Appeared in: Statistics and Computing, Vol. 16, No. 3, 2006, pp. 293–308.

Approximating Probability Density Functions in Hybrid Bayesian Networks with Mixtures of Truncated Exponentials*

Barry R. Cobb[1]
cobbbr@vmi.edu

Prakash P. Shenoy[2]
pshenoy@ku.edu

Rafael Rumí[3]
rrumi@ual.es

[1]Virginia Military Institute
Department of Economics and Business
Lexington, VA 24450

[2]University of Kansas School of Business
1300 Sunnyside Ave., Summerfield Hall
Lawrence, KS 66045–7585

[3]Universidad de Almería
Departamento de Estadística y Matemática Aplicada
Ctra. Sacramento s/n, La Cañada de San Urbano
04120 - Almería (Spain)

May 17, 2006

Abstract

Mixtures of truncated exponentials (MTE) potentials are an alternative to discretization and Monte Carlo methods for solving hybrid Bayesian networks. Any probability density function (PDF) can be approximated by an MTE potential, which can always be marginalized in closed form. This allows propagation to be done exactly using the Shenoy-Shafer architecture for computing marginals, with no restrictions on the construction of a join tree. This paper presents MTE potentials that approximate standard PDF's and applications of these potentials for solving inference problems in hybrid Bayesian networks. These approximations will extend the types of inference problems that can be modeled with Bayesian networks, as demonstrated using three examples.

Keywords: Graphs and networks, probabilistic computation, modeling methodologies, Bayesian networks.

*To appear in *Statistics and Computing* 16:293–308.

List of Figures

1 Introduction

Bayesian networks model knowledge about propositions in uncertain domains using graphical and numerical representations (Spiegelhalter *et al.* 1993). At the qualitative level, a Bayesian network is a directed acyclic graph where nodes represent variables and the (missing) edges represent conditional independence relations among the variables. At the numerical level, a Bayesian network consists of a factorization of a joint probability distribution into a set of conditional distributions, one for each variable in the network. Hybrid Bayesian networks contain both discrete probability mass functions (PMF's) and continuous conditional probability density functions (PDF's) as numerical inputs. This paper presents a method of modeling non-Gaussian standard PDF's in hybrid Bayesian networks and demonstrates that such a method can extend the applications to which Bayesian networks can be applied.

Poland (1994) proposes using a finite mixture of Gaussians to fit arbitrary continuous distributions for chance variables in hybrid Bayesian networks. One advantage of using mixtures of Gaussians is that marginals can be computed exactly using the technique of Lauritzen and Jensen (2001) because the network can be reduced to a Conditional Linear Gaussian (CLG) model (Lauritzen 1992, Cowell *et al.* 1999). CLG models are solved using operations from multivariate normal probability theory.

Logspline density estimation methods (Kooperberg and Stone 1991) divide sample data from an unknown density f into subsets, then estimate $\ell = log(f(x))$ by a function of the form $\hat{\ell}(x, \theta) = \mathbf{B}(x)\theta$. In this estimate, the basis functions, $\mathbf{B}(x)$, are cubic polynomials and θ is a suitably-chosen column-vector of constants. After a normalization step, the corresponding density estimate $\hat{f} = \exp\{\hat{l}\}$ is positive and integrates to one. Use of density estimators, such as logspline or kernel density estimates, in hybrid Bayesian networks presents difficulties because the resulting estimates are neither Gaussian, nor guaranteed to be integrable in closed form. These limitations prohibit the use of general purpose algorithms for calculating marginals.

An alternative to using mixtures of Gaussians or other density estimation methods for approximating continuous chance variables in hybrid Bayesian networks is mixtures of truncated exponentials (MTE) potentials (Moral *et al.* 2001, Rumí 2003). The class of MTE potentials is closed under combination and marginalization, and an MTE potential can always be integrated in closed form, allowing use of the Shenoy-Shafer architecture for calculating marginals. Previous work presents MTE approximations to the normal PDF (Cobb and Shenoy 2006) and demonstrates that MTE potentials can be used to solve augmented CLG models (Lerner *et al.* 2001), where discrete nodes have continuous parents with normal distributions.

In this paper, we describe MTE approximations for seven standard probability distributions and a method for approximating any standard PDF. These approximation methods allow the numerical specification of a hybrid Bayesian network to use parameters for standard PDF's which can then be modeled by MTE potentials in the solution phase. The remainder of this paper is organized as follows. Section 2 contains notation and definitions used throughout the paper. Section 3 describes a method of estimating parameters for MTE potentials. Section 4 presents MTE approximations to standard PDF's. Section 5 demonstrates inference in hybrid Bayesian networks using MTE potentials. Section 6 provides some discussion of the approach presented in the paper.

2 Notation and Definitions

2.1 Notation

Random variables in a hybrid Bayesian network will be denoted by capital letters, e.g. A, B, C. Sets of variables will be denoted by boldface capital letters, $\mathbf{Y}$ if all variables are discrete, $\mathbf{Z}$ if all variables are continuous, or $\mathbf{X}$ if some of the components are discrete and some are continuous. If $\mathbf{X}$ is a set of variables, $\mathbf{x}$ is a configuration of specific states of those variables. The discrete, continuous, or mixed state space of $\mathbf{X}$ is denoted by $\Omega_{\mathbf{X}}$.

MTE probability potentials and discrete probability potentials are denoted by lower-case greek letters, e.g. α, β, γ. Subscripts are used for fragments of MTE potentials or conditional probability tables when different parameters or values are required for each configuration of a variable's discrete parents, e.g. α_1, β_2, ${\gamma_3}^1$. Discrete probabilities for a specific element of the state space are denoted as an argument to a discrete potential, e.g. $\delta(0) = P(D = 0)$.

In graphical representations, continuous nodes in hybrid Bayesian networks are represented by double-border ovals, whereas discrete nodes are represented by single-border ovals. Continuous nodes that are deterministic functions of their parents are represented by triple-border ovals.

2.2 MTE Potentials

A mixture of truncated exponentials (MTE) potential has the following definition (Moral *et al.* 2001, Rumí 2003).

MTE potential. Let $\mathbf{X}$ be a mixed n-dimensional random variable. Let $\mathbf{Y} = (Y_1, \ldots, Y_d)$ and $\mathbf{Z} = (Z_1, \ldots, Z_c)$ be the discrete and continuous parts of $\mathbf{X}$, respectively, with $c + d = n$. A function $\phi : \Omega_{\mathbf{X}} \mapsto \mathbb{R}^+$ is an MTE potential if one of the next two conditions holds:

1. The potential ϕ can be written as

$$\phi(\mathbf{x}) = \phi(\mathbf{y}, \mathbf{z}) = a_0 + \sum_{i=1}^{m} a_i \exp\left\{ \sum_{j=1}^{d} b_i^{(j)} y_j + \sum_{k=1}^{c} b_i^{(d+k)} z_k \right\} \tag{1}$$

for all $\mathbf{X} \in \Omega_{\mathbf{X}}$, where $a_i, i = 0, \ldots, m$ and $b_i^{(j)}$, $i = 1, \ldots, m$, $j = 1, \ldots, n$ are real numbers.

2. There is a partition $\Omega_1, \ldots, \Omega_k$ of $\Omega_{\mathbf{X}}$ verifying that the domain of continuous variables, $\Omega_{\mathbf{Z}}$, is divided into hypercubes, the domain of the discrete variables, $\Omega_{\mathbf{Y}}$, is divided into arbitrary sets, and such that ϕ is defined as

$$\phi(\mathbf{x}) = \phi_i(\mathbf{x}) \qquad \text{if } \mathbf{x} \in \Omega_i, \tag{2}$$

where each $\phi_i, i = 1, ..., k$ can be written in the form of equation (1) (i.e. each ϕ_i is an MTE potential on Ω_i).

In the definition above, k is the number of *pieces* and m is the number of exponential *terms* in each piece of the MTE potential. In this paper, all MTE potentials are equal to zero in unspecified regions.

Estimating the parameters of MTE potentials is an open research problem. An iterative algorithm based on least squares approximation has been proposed to estimate MTE potentials from data (Moral *et al.* 2002). Moral *et al.* (2003) describes a method to approximate conditional MTE potentials using a mixed tree structure.

2.3 Propagation in MTE Networks

The operations of restriction, marginalization, and combination from (Moral *et al.* 2001) required for propagation with MTE potentials in hybrid Bayesian networks are included in this section for completeness.

2.3.1 Restriction

Restriction—or entering evidence—involves dropping coordinates to define a potential on a smaller set of variables. During propagation, restriction is performed by substituting values for known variables into the appropriate MTE potentials and simplifying the potentials accordingly.

Let ϕ be an MTE potential for $\mathbf{X} = \mathbf{Y} \cup \mathbf{Z}$. Assume a set of variables $\mathbf{X}' = \mathbf{Y}' \cup \mathbf{Z}' \subseteq \mathbf{X}$, whose values $\mathbf{x}^{\downarrow \Omega_{\mathbf{X}'}}$ are fixed to values $\mathbf{x}' = (\mathbf{y}', \mathbf{z}')$. The restriction of ϕ to the values $(\mathbf{y}', \mathbf{z}')$ is a new potential defined on $\Omega_{\mathbf{X} \setminus \mathbf{X}'}$ according to the following expression:

$$\phi^{R(\mathbf{X}'=\mathbf{x}')}(\mathbf{w}) = \phi^{R(\mathbf{Y}'=\mathbf{y}', \mathbf{Z}'=\mathbf{z}')}(\mathbf{w}) = \phi(\mathbf{x}) \tag{3}$$

for all $\mathbf{w} \in \Omega_{\mathbf{X} \setminus \mathbf{X}'}$ such that $\mathbf{x} \in \Omega_{\mathbf{X}}$, $\mathbf{x}^{\downarrow \Omega_{\mathbf{X} \setminus \mathbf{X}'}} = \mathbf{w}$ and $\mathbf{x}^{\downarrow \Omega_{\mathbf{X}'}} = \mathbf{x}'$. In this definition, each occurrence of $\mathbf{X}'$ in ϕ is replaced with $\mathbf{x}'$.

2.3.2 Combination

Combination of MTE potentials is pointwise multiplication. Let ϕ_1 and ϕ_2 be MTE potentials for $\mathbf{X}_1 = \mathbf{Y}_1 \cup \mathbf{Z}_1$ and $\mathbf{X}_2 = \mathbf{Y}_2 \cup \mathbf{Z}_2$. The combination of ϕ_1 and ϕ_2 (denoted by $\phi_1 \otimes \phi_2$) is a new MTE potential for $\mathbf{X} = \mathbf{X}_1 \cup \mathbf{X}_2$ defined as follows

$$\phi(\mathbf{x}) = \phi_1(\mathbf{x}^{\downarrow \mathbf{X}_1}) \cdot \phi_2(\mathbf{x}^{\downarrow \mathbf{X}_2}) \text{ for all } \mathbf{x} \in \Omega_{\mathbf{X}}. \tag{4}$$

Normalization is implicit in the definition of combination (in the sense that instead of normalizing every time combination is done, we omit it and normalize just once at the end of propagation).

