Accepted for publication in: International Journal of Approximate Reasoning, 2012, <http://dx.doi.org/10.1016/j.ijar.2012.01.008>

Two Issues in Using Mixtures of Polynomials for Inference in Hybrid Bayesian Networks

Prakash P. Shenoy

University of Kansas School of Business

1300 Sunnyside Ave., Summerfield Hall, Lawrence, KS 66045-7601 USA

pshenoy@ku.edu

February 14, 2012

Abstract

We discuss two issues in using mixtures of polynomials (MOPs) for inference in hybrid Bayesian networks. MOPs were proposed by Shenoy and West for mitigating the problem of integration in inference in hybrid Bayesian networks. First, in defining MOP for multi-dimensional functions, one requirement is that the pieces where the polynomials are defined are hypercubes. In this paper, we discuss relaxing this condition so that each piece is defined on regions called hyper-rhombuses. This relaxation means that MOPs are closed under transformations required for multi-dimensional linear deterministic conditionals, such as $Z = X + Y$, etc. Also, this relaxation allows us to construct MOP approximations of the probability density functions (PDFs) of the multi-dimensional conditional linear Gaussian distributions using a MOP approximation of the PDF of the univariate standard normal distribution. Second, Shenoy and West suggest using the Taylor series expansion of differentiable functions for finding MOP approximations of PDFs. In this paper, we describe a new method for finding MOP approximations based on Lagrange interpolating polynomials (LIP) with Chebyshev points. We describe how the LIP method can be used to find efficient MOP approximations of PDFs. We illustrate our methods using conditional linear Gaussian PDFs in one, two, and three dimensions, and conditional log-normal PDFs in one and two dimensions. We compare the efficiencies of the hyper-rhombus condition with the hypercube condition. Also, we compare the LIP method with the Taylor series method.

1 Introduction

An hybrid Bayesian network (BN) is a BN with 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. Each variable in a BN is associated with a conditional distribution function (conditional, in short) for the variable given its parents. A conditional is said to be *deterministic* if its conditional variance is zero for each state of its parents.

Marginalizing a continuous variable involves integration of the product of all potentials that contain the variable in their domains. Often, these potentials are not integrable in closed form. This is a major problem in making inferences in hybrid BNs. We will call this the *integration* problem.

Literature Review. A traditional approximate method for dealing with the integration problem is to discretize all continuous variables. If the number of bins used for discretization is large (to increase the accuracy of the results), the computational effort required to find marginals can be large. A priori, we may not know the regions of the continuous variables where the posterior density lies. Kozlov and Koller [1] have proposed a dynamic discretization technique where one starts with a uniform coarse discretization, and then iteratively refines the discretization based on the location of the probability masses.

Another approximate method for dealing with the integration method is to use Monte Carlo sampling methods. There are a host of methods including importance sampling (e.g., [2, 3]) and Markov chain Monte Carlo (e.g., [4]). The idea is to sample from the posterior distribution. In the presence of deterministic conditionals, convergence can be a problem.

One exact solution to the integration problem proposed by Lauritzen and Jensen [5] is to restrict conditionals of continuous variables to the conditional linear Gaussian (CLG) family, and for discrete variables to not have continuous parents. Such BNs are called *mixture of Gaussians* BNs. In this case, we can avoid the integration problem as marginals of multivariate normal distributions are multivariate normal and no integration needs to be done. However, restricting conditionals to the CLG family can be too restrictive. Also, the requirement that discrete variables not have continuous parents can also be too restrictive. Finally, in finding marginals, all continuous variable have to be marginalized before marginalizing discrete ones, and this restriction can lead to large cliques making inference intractable [6].

If a BN has discrete variables with continuous parents, Murphy [7] uses a variational approach to approximate the product of the potentials associated with a discrete variable and its parents with a CLG distribution. Lerner *et al.* [8] uses a numerical integration technique called Gaussian quadrature to approximate non-CLG distributions with CLG distributions, and this same technique can be used to approximate the product of potentials associated with a discrete variable and its continuous parents. Murphy's and Lerner's approach is then embedded in the Lauritzen-Jensen [5] algorithm to solve the resulting mixtures of Gaussians BN.

Shenoy [9] proposes approximating non-CLG distributions by mixtures of Gaussians using a nonlinear optimization technique, and using arc reversals to ensure discrete variables do not have continuous parents. The resulting mixture of Gaussians BN is then solved using the Lauritzen-Jensen [5] algorithm.

Another solution to the integration problem is to approximate conditional PDFs by a functions called *mixtures of truncated exponentials* (MTEs) [10]. MTE functions are piecewise functions that are defined on regions called hypercubes, and the functions themselves are exponential functions of a linear function of the variables. Such functions are easy to integrate, and the family of MTE functions is closed under multiplication, addition, and integration, three operations that are used in finding marginals using the extended Shenoy-Shafer architecture [11]. Cobb *et al.* [12] describe MTE approximations of several commonly

used one-dimensional PDFs. Moral *et al.* [13] describe a mixed-tree method for representing an MTE approximation of the 2-dimensional CLG distribution. Parameter learning in MTE networks are discussed in [14, 15]. Rumí and Salmerón [16] discuss approximate inference in MTE hybrid BNs that do not contain deterministic conditionals.

Another method that is similar in principle to the MTE method is the mixture of polynomials (MOP) method proposed by Shenoy and West [17]. Instead of using piecewise exponential functions, the MOP method uses piecewise polynomials. Although a detailed comparison of MTE and MOP methods has yet to be done, an advantage of the MOP method is that one can easily find MOP approximations of differentiable PDFs using the Taylor series expansion of the PDF. Shenoy and West [17] describe a MOP approximation of a two-dimensional CLG distribution using the Taylor series method.

Contributions. In both the MTE and the MOP methods, the multi-dimensional piecewise functions are defined on regions called hypercubes. One advantage of this restriction is that such multi-dimensional piecewise functions are easy to integrate. However, the hypercube restriction poses two limitations. First, it is difficult to find an MTE or a MOP approximation of a multi-dimensional conditional PDF for dimensions greater than two. The mixed-tree method proposed by Moral *et al.* [13] and the Taylor series method proposed by Shenoy and West [17] do not scale up to higher dimensions in practice, i.e., the approximations using these methods have too many pieces or too many terms or have too high a degree for practical use.

The second limitation is that in the presence of multi-dimensional linear deterministic conditionals, the family of MTE and MOP functions are not closed. For example, suppose X has PDF $f_X(x)$ and suppose Y has conditional PDF $f_{Y|x}(y)$, and suppose Z has a deterministic conditional given by the linear function $Z = X + Y$. To find the marginal distribution of Z, we need to combine $f_X(x)$ and $f_{Y|x}(z-x)$ and then integrate x out of the combination. The problem is that even if $f_{Y|x}(y)$ was defined on hypercubes, $f_{Y|x}(z-x)$ is no longer defined on hypercubes. This problem applies equally to the MTE and MOP methods.

In this paper, we suggest replacing the hypercube condition with a more general hyper-rhombus condition. For one-dimensional functions, the two conditions coincide. However, for dimensions two or greater, the hyper-rhombus condition is a generalization of the hypercube condition. The hyper-rhombus condition has three important advantages. First, MOP functions defined on hyper-rhombuses are closed under operations required for multi-dimensional linear deterministic conditionals. Second, it allows us to define MOP approximations of high-dimensional CLG distributions using a MOP approximation of the one-dimensional standard normal PDF. Third, the hyper-rhombus condition allows us to find MOP approximations of multi-dimensional conditional PDFs that have fewer pieces and lower degrees than MOP approximations that are restricted to hypercubes.

Another contribution of this paper is a method for finding MOP approximations of PDFs based on Lagrange interpolating polynomials (LIP) with Chebyshev points. We describe this method, and compare it with the Taylor series method. The LIP method produces MOP approximations that have a better fit than the Taylor series method assuming the same number of pieces and same degree. The LIP method does not require a PDF to be differentiable. For multi-dimensional conditional PDFs, the LIP method with Chebyshev

points coupled with the hyper-rhombus condition allows us to find MOP approximations that have fewer pieces and lower degrees than MOP approximations found using Taylor series method.

Limitations. The hyper-rhombus condition has some disadvantages compared to hypercubes. First, integrating MOPs defined on hyper-rhombuses takes longer than integrating MOPs on hypercubes. Second, after integration, the degrees of MOPs defined on hyper-rhombuses tend to increase, whereas for MOPs defined on hypercubes the degrees always decrease.

The hyper-rhombus condition does not help for MTE functions, i.e., if the definition of MTE functions were generalized so that the hypercube condition was replaced by the hyper-rhombus condition, then MTE functions would not be closed under operations required for multi-dimensional linear deterministic functions. For example, the sum of two independent variables with exponential PDFs (which are MTEs) has a gamma PDF, which is not a MTE function.

One downside of the LIP method is that when one is restricted to doing exact analysis (using integers for non-transcendental numbers) for stability reasons in the presence of high degree polynomials, the MOP functions produced by the LIP method have significantly more terms (compared to the Taylor series method), and this can slow down the computation of marginals.

Outline. An outline of the remainder of this article is as follows. In Section 2, we provide a re-definition of high-dimensional MOP functions that are defined on regions called hyper-rhombuses. Also, by means of a simple example, we describe the process of integrating a MOP defined on a hyper-rhombus. In Section 3, we describe a general process for finding MOP approximations of PDFs using the LIP method with Chebyshev points. Also, we illustrate this general process to find MOP approximations of the standard normal PDF, a one-dimensional log-normal PDF, and a two-dimensional conditional log-normal PDF. Also, we show how we can use a MOP approximation of the standard normal PDF to find MOP approximations of the PDFs of two- and three-dimensional CLG distributions. In Section 4, we compare the practical implications of the hyper-rhombus condition with the hypercube condition. We compare the time required for computation of marginals for a couple of simple Bayesian networks, and also the accuracy of the computed marginals. In Section 5, we compare the LIP method with the Taylor series method. Finally in Section 6, we summarize our findings and discuss some issues for further research.

2 Mixture of Polynomials Functions

In this section, we define MOP functions. The definition we provide here is slightly more general than the definition provided in Shenoy and West [17] for the case of multi-dimensional functions.

2.1 MOP Functions

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.

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

$$g_1(x) = \begin{cases} 0.41035 + 0.09499x - 0.09786x^2 - 0.02850x^3 & \text{if } -3 < x < 0, \\ 0.41035 - 0.09499x - 0.09786x^2 + 0.02850x^3 & \text{if } 0 \leq x < 3 \\ 0 & \text{otherwise} \end{cases} \tag{2.2}$$

$g_1(\cdot)$ 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, which will be discussed in Section 3.4

The definition given in Equation (2.1) is exactly the same as in Shenoy and West [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:

$$\begin{aligned} P(x_1, x_2) = a_{00} + a_{10}x_1 + a_{01}x_2 + a_{11}x_1x_2 + a_{20}x_1^2 + a_{02}x_2^2 \\ + a_{21}x_1^2x_2 + a_{12}x_1x_2^2 + a_{22}x_1^2x_2^2 \end{aligned} \tag{2.3}$$

The degree of the polynomial in Equation (2.3) 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

$$f(x_1, x_2, \ldots, x_m) \quad = \quad \begin{cases} P_i(x_1, x_2, \ldots, x_m) & \text{for } (x_1, x_2, \ldots, x_m) \in A_i, i = 1, \ldots, k, \\ 0 & \text{otherwise} \end{cases} \tag{2.4}$$

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

$$\begin{aligned} l_{1i} \leq x_{\pi(1)} \leq u_{1i}, \\ l_{2i}(x_{\pi(1)}) \leq x_{\pi(2)} \leq u_{2i}(x_{\pi(1)}), \\ \vdots \\ l_{mi}(x_{\pi(1)}, \ldots, x_{\pi(m-1)}) \leq x_{\pi(m)} \leq u_{mi}(x_{\pi(1)}, \ldots, x_{\pi(m-1)}) \end{aligned} \tag{2.5}$$

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 Equation (2.5) as a *hyper-rhombus*. Although we have defined the hyper-rhombus as a closed region in Equation (2.5), each of the $2m$ inequalities can be either strictly $<$ or $\leq$. Notice that the hyper-rhombus region in Equation (2.5) is a generalization of the condition that A_i are intervals for the one-dimensional case.

