Maximum likelihood estimation of ordered multinomial parameters

NICHOLAS P. JEWELL
Division of Biostatistics, School of Public Health, University of California, Berkeley, USA

JOHN D. KALBFLEISCH
Department of Biostatistics, School of Public Health, University of Michigan, USA

SUMMARY
The pool adjacent violator algorithm (Ayer et al., 1955) has long been known to give the maximum likelihood estimator of a series of ordered binomial parameters, based on an independent observation from each distribution (see Barlow et al., 1972). This result has immediate application to estimation of a survival distribution based on current survival status at a set of monitoring times. This paper considers an extended problem of maximum likelihood estimation of a series of ‘ordered’ multinomial parameters \( p_i = (p_{i1}, p_{i2}, \ldots, p_{im}) \) for \( 1 \leq i \leq k \), where ordered means that \( p_{j1} \leq p_{j2} \leq \cdots \leq p_{jk} \) for each \( j \) with \( 1 \leq j \leq m - 1 \). The data consist of \( k \) independent observations \( X_1, \ldots, X_k \) where \( X_i \) has a multinomial distribution with probability parameter \( p_i \) and known index \( n_i \geq 1 \). By making use of variants of the pool adjacent violator algorithm, we obtain a simple algorithm to compute the maximum likelihood estimator of \( p_1, \ldots, p_k \), and demonstrate its convergence. The results are applied to nonparametric maximum likelihood estimation of the sub-distribution functions associated with a survival time random variable with competing risks when only current status data are available (Jewell et al., 2003).

Keywords: Competing risks; Current status data; Isotonic regression; Nonparametric maximum likelihood estimation.

1. INTRODUCTION

Suppose that \( X_1, \ldots, X_k \) are independent random variables where, for \( 1 \leq i \leq k \), \( X_i \) has a binomial distribution with known index \( n_i \) and probability \( p_i \), where \( 0 \leq p_1 \leq p_2 \leq \cdots \leq p_k \leq 1 \), and consider estimation of \( p_1, \ldots, p_k \). This problem has a rich history with many applications. Least squares and maximum likelihood yield the same estimator (Barlow et al., 1972), and the well-known pool-adjacent violator (PAV) algorithm (Ayer et al., 1955) provides a fast and straightforward method for computing it. Variations on this algorithm have been considered by several authors such as the Kruskal (1964) ‘up-and-down-blocks’ algorithm; see also Wu (1982). Current status observations of a survival random variable, \( T \), at an ordered sequence of independent monitoring times, \( C_1 < \cdots < C_k \), provides a specific example of this data structure. Nonparametric maximum likelihood estimation of the distribution function \( F \) of \( T \) corresponds to maximum likelihood estimation of \( p_i = F(C_i), \ i = 1, \ldots, k \) as above if, at each \( C_i \), we observe the number \( X_i \) out of \( n_i \) independent individuals, to have failed by time \( C_i \). These methods have been widely applied to estimation problems in such divergent fields as carcinogenicity testing, demography, economics, and epidemiology (Jewell and van der Laan, 2004).

We consider a more general problem in which, for each \( i, X_i = (X_{i1}, \ldots, X_{im}) \) is an independent multinomial variable with known index \( n_i \) and probability \( p_i = (p_{i1}, p_{i2}, \ldots, p_{imi}), \sum_{j=1}^{m} p_{ji} = 1 \),
where the \( p_j \) are known to satisfy the constraints

\[
0 \leq p_{j1} \leq p_{j2} \leq \cdots \leq p_{jk} \leq 1, \quad 1 \leq j \leq m - 1.
\]  

The log likelihood function is

\[
\ell = \sum_{i=1}^{k} \left[ X_{1i} \log(p_{i1}) + \cdots + X_{mi} \log(p_{mi}) \right],
\]

and the proposed maximum likelihood estimator (MLE) maximizes (2) subject to the constraints (1) and \( \sum_{j=1}^{m} p_{j1} = 1 \) for each \( i \). The problem of maximum likelihood estimation of a sequence of \( k \) ordered binomial parameters is the special case \( m = 2 \).

A key motivating application is provided by the need to estimate the properties of a survival random variable, \( T \), in the presence of \( m - 1 \) competing risks. This is described in Section 2, and illustrated by a simple example on characteristics of age at menopause, due to either operative or natural causes; the MLE for a data set for this situation is illustrated in Section 6.

For a given \( j \), the PAV yields a ‘naive MLE’ of the parameters \( p_{j1}, \ldots, p_{jk} \) by maximizing the product binomial likelihood for \( p_{j1}, \ldots, p_{jk} \) subject to the constraint \( 0 \leq p_{j1} \leq p_{j2} \leq \cdots \leq p_{jk} \leq 1 \). By the theory of isotonic regression, this estimator is consistent, but it is not generally the full MLE of \( p_{j} : 1 \leq i \leq k \) from the log likelihood (2). (For example, there is no guarantee that these naive estimators satisfy \( \sum_{j=1}^{m-1} \hat{p}_{ji} \leq 1 \) for all \( i, 1 \leq i \leq k \).) In this article, we derive a simple algorithm to compute the full MLE. In Section 7, we make some further comparison of the full MLE to the naive univariate isotonic MLEs.

Section 3 describes an iterative algorithm that, subject to the constraints, maximizes (2) over \( p_{j1}, p_{j2}, \ldots, p_{jk} \) for each \( j = 1, \ldots, m - 1 \) in turn, when the other parameters \( p_{l1}, p_{l2}, \ldots, p_{lk} \) for \( l \neq j, m \) are fixed. It is shown that this algorithm, with a slight modification, converges to the full MLE from (2). Section 4 develops a modified PAV algorithm to implement each of the univariate isotonic maximizations used in Section 3. Section 5 presents simple ‘toy’ examples; we encourage the reader to glance at this section first, and have the examples at hand as an aid to following the description of the MLE algorithm in Sections 3 and 4. A small simulation study is reported in Section 7. Convergence proofs are given in the Appendices.

For the rest of the paper, we restrict attention to the case where \( m = 3 \) for simplicity. It is straightforward to extend the algorithm and ideas to larger values of \( m \).

2. Motivating Example——Current Status Observation of a Survival Variable with Competing Risks

Consider a survival random variable, \( T \), where ‘failure’ can occur due to one of \( m - 1 \) competing risks. If \( J \) is the random variable that measures cause of failure, the sub-distribution functions of interest are defined by

\[
F_j(t) = \Pr(T < t, J = j),
\]

for \( j = 1, \ldots, m - 1 \), with the overall survival function given by

\[
S(t) = 1 - \sum_{j=1}^{m-1} F_j(t).
\]

Suppose, for each individual under study, only current status information on survival status is available at a single monitoring time \( C \), where it is assumed that, if an individual is known to have failed by the
Maximum likelihood estimation of ordered multinomial parameters

Table 1. Data on menopausal current status from MacMahon and Worcester (1966). Ci = age, ni = # respondents, xi =# operative menopause, yi=# natural menopause, zi=# non-menopausal

| Ci  | n_i | x_i | y_i | z_i | Ci  | n_i | x_i | y_i | z_i |
|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|
| 27.5 | 380 | 4  | 0  | 376 | 46.5 | 76 | 16 | 11 | 49 |
| 32.5 | 359 | 21 | 0  | 338 | 47.5 | 75 | 18 | 16 | 41 |
| 35.5 | 89  | 7  | 0  | 82  | 48.5 | 80 | 19 | 18 | 43 |
| 36.5 | 87  | 5  | 0  | 82  | 49.5 | 66 | 20 | 19 | 27 |
| 37.5 | 61  | 5  | 1  | 55  | 50.5 | 72 | 18 | 32 | 22 |
| 38.5 | 83  | 11 | 2  | 70  | 51.5 | 66 | 10 | 38 | 18 |
| 39.5 | 99  | 11 | 2  | 86  | 52.5 | 54 | 16 | 30 | 8  |
| 40.5 | 78  | 8  | 1  | 69  | 53.5 | 67 | 18 | 40 | 9  |
| 41.5 | 66  | 7  | 1  | 58  | 54.5 | 50 | 18 | 28 | 4  |
| 42.5 | 80  | 16 | 4  | 60  | 55.5 | 45 | 19 | 25 | 1  |
| 43.5 | 74  | 11 | 5  | 58  | 56.5 | 50 | 13 | 36 | 1  |
| 44.5 | 67  | 10 | 3  | 54  | 57.5 | 54 | 13 | 40 | 1  |
| 45.5 | 99  | 20 | 12 | 67  | 58.5 | 46 | 13 | 33 | 0  |
differential mortality or secular changes are unlikely to change estimates of incidence probabilities for natural or operative menopause in practice.

