To appear in *International Journal of Intelligent Systems*, 2017, DOI: **10.1002/int.21897**

Inference in Hybrid Bayesian Networks with Nonlinear Deterministic Conditionals*

Barry R. Cobb[1]
BarryCobb@MissouriState.edu

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

[1]Department of Marketing
Missouri State University
Springfield, MO 65897 USA

[2]University of Kansas School of Business
1654 Naismith Dr., Capitol Federal Hall
Lawrence, KS 66045 USA

February 7, 2017

*A small portion of this paper appeared in [3].

Contents

1 Introduction **3**

2 Definitions **5**
2.1 Extended Shenoy-Shafer Architecture 5
2.2 Mixtures of Polynomials 6
2.3 Quality of MOP Approximations 7

3 Piecewise Linear Approximations of Nonlinear Functions **8**
3.1 One-Dimensional Functions 9
3.1.1 The Quadratic Function $Y = X^2$ 9
3.1.2 The Exponential Function $Y = e^X$ 12
3.1.3 Approximation Algorithm 14
3.2 Multi-Dimensional Functions 14
3.2.1 The Product Function $W = X \cdot Y$ 15
3.2.2 The Quotient Function $W = 3X/Y$ 17

4 Two Examples **20**
4.1 Crop Problem 20
4.2 Risk of Fatal Crashes 22

5 Summary and Conclusions **23**

Inference in Hybrid Bayesian Networks with Nonlinear Deterministic Conditionals

Abstract

To enable inference in hybrid Bayesian networks containing nonlinear deterministic conditional distributions, 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 a penalized MSE heuristic, which consists of minimizing a penalized MSE function subject to two principles, domain and symmetry. 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: deterministic variable, hybrid Bayesian networks, mixtures of polynomials, deterministic conditional distributions, nonlinear functions, penalized mean square error heuristic

1 Introduction

The primary goal of this paper is to describe a method for computing marginal distributions in hybrid Bayesian networks (BNs) containing nonlinear deterministic conditionals for some continuous variables. To achieve this goal, we approximate nonlinear functions with piecewise linear functions, then use the results in combination with mixtures of polynomial functions to perform inference. Hybrid BNs are 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 BNs was proposed by Lauritzen and Jensen [10, 11] 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.

Another method for dealing with the integration problem is the mixture of truncated exponentials (MTE) technique proposed by Moral et al. [13]. 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 [4, 8]. The MTE method cannot be used directly in hybrid BNs containing nonlinear conditionals. However, by approximating nonlinear functions by piecewise linear (PL) ones, the MTE method can be used for hybrid BNs containing nonlinear deterministic conditionals [2]. However, using MTE functions to perform the required operations with PL functions requires some manipulation of the results to ensure the resulting function remains an MTE potential.

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. MOP approximations of PDFs can be easily found by using Lagrange interpolating polynomials with Chebyshev points [18]. This method can also be used with two-dimensional conditional PDFs. Also, MOP functions are naturally closed under transformations for multi-dimensional linear deterministic functions, e.g., $W = X + Y$, etc. Langseth et al. [9] introduced mixtures of truncated basis functions (MOTBFs) as a generalization of the MTE and MOP frameworks that provide a flexible tradeoff between accuracy and complexity when approximating empirical PDFs.

In this paper, we demonstrate the use of MOP functions for BNs with nonlinear deterministic conditionals and improve on the approach previously used to find PL linear approximations to nonlinear functions [2]. We propose a method for finding PL approximations of nonlinear functions based on a penalized MSE heuristic, which consists of minimizing a penalized mean square error function subject to two basic principles. 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 $W = X/Y$). For reasons of space, we limit our approximations of the PDFs in the optimization and inference problems to MOP functions. Similar ideas can be applied to MTE or MOTBF functions.

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 and define mixtures of polynomial functions. Also, we describe some numerical measures of goodness of an approximation of a PDF/CDF. In Section 3, 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 4, we describe two examples of hybrid BNs containing nonlinear deterministic conditionals. Finally, in Section 5, we summarize our contributions and describe some issues for further research.

2 Definitions

In this section, we briefly sketch the extended Shenoy-Shafer architecture [21] for inference in hybrid BNs containing deterministic conditionals, define mixture of polynomial functions, and discuss some methods for measuring the quality of MOP approximations to PDFs.

2.1 Extended Shenoy-Shafer Architecture

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 δ 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 are 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 are the same as the units of the potential being marginalized. Marginalizing a continuous variable from a potential is achieved by integrating the potential over the state space of the continuous variable. If the potential being marginalized does not contain Dirac delta functions, 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 consist 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 [17], and then propagate messages in the binary join tree resulting in the marginals of variables of interest.

2.2 Mixtures of Polynomials

The definition of mixture of polynomials given here is taken from [18]. 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:

$$f(x) = \begin{cases} a_{0i} + a_{1i}x + \cdots + a_{ni}x^n & \text{for } x \in A_i, i = 1, \ldots, k, \\ 0 & \text{otherwise.} \end{cases} \tag{2.1}$$

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:

$$g_1(x) = \begin{cases} 0.424 + 0.128x - 0.085x^2 - 0.028x^3 & \text{if } -3 < x < 0, \\ 0.424 - 0.128x - 0.085x^2 + 0.028x^3 & \text{if } 0 \leq x < 3 \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

$g_1(\cdot)$ 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 [18]. ■

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 BNs. The requirement that each piece is defined on an interval A_i is also designed to ease the burden of integrating MOP functions. Pieces of a MOP function can also be defined on a region called a *hyper-rhombus* that includes a linear function of one or more variables [18]. In general, integration of hyper-rhombus MOP functions is slower than hypercube MOP functions, so we utilize the hypercube formulation exclusively in this paper.

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 BNs. They are also closed under transformations needed for linear deterministic functions. We will illustrate this by a small example.

Example 2. 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. (2.2)). Then $\xi(x) = g_1(x - 3)$ is a MOP approximation of the PDF of X, and $\psi(x, y)$ as defined in Eq. (2.3) is a MOP approximation of the conditional PDF of $Y \mid x$.

$$\psi(x, y) = \frac{g_1(\frac{y-6-2x}{2})}{2} \tag{2.3}$$

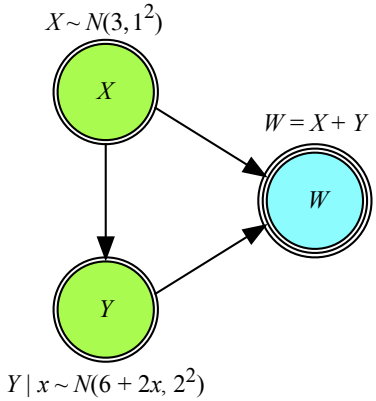

Figure 1: A BN with a Sum Deterministic Conditional.