2.3.3 Marginalization

Marginalization in a network with MTE potentials corresponds to summing over discrete variables and integrating over continuous variables. Let ϕ be an MTE potential for $\mathbf{X} = \mathbf{Y} \cup \mathbf{Z}$. The marginal of ϕ for a set of variables $\mathbf{X}' = \mathbf{Y}' \cup \mathbf{Z}' \subseteq \mathbf{X}$ is an MTE potential computed as

$$\phi^{\downarrow \mathbf{X}'}(\mathbf{y}', \mathbf{z}') = \sum_{\mathbf{y} \in \Omega_{\mathbf{Y} \setminus \mathbf{Y}'}} \left(\int_{\Omega_{\mathbf{Z}''}} \phi(\mathbf{y}, \mathbf{z}) \, d\mathbf{z}'' \right) \tag{5}$$

where $\mathbf{z} = (\mathbf{z}', \mathbf{z}'')$, and $(\mathbf{y}', \mathbf{z}') \in \Omega_{\mathbf{X}'}$. Although we show the continuous variables being marginalized before the discrete variables in (5), the variables can be marginalized in any sequence, resulting in the same final MTE potential.

2.4 Shenoy-Shafer Architecture

Moral *et al.* (2001) shows that the class of MTE potentials is closed under the operations of restriction, marginalization, and combination. Thus, MTE potentials can be propagated using the Shenoy-Shafer architecture (Shenoy and Shafer 1990), since only restrictions, marginalizations, and combinations are perfomed.

The Shenoy-Shafer architecture relies on three axioms—consonance of marginalization, commutativity and associativity of combination, and distributivity of marginalization over combination—that enable efficient local computation of marginals of the joint distribution of variables in a Bayesian network. To complete the algorithm, each node in the join tree sends a message to each of its neighbors that is the combination of its own potential and all incoming messages—except the message from the receiving node—followed by marginalization to the intersection with the receiving node. The combination of a node's own potential and all incoming messages is the posterior distribution of the variables in the node conditioned on the evidence. A binary join tree (Shenoy 1997) contains a node for each singleton subset of variables, so using the algorithm with a binary join tree results in marginals for all variables in the network.

3 Estimating Parameters for Mixtures of Truncated Exponentials (MTE) Potentials

This section describes a method for estimating the parameters of MTE potentials which approximate standard PDF's.

3.1 Kullback-Leibler (KL) Divergence

When approximating a standard PDF with an MTE potential, we measure the Kullback-Leibler (KL) divergence introduced by the approximation and minimize this measure in the process of finding parameters for the MTE potential, subject to certain constraints.

The relative entropy or Kullback-Leibler (KL) divergence (Kullback and Leibler 1951, MacKay 2003) between a standard PDF $f_X(x)$ and its MTE approximation $\tilde{f}_X(x)$ is defined as

$$D_{KL}(f_X(x) \mid\mid \tilde{f}_X(x)) = \int_S f_X(x) \log \frac{f_X(x)}{\tilde{f}_X(x)}\, dx. \tag{6}$$

Define $p_{f_{Xi}}$ and $q_{\tilde{f}_{Xi}}$ as the probability masses of $f_X(x)$ and $\tilde{f}_X(x)$, respectively, in the interval $(x_{i-1}, x_i]$. A discrete approximation to the KL divergence statistic over a set of points $x_i, i = 0, ..., n$ can be calculated as follows:

$$D'_{KL}(f_X(x) \mid\mid \tilde{f}_X(x)) = \sum_{i=1}^{n} p_{f_{Xi}} \log \frac{p_{f_{Xi}}}{q_{\tilde{f}_{Xi}}}. \tag{7}$$

The function $g(x) = \log\left(f_X(x)/\tilde{f}_X(x)\right)$ can be interpreted as the information contained in x for distinguishing between $f_X(x)$ and $\tilde{f}_X(x)$. Thus, KL divergence is the expectation of the information content over the domain S taken with respect to the distribution $f_X(x)$.

By minimizing this expectation when determining parameters for MTE approximations to standard PDF's—subject to probability mass constraints—we ensure a small chance of distinguishing between results obtained from inference with standard PDF's and corresponding MTE approximations.

3.2 Estimation Procedure

The numerical representation of a hybrid Bayesian network requires a conditional probability potential for each variable in the network, given its parents. We first consider the problem of estimating parameters for an MTE potential approximating a marginal PDF. This technique can be extended in a straightforward way to estimate the parameters for a conditional MTE potential by using the mixed tree structure in (Moral *et al.* 2003).

3.2.1 Partitioning the Domain

To estimate the parameters of an MTE potential for a continuous variable X, a partition $\Omega_1, \ldots, \Omega_k$ of Ω_X must be determined. Typically, in each interval of the partition, the PDF to be approximated should show no changes in concavity/convexity or increase/decrease. For example, the normal PDF with parameters μ and σ^2 can be divided into four such intervals: 1) increasing and convex on $(-\infty, \mu - \sigma)$; 2) increasing and concave on $[\mu - \sigma, \mu)$; 3) decreasing and concave on $[\mu, \mu + \sigma)$; and 4) decreasing and convex on $[\mu + \sigma, +\infty)$. To increase efficiency in the inference process, we may choose to exclude a small amount of probability density in the tails when approximating a PDF. PDF's whose basic shape does not change dramatically when the distribution parameters change, such as the normal PDF, can be fit by partitioning the domain with respect to changes in increase/decrease only.

Suppose the domain of the continuous variable has been divided into K intervals denoted $D_1, ..., D_K$. To estimate parameters for an MTE potential which approximates a standard PDF within a given interval D_k, we choose a set of points $x = (x_0, ..., x_n)$ by evenly dividing the portion of the domain of the PDF represented by D_k. A set of points $y = (y_0, ..., y_n)$ is determined by calculating the value of the PDF at each point x_i, $i = 0, ..., n$.

3.2.2 Approximation by Nonlinear Optimization

Defining an MTE approximation to a PDF $f_X(x; \mathbf{\Theta}_m)$ (abbreviated $f_X(x)$) requires estimating constants a_{0k}, a_{ik} and $b_{ik}^{(j)}$ in (1) for each interval D_k. We assume $\mathbf{\Theta}_m$ is an arbitrary vector of parameters of a standard PDF and the MTE approximation will be fitted for potential parameter vectors, $m = 1, ..., M$. The formulation in (1) allows an independent term (a_{0k}) and an unlimited number of exponential terms in each interval D_k; however, we will restrict the MTE potential to three exponential terms in each interval to increase efficiency during the inference process. Additionally, we assume one MTE potential will be defined for each configuration of a variable's discrete parents, so in each exponential term, parameters $b_{ik}^{(j)}$ are only defined for $j = d+1, ..., n$ in (1). Thus, the parameters to be estimated are a_{0k}, a_{1k}, a_{2k},a_{3k}, $b_{1k}^{(j)}$,$b_{2k}^{(j)}$ and $b_{3k}^{(j)}$.

Define $\hat{\phi}^{(k)}(x; \theta_{mk})$ (abbreviated $\hat{\phi}^{(k)}(x)$) as the initial MTE approximation for PDF $f_X(x)$ in interval D_k. To estimate the parameters $\theta_{mk} = \{a_{0mk}, a_{1mk}, a_{2mk}, a_{3mk}, b_{1mk}^{(j)}, b_{2mk}^{(j)}, b_{3mk}^{(j)}\}$ in

(1), the following general optimization problem is solved for each selected paremeter vector $\mathbf{\Theta}_m,\ m = 1, ...M$:

$$\begin{array}{cc} \underset{\theta_{mk}}{\text{argmin}} & \sum_{x_i \in D_k} D(f_X(x_i) \mid\mid \hat{\phi}^{(k)}(x_i)) \\ \text{subject to} & \text{Continuity Constraints} \\ & \text{Probability Mass Constraints} \\ & \text{Non-negativity Constraints.} \end{array}$$

In words, a discrete measure of divergence between the standard PDF and the MTE approximation is minimized subject to continuity, probability mass and non-negativity constraints.

To speed convergence of the optimization problem, we can implement the technique in Moral *et al.* (2002) to choose starting values for the parameters. Alternatively, parameters for an existing MTE approximation to a different standard PDF over similar intervals of concavity/convexity and/or increase/decrease can be used as starting values. The parameters obtained by the nonlinear optimization technique may be sensitive to starting values; thus, the parameters selected are not guaranteed to be global optima. However, multiple MTE approximations can be obtained for any PDF, such as those produced by obtaining different local optima for the same nonlinear optimization problem. For the distributions presented in this paper, each of these possible approximations corresponds to an extremely small KL divergence statistic.

To create the approximation to the gamma, beta and lognormal PDF's in Section 4, the discrete approximation to the KL divergence statistic is used as follows[2],

$$\begin{array}{cc} \underset{\theta_{mk}}{\text{argmin}} & \sum_{i=1}^{n} p_{f_{Xi}} \log \frac{p_{f_{Xi}}}{q_{\hat{\phi_{X_i}}^{(k)}}} \\ \text{subject} & f_X(x_0) = \hat{\phi}^{(k)}(x_0) \\ \text{to} & f_X(x_n) = \hat{\phi}^{(k)}(x_n) \\ & \int_{x_0}^{x_n} \left(f_X(x) - \hat{\phi}^{(k)}(x) \right) dx = 0 \\ & \hat{\phi}^{(k)}(x_i) \geq 0,\ i = 0, ..., n, \end{array}$$

where $p_{f_{Xi}}$ and $q_{\hat{\phi_{X_i}}^{(k)}}$ are the probability masses between x_i and x_{i-1} for $f_X(x)$ and $\hat{\phi}^{(k)}(x)$, respectively. The solution to the above optimization problem is defined as $\hat{\theta}_{mk}$. The first and second constraints ensure that the end points in adjacent regions of the MTE potential are equal. The first constraint can be relaxed in region D_1 and the second constraint can be relaxed in region D_K.

Next, each MTE parameter can be stated as a function of the standard PDF parameters, $\mathbf{\Theta}$, by solving the following optimization problem:

$$\underset{\mathbf{\Lambda}_{ki}}{\text{argmin}} \quad \sum_{m=1}^{M} \left(\hat{\theta}_{mk}^{(i)} - h(\mathbf{\Theta}_m; \mathbf{\Lambda}_{ki}) \right)^2 ,$$

where $\mathbf{\Lambda}_{ki}$ is a vector of parameters for the function $h(\mathbf{\Theta}_m; \mathbf{\Lambda}_{ki})$ required to state the ith parameter of the MTE approximation in interval D_k as a function of the standard PDF parameters. If we simply want to maintain a table of values for the MTE parameters for each parameter vector $\mathbf{\Theta}_m$, we can define $\mathbf{\Lambda}_{ki} = \{\emptyset\}$ and $h(\mathbf{\Theta}_m) = \hat{\theta}_{mk}^{(i)}$ for $m = 1, ..., M$.

