Inference in Hybrid Bayesian Networks with Nonlinear Deterministic Conditionals*

Barry R. Cobb\textsuperscript{1} \quad Prakash P. Shenoy\textsuperscript{2}
cobbbr@vmi.edu \quad pshenoy@ku.edu

\textsuperscript{1}Department of Economics and Business\newline Virginia Military Institute\newline Lexington, VA 24450 USA

\textsuperscript{2}University of Kansas School of Business\newline 1300 Sunnyside Ave., Summerfield Hall\newline Lawrence, KS 66045-7601 USA

February 12, 2013

*\textsuperscript{}A small portion of this paper appeared in [3]. Comments and suggestions for improvement are welcome.
# Contents

1 Introduction 3

2 Extended Shenoy-Shafer Architecture 5

3 Mixtures of Polynomials and Truncated Exponentials 6
   3.1 Mixtures of Polynomials ................................. 6
   3.2 Mixtures of Truncated Exponentials ..................... 9
   3.3 Quality of MOP/MTE Approximations ..................... 10

4 Piecewise Linear Approximations of Nonlinear Functions 13
   4.1 One-Dimensional Functions ............................... 13
      4.1.1 The Quadratic Function $Y = X^2$ .................... 13
      4.1.2 The Exponential Function $Y = e^X$ ................. 16
   4.2 Multi-Dimensional Functions ............................. 18
      4.2.1 The Product Function $W = X \cdot Y$ .................. 18
      4.2.2 The Quotient Function $W = 3X/Y$ ..................... 23

5 Two Examples 25
   5.1 Crop Problem ........................................... 25
   5.2 Foreign Security Valuation ................................ 29

6 Summary and Conclusions 31
Inference in Hybrid Bayesian Networks with Nonlinear Deterministic Conditionals

Abstract

To enable inference in hybrid Bayesian networks containing nonlinear deterministic conditional distributions using mixtures of polynomials or mixtures of truncated exponentials, Cobb and Shenoy in 2005 propose approximating nonlinear deterministic functions by piecewise linear ones. In this paper, we describe a method for finding piecewise linear approximations of nonlinear functions based on two basic principles and an AIC-like heuristic. We illustrate our method for some commonly used one-dimensional and two-dimensional nonlinear deterministic functions such as $W = X^2$, $W = e^X$, $W = X \cdot Y$, and $W = X/Y$. Finally, we solve two small examples of hybrid Bayesian networks containing nonlinear deterministic conditionals that arise in practice.

Key Words: hybrid Bayesian networks, mixtures of truncated exponentials, mixtures of polynomials, deterministic conditional distributions

1 Introduction

This paper is concerned with inference in hybrid Bayesian networks containing nonlinear deterministic conditional distributions for some continuous variables. Hybrid Bayesian networks are Bayesian networks (BNs) containing a mix of discrete and continuous random variables. A random variable is said to be discrete if its state space is countable, and continuous otherwise.

In a BN, each variable is associated with a conditional probability distribution (or a conditional, in short), one for each state of its parent variables. A conditional for a variable is said to be deterministic if the variances of the conditional are all zeroes (for all states of the variable’s parents). If a discrete variable has a deterministic conditional, this does not cause any difficulties in the propagation algorithm. However, if a continuous variable has a deterministic conditional, then the joint probability density function for all continuous variables does not exist, and this must be taken into account in a propagation algorithm for computing posterior marginals. Recently, Shenoy and West [21] have proposed an extension of the Shenoy-Shafer architecture for discrete BNs [20] where deterministic conditionals for continuous variables are represented by Dirac delta functions [5]. Henceforth, when we talk about deterministic conditionals, we implicitly mean for continuous variables.

A major problem in inference in hybrid BNs is marginalizing continuous variables, which involves integration. Often, there are no closed form solutions for the result of the integration, making representation of the intermediate functions difficult. We will refer to this as the integration problem.

One of the earliest non-Monte Carlo methods for inference in hybrid Bayesian networks was proposed by Lauritzen and Jensen [8, 9] for the special case where all continuous variables have the so-called conditional linear Gaussian (CLG) probability distributions, and
discrete variables do not have continuous parents. A CLG distribution is a Gaussian distribution whose mean is a linear function of its continuous parents, and whose variance is a non-negative constant. Such BNs are called mixture of Gaussians BNs since the joint conditional distribution of all continuous variables is a multivariate Gaussian distribution for each combination of states of the discrete variables. Since marginals of multivariate Gaussian distributions are multivariate Gaussian distributions whose parameters can be easily found from the parameters of the original distribution, this obviates the need to do integrations. However, the requirement that all continuous conditional distributions are CLG, and the topological restriction that discrete variables have no continuous parents, restrict the class of hybrid BNs that can be represented using this method. Also, during the inference process, all continuous variables have to be marginalized before marginalizing discrete ones, and this restriction can lead to large cliques making inference intractable [10].

Another method for dealing with the integration problem is the mixture of truncated exponentials (MTE) technique proposed by Moral et al. [12]. The main idea here is to approximate conditional probability density functions (PDFs) by piecewise exponential functions, whose exponents are linear functions of the variables in the domain, and where the pieces are defined on hypercubes, i.e., intervals for each variable. Such functions are called MTEs, and this class of functions is closed under multiplication, addition, and integration, operations that are done in the propagation algorithm. Thus, the MTE method can be used for hybrid BNs that do not contain deterministic conditionals.

The MTE method does not pose restrictions such as the limitation that discrete variables cannot have continuous parents, and any conditional distribution can be used as long as they can be approximated by MTE functions. Cobb et al. [4] show that many commonly used univariate distributions can be approximated by MTE functions. Langseth et al. [7] show that empirical distributions can be approximated by MTE functions.

The family of MTE functions is not closed for operations required for multi-dimensional linear deterministic functions, such as, e.g., $W = X + Y$, because of the loss of the hypercube condition. However, Shenoy et al. [18] has shown that by using mixed tree hypercube approximation proposed by Moral et al. [13], we can recover the hypercube condition. Thus, the MTE method can be used for hybrid BNs containing linear deterministic conditionals. The MTE method cannot be used directly in hybrid Bayesian networks containing nonlinear conditionals. However, by approximating nonlinear functions by piecewise linear (PL) ones, the MTE method can be used for hybrid Bayesian networks containing nonlinear deterministic conditionals [2].

Recently, Shenoy and West [22] have proposed another method called mixture of polynomials (MOP) to address the integration problem. The main idea is to approximate conditional PDFs by piecewise polynomials defined on hypercubes. In all other respects, the MOP method is similar in spirit to the MTE method.

Recently, Shenoy [17] has proposed a generalization of the MOP function by allowing the pieces to be defined on regions called hyper-rhombases, which are a generalization of hypercubes. One advantage of MOPs defined on hyper-rhombases is that such functions are closed under transformations needed for multi-dimensional linear deterministic functions. Another advantage is that one can find accurate MOP approximations of high-dimensional CLG PDFs from a MOP approximation of a one-dimensional standard normal PDF. A disadvantage of hyper-rhombus MOPs is that such functions take longer to integrate than
hypercube MOPs. A detailed comparison of hypercube versus hyper-rhombus MOPs appears in [17, 19].

While a detailed comparison of the MTE and MOP methods has yet to be done, MOP approximations of PDFs can be easily found by using Lagrange interpolating polynomials with Chebyshev points [17]. This method can also be used with 2-dimensional conditional PDFs. Also, MOP functions are naturally closed under transformations for multi-dimensional linear deterministic functions, e.g., $W = X + Y$, etc., whereas MTEs are closed only if we use mixed-tree approximations to recover the hypercube condition which may be lost during the transformations needed for such functions. In the examples presented later in the paper, we will compare some results of using MTE functions versus MOP functions to model BNs with nonlinear deterministic variables.

Cobb and Shenoy [2] extend the applicability of MTE and MOP methods to hybrid Bayesian networks containing nonlinear deterministic conditionals. The main idea is to approximate a nonlinear function by a PL function, and then apply the usual MTE/MOP methods.

In this paper, we propose a method for finding PL approximations of nonlinear functions based on two basic principles and an AIC-like heuristic. We illustrate our method for some one-dimensional functions (such as $W = X^2$, and $W = e^X$), and some two-dimensional functions (such as $W = X \cdot Y$, and $W = X/Y$).

An outline of the remainder of the paper is as follows. In Section 2, we briefly sketch the extended Shenoy-Shafer architecture [21] for inference in hybrid BNs containing deterministic conditionals. In Section 3, we define mixtures of polynomials and mixtures of truncated exponentials functions. Also, we describe some numerical measures of goodness of an approximation of a PDF/CDF. In Section 4, we describe two basic principles and a heuristic for finding a PL approximation of a nonlinear function in one and two dimensions, and we illustrate these principles and heuristic for the functions $W = X^2$ and $W = e^X$ in the one-dimensional case, and $W = X \cdot Y$ and $W = X/Y$ for the two-dimensional case. In Section 5, we describe two examples of hybrid BNs containing nonlinear deterministic conditionals. Finally, in Section 6, we summarize our contributions and describe some issues for further research.

2 Extended Shenoy-Shafer Architecture

In this section, we briefly sketch the extended Shenoy-Shafer architecture [21] for inference in hybrid BNs containing deterministic conditionals.

Conditionals for discrete variables are represented by functions called discrete potentials, whose values are in units of probability (dimension-less quantities). Conditionals for continuous variables are represented by functions called continuous potentials, whose units are in units of probability density (probability per unit $X$, where $X$ is a continuous variable). If $X$ is a continuous variable with a deterministic conditional represented by the deterministic function $X = g(Y_1, \ldots, Y_n)$, where $Y_1, \ldots, Y_n$ are the continuous parents of $X$, then such a conditional is represented by $\delta(x - g(y_1, \ldots, y_n))$, where $\delta$ denotes the Dirac delta function [5].

