Proceedings of the 1996 Winter Simulation Conference
ed. J. M. Charnes, D. J. Morrice, D. T. Brunner, and J. J. Swain

RECENT DEVELOPMENTS IN INPUT MODELING WITH BÉZIER DISTRIBUTIONS

Mary Ann Flanigan Wagner

Boeing Information Services
7990 Boeing Court MS CV-82
Vienna, VA 22183-7000, U.S.A.

James R. Wilson

Department of Industrial Engineering
North Carolina State University
Raleigh, NC 27695-7906, U.S.A.

ABSTRACT

New methods are presented for estimating univariate and bivariate Bézier distributions. A likelihood ratio test is used to estimate the number of control points for a univariate Bézier distribution fitted to sample data. To estimate the control points of a bivariate Bézier distribution with fixed marginals based on either sample data or subjective information about the joint dependency structure, a linear-programming approach is formulated. These methods are implemented in the Windows-based software system called PRIME—PRobabilistic Input Modeling Environment. Several examples illustrate the application of these estimation procedures.

1 INTRODUCTION

One of the central problems in the design and construction of stochastic simulation experiments is the selection of valid input models—that is, probability distributions that accurately mimic the behavior of the random input processes driving the system. In many applications, it is critical not only to capture the shape of the marginal distribution of each major input random variable but also to accurately represent the stochastic dependencies between those variates.

Although many practitioners appreciate the need for valid models of multivariate simulation inputs, they lack effective and widely available tools for building such input models. Stanfield et al. (1996) developed a technique for fitting a multivariate distribution when the correlation matrix and the first four moments for each marginal distribution have been specified or estimated by the user. Because the fitted joint distribution is built from univariate marginals belonging to the Johnson translation system (Johnson 1949a; Swain, Venkatraman, and Wilson 1988), the multivariate input-modeling technique of Stanfield et al. has substantial flexibility. Unfortunately, the fitted joint distribution does not belong to the multivariate Johnson translation system (Johnson 1949b, Johnson 1987); moreover, the corresponding conditional distributions do not belong to the Johnson system—and this lack of "closure" makes it impossible to obtain convenient closed-form expressions for the conditional distributions that naturally arise in some applications.

Other approaches to multivariate input modeling can be based on TES (Transform-Expand-Sample) processes (Jagerman and Melamed 1992a, 1992b; Melamed, Hill, and Goldsman 1992) and ARTA (AutoRegressive To Anything) processes (Cario and Nelson 1996). Both methodologies enable the user to specify the autocorrelation function out to an arbitrary lag for a univariate stochastic process with a user-specified marginal distribution, but ARTA processes seem to be substantially easier to use. Unfortunately the conditional distributions associated with TES and ARTA processes do not appear to possess any advantages in analytical or numerical tractability when compared to multivariate processes based on the Johnson translation system.

In this paper we extend the input-modeling methodology of Wagner and Wilson (1993, 1994, 1995, 1996) for representing continuous univariate and bivariate populations using Bézier distributions. The remainder of this paper is organized as follows. In Section 2 we summarize the main properties of univariate and bivariate Bézier distributions, and we establish some basic notation that is used throughout the paper. In Section 3 we develop a likelihood ratio test to estimate the number of control points (that is, the number of parameters) for a univariate Bézier distribution fitted to sample data. To estimate a configuration of control points for a bivariate Bézier distribution with fixed marginals that will yield a desired covariance structure based either on sample data or subjective information, in Section 4 we present a fitting procedure that is formulated as a linear-programming problem. These methods are implemented in the Windows-based software system called PRIME—PRobabilistic Input Modeling Environment. A public-domain version of PRIME is available upon request; and in Section 5 an application of PRIME illustrates the effectiveness of these distribution-fitting techniques. Finally in Section 6 we summarize the main contributions of this work, and we make recommendations for future research.

2 OVERVIEW OF BÉZIER DISTRIBUTIONS

2.1 Definition of Bézier Curves

A Bézier curve of degree n with *control points* $\{\boldsymbol{p}_i \equiv (x_i, z_i)^{\mathrm{T}} : i = 0, 1, \ldots, n\}$ is given parametrically by

$$\boldsymbol{P}(t) = \boldsymbol{P}(t; n, \boldsymbol{x}, \boldsymbol{z}) = \sum_{i=0}^{n} B_{n,i}(t)\, \boldsymbol{p}_i \qquad (1)$$

for $t \in [0, 1]$, where $\boldsymbol{x} \equiv (x_0, x_1, \ldots, x_n)^{\mathrm{T}}$ and $\boldsymbol{z} \equiv (z_0, z_1, \ldots, z_n)^{\mathrm{T}}$ respectively denote the vectors of x- and z-coordinates of the control points $\{\boldsymbol{p}_i\}$, and where the *blending function* $B_{n,i}(t)$ is the Bernstein polynomial

$$B_{n,i}(t) \equiv \frac{n!}{i!\,(n-i)!}\, t^i (1-t)^{n-i} \qquad (2)$$

for $t \in [0, 1]$ and $i = 0, 1, \ldots, n$. (Throughout this paper, all vectors will be column vectors unless otherwise stated; and the roman superscript $^{\mathrm{T}}$ will denote the transpose of a vector or matrix so that each control point is understood to be a column vector. Moreover, we will use the simpler the notation $\boldsymbol{P}(t)$ in (1) when no confusion can arise from this usage.)

