Nonparametric Estimation of a Multivariate Distribution in the Presence of Censoring

James A. Hanley; Milton N. Parnes

Biometrics, Vol. 39, No. 1. (Mar., 1983), pp. 129-139.

Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28198303%2939%3A1%3C129%3ANEOAMD%3E2.0.CO%3B2-Q

Biometrics is currently published by International Biometric Society.

Biometrics 39, 129–139
March 1983

Nonparametric Estimation of a Multivariate Distribution in the Presence of Censoring

James A. Hanley[1]
Sidney Farber Cancer Institute and Harvard School of Public Health, 44 Binney Street,
Boston, Massachusetts 02115, U. S. A.
and
Milton N. Parnes
Department of Statistics, Temple University, Philadelphia, Pennsylvania 19122, U.S.A.

Summary

This paper presents examples of situations in which one wishes to estimate a multivariate distribution from data that may be right-censored. A distinction is made between what we term 'homogeneous' and 'heterogeneous' censoring. It is shown how a multivariate empirical survivor function must be constructed in order to be considered a (nonparametric) maximum likelihood estimate of the underlying survivor function. A closed-form solution, similar to the product-limit estimate of Kaplan and Meier, is possible with homogeneous censoring, but an iterative method, such as the EM algorithm, is required with heterogeneous censoring. An example is given in which an anomaly is produced if censored multivariate data are analyzed as a series of univariate variables; this anomaly is shown to disappear if the methods of this paper are used.

1. Introduction

The time until the occurrence of an event is a commonly studied variable. Frequently, for example in clinical trials, data must be analyzed before all of the subjects have experienced the event. Methods of estimating an underlying distribution from such 'censored' data are well established (Kalbfleisch and Prentice, 1980). However, one is often interested in *several* endpoints—for example, when various teeth erupt, various sexual features develop, various behaviours (e.g. smoking and drinking) are adopted, or certain disease symptoms are first experienced. Unfortunately, very few multivariate distributions other than the multivariate normal (Bhattacharya, 1954; Cohen, 1955; Nath, 1974) have been suggested to help model such phenomena. The univariate negative exponential is the only other distribution for which a multivariate analogue has received serious study; even then, there are several versions (Gumbel, 1960; Freund, 1961; Marshall and Olkin, 1967). The difficulty in extending univariate time-to-onset distributions to many dimensions is compounded by the increased opportunity for complex censoring; subjects may have reached some, but not all, of the endpoints by the time of the analysis, and different endpoints may be checked at different intervals or measured from different origins.

As a result of these difficulties, the various endpoints are often analyzed separately by means of well-understood univariate actuarial techniques. However, analysis of each marginal distribution ignores valuable information about the inter-relationships among endpoints, and indeed can even lead to paradoxical results as will be illustrated later. Rather than remedy

[1] Present address: Department of Epidemiology and Health, McGill University, 3775 University Street, Montreal, Quebec, Canada H3A 2B4.

Key words: Multivariate; Censoring; Estimation; Self-consistency; EM algorithm.

this by attempting to build complex parametric multivariate models, it would seem more logical to begin by constructing an empirical multivariate distribution. In this way, the data can be summarized in a multivariate way and parametric representations can be suggested. Barlow and Proschan have suggested this approach in two papers, although they have limited the discussion to uncensored data. Their 'multivariate life-table analysis' (Technical Report ORC 76-9, Operations Research Center, University of California, Berkeley, 1976) assumes complete observations and deals only with asymptotic properties of the resulting empirical distribution function, while their second article (Barlow and Proschan, 1977) shows how to estimate a 'multivariate hazard gradient' using a piecewise exponential model. Again the treatment assumes complete observations although it could be extended to include censored data.

In the present paper we show how to construct a multivariate empirical survivor function (mesf) from censored data. We distinguish between types of censoring where the mesf can be constructed explicitly and where an iterative procedure, such as the EM algorithm, is necessary. To simplify the presentation, but without loss of generality, we consider bivariate data. Higher dimensions pose no additional difficulties other than notational ones. The problem is formulated in §2, closed-form mesfs for a frequently encountered type of right censoring are presented in §3, and a more general form of right censoring, requiring iterative estimates, is discussed in §4. Section 5 concludes with a discussion.

2. Formulation

Let $\mathbf{T} = (T_1, T_2)$ represent a bivariate random variable denoting the durations before two events occur. The components could refer either to some event in each of two paired organs, e.g. blindness in an individual, to an event in each of two related individuals, e.g. menarche in twin girls, or to two different events in the same individual, e.g. smoking and drinking. Let $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_n$ represent n independent realizations of $\mathbf{T}$. Then by direct enumeration one can form the empirical 'survivor function' $\hat{S}(\mathbf{t})$ as an estimate of the underlying 'survivor' function $S(\mathbf{t}) = S(t_1, t_2) = \text{pr}(T_1 > t_1;\ T_2 > t_2)$. See Barlow and Proschan (1977) for a full discussion.