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

$$\begin{aligned}((\psi \otimes \omega)^{-Y})(x, w) &= \int_{-\infty}^{\infty} \psi(x, y)\, \omega(x, y, w)\ \mathrm{d}y \\ &= \int_{-\infty}^{\infty} \psi(x, y)\, \delta(w - x - y)\, \mathrm{d}y \\ &= \psi(x, w - x) \end{aligned} \tag{2.4}$$

The result in Eq. (2.4) follows from the sampling property of Dirac delta function: If f is continuous in a neighborhood of a, then

$$\int_{-\infty}^{\infty} f(x)\, \delta(x - a)\, \mathrm{d}x = f(a) \tag{2.5}$$

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

$$((\xi \otimes (\psi \otimes \omega)^{-Y})^{-X})(w) = \int_{-\infty}^{\infty} \xi(x)\, \psi(x, w - x)\, \mathrm{d}x \tag{2.6}$$

Since $\xi(x)$, and $\psi(x, w - x)$ are MOPs, the marginal distribution of W computed in Eq. (2.6) is a MOP. ■

2.3 Quality of MOP Approximations

In this section, we discuss some quantitative ways to measure the accuracy of a MOP 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 of the mean, and absolute error of 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 ([7]):

$$KL(f, g) = \int_a^b \ln\left(\frac{f(x)}{g(x)}\right) f(x)\, dx. \tag{2.7}$$

$KL(f, g) \geq 0$, and $KL(f, g) = 0$ if and only if $g(x) = f(x)$ for all $x \in (a, b)$. The statistic $KL(f, g)$ has information theory semantics as the # bits needed to encode g if it is optimized to encode f. It has been shown in [23] that for any measurable event, the error in its probability (by using PDF g instead of f) is bounded by $\sqrt{KL(f, g)/2}$.

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

$$MAD(f, g) = sup\{|f(x) - g(x)| : a < x < b\} \tag{2.8}$$

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 $\int_c^d g(x)\, \mathrm{d}x$, 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

$$MAD(F, G) = sup\{|F(x) - G(x)| : a < x < b\} \tag{2.9}$$

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:

$$AEM(f, g) = |E(f) - E(g)| \tag{2.10}$$

$$AEV(f, g) = |V(f) - V(g)| \tag{2.11}$$

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. (2.2). 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: $KL(f, g_1) \approx 0.0051$, $MAD(f, g_1) \approx 0.0248$, $MAD(F, G_1) \approx 0.0028$, $AEM(f, g_1) \approx 0.0000$, $AEV(f, g_1) \approx 0.0239$.

3 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

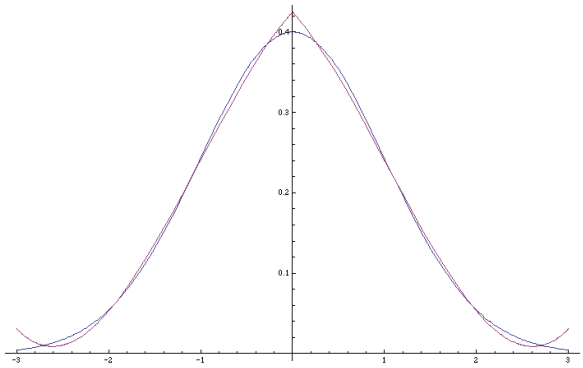

Figure 2: A Graph of $g_1(z)$ (in red) Overlaid on the Graph of $f(z)$ (in blue).

needed for linear deterministic functions.

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.

3.1 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.

3.1.1 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. (2.2) on the domain $(-3, 3)$ for the MOP 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. By domain, we mean the region of the real line where the PDF is positive. We will refer to this principle as the *domain* principle. Given a random variable, we have two questions: 1) What are its possible values? 2) What are the probabilities (or densities) of these values? It is clear that the first question is more fundamental than the second. It makes little sense to compare two PDFs that do not agree on the domain. The domain principle requires that in finding a PL approximation, the domain of the deterministic variable should not be changed. 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 $(0, 9)$.

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 approxima-

tion retains the symmetry. We will refer to this principle as the *symmetry* principle. 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. The main idea behind the symmetry principle is to reduce the # parameters required to describe a PL approximation. As we will see subsequently, it is important to keep the # parameters small in finding a PL approximation. Exploiting any symmetry is one way to do so. 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 $X = 0$.

Penalized MSE Heuristic In the statistics literature, there are several heuristics (such as Akaike's information criterion (AIC) [1] and Bayes information criterion (BIC) [16]) 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 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 $\widehat{\sigma}^2$ is an estimate of σ^2 in the regression model, the AIC heuristic is to minimize $n \times \ln(\widehat{\sigma}^2) + 2p$, 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 are 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 pieces (and therefore more parameters) may lead to a worse result for the marginal distribution of the target variable. One reason for this is that we are approximating PDFs by MOPs, and having a lot of pieces in a PL approximation means a lot of parameters, and the marginal of the target variable is a convoluted sum of the pieces, and the errors do not always cancel out. In our empirical tests, we have observed that increasing the # pieces decreases errors in the marginal up to a point, and beyond this point, increasing # pieces causes errors in the marginal to increase. Also, making inferences with MOPs/MTEs that have many pieces can be intractable [19]. For these two reasons, 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 nonlinear 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

$$MSE(r_1) = \int_{-\infty}^{\infty} f_X(x) \left(r(x) - r_1(x)\right)^2 \mathrm{d}x. \tag{3.1}$$

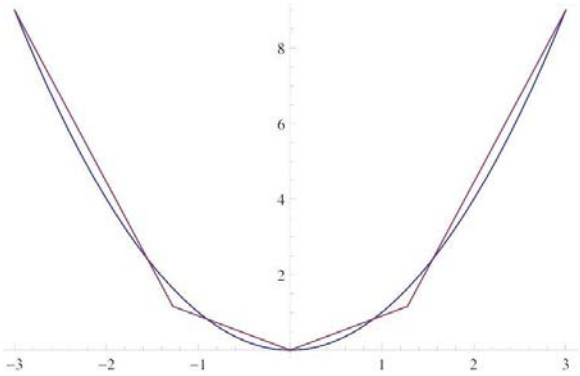

Figure 3: A Graph of $Y = r_1(X)$ (in red) Overlaid on the Graph of $Y = X^2$ (in blue).

The penalized MSE heuristic finds a PL approximation $Y = r_1(X)$ with p free parameters such that the $pMSE(r_1) = \ln(MSE(r_1)) + p$ 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 penalized MSE heuristic as follows:

$$pMSE(r_1) = d\ \ln(MSE(r_1)) + p \tag{3.2}$$

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

The PL approximation $Y = r_1(X)$ is as follows (see Fig. 3):

$$Y = \begin{cases} -4.66 - 4.55X & \text{if } X < -1.28 \\ -0.91X & \text{if } -1.28 \leq X < 0 \\ 0.91X & \text{if } 0 \leq X < 1.28 \\ -4.66 + 4.55X & \text{if } X \geq 1.28 \end{cases} \tag{3.3}$$

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 $pMSE(r_2)$ is -1.1235, which is higher than $pMSE(r_1)$. Similarly, if we use an 8-piece approximation (with 6 free

Table 1: Goodness of Fit and Run Time Results for the Quadratic Function Example.