The method proposed here requires the PDF to be approximated to have a known functional form. To approximate a PDF from data by an MTE potential, a density estimation method (such as the logspline method of Kooperberg and Stone (1991)) can first be applied to the data to determine a partition and provide a functional form, then the nonlinear optimization method can be applied to create an MTE approximation.

4 MTE Approximations to Standard PDF's

An MTE potential can be used to approximate any PDF. In this section, we present MTE approximations to seven standard PDF's.

4.1 Uniform PDF

The uniform PDF can be expressed as a trivial case of the MTE formulation in (1) where the potential is defined over one region and the constant $a_0 = \frac{1}{b-a}$, where a and b are the minimum and maximum values, respectively, of the uniform PDF. All other parameters $a_i,\ i = 1, \ldots, m$ in (1) equal zero.

4.2 Exponential PDF

Suppose we have a Poisson process with constant rate λ per unit of time. Let X denote the time between two consecutive events. The variable X has an exponential distribution with parameter λ, i.e. $f_X(x) = \lambda \exp\{-\lambda x\}$ for $x > 0$. This PDF can be expressed as a special case of the MTE formulation in (1) where the potential is defined over the region $(0, \infty)$, the constant $a_0 = 0$ and coefficient $a_1 = \lambda$ and $b_1^{(1)} = -\lambda$.

4.3 Normal PDF

MTE approximations to the normal PDF are presented in Cobb and Shenoy (2006). Consider a normally distributed variable X with mean μ and variance $\sigma^2 > 0$. The PDF for the normal distribution is

$$f_X(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2 \right\} .$$

The general formulation for a 2-piece, 3-term MTE potential which approximates the normal PDF is

$$\phi(x) = \begin{cases} \sigma^{-1}(-0.010564 + 197.055720 \exp\{2.2568434(\frac{x-\mu}{\sigma})\} \\ -461.439251 \exp\{2.3434117(\frac{x-\mu}{\sigma})\} + 264.793037 \exp\{2.4043270(\frac{x-\mu}{\sigma})\}) \\ \qquad\qquad \text{if } \mu - 3\sigma \leq x < \mu \\ \sigma^{-1}(-0.010564 + 197.055720 \exp\{-2.2568434(\frac{x-\mu}{\sigma})\} \\ -461.439251 \exp\{-2.3434117(\frac{x-\mu}{\sigma})\} + 264.793037 \exp\{-2.4043270(\frac{x-\mu}{\sigma})\}) \\ \qquad\qquad \text{if } \mu \leq x \leq \mu + 3\sigma \end{cases} . \tag{8}$$

In this formulation, the mean, μ, of X may be represented by a linear function of its continuous parents, as in the CLG model. Each element $\hat{\theta}_{mk}^{(i)}$ is stated as a function of the standard PDF parameters. For instance, $\mathbf{\Lambda}_{11} = \{-0.010564\}$ and $a_{01} = -0.010564/\sigma$.

The KL divergence of the normal PDF (with any μ and $\sigma^2 > 0$) and the MTE approximation to the normal PDF is 0.000346. Additional properties of this MTE approximation to the normal PDF are presented in Cobb and Shenoy (2006). Additionally, three examples are solved, including an example of a hybrid Bayesian network where a discrete node has a continuous parent.

4.4 Gamma PDF

4.4.1 Function Characteristics

Suppose we have a Poisson process with constant rate λ per unit of time. Let the random variable X denote the waiting time for r events. The variable X has the *gamma distribution with parameters r and λ* where

$$f_X(x) = \frac{\lambda^r}{\Gamma(r)} x^{r-1} \exp\{-\lambda x\} \ , \ x > 0 \ ,$$

for any $\lambda > 0$, where $\Gamma(r)$ is the *gamma function*, defined by

$$\Gamma(r) = \int_0^\infty t^{r-1} \exp\{-t\} \, dt \ ,$$

for any real number $r > 0$.

For $r > 1$, the gamma PDF has an absolute maximum where its first derivative equals zero. The first derivative of the gamma PDF is

$$f_X'(x) = \frac{\exp\{-\lambda x\} x^{r-2} \lambda^r [(r-1) - x\lambda]}{\Gamma(r)} \ . \tag{9}$$

The absolute maximum, m, is defined where (9) equals zero, or

$$m = (r-1)/\lambda \ . \tag{10}$$

For $r > 2$, the gamma PDF has inflection points (changes in concavity) where its second derivative equals zero. The second derivative of the gamma PDF is

$$f''_X(x) = \frac{\exp\{-\lambda x\} x^{r-3} \lambda^r [(r-2)(r-1) - 2(r-1)x\lambda + x^2\lambda^2]}{\Gamma(r)} \ . \tag{11}$$

The inflection points are defined where (11) equals zero, or

$$x = \frac{(r-1)}{\lambda} \pm \frac{\sqrt{(r-1)}}{\lambda} \ . \tag{12}$$

Define $d = \sqrt{(r-1)}/\lambda$ so that the inflection points are defined as $x = m \pm d$. The gamma PDF has two inflection points and one critical point (which is always a maximum) when $r \geq 3$. When $1 < r < 3$, the gamma PDF is a concave down function to the left of the critical point. When $r = 1$, the gamma pdf is a special case of the exponential PDF and is a monotonically decreasing, concave up function. For $0 < r \leq 1$, we approximate the gamma PDF with the exponential PDF.

4.4.2 MTE Approximation

A 4-piece MTE approximation to the gamma PDF is defined as

$$\phi(x) =$$

$$\begin{cases} \lambda(a_{01} + a_{11}\exp\{b_{11}\lambda(x-m)\} + a_{21}\exp\{b_{21}\lambda(x-m)\}) & \\ \qquad \text{if } ((m - 1.414d \leq x < m - d) \cap (3 \leq r < 5)) & \\ \qquad \cup ((m - 2d \leq x < m - d) \cap (5 \leq r \leq 10)) & \\ \qquad \cup ((m - 2.5d \leq x < m - d) \cap (r > 10)) & \\ \lambda(a_{02} + a_{12}\exp\{b_{12}\lambda(x-m)\} + a_{22}\exp\{b_{22}\lambda(x-m)\}) & \\ \qquad \text{if } \text{Max}[0, m-d] \leq x < m & \\ \lambda(a_{03} + a_{13}\exp\{b_{13}\lambda(x-m)\} + a_{23}\exp\{b_{23}\lambda(x-m)\}) & \\ \qquad \text{if } m \leq x < m + d & \\ \lambda(a_{04} + a_{14}\exp\{b_{14}\lambda(x-m)\} + a_{24}\exp\{b_{24}\lambda(x-m)\}) & \\ \qquad \text{if } (m + d \leq x < m + 6d) \cap (1 < r \leq 10) & \\ \qquad \cup (m + d \leq x < m + 4d) \cap (r > 10) & \\ \lambda\exp\{-\lambda x\} \qquad \text{if } 0 < r \leq 1 \ . & \end{cases} \tag{13}$$

We set $h(\mathbf{\Theta}_r m) = \hat{\theta}^{(i)}_{mk}$ and define a table of constants[3] for values of $r = 1.5, 2.0, 2.5, 3.0, \ldots, 99.5, 100.0$. For $r > 100$ we approximate the gamma PDF with the MTE approximation to the normal PDF in (8). The MTE approximations to the gamma PDF with parameters $r = 6, 8$ and 11 and $\lambda = 1$ are displayed graphically in Figure 1 with the parameters and KL divergence statistics listed in Table 1.

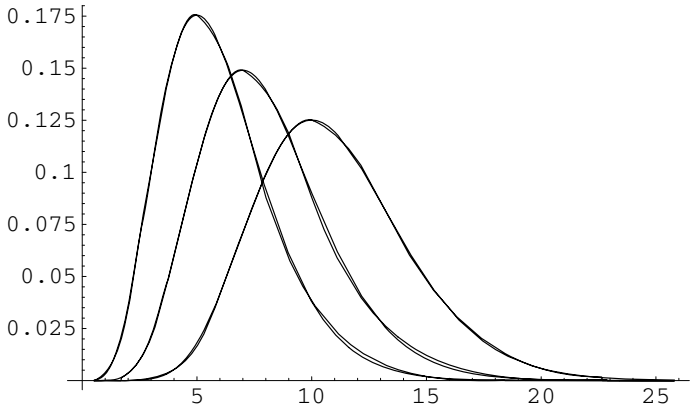

Figure 1: The MTE approximations to gamma PDF's with parameters $r = 6, 8$ and 11 and $\lambda = 1$ overlayed on the graph of the gamma PDF's.

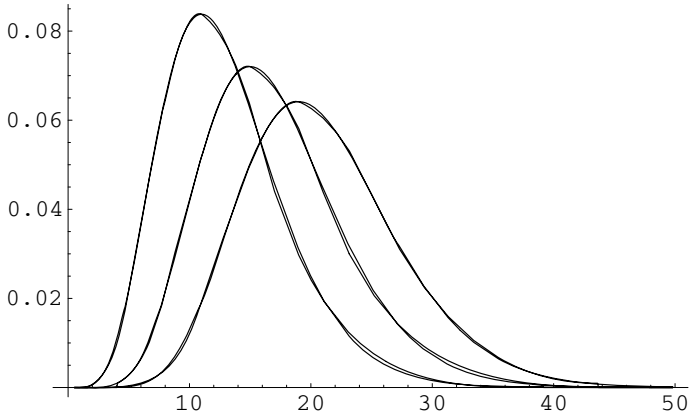

Figure 2: The MTE approximations to chi–square PDF's with parameters $n = 13, 17$ and 21 overlayed on the graph of the chi–square PDF's.

4.5 Chi-Square PDF

The chi-square distribution with n degrees of freedom is a special case of the gamma PDF where $r = n/2$ and $\lambda = 0.5$. Thus, the chi-square distribution can be approximated by the MTE approximation to the gamma PDF in (13). Figure 2 shows MTE approximations to chi–square distributions with $n = 13, 17$ and 21.

Consider the chi–square distribution with 21 degrees of freedom. The critical values in a chi–square table for the upper percentile values of this distribution are 32.671 for $p = 0.95$, 35.479 for $p = 0.975$ and 38.932 for $p = 0.99$. Integrating the MTE potential for $n = 21$ in Figure 2 over the intervals from $[0, 32.671]$, $[0, 35.479]$ and $[0, 38.932]$ gives probabilities of 0.9750, 0.9489 and 0.9903, respectively.