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

$$l_{1i} \leq x_1 \leq u_{1i}, \quad l_{2i} \leq x_2 \leq u_{2i}, \quad \ldots, \quad l_{mi} \leq x_m \leq u_{mi} \tag{2.6}$$

where $l_{1i}, \ldots, l_{mi}, u_{1i}, \ldots, u_{mi}$ are all constants. We refer to the region defined in Equation (2.6) as a *hypercube* (in m dimensions).

An example of a 2-piece, 3-degree MOP $h_1(\cdot, \cdot)$ defined on a two-dimensional hyper-rhombus region is as follows:

$$h_1(x, y) = \begin{cases} 0.41035 + 0.09499(y - x) - 0.09786(y - x)^2 - 0.02850(y - x)^3 & \text{if } x - 3 < y < x, \\ 0.41035 - 0.09499(y - x) - 0.09786(y - x)^2 + 0.02850(y - x)^3 & \text{if } x \leq y < x + 3 \\ 0 & \text{otherwise} \end{cases} \tag{2.7}$$

$h_1(x, y)$ is a two-dimensional MOP approximation of the PDF of the CLG distribution of $Y|x \sim N(x, 1^2)$ on the domain $-\infty < x < \infty$, $x - 3 < y < x + 3$. Notice that $h_1(x, y) = g_1(y - x)$, where $g_1(\cdot)$ is as defined in Equation (2.2).

The definition of a m-dimensional MOP function stated in Equation (2.4) is more general than the corresponding definition stated in Shenoy and West [17], which is as follows:

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

$$f(x_1, \ldots, x_m) = f_1(x_1) \cdot f_2(x_2) \cdots f_m(x_m) \tag{2.8}$$

where each $f_i(x_i)$ is a one-dimensional MOP function as defined in Equation (2.1).

It is easy to see that an m-dimensional function satisfying the condition in Equation (2.8) will also satisfy the condition in Equation (2.4), but the converse is not true. Thus, a function as follows:

$$f(x_1, x_2) = \begin{cases} x_1 x_2^2 + x_1^2 x_2 & \text{for } -3 \leq x_1 \leq 3 \text{ and } x_1 - 3 \leq x_2 \leq x_1 + 3 \\ 0 & \text{otherwise} \end{cases} \tag{2.9}$$

satisfies Equation (2.4) but not Equation (2.8) for two reasons. First, $x_1 x_2^2 + x_1^2 x_2$ cannot be obtained by a product of two one-dimensional polynomials. Second, the function f is defined on the region $-3 \leq x_1 \leq 3$, $x_1 - 3 \leq x_2 \leq x_1 + 3$, which is not a hypercube, but is a hyper-rhombus.

Finally, high-dimensional MOP functions defined on hyper-rhombuses remain MOP functions after integration. Thus, the family of MOP functions are closed under multiplication, addition, and integration. They are also closed under operations needed for multi-dimensional linear deterministic conditionals.

There are some advantages and some disadvantages of the hyper-rhombus condition compared to hypercubes. First, the family of MOP functions are closed under operations needed

for multi-dimensional linear deterministic conditionals. Second, we can more easily construct MOP approximations of high dimensional conditional PDFs such as the conditional linear Gaussian distributions. Third, the hyper-rhombus condition allows us to construct MOP approximations of conditional PDFs that have fewer pieces and lower degrees than MOP approximations on hypercubes. A disadvantage of the new definition is that it is more difficult to integrate MOP functions on hyper-rhombuses compared to MOP functions on hypercubes. Also, after integration, the degrees of MOP functions defined on hyper-rhombuses tend to increase, whereas for MOP functions defined on hypercubes, they tend to decrease. The added complexity of integrating a MOP function defined on a hyper-rhombus is described in the next subsection. Also, the advantages and disadvantages of hyper-rhombus versus hypercube are discussed further in greater detail in Section 4.

2.2 Integrating a MOP on a Hyper-rhombus

In this subsection, we investigate the complexity of integrating a MOP defined on a hyper-rhombus by means of a small example.

Suppose $X \sim U(0,1)$, $Y \sim U(0,2)$, X and Y are independent, and $W = X + Y$. Let $f_X(\cdot)$ and $f_Y(\cdot)$ denote the PDFs of X and Y, respectively. Thus,

$$f_X(x) = \begin{cases} 1 & \text{if } 0 < x < 1 \\ 0 & \text{otherwise} \end{cases}$$

$$f_Y(y) = \begin{cases} \frac{1}{2} & \text{if } 0 < y < 2 \\ 0 & \text{otherwise} \end{cases}$$

Both $f_X(\cdot)$ and $f_Y(\cdot)$ are 1-piece, 0-degree MOPs. Suppose we wish to find the marginal of W. Let $f_W(\cdot)$ denotes the marginal PDF of W. Then, $f_W(w)$ is given by the convolution formula:

$$f_W(w) = \int_{-\infty}^{\infty} f_X(x)\, f_Y(w - x)\, \mathrm{d}x \tag{2.10}$$

Notice that the potential $f_Y(w - x)$ is a two-dimensional MOP defined on a hyper-rhombus. The integrand, $f_X(x)\, f_Y(w - x)$, is a 1-piece, 0-degree two-dimensional MOP in x and w defined on the hyper-rhombus $0 < x < 1$, $0 < (w - x) < 2$. In integrating with respect to x, the lower and upper limits of integration depend on w. Thus, the integrand is equal to $\frac{1}{2}$ if $\max\{0, w - 2\} < x < \min\{1, w\}$, and 0 otherwise. So, if $w \leq 0$ or $w \geq 3$, the integrand is 0, and non-zero otherwise. The two upper bounds on x are equal when $w = 1$, and the two lower bounds on x are equal when $w = 2$. Thus, if $0 < w \leq 1$, the limits of integration are from 0 to w. If $1 < w \leq 2$, the limits of integration are 0 to 1, and if $2 < w < 3$, the limits of integration are $(w - 2)$ to 1. Therefore, the result of the integration is a 3-piece, 1-degree polynomial as follows:

$$f_W(w) = \begin{cases} \frac{w}{2} & \text{if } 0 < w \leq 1 \\ \frac{1}{2} & \text{if } 1 < w \leq 2 \\ \frac{3-w}{2} & \text{if } 2 < w < 3 \\ 0 & \text{otherwise} \end{cases} \tag{2.11}$$

From this simple example, we see that in integrating a MOP defined on a hyper-rhombus, the limits of integration are the results of solutions of linear inequalities, which may be a linear function of the remaining variables, and thus, may not be constants. In comparison, in the case of integrating a MOP defined on a hypercube, while the limits of integration may depend on the remaining variables, they are always constants.

3 Finding MOP Approximations of PDFs

In this section, we describe finding MOP approximations of PDFs using Lagrange interpolating polynomials (LIP) with Chebyshev points. Given an approximation of a PDF, we describe some measures of goodness of fit. Next, we describe a general process for finding a MOP approximation of a PDF (in one or higher dimensions) using LIP with Chebyshev points. We illustrate this general process for finding a MOP approximation of the standard normal PDF, an univariate log-normal PDF, and a two-dimensional conditional log-normal PDF. Also, we describe how we can use a MOP approximation of the univariate standard normal PDF to find MOP approximations of CLG PDFs in two or higher dimensions.

3.1 Lagrange Interpolating Polynomials with Chebyshev Points

Suppose we need to fit a polynomial for a given one-dimensional function $f(x)$. Given a set of n points $\{(x_1, f(x_1)), ..., (x_n, f(x_n))\}$, a Lagrange interpolating polynomial in x is a function $P(x)$ given by

$$P(x) = \sum_{j=1}^{n} \left[f(x_j) \prod_{k=1, k\neq j}^{n} \frac{x - x_k}{x_j - x_k} \right] \tag{3.1}$$

Although the formula in Equation (3.1) is attributed to Lagrange, who published it in 1795 in his book *Leçons Elémentaires sur les Mathématiques*, it was first published by Waring [18] in 1779.

$P(x)$ has the following properties [19]. It is a polynomial of degree $\leq (n-1)$ that passes through the n points $\{(x_1, f(x_1)), ..., (x_n, f(x_n))\}$, i.e., $P(x_j) = f(x_j)$ for $j = 1, ..., n$. If $f(x)$ is continuous and n-times differentiable in an interval $[a, b]$, and $x_1, \ldots, x_n$ are distinct points in $[a, b]$ such that $x_1 < \ldots < x_n$, then for each $x \in [a, b]$, there exists a number $\xi(x)$ (generally unknown) between x_1 and x_n such that

$$f(x) = P(x) + \frac{f^{(n+1)}(\xi(x))}{n!}(x - x_1)(x - x_2) \cdots (x - x_n) \tag{3.2}$$

When constructing a polynomial to fit a PDF, there is a tradeoff between the fit and the smoothness of the fit. The more data points that are used to construct the polynomial, the higher the degree of the polynomial, and therefore, the greater the oscillation it will exhibit between the data points. One solution to this problem is to divide the range in which we wish to fit a function into several intervals, select a small number of points in each interval, and then fit a low-degree polynomial in each interval. Another solution to this problem is to choose the points as per Chebyshev's theory so as to minimize the deviation between the target function and the interpolating polynomial. We will use both of these strategies to find MOP approximation using a small number of pieces and low degrees.

Chebyshev Points. Suppose we wish to fit a $(n-1)$-degree polynomial for the function $f(x)$ on the domain (a, b). How should we choose the n points? For the interval (a, b), the n Chebyshev points are given by [19]:

$$x_i = \frac{1}{2}(a+b) + \frac{1}{2}(b-a)\,cos\left(\frac{2i-1}{2n}\pi\right), \text{for } i = 1, \ldots, n. \tag{3.3}$$

The Chebyshev points are often used with Lagrange interpolating polynomials because the resulting polynomial approximation $P(x)$ minimizes the quantity $|(x-x_1)\cdots(x-x_n)|$ for all $x \in [a, b]$, which is proportional to the absolute error between the function $f(x)$ and the interpolating polynomial $P(x)$ (see Equation (3.2)). The minimum value of $|(x-x_1)\cdots(x-x_n)|$ is $\frac{1}{2^{n-1}}$. Thus, as n increases, the maximum absolute deviation decreases, thus avoiding the oscillating behavior of Lagrange interpolating polynomials. An example of the use of Chebyshev points will be given in the Section 3.4.

3.2 Some Measures of Goodness of Fit

Suppose we find an approximation $g(\cdot)$ of a PDF $f(\cdot)$. There are several ways of measuring the goodness of fit of $g(\cdot)$ with respect to $f(\cdot)$.