# pieces	4	6	8
# free parameters	**2**	4	6
MSE	0.0433	0.0060	**0.0016**
$pMSE$	**−1.1405**	−1.1235	−0.4205
KL	0.2498	0.1532	**0.1098**
MAD of PDF	35.3322	35.0924	**34.7124**
MAD of CDF	0.1639	0.1332	**0.1010**
AEM	**0.0589**	0.1886	0.1683
AEV	**0.0651**	0.1055	0.1864
CPU (in secs.)	**0.50**	0.75	1.05

parameters), then the value of $MSE(r_3)$ is 0.0016, and the value of $pMSE(r_3)$ is -0.4205, which is higher than $pMSE(r_1)$ and $pMSE(r_2)$. Thus, the penalized MSE heuristic suggests a 4-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 CPU row gives the run time required by Mathematica 10.1 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 g_1 and $Y = X^2$. One reason for doing this is to capture the errors caused by the PL approximation without confounding by the errors caused by using a MOP approximation g_1 of the PDF of X.

In Table 1, notice that the MSE decreases monotonically with the # of pieces. As the # free parameters increases, the penalized MSE score will eventually increase after bottoming out at some point (here at # pieces = 4). We report the results for the various measures of accuracy. The minimum values of KL divergence, MAD of PDF, and MAD of CDF are attained for # of pieces = 8. The minimum values of AEM and AEV are attained at # pieces = 4. The CPU time is monotonic in the # pieces.

Let $g_2(\cdot)$ denote the marginal PDF of Y computed using $g_1(\cdot)$ and $Y = X^2$ (the CDF corresponding to this PDF is denoted by G_2). Let $g_{11}(\cdot)$ denote the marginal PDF of Y using $g_1(\cdot)$ and $Y = r_1(X)$. Let $G_{11}(\cdot)$ denote the CDF corresponding to PDF $g_{11}(\cdot)$. A graph of $G_{11}(\cdot)$ overlaid on the graph of $G_2(\cdot)$ is shown in Fig. 4.

3.1.2 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

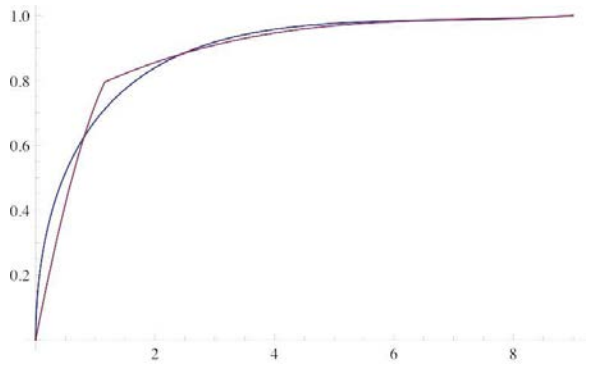

Figure 4: A Graph of G_{11} (in red) Overlaid on the Graph of G_2 (in blue).

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:

$$\begin{gathered} \text{Find } a_1, b_1 \text{ so as to} \\ Minimize \int_{-\infty}^{\infty} g_1(x)(r(x) - r_1(x))^2 \, \mathrm{d}x \\ \text{subject to} : -3 < a_1 < 3, \text{and} \\ e^{-3} < b_1 < e^3. \end{gathered} \tag{3.4}$$

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. 5. The minimum value of MSE is 0.4566, and the corresponding value of pMSE 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. 6 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 pMSE 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 PDF of Y, which may explain why the CPU time is lowest for the case of the 2-piece PL approximation.

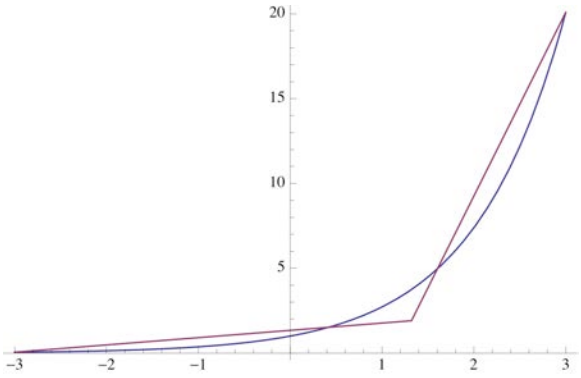

Figure 5: A Graph of $Y = r_1(X)$ (in red) Overlaid on the Graph of $Y = e^X$ (in blue).

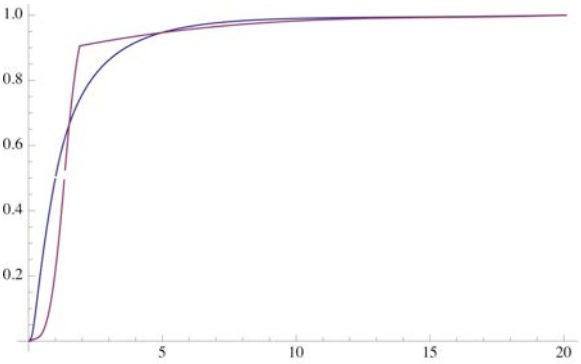

Figure 6: A Graph of the CDF of Y using $Y = r_1(X)$ (in red) Overlaid on the Graph of the CDF of Y using $Y = e^X$ (in blue).

3.1.3 Approximation Algorithm

An algorithm for approximating a one-dimensional nonlinear function with a PL functions is summarized in Figure 7. The algorithm begins by establishing a set of knots that define the PL function. This set includes the endpoints of the domain of the independent variable and the decision variables determined in the optimization process. If the nonlinear function is symmetric, this set is expanded to include the points where the result of the function $r(x)$ is identical on the opposite of the point of symmetry x_s. The process is repeated until the pMSE value is no longer improved.

3.2 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.

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

# pieces	2	3	4
# free parameters	2	4	6
MSE	0.4566	0.0623	**0.0311**
$pMSE$	**1.2161**	1.2239	2.5298
KL	0.8437	0.3747	**0.1411**
MAD of PDF	**0.6233**	1.1312	2.3453
MAD of CDF	0.3035	0.1484	**0.0442**
AEM	0.1429	0.0305	**0.0163**
AEV	0.3164	**0.1073**	0.1182
CPU (in secs.)	**0.16**	0.20	0.23

3.2.1 The Product Function $W = X \cdot Y$

Consider a BN: $X \sim N(5, 0.5^2)$, $Y \sim N(15, 4^2)$, X and Y are independent, and $W = r(X, Y) = X \cdot Y$. We have a 2-piece, 3-degree MOP $g_X(x) = g_1(\frac{x-5}{0.5})/0.5$ of the PDF of X on the domain $(3.5, 6.5)$, and a 2-piece, 3-degree MOP $g_Y(y) = g_1(\frac{y-15}{4})/4$ 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. 2.2).

Using these two MOP approximations of the PDFs of X and Y, we can find an "exact" marginal PDF of W as follows:

$$g_W(w) = \int_{-\infty}^{\infty} g_X(x) \left(\int_{-\infty}^{\infty} g_Y(y)\, \delta(w - x \cdot y)\, \mathrm{d}y \right) \mathrm{d}x \tag{3.5}$$

$g_W(\cdot)$ 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.

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, 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 in this case.