However, constraints of time and experimental design often force one to analyze the data before all $2n$ events have occurred. Suppose that one can only spend a maximum of L_{ij} time units waiting for the jth ($j = 1, 2$) of the two events to occur in Subject i ($i = 1, \ldots, n$) and that the L_{ij} are independent of the t_{ij} (L_{i1} need not be independent of L_{i2}, but the $\mathbf{L}_i$ are assumed independent of each other). Then the observable data for Subject i are contained in the two vectors $\mathbf{t}_i^*$ and $\mathbf{z}_i$, where, for $j = 1, 2$, $= t_{ij}^* = \min(t_{ij}, L_{ij})$ and $z_{ij} = 1$ if $t_{ij}^* = t_{ij}$ and 0 otherwise. Each t_{ij}^* for which $z_{ij} = 0$ is called a censored observation, since the limit of observation, L_{ij}, 'censors' the actual t_{ij}. Equivalently, one can represent the data on Subject i by noting that $\mathbf{t}_i$ belongs to a region or subset $\mathscr{R}_i$ of the space of $\mathbf{T}$. The region $\mathscr{R}_i$ will be an elemental rectangle, a horizontal or vertical strip, or an open quadrant, depending on whether $\mathbf{z}_i = (1, 1), (0, 1), (1, 0)$ or $(0, 0)$ (see Fig. 1 for examples).

Consider choosing, from among *all* admissible survivor functions $S(\mathbf{t})$, one, denoted by $\hat{S}(\mathbf{t})$, which maximizes the likelihood $\mathscr{L}$ of the observed data. [We remark in passing that the concept of nonparametric maximum likelihood estimation poses certain measure-theoretic difficulties, which are mentioned by Kalbfleisch and Prentice (1980, pp. 12–13) and discussed more fully by Scholz (1980), who extends the classical definition of maximum likelihood so that it provides a consistent approach to nonparametric maximum likelihood problems.] For any specified probability distribution $p(\mathbf{t}) = dS(\mathbf{t})$ on $\mathbf{T}$,

$$\mathscr{L} \propto \prod_i \int_{\mathscr{R}_i} p(\mathbf{t}) \equiv \prod_i P_i, \tag{1}$$

integrals and differentials being used for both discrete and continuous-type random variables. Although the L_{ij} are stochastic, they provide no information about $S(\mathbf{t})$ and so are considered

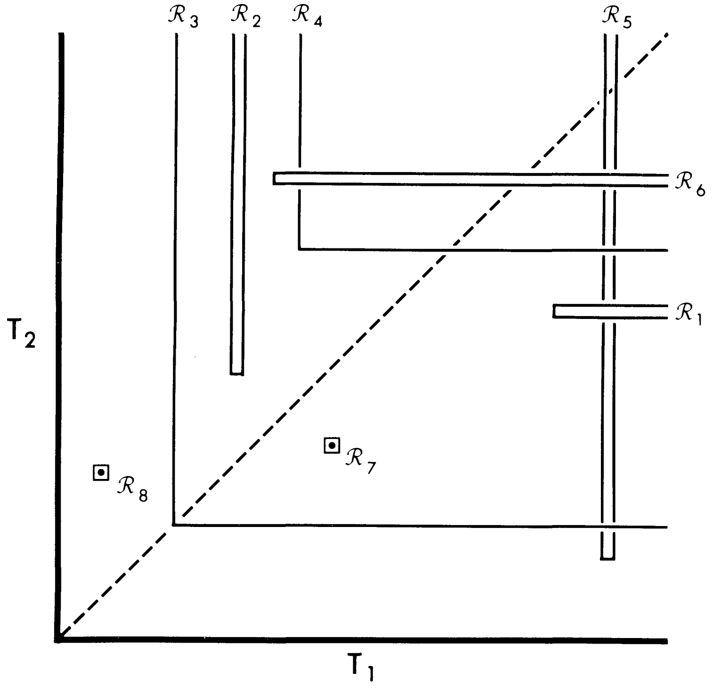

Figure 1. Data-defined regions corresponding to complete (Regions 7 and 8), half-censored (Regions 1, 2, 5 and 6) and doubly-censored (Regions 3 and 4) observations. Regions 4, 5 and 6 arose from heterogeneous censoring.

'constants' in $\mathscr{L}$. In order to maximize $\mathscr{L}$, $\hat{S}$ or equivalently $\hat{p}(\mathbf{t})$ must be constructed as follows:

(i) the entire probability mass must be distributed within $\cup_i \mathscr{R}_i$; mass placed outside the data-defined regions $\mathscr{R}_i$ will not contribute to any of the terms of $\mathscr{L}$, and will not help to maximize it;

(ii) each $\mathscr{R}_i$ must receive some probability mass, otherwise $\mathscr{L}$ will vanish;

(iii) if either component of $\mathbf{t}_i$ is censored, its contribution $P_i = \int_{\mathscr{R}_i} p(\mathbf{t})$ to $\mathscr{L}$ is not affected by how $p(\cdot)$ is distributed within $\mathscr{R}_i$; thus, P_i should be arranged so that it is maximally shared by other regions, $\mathscr{R}_j$, that are contained in, or intersect with, $\mathscr{R}_i$. In this way, the contributions P_j of these other regions will be increased without changing P_i.