We can use the Kullback-Liebler (KL) divergence [20] as a measure of the goodness of fit. If f is a PDF on the range (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

$$KL(f, g) = \int_a^b \ln\left(\frac{f(x)}{g(x)}\right) f(x)\,\mathrm{d}x. \tag{3.4}$$

$KL(f, g) \geq 0$, and $KL(f, g) = 0$ if and only if $g(x) = f(x)$ for all $x \in (a, b)$.

Another measure of goodness of a fit is the maximum absolute deviation. Thus, if f is a PDF on the range (a, b), and g is a PDF that is an approximation of f, then the maximum absolute deviation between f and g, denoted by $MAD(f, g)$, is given by:

$$MAD(f, g) = sup\{|f(x) - g(x)| : a < x < b\} \tag{3.5}$$

Finally, other measures of goodness of fit are the absolute errors in the means and variances. Thus, 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:

$$AEM(f, g) = |E(f) - E(g)| \tag{3.6}$$

$$AEV(f, g) = |V(f) - V(g)| \tag{3.7}$$

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

3.3 A General Procedure for Finding MOP Approximations of PDFs

In this subsection, we will describe a general procedure for finding a MOP approximation of a conditional PDF. In the succeeding subsections, we will illustrate this procedure for some

CLG and non-CLG PDFs, in one and higher dimensions. We will describe the procedure for a PDF in one-dimension. The procedure is the same for higher dimensions.

Suppose we wish to find a MOP approximation of a PDF $f(\cdot)$. First, if $f(x)$ is defined on a infinite domain, we need to decide on a finite domain for the MOP approximation. In general, depending on the context of the application, we can fit a distribution from 0.5 to 99.5 percentiles giving us 99% of the total probability. If more precision is required, we can choose, e.g., 0.05 to 99.95 percentiles giving us 99.9% of the total probability.

Suppose we have fixed the domain (a, b) for the MOP approximation. Next, we fit the PDF $f(x)$ on the domain (a, b) with a 3-degree polynomial, say $g_u(x)$, using the LIP method with Chebyshev points. Second, we make sure the PDF is non-negative on the entire range (a, b) (by computing the minimum of $g_u(x)$ on the range (a, b) and making sure it is non-negative). If not, we increase the degree of the fitted polynomial until we get non-negativity. Since we are using Chebyshev points, we are guaranteed to get non-negativity for some n assuming $f(x) > 0$ for all $x \in (a, b)$. If the smallest degree n for which we obtain non-negativity is too high (e.g., > 5 for a one-dimensional MOP), then we partition the domain (a, b) into more pieces and restart. Currently, we have no theory for how to partition the global domain. However, we can use heuristics such as choosing the mode of the PDF, local optima, inflection points, equal widths, or some combination of these. Ideally, we would like to keep the number of pieces and degrees as small as possible.

Next, after we have a non-negative MOP approximation $g_u(x)$, we normalize it so that it integrates to 1. Then we check the goodness of fit statistics of the normalized MOP approximation. If these are acceptable, we are done. If not, we can increase the degree of the approximation, or increase the number of pieces, or both. In all cases, we need to ensure that the MOP approximation is non-negative over the entire domain, and that it integrates to 1.

Regarding the number of pieces and the degree of a MOP approximation, there is a tradeoff between number of pieces and degree. Using more pieces will, in general, allow us to fit a lower degree MOP. Thus, e.g., we can approximate any PDF with a 1-degree (linear) MOP function if we have sufficient number of pieces to ensure a good fit. Our strategy is to keep the number of pieces as small as possible, subject to the constraint that the degree is below some threshold. We will illustrate this general procedure in Sections 3.4, 3.5, and 3.7.

3.4 Fitting MOPs to One-dimensional Gaussian PDFs

Consider the PDF of the one-dimensional standard normal distribution, $\varphi(z) = (1/\sqrt{2\pi})e^{-z^2/2}$ for $-\infty < z < \infty$. To approximate this function with a MOP, we must first decide on a range on which we find an approximation. For the standard normal PDF, we can use the interval $(-3, 3)$, since this interval contains 99.73% of the total probability.

If we try to fit a 1-piece MOP on the interval $(-3, 3)$, the smallest n (# of Chebyshev points) for which we get non-negativity is $n = 9$, thus resulting in a 1-piece, 8-degree MOP. For many applications, degree 8 is too high. So, we divide the interval $(-3, 3)$ into two sub-intervals $(-3, 0), [0, 3)$. Why these sub-intervals? The PDF φ is symmetric around 0, and 0 is also the mode of φ.

For each of the pieces $(-3, 0)$ and $[0, 3)$, we find n-Chebyshev points using Equation (3.3). The smallest n for which we obtain non-negativity is $n = 4$. The four Chebyshev points for

the interval $(-3, 0)$ are $x_1 = -2.886$, $x_2 = -2.074$, $x_3 = -0.926$, and $x_4 = -0.114$, and for the interval $[0, 3)$, the four Chebyshev points are: $x_1 = 0.114$, $x_2 = 0.926$, $x_3 = 2.074$, and $x_4 = 2.886$. After normalization, the resulting 2-piece, 3-degree MOP $g_1(\cdot)$ is as described in Equation 2.2.

Let $\Phi(\cdot)$ denote the CDF corresponding to PDF $\varphi(\cdot)$, and let $G_1(\cdot)$ denote the CDF corresponding to PDF $g_1(\cdot)$. For example, we can obtain $G_1(\cdot)$ from $g_1(\cdot)$ as follows:

$$G_1(z) = \int_{-\infty}^{z} g_1(y)\,\mathrm{d}y \tag{3.8}$$

Since MOPs are closed under integration, $G_1(\cdot)$ is computed as a 3-piece, 4-degree MOP as follows:

$$G_1(z) = \begin{cases} 0 & \text{if } z \leq 0 \\ 0.5 + 0.41035z^1 + 0.04750z^2 - 0.03262z^3 - 0.00713z^4 & \text{if } -3 < z < 0 \\ 0.5 + 0.41035z^1 - 0.04750z^2 - 0.03262z^3 + 0.00713z^4 & \text{if } 0 \leq z < 3 \\ 1 & \text{if } z \geq 3 \end{cases}$$

Figure 1 shows a graph of CDF $G_1(\cdot)$ overlaid on the graph of $\Phi(\cdot)$. Also, Figure 2 shows a graph of the PDF $g_1(\cdot)$ overlaid on the graph of the PDF $\varphi(\cdot)$. The goodness of fit statistics are as follows:

$$\begin{aligned} KL(\varphi, g_1) &\approx 0.0086 \\ MAD(\varphi, g_1) &\approx 0.0140 \\ MAD(\Phi, G_1) &\approx 0.0053 \\ AEM(\varphi, g_1) &\approx 0.0000 \\ AEV(\varphi, g_1) &\approx 0.0203 \end{aligned}$$

We make several comments about the values of the various goodness of fit statistics.

1. $MAD(\Phi, G_1)$ is in units of probability, whereas $MAD(\varphi, g_1)$ is in units of probability density. Thus, the two values cannot be compared with each other.

2. Since the functions G_1 and g_1 represent exactly the same distribution, the "accuracies" of the two functions are exactly the same. Thus, if we use these two functions to compute the probability of any interval $(c, d) \subseteq (-3, 3)$ (using $G_1(d) - G_1(c)$, and $\int_c^d g_1(x)\,\mathrm{d}x$), then we get exactly the same result.

3. Since $P(Z \in (c, d)) = G_1(d) - G_1(c)$, the maximum absolute error of this probability is $\leq 2 \times MAD(\Phi, G_1) = 0.0106$.

4. Since $P(Z \in (c, d)) = \int_c^d g_1(x)\,\mathrm{d}x$, the maximum absolute error of this probability is $\leq MAD(\varphi, g_1) \cdot (d - c) = 0.0140 \cdot (d - c)$. Thus, if $(d - c)$ is small, we may get smaller error bounds on the probability compared to the bounds obtained from $MAD(\Phi, G_1)$.

5. Based on the graphs in Figures 1 and 2, it may appear visually that G_1 is more accurate than g_1. But as discussed above, this is an incorrect conclusion. The y-axis on these two graphs are in different units, and different scales.

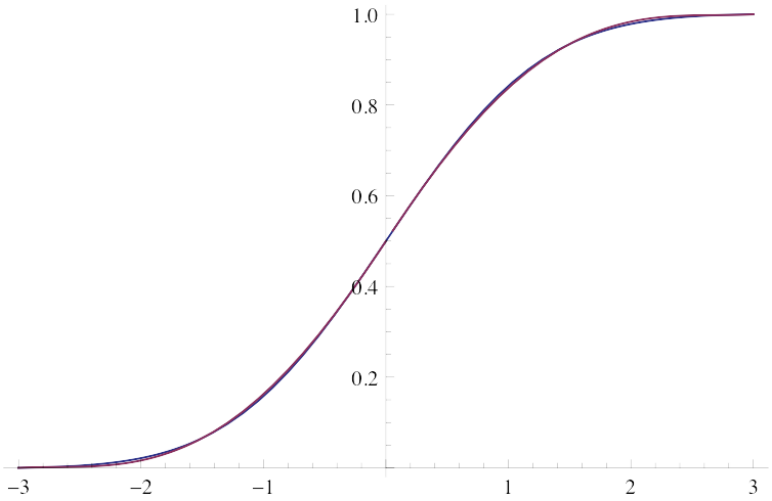

Figure 1: A graph of the CDF corresponding to $g_1(z)$ (in red) overlaid on the graph of the CDF corresponding to $\varphi(z)$ (in blue)

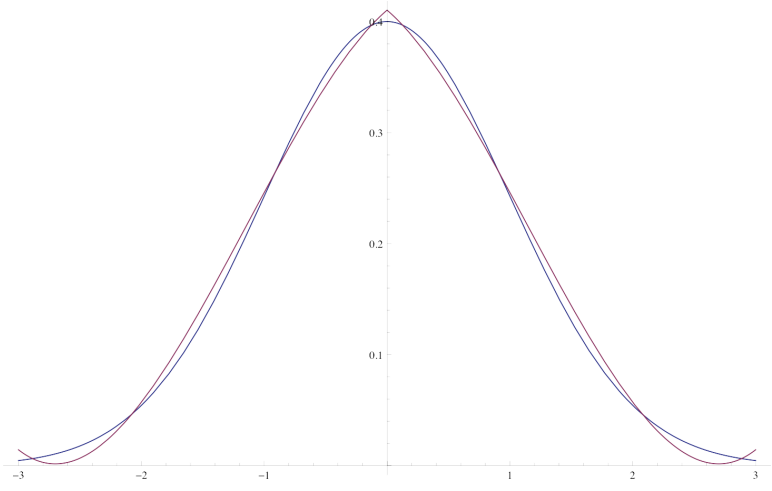

Figure 2: A graph of $g_1(z)$ (in red) overlaid on the graph of $\varphi(z)$ (in blue)

To find a MOP approximation of the PDF of the general $N(\mu, \sigma^2)$ distribution, where μ and σ are real constants such that $\sigma \neq 0$, we exploit the fact that MOP functions are closed under linear transformations. Thus, if $f(x)$ is a MOP function, then given any constants a and b, $f(ax+b)$ is also a MOP function. If $Z \sim N(0,1)$, its PDF is approximated by a MOP function $g(z)$, and $X = \sigma Z + \mu$, then $X \sim N(\mu, \sigma^2)$, and a MOP approximation of the PDF of X is given by $\frac{1}{|\sigma|}\, g(\frac{x-\mu}{\sigma})$. Notice that $\frac{1}{|\sigma|}\, g(\frac{x-\mu}{\sigma})$ remains a MOP even if μ is a variable (and not a constant) as long as σ is a non-zero constant. In the Subsection 3.6, we will exploit this fact to find MOP approximations of CLG PDFs in two and higher dimensions from a MOP approximation of the univariate standard normal PDF.