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

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

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

$$\begin{aligned} & \text{Find } a_1, b_1, c_1, a_2, b_2, c_2, s_X \text{ so as to} \\ & \quad Minimize \quad MSE(a_1, b_1, c_1, a_2, b_2, c_2, s_X) \\ \text{subject to :} \quad & r_1(3.5, 3) = 10.5, r_1(s_X, 27) \leq 175.5, \\ & r_1(s_X, 3) \geq 10.5, r_1(6.5, 27) = 175.5, \\ & 3.5 < s_X < 6.5, \\ \text{and} \quad & a_1, b_1, a_2, b_2 \geq 0. \end{aligned} \tag{3.6}$$

Solving the optimization problem in (3.6), we obtain a PL approximation r_1 as follows:

$$r_1(X, Y) = \begin{cases} 8.27X + 4.26Y - 31.32 & \text{if } X < 5.26 \\ 24.43X + 5.61Y - 134.90 & \text{if } X \geq 5.26 \end{cases} \tag{3.7}$$

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 pMSE value is $pMSE(r_1) = 2 \log(11.5121) + 4 = 8.8869$.

Let $g_{W_1}(\cdot)$ denote the marginal PDF of W computed using $g_X(x)$, $g_Y(y)$, and $\delta(w - r_1(x, y))$. $g_{W_1}(\cdot)$ is computed as a 19-piece, 7-degree MOP on the domain $(10.5, 175.5)$. A graph of $g_{W_1}(\cdot)$ 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.

One way to reduce the pMSE value of a PL approximation is to reduce its number of parameters. Notice that in the solution of the optimization problem (3.6), $c_1 \approx -a_1 \cdot b_1$, 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 (3.6), we obtain a PL approximation r_2 as follows:

$$r_2(X, Y) = \begin{cases} 3X + 4.69Y - 14.08 & \text{if } X < 5.15 \\ 27X + 5.35Y - 144.55 & \text{if } X \geq 5.15 \end{cases} \tag{3.8}$$

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 penalized MSE heuristic is $pMSE(r_2) = 7.6209$, which is lower than $pMSE(r_1) = 8.8869$. The pMSE value of r_2 is lower than the pMSE 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} (computed using r_2) are better than the corresponding ones for g_{W_1} (computed using r_1). $pMSE(r_2) = 7.6209$ is lower than $pMSE(r_1) = 8.8869$. At least in this case, the penalized MSE heuristic chooses a PL approximation that has lower errors. Of course, there are no guarantees that this will always happen.

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}, 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 $MSE(r_2) = 16.6193$, $pMSE(r_3) = 7.9791$ is higher than $pMSE(r_2) = 7.6209$ 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 ($x < s_X$, $x \geq s_X$, 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 penalized MSE heuristic is the 2-piece PL approximation described in Eq. (3.8), where we assume $c = -ab$, and that has the best goodness of fits statistics for KL, $MAD\,(PDF)$, and AEV.

3.2.2 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.

Table 3: Goodness of Fit and Run Time Results for the Product Function Example.

Split on	X	X	X	Y	Y	Y	Y
$c = -a \cdot b$?	No	Yes	Yes	No	Yes	Yes	Yes
# pieces	2	2	3	2	2	3	4
# free parameters	4	2	5	5	3	6	9
MSE	11.5121	16.6193	4.8312	8.9383	14.5624	2.6470	**1.10**
$pMSE$	8.8869	**7.6209**	8.1502	9.3807	8.3568	7.9469	9.19
KL	0.0095	**0.0011**	0.0029	0.0369	0.0453	0.0061	0.0032
$MAD\,(PDF)$	0.0021	**0.0006**	0.0006	0.0067	0.0069	0.0025	0.0015
AEM	1.0478	0.6677	0.2687	0.5660	**0.0000**	**0.0000**	**0.0000**
AEV	48.9341	**0.4223**	29.2187	11.1167	11.6572	1.1496	0.6330
CPU (in secs)	17.86	21.08	25.71	**17.04**	20.14	25.65	32.12

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

$$f_W(w) = \int_{-\infty}^{\infty} |y|\, f_X(w\, y)\, f_Y(y)\, \mathrm{d}y \tag{3.9}$$

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

$$r_1(X, Y) = \begin{cases} a_1 X + b_1 Y + c_1 & \text{if } Y < s_Y \\ a_2 X + b_2 Y + c_2 & \text{if } Y \geq s_Y \end{cases} \tag{3.10}$$

The PL approximation $r_1(X, Y)$ has 7 parameters (a_1, b_1, c_1, a_2, b_2, c_2, and 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:

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

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

$$\begin{aligned} & \text{Find } s_Y, a_1, b_1, c_1, a_2, b_2, c_2 \text{ so as to} \\ & \textit{Minimize } \ MSE(a_1, b_1, c_1, a_2, b_2, c_2, s_Y) \\ \text{subject to : } \ & r_1(0.41, 32.80) = 0.04, r_1(16.75, s_Y) \leq 10.92, \\ & r_1(0.41, s_Y) \geq 0.04, r_1(16.75, 4.60) = 10.92, \\ & 4.60 \leq s_Y \leq 32.80, \\ \text{and} \quad & a_1, a_2 \geq 0, b_1, b_2 \leq 0. \end{aligned} \tag{3.11}$$

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

$$r_1(X, Y) = \begin{cases} 0.61X - 0.37Y + 2.41 & \text{if } Y < 7.02 \\ 0.18X - 0.016Y + 0.50 & \text{if } Y \geq 7.02 \end{cases} \tag{3.12}$$

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. (3.12) 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 pMSE 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 χ^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.

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 pMSE value is $pMSE(r_2) = 0.8398$, which is higher than the pMSE value $pMSE(r_1) = 0.3349$ of the 2-piece PL approximation described in Eq (3.12).

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 pMSE 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. (3.12) is the PL approximation suggested by the penalized MSE heuristic. This is one example where the penalized MSE heuristic doesn't suggest the approximation that leads to the best goodness of fit statistics for the resulting marginal PDF; however, the trade-off between accuracy and computational complexity seems reasonable given the reasonable comparison of the marginal PDF to the actual model.

Table 4: Goodness of Fit and Run Time Results for the Quotient Function Example.

Split on	Y	Y	X	X
# pieces	2	3	2	3
# free parameters	4	7	5	9
MSE	0.1600	**0.0460**	0.2226	0.1232
$pMSE$	**0.3349**	0.8398	1.9957	4.8119
KL	0.2449	**0.0605**	0.8071	0.4393
$MAD\,(PDF)$	0.4834	**0.3054**	0.5292	0.4618
AEM	0.0951	**0.0524**	0.0868	0.0559
AEV	0.1644	**0.1021**	0.4202	0.2488
CPU (in secs.)	33.03	46.35	**28.15**	47.58

4 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 BNs in domains such as business.

4.1 Crop Problem

This example is similar to one used by Lerner [12] and Murphy [15]. 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. $V = C \cdot P$.

The BN 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 [18]. 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.

Next we found a MOP approximation of the conditional PDF of $P|c$ using the mixed tree technique proposed in [14]. 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 [18]). 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 $P|c$.

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.