In the definition (1) of the Bézier curve, we note that the control points act like "magnets"; and the "magnetic attraction" exerted on the Bézier curve $\{\boldsymbol{P}(t) : t \in [0, 1]\}$ by the ith control point $\boldsymbol{p}_i$ is strongest at the value $t = i/n$ for the parameter t so that the corresponding point $\boldsymbol{P}(t)$ on the curve is "in the vicinity" of $\boldsymbol{p}_i$. If the weight (magnetic attraction) of a control point is 1, then the Bézier curve is forced to pass through that control point exactly (Farin 1990).

2.2 Definition of Bézier Surfaces

If selected control points are represented by the column vectors $\{\boldsymbol{q}_{i,j} \equiv (x_{i,j},\ y_{i,j},\ z_{i,j})^{\mathrm{T}} : i = 0, 1, \ldots, n_x;\ j = 0, 1, \ldots, n_y\}$, then the corresponding two-dimensional Bézier surface in three-dimensional Euclidean space is given parametrically as

$$\begin{aligned} \boldsymbol{Q}(t_x, t_y) &= \boldsymbol{Q}(t_x, t_y; n_x, n_y, \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z}) \\ &= \sum_{i=0}^{n_x} \sum_{j=0}^{n_y} B_{n_x,i}(t_x) B_{n_y,j}(t_y)\, \boldsymbol{q}_{i,j} \end{aligned} \qquad (3)$$

for all $t_x, t_y \in [0, 1]$, where $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ are the $(n_x + 1) \times (n_y + 1)$ matrices of the x-, y-, and z-coordinates of the given control points $\{\boldsymbol{q}_{i,j}\}$. As in (1) and elsewhere in this paper, we will suppress the dependence of Bézier functions like $\boldsymbol{Q}(t_x, t_y)$ on the parameters n_x, n_y, $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ when no confusion can arise from this simplification.

2.3 Univariate Bézier Distributions

In this subsection we summarize briefly some key properties of univariate Bézier distributions. For a detailed development of these properties, see Wagner and Wilson (1993, 1996). Given a continuous random variable X with bounded support $[x_*, x^*]$ and unknown cumulative distribution function (c.d.f.) $F_X(\cdot)$, we can approximate $F_X(\cdot)$ arbitrarily closely by a Bézier curve of the form (1) with sufficiently high degree n, where

$$\left.\begin{aligned} x(t) &= \sum_{i=0}^{n} B_{n,i}(t)\, x_i \\ F_X[x(t)] &= \sum_{i=0}^{n} B_{n,i}(t)\, z_i \end{aligned}\right\}, \qquad (4)$$

for all $t \in [0, 1]$. If $F_X(\cdot)$ is given parametrically by (4), then the corresponding probability density function (p.d.f.) $f_X(\cdot)$ is given parametrically by $x(t)$ in (4) together with

$$f_X[x(t)] = \frac{\sum_{i=0}^{n-1} B_{n-1,i}(t) \Delta z_i}{\sum_{i=0}^{n-1} B_{n-1,i}(t) \Delta x_i} \qquad (5)$$

for all $t \in [0, 1]$, where

$$\left.\begin{aligned} \Delta x_i &\equiv x_{i+1} - x_i \\ \Delta z_i &\equiv z_{i+1} - z_i \end{aligned}\right\} \text{ for } i = 0, 1, \ldots, n-1. \qquad (6)$$

In Wagner and Wilson (1993, 1996), we present several applications of the Bézier family of univariate distributions for modeling simulation inputs.

2.4 Bivariate Bézier Distributions

In this subsection we summarize briefly some key properties of bivariate Bézier distributions. For a detailed development of these properties, see Wagner and Wilson (1994, 1995). If $(X, Y)^{\mathrm{T}}$ is a continuous random vector with bounded support $[x_*, x^*] \times [y_*, y^*]$, unknown c.d.f. $F_{XY}(\cdot, \cdot)$, and unknown p.d.f. $f_{XY}(\cdot, \cdot)$, then we can approximate $F_{XY}(\cdot, \cdot)$ arbitrarily closely with an appropriate Bézier surface of the form (3) that has sufficiently large values of n_x and n_y (Farin 1990), where

$$x(t_x) = \sum_{i=0}^{n_x} B_{n_x,i}(t_x)\, x_i \quad \text{for all } t_x \in [0,\, 1], \tag{7}$$

$$y(t_y) = \sum_{j=0}^{n_y} B_{n_y,j}(t_y)\, y_j \quad \text{for all } t_y \in [0,\, 1], \tag{8}$$

and

$$F_{XY}[x(t_x),\, y(t_y)] = \sum_{i=0}^{n_x}\sum_{j=0}^{n_y} B_{n_x,i}(t_x) B_{n_y,j}(t_y)\, z_{i,j} \tag{9}$$

for all $t_x, t_y \in [0,\, 1]$. In Wagner and Wilson (1994, 1995) we describe special properties of the $\boldsymbol{x}$, $\boldsymbol{y}$, and $\boldsymbol{z}$ matrices that are required to ensure the validity of the parametric representation (7)–(9) of a bivariate probability distribution.