In the process of making inferences, we use two operations called combination and
marginalization. Combination of potentials consists of pointwise multiplication. The units of the combined potential is the product of the units of the component potentials. Marginalizing a discrete variable from a potential is by addition over the state space of the discrete variable. The units of the marginal is the same as the units of the potential being marginalized. Marginalizing a continuous variable from a potential is by integrating the potential over the state space of the continuous variable. If the potential being marginalized does not contain Dirac delta functions, the the usual rules of Riemann integration apply. If the potential being marginalized contains Dirac delta functions, then we use the properties of Dirac delta functions [21]. In either case, the units of the marginal consists of the units of the potential multiplied by the units of continuous variable $X$.

In all other respects, the extended Shenoy-Shafer architecture is the same as the Shenoy-Shafer architecture [20]. Given a hybrid BN with evidence potentials, we first construct a binary join tree [16], and then propagate messages in the binary join tree resulting in the marginals of variables of interest.

3 Mixtures of Polynomials and Truncated Exponentials

In this section, we will define mixture of polynomials and mixture of truncated exponentials functions. We will explain why mixtures of truncated exponentials are not closed under transformations needed for multi-dimensional linear deterministic conditionals, and we show how we can recover the mixture of truncated exponentials property using mixed trees.

3.1 Mixtures of Polynomials

The definition of mixture of polynomials given here is taken from [17].

A one-dimensional function $f : \mathbb{R} \to \mathbb{R}$ is said to be a mixture of polynomials (MOP) function if it is a piecewise function of the form:

$$f(x) = \begin{cases} a_{0i} + a_{1i}x + \cdots + a_{ni}x^n & \text{for } x \in A_i, i = 1, \ldots, k, \\ 0 & \text{otherwise.} \end{cases} \quad (3.1)$$

where $A_1, \ldots, A_k$ are disjoint intervals in $\mathbb{R}$ that do not depend on $x$, and $a_{0i}, \ldots, a_{ni}$ are constants for all $i$. We will say that $f$ is a $k$-piece (ignoring the 0 piece), and $n$-degree (assuming $a_{ni} \neq 0$ for some $i$) MOP function.

Example 1. An example of a 2-piece, 3-degree MOP function $g_1(\cdot)$ in one-dimension is as follows:

$$g_1(x) = \begin{cases} 0.424 + 0.128x - 0.085x^2 - 0.028x^3 & \text{if } -3 < x < 0, \\ 0.424 - 0.128x - 0.085x^2 + 0.028x^3 & \text{if } 0 \leq x < 3 \\ 0 & \text{otherwise} \end{cases} \quad (3.2)$$

$g_1(\cdot)$ is a MOP approximation of the PDF of the standard normal distribution on the domain $(-3, 3)$, and was found using Lagrange interpolating polynomial with Chebyshev points [17].
The main motivation for defining MOP functions is that such functions are easy to integrate in closed form, and that they are closed under multiplication, integration, and addition, the main operations in making inferences in hybrid Bayesian networks. The requirement that each piece is defined on an interval $A_i$ is also designed to ease the burden of integrating MOP functions.

A multivariate polynomial is a polynomial in several variables. For example, a polynomial in two variables is as follows:

$$P(x_1, x_2) = a_{00} + a_{10}x_1 + a_{01}x_2 + a_{11}x_1x_2 + a_{20}x_1^2 + a_{02}x_2^2 + a_{21}x_1x_2^2 + a_{12}x_1^2x_2 + a_{22}x_1x_2^2$$

The degree of the polynomial in Eq. (3.3) is 4 assuming $a_{22}$ is a non-zero constant. In general, the degree of a multivariate polynomial is the largest sum of the exponents of the variables in the terms of the polynomial.

An $m$-dimensional function $f : \mathbb{R}^m \to \mathbb{R}$ is said to be a MOP function if

$$f(x_1, x_2, \ldots, x_m) = \begin{cases} P_i(x_1, x_2, \ldots, x_m) & \text{for } (x_1, x_2, \ldots, x_m) \in A_i, i = 1, \ldots, k, \\ 0 & \text{otherwise} \end{cases} \quad (3.4)$$

where $P_i(x_1, x_2, \ldots, x_m)$ are multivariate polynomials in $m$ variables for all $i$, and the regions $A_i$ are disjoint and as follows. Suppose $\pi$ is a permutation of $\{1, \ldots, m\}$. Then each $A_i$ is of the form:

$$l_{1i} \leq x_{\pi(1)} \leq u_{1i},
$$

$$l_{2i}(x_{\pi(1)}) \leq x_{\pi(2)} \leq u_{2i}(x_{\pi(1)}),
$$

$$l_{mi}(x_{\pi(1)}, \ldots, x_{\pi(m-1)}) \leq x_{\pi(m)} \leq u_{mi}(x_{\pi(1)}, \ldots, x_{\pi(m-1)})$$

where $l_{1i}$ and $u_{1i}$ are constants, and $l_{ji}(x_{\pi(1)}, \ldots, x_{\pi(j-1)})$ and $u_{ji}(x_{\pi(1)}, \ldots, x_{\pi(j-1)})$ are linear functions of $x_{\pi(1)}, x_{\pi(2)}, \ldots, x_{\pi(j-1)}$ for $j = 2, \ldots, m$, and $i = 1, \ldots, k$. We will refer to the nature of the region described in Eq. (3.5) as a hyper-rhombus. Although we have defined the hyper-rhombus as a closed region in Eq. (3.5), each of the $2m$ inequalities can be either strictly $<$ or $\leq$.

A special case of a hyper-rhombus is a region of the form:

$$l_{1i} \leq x_1 \leq u_{1i},
$$

$$l_{2i} \leq x_2 \leq u_{2i},
$$

$$\vdots$$

$$l_{mi} \leq x_m \leq u_{mi}$$

where $l_{1i}, \ldots, l_{mi}, u_{1i}, \ldots, u_{mi}$ are all constants. We will refer to the region described in Eq. (3.6) as a hypercube (in $m$-dimensions).
Example 2. An example of a 2-piece, 3-degree MOP \( g_2(\cdot, \cdot) \) defined on a two-dimensional hyper-rhombus is as follows:

\[
g_2(x, y) = \begin{cases} 
0.425 + 0.128(y - x) - 0.085(y - x)^2 - 0.028(y - x)^3 & \text{if } x - 3 < y < x, \\
0.424 - 0.128(y - x) - 0.085(y - x)^2 + 0.028(y - x)^3 & \text{if } x \leq y < x + 3 \\
0 & \text{otherwise}
\end{cases} 
\]  

(3.7)

\( g_2(x, y) \) is a two-dimensional MOP approximation of the PDF of the CLG distribution of \( Y \) given \( x \) with mean \( x \) and variance 1 on the domain \(-\infty < x < \infty, x - 3 < y < x + 3\). Notice that \( g_2(x, y) = g_1(y - x) \), where \( g_1(\cdot) \) is as defined in Eq. (3.2).

The family of MOP functions is closed under multiplication, addition and integration, the operations that are done during propagation of potentials in the extended Shenoy-Shafer architecture for hybrid Bayesian networks. They are also closed under transformations needed for linear deterministic functions. We will illustrate this by a small example.

![Figure 1: A BN with a Sum Deterministic Conditional.](image)

Example 3. Consider the BN shown in Fig. 1. In this BN, \( X, Y, \) and \( W \) are all continuous, and \( W \) has a deterministic conditional, \( W = X + Y \). Suppose we are interested in computing the marginal PDF of \( W \). Suppose \( g_1(\cdot) \) is a MOP approximation of the PDF of the standard normal distribution (as described in Eq. (3.2)). Then \( \xi(x) = g_1(x - 3) \) is a MOP approximation of the PDF of \( X \), and \( \psi(x, y) \) as defined in Eq. (3.8) is a MOP approximation of the conditional PDF of \( Y \mid x \).

\[
\psi(x, y) = \frac{g_1\left(\frac{y-6-2x}{2}\right)}{2} 
\]

(3.8)

The deterministic conditional of \( W \) is represented by \( \omega(x, y, w) = \delta(w - x - y) \), where \( \delta \)
is the Dirac delta function. First, we marginalize $Y$:

$$(((\psi \otimes \omega)^{-Y})(x, w) = \int_{-\infty}^{\infty} \psi(x, y) \omega(x, y, w) \, dy$$

$$= \int_{-\infty}^{\infty} \psi(x, y) \delta(w - x - y) \, dy$$

$$= \psi(x, w - x)$$ (3.9)

The result in Eq. (3.9) follows from the sampling property of Dirac delta function: If $f$ is continuous in a neighborhood of $a$, then

$$\int_{-\infty}^{\infty} f(x) \delta(x - a) \, dx = f(a)$$ (3.10)

Since $\psi(x, y)$ is a MOP, $\psi(x, w - x)$ is a MOP. Next, we marginalize $X$:

$$(((\xi \otimes (\psi \otimes \omega)^{-Y}))^{-X})(w) = \int_{-\infty}^{\infty} \xi(x) \psi(x, w - x) \, dx$$ (3.11)

Since $\xi(x)$, and $\psi(x, w - x)$ are MOPs, the marginal distribution of $W$ computed in Eq. (3.11) is a MOP. ■

### 3.2 Mixtures of Truncated Exponentials

An $m$-dimensional function $f : \mathbb{R}^m \rightarrow \mathbb{R}$ is said to be a mixture of truncated exponentials (MTE) function if

$$f(x_1, x_2, \ldots, x_m) = \begin{cases} 
    a_{i0} + \sum_{j=1}^{n} a_{ij} e^{\sum_{l=1}^{m} b_{ijl} x_l} & \text{for } (x_1, x_2, \ldots, x_m) \in A_i, i = 1, \ldots, k, \\
    0 & \text{otherwise} 
\end{cases}$$ (3.12)

where $a_{i0}$, $a_{ij}$, and $b_{ijl}$ are constants, and $A_i$ are disjoint hypercubes in $m$-dimensions. We say $f$ is a $k$-piece, $n$-term MTE function.