Similar examples arise from other applications where current status data occurs with more than one 'competing' outcome. An application from epidemiology occurs in studies of the time to HIV infection when two or more distinct HIV-1 subtypes are prevalent and when only current status information on infection status is available at a single monitoring time for each study participant; see Hudgens et al. (2001).

3. ITERATIVE MAXIMIZATION OVER COMPONENTS OF \( p \)

To avoid unnecessary use of subscripts, it is convenient to introduce a slightly different notation for the case \( m = 3 \). Let \((X_i, Y_i, Z_i)\) be a trinomial variate with index \( n_i \) and probabilities \( p_i, q_i, 1 - p_i - q_i \), independently for \( i = 1, \ldots, k \). We wish to maximize the log likelihood function

\[
\ell(p, q) = \sum_{i=1}^{k} [x_i \log p_i + y_i \log q_i + z_i \log (1 - p_i - q_i)],
\]

where \( p = (p_1, \ldots, p_k) \) and \( q = (q_1, \ldots, q_k) \). The parameter space,

\[
\Theta = \{(p, q) : 0 \leq p_1 \leq \cdots \leq p_k; 0 \leq q_1 \leq \cdots \leq q_k; 1 - p - q \geq 0\},
\]
is a compact convex set in \( \mathbb{R}^{2k} \).

A few general remarks follow easily at this stage.

R\(_1\) If for each \( i, 1 \leq i \leq k, x_i < n_i \) and \( y_i < n_i \), then \( \ell \) is a strictly concave function of \((p, q)\). As a consequence, there exists a unique MLE \((\hat{p}, \hat{q})\) of \((p, q)\) in \( \Theta \).

R\(_2\) If \( x_i = n_i \) for some \( i \), then the corresponding \( q_i \) does not enter the likelihood and the MLE of \( q_i \) will not be uniquely determined in general. It is possible however, to reduce the problem by considering (3) as a likelihood only in the remaining elements of \((p, q)\). With respect to these variables, the likelihood is again strictly concave and there is a unique MLE. We adopt a convention that, when \( q_i \) is missing from the likelihood, we take its estimate to be the same as that for \( q_{i-1} \) where we interpret \( q_0 = 0 \). There is a unique such MLE. A similar issue arises when \( y_i = n_i \) for some \( i \) where now \( p_i \) disappears from the likelihood. This is accommodated in an identical fashion.

R\(_3\) In general, \((\hat{p}, \hat{q})\) is an MLE if and only if the directional derivative from \((\hat{p}, \hat{q})\) toward any other point in \( \Theta \) is non-positive; that is, if and only if

\[
D_x[\ell(\hat{p}, \hat{q}); (p, q)] = \lim_{\epsilon \to 0^+} \frac{\ell((1 - \epsilon)(\hat{p}, \hat{q}) + \epsilon(p, q)) - \ell(p, q)}{\epsilon} \leq 0,
\]

for all \((p, q) \in \Theta \).

Consider now the one-dimensional problem of maximizing the likelihood with respect to \( p \) for given \( q \). This log likelihood can be written

\[
\ell^q(p; q) = \sum_{i=1}^{k} [x_i \log p_i + z_i \log (1 - p_i - q_i)]
\]

with \( q_i \) fixed. This is a one-dimensional isotonic problem, and we obtain the order restricted estimate \( \hat{p}(q) \) by a variation of the PAV algorithm as described in Section 4. Similarly, the isotonic estimate of \( q \) given \( p \) is \( \hat{q}(p) \). With a view to maximizing the joint likelihood (3), consider the following algorithm.
Cyclical Algorithm:

1. Let \( q^{(0)} \) be an initial estimate that satisfies the restriction \( q_1^{(0)} \leq \cdots \leq q_k^{(0)} \). Set \( j = 0 \).
2. Find \( p^{(j+1)} = \hat{p}(q^{(j)}) \); find \( q^{(j+1)} = \hat{q}(p^{(j+1)}) \).
3. Check for convergence. If not, then set \( j = j + 1 \) and go to 2.

**Theorem 1** If \( z_k > 0 \), the cyclical algorithm converges to \((\hat{p}, \hat{q})\), the unique MLE of \((p, q)\).

The proof of this theorem is given in Appendix A. When \( z_k = 0 \), the situation is more complicated since the successive iterated values of \( q_k \) and \( p_k \) do not change. If \( z_k = 0 \) and \( z_{k-1} > 0 \), Lemma 2 of Appendix A establishes that the algorithm converges to a constrained MLE—the constraint being that \( q_k \) is fixed at its starting value. To find the overall MLE, one simply computes the constrained maxima for the univariate maximizations needed in Step 2 of the cyclical algorithm of Section 3. Consider these terms.

Maximum likelihood estimation of ordered multinomial parameters

Theorem 2 of Appendix A establishes that the algorithm converges to a constrained MLE—the constraint being that \( q_k \) is small. A numerical one-dimensional Newton algorithm in place of \( \Delta_1 \) is small. A numerical one-dimensional Newton algorithm in place of \( \Delta_1 \) provides a further alternative.

Finally, if \( z_{k-1} > 0 \) and \( z_i = 0 \) for \( k - r \leq i \leq k \) with \( r \geq 1 \), the situation is similar. With starting value \( q^{(0)} \) the algorithm converges to an estimator that has \( \hat{q}_i = q_k^{(0)} \), \( k - r \leq i \leq k \), fixed. These values in fact are fixed throughout the iteration. The overall MLE can again be found by exploration in one dimension.

4. The Univariate Isotonic Problem

In order to fully implement the algorithm described in Section 3, we now need to develop an algorithm for the univariate maximizations needed in Step 2 of the cyclical algorithm of Section 3. Consider maximization of the function of \( p \) given by

\[
\phi(p) = \sum_{i=1}^{k} \left[ x_i \log p_i + z_i \log (1 - p_i - q_i) \right] = \sum_{i=1}^{k} \left[ x_i \log p_i + z_i \log (c_i - p_i) \right],
\]

subject to the constraints \( 0 \leq p_i \leq c_i \), where \( c_1 = 1 - q_1, \ldots, c_k = 1 - q_k \) are constants satisfying \( 1 \geq c_1 \geq c_2 \geq \cdots \geq c_k > 0 \), and the isotonic condition \( p_1 \leq p_2 \leq \cdots \leq p_k \). Note that the combination of these two sets of inequalities implies that \( p_j \leq c_k \) for \( 1 \leq j \leq k \). An identical situation arises when maximizing \( \ell \) in \( q \), holding \( p \) fixed.

As in the overall maximization, we must be careful at the boundary of the parameter space. First, we assume that \( x_j + z_j > 0 \); this corresponds to \( y_i < n_i \) as discussed in the second remark in Section 2. If \( x_j + z_j = 0 \), then the corresponding \( p_j \) does not appear in \( \phi \) and so cannot be identified, in general. We reiterate the convention of Section 2: in such circumstances we take \( \hat{p}_j \) to equal \( \hat{p}_{j-1} \) (or 0 if \( i = 1 \)). Second, we restrict attention to the case where \( x_1 > 0 \); for, if \( x_1 = x_2 = \cdots = x_j = 0 \), then \( \hat{p}_1 = \cdots = \hat{p}_j = 0 \) maximizes the first \( j \) terms in (6) without additional constraints on the remaining \( k - j \) terms.