To represent the control points for each marginal distribution of $F_{XY}(\cdot, \cdot)$, we observe that the $\{x_i\}$ in (7) and the $\{y_j\}$ in (8) respectively serve as the x- and y-coordinates of the control points for the marginal c.d.f.'s $F_X(\cdot)$ and $F_Y(\cdot)$. We reserve the symbols $\{z_i^{(X)} : i = 0, 1, \ldots, n_x\}$ to denote the z-coordinates of the control points of $F_X(\cdot)$; and we reserve the symbols $\{z_j^{(Y)} : j = 0, 1, \ldots, n_y\}$ to denote the z-coordinates of the control points of $F_Y(\cdot)$.

If $F_{XY}(\cdot, \cdot)$ is given parametrically by Equations (7)–(9), then the corresponding bivariate density function $f_{XY}(\cdot, \cdot)$ is given parametrically by $x(t_x)$ in (7) and $y(t_y)$ in (8) together with

$$f_{XY}[x(t_x),\, y(t_y)] = \frac{\displaystyle\sum_{i=0}^{n_x-1}\sum_{j=0}^{n_y-1} B_{n_x-1,i}(t_x) B_{n_y-1,j}(t_y)\, \Delta_i \Delta_j z_{i,j}}{\left[\displaystyle\sum_{i=0}^{n_x-1} B_{n_x-1,i}(t_x)\, \Delta x_i\right]\left[\displaystyle\sum_{j=0}^{n_y-1} B_{n_y-1,j}(t_y)\, \Delta y_j\right]} \tag{10}$$

for all $t_x, t_y \in [0,\, 1]$, where

$$\left.\begin{aligned} \Delta_j z_{i,j} &= z_{i,j+1} - z_{i,j} \\ \Delta_i \Delta_j z_{i,j} &= z_{i+1,j+1} - z_{i,j+1} - z_{i+1,j} + z_{i,j} \end{aligned}\right\} \tag{11}$$

for $i = 0, 1, \ldots, n_x - 1$ and $j = 0, 1, \ldots, n_y - 1$.

2.5 Estimating Bézier Distributions Using PRIME

PRIME is a graphical Windows-based software system that is used to construct both univariate and bivariate Bézier distributions. PRIME is designed to be easy and intuitive to use. The construction of a Bézier distribution is performed through the actions of the mouse, and several options are conveniently available through menu selections. To manipulate a c.d.f., the user may move any of the control points by clicking on a chosen control point and then dragging that control point to the desired location by moving the mouse. Control points are represented as small black squares, and each control point is given a label corresponding to its index i in Equation (4). The user may also add or delete control points via the mouse and the keyboard. Any movement, addition or deletion of a control point causes the displayed distribution to be updated (nearly) instantaneously so that the user gets immediate feedback on the effects of editing that distribution. See Wagner and Wilson (1993, 1994, 1995, 1996) for more information on PRIME.

3 ESTIMATING THE NUMBER OF CONTROL POINTS

A likelihood ratio test is used to estimate the number of control points of a univariate Bézier distribution. If the random sample $\boldsymbol{X} \equiv \{X_j : j = 1, \ldots, m\}$ has been taken from a univariate Bézier p.d.f. $f_X(\cdot; n, \boldsymbol{x}, \boldsymbol{z})$ that is given parametrically by (5) with a known value of n, then the maximum likelihood estimates $\widehat{\boldsymbol{x}}_n$, $\widehat{\boldsymbol{z}}_n$ of the x- and z-coordinates of the control points $\{\boldsymbol{p}_i : i = 0, \ldots, n\}$ in (1) are obtained by solving the following nonlinear optimization problem:

$$\left.\begin{array}{ll} \underset{\boldsymbol{x},\,\boldsymbol{z}}{\text{min.}} & \mathcal{L}_n(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{X}) \equiv \displaystyle\prod_{j=1}^{m} f_X(X_j; n, \boldsymbol{x}, \boldsymbol{z}) \\ \text{s. t.} & f_X[x(t); n, \boldsymbol{x}, \boldsymbol{z}] > 0 \text{ for } t \in (0,\, 1) \\ & z_0 = 0 \\ & z_n = 1 \\ & x_0 < X_{(1)} \\ & x_n > X_{(m)} \end{array}\right\}, \tag{12}$$

where

$$X_{(1)} \le X_{(2)} \le \cdots \le X_{(m)}$$

are the order statistics for the sample $\{X_j\}$. Thus the optimal value of the likelihood function $\mathcal{L}_n(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{X})$ for the given sample $\boldsymbol{X}$ is $\mathcal{L}_n\left(\widehat{\boldsymbol{x}}_n, \widehat{\boldsymbol{z}}_n \middle| \boldsymbol{X}\right)$.

Next we consider the hypothesis that the sample $\boldsymbol{X}$ was actually taken from a Bézier distribution with one additional control point—that is, with a total of $n+2$ control points rather than the $n+1$ control points postulated in (12). For a hypothesized Bézier p.d.f. $f_X(\cdot; n+1, \boldsymbol{x}, \boldsymbol{z})$ involving one additional control point, let $\widehat{\boldsymbol{x}}_{n+1}$ and $\widehat{\boldsymbol{z}}_{n+1}$ respectively denote the maximum likelihood estimators of the x- and z-coordinates of the corresponding control points; and let $\mathcal{L}_{n+1}\left(\widehat{\boldsymbol{x}}_{n+1}, \widehat{\boldsymbol{z}}_{n+1} \middle| \boldsymbol{X}\right)$ denote the corresponding maximum value of the likelihood function. Using the matrix representation of Bézier curves of degree n and $n+1$, we can show that any nth degree Bézier curve can be represented as an $(n+1)$st degree Bézier curve whose control-point coordinates are subject to two linear constraints (see §4.6 of Farin 1990). Thus under the null hypothesis that n is the true degree of the underlying Bézier distribution from which the sample $\boldsymbol{X}$ was taken, Theorem 4.4.4 of Serfling (1980) ensures that the likelihood ratio test statistic asymptotically has a chi-square distribution with two degrees of freedom as the sample size m grows large:

$$2\left[\mathcal{L}_{n+1}\left(\widehat{\boldsymbol{x}}_{n+1}, \widehat{\boldsymbol{z}}_{n+1} \middle| \boldsymbol{X}\right) - \mathcal{L}_n\left(\widehat{\boldsymbol{x}}_n, \widehat{\boldsymbol{z}}_n \middle| \boldsymbol{X}\right)\right] \underset{m\to\infty}{\xrightarrow{\mathcal{D}}} \chi^2(2). \qquad (13)$$

We exploit (13) to assess the importance of successive increments of the likelihood function as the number of control points is repeatedly incremented by one. The final estimate of the number of control points for the Bézier distribution fitted to the sample $\boldsymbol{X}$ is determined to be the smallest value of n for which the difference on the left-hand side of (13) is not significant at a prespecified level of significance. The corresponding vectors $\widehat{\boldsymbol{x}}_n$ and $\widehat{\boldsymbol{z}}_n$ respectively provide the final estimates of the x- and z-coordinates for the fitted Bézier c.d.f.

4 ESTIMATING BIVARIATE BÉZIER DISTRIBUTIONS

Wagner and Wilson (1994, 1995) present a methodology for estimation of bivariate Bézier distributions. This technique requires fitting each marginal distribution separately; then the user must employ trial and error in manipulating the control points either for the joint p.d.f. or for selected conditional distributions in order to match the joint dependency structure between the two components of the target random vector $(X,Y)^{\mathrm{T}}$. This trial-and-error approach, however, is extremely difficult to perform in practice; and this difficulty motivated the development of the bivariate fitting procedure given in this section.

A detailed derivation of the covariance between two Bézier variates X and Y is given in Flanigan (1993) and Wagner and Wilson (1995). For completeness, we summarize the final result. The covariance between X and Y, Cov(X, Y), is

$$\mathrm{Cov}(X, Y) = \sum_{i=0}^{n_x-1} \sum_{j=0}^{n_y-1} \vartheta_i^{(X)} \vartheta_j^{(Y)} \Delta_i \Delta_j z_{i,j}\,, \qquad (14)$$

where

$$\vartheta_i^{(X)} = \left[\frac{1}{2} \sum_{l=0}^{n_x} \frac{\binom{n_x}{l}\binom{n_x-1}{i}}{\binom{2n_x-1}{i+l}} x_l\right] - \mathrm{E}[X] \qquad (15)$$

for $i = 0, 1, \ldots, n_x$ and

$$\vartheta_j^{(Y)} = \left[\frac{1}{2} \sum_{k=0}^{n_y} \frac{\binom{n_y}{k}\binom{n_y-1}{j}}{\binom{2n_y-1}{j+k}} y_k\right] - \mathrm{E}[Y] \qquad (16)$$

for $j = 0, 1, \ldots, n_y$. Notice that the covariance defined by (14)–(16) depends on the coordinates of the control points $\{\boldsymbol{q}_{i,j} : i = 0, 1, \ldots, n_x; = 0, 1, \ldots, n_y\}$.

If we are given a target covariance $\widehat{C}$ that has been elicited from experts or has been estimated from sample data and if we have previously estimated the marginals so that all of the following quantities have fixed values

$$\left\{x_i,\, z_i^{(X)},\, \vartheta_i^{(X)} : i = 0, 1, \ldots, n_x\right\}$$

and

$$\left\{y_j,\, z_j^{(Y)},\, \vartheta_j^{(Y)} : j = 0, 1, \ldots, n_y\right\},$$

then we can match the value $\widehat{C}$ as closely as possible by minimizing the absolute deviation

$$\left|\left\{\sum_{i=0}^{n_x-1} \sum_{j=0}^{n_y-1} \vartheta_i^{(X)} \vartheta_j^{(Y)} \Delta_i \Delta_j z_{i,j}\right\} - \widehat{C}\right| \qquad (17)$$

as a linear function of the $\{z_{i,j} : i = 1, \ldots, n_x - 1; j = 1, \ldots, n_y - 1\}$ subject to linear constraints that ensure a valid joint distribution. In particular the following linear programming problem must be solved:

$$\underset{\substack{W_1, W_2;\ \text{all } z_{i,j}:\\ 1 \le i \le n_x - 1\\ 1 \le j \le n_y - 1}}{\text{minimize}} \quad W_1 + W_2 \qquad (18)$$

