Journal of Applied Statistics
Vol. 33, No. 2, 175–187, March 2006

Routledge
Taylor & Francis Group

Modelling Multi-stage Processes through Multivariate Distributions

ASHIS SENGUPTA* & FIDELIS I. UGWUOWO**

**Applied Statistics Unit, Indian Statistical Institute, Kolkata, W.B., 700108, India, **Department of Statistics, University of Nigeria, Nsukka, Enugu State, Nigeria*

ABSTRACT *A new model combining parametric and semi-parametric approaches and following the lines of a semi-Markov model is developed for multi-stage processes. A Bivariate sojourn time distribution derived from the bivariate exponential distribution of Marshall & Olkin (1967) is adopted. The results compare favourably with the usual semi-parametric approaches that have been in use. Our approach also has several advantages over the models in use including its amenability to statistical inference. For example, the tests for symmetry and also for independence of the marginals of the sojourn time distributions, which were not available earlier, can now be conveniently derived and are enhanced in elegant forms. A unified Goodness-of-Fit test procedure for our proposed model is also presented. An application to the human resource planning involving real-life data from University of Nigeria is given.*

KEY WORDS: Bivariate exponential, multi-stage processes, semi-Markov, semi-parametric, human resource planning

Introduction and Motivations

Population models of multi-grade systems have been discussed by a number of authors and have also been applied in a number of ways. The grades normally correspond to recognized divisions within the system like grades of staff in a manpower system, level of commitment to a job, etc, as shown in McClean (1980), Gani (1963) and the references therein. References on their applications to biological systems, pharmacokinetic processes, epidemiology, etc may be found in McClean (1978). However, it seems that in all these areas, no work has so far been done using a multivariate modelling approach.

This paper is aimed at unifying the existing models by employing a joint distribution function in estimating the sojourn time of individuals in a multi-stage process. With this model, it will now be possible to evaluate the conditional probability of sojourn time in any state given the sojourn time in the previous state. This model also enhances the use of statistical tests such as tests for independence and for symmetry of the sojourn times. The model assumes naturally that the sojourn times in different states are dependent on their immediate past states. Estimation of sojourn times for event history can throw

Correspondence Address: A. SenGupta, Applied Statistics Unit, Indian Statistical Institute, 203 B.T. Road, Kolkata, W.B. 700108, India. Email: ashis@isical.ac.in

0266-4763 Print/1360-0532 Online/06/020175–13
DOI: 10.1080/02664760500250586

light on how the process works and help in planning for the future, e.g. as of the sojourn times in different stages of a disease like cancer or AIDS.

Research has shown that for a chronic disease like breast cancer, women under the age of 50 years have a shorter sojourn time than women aged between 50–74 years (Tabar *et al*., 1995). Some methods for the estimation of sojourn time using a specific sojourn time distribution can be found in Day & Walter (1984), and Paci & Duffy (1991).

Chen & Porok (1983) used a non-parametric method and split time into discrete intervals. The use of Markov models for the natural history of a disease process from disease free state to the preclinical-screen detectable phase (PCDC) and then to the clinical phase can be seen in Duffy *et al*. (1995) and Chen *et al*. (2000).

Many authors have proposed the use of multivariate exponential distributions to model lifetimes of components of a multivariate system, see, for example, Marshall & Olkin (1967) (MO henceforth), Block & Savits (1981), Basu (1988). Several tests of the parameters of these multivariate models have also been developed in SenGupta (1995) and these have enhanced the usefulness of these models in statistical inference. Here, based on both practical and theoretical justifications, we enhance the multivariate exponential model of MO to model multi-stage processes in general and data from human resource domain in particular. Some of the justifications for this choice are as follows. First, exponential distribution is a commonly adopted model for the promotional time in each category. We further believe that promotions are usually based on merit and efficiency rather than on the duration of service. This implies that the lack of memory for promotional time is only reasonable to enforce and this is a characterizing property of our chosen marginal distributions, i.e. the exponential distributions. Second, promotions for each category are usually (save out-of-turn merit promotions) given after the mandatory eligibility period at certain intervals of time at a pre-fixed date, say January 1. This gives positive probability of exactly equal lengths of time for successive or several promotions. The chosen model encompasses such situations since it gives $P\ (X = Y) > 0$ for the two marginal random variables X and Y. This is not true for the other familiar generalizations of the univariate exponential distribution. Third, as pointed out by a referee, the existences of candidates who are high-fliers is only to be expected. In academic cases, the proportion of such cases may be quite high implying higher probability of promotion at the earlier years rather than at some distant year – again a property possessed by the exponential distribution. Finally, we note that usually the professional characteristics of an individual tends to persist and, as driving forces, should yield similar results for the transition of the individual from one category to another and to the next, etc – i.e. incumbents with early (late) promotions at the initial categories are expected to receive early (late) promotions at the subsequent categories too. This fact establishes that the correlation of promotional times for different categories should be taken to be positive. Here again, our chosen model guarantees this requirement of the relevant correlation.