**Example 4.** An example of a 2-piece, 3-term MTE function $h_1(\cdot)$ in one-dimension is as follows:

$$h_1(x) = \begin{cases} 
    0.011 + 197.588 e^{2.257 x} - 462.686 e^{2.343 x} + 265.508 e^{2.404 x} & \text{if } -3 < x < 0, \\
    0.011 + 197.588 e^{-2.257 x} - 462.686 e^{-2.343 x} + 265.508 e^{-2.404 x} & \text{if } 0 \leq x < 3 \\
    0 & \text{otherwise} 
\end{cases}$$ (3.13)

$h_1(\cdot)$ is a MOP approximation of the PDF of the standard normal distribution on the domain $(-3, 3)$, and was found by solving a nonlinear optimization problem [4]. ■

Unlike MOPs, MTEs cannot be defined on hyper-rhombus regions since the family of such functions would not be closed under the integration operation. If we lose the hypercube property after some operation, we can approximate such functions using mixed-tree
approximation as suggested in [13]. We will illustrate mixed-tree approximations for the BN in Example 3.

**Example 5.** Consider the BN as shown in Fig. 1. Using \( h_1(\cdot) \) described in Eq. (3.13), an MTE approximation of the PDF \( X \) is given by \( \xi(x) = h_1(x - 3) \). Unfortunately, \( \psi(x, y) = \frac{h_1(y - \frac{6 - 2x}{2})}{y - \frac{6 - 2x}{2}} \) is not an MTE function since the pieces are now defined on regions such as \(-3 < y - \frac{6 - 2x}{2} < 0\), etc., which are not hypercubes. So we partition the domain of \( X \) into several pieces, and assume that the \( x \) is a constant in each piece equal to the mid-point of the piece. Thus \( \psi_c(x, y) \) as described in Eq. (3.14) is a 3-point mixed tree MTE approximation of \( \psi(x, y) \) representing the conditional PDF of \( Y \mid x \).

\[
\psi_c(x, y) = \begin{cases} 
  h_1(y - 8)/2 & \text{if } 0 < x < 2, \\
  h_1(y - 12)/2 & \text{if } 2 \leq x < 4, \\
  h_1(y - 16)/2 & \text{if } 4 \leq x < 6.
\end{cases}
\]  

(3.14)

Notice that the pieces of \( \psi_c(x, y) \) are defined on hypercubes. After marginalizing \( y \), we obtain the potential \( \phi(x, w) = \psi_c(x, w - x) \), which is no longer defined on hypercubes (since we have pieces \(-3 < w - x < 0\), etc.). So we approximate \( \phi(x, w) \) by \( \phi_c(x, w) \) using mixed trees as follows:

\[
\phi_c(x, w) = \begin{cases} 
  \psi_c(1, w - 1) & \text{if } 0 < x < 2, \\
  \psi_c(3, w - 3) & \text{if } 2 \leq x < 4, \\
  \psi_c(5, w - 5) & \text{if } 4 \leq x < 6.
\end{cases}
\]  

(3.15)

\( \phi_c(x, w) \) is an MTE function since it is defined on hypercubes. After we marginalize \( X \), we obtain the marginal for \( W \) as given in Eq. (3.9) where \( \phi(x, w) = \psi(x, w - x) \) is now replaced by \( \phi_c(x, w) \). ■

Notice that the mixed tree method can also be used for MOPs, in which case we obtain MOPs defined on hypercubes. Although we lose some accuracy, MOPs defined on hypercubes can be integrated faster than MOPs defined on hyper-rhombuses. The accuracy of a mixed tree approximation can be improved by using more points. However, using more points means more pieces, which slows down the computation of integrals. A key point is that with hypercube MOPs/MTEs, we have a choice between fast computation and accuracy, which we don’t have with hyper-rhombus MOPs.

### 3.3 Quality of MOP/MTE Approximations

In this section, we discuss some quantitative ways to measure the accuracy of a MOP/MTE approximation of PDFs.

We will measure the accuracy of a PDF with respect to another defined on the same domain by four different measures, the Kullback-Leibler (KL) divergence, maximum absolute deviation, absolute error in the mean, and absolute error in the variance.

If \( f \) is a PDF on the interval \((a, b)\), and \( g \) is a PDF that is an approximation of \( f \) such that \( g(x) > 0 \) for \( x \in (a, b) \), then the KL divergence between \( f \) and \( g \), denoted by \( KL(f, g) \),
is defined as follows ([6]):

\[
KL(f, g) = \int_a^b \ln \left( \frac{f(x)}{g(x)} \right) f(x) \, dx. \tag{3.16}
\]

\(KL(f, g) \geq 0\), and \(KL(f, g) = 0\) if and only if \(g(x) = f(x)\) for all \(x \in (a, b)\). We do not know the semantics associated with the statistic \(KL(f, g)\).

The maximum absolute deviation between \(f\) and \(g\), denoted by \(MAD(f, g)\), is given by:

\[
MAD(f, g) = \sup \{|f(x) - g(x)| : a < x < b\} \tag{3.17}
\]

One semantic associated with \(MAD(f, g)\) is as follows. If we compute the probability of some interval \((c, d) \subseteq (a, b)\) by computing \(\int_c^d g(x) \, dx\), then the error in this probability is bounded by \((d - c) \cdot MAD(f, g)\).

The maximum absolute deviation can also be applied to CDFs. Thus, if \(F(\cdot)\) and \(G(\cdot)\) are the CDFs corresponding to \(f(\cdot)\), and \(g(\cdot)\), respectively, then the maximum absolute deviation between \(F\) and \(G\), denoted by \(MAD(F, G)\), is

\[
MAD(F, G) = \sup \{|F(x) - G(x)| : a < x < b\} \tag{3.18}
\]

The value \(MAD(F, G)\) is in units of probability, whereas the value \(MAD(f, g)\) is in units of probability density, and the two values cannot be compared to each other. The semantic associated with \(MAD(F, G)\) is as follows. If we compute the probability of some interval \((c, d) \subseteq (a, b)\) using \(G(d) - G(c)\), then the error in this probability is bounded by \(2 \cdot MAD(F, G)\).

The absolute error of the mean, denoted by \(AEM(f, g)\), and the absolute error of the variance, denoted by \(AEV(f, g)\), are given by:

\[
AEM(f, g) = |E(f) - E(g)| \tag{3.19}
\]

\[
AEV(f, g) = |V(f) - V(g)| \tag{3.20}
\]

where \(E(\cdot)\) and \(V(\cdot)\) denote the expected value and the variance of a PDF, respectively.

To illustrate these definitions, let \(f(\cdot)\) denote the PDF of the standard normal distribution truncated to \((-3, 3)\). Consider \(g_1(\cdot)\), the 2-piece, 3-degree MOP approximation of \(f(\cdot)\) as described in Eq. (3.2). Also, let \(F(\cdot)\) and \(G_1(\cdot)\) denote the CDFs corresponding to \(f\) and \(g_1\), respectively. Fig. 2 shows a graph of \(g_1(z)\) overlaid on the graph of \(f(z)\). The goodness of fit statistics for \(g_1\) are as follows: \(KL(f, g_1) \approx 0.0051\), \(MAD(f, g_1) \approx 0.0248\), \(MAD(F, G_1) \approx 0.0028\), \(AEM(f, g_1) \approx 0.0000\), \(AEV(f, g_1) \approx 0.0239\).

Consider \(h_1\), the 2-piece, 3-term MTE approximation of \(f(\cdot)\) as described in Eq. (3.13). Let \(H_1\) denote the CDF corresponding to \(h_1\). Fig. 3 shows a graph of \(h_1(z)\) overlaid on the graph of \(f(z)\). The goodness of fit statistics for \(h_1\) are as follows: \(KL(f, h_1) \approx 0.0004\), \(MAD(f, h_1) \approx 0.0056\), \(MAD(F, H_1) \approx 0.0020\), \(AEM(f, h_1) \approx 0.0000\), \(AEV(f, h_1) \approx 0.0085\).
Figure 2: A Graph of $g_1(z)$ (in red) Overlaid on the Graph of $f(z)$ (in blue).

Figure 3: A Graph of $h_1(z)$ (in red) Overlaid on the Graph of $f(z)$ (in blue).
4 Piecewise Linear Approximations of Nonlinear Functions

When we have nonlinear deterministic conditionals, our strategy is to approximate these functions by PL functions. The family of MOP functions is closed under the operations needed for linear deterministic functions. The family of MTE functions is also closed for linear deterministic functions assuming each time we lose the hypercube condition, we replace the function by a mixed-tree MTE approximation as described in Section 3.2.

There are many ways in which we can approximate a nonlinear function by a PL function. In this section, we examine two basic principles and a heuristic with the goal of minimizing the errors in the marginal distribution of the variable with the deterministic conditional represented by the PL approximation.

4.1 One-Dimensional Functions

In this subsection, we will describe PL approximations of two one-dimensional functions $Y = X^2$ and $Y = e^X$ using two basic principles and a heuristic.

4.1.1 The Quadratic Function $Y = X^2$

Consider a simple BN as follows: $X \sim N(0, 1)$, $Y = X^2$. The exact marginal distribution of $Y$ is chi-square with 1 degree of freedom. We will use the 2-piece, 3-degree MOP $g_1(\cdot)$ defined in Eq. (3.2), and the 2-piece, 3-term MTE function $h_1(\cdot)$ defined in Eq. (3.13) on the domain $(-3, 3)$, for the MOP/MTE approximation of the PDF of $N(0, 1)$.