3.5 Fitting MOPs to One-dimensional Log-normal PDFs

In this subsection, we find a MOP approximation of a one-dimensional log-normal PDF. S_1 is said to have log-normal PDF with parameters μ and σ^2, written as $S_1 \sim LN(\mu, \sigma^2)$, if $\ln(S_1) \sim N(\mu, \sigma^2)$.

Suppose $S_1 \sim LN(\mu, \sigma^2)$, where $\mu = \ln(40) + 0.00074$, and $\sigma^2 = 0.13229^2$ (these parameters are taken from an *American Put Option* problem described in [21]). We will find a MOP approximation of the PDF of S_1 on the domain $(e^{\mu-3\sigma}, e^{\mu+3\sigma}) = (27.03, 59.28)$ (that will capture 99.73% of the total probability). If we try to fit a 1-piece MOP on the domain $(27.03, 59.28)$, we end up with a 8-degree MOP using the procedure described in Section 3.3. So we partition the domain into two pieces $(27.03, 39.34)$, $[39.34, 59.28)$, where $39.34\,(= e^{\mu-\sigma^2})$ is the mode of the PDF. Given these two intervals, the procedure described in Section 3.3 results in a 2-piece, 5-degree MOP as follows:

$$g_2(x) = \begin{cases} -31.17 + 4.75z - 0.29z^2 + 0.0085z^3 - 0.00012z^4 + 7.11 \times 10^{-7} z^5 & \text{if } 27.03 < x < 39.34 \\ -49.56 + 4.85z - 0.19z^2 + 0.0036z^3 - 0.000034z^4 + 1.28 \times 10^{-7} z^5 & \text{if } 39.34 \leq x < 59.28 \end{cases} \tag{3.9}$$

Let $f_{S_1}(\cdot)$ denote the PDF of $LN(\ln(40) + 0.00074, 0.13229^2)$ truncated to $(27.03, 59.28)$. A graph of $g_2(\cdot)$ overlaid on the graph of $f_{S_1}(\cdot)$ is as shown in Figure 3.

The goodness of fit statistics for $g_2(\cdot)$ are as follows (G_2 is the CDF corresponding to PDF g_2, and , and F_{S_1} is the CDF corresponding to f_{S_1}).

$$\begin{aligned} KL(f_{S_1}, g_2) &\approx 8.65 \times 10^{-6} \\ MAD(f_{S_1}, g_2) &\approx 0.00030 \\ MAD(F_{S_1}, G_2) &\approx 0.00034 \\ AEM(f_{S_1}, g_2) &\approx 0.00096 \\ AEV(f_{S_1}, g_2) &\approx 0.00398 \end{aligned}$$

3.6 Fitting MOPs to Higher-dimensional CLG PDFs

In this subsection, we will find MOP approximations of the PDFs of two- and three-dimensional conditional linear Gaussian (CLG) distributions. Our revised definition of

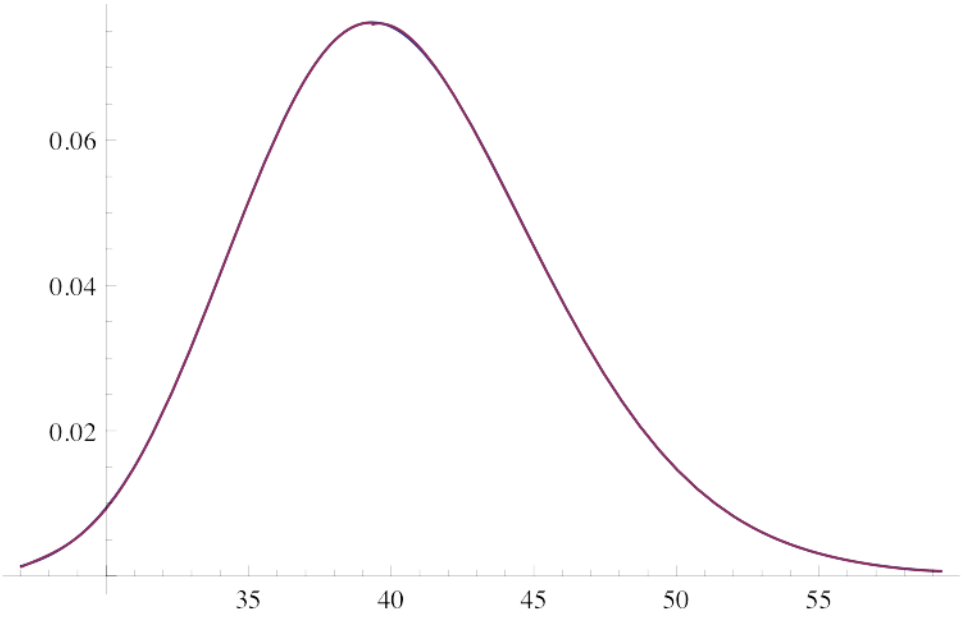

Figure 3: A graph of $g_2(\cdot)$ (in red) overlaid on the graph of $f_{S_1}(\cdot)$ (in blue)

multi-dimensional MOP functions in Equation (2.4) facilitates the task of finding MOP approximations of the PDFs of CLG conditional distributions.

3.6.1 Two-dimensional CLG Distributions

Consider the CLG conditional distribution $Y|z \sim N(z, 1)$, where $Z \sim N(0, 1)$. As in the one-dimensional case, we will find a MOP approximation of the conditional PDF of $Y|z$ on the two-dimensional region $z - 3 < y < z + 3$.

In Shenoy and West [17], a 12-piece, 14-degree MOP approximation is found by covering the two-dimensional region $-3 < z < 3, z - 3 < y < z + 3$ by 12 squares (hypercubes in two dimensions), and then by using two-dimensional Taylor series approximation at the mid-point of each square.

Here, we can use the one-dimensional 2-piece, 3-degree MOP approximation $g_1(z)$ of the standard normal distribution as follows. Let $h_1(z, y)$ denote a MOP approximation of the conditional PDF of $Y|z$. Then,

$$h_1(z, y) = g_1(y - z) \tag{3.10}$$

It follows from the remark at the end of Subsection 3.4, that $h_1(z, y)$ as defined in Equation (3.10) represents a MOP approximation of the PDF of $N(z, 1)$. Since $g_1(z)$ is a PDF, it follows that $h_1(z, y)$ is a PDF, i.e., $h_1(z, y) \geq 0$, and $\int_{-\infty}^{\infty} h_1(z, y)\, \mathrm{d}y = 1$ for all z. Notice that the two pieces of $h_1(z, y)$ are not defined on hypercubes, but rather on hyper-rhombuses (since we now have regions such as $-3 < y - z \leq 0$, etc). A three-dimensional plot of $h_1(z, y)$ is shown in Figure 4.

Since we are using the one-dimensional MOP approximation $g_1(z)$, the goodness of fit of $h_1(z, y)$ is same as that of $g_1(z)$. One question is how long does it take to integrate MOPs that are defined on hyper-rhombuses? To test this, we do two simple experiments.

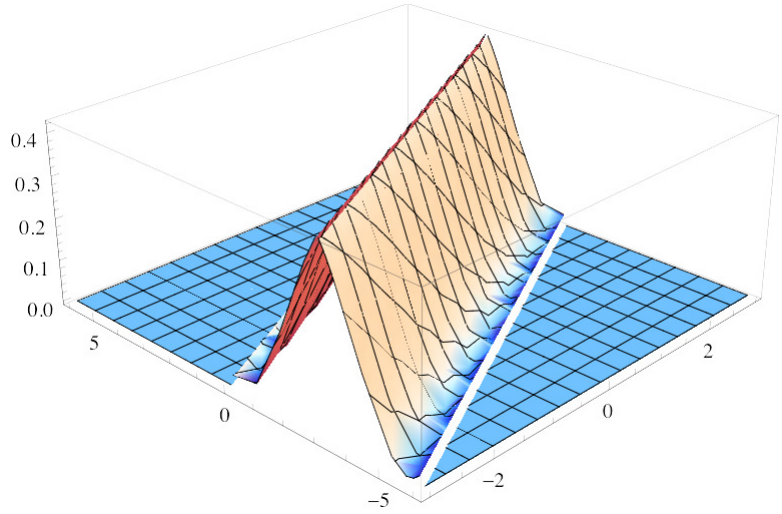

Figure 4: A three-dimensional plot of $h_1(z, y)$

All times reported here were obtained using the *Timing* command in *Mathematica*©, v. 8.0.4.0, running on a MacBook Air laptop computer (with 1.8 GHz Intel Core i7 CPU and 4 GB 1333 MHz DDR3 RAM). Also, the integrations were done exactly (and not numerically) resulting in a MOP function that can be used for down-stream calculations.

First, we compute the marginal PDF of Y as follows. $g_1(z)\, h_1(z, y)$ represents a MOP approximation of the joint PDF of (Z, Y). To find the marginal PDF of Y, we integrate Z out of the joint. Thus, a MOP approximation of the marginal PDF of Y is given by:

$$h_2(y) = \int_{-\infty}^{\infty} g_1(z)\, h_1(z, y)\, \mathrm{d}z \tag{3.11}$$

It takes Mathematica© ≈ 3.2 seconds to do the multiplication and integration in Equation (3.11), and $h_2(y)$ is computed as a 5-piece, 7-degree MOP function on the domain $(-6, 6)$. The exact joint distribution of Y is $N(0, 2)$. Let $f_Y(y)$ denote the exact PDF of $N(0, 2)$ truncated to $(-6, 6)$. A plot of $h_2(y)$ overlaid on the plot of $f_Y(y)$ is shown in Figure 5. The goodness of fit statistics between $f_Y(\cdot)$ and $h_2(\cdot)$ are as follows:

$$\begin{aligned} KL(f_Y, h_2) &\approx 0.0017 \\ MAD(f_Y, h_2) &\approx 0.0030 \\ AEM(f_Y, h_2) &\approx 0.0000 \\ AEV(f_Y, h_2) &\approx 0.0931 \end{aligned}$$

Second, consider the Bayesian network as shown in Figure 6 that includes W with a deterministic conditional, $W = Z + Y$. Suppose we use $g_1(z)$ as a MOP approximation of $N(0, 1)$, and $h_1(z, y)$ as a MOP approximation of $N(z, 1)$. The marginal distribution of W is then given by the convolution formula:

$$h_3(w) = \int_{-\infty}^{\infty} g_1(z)\, h_1(z, w - z)\, \mathrm{d}z \tag{3.12}$$