Several semi-Markov models for human resource planning are available. These models were developed using different approaches and applied to different aspects of human resource planning. The continuous time semi-Markov modelling approach may be found in Mehlmann (1979), Bartholomew (1982) and McClean (1993). They defined the force of transition or hazard rate from one grade to another given the duration in the first grade and then used it to derive the probability that an entry into a grade will move to the next grade given the holding time in the earlier grade. They also used the method of maximum likelihood estimate to obtain the probability of eventual transition from one grade to the other. Mehlmann (1979), McClean (1980) and Bartholomew *et al*. (1991) discussed the other version of a semi-Markov model as a renewal type equation.

Their approach defines the probability of an individual being in a state at time t given that the individual was in the earlier state at time zero. They used it to derive a renewal type equation for predicting future manpower structure. There have also been generalizations to non-homogeneous semi-Markov models in Vassiliou & Papadopoulou. (1992) and McClean *et al*. (1998). Some of these models assume time homogeneity while the non-homogeneous ones divide the calendar time into a succession of time windows. The approach in this paper does not require those assumptions since it is based on observed sojourn times. However, we recall the result from MO that an underlying multivariate Poisson process yields their multivariate exponential distribution. Hence the conditions driving such a process are being implicitly assumed here.

The preliminary notions for the parametric and semi-parametric models are derived in the next section. The section after discusses the models and estimation of parameters including some statistical tests, such as the test of goodness-of-fit for a sparsely distributed contingency table, test for independence and the symmetry test for the marginal of the Bivariate exponential distribution. In the fourth section, the model and methodologies developed here are applied to the real-life data on promotion times for faculty members in University of Nigeria. The fifth section contains the suggestions for further generalizations and concluding remarks.

Preliminary Notions

Consider a system with grades $S_1, \ldots, S_m$, where the length of stay in S_i conditional on eventually making the transition to S_j, has a probability density function (p.d.f.) $f_{ij}(t)$, with distribution function $F_{ij}(t)$ and survivor function

$$G_{ij}(t) = 1 - F_{ij}(t) = \int_t^\infty f_{ij}(x)\mathrm{d}x; \qquad i \neq j; \quad i,j = 1, \ldots, m. \tag{1}$$

(In many applications, e.g. in promotional data where demotion is ruled out, we will have $i < j$.) The corresponding p.d.f. of time spent in S_i is $f_i(t)$ with distribution function $F_i(t)$ and survivor function

$$G_i(t) = 1 - F_i(t) = \int_t^\infty f_i(x)\mathrm{d}x \tag{2}$$

Consider now the case of grouped data, as in a contingency table, with 2-way classification first. Let the random variable X_l denote the sojourn time in S_l, $l = i,j$. Let there be R and K 'time-intervals', defining the classes in the contingency table, for X_i and X_j respectively. The data may then be visualized as a $R \times K$ contingency table of the two factors S_i and S_j at levels R and K respectively. This table yields the RK 'cells' for the joint distribution of X_i and X_j, the ijth cell corresponding to the joint event that X_i and X_j are in the time-intervals $[t_r^{(i)}, t_{r+1}^{(i)})$ and $[t_k^{(j)}, t_{k+1}^{(j)})$ respectively, $r = 1, \ldots, R, k = 1, \ldots, K$, and R and K, $i,j = 1, \ldots, m$. Then,

$$p[x_j \geq t_k^{(j)} | x_i \geq t_r^{(i)}] = p[x_i \geq t_r^{(i)}, x_j \geq t_k^{(j)}]/p[x_i \geq t_r^{(i)}]$$

Similarly,

$$p[x_j \leq t_k^{(j)} | x_i \leq t_r^{(i)}] = p[x_i \geq t_r^{(i)}, x_j \leq t_k^{(j)}]/p[x_i \leq t_r^{(i)}]$$

But

$$p[x_i \le t_r^{(i)}, x_j \le t_k^{(j)}] = 1 - p[x_i \ge t_r^{(i)}] - p[x_j \ge t_k^{(j)}] + p[x_i \ge t_r^{(i)}, x_j \ge t_k^{(j)}]$$

$$\therefore \quad p[x_j \ge t_k^{(j)} | x_i \le t_r^{(i)}] = \{1 - p[x_i \ge t_r^{(i)}] - p[x_j \ge t_k^{(j)}] + p[x_i \ge t_r^{(i)}, x_j \ge t_k^{(j)}]\} / p[x_i \le t_r^{(i)}] \tag{3}$$

Let the joint p.d.f. of the length of time spent in S_i and S_j be given by $h_{i,j}(t_r^{(i)}, t_k^{(j)})$ with distribution function $H_{i,j}(t_r^{(i)}, t_k^j)$ and survivor function

$$G_{i,j}(t_r^{(i)}, t_k^{(j)}) = 1 - H_{i,j}(t_r^{(i)}, t_k^{(j)}) = \int_{t_r^{(i)}}^{\infty} \int_{t_k^{(j)}}^{\infty} h_{i,j}(x_i, x_j) \mathrm{d}x_i \mathrm{d}x_j \tag{4}$$

Define $p_{i,j}(t_r^{(i)}, t_k^{(j)})$ to be the probability that an individual will spend less than or equal to $t_r^{(i)}$ time in S_i and less than or equal to $t_k^{(j)}$ time in S_j. Then

$$p_{i,j}(t_r^{(i)}, t_k^{(j)}) = p[x_i \le t_r^{(i)}, x_j \le t_k^{(j)}] \tag{5}$$

