University of California, Berkeley
U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2001 *Paper* 100

Maximum Likelihood Estimation of Ordered Multinomial Parameters

Nicholas P. Jewell* John D. Kalbfleisch†

*Division of Biostatistics, School of Public Health, University of California, Berkeley

†Dept. of Statistics & Actuarial Science, University of Waterloo, Ontario, Canada

Maximum Likelihood Estimation of Ordered Multinomial Parameters

Nicholas P. Jewell and John D. Kalbfleisch

Abstract

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. By making use of variants of the pool adjacent violator algorithm, we obtain a simple algorithm to compute the maximum likelihood estimator 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., 2001).

1 Introduction

Suppose that $X_1, ..., X_k$ 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 $p_1, ..., p_k$. 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 violators (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, $C_1 < \ldots < C_k$, 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 $p_i = F(C_i),\ i = 1, ..., k$ as above if, at each C_i, we observe the number X_i out of n_i independent individuals who 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, 2003).

We consider a more general problem in which, for each i, $\mathbf{X}_i = (X_{1i}, ..., X_{mi})$ is an independent multinomial variable with known index n_i and probability $\mathbf{p}_i = (p_{1i}, p_{2i}, \ldots, p_{mi})$, $\sum_{j=1}^{m} p_{ji} = 1$, where the $\mathbf{p}_i$s are known to satisfy the constraints:

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

The log likelihood function is

$$\ell = \sum_{i=1}^{k} \left[X_{1i} \log(p_{1i}) + \cdots + X_{mi} \log(p_{mi}) \right], \tag{2}$$

and the proposed maximum likelihood estimator (MLE) maximizes (2) subject to the constraints (1) and $\sum_{j=1}^{m} p_{ji} = 1$ 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 $p_{j1}, ..., p_{jk}$ 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 $\mathbf{p}_i : 1 \leq i \leq k$ 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 $j = 1, ..., m-1$ 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

$$F_j(t) = \text{pr}(T \leq t, J = j),$$

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 observation time, the cause of failure is also observed. Thus, the data from a single observation is simply $(I(T \leq C), \Delta)$, where $I(\cdot)$ is the indicator function, and $\Delta = j$ if $T \leq C$ and $J = j$ $(j = 1, \ldots, m-1)$, and $\Delta = m$ if $T > C$. Assuming that monitoring times are independent of T and are uninformative, it is easy to see that the log likelihood of n independent observations of this kind, conditional on the observed Cs, is

Table 1: Data on menopausal current status from McMahon and Worcester (1966). C_i = age, n_i= # respondents, x_i=# operative menopause, y_i=# natural menopause, z_i=# non menopausal

C_i	n_i	x_i	y_i	z_i	C_i	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

-

given by (2). In particular, we assume that there are k distinct ordered monitoring times, $C_1 < \ldots < C_k$ and X_{ji} is the number of observations monitored at time C_i for whom $\Delta = j$; then equivalence of the likelihoods is obtained by setting $p_{ji} = F_j(C_i)$ for $j = 1, \ldots, m-1$ and $p_{mi} = S(C_i)$.

A particularly simple version of this data structure arises from the National Center for Health Statistics' Health Examination Survey, discussed by Krailo and Pike (1983) and originally analyzed by McMahon and Worcester (1966). These papers focus on the menopausal history of 3,581 female respondents from 1960-1962 who provided cross-sectional information on their age and their menopausal status. For those who had experienced menopause, McMahon and Worcester deemed further retrospective information on the age of onset unreliable. Thus, Krailo and Pike (1983) concentrated on data on current menopausal status along with supplementary information on whether or not menopause was the result of surgery. The time variable, T, is age and the competing risks, $J = 1, 2$, are natural menopause and operative menopause. The summarized data are in Table 1. Previous analyses examined parametric estimates of the sub-distribution functions F_j. In Section 6 we provide the nonparametric maximum likelihood estimates.