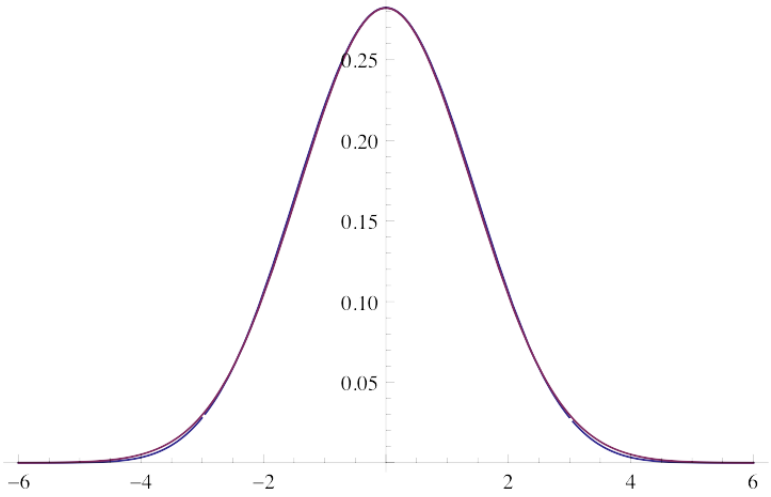

Figure 5: A graph of $h_2(y)$ (in red) overlaid on the graph of $f_Y(y)$ (in blue)

It takes Mathematica© ≈ 5.9 seconds to do the multiplication and integration in Equation (3.12). h_3 is computed as a 8-piece, 7-degree MOP function on the domain $(-9, 9)$. The exact marginal distribution of W is $N(0, 5)$. Let $f_W(w)$ denote the exact PDF of $N(0, 5)$ truncated to $(-9, 9)$. A plot of $h_3(w)$ overlaid on the the plot of $f_W(w)$ is shown in Figure 7. The goodness of fit statistics between $f_W(\cdot)$ and $h_3(\cdot)$ are as follows:

$$\begin{aligned} KL(f_W, h_3) &\approx 0.0027 \\ MAD(f_W, h_3) &\approx 0.0030 \\ AEM(f_W, h_3) &\approx 0.0000 \\ AEV(f_W, h_3) &\approx 0.2299 \end{aligned}$$

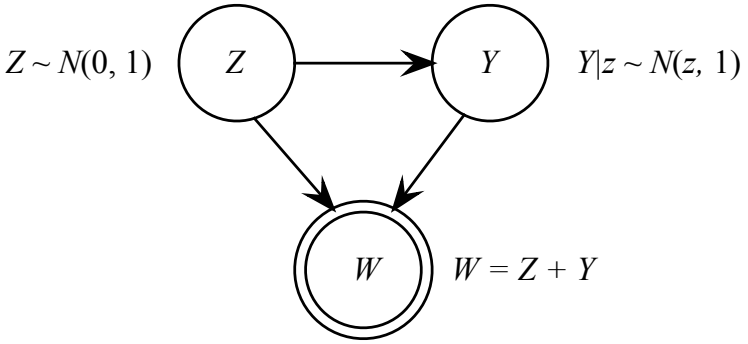

Figure 6: A Bayesian network with a sum deterministic conditional

3.6.2 Three-dimensional CLG Distributions

Suppose $Z \sim N(0, 1), Y|z \sim N(z, 1)$, and $X|(z, y) \sim N(z + y, 1)$. Notice that the conditional PDF of X is in three dimensions. As in the two-dimensional case, we find a

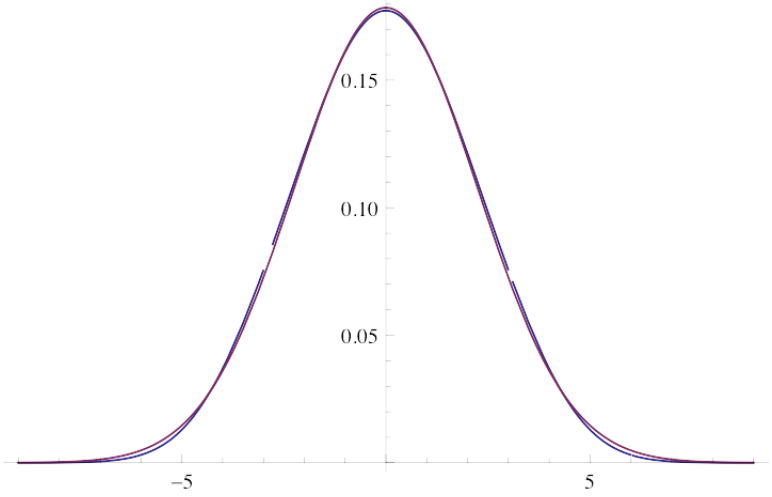

Figure 7: A graph of $h_3(w)$ (in red) overlaid on the graph of $f_W(w)$ (in blue)

MOP approximation $h_4(z, y, x)$ of the PDF of $N(z + y, 1)$ in the three-dimensional region $z+y-3 < x < z+y+3$ by using the 2-piece, 3-degree MOP approximation $g_1(z)$ for $N(0, 1)$ as follows:

$$h_4(z, y, x) = g_1(x - (z + y)) \tag{3.13}$$

Notice that the 2 pieces of h_4 are defined on regions $-3 < x - (z + y) < 0$, etc. Therefore, h_4 is a MOP by our definition in Equation (2.4).

As in the two-dimensional case, we will investigate how long it takes to integrate a MOP that is defined on a hyper-rhombus. First, we will compute the marginal PDF of X as follows. $g_1(z)$ denotes a MOP approximation of the marginal PDF of Z, $h_1(z, y)$ denotes a MOP approximation of the conditional PDF of $Y|z$, and $h_4(z, y, x)$ denotes a MOP approximation of the conditional PDF of $X|(y, z)$. Thus, $g_1(z)\, h_1(z, y)\, h_4(z, y, x)$ denotes a MOP approximation of the joint PDF of (Z, Y, X). Thus, a MOP approximation of the marginal PDF of X is given by:

$$\begin{aligned} h_6(x) = &\quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g_1(z)\, h_1(z, y)\, h_4(z, y, x)\, \mathrm{d}y\, \mathrm{d}z \\ = &\quad \int_{-\infty}^{\infty} g_1(z) \left(\int_{-\infty}^{\infty} h_1(z, y)\, h_4(z, y, x)\, \mathrm{d}y \right) \mathrm{d}z \end{aligned} \tag{3.14}$$

The integration in Equation (3.14) was done in two stages in Mathematica©. The inner integral (with respect to y) required $\approx$ 38.8 seconds, and the outer integral (with respect to z) required $\approx$ 24.8 seconds, and resulted in a 12-piece, 11-degree MOP on the interval $(-12, 12)$. Thus, the two multiplications and the two integrations in Equation (3.14) require a total of approximately 63.6 seconds. The exact distribution of X can be shown to be $N(0, 6)$. Let $f_X(\cdot)$ denote the PDF of $N(0, 6)$ truncated to the region $(-12, 12)$. A graph of $h_6(x)$ overlaid on the graph of $f_X(x)$ is shown in Figure 8. The goodness of fit statistics for

h_6 with respect to f_X are as follows:

$$\begin{aligned} KL(f_X, h_6) &\approx 0.0017 \\ MAD(f_X, h_6) &\approx 0.0018 \\ AEM(f_X, h_6) &\approx 0.0000 \\ AEV(f_X, h_6) &\approx 0.2816 \end{aligned}$$

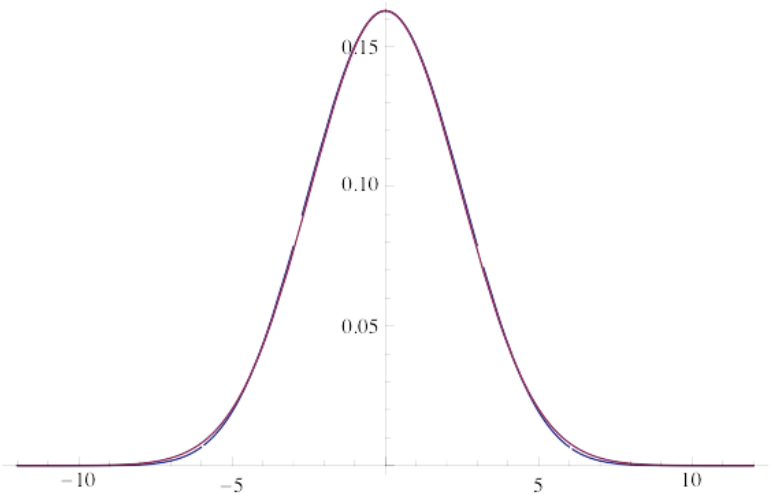

Figure 8: A graph of $h_6(x)$ (in red) overlaid on the graph of $f_X(x)$ (in blue)

Second, consider the Bayesian network as shown in Figure 9 that includes V with a deterministic conditional, $V = Z+Y+X$. Suppose we use $g_1(z)$ as a MOP approximation of $N(0,1)$, $h_1(z,y)$ as a MOP approximation of $N(z,1)$, and $h_4(z,y,x)$ as a MOP approximation of $N(z+y,1)$. The marginal distribution of V is then given by the convolution formula:

$$\begin{aligned} h_8(v) = &\quad \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g_1(z)\, h_1(z,y)\, h_4(z,y,v-z-y)\, \mathrm{d}y\, \mathrm{d}z \\ = &\quad \int_{-\infty}^{\infty} g_1(z) \left(\int_{-\infty}^{\infty} h_1(z,y)\, h_4(z,y,v-z-y)\, \mathrm{d}y\right) \mathrm{d}z \end{aligned} \tag{3.15}$$

The integration in Equation (3.15) was done in two stages in Mathematica©. The inner integral (with respect to y) required approximately 59.6 seconds, and resulted in a 9-piece, 7-degree MOP. The outer integral (with respect to z) required 38.7 seconds, and resulted in a 21-piece, 11-degree, MOP on the interval $(-21, 21)$. Thus, the two multiplications and the two integrations in Equation (3.15) require a total of approximately 98.3 seconds. The exact marginal distribution of V is $N(0, 21)$. Let $f_V(v)$ denote the exact PDF of $N(0, 21)$ truncated to $(-21, 21)$. A plot of $h_8(w)$ overlaid on the the plot of $f_V(v)$ is shown in Figure 10. The goodness of fit statistics for $h_8(v)$ with respect to $f_V(v)$ are as follows:

$$\begin{aligned} KL(f_V, h_8) &\approx 0.0024 \\ MAD(f_V, h_8) &\approx 0.0004 \\ AEM(f_V, h_8) &\approx 0.0000 \\ AEV(f_V, h_8) &\approx 0.9841 \end{aligned}$$

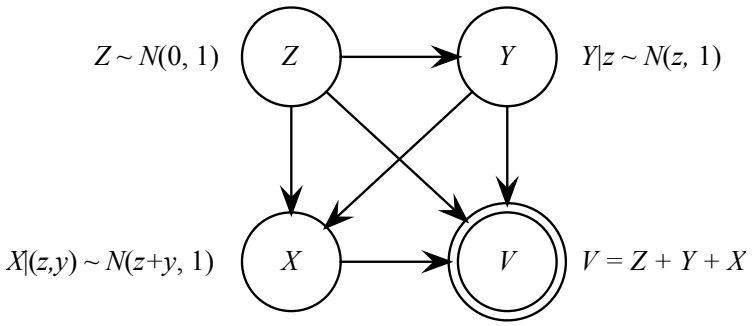

Figure 9: A Bayesian network with a three-dimensional conditional PDF

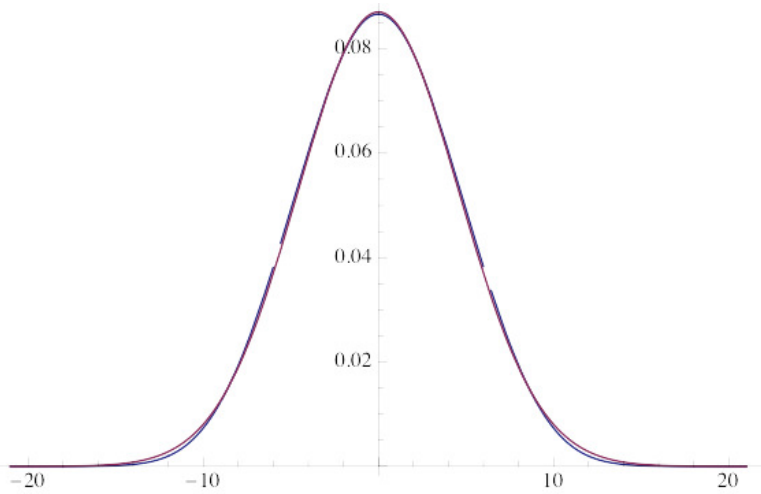

Figure 10: A graph of $h_8(v)$ (in red) overlaid on the graph of $f_V(v)$ (in blue)

The characteristics and approximate times required for computation of h_2, h_3, h_6, and h_8 are summarized in the following table.