It can easily be shown that equation (5) is equivalent to

$$= 1 - p[x_i \ge t_r^{(i)}] - p[x_j \ge t_k^{(j)}] + p[x_i \ge t_r^{(i)}, x_j \ge t_k^{(j)}] \tag{6}$$

Then, the cell probabilities can be expressed as follows

$$\begin{aligned} & p(t_r^{(i)} \le X_i \le t_{r+1}^{(i)}, t_k^{(j)} \le X_j \le t_{k+1}^{(j)}) \\ & = p[x_i \le t_{r+1}^{(i)}, x_j \le t_{k+1}^{(j)}] - p[x_i \le t_r^{(i)}, x_j \le t_{k+1}^{(j)}] - p[x_i \le t_{r+1}^{(i)}, x_j \le t_k^{(j)}] \\ & + p[x_i \le t_r^{(i)}, x_j \le t_k^{(j)}] \qquad (7) \\ & \equiv F(t_{r+1}^{(i)}, t_{k+1}^{(j)}) - F(t_r^{(i)}, t_{k+1}^{(j)}) - F(t_{r+1}^{(i)}, t_k^{(j)}) + F(t_r^{(i)}, t_k^{(j)}) \qquad (8) \end{aligned}$$

where $F(a,b) \equiv p(X_i \le a, X_2 \le b)$ is the cumulative distribution and can easily be obtained from equation (6).

The Models

We consider below both the parametric and the semi-parametric basic models.

(1) *The Parametric Model*
Given the states of a multi-grade process, the probability that an individual will spend less than or equal to $t_r^{(i)}$ time in S_i and less than or equal to $t_k^{(j)}$ time in S_j gives such a bivariate distribution as

$$p_{i,j}(t_r^{(i)}, t_k^{(j)}) = \int_0^{t_r^{(i)}} \int_0^{t_k^{(j)}} h_{i,j}(x_i, x_j) \mathrm{d}x_i \mathrm{d}x_j \tag{9}$$

This can easily be evaluated using equations (5) and (6).

(2) *The Semi-parametric Model*
There is a probability that an individual spends $t_r^{(i)}$ time in S_i given such an individual eventually moves to S_j with a certain force of transition and then spends time $t_k^{(j)}$ in S_j.

We assume that the time spent in S_j is independent of the probability dictating the force of transition from S_i to S_j but it possibly depends on the duration of the individual's stay in S_i. Then the probability of such an occurrence is given by,

$$p[x_i \leq t_r^{(i)}]\, p[s_j/x_i]\, p[x_j \leq t_k^{(j)}/x_i, s_j] \equiv p[x_i \leq t_r^{(i)}, x_j \leq t_k^{(j)}]\, p(s_j/x_i)$$
$$= z_{i,j}(t_r^{(i)}) \int_0^{t_r^{(i)}} \int_0^{t_k^{(j)}} h_{i,j}(x_i, x_j)\mathrm{d}x_i \mathrm{d}x_j \qquad (10)$$

where $p[s_j/x_i] = z_{i,j}(t_r^{(i)})$ is the force of transition from S_i to S_j at duration $t_r^{(i)}$ and its Kaplan–Meier estimate (McClean, 1980) is given by

$$z_{i,j}(t_r^{(i)}) = \frac{n_{i,j}(t_r^{(i)})}{n..} \quad \text{and} \quad n.. = \sum_{r(i)} \sum_{r(j)} n_{i,j}(t_r^{(i)})$$

where $n_{i,j}(t_r^{(i)})$ is the observed frequency in ijth cell at duration $t_r^{(i)}$.

We note that the force of transition depends on the conglomeration of candidates as a totality and is normally evaluated from a priori information on this group whereas the time spent by individuals in S_j may very well depend on the amount of time spent in S_i by those very individuals. Furthermore, if we have Markov transitions between grades then the duration of stay in each grade is exponential and does not depend on the destination. If instead, we consider semi-Markov transitions, then we may include our knowledge of the distribution of length of service before leaving, and also allow for the fact that the length of time spent in a state may depend on the destination as well as the present grade of the individual (McClean, 1976).

The Bivariate Exponential Distribution

Several bivariate exponential models may be considered given the basic assumption of dependent sojourn times, and the usual practice of modelling univariate sojourn time by an exponential distribution: it was just appropriate to adopt the model given by MO. In justifying the use of this model, we considered that $p(X_1 = X_2) > 0$. If X_1 and X_2 are sojourn times before promotion then it is much more likely since management normally meet at a particular time of the year to take decisions on such matters. The distribution function of this bivariate exponential model (henceforth, BVE) is then given as follows:

$$P[X_1 > x_1, X_2 > x_2] = \exp\left(-\lambda_1 x_1 - \lambda_2 x_2 - \lambda_{12} \max\left(x_1, x_2\right)\right) \qquad (11)$$

The bivariate model proposed in MO has the marginal exponential distributions given by

$$\exp\left(\lambda_\ell + \lambda_{12}\right), \ell = 1,2$$

Let,

$$\lambda = \lambda_1 + \lambda_2 + \lambda_{12} \qquad (12)$$

Then,

$$\mathrm{E}(X_l) = \lambda_\ell + \lambda_{12}, \ell = 1,2 \quad \text{and}$$
$$\mathrm{Cov}(\mathrm{X}_1, \mathrm{X}_2) = \lambda_{12}[\lambda(\lambda_1 + \lambda_{12})(\lambda_2 + \lambda_{12})]^{-1} > 0$$