Two implicit assumptions are required to apply the general structure to this particular

example and others with similar characteristics. First, selective mortality effects are being ignored. That is, the proposed analysis is, in fact, estimating age-specific probabilities related to menopause *assuming survival to that particular age*. In particular, we need here that the risk of death at a given age is no different after menopause than before. While this may be appropriate for natural menopause, it should be treated with caution for operative menopause. Second, since the data for all different ages was collected at a single calendar time, it is assumed that the distributions of age at menopause are stationary in calendar time. From another point of view, if date of birth is a predictor of age at menopause, this covariate (along with others for that matter) is being ignored in this marginal analysis. Krailo and Pike (1983) argue that both 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 $i = 1, ..., k$. We wish to maximize the log likelihood function

$$\ell(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^{k} \{x_i \log p_i + y_i \log q_i + z_i \log[1 - p_i - q_i]\}, \tag{3}$$

where $\mathbf{p} = (p_1, ..., p_k)$ and $\mathbf{q} = (q_1, ..., q_k)$. The parameter space,

$$\Theta = \{(\mathbf{p}, \mathbf{q}) : 0 \le p_1 \le ... \le p_k;\ 0 \le q_1 \le ... \le q_k; \mathbf{1} - \mathbf{p} - \mathbf{q} \ge 0\},$$

is a compact convex set in $\mathcal{R}^{2k}$.

A few general remarks follow easily at this stage.

R_1 If for each i, $1 \le i \le k$, $x_i < n_i$ and $y_i < n_i$, then ℓ is a strictly concave function of $(\mathbf{p}, \mathbf{q})$. As a consequence, there exists a unique MLE $(\hat{\mathbf{p}}, \hat{\mathbf{q}})$ of $(\mathbf{p}, \mathbf{q})$ in Θ.

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 $(\mathbf{p}, \mathbf{q})$. 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, $(\hat{\mathbf{p}}, \hat{\mathbf{q}})$ is an MLE if and only if the directional derivative from $(\hat{\mathbf{p}}, \hat{\mathbf{q}})$ toward any other point in Θ is non-positive. That is, if and only if

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

for all $(\mathbf{p}, \mathbf{q}) \in \Theta$.

Consider now the one dimensional problem of maximizing the likelihood with respect to $\mathbf{p}$ for given $\mathbf{q}$. This log likelihood can be written

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

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

CYCLICAL ALGORITHM:

1. Let $\mathbf{q}^{(0)}$ be an initial estimate that satisfies the restriction $q_1^{(0)} \leq ... \leq q_k^{(0)}$. Set $j = 0$.
2. Find $\mathbf{p}^{(j+1)} = \tilde{\mathbf{p}}(\mathbf{q}^{(j)})$; find $\mathbf{q}^{(j+1)} = \tilde{\mathbf{q}}(\mathbf{p}^{(j+1)})$.
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 $(\hat{\mathbf{p}}, \hat{\mathbf{q}})$, the unique M¿E of $(\mathbf{p}, \mathbf{q})$.

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 each choice of starting value for q_k over the interval $[0, 1]$. A tabulation of the maximized constrained likelihood (3) then identifies the global maximum. An alternative approach is to place a small mass Δ in place of $z_k = 0$ in the likelihood and apply the algorithm to the revised likelihood. As $\Delta \to 0$, the resulting estimates should converge to the MLEs in the original problem with $z_k = 0$. The algorithm tends to become slow, however, when Δ is small. A numerical one-dimensional Newton algorithm in q_k provides a further alternative.

Finally, if $z_{k-r-1} > 0$ and $z_i = 0$ for $k - r \le i \le k$ with $r \ge 1$, the situation is similar. With starting value $\mathbf{q}^{(0)}$ the algorithm converges to an estimator that has $\bar{q}_i = q_k^{(0)}$, $k-r \le i \le k$, 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 $\mathbf{p}$ given by

$$\phi(\mathbf{p}) = \sum_{i=1}^{k} [x_i \log p_i + z_i \log(1 - p_i - q_i)] \equiv [x_i \log p_i + z_i \log(c_i - p_i)], \tag{4}$$

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

As in the overall maximization, we must be careful at the boundary of the parameter space. First, we assume that $x_i + z_i > 0$; this corresponds to $y_i < n_i$ as discussed in the second remark in Section 2. If $x_i + z_i = 0$, then the corresponding p_i does not appear in ϕ and so cannot be identified, in general. We reiterate the convention of Section 2: in such circumstances we take $\hat{p}_i$ to equal $\hat{p}_{i-1}$ (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 $\hat{p}_1 = \hat{p}_2 = \cdots = \hat{p}_j = 0$ 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 $j = 1, ..., k - 1$. 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 $p_1, ..., p_{k-1}$ 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 $\hat{p}_{u+1}, ..., \hat{p}_k = c_k$ for some $u \leq k-1$. Having done this, we can proceed with maximization over $p_1, ..., p_r$ "away" from the boundary constraint.