Stated in set-theoretic terms, this implies that the total probability mass should be distributed over the maximal intersections $\mathscr{A}_1, \ldots, \mathscr{A}_m$ of the $\mathscr{R}_i$ (by a maximal intersection $\mathscr{A}$ we mean a nonempty finite intersection of the $\mathscr{R}_i$ such that for each i, $\mathscr{A} \cap \mathscr{R}_i = \phi$ or $\mathscr{A}$). Some of the maximal intersections will each contain just one point, which is either an observed (uncensored) $\mathbf{t}$ or possibly an intersection of two 'half-censored' observations. The single points in these sets form unambiguous support points for $\hat{p}(\cdot)$. In the case where a maximal intersection $\mathscr{A}$ consists of more than a single point, there is no unique choice of specific support points from $\mathscr{A}$, and we can without any loss of generality refer either to $\mathscr{A}$ itself or choose a point $\mathbf{a}$ from $\mathscr{A}$ to represent it. However, because we find it easier to focus on a probability distribution over specific points rather than over sets, we will write $p(\mathbf{a})$ rather than $p(\mathscr{A})$.

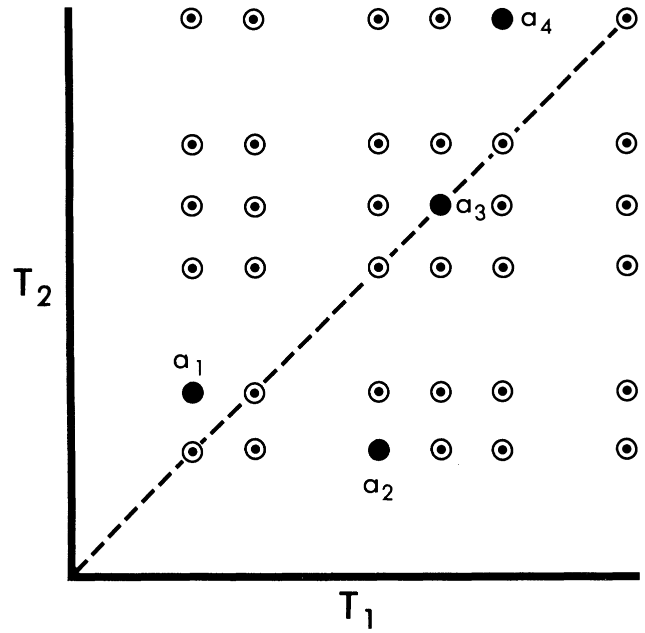

Figure 2. Augmentation and relabeling of original support points a_1, a_2, a_3 and a_4 (solid circles) as a rectangular grid (open and solid circles). Grid formed from (i) intersections of vertical and horizontal lines through original points and (ii) intersections of vertical and horizontal lines through points where diagonal line crosses the lines through these original points. The augmented set of points is then relabeled a_{11} to a_{66}, with the first subscript referring to T_1 and the second to T_2.

Although these guidelines cannot be formalized into any obvious algorithm, we think they are readily illustrated by the example in Fig. 1. From the $n = 8$ regions $\mathscr{R}_i$ shown there, we can construct a support consisting of $m = 5$ sets $\mathscr{A}_1, \ldots, \mathscr{A}_5$. From the previous discussion, it is clear that the observations which generated $\mathscr{R}_7$ and $\mathscr{R}_8$ must form two of the support points, which we arbitrarily label $\mathbf{a}_1$ and $\mathbf{a}_2$, and that the probability mass $p(\mathbf{a}_1)$ will contribute to both P_7 and P_3 and thus to $\mathscr{L}$. A third, $\mathbf{a}_3 = \mathscr{R}_5 \cap \mathscr{R}_6$, will contribute to P_4, P_5 and P_6, while a fourth, $\mathbf{a}_4 = \mathscr{R}_5 \cap \mathscr{R}_1$, will contribute to both P_5 and P_1. For the fifth component of the support for $\hat{p}(\cdot)$, one can take either the entire $\mathscr{R}_2$ region or any arbitrary point $\mathbf{a}_5 \in \mathscr{R}_2$; in any event, $p(\mathbf{a}_5)$ will contribute to both P_3 and P_2.

Once a support set $A = \{\mathbf{a}_1, \ldots, \mathbf{a}_m\}$ has been chosen for $\hat{p}(\cdot)$, the likelihood $\mathscr{L}$ can be written as

$$\mathscr{L} \propto \prod_i \left\{ \sum_{\mathbf{a}_k \in \mathscr{R}_i} p(\mathbf{a}_k) \right\}, \qquad (2)$$

or, in the example in Fig. 1 [with the $m = 5$ support points $a_1, \ldots, a_5$ receiving probability masses of $p(\mathbf{a}_1) = p_1, \ldots, p(\mathbf{a}_5) = p_5$], as

$$\mathscr{L} \propto p_4 p_5 (p_3 + p_4 + p_5) p_3 (p_3 + p_4) p_3 p_1 p_2 .$$

The remaining task, then, is to determine the magnitudes of $p(\mathbf{a}_1), \ldots, p(\mathbf{a}_m)$. To do this it is helpful to distinguish two censoring patterns which we will call 'homogeneous' and 'heterogeneous'; we deal with these in §3 and §4, respectively.