4.6 Beta PDF

4.6.1 Function Characteristics

A distribution of a random proportion, such as the proportion of defective items in a shipment, can be represented by the beta PDF. The beta PDF for a random variable X which

Table 1: Parameters and KL divergence statistics for MTE approximations to the gamma PDF with $\lambda = 1$.

	r		
$\hat{\theta_{mk}}$	**6**	**8**	**11**
a_{01}	-0.002704	-0.003287	-0.000668
a_{11}	33.449113	55.955116	55.952009
b_{11}	2.058104	1.571871	1.324635
a_{21}	-92.962972	-70.339661	70.342622
b_{21}	2.650266	1.694115	1.422907
a_{02}	-2.284046	-2.375417	-2.451516
a_{12}	2.712855	2.702941	2.694400
b_{12}	0.045703	0.028784	0.017576
a_{22}	-0.253342	-0.178522	-0.117775
b_{22}	0.512581	0.454392	0.420215
a_{03}	-1.336838	-1.368414	-1.394603
a_{13}	1.704404	1.695418	1.687299
b_{13}	0.023885	0.018704	0.014568
a_{23}	-0.192098	-0.178001	-0.167586
b_{23}	0.257974	0.213672	0.172972
a_{04}	-5.991587	-5.987559	-6.011745
a_{14}	0.297779	0.262770	0.288893
b_{14}	-0.350223	-0.320715	-0.169005
a_{24}	5.971677	5.972495	5.907099
b_{24}	0.000216	0.000147	0.000962
D_{KL}	0.002095	0.000856	0.000283

represents a random proportion is

$$f_X(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)} \ , \ 0 < x < 1 \ ,$$

where $B(\alpha,\beta)$ is the *beta function*, defined by

$$B(\alpha,\beta) = \int_0^1 \ t^{\alpha-1}(1-t)^{\beta-1} \ dt \ ,$$

for any real numbers $\alpha > 0$ and $\beta > 0$. Thus, $\alpha > 0$ and $\beta > 0$ are the shape parameters of the beta PDF.

For most parameters α and β, the beta PDF has a critical point (either an absolute maximum or minimum) where its first derivative equals zero. The first derivative of the beta PDF is

$$f_X^{'}(x) = \frac{(1-x)^{\beta-2}x^{\alpha-2}[(1-x)(\alpha-1)-x(\beta-1)]}{B(\alpha \ , \ \beta)}. \tag{14}$$

The critical point, m, is defined where (14) equals zero, or

$$m = (1 - \alpha)/(2 - \alpha - \beta). \tag{15}$$

The beta PDF is monotonic (thus the critical point does not exist) in the following cases:

- $(\alpha < 1) \cap (\beta > 1)$
- $(\beta < 1) \cap (\alpha > 1)$

For some parameters α and β, the beta PDF has inflection points (changes in concavity), $d^{\pm}$, where its second derivative equals zero, or

$$d^{\pm} = \frac{(\alpha-1)(\alpha+\beta-3) \pm \sqrt{(\beta-1)(\alpha-1)(\alpha+\beta-3)}}{(\alpha+\beta-3)(\alpha+\beta-2)}. \tag{16}$$

The distribution has just one inflection point in the following cases:

- $(\alpha < 1) \cap (1 < \beta < 2)$
- $(\beta < 1) \cap (1 < \alpha < 2)$
- $(1 < \alpha < 2) \cap (\beta > 2)$
- $(1 < \beta < 2) \cap (\alpha > 2)$

None of the inflection points exist in the following cases:

- $(\beta \leq 1) \cap (\alpha \leq 1)$
- $(\alpha \leq 1) \cap (\beta \geq 2)$
- $(\beta \leq 1) \cap (\alpha \geq 2)$
- $(1 \leq \alpha \leq 2) \cap (1 \leq \beta \leq 2)$

The behavior of the beta PDF is summarized in Figure 3.

4.6.2 MTE Approximation

For each of the regions defined by the parameters in Figure 3, we could define an MTE approximation, but the symmetry of the beta PDF allows us to reduce the parameter space. If $\pounds(X) \sim Beta(\alpha, \beta)$, then

$$f_X(x) = f_Y(1 - x), \tag{17}$$

where $\pounds(Y) \sim Beta(\beta, \alpha)$.

This property allows us to define an MTE approximation for parameters (α, β) fulfilling the property $\alpha \geq \beta$. The MTE approximation will have a different number of pieces, depending on the critical point and the existence of inflection points, as follows:

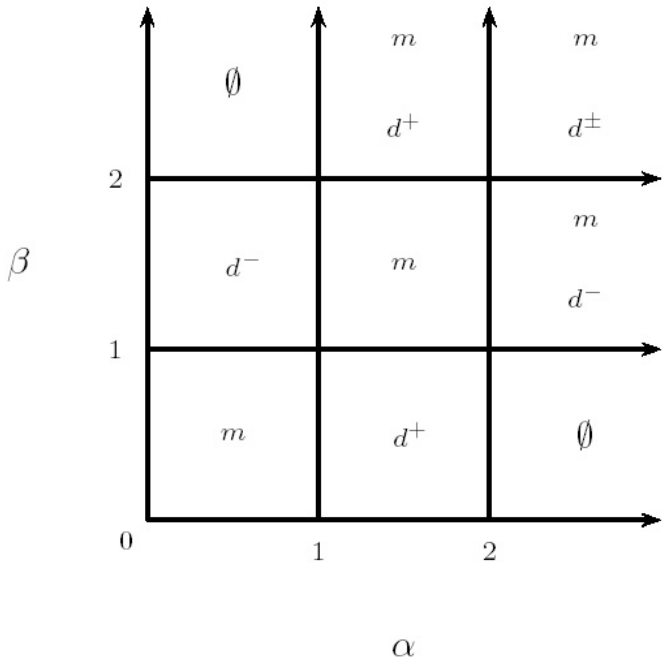

Figure 3: Critical and inflection points for the different parameters of the beta distribution.

$$\phi(x) =$$

$$\begin{cases} a_{01} + a_{11}\exp\{b_{11}x\} + a_{21}\exp\{b_{21}x\} \\ \quad \text{if } ((0 < x < m) \cap (\alpha \leq 1) \cap (\beta \leq \alpha)) \cup ((0 < x < m) \cap ((1 < \alpha \leq 2) \cap (1 < \beta \leq \alpha))) \\ \quad \cup ((0 < x < d+) \cap ((1 < \alpha < 2) \cap (\beta < 1))) \cup ((0 < x < d-) \cap ((\alpha > 2) \cap (1 < \beta \leq \alpha))) \\ \quad \cup ((0 < x < 1) \cap ((\alpha \geq 2)(\beta \leq 1))) \cup ((0 < x < 1) \cap ((\beta = 1) \cap (1 \leq \alpha \leq 2))) \\ a_{02} + a_{12}\exp\{b_{12}x\} + a_{22}\exp\{b_{22}x\} \\ \quad \text{if } ((0 < x < m) \cap (\alpha \leq 1) \cap (\beta \leq \alpha)) \cup ((0 < x < m) \cap ((1 < \alpha \leq 2) \cap (1 < \beta \leq \alpha))) \\ \quad \cup ((d- \leq x < m) \cap ((\alpha > 2) \cap (1 < \beta \leq \alpha))) \cup ((d+ \leq x < 1) \cap ((1 < \alpha < 2) \cap (\beta < 1))) \\ a_{03} + a_{13}\exp\{b_{13}x\} + a_{23}\exp\{b_{23}x\} \\ \quad \text{if } ((m \leq x < 1) \cap ((\alpha > 2) \cap (1 < \beta \leq 2))) \cup ((m \leq x < d+) \cap ((\alpha > 2) \cap (2 < \beta \leq \alpha)) \\ a_{04} + a_{14}\exp\{b_{14}x\} + a_{24}\exp\{b_{24}x\} \\ \quad \text{if } (d+ \leq x < 1) \cap ((\alpha > 2) \cap (2 < \beta \leq \alpha)) \, . \end{cases} \tag{18}$$

The MTE approximations to the beta PDF with parameters $(\alpha, \beta) = (2, 2), (2.7, 1.3)$ and $(1.3, 2.7)$ are displayed graphically in Figure 4 with the parameters and KL divergence statistics listed in Table 2. The MTE parameters for $Beta(1.3, 2.7)$ are obtained from $Beta(2.7, 1.3)$ as shown in (17).

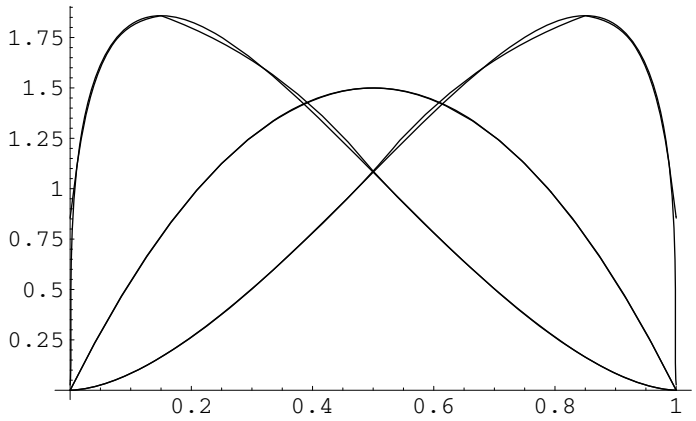

Figure 4: The MTE approximations to beta PDF's with parameters $(\alpha, \beta) = (2, 2), (2.7, 1.3)$ and $(1.3, 2.7)$ overlayed on the graph of the beta PDF's.

Table 2: Parameters and KL divergence statistics for the MTE approximations to the beta PDF.

	(α, β)		
$\hat{\theta}_{mk}$	$(2, 2)$	$(2.7, 1.3)$	$(1.3, 2.7)$
a_{01}	-1299.228439	-5.951668	1.823067
a_{11}	-545.789594	5.573315	-1.029580
b_{11}	-0.177215	0.461387	-26.000040
a_{21}	1845.018033	0.378353	0.060778
b_{21}	-0.049140	-6.459391	-0.529991
a_{02}	-1299.228439	0.473653	0.473653
a_{12}	-457.153000	-6.358482	-0.453988
b_{12}	0.177215	-2.639473	2.639473
a_{22}	1756.540000	2.729394	1.959340
b_{22}	0.049140	-0.331471	0.331471
a_{03}	—	1.823067	-5.951668
a_{13}	—	$-5.26E{-}12$	8.840810
b_{13}	—	26.000040	-0.461387
a_{23}	—	0.035774	0.000592
b_{23}	—	0.529991	6.459391
D_{KL}	$2.62118E{-}6$	0.000330	0.000330

4.7 Lognormal PDF

4.7.1 Function Characteristics

A random variable X is lognormal, i.e. $\pounds(X) \sim LN(\mu, \sigma^2)$, if and only if $\pounds(\ln X) \sim N(\mu, \sigma^2)$. A lognormal random variable has the PDF

$$f_X(x) = \frac{1}{x\sqrt{2\pi\sigma^2}} \exp \frac{-(\ln x - \mu)^2}{2\sigma^2} \ , \ x > 0 \ ,$$

for any $\sigma^2 > 0$.