Adopting the method of moments to estimate the parameters of equation (11), we have the following three equations,

$$\bar{x}_1 = \lambda_\ell + \lambda_{12}, \ell = 1,2$$
$$\overline{x_1 x_2} \equiv E(x_1 x_2) = \lambda^{-1}\{(\lambda_1 + \lambda_{12})^{-1} + (\lambda_2 + \lambda_{12})^{-1}\} \tag{13}$$

where we are using the notation $\hat{EY} = \overline{Y} \equiv EY$ to denote the sample average of $y_i, i = 1, \ldots, n$, which estimates the corresponding population moment as is done in the method of moments.

Then it follows that

$$\hat{\lambda} = (\bar{x}_1 + \bar{x}_2)/\overline{x_1 x_2};\ \hat{\lambda}_{12} = \bar{x}_1^{-1} + \bar{x}_2^{-1} - \hat{\lambda};\ \hat{\lambda}_1 = \hat{\lambda} + \bar{x}_2^{-1};\ \hat{\lambda}_2 = \hat{\lambda} + \bar{x}_1^{-1} \tag{14}$$

Some Statistical Tests: Goodness-of-Fit Test

To determine the adequacy of the BVE distribution, the power divergence Goodness-of-Fit test of Cressie & Read (1984) can be used.

The test statistic is given by:

$$\tau_\eta^2 = \frac{2}{\eta(\eta+1)} \sum_i^m \sum_j^m n_{ij}\left[\left(\frac{n_{ij}}{E_{ij}}\right)^\eta - 1\right],\ -\infty < \eta < \infty \tag{15}$$

where n_{ij} is the observed frequency, E_{ij} is the expected frequency in the ijth cell. The expected frequency $E_{ij} = n\,.\,.\,\Pi_{ij}$ and $\prod_{ij}$ is the estimated probability of an individual who spends $t_r^{(i)}$ time in S_i and $t_k^{(j)}$ time in S_j, $r = 1, \ldots, R$ and $k = 1, \ldots, K$.

i.e. $\Pi_{ij} = p(t_r^{(i)} \le X_i \le t_{r+1}^{(i)}, t_k^{(j)} \le X_j \le t_{k+1}^{(j)})$

where $\hat{\Pi}_{ij}$ is obtained from equation (8) by using the estimated $\hat{\lambda}_1, \hat{\lambda}_2$ and $\hat{\lambda}_{12}$ and then replacing E_{ij} by $\hat{E}_{ij} = n\,.\,.\,\hat{\Pi}_{ij}$. For this test, τ_η^2 is equivalent to χ^2 when $\eta = 1$. They recommended the statistic with $\eta = 2/3$, which they found less susceptible than χ^2 to the effects of sparsely distributed data.

Test for Independence

The test for independence of X_1 and X_2 is conducted using the test statistic:

$$\chi^2 = \sum_{ij} \frac{(n_{ij} - \hat{E}_{ij})^2}{\hat{E}_{ij}} \sim \chi^2_{n-\upsilon} \tag{16}$$

where υ is the number of estimated parameters. We do this under both the parametric and the semi-parametric scenarios.

(a) *The Parametric BVE Approach*
The probabilities $\hat{\prod}_{i.}$ and $\hat{\prod}_{.j}$ of the exponential marginal distributions for the various time intervals under independence were first computed. These estimated probabilities from the marginal distributions are given respectively as $\hat{\Pi}_{i.} = \int_{t_r^{(i)}} \hat{\lambda}_1 e^{-\hat{\lambda}_1 x_1}\, dx_1$ and $\hat{\Pi}_{.j} = \int_{t_k^{(j)}} \hat{\lambda}_2 e^{-\hat{\lambda}_2 x_2}\, dx_2$ for the respective time intervals.
Then the expected frequencies were obtained as $\hat{E}_{ij} = n\,..\,\hat{\Pi}_{i.}\hat{\Pi}_{.j}$.

(b) *The Non-parametric Contingency Table Approach*.
The usual non-parametric test is implemented with the expected frequencies given by $E_{ij} = n_{i.}n_{.j}/n\,..$, where $n_{i.}$ and $n_{.j}$ are the marginal totals in the ith row and jth column.

Test for Symmetry of the Marginal (under Dependency)

The test for symmetry could be done by first using the method of moments on equation (11) to obtain the estimates for the common (under symmetry) parameter $\lambda_1 = \lambda_2 = \lambda^*$ and λ_{12}, where,

$$\hat{\lambda} = \overline{x_1 x_2}^{-1}/(\bar{x}_1^{-1} + \bar{x}_2^{-1});\ \hat{\lambda}_{12} = (\bar{x}_1^{-1} + \bar{x}_2^{-1}) - \hat{\lambda};$$
$$\lambda^* = \hat{\lambda} - \{(\bar{x}_1^{-1} + \bar{x}_2^{-1})^2/2(\bar{x}_1^{-1} + \bar{x}_2^{-1})\} \tag{17}$$

Furthermore, Π_{ij} for the given data was obtained by using the method of evaluating ijth cell probabilities as in equations (7) and (8) and using the estimates $\hat{\lambda}_1 = \hat{\lambda}_2 = \hat{\lambda}^*$ and $\hat{\lambda}_{12}$. Finally, the expected frequency is $\hat{E}_{ij} = n\,..\,\hat{\Pi}_{ij}$.