subject to

$$\left.\begin{array}{l} W_1 - W_2 \\ \quad -\sum_{i=1}^{n_x-1} \sum_{j=1}^{n_y-1} \left[\Delta_i \Delta_j \vartheta_{i-1}^{(X)} \vartheta_{j-1}^{(Y)}\right] z_{i,j} \\ \quad = \widehat{C} - \vartheta_0^{(X)} \vartheta_{n_y-1}^{(Y)} z_1^{(X)} - \vartheta_{n_x-1}^{(X)} \vartheta_0^{(Y)} z_1^{(Y)} \\ \quad -\vartheta_{n_x-1}^{(X)} \vartheta_{n_y-1}^{(Y)} \left[1 - z_{n_x-1}^{(X)} - z_{n_y-1}^{(Y)}\right] \\ \quad - \displaystyle\sum_{j=1}^{n_y-2} \vartheta_{n_x-1}^{(X)} \vartheta_j^{(Y)} \Delta_j z_j^{(Y)} \\ \quad - \displaystyle\sum_{i=1}^{n_x-2} \vartheta_i^{(X)} \vartheta_{n_y-1}^{(Y)} \Delta_i z_i^{(X)} \end{array}\right\} \quad (19)$$

as well as

$$\left.\begin{array}{l} z_{1,j+1} - z_{1,j} \geq 0 \\ z_{1,n_y-1} \leq z_1^{(X)} \\ z_{i+1,1} - z_{i,1} \geq 0 \\ z_{n_x-1,1} \leq z_1^{(Y)} \\ z_{i+1,j+1} - z_{i,j+1} - z_{i+1,j} + z_{i,j} \geq 0 \\ z_{n_x-1,j+1} - z_{n_x-1,j} \leq z_{j+1}^{(Y)} - z_j^{(Y)} \\ z_{i+1,n_y-1} - z_{i,n_y-1} \leq z_{i+1}^{(X)} - z_i^{(X)} \\ z_{n_x-1,n_y-1} \geq z_{n_x-1}^{(X)} + z_{n_y-1}^{(Y)} - 1 \end{array}\right\} \quad (20)$$

for $i = 1, \ldots, n_x - 2$ and $j = 1, \ldots, n_y - 2$; and finally

$$W_1, W_2, z_{i,j} \geq 0 \; . \quad (21)$$

for $i = 1, \ldots, n_x - 1$ and $j = 1, \ldots, n_y - 1$. In this formulation W_1 and W_2 respectively represent the positive and negative parts of the estimation error as defined in (19); the constraints in (20) ensure that the joint density $f_{XY}(\cdot, \cdot)$ is positive; and finally (21) represents the standard nonnegativity constraints. Notice that in (19), the first and second differences $\Delta_i z_i^{(X)}$, $\Delta_j z_j^{(Y)}$, and $\Delta_i \Delta_j \vartheta_{i-1}^{(X)} \vartheta_{j-1}^{(Y)}$ are defined in the same way as the analogous quantities $\Delta_i x_i$, $\Delta_j y_j$, and $\Delta_i \Delta_j z_{i,j}$ in (6) and (11).

The linear programming problem (18)-(21) must be solved to complete the estimation of a bivariate Bézier distribution. This problem consists of $(n_x - 2)(n_y - 2) + 2(n_x - 2) + 2(n_y - 2) + 4 = n_x n_y$ structural constraints involving $(n_x - 1)(n_y - 1) + 2 = n_x n_y - n_x - n_y + 3$ decision variables. Thus it follows that the target covariance $\widehat{C}$ may not be exactly achieved by the final fitted bivariate Bézier distribution, but the target covariance $\widehat{C}$ will be matched as closely as is mathematically possible.

5 EXAMPLES

To illustrate the application of the likelihood ratio test (13) and the covariance-matching procedure based on (18)-(21), we consider a sample data set consisting of $m = 672$ bivariate observations that were generated in a simulation-based forestry study. Figures 1 and 2 depict the empirical and fitted marginal distributions for X and Y, respectively. Table 1 displays the sample statistics for each marginal distribution together with the corresponding population characteristics for the marginal Bézier distributions that were fitted using PRIME.

The histogram in Figure 1 clearly reveals that the distribution of the first coordinate X has two modes; and in our experience some manual intervention beyond routine application of the likelihood ratio test (13) is often required to obtain adequate fits to multimodal data sets. The fitted c.d.f. in Figure 1 was obtained by using PRIME to position 11 control points interactively so that $n_x = 10$. An alternative approach is to automatically obtain a bimodal Bézier p.d.f. via the likelihood ratio procedure (13). Using this approach, we had to start the test procedure with 11 control points; and with a significance level of 20% for each iteration, the likelihood ratio test procedure yielded a final estimate of 13 control points whose associated c.d.f. and p.d.f. closely resemble their counterparts in Figure 1. On a 66 Mhz 80486-based microcomputer running Windows 3.1, four iterations of the likelihood ratio test procedure (13) starting with 11 control points ($n_x = 10$) and stopping with 13 control points ($n_x = 12$) required 4.6 minutes of execution time for this univariate data set. Note that the execution time reported here is for a nonoptimized, debugger-enabled version of PRIME rather than a production version of the software; and substantially better execution times are expected for the final production version of the software.