The lognormal PDF has an absolute maximum where its first derivative equals zero. The first derivative of the lognormal PDF is

$$f_X^{'}(x) = -\frac{\exp \frac{-(\ln x - \mu)^2}{2\sigma^2}(\ln x - \mu)}{x\sigma^2\sqrt{2\pi\sigma^2}} \ . \tag{19}$$

The absolute maximum, m, is defined where (19) equals zero, or $m = \exp\{\mu - \sigma^2\}$.

The inflection points, $d^{\pm}$, are defined where the second derivative of the lognormal PDF equals zero, or

$$d^{\pm} = \exp\left\{\frac{1}{2}(2\mu - 3\sigma^2 \pm \sigma\sqrt{4 + \sigma^2})\right\}. \tag{20}$$

4.7.2 MTE Approximation

To define upper and lower bounds for the MTE approximation to the lognormal PDF, we use the normal PDF as a benchmark and construct a potential containing the same probability mass in the lognormal PDF as contained in the normal PDF over the interval $[\mu - 3\sigma, \mu + 3\sigma]$. This probability mass—which equals 0.9973—is contained in the interval $[\exp\{\mu - 3\sigma\}, \exp\{\mu + 3\sigma\}]$ of the lognormal PDF.

A 4-piece, 2-term MTE approximation to the lognormal PDF is defined as

$$\phi(x) =$$

$$\begin{cases} a_{01} + a_{11} \exp\{b_{11}(x - m)\} + a_{21} \exp\{b_{21}(x - m)\} & \text{if } \exp\{\mu - 3\sigma\} \le x < d^- \\ a_{02} + a_{12} \exp\{b_{12}(x - m)\} + a_{22} \exp\{b_{22}(x - m)\} & \text{if } d^- \le x < m \\ a_{03} + a_{13} \exp\{b_{13}(x - m)\} + a_{23} \exp\{b_{23}(x - m)\} & \text{if } m \le x < d^+ \\ a_{04} + a_{14} \exp\{b_{14}(x - m)\} + a_{24} \exp\{b_{24}(x - m)\} & \text{if } d^+ \le x \le \exp\{\mu + 3\sigma\} \ . \end{cases} \tag{21}$$

The MTE approximations to the lognormal PDF with parameters $\mu = 0$ and $\sigma^2 = 0.25, 0.50$ and 1 are displayed graphically in Figure 5 with the parameters and KL divergence statistics listed in Table 3.

5 Applications

This section presents three applications of MTE potentials to inference problems in hybrid Bayesian networks.

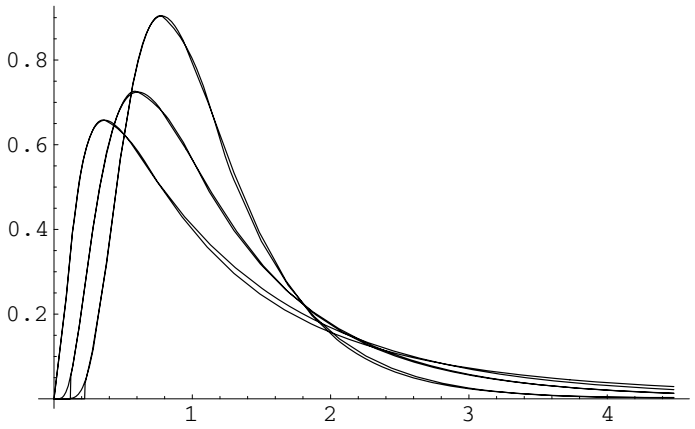

Figure 5: The MTE approximations to lognormal PDF's with parameters $\mu = 0$ and $\sigma^2 = 0.25, 0.5$ and 1.0 overlayed on the graph of the lognormal PDF's.

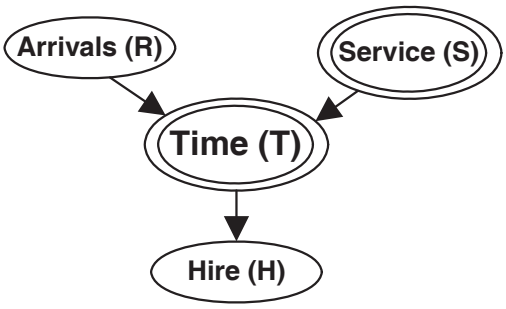

Figure 6: Hybrid Bayesian network for the Bank example.

5.1 Bank Example

A small town has 50 residents and one bank. The number of daily arrivals to the bank (R) follows a Poisson distribution with rate $\lambda = 0.24$, i.e. $\pounds(R) \sim Poisson$ (12). Let R denote a variable for number of arrivals to the bank with the following probability potential

$$\rho(r) = p_R(r) = \frac{e^{-0.24(50)}(0.24(50))^r}{r!}, \quad 0 \leq r \leq 20.$$

We assume the maximum arrivals in one day is 20, so the probability potential is truncated.

The service rate of customers (S) is normally distributed with a mean of 3.0 per hour and a standard deviation of 0.25, i.e. $\pounds(S) \sim N(3.0, 0.0625)$. The time to serve all customers arriving in one day (T) has a gamma distribution that is conditional on random variables R and S. The bank manager has established a "soft threshold" of five hours of total daily service time before hiring an additional teller. Thus, H is a binary, discrete random variable representing whether or not the bank manager hires an additional teller. The hybrid Bayesian network for the Bank example is depicted in Figure 6.

The potential φ for S is the MTE approximation to the normal PDF in (8) with $\mu = 3$ and $\sigma^2 = 0.0625$. Variable H is a discrete node with a continuous parent and is modeled with a binary sigmoid function. This sigmoid function is approximated using a general MTE formulation stated in terms of its two parameters w and g (Cobb and Shenoy 2006). Parameter w determines the steepness of the "soft threshold" and g is the offset of the threshold

Table 3: Parameters and KL divergence statistics for MTE approximations to the lognormal PDF with $\mu = 0$.

	σ^2		
$\hat{\theta_{mk}}$	**0.25**	**0.5**	**1.0**
a_{01}	-0.303932	-0.440613	-1.130046
a_{11}	3.036290	2.739898	2.666958
b_{11}	4.387105	4.247658	3.856545
a_{21}	0.177082	0.309889	0.710654
b_{21}	1.588571	1.573502	1.530143
a_{02}	5.716321	6.282564	5.900258
a_{12}	-3.364090	-3.134446	-0.290081
b_{12}	-0.880178	-0.978911	-4.780600
a_{22}	-1.448109	-2.423684	-4.952431
b_{22}	2.108769	1.332677	0.298929
a_{03}	-0.267558	-0.446017	-0.307309
a_{13}	-0.778641	-0.602649	-0.797981
b_{13}	1.652415	1.415704	1.400388
a_{23}	1.950321	1.773100	1.763035
b_{23}	0.528256	0.383786	0.541464
a_{04}	-0.678707	-0.698647	-0.724208
a_{14}	1.254983	0.886631	0.715287
b_{14}	-1.624085	-1.177871	-0.906293
a_{24}	0.657170	0.705372	0.729869
b_{24}	0.008725	-0.001153	-0.000395
D_{KL}	0.000330	0.000099	0.006467

from zero. In this example, we assume $w = -1$ and $g = 5$, so the MTE approximation to the binary sigmoid function representing the potential fragment for $\{H = 1, T\}$ in this example is

$$\psi_1(t) = P(H = 1 \mid T = t) = \begin{cases} 0 & \text{if } t < 0 \\ -0.021704 + 0.021804 \exp\{0.635t\} & \text{if } 0 \leq t < 5 \\ 1.021704 - 12.4827 \exp\{-0.635t\} & \text{if } 5 \leq t \leq 10 \\ 1 & \text{if } t > 10. \end{cases}$$

Since the variable is binary, $\psi_0(t) = P(H = 0 \mid T = t) = 1 - P(H = 1 \mid T = t)$. The MTE potential fragments ψ_0 and ψ_1 constitute the MTE potential ψ for $\{H, T\}$. The MTE potential fragment for $\{H = 1, T\}$ is shown graphically in Figure 7.

The time until all customers have arrived (T) depends on the service time in customers per hour (S) and the number of arrivals (R). Thus, $\pounds(T \mid S = s, R = r) \sim \Gamma(r, s)$, which is represented by the MTE potential ϑ using the formulation in (13). Solving the problem of calculating marginal distributions for each variable in the network will require the estimate of the parameter vectors $\hat{\theta}_{mk} = \hat{\theta}_{rk}$ for $r = 2, ..., 20$, where $\hat{\theta}_{2k}$ represents parameters needed to approximate the gamma PDF with $r = 2$, etc.

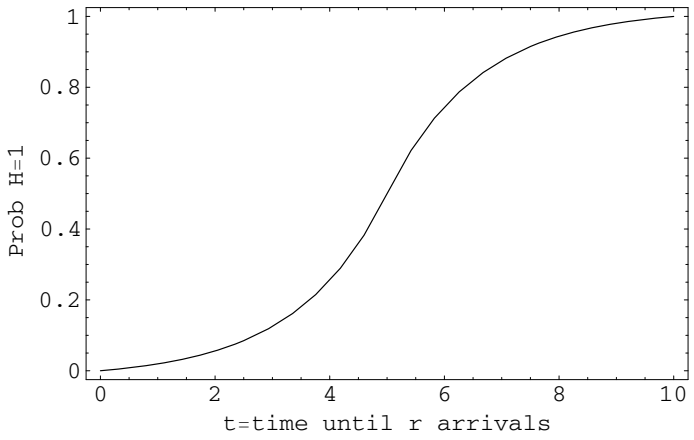

Figure 7: The MTE approximation to the sigmoid function representing $P(H = 1 \mid T = t)$ in the Bank network.

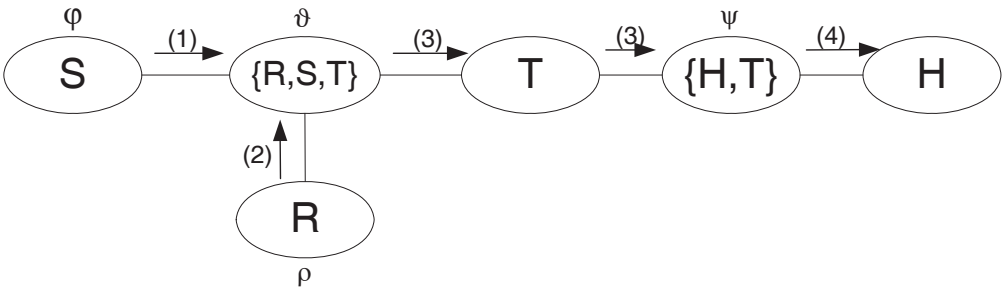

Figure 8: The binary join tree for the Bank example.

