DAG.

To illustrate that recursive combination is necessary for the initialization process, we consider two simple examples.

Example 1 Consider the DAG in Figure 1. When potentials are assigned to cliques, the nodes c and e must be assigned to $\{b, c, e\}$ and the remaining nodes to $\{a, b, c, d\}$.

Combining the potentials in the two cliques leads to potentials with head and tail $(\{c, e\} \,|\, \{b\})$ in $\{b, c, e\}$ and $(\{b, d\} \,|\, \{c\})$ in $\{a, b, c, d\}$.

When the first of these is marginalized to the separator $\{b, c\}$, the result has head and tail $(\{c\} \,|\, \{b\})$, which cannot be directly combined with the potential on the root clique $\{a, b, c, d\}$.

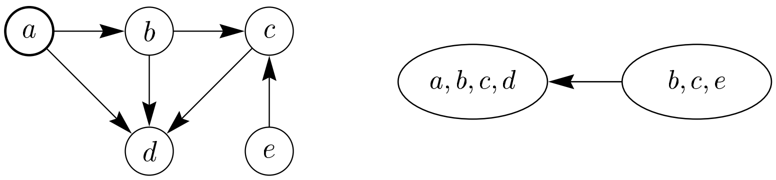


Figure 1: A mixed Bayesian network with associated junction tree. The variable a is the only discrete variable and the strong root is $\{a, b, c, d\}$.

The root clique potential is then decomposed into potentials with head and tail $(\{d\} \,|\, \{b, c\})$ and $(\{b\} \,|\, \{c\})$. But the latter can be reduced to $(\{b\} \,|\, -)$ as the dependence on c is spurious. The potentials can now be combined directly. □

In the example above, it was the potential in the receiving clique that was decomposed. And had we not combined the potentials in the receiving clique before combining with the incoming message, the combination could have been performed directly. The next example illustrates that it may be the incoming message which needs to be decomposed and there is no way to avoid computation during the decomposition.

Example 2 Consider the DAG in Figure 2. When potentials are assigned to cliques, the nodes d, e and f must be assigned to $\{c, d, e, f\}$, c to $\{a, c, d, e\}$, and b to $\{a, b, d\}$. There are two choices for the node a and we choose to assign it to the clique $\{a, b, d\}$, which is also chosen as root.

When Collecting towards the root, the first message is the $\{c, d, e\}$-marginal of the potential in $\{c, d, e, f\}$. This must be calculated by combining the assigned potentials to one with head and tail $(\{d, e, f\} \,|\, \{c\})$ and then marginalizing to $(\{d, e\} \,|\, \{c\})$.

Again this cannot be directly combined with the potential on the neighbouring clique which has head and tail $(\{c\} \,|\, \{e\})$.

The incoming potential is then decomposed into potentials with head and tail $(\{d\} \,|\, \{c, e\})$ and $(\{e\} \,|\, \{c\})$. But the latter can be reduced to $(\{e\} \,|\, -)$ as the dependence on c is spurious. The potentials can now be combined directly. □

After the root has received messages from all its neighbours, the root potential contains the correct root marginal and its tail is then empty. If evidence has been incorporated, a normalization of the discrete part of the root potential may be necessary, see Section 7.

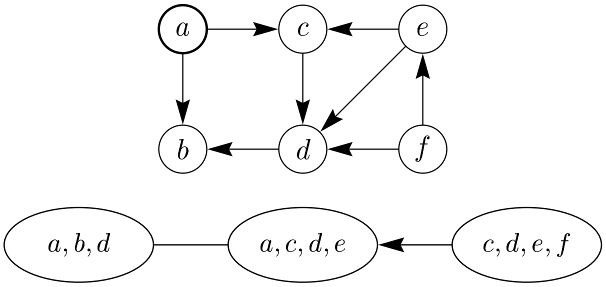


Figure 2: A mixed Bayesian network with associated junction tree. The variable a is the only discrete variable and the strong root could be chosen to be either $\{a, b, d\}$ or $\{a, c, d, e\}$.