Two Basic Principles In constructing PL approximations, we will adhere to two basic principles. First, the domain of the marginal PDF of the variable with the deterministic conditional should remain unchanged. Thus, in the chi-square example, since the PDF of $X$ is defined on the domain $(-3, 3)$, and $Y = X^2$, the domain of $Y$ is $(0, 9)$, and we need to ensure that any PL approximation of the function $Y = X^2$ results in the marginal PDF of $Y$ on the domain $(0, 9)$. We will refer to this principle as the domain principle.

Second, if the PDF of $X$ is symmetric about some point, and the deterministic function is also symmetric about the same point, then we need to ensure that the PL approximation retains the symmetry. In the chi-square example, the PDF of $X$ is symmetric about the point $X = 0$, and $Y = X^2$ is also symmetric about the point $X = 0$ on the domain $(-3, 3)$. Therefore, we need to ensure that the PL approximation is also symmetric about the point $X = 0$. We will refer to this principle as the symmetry principle. Notice that the symmetry principle applies only when we have symmetry of the deterministic function and symmetry of the joint PDF of the parent variables about a common point.

AIC-like Heuristic In the statistics literature, there are several heuristics (such as Akaike’s information criterion (AIC) [1] and Bayes information criterion (BIC) [15]) for building statistical models from data. For example, in a multiple regression setting, if we have a data set with $p$ explanatory variables and a response variable, we could always decrease the sum
of squared errors in the model by using more explanatory variables. However, this could lead to over-fitting and lead to poor predictive performance. Thus, we need a measure that has a penalty factor for including more explanatory variables than is necessary. If we have a model with $p$ explanatory variables, and $\hat{\sigma}^2$ is an estimate of $\sigma^2$ in the regression model, the AIC heuristic is to minimize $n \times \ln(\hat{\sigma}^2) + 2p$, where the $2p$ term acts like a penalty factor for using more explanatory variables than are necessary.

Our context here is slightly different from statistics. In statistics, we have data, and the true model is unknown. In our context, there is no data and the true model is known (the true model could be a nonlinear model estimated from data). However, there are some similarities. We could always decrease the error in the fit between the nonlinear function and the PL approximation by using more parameters (pieces), but doing so does not always guarantee that the error in the marginal distribution of the deterministic variable with the nonlinear function will be minimized. We will demonstrate that using more parameters may lead to a worse result for the marginal distribution of the target variable. Also, making inferences with MOPs/MTEs that have many pieces can be intractable [18]. For this reason, we need to keep the number of pieces as small as possible.

Suppose $f_X(x)$ denotes the PDF of $X$ and suppose we approximate a non-linear deterministic function $Y = r(X)$ by a PL function, say $Y = r_1(X)$, that has $p$ free parameters. By free parameters, we mean parameters that are used in specifying $Y = r_1(X)$ that are not already included in the specifications of $f_X(x)$ and $Y = r(X)$, and those that can vary freely without violating the domain and symmetry principles. The mean square error (MSE) of the PL approximation $r_1$, denoted by $MSE(r_1)$, is given by

$$MSE(r_1) = \int_{-\infty}^{\infty} f_X(x) (r(x) - r_1(x))^2 \, dx. \quad (4.1)$$

The AIC-like heuristic finds a PL approximation $Y = r_1(X)$ with $p$ free parameters such that the $AIC(r_1) = \ln(MSE(r_1)) + p$ is minimized subject to the domain and symmetry principles.

In approximating higher dimensional nonlinear functions, we need to balance the weight between the $\ln(MSE(r_1))$ factor and the penalty factor $p$. For example, in approximating $W = r(X,Y)$, introducing an additional PL piece, of the form $W = aX + bY + c$, may cost as much as 3 additional free parameters. So to balance the two competing factors, one possibility is to define the AIC-like heuristic as follows:

$$AIC(r_1) = d \ln(MSE(r_1)) + p \quad (4.2)$$

where $d$ is the dimension of the PL function $r_1$. For one-dimensional functions, e.g., $Y = X^2$, $d = 1$. For two-dimensional function, e.g., $W = X \cdot Y$, $d = 2$.

For the chi-square BN, the domain and symmetry principles require use of $(-3,9), (0,0)$, and $(3,9)$ as knots of a PL approximation. The knots are the endpoints of the intervals of the domain of $X$ and, along with the corresponding values for $Y = X^2$, completely determine the PL approximation. Suppose we wish to find a 4-piece PL approximation. Let $(x_1, y_1)$ and $(-x_1, y_1)$ denote the two additional knots where $-3 < x_1 < 0$, and $0 < y_1 < 9$. Such a PL approximation would consist of 2 free parameters (where the parameters are $x_1$ and $y_1$). Solving for the minimum $MSE(r_1)$ with $g_1(x)$ as the PDF of $X$ results in the solution:
Figure 4: A Graph of $Y = r_1(X)$ (in red) Overlaid on the Graph of $Y = X^2$ (in blue).

$x_1 = -1.28$, $y_1 = 1.16$, the minimum value of $MSE(r_1) = 0.0433$, and the corresponding value of $AIC(r_1)$ is $-1.1405$.

The PL approximation $Y = r_1(X)$ is as follows (see Fig. 4):

$$Y = \begin{cases} 
-4.66 - 4.55X & \text{if } X < -1.28 \\
-0.91X & \text{if } -1.28 \leq X < 0 \\
0.91X & \text{if } 0 \leq X < 1.28 \\
-4.66 + 4.55X & \text{if } X \geq 1.28 
\end{cases} \quad (4.3)$$

If we approximate $Y = X^2$ by a PL approximation $Y = r_2(X)$ with, say 6 pieces (4 parameters), then the value of $MSE(r_2)$ is 0.0060, and the value of $AIC(r_2)$ is $-1.1235$, which is higher than $AIC(r_1)$. Similarly, if we use a 8-piece approximation (with 6 free parameters), then the value of $MSE(r_3)$ is 0.0016, and the value of $AIC(r_3)$ is $-0.4205$, which is higher than $AIC(r_1)$ and $AIC(r_2)$. Thus, the AIC heuristic suggests a 2-piece PL approximation $Y = r_1(X)$. The accuracies of the marginal PDF of $Y$ computed using MOP $g_1(x)$ for the PDF of $X$, and the three PL approximations $r_1$, $r_2$, and $r_3$ are shown in Table 1 (best values are shown in boldface). The alternate model used as the actual PDF to calculate the goodness of fit statistics is the marginal PDF of $Y$ found using $g_1$ and $Y = X^2$.

Solving for the minimum $MSE(r_4)$ with the MTE function $h_1(x)$ as the PDF of $X$ and a 2-piece PL approximation of $Y = X^2$ results in the solution: $x_1 = -1.29$, $y_1 = 1.17$, minimum value of $MSE(r_4) = 0.0454$, and the corresponding value of $AIC(r_4)$ is $-1.0931$.

The accuracy of the marginal PDF of $Y$ computed using MTE $h_1(x)$ for the PDF of $X$, and the PL approximation $r_4$ is shown in Table 1 (see the rightmost column). The CPU row gives the run time required by Mathematica 9.0 to compute the marginal PDF of $W$ using the PL approximation. All experiments were run on a desktop computer under identical conditions.

The alternate model used as the actual PDF to calculate the goodness of fit statistics is the marginal PDF of $Y$ found using $h_1$ and $Y = X^2$. We did not calculate PL approximations with 6 or 8 pieces by using the MTE function to solve the optimization problem because using MTE functions to find parameters for such functions has proven to be intractable in many of our experiments.

Let $g_2(\cdot)$ denote the marginal PDF of $Y$ computed using $g_1(\cdot)$ and $Y = X^2$, and Let $h_2(\cdot)$
Table 1: Goodness of Fit and Run Time Results for the Quadratic Function Example.

<table>
<thead>
<tr>
<th>Model</th>
<th>MOP</th>
<th>MOP</th>
<th>MOP</th>
<th>MTE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td># pieces</td>
<td># free parameters</td>
<td>MSE</td>
<td>AIC</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2</td>
<td>0.0433</td>
<td>-1.1405</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>4</td>
<td>0.0060</td>
<td>-1.1235</td>
</tr>
<tr>
<td></td>
<td>8</td>
<td>6</td>
<td><strong>0.0016</strong></td>
<td>-0.4205</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>2</td>
<td>0.0454</td>
<td>-1.0931</td>
</tr>
</tbody>
</table>

denote the marginal PDF of $Y$ computed using $h_1(\cdot)$ and $Y = X^2$ (the CDFs corresponding to these PDFs are denoted by $G_2$ and $H_2$, respectively). Let $g_{11}(\cdot)$ denote the marginal PDF of $Y$ using $g_1(\cdot)$ and $Y = r_1(X)$, and let $h_{14}(\cdot)$ denote the marginal PDF of $Y$ using $h_1(\cdot)$ and $Y = r_4(X)$. Let $G_{11}(\cdot)$ and $H_{14}(\cdot)$ denote the CDFs corresponding to PDFs $g_{11}(\cdot)$, and $h_{14}(\cdot)$, respectively. A graph of $G_{11}(\cdot)$ overlaid on the graph of $G_2(\cdot)$, and a graph of $H_{14}(\cdot)$ overlaid on the graph of $H_2(\cdot)$, are shown in Fig. 5.

Figure 5: Left: A Graph of $G_{11}$ (in red) Overlaid on the Graph of $G_2$ (in blue). Right: A Graph of $H_{14}$ (in red) Overlaid on the Graph of $H_2$ (in blue).

4.1.2 The Exponential Function $Y = e^X$

Consider the problem where $X \sim N(0, 1)$, $Y = e^X$, and we wish to compute the marginal PDF of $Y$. The theoretical marginal distribution of $Y$ is log-normal with parameters $\mu = 0$ and $\sigma^2 = 1$. As the PDF of $X$ is approximated on the domain $(-3, 3)$, the domain principle requires that the marginal for $Y$ be defined on the domain $(e^{-3}, e^3)$. Thus, in finding a PL approximation of $Y = e^X$, we need to use the knots $(-3, e^{-3})$ and $(3, e^3)$. Although the PDF
of $X$ is symmetric about the axis $X = 0$, the function $Y = e^X$ is not symmetric about any axis. Therefore, the symmetry principle does not apply for this problem.