5.1.1 Join Tree Initialization

When the conditional probability distributions in hybrid Bayesian networks are approximated by MTE potentials, we can use any join tree for propagation with no restrictions placed on the initialization phase. This is in contrast to the architecture of Lauritzen and Jensen (2001), which requires a strong junction tree so that continuous variables are marginalized before discrete ones since the algorithm uses properties of Gaussian distributions to achieve marginalization. In the MTE approach, there are no constraints on the order in which variables are marginalized. Avoiding the use of a strong junction tree improves the efficiency of the solution phase because strong junction trees often contain larger cliques.

Other algorithms developed for inference in hybrid Bayesian networks with discrete children of continuous parents place special restrictions on the process of initializing the network. For instance, Lerner *et al.* (2001) requires a preprocessing phase where all potentials except those for the discrete children of continuous parents are inserted. The algorithm suggested by Murphy (1999) requires any logistic or softmax functions to be converted to Gaussian potentials by using a variational lower bound.

A binary join tree for the Bank example is shown in Figure 8.

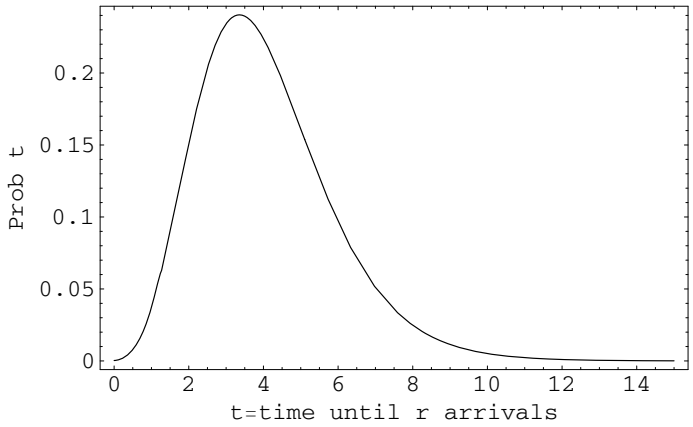

Figure 9: The prior marginal distribution for T in the Bank example.

5.1.2 Computing Messages

The following messages are required to compute the marginal distributions for T and H in the Bank example:

(1) φ from $\{S\}$ to $\{R, S, T\}$
(2) ρ from $\{R\}$ to $\{R, S, T\}$
(3) $(\varphi \otimes \rho \otimes \vartheta)^{\downarrow T}$ from $\{R, S, T\}$ to $\{T\}$
(4) $((\varphi \otimes \rho \otimes \vartheta)^{\downarrow T} \otimes \psi)^{\downarrow H}$ from $\{H, T\}$ to $\{H\}$.

5.1.3 Prior Marginals

1. `Prior Marginal for` T

   The message sent from $\{R, S, T\}$ to $\{T\}$ is the marginal distribution for T and is calculated as follows

   $$\tau(t) = \int_{\Omega_S} \left(\varphi(s) \left(\sum_{r=0}^{20} \rho(r) \cdot \vartheta(s,t) \right) \right) \, ds.$$

   The expected value and variance of the marginal distribution for T are computed as 4.0782 and 3.2859, respectively. The prior marginal distribution for T is shown graphically in Figure 9.

2. `Prior Marginal for` H
   To calculate the prior marginal probabilities for H, the marginal distribution for T is combined with the conditional MTE potential fragments ψ_0 and ψ_1. The joint potential for $\{H = 1, T\}$ is calculated as $\varrho_1(t) = \psi_1(t) \cdot \tau(t)$ and the joint potential for $\{H = 0, T\}$ is calculated as $\varrho_0(t) = \psi_0(t) \cdot \tau(t)$.

   The marginal probabilities for H are found by removing T from ϱ_0 and ϱ_1 by integration. The marginal probability of the bank manager hiring an additional teller is $P(H = 1) = 33.6\%$.

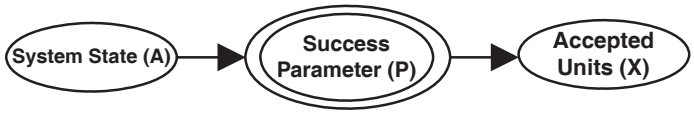

Figure 10: The hybrid Bayesian network for the Quality Control example.

5.2 Quality Control Example

In a quality control process, a random sample of output is taken and evaluated on whether or not each unit meets a pre–determined standard. Suppose the prior distribution for the success parameter P of the binomial distribution (where $0 < p < 1$) characterizing the sample output has a beta distribution which depends on the state of the system (A) with $\Omega_A = \{0 = poor,\ 1 = average,\ 2 = good\}$. A discrete random variable X represents the number of successes in 5 trials, i.e. $\pounds(X) \sim Binomial\ (5, P)$. The Bayesian network for this example is shown in Figure 10.

Assume the following discrete distribution for A:

$$\begin{aligned}\varphi(0) &= P(A = 0) = 0.05,\\ \varphi(1) &= P(A = 1) = 0.15,\\ \varphi(2) &= P(A = 2) = 0.80.\end{aligned}$$

The potential fragment for $\{P, A = 2\}$ is an MTE approximation to the beta PDF with parameters $\alpha = 1.3$ and $\beta = 2.7$:

$$\rho_2(p, A = 2) = f_{P|A=2}(p) =$$

$$\begin{cases} -5.951669 + 5.573316\ \exp\{0.461388p\} - 0.378353\ \exp\{-6.459391p\} \\ \qquad\qquad \text{if } 0 < p < 0.492929 \\ 0.473654 - 6.358483\ \exp\{-2.639474p\} + 2.729395\ \exp\{-0.331472p\} \\ \qquad\qquad \text{if } 0.492929 \leq p < 0.85 \\ 1.823067 - (5.26\text{E} - 12)\ \exp\{26.000041p\} + 0.035775\ \exp\{0.529991p\} \\ \qquad\qquad \text{if } 0.85 \leq p < 1\ . \end{cases}$$

If the system is in state $A = 0$, P has a beta distribution with $\beta = 1.3$ and $\alpha = 2.7$. Due to the symmetry of the beta PDF, the potential fragment for $\{P, A = 0\}$ is approximated as

$$\rho_0(p, A = 0) = f_{P|A=0}(p) = \rho_2(1 - p, A = 2).$$

The potential fragment for $\{P, A = 1\}$ is an MTE approximation to the beta PDF with parameters $\alpha = 2$ and $\beta = 2$. The numerical details of this potential are omitted. The potential fragments constituting the potential ρ for $\{P, A\}$ are shown graphically in Figure 4 in Section 4.

The potential for $\{X, P\}$ is

$$\psi(x, p) = P(X = k \mid P = p) = \left(\begin{array}{c} 5 \\ k \end{array} \right) p^k (1 - p)^{5-k}.$$

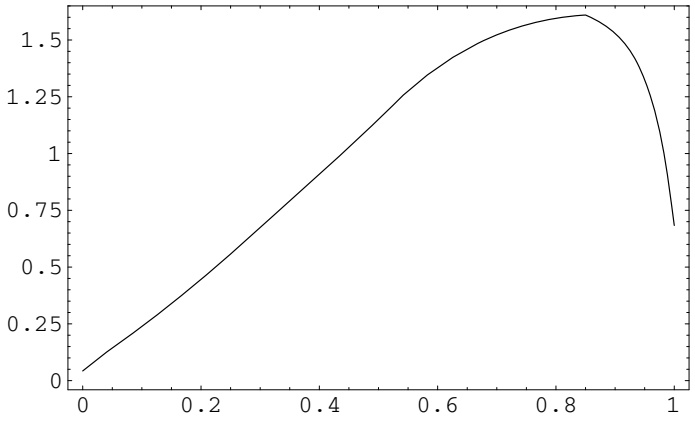

Figure 11: The marginal distribution for (P) in the Quality Control example.

The marginal distribution for P obtained after propagation is shown graphically in Figure 11. The expected value and variance of this distribution are 0.6308 and 0.0536, respectively. Marginal probabilities for X are as follows: $P(X = 0) = 0.0563$; $P(X = 1) = 0.1081$; $P(X = 2) = 0.1604$; $P(X = 3) = 0.2068$; $P(X = 4) = 0.2374$; and $P(X = 5) = 0.2310$.

Suppose a sample of output from the system is taken and only one unit meets the quality standard. The potential $\psi'(p) = \psi(1, p)$ is calculated and new potential fragments for $\{P, A\}$ are determined as

$$\begin{aligned}
\pi_0(p, A = 0) &= \psi'(p) \cdot \rho_0(p, A = 0) \\
\pi_1(p, A = 1) &= \psi'(p) \cdot \rho_1(p, A = 1) \\
\pi_2(p, A = 2) &= \psi'(p) \cdot \rho_2(p, A = 2).
\end{aligned}$$

The revised probabilities for A given the evidence are determined by integrating these potentials over P and combining them with the prior probabilities for A as follows:

$$\gamma(0) = P(A = 0) = K^{-1} \cdot \varphi(0) \cdot \int_0^1 \pi_0(p, A = 0)\, dp = 0.1193$$
$$\gamma(1) = P(A = 1) = K^{-1} \cdot \varphi(1) \cdot \int_0^1 \pi_1(p, A = 1)\, dp = 0.2477$$
$$\gamma(2) = P(A = 2) = K^{-1} \cdot \varphi(2) \cdot \int_0^1 \pi_2(p, A = 2)\, dp = 0.6330.$$

The normalization constant, K, is the prior probability of the observed evidence, or $K = P(X = 1) = 0.1081$.

The revised marginal distribution for P is shown in Figure 12 and has an expected value and variance of 0.3735 and 0.0260, respectively.

5.3 Extended Crop Example

A diagram of the hybrid Bayesian network for this example appears in Figure 13. In this model, the price (P) of a crop is assumed to decrease with the amount of crop (C) produced. Prices (P) will also be higher if the government subsidizes prices $(S = 1)$. The consumer

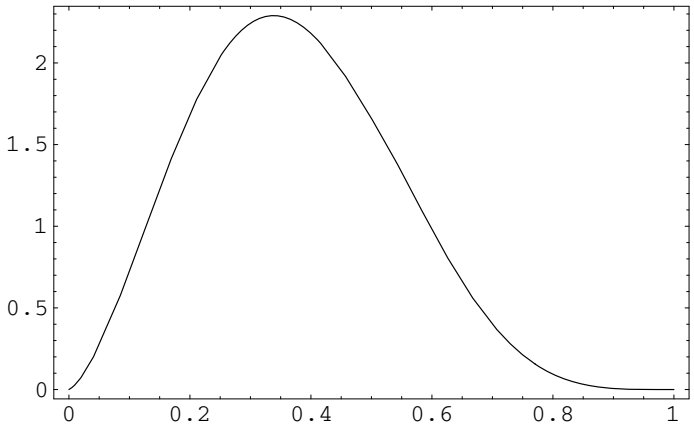

Figure 12: The revised marginal distribution for (P) incorporating the evidence $X = 1$.