A similar, but more complicated, situation occurs when \( z_k = 0 \). If \( z_k > 0 \), then \( \hat{p}_k < c_k \) and all other estimates \( \hat{p}_j < c_k \) for \( j = 1, \ldots, k - 1 \). This situation is at the core of our analysis in Theorem 2 below. If \( z_k = 0 \), then clearly \( \hat{p}_k = c_k \). But, because of the boundary \( c_k \), we cannot simply ignore this \( k \)th term and proceed with maximization over \( p_1, \ldots, p_{k-1} \) since, even if \( z_{k-1} > 0 \), it is possible that \( \hat{p}_{k-1} = c_k \), and so on. When \( z_k = 0 \), we first peel off the ‘upper’ estimates for which \( \hat{p}_{n+1}, \ldots, \hat{p}_k = c_k \).
for some \( u \leq k - 1 \). Having done this, we can proceed with maximization over \( p_1, \ldots, p_r \) ‘away’ from the boundary constraint.

Let \( S(p) = (S_1(p), \ldots, S_k(p)) \) where

\[
S_i(p) = \frac{\partial}{\partial p_i} \phi(p) = \frac{x_i}{p_i} - \frac{z_i}{c_i - p_i}, \quad i = 1, \ldots, k.
\]

The following theorem, proved in Appendix B, characterizes the solution to maximization of \( \phi(p) \) in terms of these score functions:

**Theorem 2** Let \( \phi(p) \) be given by (4) with \( x_1 > 0 \) and \( z_k > 0 \). There is a unique value of \( p \) that maximizes \( \phi \) subject to the constraints \( 0 \leq p_i \leq c_i \) and \( p_1 \leq p_2 \leq \cdots \leq p_k \). The value \( p \) is determined by the properties

\[
\sum_{i=1}^{k} p_i S_i(p) = 0
\]

and

\[
\sum_{i \neq j} S_i(p) \leq 0, \quad \text{for} \ 1 \leq j \leq k.
\]

Before we return to the boundary problem, we describe a modified PAV algorithm for implementation of the maximization in Theorem 2. For \( s \leq t \in \{1, 2, \ldots, k\} \), let \( p^{(s,t)} \) maximize the log likelihood \( \phi^{(s,t)} = \sum_{i=s}^{t} [x_i \log p + z_i \log(c_i - p)] \). Let \( p_{\text{min}}^{(s,t)} = \min_{t \geq i} p^{(s,i)} \) for \( s \leq i \) and define

\[
\{ i : p_{\text{min}}^{(i,i)} = \max_{t \geq i} p^{(t,i)} \} \equiv \{ k_1^{*}, \ldots, k_r^{*} \},
\]

say, where \( k_1^{*} < \cdots < k_r^{*} \). Note that \( k_1^{*} = 1 \), and let \( k_{r+1}^{*} = k + 1 \). The specific indices \( k_1^{*}, \ldots, k_r^{*} \) define blocks of indices \( i \) for which the maximum likelihood estimator \( \hat{p}_i \) is constant. In particular, the value of \( p \) that maximizes \( \phi \), as characterized in Theorem 2, is \( \hat{p} = (\hat{p}_1, \ldots, \hat{p}_k)^T \) where

\[
\hat{p}_i = \hat{\theta}_j, \quad k_j^{*} \leq i < k_{j+1}^{*}, \quad j = 1, \ldots, r,
\]

and \( \hat{\theta}_j \) maximizes \( \phi^{(k_j^{*}, k_{j+1}^{*} - 1)} \), \( 1 \leq i \leq k; \ 1 \leq j \leq r \).

In the simple binomial isotonic situation, the analogues of the estimated values \( (\theta_1, \ldots, \theta_r) \) are simply grouped proportions over the relevant blocks. Here, it is still straightforward to obtain the estimates but now \( \theta_j \) is the solution to a polynomial equation of order \( t \) if the block corresponding to \( \theta_j \) contains \( t \) indices.

We now put the boundary estimate and the MLE in Theorem 2 together by defining \( u = \max \{ j : \hat{p}_j < c_k \} \). Then, \( \hat{p}_i = c_k \) for \( i = u + 1, \ldots, k \) and, for \( i \leq u \), \( \hat{p}_i \) is determined via Theorem 2 (note that, necessarily \( z_u > 0 \)). It is, in fact, not necessary to consecutively search for the value of \( u \) and then the estimates \( \hat{p}_i \) for \( i \leq u \). We can identify the entire maximum likelihood estimate using the modified PAV algorithm implemented, for example, by passing through the data from right to left. At each stage in the algorithm, there is a group of blocks denoted \( a_0, a_1, \ldots, a_r \), each block comprising sequential indices. Denote the entries of an arbitrary set of blocks by \( a_j = \{ i : k_j \leq i < k_{j+1} \}, \quad j = 0, \ldots, r \) where \( 1 \leq k_0 < k_1 < \cdots < k_{r+1} = k \). For each \( i \in a_j \), \( p_i \) is estimated by the common value \( \hat{\theta}_i \in [0, c_k] \) which maximizes

\[
\phi_{k_j, k_{j+1} - 1} = \sum_{i \in a_j} [x_i \log p + z_i \log(c_i - p)].
\]
With this background, we define the following algorithm.

**The Modified PAV Algorithm**

P1. Define initial blocks $\alpha_0 = \{k - 1\}, \alpha_1 = \{k\}$ with corresponding estimates $\theta_0, \theta_1$. Note that $k_0 = k - 1, k_1 = k, s = 1$.

P2. For current blocks $\alpha_0, \alpha_1, \ldots, \alpha_s$, is $\theta_0 \leq \theta_1$? If so, then go to P4. If not, then go to P3.

P3. There is a violation. Pool $\alpha_0$ and $\alpha_1$, label as $\alpha_0$ and determine the new $\theta_0$. Relabel $\alpha_2, \ldots, \alpha_s$ as $\alpha_1, \ldots, \alpha_{s-1}$. Set $s = s - 1$. If $s = 1$ go to P4; if $s > 1$, go to P2.

P4. There is no violation. Is $k_0 = 1$? If so, then END. If not, then relabel $\alpha_0, \ldots, \alpha_s$ as $\alpha_1, \ldots, \alpha_{s+1}$, set $\alpha_0 = \{k_0 - 1\}$ and go to P2.

At the conclusion of the algorithm, the blocks $\alpha_0, \ldots, \alpha_s$ and probabilities $\theta_0, \ldots, \theta_s$ define the isotonic maximum likelihood estimate.

5. Simple examples

In the next section, we apply the algorithm to a set of data on the onset of menopause. First, however, we consider two simple examples that serve to illustrate some properties of the algorithm and the estimates. Let $m = 3$ and $k = 2$ and use the notation of Section 3. In the first example, the data are $(X_1, Y_1, Z_1) = (1, 0, 1)$ and $(X_2, Y_2, Z_2) = (0, 2, 0)$, whereas in the second $(X_1, Y_1, Z_1) = (1, 0, 1)$ and $(X_2, Y_2, Z_2) = (0, 1, 1)$.

In the first case, $p_2$ does not appear in the likelihood and, since $z_2 = 0$, the algorithm leaves the initial values $(p_2, q_2)$ (with $p_2 = 1 - q_2$) unchanged. For the moment take $q_2$ to be fixed; later, to find the MLE, we carry out a one-dimensional search over such fixed starting values for $q_2$ as discussed at the end of Section 3. With $q_2$ fixed, we start the iteration at any value $p_1(0)$, then we see from $\phi(q)$, defined as for $\phi(p)$ in (4), that $q_1(0) = 0$ and we proceed to find $p_1(1)$. If $q_2 > 0.5$, then $p_1(1) = 1 - q_2$, and, if $q_2 < 0.5$, $p_1(1) = 0.5$. In either case, the algorithm converges in one step. The maximum log likelihood can be computed for all $q_2$, and it immediately follows that the MLE is $(\hat{p}_1, \hat{q}_1) = (0.25, 0)$ and $(\hat{p}_2, \hat{q}_2) = (0.25, 0.75)$. In this case, note that $\hat{p}_2$ is uniquely identified by the inequalities imposed by $\hat{p}_1$ and $\hat{q}_2$.

In the second case, $z_2 > 0$, and so the algorithm determines $\hat{q}_2$ directly. Suppose we begin with $(p_1(0), p_2(0)) = (0.25, 0.25)$, the naive estimator of Section 1. We find that $(q_1(0), q_2(0)) = (0, 0.5)$ and $(p_1(1), p_2(1)) = (p, p)$ where $p \leq 0.5$ is a root of $6p^2 - 6p + 1 = 0$. Thus $p_1(1) = p_2(1) \approx 0.21$. The algorithm has converged and the MLE is thus given by $(\hat{p}_1, \hat{q}_1) = (0.21, 0)$ and $(\hat{p}_2, \hat{q}_2) = (0.21, 0.5)$.