3. Estimating S(t) from Homogeneously Censored Data

We label the censoring 'homogeneous' if every two data-defined regions $\mathcal{R}_i$ and $\mathcal{R}_j$ are either disjoint or nested one within the other. This pattern occurs for example when one follows a subject for equal lengths of time towards each endpoint, i.e. if $L_{i1} = L_{i2}$. With this type of follow-up, incomplete observations can be represented by regions which are either (i) horizontal strips lying entirely to the right of the diagonal $T_1 = T_2$, (ii) vertical strips lying entirely above this diagonal, or (iii) squares which are open to the right and have their lower left corner on the diagonal (Regions 1, 2 and 3, respectively, in Fig. 1). In describing how the probability mass $p(\cdot)$ is to be assigned over the support set A, it will simplify the presentation if we introduce additional support points, which will receive zero mass in the ML estimation, but which allow us to speak of a grid of K^2 support points a_{11} to a_{KK} (see Fig. 2).

With an augmented and relabeled A, and abbreviating $p(a_{rs})$ to p_{rs}, P_i can be written as a sum over a rectangular grid

$$P_i = \int_{\mathcal{R}_i} p = \sum_r \sum_s p_{rs}, \tag{3}$$

where the summation index r(or s) runs from the leftmost to rightmost (lowest to highest) a_{rs} in $\mathcal{R}_i$. The $K^2 - 1$ probabilities can be reparameterized into $K^2 - 1$ equivalent conditional ones, $3(K-1)$ corresponding to $K-1$ sets of quadrinomial probabilities and $(K-1)(K-2)$ to binomial probabilities. Shown schematically in Fig. 3, they correspond respectively to the $K-1$ step-by-step conditional probabilities of advancing along the diagonal until one (or possibly both) of the events has occurred, then (if necessary) proceeding parallel to the vertical or horizontal axis, in $(K-1)(K-2)$ possible steps, until the second event is reached. Letting $X(a)$ and $Y(a)$ refer to the first and second coordinate values of a, the probabilities can be written as:

$$\left.\begin{aligned}
\phi_{kk} &= \mathrm{pr}\{T_1 > X(a_{kk});\ T_2 > Y(a_{kk}) \mid T_1 \geqslant X(a_{kk});\ T_2 \geqslant Y(a_{kk})\} \\
&= \sum_{r>k} \sum_{s>k} p_{rs} \Big/ \sum_{r\geqslant k} \sum_{s\geqslant k} p_{rs}, \\
\phi_{k\cdot} &= \sum_{s>k} p_{ks} \Big/ \sum_{r\geqslant k} \sum_{s\geqslant k} p_{rs}, \\
\phi_{\cdot k} &= \sum_{r>k} p_{rk} \Big/ \sum_{r\geqslant k} \sum_{s\geqslant k} p_{rs},
\end{aligned}\right\} 1 \leqslant k \leqslant K-1, \tag{4}$$

and

$$\begin{aligned}
\psi_{kl} &= \mathrm{pr}\{T_2 > Y(a_{kl}) \mid T_1 = X(a_{kl});\ T_2 \geqslant Y(a_{kl})\} \\
&= \sum_{s>l} p_{ks} \Big/ \sum_{s\geqslant l} p_{ks}, \qquad 1 \leqslant k < l \leqslant K-1,
\end{aligned} \tag{5}$$

$$\psi_{kl} = \sum_{r>k} p_{rl} \Big/ \sum_{r\geqslant k} p_{rl}, \qquad 1 \leqslant l < k \leqslant K-1. \tag{6}$$

Each P_i can be written as a product of ϕ and ψ terms. By doing this and rearranging terms and collecting exponents, $\mathcal{L}$ is simplified to the mathematical equivalent of a chain of quadrinomial and binomial expressions, namely

$$\mathcal{L} = \prod_k \Bigg[\phi_{kk}^{N_{kk}} \phi_{k\cdot}^{n_{k\cdot}} \phi_{\cdot k}^{n_{\cdot k}} (1 - \phi_{kk} - \phi_{k\cdot} - \phi_{\cdot k})^{n_{kk}} \left\{ \prod_{l>k} \psi_{kl}^{m_{kl}} (1 - \psi_{kl})^{n_{kl}} \right\} \left\{ \prod_{k>l} \psi_{kl}^{m_{kl}} (1 - \psi_{kl})^{n_{kl}} \right\} \Bigg], \tag{7}$$

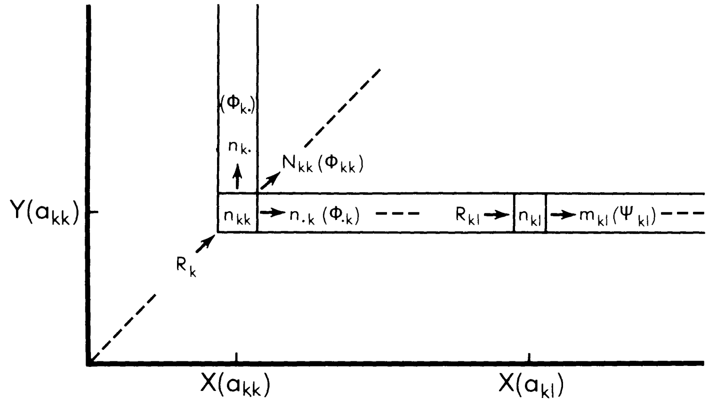

Figure 3. Illustration of quadrinomial and binomial reparameterization. *R*, *N*, *n* and *m* refer to observed counts; ϕ and ψ refer to expected proportions or probabilities. For full description, see text.

where, as is depicted in Fig. 3, the exponents refer to the following counts of sample members:

n_{kl}, those where the two events occurred at $X(a_{kl})$ and $Y(a_{kl})$;
N_{kk}, those proceeding through a_{kk} without either event taking place;
$n_{k\cdot}$, those where the '*X*' event occurs at $X(a_{kk})$ but the other occurs after $Y(a_{kk})$;
$n_{\cdot k}$, the converse of $n_{\cdot k}$;
m_{kl} $(l > k)$, those where, the '*X*' event already having taken place at $X(a_{kk})$, the '*Y*' component proceeds through $Y(a_{kl})$;
m_{kl} $(l < k)$, the converse.