Let $\mathbf{S}(\mathbf{p}) = (S_1(\mathbf{p}), \ldots, S_k(\mathbf{p}))$ where

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

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

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

$$\sum_{i=1}^{k} p_i S_i(\mathbf{p}) = 0,$$

and

$$\sum_{i \geq j}^{k} S_i(\mathbf{p}) \leq 0, \text{ for } 1 \leq j \leq k.$$

Before we return to the boundary problem, we describe a modified pool adjacent violators 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 $\phi^{(s,t)} = \sum_{l=s}^{t} [x_l \log p + z_l \log(c_l - p)]$. Let ${p_{min}}^{(s,i)} = \min_{t \geq i} p^{(s,t)}$ for $s \leq i$ and define

$$\left\{ i : p_{min}^{(i,i)} = \max_{s \leq i} p_{min}^{(s,i)} \right\} \equiv \{k_1^*, \ldots, k_r^*\},$$

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 $\mathbf{p}$ that maximizes ϕ, as characterized in Theorem 2, is $\hat{\mathbf{p}} = (\hat{p}_1, \ldots, \hat{p}_k)^T$ where

$$\hat{p}_i = \hat{\theta}_j, \quad k_j^* \leq i < k_{j+1}^*, j = 1, \ldots, r, \tag{5}$$

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 $(\theta_1, ..., \theta_r)$ are simply grouped proportions over the relevant blocks. Here, it is still straightforward to

obtain the estimates but now, θ_j is the solution to a polynomial equation of order t if the block corresponding to θ_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 $i = u+1, ..., k$ 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 $\alpha_0, \alpha_1, ..., \alpha_s$, each block comprising sequential indices. Denote the entries of an arbitrary set of blocks by $\alpha_j = \{i : k_j \leq i < k_{j+1}\},\ j = 0, ..., r$ where $1 \leq k_0 < k_1 < ... < k_{r+1} = k+1$. For each $i \in \alpha_j$, p_i is estimated by the common value $\theta_i \in [0, c_k]$ which maximizes

$$\phi_{k_j, k_{j+1}-1} = \sum_{i \in \alpha_j} \{x_i \log p + z_i \log(c_i - p)\}.$$

With this background, we define:

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

P3. There is a violation. Pool α_0 and α_1, label as α_0 and determine the new θ_0. Relabel $\alpha_2, ..., \alpha_s$ as $\alpha_1, ..., \alpha_{s-1}$. 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 $\alpha_0, ..., \alpha_s$ as $\alpha_1, ..., \alpha_{s+1}$, set $\alpha_0 = \{k_0 - 1\}$, and go to P2.

At the conclusion of the algorithm, the blocks $\alpha_0, ...\alpha_s$ and probabilities $\theta_0, ..., \theta_s$ 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 $\phi(\mathbf{q})$, defined as for $\phi(\mathbf{p})$ 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 $q_2 \leq 0.5$, ${p_1}^{(1)} = 0.5$. 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 $p \leq .5$ 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 $F_j(t) = Pr(T < t, J = j)$ 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 $(C_1, ..., C_{26}) = (27.5, 32.5, 35.5, 36.5, ...58.5)$.

Note that $\max(x_i, y_i) < n$ for all i, so that all p_is and q_is appear in the likelihood. On the other hand, $z_{26} = 0$ so that $q_{26} = 1 - p_{26}$ remains fixed throughout the algorithm. Subsequently, $\hat{q}_{26}$ 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, for example, $\{q_{26} : l_{max}(q_{26}) - l_{max}(\hat{q}_{26}) > -1.92\} = [.618, .742]$. 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 $F_1(C_i) = \hat{p}_i$ 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 $t \in [C_{j-1}, C_j)$, $j = 1, ..., k$, where $C_0 = 0$.