The Trivariate Exponential Distribution

The BVE model extends easily to the multivariate situation. In particular, we illustrate this for a three-stage sojourn time with the survival function for the trivariate distribution given as:

$$\begin{aligned} P[X_1 > x_1, X_2 > x_2, X_3 > x_3] = \exp\,(&-\lambda_1 x_1 - \lambda_2 x_2 - \lambda_{12}\max\,(x_1, x_2) \\ &- \lambda_{13}\max\,(x_1, x_3) - \lambda_{23}\max\,(x_2, x_3) \\ &- \lambda_{123}\max\,(x_1, x_2, x_3) \\ \lambda_i, \lambda_{ij}, \lambda_{123} \geq 0;\ &i \neq j;\ i, j = 1, 2, 3. \end{aligned} \tag{18}$$

We note that all the lower dimensional marginals will follow the exponential distribution and, in particular, the two-dimensional marginals of the above distribution are BVEs and the one-dimensional marginals are exponentials.

Estimation of Parameters

Let, $\lambda = \lambda_1 + \lambda_2 + \lambda_3 + \lambda_{12} + \lambda_{13} + \lambda_{23} + \lambda_{123}$.

We shall adopt the method of moments in estimating the parameters of equation (18). To obtain a more compact notation for this distribution, let S denote the set of vectors (s_1, s_2, s_3) where each $s_i = 0$ or 1 but $(s_1, s_2, s_3) \neq (000)$. To do this, it is convenient to replace the parameters λ_s by the new parameters g_s, s ∈ S, defined by $g_s = \sum_{rs \neq 0} \lambda_r$, i.e. g_s is the sum of all λ_s such that some coordinates are 1 in both r and s. For example, with $n = 3$, g_{100} is the sum over all λ_s where s_1 equals 1, then,

$$g_{100} = \lambda_{111} + \lambda_{110} + \lambda_{101} + \lambda_{100} \equiv \lambda_{123} + \lambda_{12} + \lambda_{13} + \lambda_1$$

So, $\lambda = g_{111}$.

It then follows that,

$$(g_{100}, g_{010}, g_{001}, g_{110}, g_{101}, g_{011}, g_{111}) = \begin{bmatrix} 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ 0 & 0 & 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{bmatrix} \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \lambda_3 \\ \lambda_{12} \\ \lambda_{13} \\ \lambda_{23} \\ \lambda_{123} \end{bmatrix}$$

i.e. $g = M\lambda'$, say.

Since M is a non-singular matrix, we have,

$$\underset{\sim}{\lambda}' = M^{-1} \underset{\sim}{g} \tag{19}$$

We recall that the moment generating function of a trivariate exponential distribution function is given by

$$\begin{aligned} \Phi(s_1, s_2, s_3) = \frac{s_1 s_2 s_3}{g_{111} + s_1 + s_2 + s_3} \big[& (g_{110} + s_1 + s_2)^{-1} \{(g_{100} + s_1)^{-1} \\ & + (g_{010} + s_2)^{-1}\} + (g_{101} + s_1 + s_3)^{-1} \{(g_{100} + s_1)^{-1} + (g_{001} + s_3)^{-1}\} \\ & + (g_{011} + s_2 + s_3)^{-1} \{(g_{010} + s_2)^{-1} + (g_{001} + s_3)^{-1}\} \big] \end{aligned} \tag{20}$$

We obtain seven equations to estimate the seven parameters by the method of moments.

The first three sets of equation (21) are obtained from the univariate exponential marginals, while the next three are obtained from the BVE marginals and given in equations (22) and (23). The seventh equation, equation (24), is obtained from the full trivariate exponential distribution.

$$\bar{x}_1 \equiv EX_1 = \frac{1}{g_{100}}; \bar{x}_2 \equiv EX_2 = \frac{1}{g_{010}}; \ \bar{x}_3 \equiv EX_3 = \frac{1}{g_{001}} \tag{21}$$

$$\overline{x_1 x_2} \equiv EX_1 X_2 = \left. \frac{\partial^2 \Phi}{\partial s_1 \partial s_2} \right|_{s_1 = s_2 = 0} = \frac{1}{g_{111}} \left(\frac{1}{g_{100}} + \frac{1}{g_{010}} \right) \tag{22}$$

Similarly,

$$\overline{x_1 x_3} \equiv EX_1 X_3 = \frac{1}{g_{111}} \left(\frac{1}{g_{100}} + \frac{1}{g_{001}} \right); \overline{x_2 x_3} \equiv EX_2 X_3 = \frac{1}{g_{111}} \left(\frac{1}{g_{010}} + \frac{1}{g_{001}} \right) \tag{23}$$