MOP	*Description*	*# pieces*	*Degree*	*Computation time* (secs)
h_2	marg. of Y in Fig. 6	5	7	3.2
h_3	marg. of W in Fig. 6	8	7	5.9
h_6	marg. of X in Fig. 9	12	11	63.6
h_8	marg. of V in Fig. 9	21	11	98.3

In summary, the hyper-rhombus condition enables us to easily represent CLG conditionals in high dimensions. The computational cost of integrating a high-dimensional MOP function with a hyper-rhombus condition does not seem high for two or three-dimensional CLG distributions, and there is no loss of precision compared to one-dimensional conditionals. In Section 4, we compare the efficiencies of MOPs defined on hyper-rhombus regions with MOPs defined on hypercube regions.

3.7 Fitting MOPs to a Two-dimensional Log-normal PDF

In this section, we describe the construction of a MOP for a two-dimensional log-normal PDF. Suppose $S_2|s_1 \sim LN(\ln(s_1) + 0.00074, 0.13229^2)$, where $S_1 \sim LN(\ln(40) + 0.00074, 0.13229^2)$ (these parameters are from the *American Put Option* problem described in [21]).

A MOP approximation of the PDF of S_1 was described in Section 3.5 on the domain $(26.92, 39.34) \cup [39.34, 59.53)$. If $S_1 = 26.92$, the $(e^{\mu-3\sigma}, e^{\mu+3\sigma})$ domain of S_2 is $(18.11, 40.06)$. If $S_1 = 39.34$, the corresponding domain of S_2 is $(26.47, 58.54)$, and if $S_1 = 59.53$, it is $(40.06, 88.60)$. We divide the entire domain of $S_2|s_1$ into eight regions as shown in Figure 11. Notice that four of the corner regions are triangular-shaped hyper-rhombus regions, and the other four in the middle are equal-height hypercubes. These regions were selected by trial and error. We initially fitted the domain of $S_2|s_1$ by eight hypercubes. The degree of the resulting MOP was too high (18) since the four corner hypercube regions extend far beyond the 6-sigma limits (on a log-scale). Using the hyper-rhombus condition to stay within the 6-sigma limits allows us to find a MOP with a much lower-degree. Using the general procedure described in Section 3.3, we found an 8-piece, 5-degree MOP $g_3(s_1, s_2)$ for the PDF of $S_2|s_1$. A three-dimensional plot of the MOP approximation $g_3(s_1, s_2)$ is shown in Figure 12.

How good is this MOP approximation? We compute the marginal PDF of S_2 as follows:

$$g_4(s_2) = \int_{-\infty}^{\infty} g_2(s_1)\, g_3(s_1, s_2)\, \mathrm{d}s_1, \tag{3.16}$$

where g_2 is the MOP approximation of the PDF of S_1 as described in Equation (3.9). It takes Mathematica© $\approx$ 12.4 seconds to do the multiplication and integration in Equation (3.16). $g_4(\cdot)$ is computed as a 8-piece, 11-degree MOP on the domain $(18.11, 88.60)$. The exact marginal distribution of S_2 is $LN(\ln(40) + 2 \times 0.00074, 2 \times 0.13229^2)$. Let $f_{S_2}(\cdot)$ denote the exact PDF of S_2 truncated to $(18.11, 88.60)$. A plot of $g_4(\cdot)$ overlaid on the plot of $f_{S_2}(\cdot)$

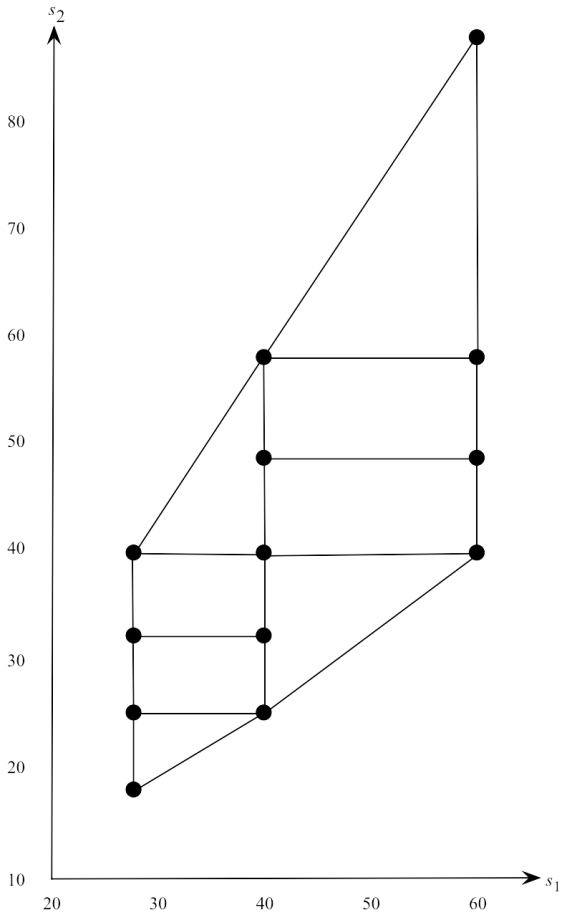

Figure 11: The region over which a MOP approximation of $S_2|s_1$ is computed

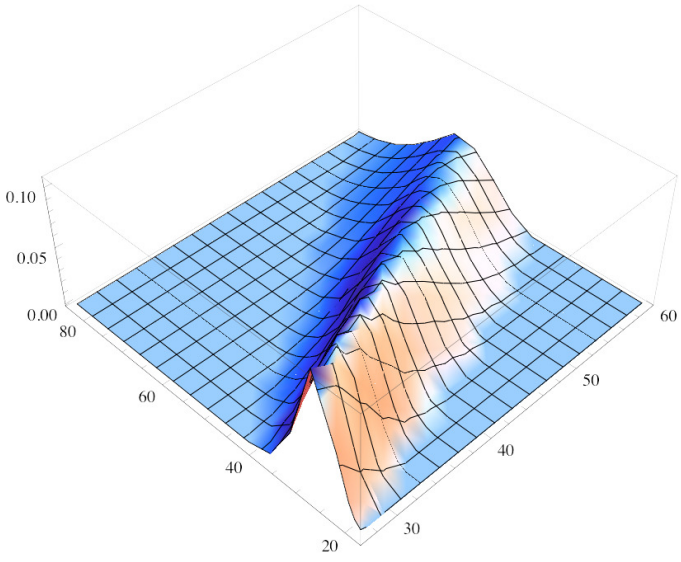

Figure 12: A three-dimensional plot of $g_3(s_1, s_2)$

is shown in Figure 13. The goodness of fit statistics between g_4 and f_{S_2} are as follows:

$$\begin{aligned} KL(f_{S_2}, g_4) &\approx 0.0007 \\ MAD(f_{S_2}, g_4) &\approx 0.0020 \\ AEM(f_{S_2}, g_4) &\approx 0.0141 \\ AEV(f_{S_2}, g_4) &\approx 1.6351 \end{aligned}$$

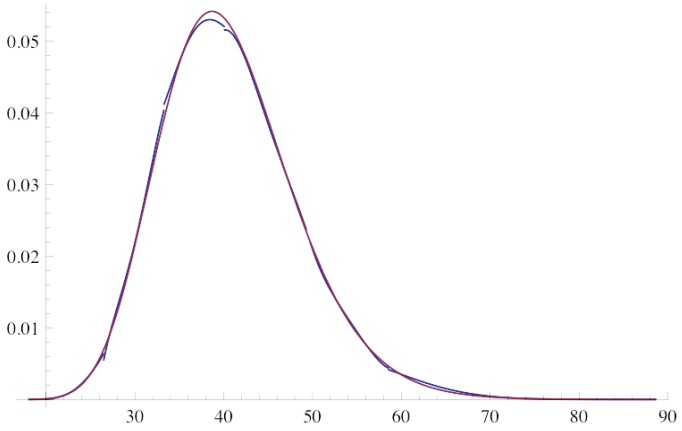

Figure 13: A plot of g_4 (in blue) overlaid on the plot of f_{S_2} (in red)

Unlike the CLG PDFs, the parameters μ and σ^2 of the log-normal PDF are shape parameters, and thus, each log-normal PDF has to approximated separately, i.e., we cannot use a linear transformation of one to find another. The hyper-rhombus condition enables us to find a MOP approximation of the conditional log-normal PDF using fewer pieces and lower degrees by avoiding regions far beyond the 6-sigma limits (on a log scale) where the PDF has extremely small values.

In Section 4, we discuss the tradeoffs between the hyper-rhombus and hypercube conditions in greater detail.

4 Comparing the Hyper-rhombus Condition with the Hypercube Condition

In this section, we will compare the hyper-rhombus condition defined in Equation (2.5) to define multi-dimensional MOP functions with the hypercube condition used by Shenoy and West [17]. As we saw in Subsection 3.6, the hyper-rhombus condition allows us to use a MOP approximation of one-dimensional standard normal PDF to define MOP approximations of higher-dimensional CLG PDFs.

Shenoy and West [17] describe a 12-piece, 14-degree, MOP approximation of the CLG distribution $Y \sim N(z, 1)$, where $Z \sim N(0, 1)$. Using 6-sigma limits, the CLG distribution of $Y|z$ is defined on the two-dimensional hyper-rhombus region $-3 < z < 3, z - 3 < y < z + 3$. This region is covered by 12 rectangles as shown in Figure 14, and a MOP approximation

is found for each rectangle using the Taylor series expansion at the mid-point of each rectangle to degree 7 in y and degree 7 in z. Why do we need such high degrees? Since the rectangles extend far outside the 6-sigma limits, we need to find a 14-degree Taylor series approximation to ensure non-negativity. For example, at the point $(z, y) = (-3, 2)$, we are 5 standard deviations away from the mean $z = -3$. Why 12-pieces? If we use fewer pieces, then we extend more outside the 6-sigma limits and the degree of the MOP approximation increases. We could use more pieces to lower the degree of the approximation, but this increases the computational effort of integrating such functions. Finally, it is not possible to normalize the 12-piece, 14-degree MOP approximation since the normalization constant would be a function of z, and normalizing it would make the normalized function a non-polynomial. Therefore, Shenoy and West [17] describe a "partial" normalization of this MOP approximation, which suffices for all purposes. In the case of MOP approximation using the hyper-rhombus condition, there is no need for normalization since we use the MOP approximation of the one-dimensional standard normal PDF, which is already normalized.