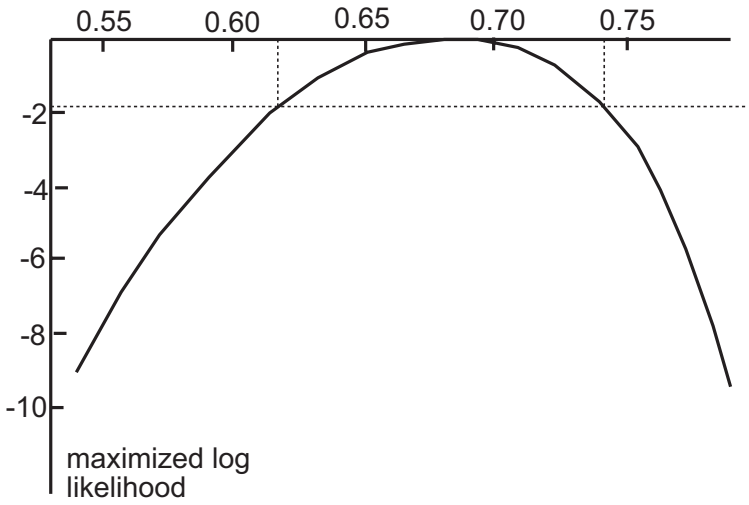

Figure 1: Maximized log likelihood (standardized to have maximum value 0) as a function of q_{26} in the current menopausal status data in Table 1. Note that $\hat{q}_{26} = 0.690$.

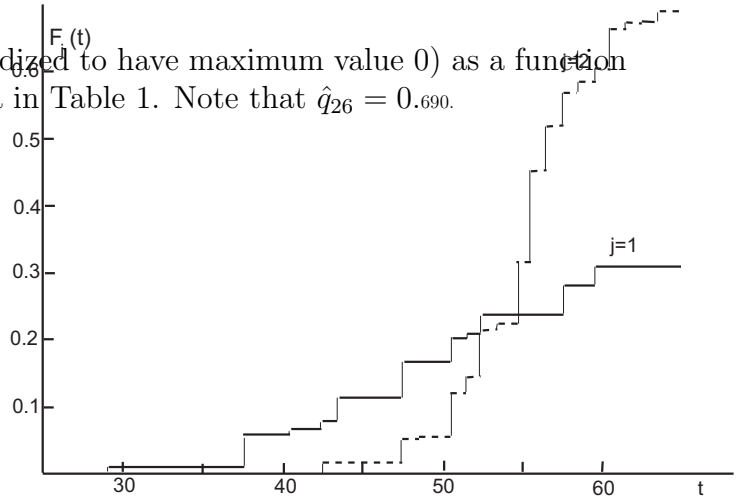

Figure 2: Bivariate isotonic estimates of the sub-distribution function $F_j(t)$, $j = 1, 2$ for the current menopausal status data in Table 1 where $j = 1$ corresponds to operative menopause and $j = 2$ to natural menopause.

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 λ_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 λ_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 (.258, .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 $F_j(t) = .5\{1 - \exp(-t)\},\ j = 1, 2$. The observation times were taken at $C_j = 2jk^{-1},\ j = 1, ..., k$. The estimates were examined at times .2, .4, . . . , 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 $\hat{F}_1^{(N)}, \hat{F}_2^{(N)}$ and $\hat{F}^{(N)}$ 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 $\hat{F}^{(N)}$ 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 .5 independently. It is intuitively clear

Table 2: Estimated mean squared errors of 'naive' MLEs ($\hat{F}^{(N)}$) and MLEs ($\hat{F}(t)$) for the CDF $F(t)$ of survival time: n = total sample size; observation times are $C_j = 2jk^{-1},\ j = 1, ...k$ 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.