Next, we found a 2-piece PL approximation $r_1(C, P)$ of the deterministic function associated with V using the constraint $c = -a \cdot b$ as described in Section 3.2.1. 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 pMSE 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 pMSE value for this PL approximation is 1.81. The details of the 2-piece PL approximation are as follows:

$$r_1(C, P) = \begin{cases} 6.96C + 2.61P - 18.14 & \text{if } C < 4.51 \\ 5.56C + 5.91P - 32.90 & \text{if } C \geq 4.51 \end{cases} \tag{4.1}$$

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) ≈ 0.0018, $AEM \approx 0.1995$, and $AEV \approx 0.4645$.

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 a pMSE 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.

Table 5: Risk of Fatal Crash per Billion Miles Traveled in Northeastern States of US in 2008

Month	*Avg. Temp.* (°F)	♯ *Fatal Crashes*	*Miles Traveled* (billions)	*Crash Rate* (per billion miles)
January	17	297	34.241	8.67
February	18	280	31.747	8.82
March	29	267	36.613	7.29
April	43	350	36.445	9.60
May	55	328	38.051	8.62
June	65	386	37.983	10.16
July	70	419	39.233	10.68
August	68	410	39.772	10.31
September	59	331	37.298	8.87
October	48	356	38.267	9.30
November	37	326	34.334	9.49
December	22	311	37.389	8.32

Source: US Department of Transportation

4.2 Risk of Fatal Crashes

This example is adapted from [6]. Table 5 describes some data on fatal crashes including average temperature, miles travelled and crash rate. A BN model for the data is as shown in Fig. 18.

To model the prior PDF of Average Temperature (T), we used a mixture of four beta distributions as follows:

$$f_T(t) = \begin{cases} \frac{9(t-40)^2}{156{,}250} & \text{if } 15 \leq t \leq 30 \\ -\frac{9(136{,}300-7440t+93t^2)}{5{,}000{,}000} & \text{if } 30 < t \leq 40 \\ -\frac{3(1{,}338{,}700-72{,}080t+901t^2)}{13{,}720{,}000} & \text{if } 40 < t \leq 50 \\ \frac{6(t-40)^2}{214{,}375} & \text{if } 50 < t \leq 75 \end{cases} \tag{4.2}$$

Notice that $f_T(t)$ is a 4-piece, 2-degree MOP on the domain $[15, 75]$. A graph of $f_T(t)$ is shown in Fig. 19. The conditional PDF of M given $T = t$ is a CLG distribution. Using the mixed-tree technique suggested by Moral et al. [14], we approximated this PDF by a 6-piece MOP, $f_{M|t}(m)$, where the mixed-tree pieces are on intervals $[15, 30]$, $(30, 50]$, $(50, 75]$. Similarly, we approximated the conditional PDF of N given $T = t$, $f_{N|t}(m)$, by a 6-piece mixed-tree MOP using the same three mixed-tree intervals as for M. The parameters for these two CLG distributions were obtained by least-squares regression with T as the independent variable and M and N as the dependent variables, respectively. The standard deviation of the error terms, ε, are defined as the standard errors from these two regression models.

The BN model assumes M and N are conditionally independent given T. This is also reflected in the regression model where if you regress, e.g., M against T and N, then the

coefficient on N is not statistically significant. The other assumption in the BN model is that given $\{N, M\}$, R is conditionally independent of T. This follows from the fact that the column *Fatal Crash Rate* in the table is the quotient $R = N/M$.

Assuming no evidence, we can compute the marginal PDF of (M, N), $f_{M,N}(m, n)$, as follows:

$$f_{M,N}(m, n) = \int_{-\infty}^{\infty} f_T(t)\, f_{M|t}(m)\, f_{N|t}(n)\, \mathrm{d}t \tag{4.3}$$

$f_{M,N}(m, n)$ is computed as a 30-piece, 6-degree MOP. Next, as discussed in Subsection 3.2.2, we found a 2-piece PL approximation of the deterministic conditional $R = \frac{N}{M}$ (where we split the domain of M in an optimal way) as follows:

$$r_1(N, M) = \begin{cases} 0.022N - 1.162M + 40.339 & \text{if } M < 31.001 \\ 0.024N - 0.192M + 8.008 & \text{if } X \geq 31.001 \end{cases} \tag{4.4}$$

Next we computed the marginal PDF of R using the PL approximation $R = r_1(N, M)$ and also using the exact conditional $R = \frac{N}{M}$. A graph showing the exact and approximate PDF of R is shown in Figure 20.

5 Summary and Conclusions

This paper is concerned with inference in hybrid BNs containing nonlinear deterministic conditionals using MOPs. MOPs are not 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 creating 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 penalized MSE heuristic for finding a PL approximation is suggested that minimizes a penalized MSE function subject to the domain and symmetry principles.

Using this heuristic, we describe a general algorithm for one-dimensional nonlinear functions. The algorithm for multidimensional functions is similar to the one-dimensional case. Also, 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.

The penalized MSE heuristic is not perfect. For the product function example, the penalized MSE 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 penalized MSE 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. In any case, the proposed penalized MSE heuristic is the first and only heuristic for finding PL approximations of a nonlinear function whose goal is to minimize the MSE of the marginal of the deterministic variable whose conditional is described by the nonlinear function.

Finally, we use our methods to solve two small hybrid BNs that contain nonlinear deterministic conditionals. The first one, called the *Crop* problem was first described by Murphy [15] and contains a product function. The second one, called *Risk of Fatal Crash*, described by Fenton and Neil [6], 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]. We are grateful to comments from three anonymous reviewers that has improved the exposition of the material. We are also grateful to Suzanna Emelio for proofreading the final version.

References

[1] H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.

[2] B. R. Cobb and P. P. Shenoy. Nonlinear deterministic relationships in Bayesian networks. In L. Godo, editor, *Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 8th European Conference, ECSQARU 2005*, Lecture Notes in Artificial Intelligence 3571, pages 27–38. Springer, Berlin, 2005.

[3] B. R. Cobb and P. P. Shenoy. Piecewise linear approximations of nonlinear deterministic conditionals in continuous Bayesian networks. In A. Cano, M. Gómez-Olmedo, and T. D. Nielsen, editors, *Proceedings of the 6th European Workshop on Probabilistic Graphical Models*, pages 59–66, Granada, Spain, 2012. DECSAI, University of Granada.

[4] B. R. Cobb, P. P. Shenoy, and R. Rumí. Approximating probability density functions in hybrid Bayesian networks with mixtures of truncated exponentials. *Statistics & Computing*, 16(3):293–308, 2006.

[5] P. A. M. Dirac. The physical interpretation of the quantum dynamics. *Proceedings of the Royal Society of London, Series A*, 113(765):621–641, 1927.

[6] N. Fenton and M. Neil. *Risk Assessment and Decision Analysis with Bayesian Networks*. CRC Press, Boca Raton, FL, 2013.

[7] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:76–86, 1951.

[8] H. Langseth, T. D. Nielsen, R. Rumí, and A. Salmerón. Parameter estimation and model selection for mixtures of truncated exponentials. *International Journal of Approximate Reasoning*, 51(5):485–498, 2010.

[9] H. Langseth, T. D. Nielsen, R. Rumí, and A. Salmerón. Mixtures of truncated basis functions. *International Journal of Approximate Reasoning*, 53(2):212 – 227, 2012.

[10] S. L. Lauritzen. Propagation of probabilities, means, and variances in mixed graphical association models. *Journal of the American Statistical Association*, 87(420):1098–1108, 1992.