6. Example—Current status data on competing risks

Consider estimation of the sub-distribution functions $F_1$ and $F_2$, defined at the beginning of Section 2 by $F_j(t) = P r(T < t, J = j)$ for $j = 1, 2$, where the random variables $T$ and $J$ measure age at onset of menopause, and cause of onset (operative ($J = 1$) and natural ($J = 2$), respectively. Note that $m = 3, k = 26$ and, in the notation of Sections 3 and 4, $p_i = F_1(C_i)$ and $q_i = F_2(C_i)$ where $(C_1, \ldots, C_{26}) = (27.5, 32.5, 35.5, 36.5, \ldots, 58.5)$.

Note that $\max(x_i, y_i) < n$ for all $i$, so that all $p_i$ and $q_i$ appear in the likelihood. On the other hand, $z_{26} = 0$ so that $q_{26} = 1 - p_{26}$ remains fixed throughout the algorithm. Subsequently, $l_{26}$ is obtained by a search. Figure 1 gives a plot of the constrained maximized likelihood as a function of $q_{26}$, and we find $\hat{q}_{26} = 0.690$ and hence $\hat{p}_{26} = 0.310$. A likelihood interval estimate could be obtained from this as,
Fig. 1. Maximized log likelihood (standardized to have maximum value 0) as a function of $q_{26}$ in the current menopausal status data in Table 1. Note that $\hat{q}_{26} = 0.690$.

For example, $\{q_{26} : l_{\text{max}}(q_{26}) - l_{\text{max}}(\hat{q}_{26}) > -1.92\} = [0.618, 0.742]$. With usual regularity conditions, this would be an approximate 95% confidence interval, though appropriate asymptotic results for this approach remain to be established.

The MLEs $\hat{F}_1(C_i) = \hat{p}_i$ and $\hat{F}_2(C_i) = \hat{q}_i$ are given in Figure 2. As before, we adopt the natural convention that $\hat{F}_j(t) = \hat{F}_j(C_{i-1})$ for $t \in (C_{j-1}, C_j)$, $j = 1, \ldots, k$, where $C_0 = 0$.

Krailo and Pike (1983) carry out a parametric analysis of the data in Table 1. Note that the overall survival function $S(t) = \Pr(T \geq t) = 1 - F_1(t) - F_2(t)$. Alternatively, $F_1(t) = \int_0^t \lambda_1(u)S(u)\,du$ where $\lambda_1$ is the cause-specific hazard function associated with operative menopause, defined by $\lambda_1(t) = \lim_{h \to \infty} h^{-1} \Pr[t \leq T < t + h, J = 1 | T \geq t]$. A similar expression exists relating $F_2$ to $\lambda_2$, the cause-specific hazard for natural menopause (Kalbfleisch and Prentice, 2002, Chapter 8.2.2). Consideration of the data (for example, by examining the nonparametric estimates given in Figure 2) suggested the use of cause-specific hazards of (i) a piecewise linear form, $\lambda_1(t) = c(t - 22)$ for $t \geq 22$, and zero otherwise, for operative menopause, and (ii) a logistic form, $\lambda_2 = b \exp(a + bt) / [1 + \exp(a + bt)]$, for natural menopause. The maximum likelihood estimates of the parameters, based on the data in Table 1, are $\hat{a} = -20.7, \hat{b} = 0.414$ and $\hat{c} = 0.000841$. These estimates, in turn, provide estimates of $F_1$ and $F_2$ (given as Figure 1 in Krailo and Pike, 1983) which can be compared with the nonparametric estimates of Figure 2. This comparison shows that this chosen parametric model fits well, supplementing a formal goodness-of-fit test of observed and expected frequencies, computed in Krailo and Pike (1983). In a similar vein, the overall probability of operative menopause ($\lim_{t \to \infty} F_1(t)$) is estimated to be 0.282 in the parametric model, as against the nonparametric estimate of 0.310 as discussed above; note that the
parametric estimate is well within the nonparametric likelihood ‘confidence interval’ of (0.258, 0.382) previously computed. These analyses all lend support to this particular parametric model for the data of Table 1.

7. Simulations

In this section, we report on a small simulation study with \( m = 3 \), undertaken to investigate some aspects of the estimates reported here. We considered total sample sizes of 100 and 400 with values of \( k \) of 10, 20 and 50. Working in the competing risk context of Section 5, the distribution of failure time was taken to be exponential with rate 1 in all simulations with the two types of failures occurring with equal rates. Thus, the sub-distribution functions were \( F_j(t) = 0.5(1 - \exp(-t)) \), \( j = 1, 2 \). The observation times were taken at \( C_j = 2jk^{-1} \), \( j = 1, \ldots, k \). The estimates were examined at times 0.2, 0.4, \ldots, 2.0.

We denote by \( \hat{F}_j(t) \), \( j = 1, 2 \) the MLEs of the sub-distribution functions and let \( \hat{F}(t) = \hat{F}_1(t) + \hat{F}_2(t) \). We denote by \( \tilde{F}_1^{(N)}(N) \), \( \tilde{F}_2^{(N)}(N) \) and \( \tilde{F}^{(N)}(N) \) the naive estimators obtained by applying the simple PAV algorithm to the type 1, type 2 and combined failures respectively. Table 2 compares the mean squared errors, based on 10 000 simulations and \( n = 100 \) for the estimators \( \hat{F}(t) \) and \( \tilde{F}(N) \) for the overall cumulative distribution function. Not surprisingly, we see almost no difference between the two estimation procedures. It should be noted that, if only combined data were available, we could define data with \( m = 3 \) simply by assigning each failure to type 1 or type 2 with probability 0.5 independently. It is intuitively clear that doing this and using the MLE based on \( m = 3 \) should not improve estimation of \( F(t) \).
Table 2. Estimated mean squared errors of ‘naive’ MLE (\(\hat{F}^{(N)}(t)\)) and MLE (\(\hat{F}(t)\)) for the CDF \(F(t)\) of survival time: \(n = \) total sample size; observation times are \(C_j = 2jk^{-1}\), \(j = 1, \ldots, k\) with \(n/k\) observations at each time; \(m = 3\) and \(F_1(t) = F_2(t) = 0.5[1 - \exp(-t)]\). Estimates are based on 10 000 simulations

<table>
<thead>
<tr>
<th>n/k</th>
<th>(t = 0.2)</th>
<th>(t = 0.4)</th>
<th>(t = 0.6)</th>
<th>(t = 0.8)</th>
<th>(t = 1.0)</th>
<th>(t = 1.2)</th>
<th>(t = 1.4)</th>
<th>(t = 1.6)</th>
<th>(t = 1.8)</th>
<th>(t = 2.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(F^{(N)})</td>
<td>(\hat{F})</td>
<td>(\hat{F})</td>
<td>(\hat{F})</td>
<td>(\hat{F})</td>
<td>(\hat{F})</td>
<td>(\hat{F})</td>
<td>(\hat{F})</td>
<td>(\hat{F})</td>
<td>(\hat{F})</td>
</tr>
<tr>
<td>100/10</td>
<td>0.0115</td>
<td>0.0104</td>
<td>0.0140</td>
<td>0.0161</td>
<td>0.0149</td>
<td>0.0092</td>
<td>0.0085</td>
<td>0.0073</td>
<td>0.0063</td>
<td>0.0059</td>
</tr>
<tr>
<td>100/20</td>
<td>0.0135</td>
<td>0.0115</td>
<td>0.0138</td>
<td>0.0139</td>
<td>0.0155</td>
<td>0.0110</td>
<td>0.0090</td>
<td>0.0076</td>
<td>0.0064</td>
<td>0.0059</td>
</tr>
<tr>
<td>100/50</td>
<td>0.0115</td>
<td>0.0115</td>
<td>0.0119</td>
<td>0.0123</td>
<td>0.0130</td>
<td>0.0102</td>
<td>0.0083</td>
<td>0.0068</td>
<td>0.0061</td>
<td>0.0057</td>
</tr>
</tbody>
</table>