Also, the potential ϕ representing the joint distribution of all the variables is now equal to the combination of all the clique potentials ϕ_C

$$\phi = \bigotimes_{C\in\mathcal{C}} \phi_C. \tag{7}$$

In fact, as all marginals computed during the Collect phase have been strong, it holds for any subset $\mathcal{C}' \subseteq \mathcal{C}$ which contains the root R and forms a connected subtree of the junction tree that

$$\phi^{\downarrow C'} = \bigotimes_{C\in\mathcal{C}'} \phi_C, \tag{8}$$

where $C' = \bigcup_{C\in\mathcal{C}'} C$. As the complements are stored in the cliques during Collect and the separators are not playing a specific role during this process, the computation is similar to the process of forming a set chain in Lauritzen and Spiegelhalter (1988). Thus the inward computation is of the type called Lauritzen–Spiegelhalter architecture in Shafer (1996), see also Lauritzen and Jensen (1997).

5.3 Distributing messages from the root

The first step in the calculation of marginals involves sending messages away from the root, similar to DistributeEvidence in the standard HUGIN architecture. The root begins by sending messages to all its neighbours, and a clique is allowed to send a message as soon as it has received one from its neighbour closer to the root. We use the term Distribute for this process which again has slightly different messages than in the standard HUGIN architecture.

When a DISTRIBUTE-message is sent from a clique C to its neighbour D further away from the root with separator $S = C \cap D$ between them, C has just received a message from its neighbour towards the root. We make the inductive assumption that the separator S' towards the root then contains the weak clique marginal of the joint potential

$$\phi_{S'} = \phi^{\downarrow S'}.$$

When sending a message, a new potential ϕ_S is created on S as follows. First the weak clique marginal at C is calculated as

$$\phi^{\downarrow C} = \phi_{S'} \dot{\otimes} \phi_C. \tag{9}$$

That this formula is correct is seen exactly as in Lauritzen (1992). Next this potential is further marginalized to the separator

$$\phi_S = (\phi^{\downarrow C})^{\downarrow S} = \phi^{\downarrow S}.$$

The combination is well defined because after the collect operation, complements were stored in the cliques so the head of ϕ_C is disjoint from S' and the (weak) marginal is well defined as the tail of ϕ_C is contained in the head of $\phi_{S'}$ implying that the combination in (9) has empty tail.

After DISTRIBUTE the separators all contain weak marginals to the separator nodes.

Note that we have chosen not to store the weak clique marginals calculated under DISTRIBUTE, but preferred to keep the original complement potentials. This is a minor variation of the Lauritzen–Spiegelhalter architecture.

The initialization process is now completed. The cliques of the junction tree contain complement potentials, the separators contain weak marginals of the joint potential, and this joint potential can be recovered by (7).

6 Computation of marginals

When the junction tree has been initialized as described in the previous section, various types of marginals can easily be calculated.

6.1 Marginals of variables in a single clique

If not stored separately, weak clique marginals can always be recalculated as in (9) when needed, and further marginalized to subsets of cliques, in particular to single nodes.

Under some circumstances, these weak marginals happen to be strong and give the correct, full marginal distribution of the variables involved. This is clearly true if the desired marginal involves discrete variables only. But there are other cases of interest when this is true.

As already mentioned, the root clique contains the correct full marginal distribution of its variables. Thus, for example, the true marginal of the set of continuous variables Y in the root clique can be easily calculated as a Gaussian discrete mixture with weights $p(i)$, i.e.

$$\mathcal{L}(Y) = \sum_i p(i) * \mathcal{N}(A(i), C(i)), \tag{10}$$

where the root potential is $[p, A, -, C](R \cap \Gamma \,|\, -)$. Further marginalization can then easily be performed.

But the same holds for a clique C that satisfies the slightly less restrictive condition that *the tail of its potential is empty.* For example, this is the case if the separator of the clique C towards the root contains discrete variables only.