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

[12] U. N. Lerner. *Hybrid Bayesian networks for reasoning about complex systems*. PhD thesis, Department of Computer Science, Stanford University, Stanford, CA, 2002.

[13] S. Moral, R. Rumí, and A. Salmerón. Mixtures of truncated exponentials in hybrid Bayesian networks. In S. Benferhat and P. Besnard, editors, *Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 6th European Conference, ECSQARU-2001*, Lecture Notes in Artificial Intelligence 2143, pages 156–167. Springer, Berlin, 2001.

[14] S. Moral, R. Rumí, and A. Salmerón. Approximating conditional MTE distributions by means of mixed trees. In T. D. Nielsen and N. L. Zhang, editors, *Symbolic and Quantitative Approaches to Reasoning with Uncertainty: 7th European Conference, ECSQARU 2003*, Lecture Notes in Artificial Intelligence 2711, pages 173–183. Springer, Berlin, 2003.

[15] K. P. Murphy. A variational approximation for Bayesian networks with discrete and continuous latent variables. In K. Laskey and H. Prade, editors, *Uncertainty in Artificial Intelligence: Proceedings of the Fifteenth Conference*, pages 457–466. Morgan Kaufmann, San Francisco, CA, 1999.

[16] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

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

[18] P. P. Shenoy. Two issues in using mixtures of polynomials for inference in hybrid Bayesian networks. *International Journal of Approximate Reasoning*, 53(5):847–866, 2012.

[19] P. P. Shenoy, R. Rumí, and A. Salmerón. Practical aspects of solving hybrid Bayesian networks containing deterministic conditionals. *International Journal of Intelligent Systems*, 30(3):265–291, 2015.

[20] P. P. Shenoy and G. Shafer. Axioms for probability and belief-function propagation. In R. D. Shachter, T. Levitt, J. F. Lemmer, and L. N. Kanal, editors, *Uncertainty in Artificial Intelligence 4*, pages 169–198. North-Holland, 1990.

[21] P. P. Shenoy and J. C. West. Extended Shenoy-Shafer architecture for inference in hybrid Bayesian networks with deterministic conditionals. *International Journal of Approximate Reasoning*, 52(6):805–818, 2011.

[22] P. P. Shenoy and J. C. West. Inference in hybrid Bayesian networks using mixtures of polynomials. *International Journal of Approximate Reasoning*, 52(5):641–657, 2011.

[23] J. Whittaker. *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester, UK, 1990.

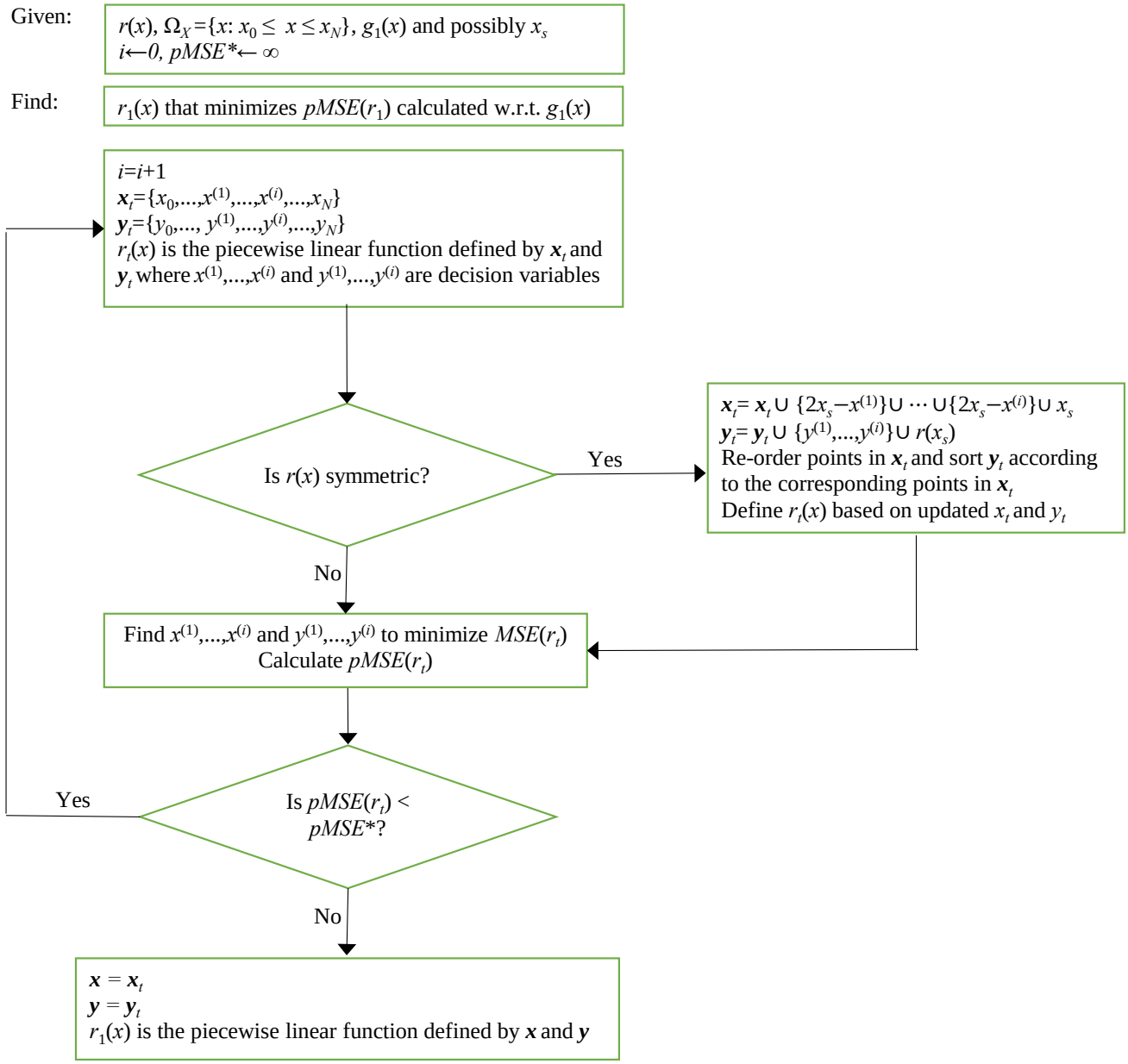

Figure 7: Algorithm in pseudo-code for approximating a one-dimensional nonlinear function with a PL function.

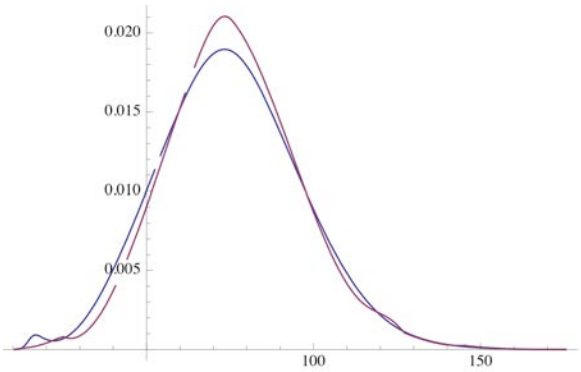

Figure 8: A Graph of $g_{W_1}(\cdot)$ (in red) Overlaid on the Graph of $g_W(\cdot)$ (in blue).