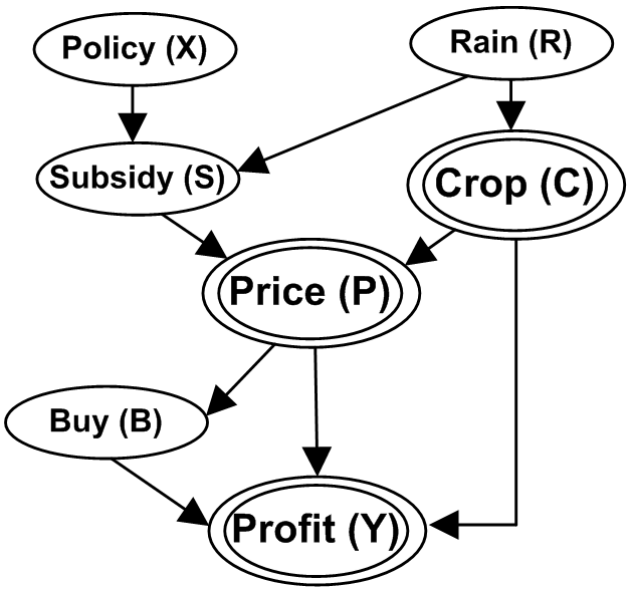

Figure 13: The hybrid Bayesian network for the Extended Crop example.

is likely to buy $(B = 1)$ if the price drops below a certain amount. Both the crop (C) produced and the subsidy (S) depend on whether the rain conditions are drought $(R = 0)$, average $(R = 1)$, or flood $(R = 2)$. Additionally, the subsidy (S) is affected by the type of government policy (X) employed. Profits (Y) are determined as a function of price (P), crop (C) produced, and the decision of the consumer on whether to buy (B) at the market price. This example uses a lognormal PDF to model the conditional distribution for price (P) given crop (C), replacing the normal PDF used in a similar example used by Lerner (2002).

The parameters of the distributions for the variables in the Extended Crop example are shown in Table 4. The normal and lognormal PDF's specified in Table 4 are approximated by normalized MTE potentials over the region containing a total probability density of 0.9973 in each distribution. For example, the MTE potential for P given $\{S = 0, 1.25 \leq c < 4.625\}$ is defined as

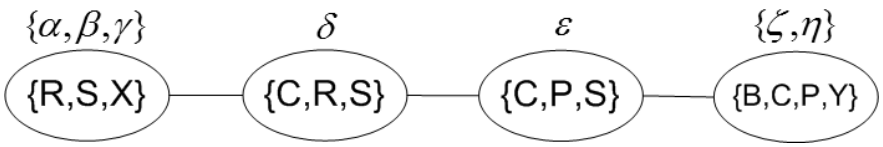

Figure 14: The join tree for the Extended Crop example.

$$\varepsilon(p, S = 0, 1.25 \leq c < 4.625) = f_{P|\{S=0,1.25\leq c<4.625\}}(p) =$$

$$\begin{cases} -0.019835 + 0.008137 \exp\{0.380496p\} - 0.000013 \exp\{1.008900p\} \\ \qquad \text{if } 2.718282 \leq p < 5.000869 \\ 0.196410 - 0.302185 \exp\{-0.131877p\} - 0.001304 \exp\{0.348873p\} \\ \qquad \text{if } 5.000869 \leq p < 9.487736 \\ 0.076381 - (6.967804\text{E}-06) \exp\{0.605724p\} + (1.548315\text{E}-11) \exp\{1.461563p\} \\ \qquad \text{if } 9.487736 \leq p < 14.018645 \\ -0.591615 + 0.365701 \exp\{-0.133531p\} + 0.589548 \exp\{0.000064p\} \\ \qquad \text{if } 14.018645 \leq p \leq 54.598150 \ . \end{cases}$$

The minimum and maximum endpoints in the domains of the MTE potentials for C given R and P given $\{C, S\}$ are used to divide the domain of the parent variables for the MTE potential for Y given $\{B, C, P\}$ into hypercubes, as in the mixed tree approach introduced by Moral *at al.* (2003); for instance, the MTE potentials for $\{C, R\}$ are defined over the range $[c_0, c_3]$, with $c_1 = c_0 + (c_3 - c_0)/3$ and $c_2 = c_0 + 2 \cdot (c_3 - c_0)/3$. The distribution for Buy (B) given Price (P) is an MTE approximation to the binary sigmoid function with parameters $w = -1$ and $g = 30$ (Cobb and Shenoy 2006).

The marginal distribution for Profit (Y) is calculated by passing messages in the join tree shown in Figure 14. The message from $\{R, S, X\}$ to $\{C, R, S\}$ is $\theta = (\alpha \otimes \beta \otimes \gamma)^{-X}$ for $\{R, S\}$. The message from $\{C, R, S\}$ to $\{C, P, S\}$ is $\vartheta = (\theta \otimes \delta)^{-R}$ for $\{C, S\}$. The potential fragment $\vartheta(c, S = 1)$ is shown graphically in Figure 15.

The discrete variable S is removed by summation after the potential ϑ for $\{C, S\}$ is combined with the potential ε for $\{C, P, S\}$ as follows:

$$\kappa(c, p) = \vartheta(c, S = 0) \cdot \varepsilon(c, p, S = 0) + \vartheta(c, S = 1) \cdot \varepsilon(c, p, S = 1) \ .$$

The potential κ for $\{C, P\}$ is the message from $\{C, P, S\}$ to $\{B, C, P, Y\}$. At $\{B, C, P, Y\}$, κ is combined with the potentials ζ for $\{B, P\}$ and η for $\{B, C, P, Y\}$. To find the marginal distribution for Profit (Y), variables C and P are removed by integration. The potential fragment for $\{Y, B = 1\}$ is calculated as

$$\nu(y, B = 1) = \int_{\Omega_P} \left(\zeta(p, B = 1) \cdot \int_{\Omega_C} (\kappa(c, p) \cdot \eta(c, p, y, B = 1)) \ dc \right) \ dp \ .$$

The potential fragment for $\{Y, B = 0\}$ is calculated similarly, then the marginal distribution for Y is calculated as

Table 4: Parameters for the distributions of the variables in the Extended Crop example.

	Variable	Distribution given parent state(s) or region		
α	Policy (X)			$(0.5, 0.5)$
β	Rain (R)			$(0.35, 0.6, 0.05)$
γ	Subsidy (S)	$R = 0$,	$S = 0$	$(0.4, 0.6)$
		$R = 0$,	$S = 1$	$(0.3, 0.7)$
		$R = 1$,	$S = 0$	$(0.95, 0.05)$
		$R = 1$,	$S = 1$	$(0.95, 0.05)$
		$R = 2$,	$S = 0$	$(0.5, 0.5)$
		$R = 2$,	$S = 1$	$(0.2, 0.8)$
δ	Crop (C)	$R = 0$		$N(3, 0.5)$
		$R = 1$		$N(5, 1)$
		$R = 2$		$N(2, 0.25)$
ε	Price (P)	$S = 0$,	$1.25 \leq c < 4.625$	$LN(2.5, 0.25)$
		$S = 0$,	$4.625 \leq c \leq 8$	$LN(2.0, 0.25)$
		$S = 1$,	$1.25 \leq c < 4.625$	$LN(3.0, 0.5)$
		$S = 1$,	$4.625 \leq c \leq 8$	$LN(2.75, 0.5)$
ζ	Buy (B)	p		$w = -1, g = 30$
η	Profit (Y)	$B = 0$,	$c_n \leq c \leq c_{n+1}$	$N(-0.5 \cdot \frac{c_n + c_{n+1}}{2} - 1, 100)$ for $n = 0, 1, 2$
		$B = 1$,	$(c_n \leq c \leq c_{n+1} \cap p_m \leq p \leq p_{m+1})$	$N((\frac{p_m + p_{m+1}}{2} - 0.5) \cdot (\frac{c_n + c_{n+1}}{2}) - 1, 100)$ for $m = 0, 1, 2$ and $n = 0, 1, 2$

$$\phi(y) = \nu(y, B = 0) + \nu(y, B = 1) \ .$$

The marginal distribution ϕ for Y is shown graphically in Figure 16. The tri-modal shape of the distribution occurs because crop produced varies under the three rain scenarios.

Suppose evidence exists that $P = 32$. The marginal distribution for Y can be updated as follows. The message $\{C, P, S\}$ to $\{B, C, P, Y\}$ remains the potential κ for $\{C, P\}$. At $\{B, C, P, Y\}$, the evidence restricts the combination of the potentials κ, ζ, and η as follows:

$$\begin{aligned} \lambda(c, y, B = 0) &= \kappa(c, 32) \cdot \zeta(32, B = 0) \cdot \eta(c, 32, y, B = 0) \ , \\ \lambda(c, y, B = 1) &= \kappa(c, 32) \cdot \zeta(32, B = 1) \cdot \eta(c, 32, y, B = 1) \ . \end{aligned}$$

The marginal distribution for Y considering the evidence is calculated as

$$\rho(y) = \int_{\Omega_C} (\lambda(c, y, B = 0) + \lambda(c, y, B = 1)) \ dc \ .$$

After a normalization step, the marginal distribution for Y appears graphically as shown in Figure 17.

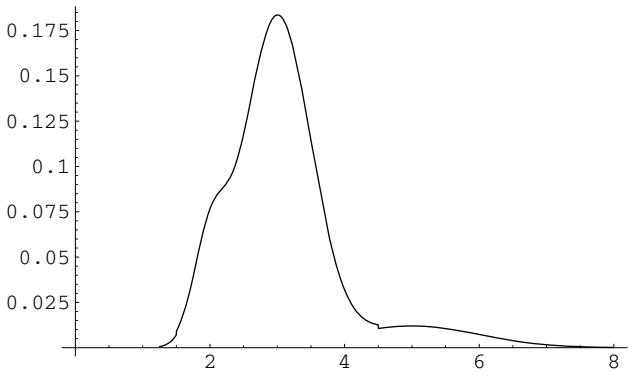

Figure 15: The potential fragment $\vartheta(c, S = 1)$ sent in the message from $\{C, R, S\}$ to $\{C, P, S\}$.