To see this we argue as follows. From (8) we have that the true marginal to the union of cliques on the path from the root to C is given by combination of the relevant potentials

$$\phi^{\downarrow D} = \bigotimes_{j=1}^{k} \phi_{C_j},$$

where

$$D = \bigcup_{j=1}^{k} C_j$$

and the cliques on the path are $R = C_1, \ldots, C_k = C$. The continuous variables in C are conditionally independent of the remaining discrete variables on this path, given the separator variables; as the tail of the potential on C is assumed empty, this also holds given just the discrete separator variables. Proposition 6.3 of Lauritzen (1996) then yields that the weak marginal to C is also equal to the full marginal and we can proceed as with the root clique.

6.2 Rearranging the junction tree

To obtain the marginal of a set of variables that is not a subset of some clique of the junction tree or to obtain strong marginals of a group of variables or a single variable that is not in a clique having a potential with an empty tail, the junction tree must be rearranged. Fortunately there is a simple operation that can be used to achieve the necessary rearrangement which we denote by PUSH. It acts on a group of variables M which are contained in a clique W with neighbour U towards the root and corresponding separator $S = U \cap W$. The operation PUSH appplied to the variables M does the following:

1. The potential ϕ_W is decomposed as
$$\phi_W = (\phi_W)^{\downarrow M\cup S} \dot{\otimes} (\phi_W)^{|M\cup S}.$$
2. The clique U is extended to $U^* = U \cup M$ and similarly $S^* = S \cup M$.
3. The potentials are changed as
$$\phi_{U^*} = \phi_U \dot{\otimes} (\phi_W)^{\downarrow M\cup S}, \quad \phi_{S^*} = \phi_S \dot{\otimes} (\phi_W)^{\downarrow M\cup S}, \quad \phi_W = (\phi_W)^{|M\cup S}.$$

After the PUSH operation the variables in M have come closer to the strong root, but the extended junction tree still represents the joint potential as after the initialization. The price that has been paid is that the clique U has increased to U^*.

Example 3 We illustrate the PUSH operation using the mixed Bayesian network in Figure 2, assuming that we have chosen $\{a, b, d\}$ as root.

After initialization the clique $\{a, c, d, e\}$ contains the potential representing the conditional distribution of variables $\{c, e\}$ given $\{a, d\}$ having head and tail $(\{c, e\} \,|\, \{d\})$.

If we use PUSH on $\{c\}$, this potential is decomposed into its marginal with head and tail $(\{c\} \,|\, \{d\})$ and complement with head and tail $(\{e\} \,|\, \{c, d\})$. The root clique is now extended with the variable c and the marginal is combined with the root potential, whereas the complement is kept in the clique $\{a, c, d, e\}$. □

6.3 Marginals of variables in different cliques

If a (weak) marginal is desired of a set of variables that is not a subset of some clique of the original junction tree, we first form the smallest connected subtree of the original junction tree that contains all the variables. Let C be the clique of the subtree that is closest to the strong root of the original junction tree. By repeated use of the PUSH operation we eventually achieve that the variables in question all become members of C. The desired weak marginal can then be computed directly using (9).

6.4 Strong marginals

If the strong marginal of a group of variables is desired, the Push operation again yields the appropriate rearrangement of the junction tree.

As in the computation of weak marginals, we first form the smallest connected subtree of the original junction tree that contains all the variables. Let C be the clique of this subtree that is closest to the strong root R of the original junction tree. Again, we use the Push operation to make the variables in question become members of C. If C, after performing the Push operations, has a potential with an empty tail, we can compute the desired strong marginal from the potential of C as in (10). Otherwise, we need to Push the variables in question closer to R until we eventually have all the variables contained in a clique having a potential with an empty tail; from the potential of this clique we can compute the desired potential as in (10). If necessary, we may need to Push the variables all the way to R.

The calculation of the strong marginal for a single continuous variable is an important special case, and from the above discussion it follows that such a marginal can be calculated with limited additional effort, since no potential of the junction tree will be extended with more than a single continuous variable as part of this calculation.

7 Incorporating evidence

At this point we assume that the initialization process has been completed so that the cliques of the junction tree contain complements and the separators contain weak marginals.