The calculation of the third moment is somewhat tedious. After some computations and simplifications we arrive at

$$\begin{aligned} \overline{x_1 x_2 x_3} \equiv EX_1 X_2 X_3 = \left. \frac{\partial^3 \Phi}{\partial s_1 \partial s_2 \partial s_3} \right|_{s_1 = s_2 = s_3 = 0} = \frac{1}{g_{111}} \Bigg\{ & \frac{1}{g_{110}} \left(\frac{1}{g_{100}} + \frac{1}{g_{010}} \right) \\ & + \frac{1}{g_{101}} \left(\frac{1}{g_{100}} + \frac{1}{g_{001}} \right) + \frac{1}{g_{011}} \left(\frac{1}{g_{010}} + \frac{1}{g_{001}} \right) \Bigg\} \end{aligned} \tag{24}$$

Table 1. Observed frequencies for the groups

	X_2				
X_1	$0 \to 1$	$1 \to 2$	$2 \to 3$	$3 \to 4$	≥ 4
$0 \to 1$	64	38	20	11	8
$1 \to 2$	54	28	16	3	14
$2 \to 3$	20	10	10	5	2
$3 \to 4$	13	3	3	2	4
≥ 4	6	4	8	6	2

where we are using the notation $EY \equiv \bar{y} = \hat{EY}$ to denote the sample average of $y_i, i = 1, \ldots, n$, as is done in the method of moments.

Then it follows that solving these seven equations simultaneously will yield the estimate of g say $\hat{g}$. Then, from equation (19), we easily get the estimates of the parameters, $\hat{\lambda}$. Finally the method of moments guarantees the optimality properties of consistency and asymptotic joint normality of these estimators.

Example

The models and methods of analysis developed above are now illustrated using the data collected from the personnel department of the University of Nigeria, Nsukka during 1970–1995. Complete data on 354 staff who have passed through the promotion process from Lecturer to Senior Lecturer and then to Associate Professor were taken. We have considered the actual waiting time in years, beyond the mandatory eligibility period, until notification for promotion is given. We define the random variable X_i to be the sojourn time in grade S_i,$i = 1,2$, where 1 and 2 refer to Lecturer and Senior Lecturer respectively.

The set of data was grouped in Table 1 to enable us to get the consolidated picture of the joint distribution. In addition, Bartholomew *et al*. (1991) have recommended grouping of even relatively small sets of such data.

Results and Discussions

In our quest for fitting a model to our data, we considered several bivariate exponential distributions. Based on the assumptions for their use, we decided on the BVE distribution that readily gave a good fit to the set of data. The striking feature in the data is the equality of both variables at several points, which unequivocally advocates the choice of the above model. The basic assumptions for the use of that model were also found to be satisfactory. The values of the estimated parameters of the distribution were obtained as $\hat{\lambda}_1 = 0.54; \hat{\lambda}_2 = 0.62, \hat{\lambda}_{12} = 0.05$. Figures 1(a) and (b) display the plots (using the values of the estimated parameters) of the cumulative and survivor distributions of the distribution respectively. A summary table of the power divergence test to determine how good the model fits the data is given in Table 2. The values show that the test is not significant at 5% levels and 21 d.f., the corresponding cut-off value being 32.67. We thus adopt this model for our data.

Adopting this model, we conducted a test for independence of X_1 and X_2. The parametric test for the null hypothesis H_0: $\lambda_{12} = 0$ gave the χ^2 value of 45.33 with d.f. $= 22$. See Table 3 for the estimated frequencies and marginal probabilities used in this test. A similar test was done using the non-parametric contingency table approach and gave the χ^2 value of 29.78