Table 3. Estimated mean squared errors of ‘naive’ MLE (\(\hat{F}^{(N)}(t)\)) and MLE (\(\hat{F}(t)\)) for sub distribution function \(F_1(t)\): \(n = \) total sample size; observation times are \(C_j = 2jk^{-1}\), \(j = 1, \ldots, k\) with \(n/k\) observations at each time; \(m = 3\) and \(F_1(t) = F_2(t) = 0.5[1 - \exp(-t)]\). Estimates are based on 10 000 simulations

<table>
<thead>
<tr>
<th>n/k</th>
<th>(t = 0.2)</th>
<th>(t = 0.4)</th>
<th>(t = 0.6)</th>
<th>(t = 0.8)</th>
<th>(t = 1.0)</th>
<th>(t = 1.2)</th>
<th>(t = 1.4)</th>
<th>(t = 1.6)</th>
<th>(t = 1.8)</th>
<th>(t = 2.0)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(\hat{F}^{(N)})</td>
<td>(\hat{F}_1)</td>
<td>(\hat{F}_1)</td>
<td>(\hat{F}_1)</td>
<td>(\hat{F}_1)</td>
<td>(\hat{F}_1)</td>
<td>(\hat{F}_1)</td>
<td>(\hat{F}_1)</td>
<td>(\hat{F}_1)</td>
<td>(\hat{F}_1)</td>
</tr>
<tr>
<td>100/10</td>
<td>0.0052</td>
<td>0.0053</td>
<td>0.0066</td>
<td>0.0062</td>
<td>0.0064</td>
<td>0.0058</td>
<td>0.0053</td>
<td>0.0053</td>
<td>0.0056</td>
<td>0.0057</td>
</tr>
<tr>
<td>100/20</td>
<td>0.0062</td>
<td>0.0064</td>
<td>0.0075</td>
<td>0.0068</td>
<td>0.0067</td>
<td>0.0064</td>
<td>0.0063</td>
<td>0.0069</td>
<td>0.0099</td>
<td>0.0326</td>
</tr>
<tr>
<td>100/50</td>
<td>0.0062</td>
<td>0.0064</td>
<td>0.0081</td>
<td>0.0075</td>
<td>0.0075</td>
<td>0.0068</td>
<td>0.0064</td>
<td>0.0063</td>
<td>0.0710</td>
<td>0.0753</td>
</tr>
<tr>
<td>400/10</td>
<td>0.0017</td>
<td>0.0018</td>
<td>0.0024</td>
<td>0.0024</td>
<td>0.0024</td>
<td>0.0022</td>
<td>0.0021</td>
<td>0.0020</td>
<td>0.0020</td>
<td>0.0023</td>
</tr>
<tr>
<td>400/20</td>
<td>0.0021</td>
<td>0.0021</td>
<td>0.0025</td>
<td>0.0024</td>
<td>0.0024</td>
<td>0.0023</td>
<td>0.0021</td>
<td>0.0021</td>
<td>0.0027</td>
<td>0.0081</td>
</tr>
<tr>
<td>400/50</td>
<td>0.0023</td>
<td>0.0023</td>
<td>0.0025</td>
<td>0.0025</td>
<td>0.0026</td>
<td>0.0025</td>
<td>0.0023</td>
<td>0.0022</td>
<td>0.0020</td>
<td>0.0020</td>
</tr>
</tbody>
</table>

Table 3 gives a similar comparison of \(\hat{F}_1(t)\) with \(\hat{F}^{(N)}(t)\). Here differences are observed that become more substantial for larger values of \(t\). Especially, toward the end of the observation period, the MLE does considerably better than the naive estimate. This is due in large part to some bias in the naive estimator of \(F_1(t)\) for values of \(t\) near the end of the observation period. The estimated means for the two estimation procedures are summarized in Table 4. It is clear that the bias decreases with increasing sample size as it should. The MLE has much better properties than the naive estimator for larger values of \(t\), at least for moderate sample sizes.

One possibility for inference is to use a bootstrap procedure. Let \(J_i, C_i\) be the data for the \(i\)th individual where \(J_i = 0\) if the \(i\)th failure time is censored and \(J_i = j\) if a failure of type \(j\) occurs before time \(C_i\). In the simplest implementation, we could consider a bootstrap sample \((J^*_i, C^*_i)\), \(i = 1, \ldots, n\) obtained by iid sampling of the observed data. The bootstrap estimates are \(\hat{F}_1^*, \hat{F}_2^*\), and \(\hat{F}^* = \hat{F}_1^* + \hat{F}_2^*\). Following
Table 4. Estimated means of ‘naive’ MLE ($\hat{F}_1^{(N)}(t)$) and MLE ($\hat{F}_1(t)$) for sub-distribution function $F_1(t)$: $n =$ total sample size; observation times are $C_j = 2^j k^{-1}$, $j = 1, \ldots, k$ with $n/k$ observations at each time; $m = 3$ and $F_1(t) = F_2(t) = .5[1 - \exp(-t)]$. Estimates are based on 10,000 simulations

<table>
<thead>
<tr>
<th>Observation Times</th>
<th>t = 0.2</th>
<th>t = 0.4</th>
<th>t = 0.6</th>
<th>t = 0.8</th>
<th>t = 1.0</th>
<th>t = 1.2</th>
<th>t = 1.4</th>
<th>t = 1.6</th>
<th>t = 1.8</th>
<th>t = 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>100/10 $\hat{F}_1^{(N)}$</td>
<td>0.0751</td>
<td>0.1499</td>
<td>0.2110</td>
<td>0.2609</td>
<td>0.3029</td>
<td>0.3381</td>
<td>0.3700</td>
<td>0.4006</td>
<td>0.4366</td>
<td>0.4932</td>
</tr>
<tr>
<td>$\hat{F}_1$</td>
<td>0.0764</td>
<td>0.1531</td>
<td>0.2158</td>
<td>0.2676</td>
<td>0.3102</td>
<td>0.3448</td>
<td>0.3746</td>
<td>0.4006</td>
<td>0.4261</td>
<td>0.4543</td>
</tr>
<tr>
<td>100/20 $\hat{F}_1^{(N)}$</td>
<td>0.0638</td>
<td>0.1439</td>
<td>0.2083</td>
<td>0.2597</td>
<td>0.3035</td>
<td>0.3408</td>
<td>0.3747</td>
<td>0.4069</td>
<td>0.4483</td>
<td>0.5413</td>
</tr>
<tr>
<td>$\hat{F}_1$</td>
<td>0.0656</td>
<td>0.1484</td>
<td>0.2164</td>
<td>0.2704</td>
<td>0.3161</td>
<td>0.3537</td>
<td>0.3868</td>
<td>0.4162</td>
<td>0.4484</td>
<td>0.4922</td>
</tr>
<tr>
<td>100/50 $\hat{F}_1^{(N)}$</td>
<td>0.0520</td>
<td>0.1352</td>
<td>0.2032</td>
<td>0.2570</td>
<td>0.3000</td>
<td>0.3371</td>
<td>0.3702</td>
<td>0.4044</td>
<td>0.4497</td>
<td>0.6165</td>
</tr>
<tr>
<td>$\hat{F}_1$</td>
<td>0.0542</td>
<td>0.1403</td>
<td>0.2101</td>
<td>0.2652</td>
<td>0.3077</td>
<td>0.3424</td>
<td>0.3716</td>
<td>0.3983</td>
<td>0.4264</td>
<td>0.4780</td>
</tr>
<tr>
<td>400/10 $\hat{F}_1^{(N)}$</td>
<td>0.0877</td>
<td>0.1610</td>
<td>0.2207</td>
<td>0.2700</td>
<td>0.3098</td>
<td>0.3436</td>
<td>0.3725</td>
<td>0.3973</td>
<td>0.4224</td>
<td>0.4582</td>
</tr>
<tr>
<td>$\hat{F}_1$</td>
<td>0.0880</td>
<td>0.1620</td>
<td>0.2222</td>
<td>0.2723</td>
<td>0.3125</td>
<td>0.3464</td>
<td>0.3744</td>
<td>0.3976</td>
<td>0.4174</td>
<td>0.4396</td>
</tr>
<tr>
<td>400/20 $\hat{F}_1^{(N)}$</td>
<td>0.0829</td>
<td>0.1588</td>
<td>0.2190</td>
<td>0.2693</td>
<td>0.3102</td>
<td>0.3442</td>
<td>0.3727</td>
<td>0.3995</td>
<td>0.4283</td>
<td>0.4827</td>
</tr>
<tr>
<td>$\hat{F}_1$</td>
<td>0.0836</td>
<td>0.1604</td>
<td>0.2212</td>
<td>0.2720</td>
<td>0.3136</td>
<td>0.3474</td>
<td>0.3749</td>
<td>0.3991</td>
<td>0.4208</td>
<td>0.4492</td>
</tr>
<tr>
<td>400/50 $\hat{F}_1^{(N)}$</td>
<td>0.0802</td>
<td>0.1561</td>
<td>0.2179</td>
<td>0.2682</td>
<td>0.3098</td>
<td>0.3434</td>
<td>0.3723</td>
<td>0.3986</td>
<td>0.4290</td>
<td>0.5212</td>
</tr>
<tr>
<td>$\hat{F}_1$</td>
<td>0.0811</td>
<td>0.1581</td>
<td>0.2207</td>
<td>0.2714</td>
<td>0.3136</td>
<td>0.3469</td>
<td>0.3748</td>
<td>0.3978</td>
<td>0.4195</td>
<td>0.4582</td>
</tr>
<tr>
<td>true $F_1(t)$</td>
<td>0.0906</td>
<td>0.1648</td>
<td>0.2256</td>
<td>0.2753</td>
<td>0.3161</td>
<td>0.3494</td>
<td>0.3767</td>
<td>0.3991</td>
<td>0.4174</td>
<td>0.4323</td>
</tr>
</tbody>
</table>