The ML estimates of the ϕ_{kk}, $\phi_{k\cdot}$ and $\phi_{\cdot k}$ are then simply the proportions N_{kk}/R_k, $n_{k\cdot}/R_k$ and $n_{\cdot k}/R_k$, respectively, where $R_k = N_{kk} + n_{k\cdot} + n_{\cdot k} + n_{kk}$ denotes the number in the 'risk set' at a_{kk}. Similarly, $\hat{\psi}_{kl} = m_{kl}/R_{kl}$, where $R_{kl} = m_{kl} + n_{kl}$. The ML estimates of the original p_{rs} are obtained from (4), (5) and (6),

$$\hat{p}_{rs} = \begin{cases} \left(\prod_{k=1}^{r-1} \hat{\phi}_{kk}\right)(1 - \hat{\phi}_{rr} - \hat{\phi}_{r\cdot} - \hat{\phi}_{\cdot r}) & \text{if } s = r, \\ \left(\prod_{k=1}^{r-1} \hat{\phi}_{kk}\right)\hat{\phi}_{r\cdot}\left(\prod_{l=r+1}^{s-1} \hat{\psi}_{rl}\right)(1 - \hat{\psi}_{rs}) & \text{if } s > r, \\ \left(\prod_{k=1}^{s-1} \hat{\phi}_{kk}\right)\hat{\phi}_{\cdot s}\left(\prod_{k=s+1}^{r-1} \hat{\psi}_{ks}\right)(1 - \hat{\psi}_{rs}) & \text{if } r > s. \end{cases} \tag{8}$$

The ML estimate of $S(t_1, t_2)$ can be obtained by summing the $\hat{p}_{rs}$ in the open rectangle $(t_1, \infty) \times (t_2, \infty)$.

If T_1 and T_2 are continuous-type variables, then $n_{kk} + n_{\cdot k} + n_{k\cdot} \leqslant 1$; thus at least two of the corresponding ϕ terms will be estimated as zero. If $n_{k\cdot} = 1$, then the m_{kl} $(l = k + 1, \ldots)$ will each be unity until the other event takes place somewhere beyond $Y(a_{kk})$, after which they will equal zero. If the observation time runs out before this second event occurs, the ψ parameters beyond the last *Y* observation on this subject cannot be uniquely estimated. This nonuniqueness is similar to the problem that occurs in the univariate Kaplan–Meier survival curve when the largest observation is censored. In the multivariate case it occurs each time a pair of values, recorded on a continuous scale, is 'half-censored', and means that one cannot supply a unique estimate for $S(t_1, t_2)$ when the region $(T_1 > t_1;\ T_2 > t_2)$ contains such observations. This shortcoming can be lessened by discretizing or grouping the data into intervals, as is commonly done in univariate life-tables, so that there are fewer ϕ and ψ parameters to be estimated, and from larger, more stable, denominators. In fact, in many studies subjects are followed up on a fixed schedule so that t actually takes discrete values.

The variance of $\hat{S}(t_1, t_2)$ can be calculated from a heuristic extension of Greenwood's formula (Kaplan and Meier, 1958). It involves computing and summing the variances and covariances of each $\hat{p}_{rs}$ in $(t_1, \infty) \times (t_2, \infty)$, i.e.

$$\operatorname{var}\{\hat{S}(\mathbf{t})\} = \sum_r \sum_s \sum_{r'} \sum_{s'} \operatorname{cov}(\hat{p}_{rs}, \hat{p}_{r's'}). \tag{9}$$

If $r < s$, the variance terms in (9) can be approximated by

$$\operatorname{var}(\hat{p}_{rs}) = \operatorname{var}\left\{\left(\prod_{k=1}^{r-1} \hat{\phi}_{kk}\right)\hat{\phi}_{r.}\left(\prod_{l=r+1}^{s-1} \hat{\psi}_{rl}\right)(1-\hat{\psi}_{rs})\right\} \tag{10}$$

$$\simeq p_{rs}^2 \left\{\sum_{k=1}^{r-1} \frac{1-\phi_{kk}}{R_k\phi_{kk}} + \frac{1-\phi_{r.}}{R_r\phi_{r.}} + \sum_{l=r+1}^{s-1} \frac{1-\psi_{rl}}{R_{rl}\psi_{rl}} + \frac{\psi_{rs}}{R_{rs}(1-\psi_{rs})}\right\}, \tag{11}$$

with the obvious converse if $r > s$. This approximation is obtained by appealing to the approximate independence of the multinomial and binomial terms in (10) and by using the fact that if $X_1, X_2, \ldots$ are independent positive random variables,

$$\operatorname{var}(\prod X_i) \simeq \{\prod \mathrm{E}(X_i)\}^2 \sum[\operatorname{var}(X_i)/\{\mathrm{E}(X_i)\}^2].$$