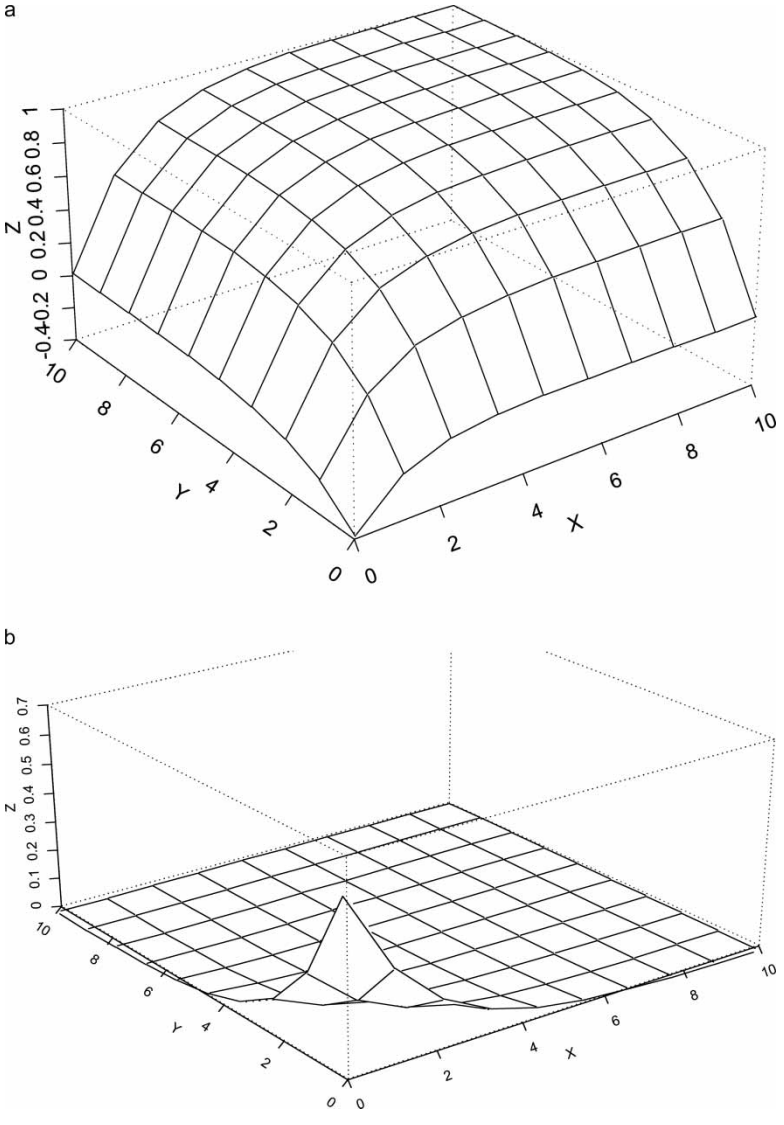

Figure 1. (a) Fitted BVE distribution function of Marshall & Olkin; (b) Survivor distribution function of the fitted model

Table 2. Summary result of test of Goodness-of-Fit using power divergence test

λ	Chi-square values
1	28.62
2/3	28.69
0.1	28.52
−0.5	30.62

Table 3. Estimated frequencies using the exponential marginals under independence

	X_2					
X_1	$0 \to 1$	$1 \to 2$	$2 \to 3$	$3 \to 4$	≥ 4	$\hat{\Pi}_{i.}$
$0 \to 1$	74.70	41.06	22.77	12.62	17.25	0.476
$1 \to 2$	39.01	21.45	11.89	6.59	9.01	0.249
$2 \to 3$	19.94	10.96	6.08	3.37	4.60	0.127
$3 \to 4$	10.17	5.59	3.10	1.72	2.35	0.065
≥ 4	13.16	7.23	4.01	2.22	3.04	0.084
$\hat{\Pi}_{.j}$	0.444	0.244	0.135	0.075	0.102	

with d.f. = 16. The observed frequencies are given in Table 4. Both these tests are significant at the 5% level. Thus we conclude that X_1 and X_2 are not independent.

The test for symmetry was also done adopting the BVE distribution. Under the hypothesis of symmetry H_0: $\lambda_1 = \lambda_2$, the estimates of the parameters were $\hat{\lambda}_1 = \hat{\lambda}_2 = 0.58$ and $\hat{\lambda}_{12} = 0.05$. These values were used for evaluating the required probabilities. The χ^2 value was 46.31, which implied significance at the 5% level of significance. We conclude that the model is not symmetric. See Table 5 for the estimated frequencies under symmetry. Tables 6 and 7 show the expected frequencies for the parametric and semi-parametric models respectively. These were obtained from the cell probabilities calculated

Table 4. Estimated frequencies using the non-parametric approach under independence

	X_2					
X_1	$0 \to 1$	$1 \to 2$	$2 \to 3$	$3 \to 4$	≥ 4	$n_{i.}$
$0 \to 1$	62.53	33.06	22.70	10.75	11.95	141
$1 \to 2$	45.81	26.96	18.52	8.77	9.78	115
$2 \to 3$	20.84	11.02	7.56	3.58	3.98	47
$3 \to 4$	11.09	5.86	4.03	1.91	2.12	25
≥ 4	11.53	6.06	4.19	1.98	2.28	26
$n_{.j}$	157	83	57	27	30	354

Table 5. Estimated frequencies under symmetry of the BVE distribution

	X_2				
X_1	$0 \to 1$	$1 \to 2$	$2 \to 3$	$3 \to 4$	≥ 4
$0 \to 1$	78.73	36.67	21.42	12.39	16.85
$1 \to 2$	52.71	31.68	15.40	8.35	9.84
$2 \to 3$	14.05	9.17	6.05	2.55	2.97
$3 \to 4$	10.96	6.17	3.46	2.19	1.02
≥ 4	9.13	7.54	4.21	2.34	2.97

Table 6. Estimated frequencies for the parametric model

	X_2				
X_1	$0 \to 1$	$1 \to 2$	$2 \to 3$	$3 \to 4$	≥ 4
$0 \to 1$	81.49	36.92	18.87	12.78	6.8
$1 \to 2$	54.52	24.50	11.01	2.51	8.92
$2 \to 3$	22.41	12.07	7.29	3.29	3.36
$3 \to 4$	12.45	6.66	3.61	2.16	1.98
≥ 4	5.03	8.07	4.32	2.36	2.69

Table 7. Estimated frequencies for the semi-parametric model

	X_2				
X_1	$0 \to 1$	$1 \to 2$	$2 \to 3$	$3 \to 4$	≥ 4
$0 \to 1$	78.59	36.82	21.42	12.57	7.08
$1 \to 2$	52.71	24.74	11.65	6.20	7.05
$2 \to 3$	18.97	11.61	7.40	3.47	3.96
$3 \to 4$	9.17	6.20	3.47	2.19	2.19
≥ 4	8.28	7.08	3.93	2.23	2.83

as described earlier. With the satisfactory result obtained from these tests and tables, it is then obvious that this model has been an appropriate choice.