Because the histogram in Figure 2 strongly suggests a unimodal distribution for the second coordinate Y, we applied the likelihood ratio test (13) starting with three control points (so that $n_y = 2$ initially) and using a 20% significance level for each iteration of the test procedure. Figure 2 also displays the final fitted c.d.f. and p.d.f. with 6 control points (so that the final estimate is $n_y = 5$). On a 66 Mhz 80486-based microcomputer running Windows 3.1, five iterations of the likelihood ratio test procedure (13) starting with 3 control points ($n_y = 2$) and stopping with 6

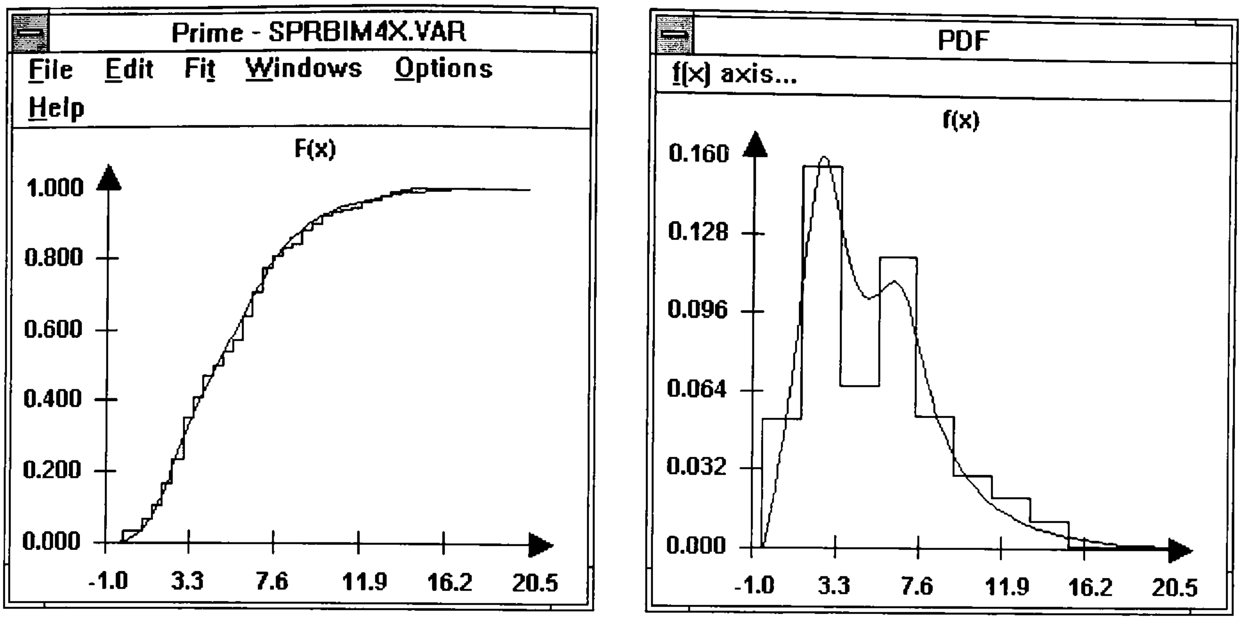

Figure 1: Fitted and Empirical Marginal Distributions of X for Sample Data Set

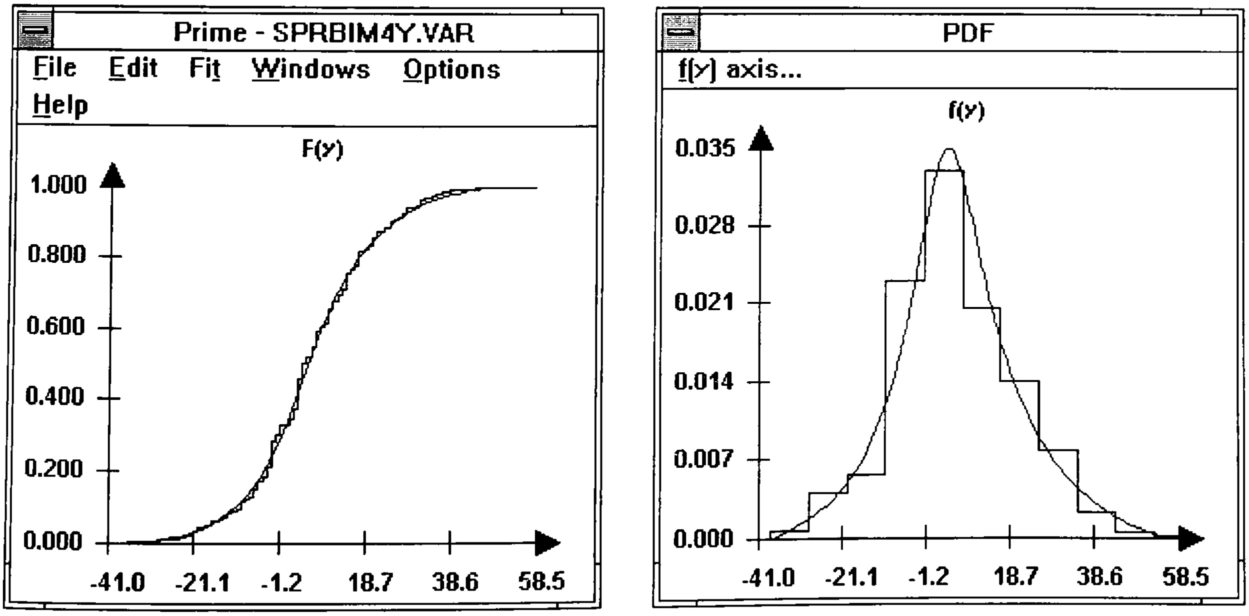

Figure 2: Fitted and Empirical Marginal Distributions of Y for Sample Data Set

Table 1: Fitted vs. Empirical Marginal Distributions of Sample Data Set

Characteristic	X Fitted	X Empirical	Y Fitted	Y Empirical
Mean	5.230	5.310	6.117	6.097
Standard Deviation	3.227	3.314	14.210	13.781
Skewness	0.997	0.845	0.252	0.155
Kurtosis	4.162	3.545	3.259	3.276
Minimum	−0.323	0.000	−37.678	−36.000
Maximum	19.824	19.500	55.000	54.000

control points ($n_y = 5$) required about 49 seconds of execution time for this univariate data set.