Discrete evidence is incorporated as usual, it does not matter where, and it is not necessary to insert discrete evidence in more than one clique.

To describe how to incorporate continuous evidence we first realize that every continuous node necessarily appears as head in exactly one clique, which is the clique where it appears closest to the strong root. In all other clique potentials where it appears, it must be a tail node.

Also, if U and W are neighbouring cliques with U closest to the root, the continuous variables in the separator $S = U \cap W$ constitute a superset of the tail of the potential (complement) that is stored in W.

It is most convenient to incorporate evidence about continuous nodes a single node at a time. Evidence that $Y_2 = y_2$ must be entered in all cliques where Y_2 appears. We assume that the clique where Y_2 appears as head has a potential with an empty tail. If this is not the case, we use the Push operation described above in Subsection 6.2 until we achieve this. We then proceed as follows:

1. In cliques where Y_2 is a tail node, the tail of the clique potential is decreased by Y_2, p and C are unchanged, and B is changed by removing the column B_2 corresponding to Y_2. A is modified to become $A^* = A + B_2 y_2$.

2. In the clique where Y_2 is a head node we partition the head nodes as under marginalization into $Y = (Y_1, Y_2)$. The potential after inserted evidence is denoted $\phi^* = [p^*, A^*, B^*, C^*](H^* \,|\, T^*)$. The head H^* is obtained from H by removing Y_2. The tail T^* (and thus B^*) is empty. We then distinguish two cases:

   (a) If there is a j with $C_{22}(j) = 0$ and $y_2 = A_2(j)$, we let for all i

   $$p^*(i) = \begin{cases} p(i) & \text{if } y_2 = A_2(i) \text{ and } C_{22}(i) = 0 \\ 0 & \text{otherwise} \end{cases}$$

   and for all i with $p^*(i) > 0$ we let

   $$A^*(i) = A_1(i), \quad C^*(i) = C_{11}(i).$$

   (b) Else we let

   $$p^*(i) = \begin{cases} \frac{p(i)}{\sqrt{2\pi C_{22}(i)}} e^{-\frac{1}{2}(y_2 - A_2(i))^2 / C_{22}(i)} & \text{if } C_{22}(i) \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

   and for all i with $p^*(i) > 0$ we let

   $$\begin{aligned} A^*(i) &= A_1(i) + C_{12}(i)(y_2 - A_2(i))/C_{22}(i) \\ C^*(i) &= C_{11}(i) - C_{12}(i) C_{21}(i)/C_{22}(i). \end{aligned}$$

Intuitively the operation reflects that any deterministic explanation of the evidence (with $C_{22}(i) = 0$) is infinitely more likely than a non-deterministic one, if it is available. The calculation for case (2a) is simply based upon the fact that

$$P(I = i \,|\, Y_2 = y_2) = \frac{P(I = i, Y = y_2)}{P(Y = y_2)} \qquad \text{if } P(Y = y_2) > 0$$

whereas a standard density calculation is appropriate in case (2b), where $P(Y = y_2) = 0$.

The correctness of the operation can be formally proved by a small calculation in (not so elementary) probability. For simplicity we only give this argument in the case where Y_1 is void, so that $Y_2 = Y$.

Let $q(i\,|\,y)$ denote the kernel obtained by normalizing p^* above, but where we have let the dependence on y be explicit, i.e.

$$q(i\,|\,y) = \frac{p^*(i)}{\sum_j p^*(j)}.$$

We need to show that for any interval D on the real line, q satisfies the relation

$$P(I = i, Y \in D) = \int_D q(i\,|\,y)\,\mu(dy),$$

where μ is the marginal distribution of Y, i.e.

$$\mu(D) = \sum_j p(j)\mu_j(D)$$

and μ_j denoting the normal distribution $\mathcal{N}\{A(j), C(j)\}$, degenerate at $A(j)$ if $C(j) = 0$.