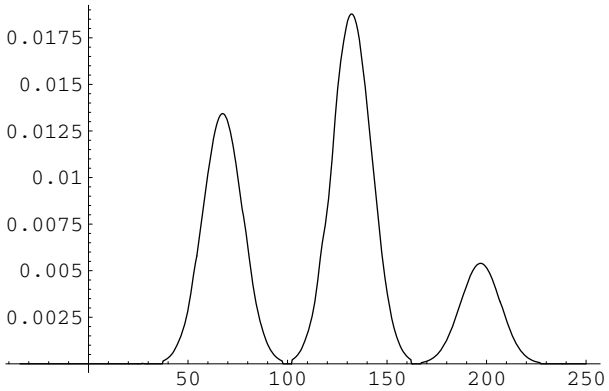

Figure 16: The marginal distribution of Profit (Y) in the Extended Crop example.

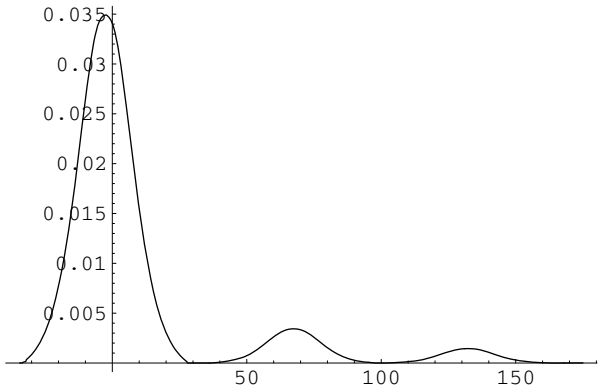

Figure 17: The marginal distribution of Profit (Y) considering the evidence $P = 32$.

6 Discussion

We have described a method of estimating the required parameters for MTE potentials which approximate standard PDF's and have presented MTE approximations to seven standard PDF's. This technique can also be extended to estimate any arbitrary PDF from data. Three examples of inference in hybrid Bayesian networks which use these MTE approximations were presented.

The approach presented in this paper allows hybrid Bayesian networks to be constructed in cases where variables are known to follow a standard probability distribution. The parameters specified for the standard PDF can be used to construct an MTE approximation to use in the solution phase. For instance, a stock price at a given time will have a lognormal distribution if it follows a Geometric Brownian Motion stochastic process. The parameters can be obtained from an expert, or via Maximum Likelihood estimates. CLG models provide the only current method of specifying such a model; however, CLG models only allow normal PDF's and impose restrictions on construction of the network. In principle, the MTE approach could apply to the problem of inference in hierarchical Bayesian statistical models, but actual application needs to be verified in practice.

The proposed technique for approximating continuous PDF's can be used for any multivariate probability model regardless of the topology and size of the Bayesian network if we can approximate each conditional PDF by an MTE potential. In the case where a continuous variable, say X has several continuous parents, say $\{Y, Z\}$, then we may have to fit a MTE approximation for the conditional density $f_{X|y,z}$, which may be a three dimensional surface. The MTE approach can be considered an "exact" method since the MTE approximations are very close to the exact densities, and there are no errors introduced during the inference process. Like all exact methods, it is not always tractable for models in which the clique sizes are large. A non-commercial implementation of Bayesian networks which uses MTE potentials is available at: `http://leo.ugr.es/~elvira` (Elvira Consortium 2002).

Other approaches for approximate inference in hybrid Bayesian networks include discretization of continuous variables and Markov Chain Monte Carlo (MCMC) methods. Compared to discretization, the MTE approach gives us a much better approximation, assuming we use the same number of pieces in the MTE approach as the number of bins in the discretization approach. Consequently, for a fixed specified accuracy, MTE would require far less time than discretization since we could use fewer pieces in the MTE approximation compared to the large number of bins that would be required by the specified accuracy. Rumí and Salmeron (2005) describe propagation methods for MTE potentials that calculate approximate messages, thereby reducing the solution time. Compared to discrete approximations, they find that using MTE potentials results in a favorable tradeoff in space/accuracy (with accuracy measured through comparisons of the marginal distributions after propagation).

The Monte Carlo methods on the other hand are always tractable, but the quality (variance) of the approximate inference depends on the sample size. In fact, computing the quality of the approximation itself can be difficult, as can detecting a steady state solution. Convergence of MCMC methods can be problematic for networks which have zero probabilities in the joint state space. Software implementations are available which use discretization and Monte Carlo methods to solve Bayesian networks. BUGS (`http://www.mrc-bsu.cam.ac.uk/bugs`) is a package which uses the Markov Chain Monte Carlo (MCMC) approach, whereas AgenaRisk (`http://www.agena.co.uk`) employs the dynamic discretization scheme of Kozlov

and Koller (1997).

Notes

[0]Barry R. Cobb, Virginia Military Institute, Department of Economics and Business
[0]Prakash P. Shenoy, University of Kansas, School of Business
[0]Rafael Rumí, Universidad de Almería, Departamento de Estadística Y Matemática Aplicada

[1]The term *potential* was introduced by Lauritzen and Spiegelhalter (1988) to describe conditional probability tables since the values in a conditional probability table do not sum to one, but to the sum of the number of states of parent variables. The term *fragment* was introduced by Demirer and Shenoy (2005) to describe a portion of a potential that is defined over a subset of the domain of the parent variables.

[2]For approximations presented in this paper, we use Microsoft Excel Solver, which implements the Generalized Reduced Gradient (GRG2) nonlinear optimization method (Lasdon and Waren 1978).

[3]A table containing these parameters is available in (Cobb 2005).

References

[1] Cobb, B.R. 2005. Inference and decision making in hybrid probabilistic graphical models, Doctoral dissertation, University of Kansas, School of Business, Lawrence, KS.

[2] Cobb, B.R. and Shenoy, P.P. 2006. Inference in hybrid Bayesian networks with mixtures of truncated exponentials, International Journal of Approximate Reasoning, 41(3):257–286.

[3] Cowell, R.G., Dawid, A.P., Lauritzen, S.L. and Spiegelhalter, D.J. 1999. Probabilistic Networks and Expert Systems, New York, Springer.

[4] Demirer, R. and Shenoy, P.P. 2005. Sequential valuation networks and asymmetric decision problems, European Journal of Operational Research, 169(1):286–309.

[5] Elvira Consortium. 2002. Elvira: An environment for probabilistic graphical models, in J.A. Gámez and A. Salmerón (eds.), Proceedings of the First European Workshop on Probabilistic Graphical Models, Cuenca, Spain, pp. 222–230.

[6] Kooperberg, C. and Stone, C.J. 1991. A study of logspline density estimation, Computational Statistics and Data Analysis, 12:327–347.

[7] A.V. Kozlov and D. Koller, D. 1997. Nonuniform dynamic discretization in hybrid networks, in D. Geiger and P.P. Shenoy (eds.), Uncertainty in Artificial Intelligence, 13:314–325, San Francisco, Morgan Kaufmann.

[8] Kullback, S. and Leibler, R.A. 1951. On information and sufficiency, Annals of Mathematical Statistics, 22:79–86.

[9] Lasdon, L.S. and Waren, A.D. 1978. Generalized reduced gradient software for linearly and nonlinearly constrained problems, in H.J. Greenberg, (ed.), Design and Implemetation of Optimization Software, Amsterdam, Sijthoff and Noordhoff, pp. 335–362.

[10] Lauritzen, S.L. 1992. Propagation of probabilities, means and variances in mixed graphical association models, Journal of the American Statistical Association 87:1098–1108.

[11] Lauritzen, S.L. and Jensen, F. 2001. Stable local computation with conditional Gaussian distributions, Statistics and Computing, 11:191–203.

[12] Lauritzen, S.L. and Spiegelhalter, D.J. 1988. Local computations with propabilities on graphical structures and their application to expert systems, Journal of the Royal Statistical Society Series B, 50(2):157–224.

[13] Lerner, U. 2002. Hybrid Bayesian networks for reasoning about complex systems, Doctoral dissertation, Stanford University, Stanford, CA.

[14] Lerner, U., Segal, E. and Koller, D. 2001. Exact inference in networks with discrete children of continuous parents, in J. Breese and D. Koller (eds.), Uncertainty in Artificial Intelligence, 17:319–328, San Francisco, Morgan Kaufmann.

[15] MacKay, D.J.C. 2003. Information Theory, Inference, and Learning Algorithms, Cambridge, United Kingdom, Cambridge University Press.

[16] Moral, S., Rumí, R. and Salmerón, A. 2001. Mixtures of truncated exponentials in hybrid Bayesian networks, in P. Besnard and S. Benferhart (eds.), Symbolic and Quantitative Approaches to Reasoning under Uncertainty, Lecture Notes in Artificial Intelligence, 2143:156–167, Berlin, Springer-Verlag.

[17] Moral, S., Rumí, R. and Salmerón, A. 2002. Estimating mixtures of truncated exponentials from data, in J.A. Gámez and A. Salmerón (eds.), Proceedings of the First European Workshop on Probabilistic Graphical Models, Cuenca, Spain, pp. 135–143.

[18] Moral, S., Rumí, R. and Salmerón, A. 2003. Approximating conditional MTE distributions by means of mixed trees, in T.D. Nielsen and N.L. Zhang (eds.), Symbolic and Quantitative Approaches to Reasoning under Uncertainty, Lecture Notes in Artificial Intelligence, 2711:173–183, Berlin, Springer-Verlag.

[19] Murphy, K. 1999. A variational approximation for Bayesian networks with discrete and continuous latent variables, in K.B. Laskey and H. Prade (eds.), Uncertainty in Artificial Intelligence, 15:467–475, San Francisco, Morgan Kaufmann.

[20] Poland, W.B. 1994. Mixtures of Gaussians and minimum relative entropy techniques for modeling continuous distributions, Ph.D. Thesis, Department of Engineering–Economic Systems, Stanford University, Stanford, CA.

[21] Rumí, R. 2003. Modelos De Redes Bayesianas Con Variables Discretas Y Continuas, Doctoral Thesis, Universidad de Almeriá, Departamento de Estadística y Matemática Aplicada, Almería, Spain.

[22] Rumí, R. and Salmerón, A. 2005. Penniless propagation with mixtures of truncated exponentials, in L. Godo (ed.), Symbolic and Quantitative Approaches to Reasoning under Uncertainty, Lecture Notes in Artificial Intelligence, 3571:39–50, Berlin, Springer-Verlag.

[23] Shenoy, P.P. 1997. Binary join trees for computing marginals in the Shenoy-Shafer architecture, International Journal of Approximate Reasoning, 17(2,3):239–263.

[24] Shenoy, P.P. and Shafer, G. 1990. Axioms for probability and belief function propagation, in R.D. Shachter, T.S. Levitt, J.F. Lemmer, L.N. Kanal (eds.), Uncertainty in Artificial Intelligence, 4:169–198, Amsterdam, North-Holland.

[25] Spiegelhalter, D.J.,Dawid, A.P.,Lauritzen, S.L. and Cowell, R.G. (1993), Bayesian analysis in expert systems, Statistical Science, 8(3):219–283.