		Observation Times									
n/k		t=.2	t=.4	t=.6	t=.8	t=1.0	t=1.2	t=1.4	t=1.6	t=1.8	t=2.0
100/10	$\hat{F}^{(N)}$	.0115	.0135	.0130	.0115	.0097	.0080	.0068	.0059	.0056	.0067
	$\hat{F}$	.0104	.0121	.0115	.0101	.0088	.0074	.0065	.0057	.0054	.0066
100/20	$\hat{F}^{(N)}$	.0140	.0148	.0138	.0119	.0102	.0083	.0068	.0060	.0065	.0105
	$\hat{F}$	.0133	.0139	.0123	.0108	.0096	.0081	.0070	.0065	.0068	.0102
100/50	$\hat{F}^{(N)}$	.0161	.0158	.0134	.0111	.0095	.0080	.0067	.0060	.0064	.0101
	$\hat{F}$	.0149	.0155	.0130	.0105	.0085	.0073	.0063	.0057	.0059	.0112

Table 3: Estimated mean squared errors of 'naive' MLEs ($\hat{F}_1^{(N)}$) and MLEs ($\hat{F}_1(t)$) for sub distribution function $F_1(t)$: n = total sample size; observation times are $C_j = 2jk^{-1},\ j = 1, ...k$ 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.

		Observation Times									
n/k		t=.2	t=.4	t=.6	t=.8	t=1.0	t=1.2	t=1.4	t=1.6	t=1.8	t=2.0
100/10	$\hat{F}_1^{(N)}$	.0052	.0065	.0067	.0064	.0061	.0058	.0059	.0063	.0081	.0163
	$\hat{F}_1$	.0053	.0066	.0067	.0062	.0058	.0053	.0053	.0056	.0067	.0093
100/20	$\hat{F}_1^{(N)}$	.0062	.0073	.0069	.0067	.0064	.0062	.0063	.0069	.0099	.0326
	$\hat{F}_1$	.0064	.0075	.0072	.0068	.0065	.0065	.0069	.0082	.0114	.0206
100/50	$\hat{F}_1^{(N)}$	.0062	.0081	.0075	.0068	.0064	.0064	.0063	.0071	.0103	.0753
	$\hat{F}_1$	.0064	.0083	.0075	.0065	.0058	.0055	.0053	.0056	.0065	.0104
400/10	$\hat{F}_1^{(N)}$	.0017	.0023	.0024	.0024	.0022	.0021	.0020	.0020	.0023	.0042
	$\hat{F}_1$	.0018	.0024	.0024	.0023	.0021	.0019	.0018	.0017	.0018	.0024
400/20	$\hat{F}_1^{(N)}$	.0021	.0025	.0026	.0024	.0023	.0022	.0021	.0021	.0027	.0081
	$\hat{F}_1$	.0021	.0025	.0025	.0024	.0022	.0020	.0018	.0018	.0020	.0032
400/50	$\hat{F}_1^{(N)}$	.0023	.0025	.0026	.0025	.0023	.0022	.0020	.0022	.0029	.0187
	$\hat{F}_1$	.0023	.0025	.0026	.0024	.0022	.0020	.0018	.0018	.0020	.0042

Table 4: Estimated means of 'naive' MLEs ($\hat{F}_1^{(N)}$) and MLEs ($\hat{F}_1(t)$) for sub distribution function $F_1(t)$: n = total sample size; observation times are $C_j = 2jk^{-1},\ j = 1, ...k$ 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.

		Observation Times									
n/k		t=.2	t=.4	t=.6	t=.8	t=1.0	t=1.2	t=1.4	t=1.6	t=1.8	t=2.0
100/10	$\hat{F}_1^{(N)}$	.0751	.1499	.2110	.2609	.3029	.3381	.3700	.4006	.4366	.4932
	$\hat{F}_1$	.0764	.1531	.2158	.2676	.3102	.3448	.3746	.4006	.4261	.4543
100/20	$\hat{F}_1^{(N)}$	.0638	.1439	.2083	.2597	.3035	.3408	.3747	.4069	.4483	.5413
	$\hat{F}_1$	.0656	.1484	.2164	.2704	.3161	.3537	.3868	.4162	.4484	.4922
100/50	$\hat{F}_1^{(N)}$	.0520	.1352	.2032	.2570	.3000	.3371	.3702	.4044	.4497	.6165
	$\hat{F}_1$	.0542	.1403	.2101	.2652	.3077	.3424	.3716	.3983	.4264	.4780
400/10	$\hat{F}_1^{(N)}$	.0877	.1610	.2207	.2700	.3098	.3436	.3725	.3973	.4224	.4582
	$\hat{F}_1$	.0880	.1620	.2222	.2723	.3125	.3464	.3744	.3976	.4174	.4396
400/20	$\hat{F}_1^{(N)}$	.0829	.1588	.2190	.2693	.3102	.3442	.3727	.3995	.4283	.4827
	$\hat{F}_1$	.0836	.1604	.2212	.2720	.3136	.3474	.3749	.3991	.4208	.4492
400/50	$\hat{F}_1^{(N)}$	.0802	.1561	.2179	.2682	.3098	.3434	.3723	.3986	.4290	.5212
	$\hat{F}_1$	.0811	.1581	.2207	.2714	.3136	.3469	.3748	.3978	.4195	.4582
true $F_1(t)$		.0906	.1648	.2256	.2753	.3161	.3494	.3767	.3991	.4174	.4323