For $C_{22}(i) = 0$, we have

$$\int_D q(i\,|\,y)\,\mu_j(dy) = \begin{cases} q\{i\,|\,A(j)\} & \text{if } A(j) \in D \text{ and } C(j) = 0 \\ 0 & \text{otherwise.} \end{cases}$$

Thus we get

$$\begin{aligned} \int_D q(i\,|\,y)\,\mu(dy) &= \begin{cases} \sum_{j:C(j)=0} p(j)q\{i\,|\,A(j)\} & \text{if } A(i) \in D \\ 0 & \text{otherwise} \end{cases} \\ &= \begin{cases} p(i) & \text{if } A(i) \in D \\ 0 & \text{otherwise} \end{cases} \\ &= P(I = i, Y \in D). \end{aligned}$$

If $C_{22}(i) > 0$ we similarly get

$$\begin{aligned} \int_D q(i\,|\,y)\,\mu(dy) &= \sum_{j:C(j)=0} p(j)q\{i\,|\,A(j)\} + \sum_{j:C(j)\neq 0} p(j)\int_D q(i\,|\,y)\,\mu_j(dy) \\ &= 0 + \sum_{j:C(j)\neq 0} p(j)\int_D q(i\,|\,y)\,\mu_j(dy) = P(I = i, Y \in D). \end{aligned}$$

When a piece of continuous evidence has been inserted, the representation is still a ‘complement’ representation, and the insertion of the next

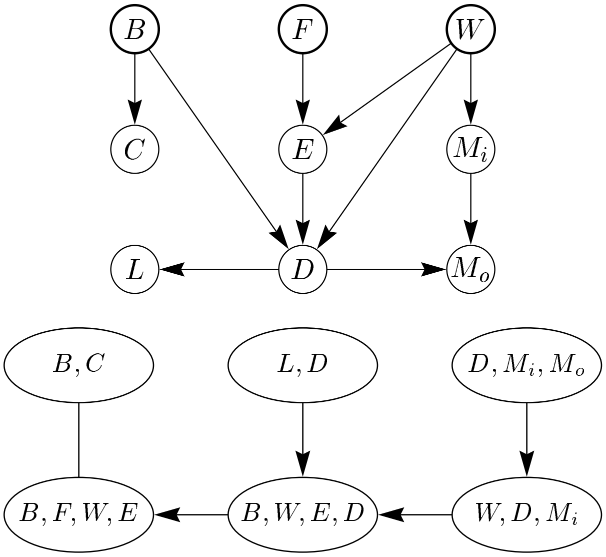


Figure 3: Bayesian network and strong junction tree for the WASTE incinerator example. The variables are W (type of waste), F (filter state), B (burning regimen), M_i (metals in waste), E (filter efficiency), C (CO_2 emission), D (emission of dust), M_o (emission of metals), and L (light penetrability). The variables W, F and B are discrete.

piece of evidence can take place. When all evidence has been inserted, we COLLECT towards the root as during initialization. This collection will only involve proper computations in the discrete part of the potentials. And the normalizer at the root clique will be equal to the joint density of the evidence.

Example 4 Our final example is the WASTE example described in Lauritzen (1992) and Cowell *et al.* (1999), Section 7.7, and we refer to either of these for the details of the numerical specifications. The example is concerned with the control of the emission of heavy metals from a waste incinerator:

> The emission from a waste incinerator differs 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 emission. The filter efficiency depends on the technical state of the electrofilter and the amount and composition of waste. The emission of heavy metals depends both on 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.

The essence of this description is represented in the Bayesian network of Figure 3, which also shows a junction tree. The strong root can be chosen either as $\{B, C\}$ or $\{B, F, W, E\}$. There is only one way to assign (the potentials corresponding to) the continuous variables to cliques of the junction tree: C is assigned to $\{B, C\}$, D to $\{B, W, E, D\}$, E to $\{B, F, W, E\}$, L to $\{L, D\}$, M_i to $\{W, D, M_i\}$, and M_o to $\{D, M_i, M_o\}$. So, there is exactly one potential involving continuous variables assigned to each clique, and the continuous components of these potentials become the corresponding continuous components of the clique potentials of the initialized strong junction tree. This is because the Collect operation — for this particular junction tree — does not change the continuous components of the clique potentials during the initialization process.