Starting from the fitted marginal distributions described above with $n_x = 10$ and $n_y = 5$, we applied the covariance-matching procedure (18)–(21) to the entire bivariate data set. The sample covariance and correlation for this data set were

$$\widehat{\text{Cov}}(X,Y) = \widehat{C} = 12.535 \quad \text{and} \quad \widehat{\text{Corr}}(X,Y) = 0.274,$$

respectively; and the corresponding values for the fitted bivariate Bézier distribution were

$$\text{Cov}(X,Y) = 12.535 \quad \text{and} \quad \text{Corr}(X,Y) = 0.273.$$

The resulting bivariate density is displayed in Figure 3. Clearly the fitted bivariate p.d.f. is also bimodal. On a 66 Mhz 80486-based microcomputer running Windows 3.1, solution of the linear programming problem (18)–(21) required about 3 seconds of execution time.

6 CONCLUSIONS AND RECOMMENDATIONS

The likelihood ratio test (13) provides a means for automatic determination of an appropriate number of control points in fitting a univariate Bézier distribution to sample data. However in our computational experience, the log-likelihood function $\mathcal{L}_n(\boldsymbol{x}, \boldsymbol{z}|\boldsymbol{X})$ corresponding to a univariate Bézier c.d.f. $F_X(\cdot; n, \boldsymbol{x}, \boldsymbol{z})$ is often extremely flat in the neighborhood of the exact maximum likelihood estimates. From the standpoint of visual closeness between the empirical c.d.f. $\widehat{F}_X(\cdot)$ and the fitted c.d.f. $\widehat{F}_X(\cdot; n, \widehat{\boldsymbol{x}}_n, \widehat{\boldsymbol{z}}_n)$, we have consistently obtained better fits with PRIME by using the methods of least squares or minimum L_1 norm estimation rather than the method of maximum likelihood to calculate the parameter estimates $\widehat{\boldsymbol{x}}_n$, $\widehat{\boldsymbol{z}}_n$ used in each stage of the test procedure (13). With this modification, the likelihood ratio test (13) yields excellent fits to many types of sample data with 4 control points ($n = 3$) so that effectively the fitted univariate Bézier distribution frequently has 6 parameters.

The covariance-matching procedure based on (18)–(21) often yields excellent fits to the joint dependency structure exhibited by bivariate sample data sets, but there is no guarantee that the target covariance or correlation between the two components of the random vector $(X, Y)^{\mathrm{T}}$ will be achieved even approximately. It appears that a relaxation of the system (20) of constraints may sometimes be required to match a given target covariance; but in such cases it is unclear how to ensure that the fitted bivariate Bézier c.d.f. has a legitimate p.d.f. Although further investigation of this issue is required, the general approach outlined in Section 4 of this paper promises to yield highly effective methods for fitting bivariate Bézier distributions to sample data or subjective information.

REFERENCES

Cario, M. C., and B. L. Nelson. 1996. Autoregressive to anything: Time series input processes for simulation. Working paper, Department of Industrial, Welding and Systems Engineering, The Ohio State University, Columbus, Ohio.

Farin, G. 1990. *Curves and surfaces for computer aided geometric design: A practical guide.* 2d ed. New York: Academic Press.

Flanigan, M. A. 1993. A flexible, interactive, graphical approach to modeling stochastic input processes. Ph.D. diss., School of Industrial Engineering, Purdue University, West Lafayette, Indiana.

Jagerman, D. L., and B. Melamed. 1992a. The transition and autocorrelation structure of TES processes, Part I: General theory. *Communication in Statistics-Stochastic Models* 8 (2): 193–219.

Jagerman, D. L., and B. Melamed. 1992b. The transition and autocorrelation structure of TES processes, Part II: Special cases. *Communication in*