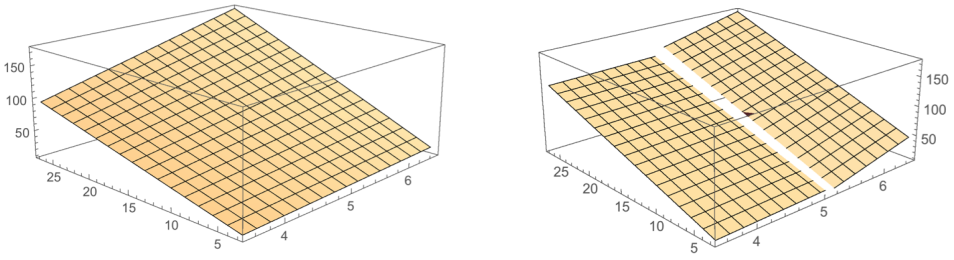

Figure 9: Left: A 3D Plot of $r(X, Y) = X \cdot Y$. Right: A 3D Plot of $r_2(X, Y)$.

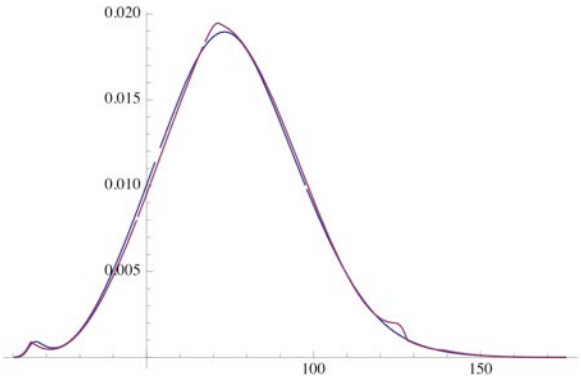

Figure 10: A graph of $g_{W_2}(\cdot)$ (in red) overlaid on the graph of $g_W(\cdot)$ (in blue)

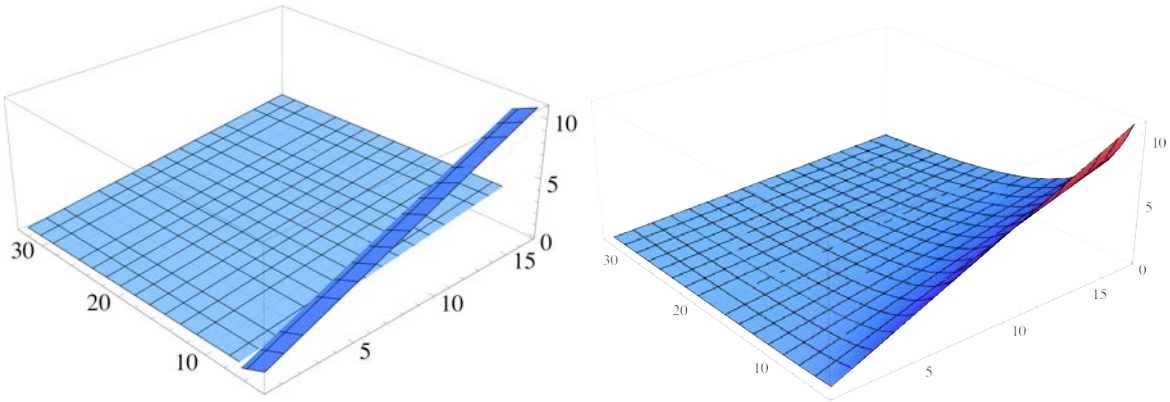

Figure 11: Left: A 3D Plot of $r_1(X, Y)$. Right: A 3D Plot of $r(X, Y)$.

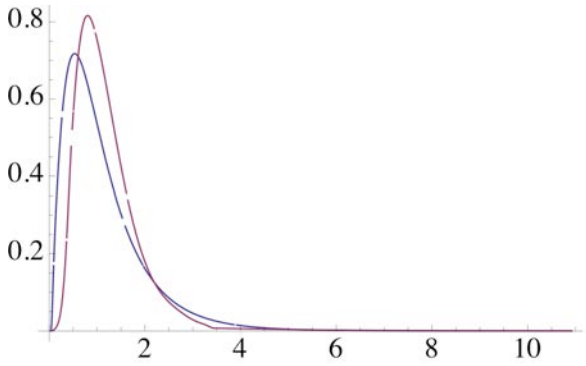

Figure 12: A Graph of $k_1(w)$ (in red) Overlaid on the Graph of $k_0(w)$ (in blue).

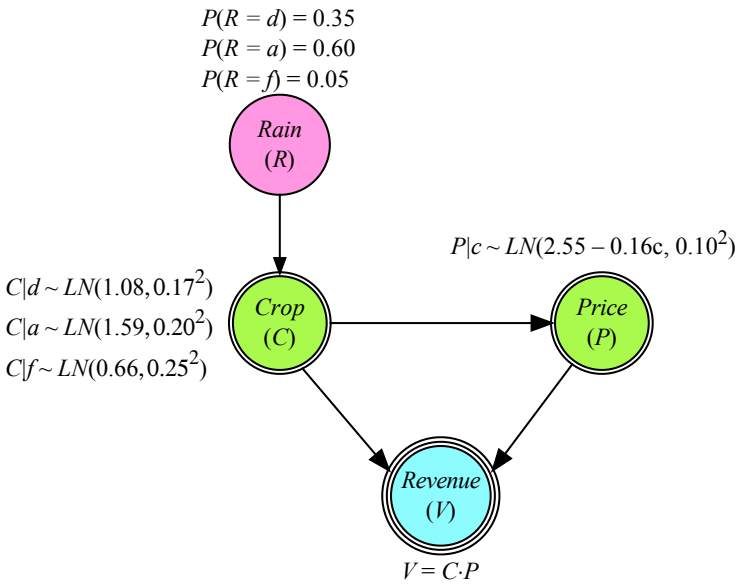

Figure 13: A BN Model for the Crop Example.

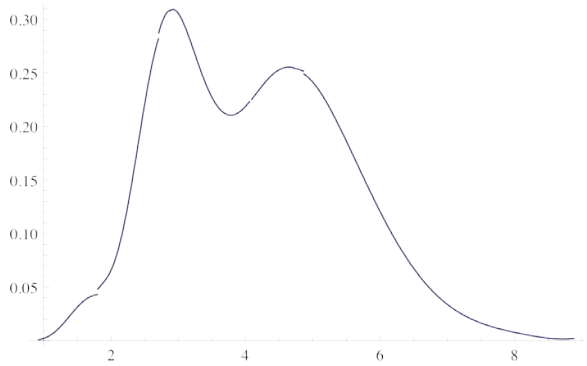

Figure 14: A MOP Approximation of the PDF of C in the Crop Example.