Suppose we wish to find, e.g., a 2-piece PL approximation of $Y = r(X) = e^X$. This involves solving an optimization problem with two parameters associated with the knots $(-3, e^{-3}), (a_1, b_1), (3, e^3)$. Let $Y = r_1(X)$ denote the PL approximation given by these knots. We solve an optimization problem as follows:

Find $a_1, b_1$ so as to

\[
\text{Minimize } \int_{-\infty}^{\infty} g_1(x)(r(x) - r_1(x))^2 \, dx \quad (4.4)
\]

subject to: $-3 < a_1 < 3,$ and $\ e^{-3} < b_1 < e^3$.

Solving this optimization problem results in the optimal solution: $a_1 = 1.32, b_1 = 1.91$. A graph of the PL approximation $Y = r_1(X)$ overlaid on $Y = r(X)$ is shown in Fig. 6. The minimum value of MSE is 0.4566, and the corresponding value of AIC is 1.2161.

Using the PL approximation $Y = r_1(X)$, and the MOP approximation $g_1(x)$ of the PDF of $N(0, 1^2)$, we computed the marginal PDF/CDF of $Y$, and compared it with the “exact” marginal PDF/CDF of $Y$ (computed using $g_1(x)$ and $Y = e^X$, which is not a MOP, but we have a representation of it). Fig. 7 shows the marginal CDF of $Y$ computed using $Y = r_1(X)$ overlaid on the marginal CDF of $Y$ computed using $Y = e^X$.

We repeated this procedure for a 3-piece and 4-piece PL approximation of $Y = e^X$. The AIC value is the smallest for the 2-piece approximation. The goodness of fit statistics for the three PL approximations are as shown in Table 2. Also shown are CPU time (in seconds) required to compute the marginal PDF of $Y$. Notice that the 2-piece PL approximation results in the smallest MAD of PDF statistic, and requires the least CPU time for computing the marginal PDF of $Y$. In the case of the 2-piece PL approximation, the marginal PDF of $Y$ is computed as a 3-piece, 3-degree MOP. In the case of the 3-piece PL approximation, the marginal PDF of $Y$ is computed as a 4-piece, 3-degree MOP, and in the case of the 4-piece PL approximation, the marginal PDF of $Y$ is a 5-piece, 3-degree MOP. Thus, the 2-piece PL approximation results in the most economical marginal representation of the marginal...
Figure 7: A Graph of the CDF of \( Y \) using \( Y = r_1(X) \) (in red) Overlaid on the Graph of the CDF of \( Y \) using \( Y = e^X \) (in blue).

Table 2: Goodness of Fit and Run Time Results for the Exponential Function Example.

<table>
<thead>
<tr>
<th># pieces</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td># free parameters</td>
<td>2</td>
<td>4</td>
<td>6</td>
</tr>
<tr>
<td>MSE</td>
<td>0.4566</td>
<td>0.0623</td>
<td>0.0170</td>
</tr>
<tr>
<td>AIC</td>
<td>1.2161</td>
<td>1.2239</td>
<td>1.9229</td>
</tr>
<tr>
<td>KL</td>
<td>0.8437</td>
<td>0.3747</td>
<td>0.1933</td>
</tr>
<tr>
<td>MAD of PDF</td>
<td>0.6233</td>
<td>1.1312</td>
<td>1.3891</td>
</tr>
<tr>
<td>MAD of CDF</td>
<td>0.3035</td>
<td>0.1484</td>
<td>1.0000</td>
</tr>
<tr>
<td>AEM</td>
<td>0.1429</td>
<td>0.0305</td>
<td>0.0107</td>
</tr>
<tr>
<td>AEV</td>
<td>0.3164</td>
<td>0.1073</td>
<td>0.0589</td>
</tr>
<tr>
<td>CPU (in secs.)</td>
<td>0.47</td>
<td>0.66</td>
<td>0.84</td>
</tr>
</tbody>
</table>

PDF of \( Y \), which may explain why the CPU time is lowest for the case of the 2-piece PL approximation.

4.2 Multi-Dimensional Functions

In this section, we find PL approximations of the two-dimensional nonlinear functions, \( W = X \cdot Y \), and \( W = X/Y \). For multi-dimensional nonlinear functions, we can use the same principles and heuristic as for the one-dimensional case.

4.2.1 The Product Function \( W = X \cdot Y \)

Consider a Bayesian network: \( X \sim N(5, 0.5^2) \), \( Y \sim N(15, 4^2) \), \( X \) and \( Y \) are independent, and \( W = r(X, Y) = X \cdot Y \). We have a 2-piece, 3-degree MOP \( g_X(x) = g_1(x - 5)/0.5 \) of the PDF of \( X \) on the domain \((3.5, 6.5)\), and a 2-piece, 3-degree MOP \( g_Y(y) = g_1(y - 15)/4 \) of the PDF of \( Y \) on the domain \((3, 27)\) (here \( g_1(\cdot) \) is the 2-piece, 3-degree MOP approximation of the standard normal PDF on the domain \((-3, 3)\) as described in Eq. 3.2).
Using these two MOP approximations of the PDFs of \( X \) and \( Y \), we can find an “exact” marginal PDF of \( W \) as follows:

\[
g_W(w) = \int_{-\infty}^{\infty} g_X(x) \left( \int_{-\infty}^{\infty} g_Y(y) \delta(w - x \cdot y) \, dy \right) \, dx \quad (4.5)
\]

\( g_W(\cdot) \) is not a MOP, but we do have a representation of it, and can compute its mean \( (E(g_W) = 75) \) and variance \( (V(g_W) = 458.96) \). Unfortunately, we cannot compute the CDF corresponding to \( g_W(\cdot) \). So we do not report any MAD for the CDFs statistics. Also, if we use MTE approximations of the PDFs of \( X \) and \( Y \), we cannot compute an “exact” marginal PDF of \( W \). Therefore we do not report any results for the MTE case.

Suppose we wish to find a 2-piece PL approximation of \( W = X \cdot Y \) as follows:

\[
r_1(x,y) = \begin{cases} 
    a_1 x + b_1 y + c_1 & \text{if } x < s_X \\
    a_2 x + b_2 y + c_2 & \text{if } x \geq s_X 
\end{cases}
\]

Notice that the number of parameters in this 2-piece PL approximation is 7 \((a_1, b_1, c_1, a_2, b_2, c_2, \text{ and } s_X)\).

The domain of the joint distribution of \( X \) and \( Y \) is a rectangle \((3.5 < X < 6.5) \times (3 < Y < 27)\). The exact domain of \( W \) is \((10.5, 175.5)\). We need to find a PL approximation \( r_1(X, Y) \) of \( r(X, Y) = X \cdot Y \) that satisfies the domain principle. The smallest value of \( W = X \cdot Y \) is 10.5 at the point \((X, Y) = (3.5, 3)\), and the largest value of \( W \) is 175.5 at the point \((X, Y) = (6.5, 27)\). To satisfy the domain principle, we impose the constraints \( r_1(3.5, 3) = 10.5 \), and \( r_1(6.5, 27) = 175.5 \). These two equality constraints reduce the number of free parameters from 7 to 5.

The function \( W = X \cdot Y \) is symmetric about the axis \( X = Y \), but the joint PDF of \((X, Y)\) is not symmetric about this axis. Therefore, the symmetry principle does not apply.

We define the following function that will be used in the optimization problem to find the parameters for the PL approximation:

\[
MSE(a_1, b_1, c_1, a_2, b_2, c_2, s_X) = \int_{-\infty}^{\infty} g_X(x) \left( \int_{-\infty}^{\infty} (r(x,y) - r_1(x,y))^2 g_Y(y) \, dy \right) \, dx
\]

To find values for the PL parameters, we solve an optimization problem as follows:

Find \( a_1, b_1, c_1, a_2, b_2, c_2, s_X \) so as to \((4.6)\) subject to:

\[
\begin{align*}
    r_1(3.5, 3) = 10.5, & \quad r_1(s_X, 27) \leq 175.5, \\
    r_1(s_X, 3) \geq 10.5, & \quad r_1(6.5, 27) = 175.5, \\
    3.5 < s_X < 6.5, & \quad a_1, b_1, a_2, b_2 \geq 0.
\end{align*}
\]

Solving the optimization problem in \((4.6)\), we obtain a PL approximation \( r_1 \) as follows:

\[
r_1(X, Y) = \begin{cases} 
    8.27X + 4.26Y - 31.32 & \text{if } X < 5.26 \\
    24.43X + 5.61Y - 134.90 & \text{if } X \geq 5.26
\end{cases}
\]

19
For this optimal solution, the constraint \( r_1(s_X, 3) \geq 10.5 \) is binding, i.e., \( r_1(3.5, 5.26) = 10.5 \). Therefore, the number of free parameters for this optimal solution is 4. The total MSE for \( r_1(X, Y) \) when compared to \( r(X, Y) \) using PDFs \( g_X(\cdot) \) and \( g_Y(\cdot) \) is 11.5121. Since \( W = X \cdot Y \) is two-dimensional, and we have 4 free parameters in the PL approximation \( r_1(X, Y) \), the AIC value is \( AIC(r_1) = 2 \log(11.5121) + 4 = 8.8869 \).

Let \( g_{W_1}\) denote the marginal PDF of \( W \) computed using \( g_X(x) \), \( g_Y(y) \), and \( \delta(w - r_1(x, y)) \). \( g_{W_1}\) is computed as a 19-piece, 7-degree MOP on the domain \((10.5, 175.5)\). A graph of \( g_{W_1}\) overlaid on the graph of \( g_W(\cdot) \) is shown in Fig. 8. The goodness of fit statistics of \( g_{W_1}\) compared to \( g_W \) are shown in the second column in Table 3. Results from additional experiments described subsequently in this section are also displayed in this table.