Efron’s percentile method, a 100(1 − α)% confidence interval for $F_1(t)$ at a fixed value of $t$ is obtained as the interval spanned by the upper and lower α/2 quantiles of the bootstrap sample $\hat{F}_1(t)$. Exploration of the properties of such a procedure could be useful, but would most naturally be done in the context of estimation with standard current status data.

In order to gain some preliminary information on the potential value of such a procedure, we carried out a small and very preliminary simulation in the context of the present paper. Specifically, we again supposed that the time to failure had a unit exponential distribution and that failures at any time $t$ were equally likely to be of types 1 or 2. The observation times were selected to be an iid sample from the $k$ point discrete uniform distribution on $2/k, 4/k, \ldots, 2$. To simplify the endpoint issue discussed earlier, for the purpose of these simulations, we added 0.5 to the frequency of survivors at the final observation time, 2, in both the original sample and in all bootstrap repetitions. We report some results in Tables 5 and 6 for the cases $n = 400, k = 20$ and $n = 1600, k = 80$. The calculations are based on 1000 replications, with 1000 (for $n = 400$) and 200 (for $n = 1600$) bootstrap samples selected for each data point. The results of the simulation are relatively encouraging with coverage probabilities of 80, 90 and 95% intervals for $F_1(t)$ and $F(t)$ being fairly accurate at least for values of $t \leq 1.6$. The results for $t = 2$ are very poor, especially for estimation of $F(t)$, but this may be due to bias induced by our convention of adding 0.5 to $z_k$ in each sample. Also given in the tables are the average lengths of the confidence intervals, and it can be seen that the ratio of the average length for $n = 400$ to $n = 1600$ is reasonably well approximated by $n^{1/3}$, consistent with the cube root asymptotics that apply in the case of current status data when the distribution of the censoring and the failure time are both continuous at time $t$. The applicability of the asymptotics could be examined further through simulations of this sort, again in the context of current status data.
For the one-dimensional isotonic maximization of components of $X_j$, careful with the choice of starting values and the fact that the algorithm is usually very slow to converge.

Kruskal (1964) and Wu (1982) to obtain some savings on the right-to-left PAV algorithm we have used. Averaging as compared to the polynomial root solving required in our algorithm. The disadvantage is that

There are alternative algorithms that can be used to compute the MLE. The EM algorithm imputes `complete’ data based on constructing hypothetical sub-categorization of the $X_j$, that take advantage of the isotonicity of the components of the $p_j$. This is most easily visualized in the context of current status competing risks data where the frequency $X_j$ is distributed across intervals defined by the observation times $C_1, \ldots, C_j$ in the E step of the algorithm. Disadvantages to this approach include the need to be careful with the choice of starting values and the fact that the algorithm is usually very slow to converge. For the one-dimensional isotonic maximization of components of $p$ discussed in Section 3, we could use a modified weighted PAV algorithm, based on Jongbloed (1995), for which the pooling involves simple averaging as compared to the polynomial root solving required in our algorithm. The disadvantage is that iterative weights need to be computed at each cycle. We suspect that the algorithms share similar speed of convergence properties. It may also be possible to generalize the up- and down-blocks algorithm of Kruskal (1964) and Wu (1982) to obtain some savings on the right-to-left PAV algorithm we have used.

### Table 5. Coverage probabilities of bootstrap confidence intervals for $F_1$ and $F$ with $n = 400$ and $k = 20$. Calculations are based on 1000 replications and 1000 bootstrap samples

<table>
<thead>
<tr>
<th>$t$</th>
<th>Estimation of $F_1(t)$</th>
<th>Estimation of $F(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>80% 90% 95%</td>
<td>80% 90% 95%</td>
</tr>
<tr>
<td>0.4</td>
<td>cov. prob. 0.806 0.898 0.944</td>
<td>0.797 0.905 0.945</td>
</tr>
<tr>
<td></td>
<td>ave. length 0.119 0.152 0.180</td>
<td>0.161 0.206 0.245</td>
</tr>
<tr>
<td>0.8</td>
<td>cov. prob. 0.850 0.923 0.963</td>
<td>0.828 0.935 0.972</td>
</tr>
<tr>
<td></td>
<td>ave. length 0.118 0.151 0.178</td>
<td>0.150 0.191 0.227</td>
</tr>
<tr>
<td>1.2</td>
<td>cov. prob. 0.853 0.924 0.964</td>
<td>0.820 0.925 0.968</td>
</tr>
<tr>
<td></td>
<td>ave. length 0.108 0.138 0.164</td>
<td>0.127 0.163 0.193</td>
</tr>
<tr>
<td>1.6</td>
<td>cov. prob. 0.825 0.931 0.970</td>
<td>0.831 0.919 0.964</td>
</tr>
<tr>
<td></td>
<td>ave. length 0.105 0.135 0.161</td>
<td>0.109 0.139 0.166</td>
</tr>
<tr>
<td>2.0</td>
<td>cov. prob. 0.803 0.900 0.947</td>
<td>0.626 0.754 0.825</td>
</tr>
<tr>
<td></td>
<td>ave. length 0.132 0.169 0.200</td>
<td>0.107 0.133 0.153</td>
</tr>
</tbody>
</table>

### Table 6. Coverage probabilities of bootstrap confidence intervals for $F_1$ and $F$ with $n = 1600$ and $k = 80$. Calculations are based on 1000 replications and 200 bootstrap samples

<table>
<thead>
<tr>
<th>$t$</th>
<th>Estimation of $F_1(t)$</th>
<th>Estimation of $F(t)$</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>80% 90% 95%</td>
<td>80% 90% 95%</td>
</tr>
<tr>
<td>0.4</td>
<td>cov. prob. 0.832 0.920 0.960</td>
<td>0.849 0.930 0.967</td>
</tr>
<tr>
<td></td>
<td>ave. length 0.076 0.098 0.116</td>
<td>0.103 0.132 0.158</td>
</tr>
<tr>
<td>0.8</td>
<td>cov. prob. 0.861 0.932 0.974</td>
<td>0.843 0.939 0.974</td>
</tr>
<tr>
<td></td>
<td>ave. length 0.076 0.097 0.115</td>
<td>0.096 0.123 0.147</td>
</tr>
<tr>
<td>1.2</td>
<td>cov. prob. 0.839 0.928 0.970</td>
<td>0.860 0.937 0.965</td>
</tr>
<tr>
<td></td>
<td>ave. length 0.068 0.087 0.104</td>
<td>0.080 0.103 0.123</td>
</tr>
<tr>
<td>1.6</td>
<td>cov. prob. 0.854 0.941 0.976</td>
<td>0.855 0.948 0.974</td>
</tr>
<tr>
<td></td>
<td>ave. length 0.062 0.080 0.096</td>
<td>0.067 0.086 0.103</td>
</tr>
<tr>
<td>2.0</td>
<td>cov. prob. 0.758 0.865 0.927</td>
<td>0.399 0.560 0.677</td>
</tr>
<tr>
<td></td>
<td>ave. length 0.099 0.127 0.152</td>
<td>0.088 0.109 0.126</td>
</tr>
</tbody>
</table>