Similarly, if $Y_1, Y_2, \ldots$ are independent positive random variables which are also independent of the X_i, and if Z_1 and Z_2 are correlated but independent of both the X_i and the Y_i, then the covariance between

$$\left(\prod_{i=1}^{p} X_i\right)Z_1\left(\prod_{i=p+1}^{q} X_i\right) \text{ and } \left(\prod_{i=1}^{p} X_i\right)Z_2\left(\prod_{j=1}^{r} Y_j\right)$$

is approximately given by the expression

$$\left[\operatorname{var}\left(\prod_{i=1}^{p} X_i\right)\mathrm{E}(Z_1Z_2) + \left\{\mathrm{E}\left(\prod_{i=1}^{p} X_i\right)\right\}^2 \operatorname{cov}(Z_1, Z_2)\right]\left\{\prod_{i=p+1}^{q} \mathrm{E}(X_i)\right\}\left\{\prod_{j=1}^{r} \mathrm{E}(Y_j)\right\}. \tag{12}$$

Thus the covariance terms in (9) can be obtained by first expressing each $\hat{p}_{rs}$ and $\hat{p}_{r's'}$ as pairs of products, as in (8), and then applying (12).

Space does not permit us to present data and results of an analysis of a real data set containing homogeneous censoring, but examples are available on request. However, the following, somewhat extreme, hypothetical example will illustrate how much more efficient a multivariate analysis can be, even if one is only interested in each variable separately. Suppose one takes a sample of n from the following bivariate distribution:

	3	$\frac{1}{3}$	0	0
T_2	2	0	0	$\frac{1}{3}$
	1	0	$\frac{1}{3}$	0
	0			
	0	1	2	3
			T_1	

and that one half of the subjects can only be followed for one time unit. If the quantity $\mathrm{pr}(T_1 > 2)$ is estimated by using $\hat{S}(\mathbf{t})$ to derive the marginal distribution $\hat{S}_1(t)$, its sampling variance[1] will average $4/(18n)$, which incidentally equals the precision one would expect

[1] In this example $\mathrm{pr}_{\mathrm{est}}(T_1 > 2)$ involves three terms $\hat{p}_{31}, \hat{p}_{32}$ and $\hat{p}_{33}$. Only $\hat{p}_{32}$ has nonzero variance, obtained from (11) with $r = 3$, $s = 2$, $p_{32} = \frac{1}{3}$, $\phi_{11} = \frac{1}{3}$, $R_1 = n$, $\phi_{.2} = 1$, $R_2 = \frac{1}{6}n$, $\psi_{32} = 0$, $R_{32} = \frac{1}{6}n$.

using the $\hat{S}_1(t)$ computed from n *complete* univariate observations on T_1. By comparison, if the quantity $\text{pr}(T_1 > 2)$ is estimated by the usual method of Kaplan and Meier (1958), its sampling variance will, according to Greenwood's formula, average $7/(18n)$. This use of T_2 as auxiliary information is similar in spirit to the more parametric approach described by Lagakos (1976).

4. Estimating S(t) from Heterogeneously Censored Data

We say that the bivariate data are heterogeneously censored if some of the data-defined regions cannot be classified into one of the three types (i), (ii) and (iii) described at the beginning of §3. Examples are Regions 4, 5 and 6 in Fig. 1, which arise when the potential follow-up times for the two events are different, such as (a) when one monitors an individual for the occurrence of two medical conditions, which are detected by two tests performed at different intervals, and (b) when one follows a pair of individuals for different durations towards a single endpoint.

In this situation, every pair of regions is not necessarily nested or disjoint. As a result, a closed-form solution can no longer be found; instead one must simultaneously estimate all p_{rs} by an iterative technique. Available methods are the Newton–Raphson method, its modification known as Fisher's method of scoring (see Kendall and Stuart, 1967, p. 49) and the derivative-free EM algorithm described by Dempster, Laird and Rubin (1977). For a number of reasons, we suggest the latter: (i) it is very easily programmed for a computer; (ii) it avoids having to constrain the $\hat{p}_{rs}$ to [0, 1]; (iii) it avoids inversion of large matrices; (iv) it corresponds to the 'selfconsistency' construction suggested for univariate life-tables by Efron (1967); and (v) it has intuitive appeal in that one can observe how the successive approximations to the $\hat{p}_{rs}$ are formed by repeating the two 'E' and 'M' steps,

E-step:Estimate how the incomplete observations would, using the current estimates of the p_{rs}, be expected to distribute among the support points $\{a_{rs}\}$ and add these to the observed counts for the complete observations;

M-step:Form revised ML multinomial estimates of the p_{rs} based on these new 'counts', until the revisions no longer change the $\hat{p}_{rs}$. The approximate variance–covariance matrix of the $\hat{p}_{rs}$ can be obtained in the usual way from the matrix of second derivatives $(\partial^2 \log \mathscr{L}/\partial p_{rs}\partial p_{r's'})$, or by a modification to the EM algorithm suggested by Louis (1982).