![Graph of \( g_{W_1}\) and \( g_W \)](image)

One way to reduce the AIC value of a PL approximation is to reduce its number of parameters. Notice that in the solution of the optimization problem \((4.6)\), \( c_1 \approx -a_1 \cdot b_1 \), and \( c_2 \approx -a_2 \cdot b_2 \). Thus, if we add the constraints \( c_1 = -a_1 \cdot b_1 \) and \( c_2 = -a_2 \cdot b_2 \), to the optimization problem in \((4.6)\), we obtain a PL approximation \( r_2 \) as follows:

\[
r_2(X, Y) = \begin{cases} 
3X + 4.69Y - 14.08 & \text{if } X < 5.15 \\
27X + 5.35Y - 144.55 & \text{if } X \geq 5.15 
\end{cases}
\]  \hspace{1cm} (4.8)

A graph of \( r_2(X, Y) \) is shown in Fig. 9 along with the exact function \( r(X, Y) \). As in the case of \( r_1(X, Y) \), the constraint \( r_2(s_X, 3) \geq 10.5 \) is binding. Notice that the approximation \( W = r_2(X, Y) \) has only 2 free parameters (compared to 4 for \( W = r_1(X, Y) \)). The approximation \( W = r_2(X, Y) \) has a MSE of 16.6193, compared to MSE of 11.5121 for \( W = r_1(X, Y) \). The corresponding value of the AIC heuristic is \( AIC(r_2) = 7.6209 \), which is lower than \( AIC(r_1) = 8.8869 \). The AIC value of \( r_2 \) is lower than the AIC value of \( r_1 \) because \( r_2 \) has 2 less parameters than \( r_1 \). Let \( g_{W_2}(\cdot) \) denote the marginal PDF of \( W \) computed using \( g_X(x) \), \( g_Y(y) \), and \( \delta(w - r_2(x, y)) \). \( g_{W_2}(\cdot) \) is computed as a 19-piece, 7-degree MOP on the domain \((10.5, 175.5)\). A graph of \( g_{W_2}(\cdot) \) overlaid on the graph of \( g_W(\cdot) \) is shown in Fig. 10. The goodness of fit statistics of \( g_{W_2}\) compared to \( g_W \) are shown in the third column of Table 3. Comparing these statistics with those obtained without the constraints \( c = -a \cdot b \), we see that even though the MSE of \( r_2 \) is higher, all four goodness of fit statistics for \( g_{W_2} \)
Figure 9: Left: A 3D Plot of $r_2(X,Y)$. Right: A 3D Plot of $r(X,Y) = X \cdot Y$

Figure 10: A graph of $g_{W_2}(\cdot)$ (in red) overlaid on the graph of $g_W(\cdot)$ (in blue)
Table 3: Goodness of Fit and Run Time Results for the Product Function Example.

<table>
<thead>
<tr>
<th>Split on</th>
<th>X</th>
<th>X</th>
<th>X</th>
<th>Y</th>
<th>Y</th>
<th>Y</th>
<th>Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>c = −a · b?</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td># pieces</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td># free parameters</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>3</td>
<td>6</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>MSE</td>
<td>11.5121</td>
<td>16.6193</td>
<td>4.4350</td>
<td>8.9383</td>
<td>14.5624</td>
<td>2.6470</td>
<td><strong>1.10</strong></td>
</tr>
<tr>
<td>KL</td>
<td>0.0095</td>
<td><strong>0.0011</strong></td>
<td>0.0029</td>
<td>0.0369</td>
<td>0.0453</td>
<td>0.0061</td>
<td>0.0032</td>
</tr>
<tr>
<td>MAD (PDF)</td>
<td>0.0021</td>
<td><strong>0.0006</strong></td>
<td>0.0007</td>
<td>0.0067</td>
<td>0.0069</td>
<td>0.0025</td>
<td>0.0015</td>
</tr>
<tr>
<td>AEM</td>
<td>1.0478</td>
<td>0.6677</td>
<td>0.0919</td>
<td>0.5660</td>
<td><strong>0.0000</strong></td>
<td><strong>0.0000</strong></td>
<td><strong>0.0000</strong></td>
</tr>
<tr>
<td>AEV</td>
<td>48.9341</td>
<td><strong>0.4223</strong></td>
<td>34.3644</td>
<td>11.1167</td>
<td>11.6572</td>
<td>1.1496</td>
<td>0.6330</td>
</tr>
<tr>
<td>CPU (in secs)</td>
<td>17.86</td>
<td>21.08</td>
<td>25.71</td>
<td><strong>17.04</strong></td>
<td>20.14</td>
<td>25.65</td>
<td>32.12</td>
</tr>
</tbody>
</table>

(computed using \( r_2 \)) are better than the corresponding ones for \( g_{W_1} \) (computed using \( r_1 \)). This is consistent with the fact that the \( AIC(r_2) = 7.6209 \) is lower than \( AIC(r_1) = 8.8869 \).

In a similar manner, we can find a 3-piece PL linear approximation (with the assumption than \( c = −ab \)) as follows:

\[
r_3(X, Y) = \begin{cases} 
  a_1X + b_1Y - a_1b_1 & \text{if } X < s_{X_1} \\
  a_2X + b_2Y - a_2b_2 & \text{if } s_{X_1} \leq X < s_{X_2} \\
  a_3X + b_3Y - a_3b_3 & \text{if } X \geq s_{X_2}
\end{cases}
\]

This 3-piece PL approximation has 8 parameters \((a_1, b_1, a_2, b_2, a_3, b_3, s_{X_1}, \text{and } s_{X_2})\). In finding an optimal solution that satisfies the domain principle, the number of free parameters is reduced to 5 because of the equality constraints \( r_3(3.5, 3) = 10.5 \), \( r_3(s_{X_1}, 3) = 10.5 \), and \( r_3(6.5, 27) = 175.5 \). Even though the \( MSE(r_3) = 4.4350 \) is lower than the \( MSE(r_2) = 16.6193 \), \( AIC(r_3) = 7.9791 \) is higher than \( AIC(r_2) = 7.6209 \) because \( r_3 \) has 3 more free parameters than \( r_2 \).

In the 2- and 3-piece solutions described above, we split the joint domain of \((X, Y)\) on \( X \) \((x < s_X, x \geq s_X, \text{etc.})\). We can also split the domain on \( Y \). The goodness of fit statistics for the case when we split on \( Y \) are not as good as compared to when we split on \( X \). We do not know why. We conjecture that in the case of the product function, splitting on \( X \) provides better results because \( X \) has a smaller variance than \( Y \). When we split on \( Y \), none of the inequality domain constraints are binding. Thus, a 2-piece PL approximation when we split on \( Y \) has 1 more free parameter compared to when we split on \( X \). The goodness of fit statistics when we split on \( Y \) are shown in Table 3. Thus, the PL approximation suggested by the AIC-like heuristic is the 2-piece PL approximation described in Eq. (4.8), where we assume \( c = −ab \), and that has the best goodness of fits statistics for \( KL, MAD(PDF) \), and \( AEV \).
4.2.2 The Quotient Function $W = 3X/Y$

In this subsection, we will consider the problem $X \sim \chi^2(5)$, $Y \sim \chi^2(15)$, $X$ and $Y$ are independent, and $W = r(X,Y) = \frac{3X}{Y}$. The exact marginal distribution of $W$ is $F(5,15)$, where $F(n,d)$ denotes the $F$-distribution with $n$ numerator, and $d$ denominator, degrees of freedom.

In [22], it is claimed that MOPs are closed under transformations needed for quotient functions. But this claim is incorrect. If $X$ and $Y$ are independent, $f_X(x)$ denotes the PDF of $X$, $f_Y(y)$ denotes the PDF of $Y$, and $W = X/Y$, then the PDF $f_W(w)$ is given by

$$f_W(w) = \int_{-\infty}^{\infty} |y| f_X(wy) f_Y(y) \, dy$$

(4.9)

Although $f_X(wy)$, is a polynomial function of $y$ and $w$, it is not a MOP because it is defined on regions such as $a < wy < b$, which is not a hyper-rhombus. Thus, $f_X(wy)$ is not a MOP. Consequently, $f_W(w)$ may not be a MOP.

The 0.5 and 99.5 percentiles of $X$ are 0.41, and 16.75, respectively, and we approximate the PDF $g_X(x)$ of $X$ on this domain. Similarly, the 0.5 and 99.5 percentiles of $Y$ are 4.60, and 32.80, respectively, and we approximate the PDF $g_Y(y)$ of $Y$ on this domain. We will describe a PL approximation of $r(X,Y)$ on the joint domain $(0.41,16.75) \times (4.60,32.80)$. Notice that the minimum value of $W$ on the joint domain is $\frac{3(0.41)}{32.80} = 0.04$, and the maximum value is $\frac{3(16.75)}{4.60} = 10.92$. Thus, as per the domain principle, we will find an approximation of the PDF of $W$ on the domain $(0.04,10.92)$.