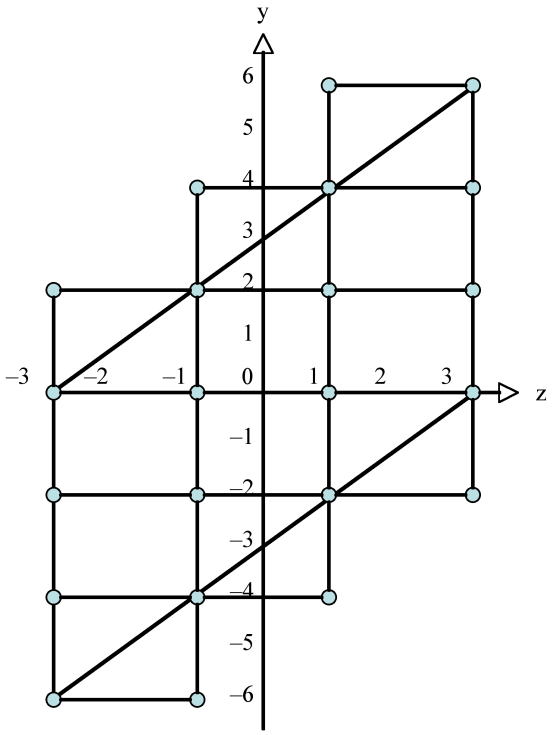

Figure 14: A graph of the domain of the MOP approximation of $Y|z$

How (in)efficient is the 12-piece, 14-degree, MOP approximation of $Y|z$ denoted by $r_1(z, y)$? We did two experiments. First, we compute the marginal PDF of Y. Second, we compute the marginal PDF of $W = Z + Y$ as shown in Figure 6. In both experiments, we note down the time required and the accuracy of the resulting MOP approximations, and compare these with the time required and the accuracy of the MOP approximations using the hyper-rhombus condition.

In the first experiment, we compute the marginal distribution of Y as described in Equation (3.11) except that we replace $h_1(z, y)$ by $r_1(z, y)$ and $h_2(y)$ by $r_2(y)$. $r_2(y)$ is computed

as a 8-piece, 7-degree MOP. Computing $r_2(y)$ takes ≈ 83.9 seconds (compared to ≈ 3.2 seconds when using $h_1(z, y)$. One reason for the increased time is the higher number of pieces in $r_1(z, y)$ (12 in $r_1(z, y)$ instead of 2 in $h_1(z, y)$). Another reason is the higher degree of the approximation (14 in $r_1(z, y)$ compared to 3 in $h_1(z, y)$).

The accuracy of the resulting MOP approximation of the marginal distribution of Y is as follows (shown with the accuracy of h_2, both with respect to f_Y).

Error	h_2	r_2
KL	0.0017	0.0004
MAD	0.0030	0.0016
AEM	0.0000	0.0000
AEV	0.0931	0.0505

In comparison, we notice that the two accuracies are the same orders of magnitude and r_2 has a slightly better accuracy than h_2, probably because of higher number of pieces and higher degree.

In the second experiment, we compute the marginal distribution of W as described in Equation (3.12), except that we replace $h_1(z, w - z)$ by $r_1(z, w - z)$, and $h_3(w)$ by $r_3(w)$. Notice that although $r_1(z, y)$ is defined on hypercubes, $r_1(z, w-z)$ is not. $r_3(w)$ is computed as a 24-piece, 18-degree MOP. Computing $r_3(w)$ takes ≈ 425.2 seconds (compared to ≈ 5.9 seconds when using $h_1(z, y)$). The accuracy of the fit of r_3 is as follows, shown with the accuracy of h_3 for comparison (both with respect to f_W):

Error	h_3	r_3
KL	0.0027	0.0020
MAD	0.0030	0.0031
AEM	0.0000	0.0000
AEV	0.2299	0.1875

The accuracy of r_3 is comparable to the corresponding accuracy of h_3.

Moral *et al.* [13] has suggested a "mixed tree" method for constructing mixture of truncated exponentials (MTE) approximation of high dimensional conditional distributions. Here we will use their method for constructing MOP approximations of the PDF of $Y|z \sim N(z, 1)$ where the regions are hypercubes, and we will compare such MOP approximations to the MOP approximation using hyper-rhombus described in Subsection 3.6.

Consider the CLG distribution of $Y|z \sim N(z, 1)$, where $Z \sim N(0, 1)$. We partition the domain of Z into 3 equal pieces: $(-3, -1], (-1, 1], (1, 3)$. In each piece, we approximate the conditional distribution of $Y|z$ as normal where the mean is constant and equal to the mid-point of the piece, and the variance is 1. Thus, we get a MOP approximation of $N(z, 1)$ as follows

$$t_1(z, y) = \begin{cases} g_1(y + 2) & \text{if } -3 < z < -1 \\ g_1(y) & \text{if } -1 \leq z < 1 \\ g_1(y - 2) & \text{if } 1 \leq z < 3 \end{cases} \tag{4.1}$$

where $g_1(\cdot)$ is the 2-piece, 3-degree, MOP approximation of $N(0, 1)$ as discussed in Section 3. Thus, $t_1(z, y)$ is a 6-piece, 3-degree MOP approximation of the PDF of $Y|z \sim N(z, 1)$

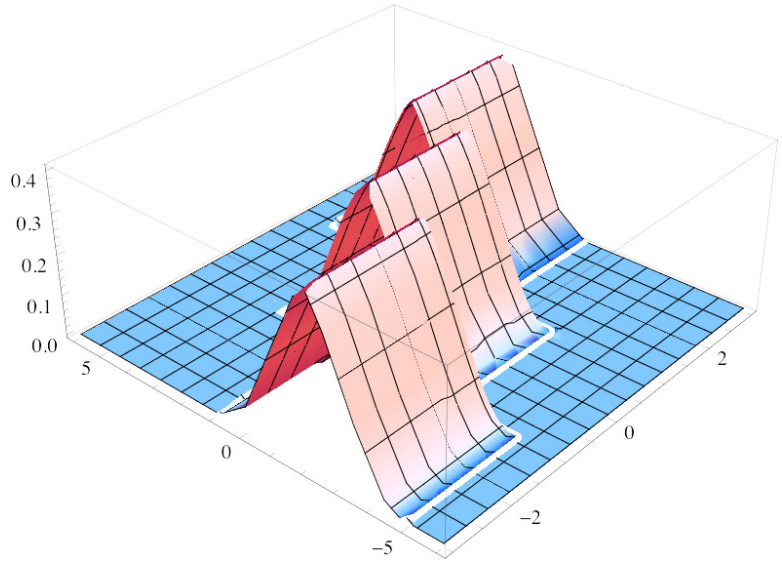

Figure 15: A three-dimensional plot of $t_1(z, y)$

assuming $Z \sim N(0, 1)$ where the pieces are defined on hypercubes. A three-dimensional plot of $t_1(z, y)$ is shown in Figure 15.

How efficient is the mixed-tree MOP approximation $t_1(z, y)$? We compute the marginal of Y as described in Equation (3.11) except that we replace $h_1(z, y)$ by $t_1(z, y)$, and replace $h_2(y)$ by $t_2(y)$. $t_2(y)$ is computed as a 8-piece, 3-degree MOP. Computation of $t_2(y)$ takes $\approx$ 2.1 seconds (compared to $\approx$ 3.5 seconds when using $h_1(z, y)$). The shorter time for multiplication and integration is probably due to the hypercube regions in $t_1(z, y)$. A graph of $t_2(y)$ overlaid on the graph of $f_Y(y)$, the marginal PDF of Y truncated on the domain $(-6, 6)$ is shown in Figure 16. The accuracy of $t_2(y)$ is as follows, which is slightly worse than the accuracy of $h_2(y)$.

Error	h_2	t_2
KL	0.0017	0.0014
MAD	0.0030	0.0138
AEM	0.0000	0.0000
AEV	0.0931	0.2651

Finally, we compute the marginal of W as described in Equation (3.12) except that we replace $h_1(z, y)$ by $t_1(z, y)$, and replace $h_3(w)$ by $t_3(w)$. $t_3(w)$ is computed as a 19-piece, 7-degree MOP. Computation of $t_3(w)$ takes $\approx$ 11.9 seconds (compared to $\approx$ 5.9 seconds when using $h_1(z, y)$). The longer time for computing $t_3(w)$ is probably due to the loss of the hypercube condition in the convolution formula in Equation (3.12), and due to the larger number of pieces in $t_1(z, y)$ (6 compared of 2 in $h_1(z, y)$). A graph of $t_3(w)$ overlaid on the graph of $f_W(w)$, the marginal PDF of W truncated on the domain $(-8, 8)$ is shown in Figure 17. The accuracy of t_3 is as follows (with respect to f_W)—in comparison with h_3, the results are mixed.

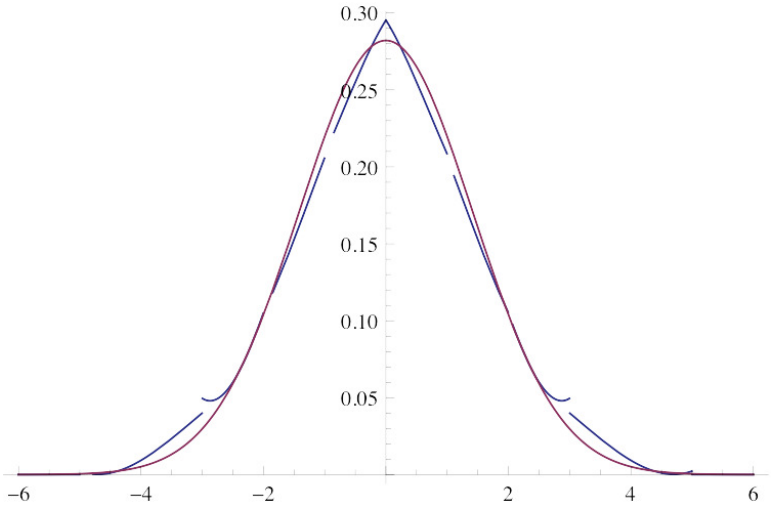

Figure 16: A graph of $t_2(y)$ (in blue) overlaid on the graph of $f_Y(y)$ (in red)

Error	h_3	t_3
KL	0.0027	0.0023
MAD	0.0030	0.0539
AEM	0.0000	0.0000
AEV	0.2299	0.1542

The following table summarizes our findings:

MOP	*Description*	*# pieces*	*Degree*	*Computation time* (secs)
h_2	marg. of Y using g_1, h_1	5	7	3.2
t_2	marg. of Y using g_1, t_1	8	3	2.1
h_3	marg. of W using g_1, h_1	19	7	5.9
t_3	marg. of W using g_1, t_1	21	11	11.9

In conclusion, the hyper-rhombus condition allows us to use the MOP approximation of the PDF of the one-dimensional standard normal distribution to construct MOP approximations of CLG distributions in two and higher dimensions. The cost of integrating a multi-dimensional MOP function on a hyper-rhombus appears to be small compared to the cost of constructing MOP approximations on hypercubes of two-dimensional CLG distributions, or by using the mixed tree approach of Moral *et al.* [13].

5 Comparing the LIP Method with the Taylor Series Method

In this section we compare the LIP method for constructing MOP approximations of PDFs with the Taylor series (TS) method suggested by Shenoy and West [17].