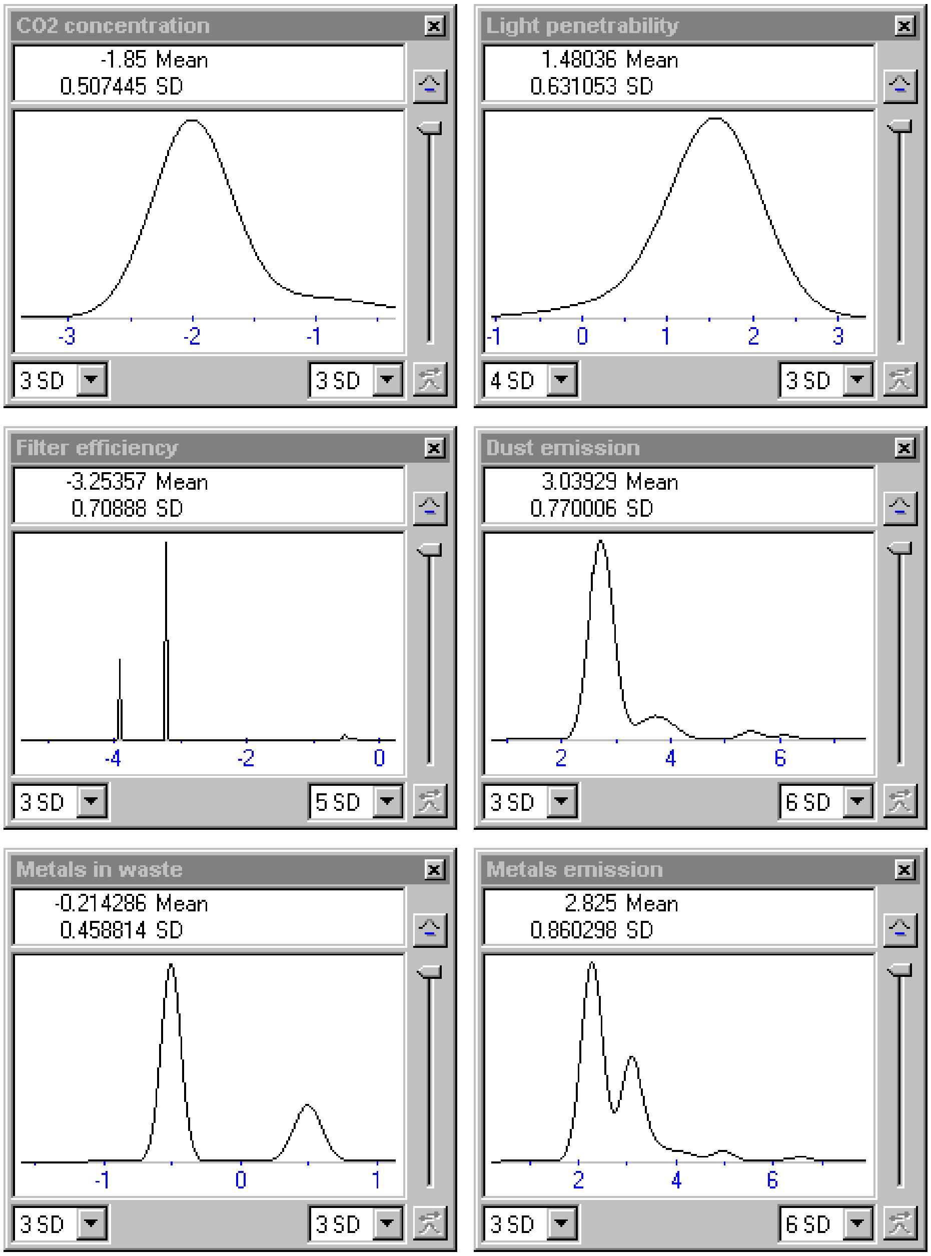
Incorporation of evidence on C or E can be done without invoking the Push operation, since these variables appear either as head in the root or in a clique with discrete separator towards the root. Incorporating evidence on D requires D to be Pushed to $\{B, F, W, E\}$ (unless evidence on E has already been incorporated). Similarly, incorporation of evidence on L will require Pushing L to $\{B, F, W, E\}$ unless some separator along the path from $\{L, D\}$ to $\{B, F, W, E\}$ has been made empty or fully discrete by incorporation of evidence on D and/or E.