8. DISCUSSION

There are alternative algorithms that can be used to compute the MLE. The EM algorithm imputes 'complete' data based on constructing hypothetical sub-categorization of the $X_j$, that take advantage of the isotonicity of the components of the $p_j$. This is most easily visualized in the context of current status competing risks data where the frequency $X_j$ is distributed across intervals defined by the observation times $C_1, \ldots, C_j$ in the E step of the algorithm. Disadvantages to this approach include the need to be careful with the choice of starting values and the fact that the algorithm is usually very slow to converge. For the one-dimensional isotonic maximization of components of $p$ discussed in Section 3, we could use a modified weighted PAV algorithm, based on Jongbloed (1995), for which the pooling involves simple averaging as compared to the polynomial root solving required in our algorithm. The disadvantage is that iterative weights need to be computed at each cycle. We suspect that the algorithms share similar speed of convergence properties. It may also be possible to generalize the up- and down-blocks algorithm of Kruskal (1964) and Wu (1982) to obtain some savings on the right-to-left PAV algorithm we have used.
Maximum likelihood estimation of ordered multinomial parameters

Other estimation criteria could also be invoked. Multivariate weighted least squares (Sasabuchi et al., 1983) is a possible alternative. In one dimension, maximum likelihood is equivalent to ordinary least squares as noted in the Introduction. In higher dimensions, the relationship between maximum likelihood and weighted least squares remains unclear.

We conclude with some brief remarks regarding inference procedures associated with the NPMLE estimates of the multinomial probabilities of Section 1, and the sub-distribution functions of Sections 2, 6 and 7. In the simple current status problem \((m = 2)\) with observation times arising from a distribution \(G\) that is absolutely continuous with respect to \(F\), it is well known that the NPMLE converges to \(F\) only at rate \(n^{-1/3}\), with a pointwise limiting distribution that is not Gaussian (and depends on \(F\) and \(G\)). Thus, standard errors are not immediately relevant even to asymptotic inference. In addition, without uniform convergence, it is still unknown whether bootstrap confidence intervals are even asymptotically correct. Under the same conditions, estimation of smooth functionals of the underlying distribution functions can be estimated at the standard \(n^{-1/2}\) rate (Groeneboom and Wellner, 1980). Asymptotic theory is thus quite delicate even in this simpler situation that has been studied for a considerable length of time. We anticipate that similar asymptotic results will apply in the current competing risk scenario, although much of this remains to be proved. Some results on smooth functionals are given in Jewell (2003). For the estimates \(F_1(t_0)\) and \(F_2(t_0)\) at a given \(t_0\), we note some possible practical approaches that at least provide some measure of variability.

First, note that we can compute the constrained MLE subject to, for example, \(p_{ij} = d\), for some \(i\) and \(j\) using essentially the same algorithm. This can be achieved by replacing \(x_i\) and \(z_j\) in (4) with large values, \(x_i^*\) and \(z_j^*\) so that \(S^*(d) = 0\) and running the algorithm as described. This allows the computation of likelihood ratios that compare the unconstrained and constrained maximum likelihoods, to yield a confidence interval. The coverage probability of such a likelihood ratio interval is unclear. If the distribution of observation times is absolutely continuous with respect to sub-distributions, however, the results of Banerjee and Wellner (2001, 2004a,b) suggest that the limiting distribution of the likelihood ratio statistic will not be chi-squared, but will not depend on \(F\) and \(G\). Thus, \(p_{ij} = d\) for some \(i\) and \(j\) using essentially the same algorithm. This can be achieved by replacing \(x_i\) and \(z_j\) in (4) with large values, \(x_i^*\) and \(z_j^*\) so that \(S^*(d) = 0\) and running the algorithm as described. This allows the computation of likelihood ratios that compare the unconstrained and constrained maximum likelihoods, to yield a confidence interval. The coverage probability of such a likelihood ratio interval is unclear. If the distribution of observation times is absolutely continuous with respect to sub-distributions, however, the results of Banerjee and Wellner (2001, 2004a,b) suggest that the limiting distribution of the likelihood ratio statistic will not be chi-squared, but will not depend on \(F_1\), \(F_2\), \(G\), or \(t_0\). These results depend on the assumed limiting form of the observational plan. In general, it is possible to embed the observed data in many different asymptotic scenarios with potentially different asymptotic results holding. For example, one plausible asymptotic view with relatively few data points assumes that observation times are fixed and lets the number of observations at each point become large. This leads to a standard likelihood with a finite number of parameters and the usual asymptotic chi-squared results. Alternatively, we could consider various other processes whereby the number of observation points approaches infinity but at a rate proportional to \(\sqrt{n}\) instead of \(n\). It is not clear in this instance what asymptotic results should apply.

Bootstrap methodology may provide the most satisfactory approach to inference. As mentioned earlier, a more detailed assessment of the asymptotics and of the bootstrap would be very useful in these problems, though as noted earlier, an evaluation in the usual case of current status data with a single failure mode would seem the most appropriate forum for initial investigation.

APPENDIX A

Proof of Theorem 1

Lemma 1 Suppose that \((p, q) \in \Theta\) and \(p_k + q_k < 1\). Then, for any \((p^*, q^*) \in \Theta\), there exists \(\epsilon > 0\) such that \(|(1 - \epsilon)p + \epsilon p^*|, q^*|, q \mid \in \Theta\).

Proof. Since \(p_k + q_k < 1\), there exists \(\epsilon > 0\) such that \((1 - \epsilon)p + \epsilon p^* + q < 1\). This is sufficient for the claim since the entries \((1 - \epsilon)p + \epsilon p^*\) and \(q\) are isotonic.

A similar result holds with \(p\) and \(q\) interchanged. \(\Box\)
THEOREM 1 If \( z_k > 0 \), the cyclical algorithm converges to \((\hat{p}, \hat{q})\), the unique MLE of \( p, q \).

**Proof.** The algorithm is monotonic in the sense that \( \ell(p^{(j+1)}, q^{(j)}) \geq \ell(p^{(j)}, q^{(j)}) \) and \( \ell(p^{(j+1)}, q^{(j+1)}) \geq \ell(p^{(j+1)}, q^{(j)}) \) for \( j = 0, 1, 2, \ldots \). Since the likelihood is bounded above, it follows that \( \ell(p^{(j)}, q^{(j)}) \) converges to \( \ell^\infty \), say. Also, the sequence \((p^{(j)}, q^{(j)})\) has a subsequence \((\bar{p}, \bar{q})\) that converges to \((\bar{p}, \bar{q})\), say, and \( \ell(\bar{p}, \bar{q}) = \ell^\infty \) by continuity.

We now want to show that \((\bar{p}, \bar{q})\) is a MLE. First, it is evident that \( \bar{q} \) maximizes the likelihood \( \ell^\dagger(\bar{p}, \bar{q}) \), since \( \bar{q} \) maximizes \( \ell^\dagger(p^{(j)}, q) \), \( p^{(j)} \to \bar{p} \), and \( q^{(j)} \to \bar{q} \), it follows that \( \bar{q} \) maximizes \( \ell^\dagger(\bar{p}, \bar{q}) \).