The two methods have very different characteristics. Using the LIP method, if we wish to get a better accuracy, we can either increase the number of pieces, or increase the degree by selecting more points, or both. With Chebyshev points, the LIP technique does not exhibit

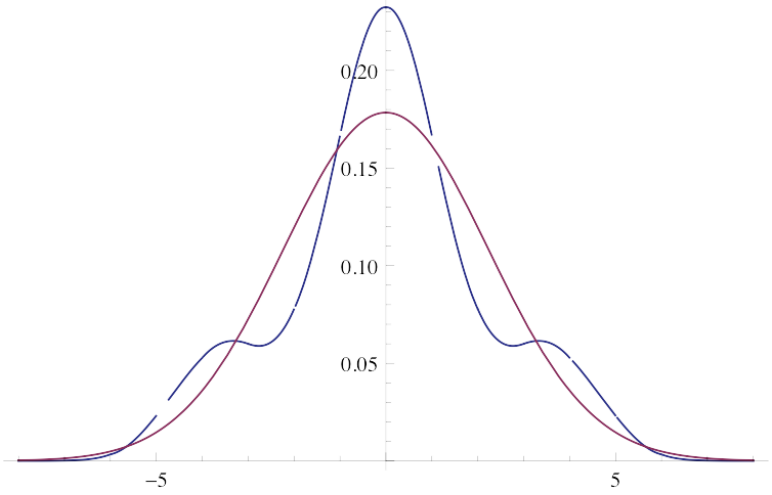

Figure 17: A graph of $t_3(w)$ (in blue) overlaid on the graph of $f_W(w)$ (in red)

oscillating behavior. On the other hand, using the TS method, if we wish to get better accuracy (around the mid-point of the interval where the Taylor series is calculated), we can just increase the degree of the approximation. But this may not always work. For example [19], consider $f(x) = \frac{1}{x}$. Taylor series polynomial of degree n at the point $x_0 = 1$ are as follows:

$$P_n(x) = \sum_{k=0}^{n} (-1)^k (x-1)^k$$

When we approximate $f(3) = 1/3$ by $P_n(3)$, the approximations become increasingly more inaccurate:

n	1	2	3	4	5	6	7
$P_n(3)$	-1	3	-5	11	-21	43	-85

When we are trying to fit a polynomial in an interval containing the tail of a PDF, non-negativity is an important requirement. We can guarantee non-negativity of a LIP polynomial approximation of a PDF by finding a one-degree LIP polynomial. A one-degree LIP polynomial is a linear interpolation between two end-points of an interval, and thus, is guaranteed to be non-negative. Of course, the accuracy may not be good, but we can improve the accuracy by having smaller pieces. On the other hand, to guarantee non-negativity of a TS polynomial, we have to increase the degree of the polynomial. This is the reason why we end up with a 14-degree MOP approximation of a two-dimensional CLG distribution. If the degree of the polynomial get very high, then we have to use exact arithmetic (integers for the non-transcendental coefficients) for stability reasons, and this increases the time required for multiplication and integration of polynomials.

An n-degree polynomial found using the LIP method may contain as many as 2^n terms in the polynomial, where as an n-degree polynomial found using TS method has at most $n+1$ terms. When we are using floating point numbers for the coefficients, we can consolidate the 2^n terms in the LIP polynomial to $n+1$ by addition. However, when we are using exact arithmetic (integers for the non-transcendental coefficients), we may not be able to

consolidate the terms, and in this case, the high number of terms may cause an increase of time needed for multiplication and integration of such polynomials.

How does the accuracy of the LIP polynomial compare with the accuracy of the TS polynomial? We will answer this question by examining the polynomials produced with the two methods under similar conditions (same regions) for the standard normal PDF. In Section 3, we described a 2-piece, 3-degree MOP approximation for the standard normal PDF using the LIP method, which is denoted by $g_1(z)$. The two regions where the pieces are defined are $(-3, 0), [0, 3)$. Using these same regions, we find a 3-degree polynomial using the TS method. In each region, we compute the TS polynomial at the mid-point of the region to degree 3. We make sure the values of polynomial pieces are non-negative, and we normalize the two pieces so that the total area of the MOP in the region $(-3, 3)$ is one. Let $s_1(z)$ denote the TS MOP. The accuracy of the two MOPs (with respect to φ) are as follows:

Error	g_1	s_1
KL	0.0086	0.0370
MAD	0.0140	0.0305
AEM	0.0000	0.0000
AEV	0.0203	0.3080

We observe that $g_1(z)$ is a more accurate approximation of $\varphi(z)$ than $s_1(z)$.

6 Summary and Discussion

A major contribution of this paper is a re-definition of multi-dimensional mixture of polynomials so that the regions where the polynomials are defined are hyper-rhombuses instead of hypercubes. This re-definition has three major advantages. First, it ensures that MOP functions are closed operations needed for multi-dimensional linear conditionals, which was not true when MOP functions were defined on regions that are hypercubes. Second, it allows us to use the MOP approximation of a one-dimensional standard normal PDF to define MOP approximations of high-dimensional CLG PDFs. Third, it allows us to construct MOP approximations of two-dimensional conditional PDFs that have fewer pieces and lower degrees by avoiding regions where the PDFs have very small values.

Another contribution is the use of Lagrange interpolation polynomial with Chebyshev points to construct MOP approximations of PDFs. There are some advantages to using the LIP method as compared to the Taylor series method. The PDFs do not have to be differentiable, and the MOP approximations produced by the LIP method have a better fit that those produced by the TS method for the case of the standard normal PDF. The Lagrange interpolation polynomial described for one-dimensional functions generalizes to multi-dimensional functions. Also, the Chebyshev points defined for one-dimensional functions have been generalized for two-dimensional functions [22].

A disadvantage of the hyper-rhombus condition is that in the process of propagating MOP potentials, the number of pieces and degrees of the intermediate potentials can get large. One solution to this (described in [23]) is to re-approximate the intermediate MOP potentials using fewer pieces and lower degrees. For example, the 21-piece, 11-degree MOP

h_8 (described in Section 3.6) can be easily re-approximated by a 2-piece, 3-degree MOP using the general procedure described in Section 3.3.

The methods described here apply for CLG PDFs in one to three dimensions, and non-CLG PDFs in one or two dimensions. Beyond that, our methods do not scale up either computationally (in the case of CLG PDFs) or for finding MOP approximations (in the case of non-CLG PDFs). Further research is needed in this regard.

Finally, an important issue is the complexity of solving hybrid BNs where the potentials are all represented by MOPs. What is the size of hybrid BNs that can be solved in a reasonable amount of time? Shenoy et al. [23] reports some results in solving a small hybrid BN with 12 variables (2 discrete, 5 continuous with non-deterministic conditionals, and 5 continuous with linear deterministic conditionals). This is yet another topic for further research.

Acknowledgements

I am grateful to Prof. Eric Neufeld of University of Saskatoon for suggesting that I look at the literature on computer graphics for modeling three-dimensional objects. Thanks also to James C. West for suggesting the use of Chebyshev points, and to Yijing Li for correcting several errors in the manuscript. This paper was revised following a sabbatical during Spring 2011 semester at the University of Almeria, where I collaborated with Antonio Salmerón and Rafael Rumí. This paper has benefitted considerably from our collaboration. A condensed version of one part of this manuscript has appeared as [24]. The paper has benefitted from comments and suggestions from two anonymous reviewers and the editor-in-chief of the *International Journal of Approximate Reasoning*.

References

[1] A. Kozlov, D. Koller, Nonuniform dynamic discretization in hybrid networks, in: D. Geiger, P. P. Shenoy (Eds.), Uncertainty in Artificial Intelligence: Proceedings of the Thirteenth Conference, Morgan Kaufmann, San Francisco, CA, 1997, pp. 302–313.

[2] V. Gogate, R. Dechter, Approximate inference for hybrid bayesian networks with discrete constraints, in: F. Bacchus, T. Jaakkola (Eds.), Uncertainty in Artificial Intelligence: Proceedings of the Twenty-First Conference, AUAI Press, Corvallis, OR, 2005, pp. 209–216.

[3] A. Salmerón, A. Cano, S. Moral, Importance sampling in Bayesian networks using probability trees, Computational Statistics and Data Analysis 34 (2000) 387–413.

[4] W. R. Gilks, S. Richardson, D. J. Spiegelhalter, Introducing Markov chain Monte Carlo, in: W. R. Gilks, S. Richardson, D. J. Spiegelhalter (Eds.), Markov Chain Monte Carlo in Practice, Chapman and Hall/CRC, Boca Raton, FL, 1996, Ch. 1, pp. 1–20.

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

[6] U. Lerner, R. Parr, Inference in hybrid networks: Theoretical limits and practical algorithms, in: J. Breese, D. Koller (Eds.), Uncertainty in Artificial Intelligence: Proceedings of the Seventeenth Conference, Morgan Kaufmann, San Francisco, CA, 2001, pp. 310–318.

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

[8] U. Lerner, E. Segal, D. Koller, Exact inference in networks with discrete children of continuous parents, in: J. Breese, D. Koller (Eds.), Uncertainty in Artificial Intelligence: Proceedings of the Seventeenth Conference, Morgan Kaufmann, San Francisco, CA, 2001, pp. 319–328.

[9] P. P. Shenoy, Inference in hybrid Bayesian networks using mixtures of Gaussians, in: R. Dechter, T. Richardson (Eds.), Uncertainty in Artificial Intelligence: Proceedings of the Twenty-Second Conference, AUAI Press, Corvallis, OR, 2006, pp. 428–436.

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

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

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

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

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

[15] A. Fernández, H. Langseth, T. D. Nielsen, A. Salmerón, Parameter learning in MTE networks using incomplete data, in: P. Myllmäki, T. Roos, T. Jaakola (Eds.), Proceedings of the Fifth European Conference on Probabilistic Graphical Models (PGM 2010), Helsinki, Finalnd, 2010, pp. 137–144.

[16] R. Rumí, A. Salmerón, Approximate probability propagation with mixtures of truncated exponentials, International Journal of Approximate Reasoning 45 (2) (2007) 191–210.

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

[18] E. Waring, Problems concerning interpolations, Philosophical Transactions of the Royal Society of London 69 (1779) 59–67.

[19] R. L. Burden, J. D. Faires, Numerical Analysis, 9th Edition, Brooks Cole, Boston, MA, 2010.

[20] S. Kullback, R. A. Leibler, On information and sufficiency, Annals of Mathematical Statistics 22 (1951) 76–86.

[21] Y. Li, P. P. Shenoy, A framework for solving hybrid influence diagrams containing deterministic conditional distributions, Decision Analysis 9 (1) (2012) in press.

[22] Y. Xu, Lagrange interpolation on Chebyshev points of two variables, Journal of Approximation Theory 87 (2) (1996) 220–238.

[23] P. P. Shenoy, R. Rumí, A. Salmerón, Some practical issues in inference in hybrid Bayesian networks with deterministic conditionals, in: S. Ventura, A. Abraham, K. Cios, C. Romero, F. Marcelloni, J. M. Benitez, E. Gibaja (Eds.), Proceedings of the 2011 Eleventh International Conference on Intelligent Systems Design and Applications, IEEE Research Publishing Services, Piscataway, NJ, 2011, pp. 605–610.

[24] P. P. Shenoy, A re-definition of mixtures of polynomials for inference in hybrid bayesian networks, in: W. Liu (Ed.), Symbolic and Quantitative Approaches to Reasoning with Uncertainty – ECSQARU 2011, Lecture Notes in Artificial Intelligence 6717, Springer, Heidelberg, 2011, pp. 98–109.