Consider a 2-piece PL approximation $r_1(X,Y)$ of $r(X,Y) = 3X/Y$ where we split the domain of $(X,Y)$ on $Y$ as follows:

$$r_1(X,Y) = \begin{cases} a_1X + b_1Y + c_1 & \text{if } Y < s_Y \\ a_2X + b_2Y + c_2 & \text{if } Y \geq s_Y \end{cases}$$

(4.10)

The PL approximation $r_1(X,Y)$ has 7 parameters $(a_1, b_1, c_1, a_2, b_2, c_2, s_Y)$. We find the values of these parameters to minimize the MSE of $r_1(X,Y)$ as compared to $r(X,Y) = 3X/Y$. To satisfy the domain principle, we impose equality constraints $r_1(0.41,32.80) = 0.04$, and $r_1(16.75,4.60) = 10.92$. The objective function to be minimized is defined as follows:

$$MSE(a_1, b_1, c_1, a_2, b_2, c_2, s_Y) = \int_{-\infty}^{\infty} g_Y(y) \left( \int_{-\infty}^{\infty} (r(x,y) - r_1(x,y))^2 g_X(x) \, dx \right) \, dy$$

To find values for the PL approximation parameters, we solve an optimization problem as follows:

Find $s_Y, a_1, b_1, c_1, a_2, b_2, c_2$ so as to

$$\text{Minimize } MSE(a_1, b_1, c_1, a_2, b_2, c_2, s_Y)$$

subject to:

$\begin{align*}
r_1(0.41,32.80) &= 0.04, \\
r_1(16.75,s_Y) &\leq 10.92, \\
r_1(0.41,s_Y) &\geq 0.04, \\
r_1(16.75,4.60) &= 10.92, \\
4.60 &\leq s_Y \leq 32.80,
\end{align*}$

and $a_1, a_2 \geq 0, b_1, b_2 \leq 0$. 

(4.11)
The constraints ensure that the domain principle is satisfied. The resulting PL approximation is as follows:

\[
  r_1(X, Y) = \begin{cases} 
  0.61X - 0.37Y + 2.41 & \text{if } Y < 7.02 \\
  0.18X - 0.016Y + 0.50 & \text{if } Y \geq 7.02 
  \end{cases} 
\] (4.13)

At the optimal solution, the constraint \( r_1(0.41, s_Y) \geq 0.04 \) is binding. Thus, the number of free parameters for the PL approximation in Eq. (4.13) is 4 \((= 7 - 3\), where 7 is the number of parameters and 3 is the number of equality constraints). The MSE of the the PL approximation \( r_1(X, Y) \) compared to \( r(X, Y) \) (with respect to a 2-piece, 5-degree MOP approximation \( g(x) \) of the PDF of \( X \), and a 2-piece, 4-degree MOP approximation \( h(y) \) of the PDF of \( Y \)) is 0.1600. This function is shown graphically in Fig. 11 along with the actual quotient function. Since there are 4 free parameters, the AIC score for this approximation is 0.3349. The resulting PDF \( k_1(\cdot) \) is shown in Fig. 12 overlaid on the PDF \( k_0(\cdot) \) found by using the actual quotient function \( r(X, Y) \) in combination with the MOP approximations to the \( \chi^2 \) PDFs. The goodness of fit statistics of \( k_1(\cdot) \) compared to \( k_0(\cdot) \) are shown in the second column of Table 4.

![Figure 11: Left: A 3D Plot of \( r_1(X, Y) \). Right: A 3D Plot of \( r(X, Y) \).](image)

![Figure 12: A Graph of \( k_1(w) \) (in red) Overlaid on the Graph of \( k_0(w) \) (in blue).](image)
Table 4: Goodness of Fit and Run Time Results for the Quotient Function Example.

<table>
<thead>
<tr>
<th>Split on</th>
<th>Y</th>
<th>Y</th>
<th>X</th>
<th>X</th>
</tr>
</thead>
<tbody>
<tr>
<td># pieces</td>
<td>2</td>
<td>3</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td># free parameters</td>
<td>4</td>
<td>7</td>
<td>5</td>
<td>9</td>
</tr>
<tr>
<td>MSE</td>
<td>0.1600</td>
<td>0.0460</td>
<td>0.2226</td>
<td>0.1232</td>
</tr>
<tr>
<td>AIC</td>
<td>0.3349</td>
<td>0.8398</td>
<td>1.9957</td>
<td>4.8119</td>
</tr>
<tr>
<td>KL</td>
<td>0.2449</td>
<td>0.0605</td>
<td>0.8071</td>
<td>0.4393</td>
</tr>
<tr>
<td>MAD (PDF)</td>
<td>0.4834</td>
<td>0.3054</td>
<td>0.5292</td>
<td>0.4618</td>
</tr>
<tr>
<td>AEM</td>
<td>0.0951</td>
<td>0.0524</td>
<td>0.0868</td>
<td>0.0559</td>
</tr>
<tr>
<td>AEV</td>
<td>0.1644</td>
<td>0.1021</td>
<td>0.4202</td>
<td>0.2488</td>
</tr>
<tr>
<td>CPU (in secs.)</td>
<td>33.03</td>
<td>46.35</td>
<td>28.15</td>
<td>47.58</td>
</tr>
</tbody>
</table>

Next, we found a 3-piece PL approximation $r_2(X,Y)$ of $r(X,Y) = 3X/Y$ in a similar way that had 7 free parameters (including two optimal split points on $Y$, and 4 equality constraints). The minimum MSE is 0.04600, and the corresponding AIC value is $AIC(r_2) = 0.8398$, which is higher than the AIC value $AIC(r_1) = 0.3349$ of the 2-piece PL approximation described in Eq (4.13).

Next, we repeated the procedure and split the joint domain of $(X,Y)$ on $X$. The results are not as good as when we split on $Y$ (see Table 4). Notice that for this example, the best goodness of fit statistics are obtained by a 3-piece PL approximation where we split on $Y$, the variable in the denominator of the quotient function. The 2-piece PL approximation with the smallest AIC score has decent goodness of fit statistics. It requires less time to compute the marginal of $W$ than the 3-piece split. Based on the results in Table 4, we conjecture that for quotient functions, splitting on the variable in the denominator will get better results than splitting on the variable in the numerator.

Thus, for the quotient example, the 2-piece PL approximation $r_1(X,Y)$ described in Eq. (4.13) is the PL approximation suggested by the AIC-like heuristic.

5 Two Examples

This section describes two examples that include a deterministic variable that is a nonlinear function of its continuous parents. Such examples may arise when constructing Bayesian networks in domains such as business.

5.1 Crop Problem

This example is similar to one used by Lerner [11] and Murphy [14]. The example differs from previous implementations because the continuous variables are assumed to have log-normal distributions (instead of normal). In this model, crop size ($C$) (in million bushels (mB)) produced depends on whether the rain conditions are drought ($R = d$), average ($R = a$), or flood ($R = f$). The price ($P$) of crop (in $$/bushel ($$/B)) is negatively correlated with crop
size \((C)\). Revenue \((V)\) (in million $ (m$)) is a deterministic function of crop size \((C)\) and price \((P)\), i.e. \(V = C \cdot P\).

\[
P(R = d) = 0.35 \\
P(R = a) = 0.60 \\
P(R = f) = 0.05
\]

The Bayesian network and the parameters of the distributions for the variables in the Crop example are shown in Fig. 13. We will describe a MOP solution to the Crop problem.

First we found 2-piece, 5-degree MOP approximations of the PDFs of \(C|d\), \(C|a\), and \(C|f\), which have log-normal distributions with parameters as specified in Fig. 13 using the Lagrange interpolating polynomials with Chebyshev points and the procedure described in [17]. After we marginalize the discrete variable \(R\), we obtain an 8-piece, 5-degree MOP approximation of the mixture PDF for \(C\) as shown in Fig. 14. The expected value and variance of this marginal PDF are 4.15 and 1.80, respectively, which are close to the theoretical expected value of 4.15 and variance of 1.83.

![Figure 13: An Bayesian Network Model for the Crop Example.](image)

![Figure 14: A MOP Approximation of the PDF of \(C\) in the Crop Example.](image)
Next we found a MOP approximation of the conditional PDF of $P|c$ using the mixed tree technique proposed in [13]. We divided the domain of $C$ into 5 equal probability intervals: $(0.93, 2.88)$, $[2.88, 3.63)$, $[3.63, 4.51)$, $[4.51, 5.34)$, $[5.34, 8.88)$. Next, we found a 2-piece, 5-degree MOP approximation of the PDF of $P|c$ at the mid-point of each interval (again using Lagrange interpolating polynomials with Chebyshev points and the procedure described in [17]). Thus, the MOP approximation of the conditional PDF of $P|c$ has 10 pieces, and 5 degrees. A 3D plot of the MOP approximation of the conditional PDF of $P|c$ is shown in Fig. 15 along with a 3D plot of the exact PDF of $P|c$.

![Figure 15: A 3D Plot of the Conditional PDF of $P|c$ (left) and its MOP Approximation (right).](image)

If we compute the marginal PDF of $P$ using the MOP approximation of the marginal PDF of $C$ and the MOP approximation of the conditional PDF of $P|c$, we obtain a 13-piece, 5-degree MOP. A plot of this MOP is shown in the left side of Fig. 16. The expected value and variance of the MOP approximation of the marginal PDF of $P$ are 6.77 and 3.73, respectively. A simulation with 5,000,000 trials produced estimates for the mean and variance of 6.78 and 2.53, respectively. A histogram for $P$ from this simulation is shown in the right side of Fig. 16.

![Figure 16: Left: A MOP Approximation of the Marginal PDF of $P$. Right: An Histogram for $P$ Found Using Monte Carlo Simulation.](image)

Next, we found a 2-piece PL approximation $r_1(C, P)$ of the deterministic function as-
associated with $V$ using the constraint $c = -a \cdot b$ as described in Section 4.2.1. Consistent with the previous examples, we split the domain of the variable $C$ because its marginal distribution has a smaller variance (1.80) than the marginal distribution of $P$ (3.73). The optimal split point was very close to the upper bound of the third region of the conditional PDF of $P|c$. To minimize the AIC score, we decided to use the upper bound of the third region 4.51 as the split point (instead of the optimal split point). The total MSE of the 2-piece PL approximation with 4.51 as the split point is 0.91. There are 2 free parameters in the 2-piece PL approximation (4 parameters $- 2$ equality constraints to satisfy the domain principle). Thus, the AIC value for this PL approximation is 1.81. The details of the 2-piece PL approximation are as follows:

$$r_1(C, P) = \begin{cases} 6.96C + 2.61P - 18.14 & \text{if } C < 4.51 \\ 5.56C + 5.91P - 32.90 & \text{if } C \geq 4.51 \end{cases} \quad (5.1)$$

Finally, using the computed MOP approximation of the marginal PDF of $C$, the fitted MOP approximation of the conditional PDF of $P|c$, and the PL approximation $V = r_1(C, P)$, we compute the marginal PDF of $V$, which is computed as a 50-piece, 11-degree MOP on the domain (6.45, 49.41). Computing the marginal PDF of $V$ required 112 seconds of computing time. A plot of the MOP approximation of the marginal PDF of $V$ is shown in Fig. 17 overlaid on the marginal PDF of $V$ using the exact nonlinear function $V = C \cdot P$ (which is not a MOP, but can be computed in Mathematica). The expected value and variance of the PDF of $V$ computed using the 2-piece PL approximation are 25.97 and 16.28, respectively, compared to corresponding values of 25.77 and 16.74, respectively, from using the exact nonlinear function $V = C \cdot P$. The measures of accuracy between these two PDFs are as follows: $KL \approx 0.0068$, $MAD$ (of PDF) $\approx 0.0018$, $AEM \approx 0.1995$, and $AEV \approx 0.4645$.