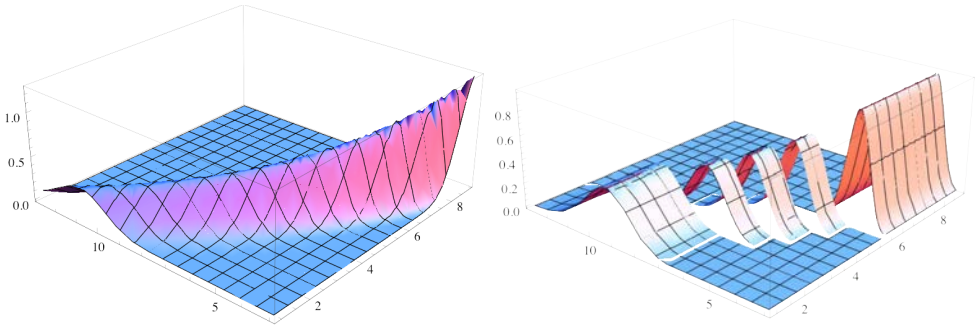

Figure 15: A 3D Plot of the Conditional PDF of $P|c$ (left) and its MOP Approximation (right).

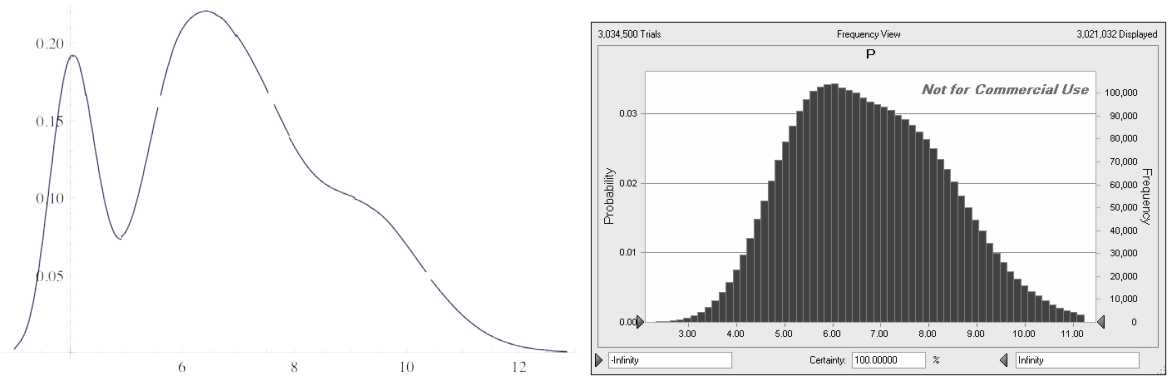

Figure 16: Left: A MOP Approximation of the Marginal PDF of P. Right: A Histogram for P Found Using Monte Carlo Simulation.

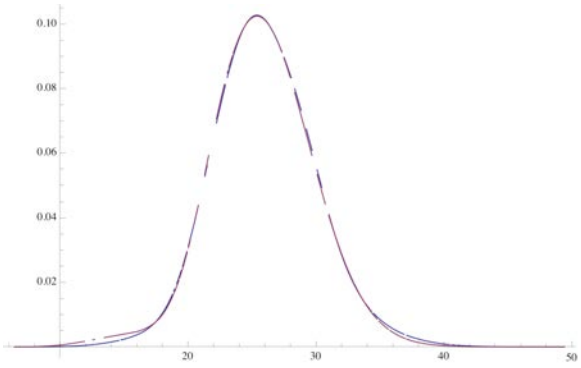

Figure 17: A MOP Approximation of the Marginal PDF of V Using a 2-piece PL Approximation of $V = C \cdot P$ (in red) Overlaid on the Marginal PDF of V Using $V = C \cdot P$ (in blue).

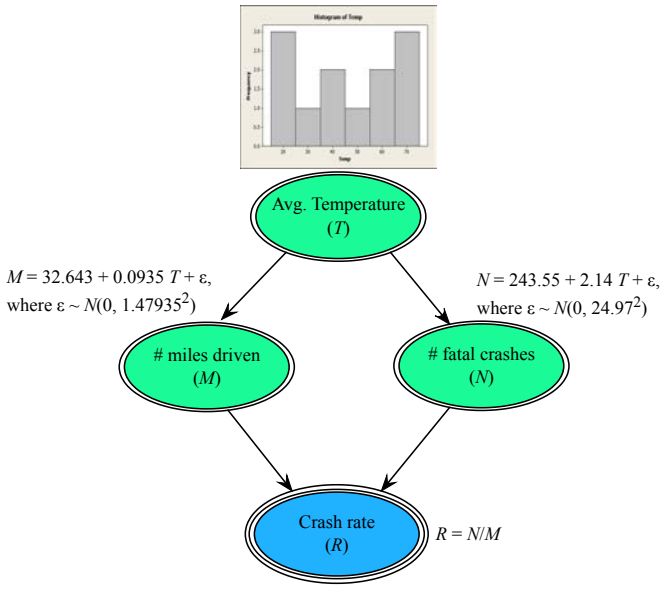

Figure 18: A BN Model for the Crash Rate Dataset.

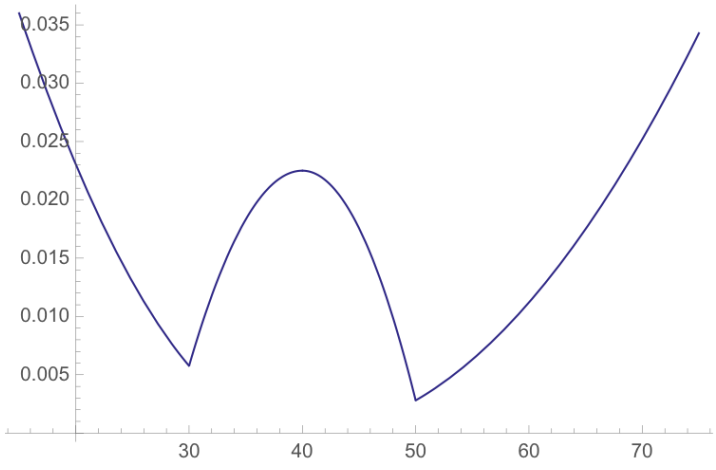

Figure 19: A Graph of the PDF of T.

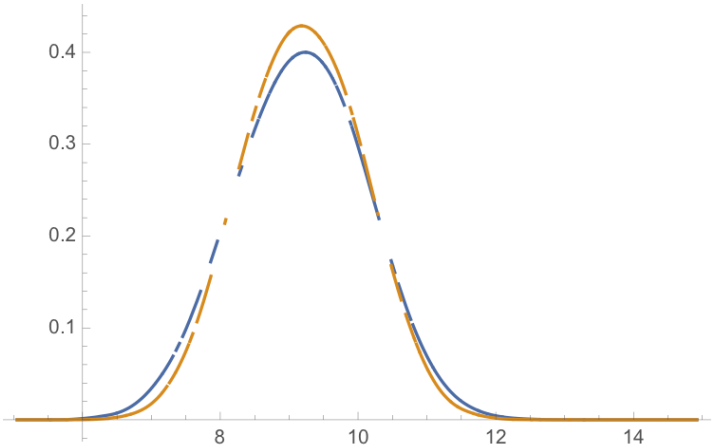

Figure 20: A Graph of the Exact (in blue) and Approximate (in orange) PDF of R.