Conclusions

By determining the conditional probabilities of length of stay in the grades, one can easily assess the level of dependency of the length of stay in the two grades and on the individual promotion prospects on entry to a grade. The above model can also be used in predicting sojourn times in different grades – this work is ongoing.

With the joint distribution function, one can also determine the expected times spent in each part of the system given the grade of entry. We may similarly use our above formulation to investigate the movement pattern prevalent in the system. Yet another important application of this approach is that we can obtain the probability of an individual's sojourn time in the present state given the sojourn time in the last state. Further, the extension of this approach to more than three grades may be considered.

Nonetheless, there are some limitations in this study. For example, this model cannot be applied if the marginals do not follow exponential distribution. We note that fitting exponential distribution to length of service has been criticised in Bartholomew *et al*. (1991). They suggested the use of lognormal since it always has a peaked distribution. This limitation can also be avoided by simply invoking other multivariate exponential or gamma distributions in our general approach. However, for promotional data the length of stay in a grade is usually shorter than the length of service. We observed that exponential distribution did give good fit to each marginal modelling this short stay in a grade before moving to the next higher grade and the test for adequacy of fit of the model confirmed that the BVE model is quite a reasonable choice.

Acknowledgements

The contributions of TWAS-UNESCO and the host, Indian Statistical Institute, Kolkata, which enabled the second author to conduct this research work as a 2001–2003 TWAS Associate is hereby acknowledged with thanks. The authors express their thanks and appreciation to a referee for constructive criticisms of this paper and for providing encouragement on the line of research enhanced herein.

References

Bartholomew, D. J. (1982) *Stochastic Models for Social Processes*, 3rd edn (New York: Wiley).

Bartholomew, D. J. *et al.* (1991) *Statistical Techniques for Manpower Planning*, 2nd edn (New York: Wiley).

Basu, A. P. (1988) Multivariate exponential distributions and their applications in Reliability, in: P. R. Krishnaiah & C. R. Rao (Eds) *Handbook of Statistics*, Vol. 7 (Amsterdam: Elsevier).

Block, H. W. & Savits, T. H. (1981) Multivariate distribution in reliability theory and life testing, in: C. Taillie *et al.* (Eds) *Statistical Distributions in Scientific Work*, Vol. 5, pp. 271–288 (Dordrecht: Reidel).

Chen, T. H. H. *et al.* (2000) Estimation of sojourn time in chronic diseases screening without data on interval class, *Biometrics*, 56, pp. 167–172.

Chen, J. S. & Porok, P. C. (1983) Lead time estimating in a controlled screening, *American Journal of Epidemiology*, 118, pp. 740–751.

Cressie, N. & Read, T. R. C. (1984) Multinomial goodness-of-fit tests, *Journal of Royal Statistical Society*, 46, pp. 440–464.

Day, N. E. & Walter, S. D. (1984) Simplified models of screening for chronic disease: estimation procedures from mass screening programs, *Biometrics*, 40, pp. 1–14.

Duffy, S. W. *et al.* (1995) Estimation of mean sojourn time in breast cancer screening using a Markov chain model of both entry to and exit from the preclinical detectable phase, *Statistics in Medicine*, 14, pp. 1531–1543.

Gani, J. (1963) Formulae for projecting enrolments and degrees awarded in universities, *Journal of Royal Statistical Society*, Ser A, 126, pp. 400–409.

Marshall A. W. & Olkin, I. (1967) A multivariate exponential distribution, *Journal of American Statistical Association*, 62, pp. 30–44.

McClean, S. I. (1978) Continuous stochastic models of a multigrade population, *Journal of Applied Probability*, 15, pp. 26–32.

McClean, S. I. (1980) A semi-Markov model for a multigrade population with Poisson recruitment, *Journal of Applied Probability*, 17, pp. 846–852.

McClean, S. I. (1993) Human resource management: Semi-Markov model for human resource modelling, *I.M.A Journal of Mathematics Applied in Business and Industry*, 4, pp. 307–315.

McClean, S. I. *et al.* (1998) Non-homogeneous continuous time Markov and semi-Markov manpower models, *Applied Stochastic Model and Data Analysis*, 13, pp. 191–198.

Mehlmann, A. (1979) Semi-Markovian manpower models in continuous time, *Journal Applied Probability*, 16, pp. 416–422.

Paci, E. & Duffy, S. W. (1991) Modeling the analyses of breast cancer screening programs: sensitivity, lead-time, and predictive value in the Florence district program (1975–1986), *International Journal of Epidemiology*, 20(4), pp. 854–858.

SenGupta, A. (1995) Optimal tests in multivariate exponential distributions, in: N Balakrishnan & A. P. Basu (Eds) *The Exponential Distribution*, Chapter 22, pp. 351–376 (USA: Gordon & Breach).

Tabar, L. *et al.* (1995) Efficacy of breast cancer screening by age: new results from the Swedish two-county trial, *Cancer*, 75, pp. 2507–2517.

Vassiliou, P. C. G. & Papadopoulou, A. A. (1992) Non-homogeneous semi-Markov systems and maintainability of the state sizes, *Journal of Applied Probability*, 29, pp. 519–534.