Figure 17: A MOP Approximation of the Marginal PDF of $V$ Using a 2-piece PL Approximation of $V = C \cdot P$ (in red) Overlaid on the Marginal PDF of $V$ Using $V = C \cdot P$ (in blue)

Consistent with our previous examples, using more pieces in the PL approximation does not improve the accuracy of the marginal distribution for $V$. For instance, a 5-piece PL approximation, one for each of the five regions of the PDF of $P|c$, has an MSE of 0.4307 but an AIC score of 6.3153. The marginal distribution is indistinguishable graphically from the
one displayed in Fig. 17 and has a less accurate mean of 25.77 and a lower variance of 16.11. With the 5-piece PL approximation, 151 seconds of computing time were required to obtain the marginal distribution.

We also solved this example using MTE functions with the conditional distribution for \( P | c \) defined using the mixed tree structure. The accuracy of the results is similar to the MOP solution, so the details are omitted. To implement the MTE solution, we need to recover the hypercube condition in the MTE function resulting from the calculation of the marginal for \( P \). This is done using the method described in Section 3.2.

### 5.2 Foreign Security Valuation

The value \((V)\) in U.S. dollars of an investment in a share of foreign stock in one month is a function of the price of the stock \((P)\) and the foreign exchange rate \((X)\). In this case, we define the exchange rate as the cost of one U.S. dollar in units of the foreign currency, so the value of the investment in U.S. dollars is determined as \(V = P/X\). In this example, we will use MOPs to find an approximation to the PDF for \( V \). We do not present any MTE results for this example because we could not obtain a PL approximation to the function \( V = P/X \) when we used MTEs to approximate the PDFs of \( P \) and \( X \).

In this example, we assume that the PDF of \( X \) is a beta distribution with parameters \( \alpha = 3 \) and \( \beta = 2 \), on the domain \((0.5, 1.2)\) as follows:

\[
f_X(x) = \begin{cases} 
34.9854(x - 0.5)^2(1.2 - x) & \text{if } 0.5 < x < 1.2 \\
0 & \text{otherwise}
\end{cases}
\]

The PDF of \( X \) is a MOP function. A graph of the PDF is shown in Fig. 18.

![Figure 18: A Graph of the PDF for Euro/U.S. Exchange Rates.](image)

Suppose we want to invest in one share of stock in the firm Deutsche Post (DHL) which is listed on the DAX stock exchange and priced in Euros. The price of the stock on March 5, 2012 was 13.43 Euros. Assuming the stock price follows a geometric Brownian motion (GBM) stochastic process, the terminal price after one year has a lognormal PDF. Using five years of daily stock prices, we estimate the GBM\((\mu, \sigma^2)\) parameters as \( \hat{\mu} = -0.04258 \) and \( \hat{\sigma}^2 = 0.34765^2 \). If the return over the one-year period, \( R \), is defined such that \( P = 13.43R \), \( R \)
is distributed as \( R \sim LN(\hat{\mu} - \hat{\sigma}^2/2, \hat{\sigma}^2) \), or \( R \sim LN(-0.10301, 0.34765^2) \). The distribution for \( P = 13.43R \) is then defined as \( P \sim LN(-0.10301 + \ln(13.43), 0.34765^2) \).

The PDF for \( P \) is approximated by a 2-piece, 6-degree MOP. The pieces of this distribution are defined on the intervals \((4.27, 10.74), [10.74, 34.38]\), where 10.74 is the mode of the lognormal PDF for \( P \). The MOP distribution is shown in Fig. 19 overlaid on the actual lognormal PDF.

As in the example in Section 4.2.2, we computed a 2-piece PL approximation \( r_1(P, X) \) by computing an optimal point to divide the domain of \( X \) that minimizes the total MSE of the PL approximation. The minimum MSE of the 2-piece PL approximation \( r_1(P, X) \) (compared to the exact nonlinear function \( r(P, X) = P/X \)) is 1.67. The 2-piece PL approximation has 5 free parameters, and therefore, an AIC value of 6.03. The 2-piece PL approximation is as follows:

\[
r_1(P, X) = \begin{cases} 
1.81P - 59.89X + 36.60 & \text{if } X < 0.68 \\
0.95P - 9.86X + 11.34 & \text{if } X \geq 0.68
\end{cases}
\]  \hspace{1cm} (5.2)

Using this 2-piece PL approximation and the MOP approximation of the PDF of \( P \) to find the marginal PDF for \( V \) results in a 15-piece, 10-degree MOP function. This PDF is shown in Fig. 20 overlaid on the PDF of \( V \) computed by using the MOP function for \( P \) in combination with the actual function \( V = P/X \). The measures of accuracy between these two PDFs are displayed in Table 5.

The expected value and variance of the marginal PDF of \( V \) are 14.72 and 25.71, respectively. Using the nonlinear function \( V = P/X \) with the MOP approximation of the PDF of \( P \), the marginal PDF of \( V \) has an expected value of 14.33 and a variance of 31.80.

For a 3-piece PL approximation, the minimum MSE is 0.36. However, with 9 free parameters, it has a higher AIC value of 6.94. Thus, the AIC heuristic suggests using the 2-piece PL approximation. Using the 3-piece PL approximation (and the MOP PDFs of \( P \) and \( X \)) results in the marginal PDF of \( V \) that is a 19-piece, 10-degree MOP. The goodness of fit statistics for the 3-piece PL approximation are shown alongside those of the 2-piece approximation in Table 5. The goodness of fit statistics improve by adding an additional piece,
but the AIC-like heuristic leads us to conclude that the 2-piece approximation is adequate, particularly given the increase in computation time that occurs when adding the third piece.

6 Summary and Conclusions

This paper is concerned with inference in hybrid Bayesian networks containing nonlinear deterministic conditionals using MTEs and MOPs. Neither MTEs nor MOPs are closed under operations needed for nonlinear deterministic conditionals. Earlier, Cobb and Shenoy [2] suggest approximating nonlinear deterministic conditionals by PL ones. However, there are many ways of doing such approximations, and a very naïve strategy was used in [2].

In this paper, we describe a principled approach to finding PL approximations of nonlinear functions. Two basic principles are the domain principle, and the symmetry principle. The domain principle states that a PL approximation should be such that the resulting domain of marginal PDF of the deterministic variable should be exactly the same as in the nonlinear case, and the symmetry principle states that a PL approximation should retain symmetry of
the nonlinear function and the symmetry of the PDFs of the parent variables, if any. Also, a simple AIC-like heuristic for finding a PL approximation is described.

Using these principles and heuristic, PL approximations of some commonly used nonlinear functions are computed. In the one-dimensional case, this include the quadratic function $Y = X^2$, and the exponential function $Y = e^X$. In the two-dimensional case, we examine the cases of the product function $W = X \cdot Y$, and the quotient function $W = 3X/Y$. For all of these nonlinear functions, we compute the marginal of the variable with the nonlinear deterministic conditional using PL approximations, and compare it with the marginal found using the exact nonlinear function, and compute the errors in the marginals using both MTE and MOP approximations of PDFs. While both the MTE and MOP methods result in marginals that are approximately the same, there are some differences.

First, the MTE approach is not closed under PL deterministic conditionals since we lose the hypercube property of MTE functions. However, by using the mixed tree approximation proposed by Moral et al. [13], we can recover the hypercube property. MOPs defined on hyper-rhombuses are closed under PL deterministic conditionals [17]. Second, in finding a PL approximation to minimize AIC with a fixed number of parameters, we need to solve an optimization problem. Such optimization problems are more easily solved using MOP approximations of PDFs than using MTE approximations. In some cases, we were unable to solve the MSE optimization problem when we were using MTE approximations of the conditional PDFs of the parent variables, while we were able to solve the same optimization problems using MOP approximations.

The AIC heuristic is not perfect. For the product function example, the AIC heuristic leads to a 2-piece PL approximation that has the smallest $KL, MAD(PDF)$, and $AEV$ statistics. However, for the quotient function example, it leads to a 2-piece PL approximation, whereas the 3-piece PL approximation has the best goodness of fit statistics. In most cases, the AIC heuristic leads to PL approximations that have few pieces (two in all the examples we did), requires less time to compute the marginal PDF of the variable with the deterministic conditional, and produces approximations of the marginal PDF with the least number of pieces.

Finally, we use our methods to solve two small hybrid Bayesian networks that contain nonlinear deterministic conditionals. The first one, called the Crop problem was first described by Murphy [14] and contains a product function. The second one, called Foreign Security Valuation, contains a quotient function. In both cases, we find the marginal PDF of the variable of interest, and compare it with the PDF obtained using the exact nonlinear function. Of course, the exact PDFs are not MOPs, and there are no guarantees that they can be used for further computation.

Acknowledgments

A small portion of this paper has been published in [3].
References