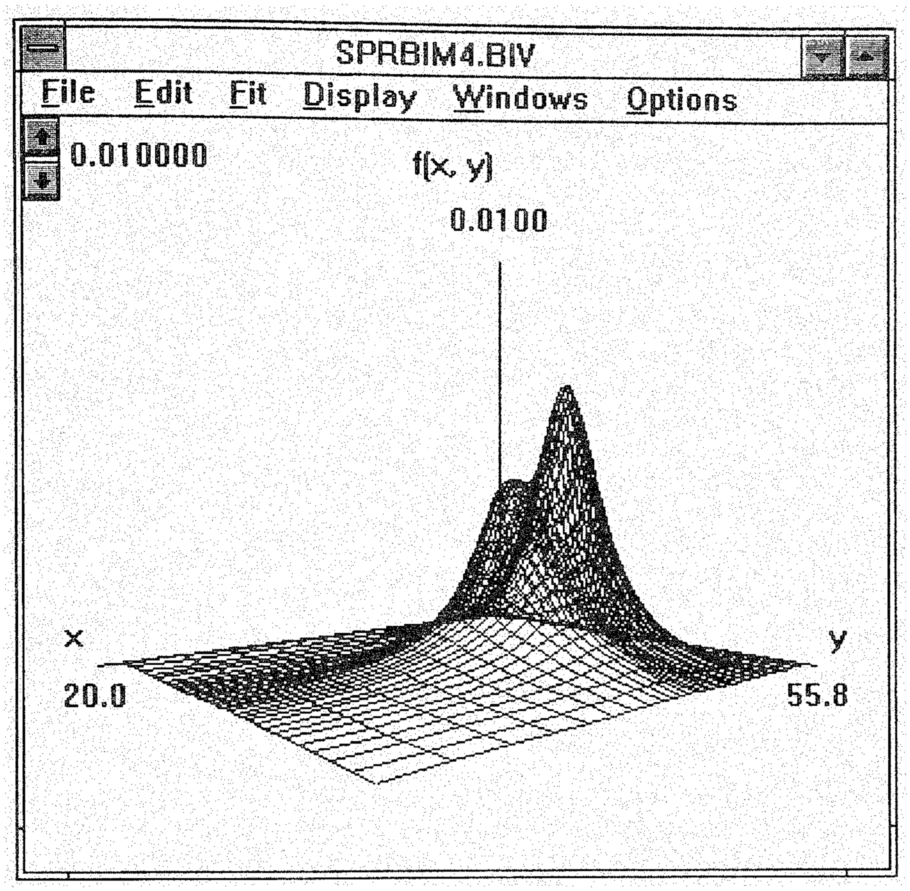

Figure 3: The Joint Density Fitted to the Sample Data Set

Statistics–Stochastic Models 8 (3): 499–527.

Johnson, M. E. 1987. *Multivariate statistical simulation.* New York: John Wiley & Sons.

Johnson, N. L. 1949a. Systems of frequency curves generated by methods of translation. *Biometrika* 36:149–176.

Johnson, N. L. 1949b. Bivariate distributions based on simple translation systems. *Biometrika* 36:297–304.

Melamed, B., J. R. Hill, and D. Goldsman. 1992. The TES methodology: Modeling empirical stationary time series. In *Proceedings of the 1992 Winter Simulation Conference*, ed. J. J. Swain, D. Goldsman, R. C. Crain, and J. R. Wilson, 135–144. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Serfling, R. J. 1980. *Approximation theorems of mathematical statistics.* New York: John Wiley & Sons.

Stanfield, P. M., J. R. Wilson, G. A. Mirka, N. F. Glasscock, J. P. Psihogios, and J. R. Davis. 1996. Multivariate input modeling with Johnson distributions. In *Proceedings of the 1996 Winter Simulation Conference*, ed. J. M. Charnes, D. J. Morrice, D. T. Brunner, and J. J. Swain, to appear. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Swain, J. J., S. Venkatraman, and J. R. Wilson. 1988. Least-squares estimation of distribution functions in Johnson's translation system. *Journal of Statistical Computation and Simulation* 29:271–297.

Wagner, M. A. F., and J. R. Wilson. 1993. Using univariate Bézier distributions to model simulation input processes. In *Proceedings of the 1993 Winter Simulation Conference*, ed. G. W. Evans, M. Mollaghasemi, E. C. Russell, and W. E. Biles, 365–373. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Wagner, M. A. F., and J. R. Wilson. 1994. Using bivariate Bézier distributions to model simulation input processes. In *Proceedings of the 1994 Winter Simulation Conference*, ed. J. D. Tew, S. Manivannan, D. A. Sadowski, and A. F. Seila, 324–331. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Wagner, M. A. F., and J. R. Wilson. 1995. Graphical interactive simulation input modeling with bivariate Bézier distributions. *ACM Transactions on Modeling and Computer Simulation* 5 (3): 163–189.

Wagner, M. A. F., and J. R. Wilson. 1996. Using univariate Bézier distributions to model simulation input processes. *IIE Transactions* to appear.

AUTHOR BIOGRAPHIES

MARY ANN FLANIGAN WAGNER is currently working at Boeing Information Services, located in Vienna, Virginia. Her responsibilities include the simulation and analysis of large-scale telecommunications models. Her principal interests are in simulation development, modeling and analysis. Her undergraduate and graduate degrees are in the field of Industrial Engineering, and in May 1993, she received her Ph.D. from Purdue University. From 1987 to 1989 she held a research position at the Regenstrief Institute, where she was responsible for the development and analysis of simulation models. Dr. Wagner is a member of: Alpha Pi Mu, ACM, IIE, INFORMS, Omega Rho, SIGGRAPH, and SIGSIM.

JAMES R. WILSON is Professor and Director of Graduate Programs in the Department of Industrial Engineering at North Carolina State University. He received a B.A. degree in mathematics from Rice University, and he received M.S. and Ph.D. degrees in industrial engineering from Purdue University. His current research interests are focused on the design and analysis of simulation experiments. He also has an active interest in applications of operations research techniques to all areas of industrial engineering. From 1988 to 1992, he served as Departmental Editor of *Management Science* for Simulation. He was *Proceedings* Editor for WSC '86, Associate Program Chair for WSC '91, and Program Chair for WSC '92. He has also held various offices in TIMS (now INFORMS) College on Simulation. He is a member of ASA, ACM/SIGSIM, IIE, and INFORMS.