In the example which follows, the variables of interest are a patient's tolerance to two cancer chemotherapy regimens, AV and CMF, studied using the Eastern Cooperative Oncology Group Protocol 2173 (1973). Patients with breast cancer were to first receive AV for a total of eight cycles, or until tolerance was reached or the disease progressed, whichever occurred first. They were then to receive as many as six cycles of CMF, again unless prohibited by toxicity or disease progression. If drugs were discontinued because of disease progression or reasons unrelated to treatment, the data on how many cycles of drug could be tolerated became censored. The data on 109 patients are shown in Table 1. Using the EM algorithm with a convergence criterion of $\sum_{rs} \{\hat{p}_{rs}(\text{new}) - \hat{p}_{rs}(\text{old})\}^2 < .001$, convergence to a maximum likelihood solution was reached in 50 iterations (the large number of iterations, compared with the Newton–Raphson method, reflects the linear rather than quadratic convergence). The estimated joint distribution is presented in Table 2. From these one can examine the dependence between the tolerance to the two regimens. Models which attempt to summarize this dependency can be tested for goodness of fit using the method of Turnbull and Weiss (1978).

5. Discussion

There are obvious benefits of analyzing censored multivariate 'failure-time' data using multivariate techniques. First, one uses *all* available data in the calculation of summaries

Table 1
Tolerance of full doses of AV (i) and CMF (j), including data censored by disease progression or protocol design

i	j						j					
	1	2	3	4	5	≥6	1	2	3	4	5	≥6
	n_{ij}						a_{ij}					
1	10	2	2	1	0	1	3	2	0	0	0	0
2	9	2	1	2	0	2	2	1	0	0	0	0
3	3	0	0	2	1	1	1	0	0	1	0	0
4	1	0	1	1	0	1	2	1	0	0	1	1
5	1	0	0	0	0	0	1	0	0	0	0	0
≥6	2	2	1	0	0	0	0	0	0	0	0	0
	b_{ij}						c_{ij}					
1	0	1	0	0	0	0	1	1	1	0	0	0
2	3	2	0	0	0	1	0	0	0	0	0	0
3	2	0	0	0	0	0	1	0	1	0	0	0
4	0	1	1	0	0	0	0	1	0	1	0	0
5	0	0	0	0	0	0	0	0	1	1	1	0
≥6	9	2	1	2	1	3	0	1	2	0	0	2

Key: n_{ij} = number tolerating exactly i and exactly j doses, a_{ij} = number tolerating exactly i and at least j doses, b_{ij} = number tolerating at least i and exactly j doses, c_{ij} = number tolerating at least i and at least j doses.

Table 2
Estimated probability distribution (from the data in Table 1)

i	j						$\hat{p}_{i\cdot}$	$\hat{p}_i$
	1	2	3	4	5	≥6		
	$\hat{p}_{ij}$							
1	.091	.025	.038	.019	.000	.019	.19	.19
2	.082	.025	.014	.029	.000	.030	.18	.18
3	.032	.000	.000	.023	.017	.017	.09	.10
4	.012	.000	.015	.017	.000	.050	.09	.10
5	.012	.015	.000	.000	.000	.000	.03	.02
≥6	.135	.068	.032	.003	.019	.132	.42	.41
$\hat{p}_{\cdot j}$	.36	.14	.10	.11	.04	.25		
$\hat{p}_j$	.36	.14	.09	.12	.04	.25		

such as correlation coefficients, times to first failure, patterns of failure etc.. Moreover, the methods we describe are distribution-free so that the data are free to suggest parametric models. Second, as a by-product of the greater precision, interrelationships between variables are preserved and gross inconsistencies produced by univariate analyses can be avoided. As a somewhat startling example, consider drawing a sample from a bivariate distribution in which one variable is stochastically larger than the other. Suppose the $n = 5$ data pairs are: (1.8, 5.2), (4.7, 11.8), (6.2+, 6.2+), (13.6+, 13.6+) and (17.0, 21.7) and that they represent T_1, the months to recurrence, and T_2, the months to cancer death, of five cancer patients (where $t+$ denotes an observation censored at t). By analyzing each variable separately by the method of Kaplan and Meier (1958), one obtains the two curves depicted in Fig. 4A, pointing up the logical inconsistency, namely that $\hat{S}_2(t) < \hat{S}_1(t)$ for some t, described by Thaler (Technical Report, Biostatistics Laboratory, Memorial Sloan-Kettering Cancer Center, New York, 1977). By using the bivariate methods of the present paper, one avoids the anomaly and, as is illustrated in Fig. 4B, the mesf produces marginal esfs which preserve the stochastic ordering.

In summary, this area of multivariate life-tables, and of arbitrarily censored multivariate

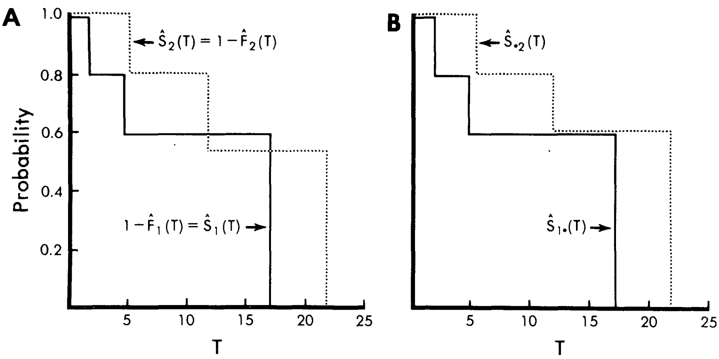

Figure 4. A, Survival curves estimated from the two separate univariate samples of T_1 and T_2. B, Survival curves estimated from the bivariate sample of (T_1, T_2).