Before incorporation of evidence on M_i and M_o the clique $\{W, D, M_i\}$ has a potential with head $\{M_i\}$ and an empty tail. Incorporating evidence on M_i at this point can therefore be done without invoking the Push operation. If, on the other hand, evidence on M_o (but not on D) has been incorporated, the potential on the clique $\{W, D, M_i\}$ will have head and tail $(\{M_i\} \mid \{D\})$ and incorporating evidence on M_i at this point will require Pushing M_i closer to $\{B, F, W, E\}$.

Incorporation of evidence on M_o requires Pushing M_o to $\{W, D, M_i\}$ unless evidence have been incorporated on both D and M_i.

Similar considerations apply to finding full mixture distributions for individual continuous variables.

Figures 4 and 5 display full mixture distributions of all the continuous variables before and after incorporation of the information that the waste has been of industrial type, L has been measured to 1.1, and C to -0.9. □



CO2 concentration

-1.85 Mean
0.507445 SD

3 SD | 3 SD

Light penetrability

1.48036 Mean
0.631053 SD

4 SD | 3 SD

Filter efficiency

-3.25357 Mean
0.70888 SD

3 SD | 5 SD

Dust emission

3.03929 Mean
0.770006 SD

3 SD | 6 SD

Metals in waste

-0.214286 Mean
0.458814 SD

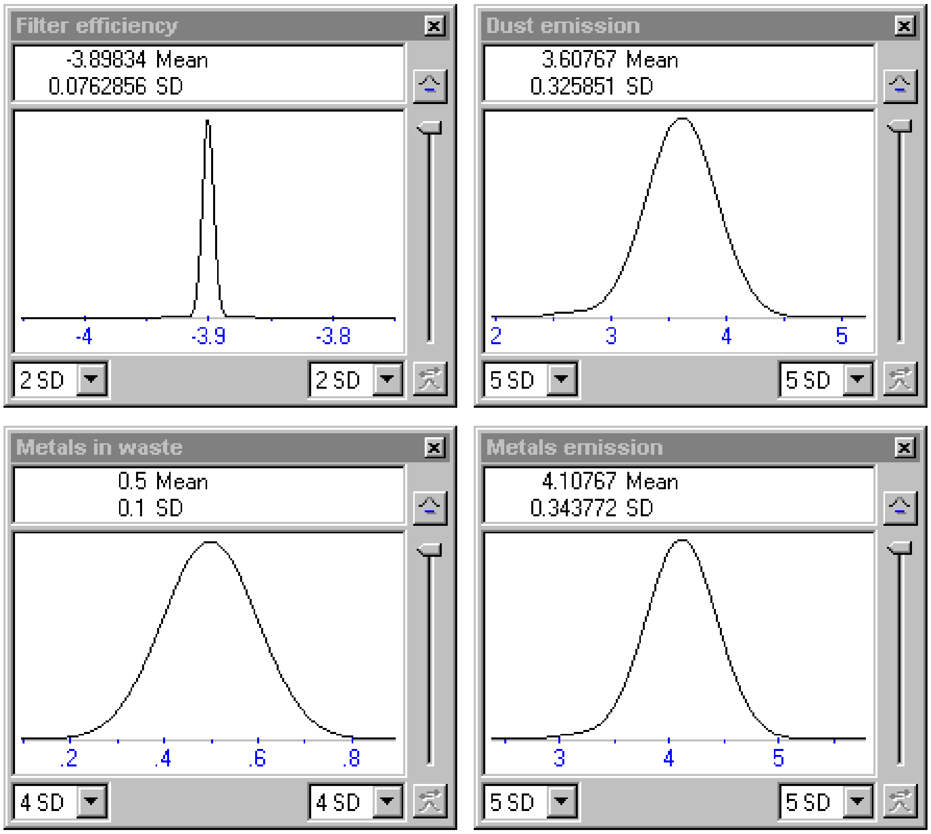
3 SD | 3 SD

Metals emission

2.825 Mean
0.860298 SD

3 SD | 6 SD

Figure 4: Screendumps from the HUGIN software displaying full marginals of all continuous variables from the WASTE incinerator example before any evidence has been incorporated.