We now turn to the same issue, but with \( p \) exchanging roles with \( q \). Unfortunately, there is not a direct symmetry since \( \bar{q} \) maximizes \( \ell^\dagger(p, q^{(j)}) \), not \( \ell^\dagger(p, q^{(j)} \, ^\dagger) \). So, from the convergent subsequence, \((p^{(j)}, q^{(j)} \, ^\dagger)\), say, we take a further convergent subsequence, \((p^{(j)} \, ^\dagger, q^{(j)} \, ^\dagger)\), that converges to \((\hat{p}, \hat{q})\), say. It follows that \( \hat{q} \) maximizes the likelihood \( \ell^\dagger(\hat{p}, \hat{q}) \). We must now show that \( \hat{q} = \bar{q} \).

Suppose that \( \hat{q} \neq \bar{q} \). Since \( \hat{q} \) maximizes \( \ell(p, q) \), it follows that the directional derivative of \( \ell \) from \((p, q)\) to \((\hat{p}, \hat{q})\) is positive. It follows that \( \ell(\hat{p}, \hat{q}) > \ell^\infty \). This contradiction indicates that \( \hat{q} = \bar{q} \) and hence that \( \hat{p} \) maximizes the likelihood \( \ell^\dagger(p, \bar{q}) \), and \( \bar{q} \) maximizes the likelihood \( \ell^\dagger(\bar{p}, q) \).

Since \( z_k > 0 \), it follows that \( \hat{p}_k + \hat{q}_k < 1 \). Consider the directional derivative \( D_\ell[(\bar{p}, \bar{q}); (\hat{p}, \hat{q})] \) of \( \ell \) from \((\bar{p}, \bar{q})\) toward an arbitrary point \((\hat{p}, \hat{q})\) in \( \Theta \). Some calculation shows that

\[
D_\ell = \sum_{i=1}^{k} \left[ x_i \frac{p_i^* - \hat{p}_i}{\bar{p}_i} + y_i \frac{q_i^* - \hat{q}_i}{\bar{q}_i} - z_i \frac{p_i^* - \hat{p}_i + q_i^* - \hat{q}_i}{1 - \hat{p}_i - \hat{q}_i} \right].
\]

(A.1)

Since \( \hat{p} \) maximizes the likelihood \( \ell^\dagger(\hat{p}, \bar{q}) \), as a consequence of Lemma 1, the directional derivative of \( \ell^\dagger \) from \( \hat{p} \) toward \( \hat{q} \) is

\[
D^\dagger_\ell(\hat{p}; \hat{q}) = \sum_{i=1}^{k} \left[ x_i \frac{p_i^* - \hat{p}_i}{\bar{p}_i} - z_i \frac{p_i^* - \hat{p}_i}{1 - \hat{p}_i - \hat{q}_i} \right] \leq 0.
\]

(A.2)

A similar inequality holds for the likelihood in \( q \) given \( \hat{p} \). The sum of these two inequalities establishes that \( D_\ell \leq 0 \), so that \((\bar{p}, \bar{q}) = (\hat{p}, \hat{q})\) is a MLE.

Thus, \( \ell^\infty = \lim_{j \to \infty} \ell(p^{(j)}, q^{(j)}) \) is the maximum of \( \ell \) over \( \Theta \). Finally, since \( \ell \) is strictly concave, \((p^{(j)}, q^{(j)})\) converges to \((\bar{p}, \bar{q})\), the unique MLE (see Rockafellar, 1970, Corollary 27.2.2). \( \square \)

Suppose now that \( z_k = 0 \) and \( z_k - 1 > 0 \). In this case, it can be seen that \( \hat{p}_k = 1 - q_k^{(0)} \) and \( \hat{q}_k = q_k^{(0)} \). As a consequence, the algorithm does not converge to the MLE in general. Let

\[
\Theta(q^{(0)}) = \{ (p, q) \in \Theta : p_k = 1 - q_k^{(0)}, q_k = q_k^{(0)} \}
\]

and let \((\hat{p}(q^{(0)}), \hat{q}(q^{(0)}))\) be the corresponding MLE in this restricted parameter space.

**Lemma 2.** If \( z_k = 0 \), \( z_k - 1 > 0 \) and \( q^{(0)} \) is the initial estimate of \( q \), then the algorithm converges to \((\hat{p}(q^{(0)}), \hat{q}(q^{(0)}))\).

**Proof.** At each iteration, \((p^{(j)}), q^{(j)}) \in \Theta(q^{(0)}) \). The directional derivative from the limit point \((\hat{p}, \hat{q})\) to any point \((p^*, q^*) \) in \( \Theta(q^{(0)}) \) is exactly as in (A.1) except that the upper limit of the sum is \( k - 1 \). An argument identical to that used in Theorem 1 gives the required result. \( \square \)
APPENDIX B

Proof of Theorem 2

Proof. First, suppose \( p \) satisfies the two properties in the statement of Theorem 2. Since \( \phi \) is concave, \( \phi(p^*) - \phi(p) \leq (S(p^*), (p^* - p)) \) for any \( p^* \) satisfying the same constraints as \( p \). (For vectors \( x, y, \langle x, y \rangle = \sum_{i=1}^{k} x_i y_i \).) By the first property, \( (S(p), p) = 0 \). Let \( e_1 = (0, 0, \ldots, 0, c_k), e_2 = (0, 0, \ldots, 0, c_k, c_k), \ldots, e_k = (c_k, c_k, \ldots, c_k) \). With \( \alpha_1 = \frac{p_k - p_{k-1}}{c_k}, \alpha_2 = \frac{p_{k-1} - p_{k-2}}{c_k}, \ldots, \alpha_k = \frac{p_1}{c_k} \), all of which are greater than or equal to zero, we have \( p^* = \sum_{j=1}^{k} \alpha_j e_j \). Then, \( (S(p), p^*) = \sum_{i=1}^{k} \alpha_i \sum_{j \geq i} c_k S_j(p) \leq 0 \), by the second property. Thus, \( \phi(p^*) \leq \phi(p) \) for all such \( p^* \), and so \( p \) maximizes \( \phi \).

On the other hand, suppose \( p \) maximizes \( \phi \) with \( 0 \leq p_i \leq c_i \) and \( p_1 \leq p_2 \leq \cdots \leq p_k \). Since \( p_1 > 0 \) and \( p_k < c_k \), \((1 + \epsilon)p\) satisfies the same constraints as \( p \) for sufficiently small \( \epsilon \). Thus, \( \lim_{\epsilon \to 0} \frac{\phi((1+\epsilon)p) - \phi(p)}{\epsilon} = \sum_{j=1}^{k} p_j S_j(p) = 0 \). Also, for \( 0 < \epsilon < \frac{c_k - p_k}{c_k} \), \( p + \epsilon e_{k-1+1} \) satisfies the same constraints as \( p \) for \( 1 \leq i \leq k \), so that

\[
\lim_{\epsilon \to 0} \frac{\phi(p + \epsilon e_{k-1+1}) - \phi(p)}{\epsilon} = \sum_{j \geq i} c_k S_j(p) \leq 0,
\]

yielding the second property.

Finally, the uniqueness of \( p \) follows since the parameter space is convex and, subject to our convention if \( a_i + b_i = 0 \) for some values of \( i \), \( \phi \) is strictly concave.

An alternative proof can be given through direct use of the Karush–Kuhn–Tucker conditions (see, for example, Proposition 14.2.3 of Lange, 1999).

APPENDIX C

Proof that (5) Describes the MLE

We show that (5) satisfies the two conditions of Theorem 2. First, from the definition of \( \hat{\theta}_j \), \( \sum_{i=k_j}^{k_{j+1}-1} S_i(\hat{p}) = 0 \) for \( 1 \leq j \leq r \). Since \( \hat{p} \) is constant over the same blocks, it is immediate that \( \sum_{i=1}^{k_j} p_i S_i(\hat{p}) = 0 \). Now consider an arbitrary integer \( l \) such that \( k_j \leq l \leq k_{j+1} \). From the definition of \( \hat{\theta}_j \), it follows that \( p^{(k_j, k_{j+1}-1)} \leq p^{(k_j, k_{j+1}-1)} = \hat{\theta}_j \). Thus,

\[
\sum_{i=l}^{k_{j+1}-1} S_i(\hat{p}) = k_{j+1}-1 \sum_{i=l}^{k_{j+1}-1} S_i(\hat{p}) \leq 0
\]

which establishes the second condition in Theorem 2.


[Received February 18, 2003; first revision May 9, 2003; second revision October 29, 2003; accepted for publication November 3, 2003]