data in general, is only just beginning to be explored. Although the estimation methods may appear to be straightforward generalizations of univariate methods, the infinitely larger range of even a bivariate random variable, compared to a univariate variable, quickly leads to problems of insufficient data and unstable estimates. To overcome them, and to make the nonparametric life tables more useful, one may need to combine the empirical distribution with a parametric prior distribution (see Susarla and Van Ryzin, 1976) or use parametric representations of the piecewise hazards. Tests of fit of such models can presumably be carried out using extensions of the test given by Turnbull and Weiss (1978) for univariate data. Presumably, the EM algorithm will easily accommodate more complex left, right and interval censoring (Turnbull, 1976), but the development of techniques which allow for systematic differences between subjects through the use of covariates poses a real challenge. With regard to the latter, the model put forward by Clayton (1976), though very specialized, is a step in this direction. Finally, we call attention to two papers on this subject by Campbell (1981) and Campbell and Földes (Mimeoseries 80-10, Department of Statistics, Purdue University, 1980) which came to our attention while the present paper was being revised.

Acknowledgements

This work was supported by a postdoctoral fellowship (M. N. P.) and by Grant Number CA-23415 (J. A. H.), awarded by the National Cancer Institute. We thank C. Begg, S. Lagakos, M. Pagano and M. Zelen for helpful comments and G. Manyan, J. MacIntyre and I. McCammon for typing the various drafts.

Résumé

Cet article présente des exemples de situations où l'on souhaite estimer une distribution multivariée pour des données qui peuvent être censurées à droite. Une distinction est faite entre des censures 'homogènes' et 'hétérogènes'. L'article montre comment construire une fonction de survie multivariée empirique, afin de construire un estimateur (non paramètrique) du maximum de vraisemblance pour la fonction de survie sous-jacente. Une solution, similaire à l'estimateur du produit limite de Kaplan et Meier, est possible avec une censure homogène, mais une méthode itérative, tel que l'algorithme EM, est nécessaire pour une censure hétérogène. Un exemple est donné, où se produit une anomalie si les données censurées multivariées sont analysées comme des séries de variables univariées; on montre que cette anomalie disparait à l'aide de la méthode présentée dans cet article.

References

Barlow, R. E. and Proschan, F. (1977). Techniques for analyzing multivariate failure data. In *Theory and Applications of Reliability with Emphasis on Bayesian and Nonparametric Method, Vol.* I, C. P. Tsokos and I. N. Shimi (eds), 373–396. New York: Academic Press.

Bhattacharya, M. N. (1954). Estimation from censored bivariate normal distributions. *Journal of the Indian Society of Agricultural Statistics* **6**, 83–92.

Campbell, G. (1981). Nonparametric bivariate estimation with randomly censored data. *Biometrika* **68**, 417–422.

Clayton, D. G. (1978). A model for association in bivariate life tables. *Biometrika* **65**, 141–151.

Cohen, A. C. (1955). Restriction and selection from bivariate normal distributions. *Journal of the American Statistical Association* **50**, 884–893.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the E.M. algorithm. *Journal of the Royal Statistical Society, Series B* **39**, 1–38.

Eastern Cooperative Oncology Group (1973). Phase II study of induction combination chemotherapy and maintenance hormonochemotherapy for metastatic breast carcinoma. Eastern Cooperative Oncology Group: 905 University Avenue, Madison, Wisconsin.

Efron, B. (1967). The two-sample problem with censored data. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Vol.* 4, L. LeCam and J. Neyman (eds), 831–853. Berkeley: University of California Press.

Freund, J. E. (1961). A bivariate extension of the exponential distribution. *Journal of the American Statistical Association* **56**, 971–977.

Gumbel, E. J. (1960). Bivariate exponential distributions. *Journal of the American Statistical Association* **55**, 698–707.

Kalbfleisch, J. D. and Prentice, R. L. (1980). *The Statistical Analysis of Failure Time Data.* New York: Wiley.

Kaplan, E. L. and Meier, P. (1958). Non-parametric estimation from incomplete observations. *Journal of the American Statistical Association* **53**, 457–481.

Kendall, M. G. and Stuart, A. (1967). *The Advanced Theory of Statistics, Vol.* 2. London: Griffin.

Lagakos, S. W. (1976). A stochastic model for censored survival data in the presence of an auxiliary variable. *Biometrics* **32**, 551–559.

Louis, T. A. (1982). Finding the observed information matrix when using the E.M. algorithm. *Journal of the Royal Statistical Society, Series B* **44**, 226–233.

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

Nath, G. B. (1974). Progressively censored bivariate normal samples. *Applied Statistics* **23**, 300–312.

Scholz, F. W. (1980). Towards a unified definition of maximum likelihood. *Canadian Journal of Statistics* **8**, 193–203.

Susarla, V. and Van Ryzin, J. (1967). Nonparametric bayesian estimation of survival curves from incomplete observations. *Journal of the American Statistical Association* **61**, 897–902.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B* **38**, 290–295.

Turnbull, B. W. and Weiss, L. (1978). A likelihood ratio statistic for testing goodness of fit with randomly censored data. *Biometrics* **34**, 367–375.

Received March 1980; *revised June and December* 1981