Filter efficiency

-3.89834 Mean
0.0762856 SD

2 SD   2 SD

Dust emission

3.60767 Mean
0.325851 SD

5 SD   5 SD

Metals in waste

0.5 Mean
0.1 SD

4 SD   4 SD

Metals emission

4.10767 Mean
0.343772 SD

5 SD   5 SD

Figure 5: Screendumps from the HUGIN software displaying full marginals of the remaining continuous variables from the WASTE incinerator example after inserting the evidence that the waste has been of industrial type, L has been measured to 1.1, and C to -0.9.

Acknowledgements

The first author has benefited from conversations with Glenn Shafer concerning the development of an abstract theory of local computation. The research has been partly supported by the Danish Research Councils through the PIFT programme.

Anders L. Madsen provided helpful comments on a draft version of this paper, and Lars M. Nielsen prepared the HUGIN screendumps shown in Figures 4 and 5.

References

Cowell, R. G., Dawid, A. P., Lauritzen, S. L., and Spiegelhalter, D. J. (1999). *Probabilistic Networks and Expert Systems*. Springer-Verlag, New York.

Jensen, F. V. (1996). *An Introduction to Bayesian Networks*. University College London Press, London.

Jensen, F. V., Lauritzen, S. L., and Olesen, K. G. (1990). Bayesian updating in causal probabilistic networks by local computation. *Computational Statistics Quarterly*, **4**, 269–82.

Lauritzen, S. L. (1992). Propagation of probabilities, means and variances in mixed graphical association models. *Journal of the American Statistical Association*, **87**, 1098–108.

Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press, Oxford.

Lauritzen, S. L. and Jensen, F. V. (1997). Local computation with valuations from a commutative semigroup. *Annals of Mathematics and Artificial Intelligence*, **21**, 51–69.

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, Series B*, **50**, 157–224.

Lauritzen, S. L. and Wermuth, N. (1984). Mixed interaction models. Technical Report R 84-8, Institute for Electronic Systems, Aalborg University.

Lauritzen, S. L. and Wermuth, N. (1989). Graphical models for associations between variables, some of which are qualitative and some quantitative. *Annals of Statistics*, **17**, 31–57.

Madsen, A. L. and Jensen, F. V. (1998). Lazy propagation in junction trees. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, (ed. G. F. Cooper and S. Moral), pp. 362–9. Morgan Kaufmann, San Mateo, California.

Pearl, J. (1986). Fusion, propagation and structuring in belief networks. *Artificial Intelligence*, **29**, 241–88.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, California.

Penrose, R. (1955). A generalized inverse for matrices. *Proceedings of the Cambridge Philosophical Society*, **51**, 406–13.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, (2 edn). John Wiley and Sons, New York.

Rao, C. R. and Mitra, S. K. (1971). *Generalized Inverse of Matrices and Its Applications.* John Wiley and Sons, New York.

Shafer, G. (1991). An axiomatic study of computation in hypertrees. Technical Report WP–232, School of Business, University of Kansas.

Shafer, G. (1996). *Probabilistic Expert Systems.* Society for Industrial and Applied Mathematics, Philadelphia.

Shenoy, P. P. and Shafer, G. (1990). Axioms for probability and belief-function propagation. In *Uncertainty in Artificial Intelligence 4*, (ed. R. D. Shachter, T. S. Levitt, L. N. Kanal, and J. F. Lemmer), pp. 169–98. North-Holland, Amsterdam, The Netherlands.