that doing this and using the MLE based on $m = 3$ should not improve estimation of $F(t)$.

Table 3 gives a similar comparison of $\hat{F}_1(t)$ with $\hat{F}_1^{(N)}$. Here differences are observed which 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 ith individual where $J_i = 0$ if the ith 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 $(J_i^*, C_i^*), i = 1, ...n$ obtained by iid sampling of the observed data. The bootstrap estimates are F_1^*, F_2^*, and $F^* = F_1^* + F_2^*$. Following Efron's percentile method, a $100(1-\alpha)\%$ confidence interval for $F_1(t)$ at a fixed value of t is obtained as the interval spanned by the upper and lower $\alpha/2$ quantiles of the bootstrap sample $F_1^*(t)$. 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.

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.

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

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.

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

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 $\mathbf{X}_j$s, that take advantage of the isotonicity of the components of the $\mathbf{p}_j$s. 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 $C_1, ..., C_j$ 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 $\mathbf{p}$ discussed in Section 3, we could use a modified weighted pool adjacent violator 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.

Other estimation criteria could also be invoked. Multivariate weighted least squares (Sasabuchi, Inutska, and Kulatunga, 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 et al (2003). For the estimates $\hat{F}_1(t_0)$ and $\hat{F}_2(t_0)$ 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, $p_{ji} = d$,

for some i and j using essentially the same algorithm. This can be achieved by replacing x_i and z_i in (4) with large values, x_i^* and z_i^* so that $S_i^*(d) = 0$ 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 subdistributions, however, the results of Banerjee and Wellner (2001, 2003a, 2003b) 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.

APPENDICES

A Proof of Theorem 1

Lemma 1: Suppose that $(\mathbf{p}, \mathbf{q}) \in \Theta$ and $p_k + q_k < 1$. Then, for any $(\mathbf{p}^*, \mathbf{q}^*) \in \Theta$, there exists $\epsilon > 0$ such that $[(1-\epsilon)\mathbf{p} + \epsilon\mathbf{p}^*, \mathbf{q}] \in \Theta$.

Proof: Since $p_k + q_k < 1$, there exists $\epsilon > 0$ such that $[(1-\epsilon)\mathbf{p} + \epsilon\mathbf{p}^*] + \mathbf{q} < 1$. This is sufficient for the claim since the entries $(1-\epsilon)\mathbf{p} + \epsilon\mathbf{p}^*$ and $\mathbf{q}$ are isotonic.
A similar result holds with $\mathbf{p}$ and $\mathbf{q}$ interchanged. $\triangleleft$

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

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

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

We now turn to the same issue, but with $\mathbf{p}$ exchanging roles with $\mathbf{q}$. Unfortunately, there is not a direct symmetry since $\mathbf{p}^{(j')}$ maximizes $\ell^\dagger(\mathbf{p}, \mathbf{q}^{(j'-1)})$, not $\ell^\dagger(\mathbf{p}, \mathbf{q}^{(j')})$. So, from the convergent subsequence, $(\mathbf{p}^{(j')}, \mathbf{q}^{(j')})$ say, we take a further convergent subsequence of $(\mathbf{p}^{(j'')}, \mathbf{q}^{(j'')})$ that converges to $(\bar{\mathbf{p}}^*, \bar{\mathbf{q}})$, say. It follows that $\bar{\mathbf{p}}^*$ maximizes the likelihood $\ell^\dagger(\mathbf{p}, \bar{\mathbf{q}})$. We must now show that $\bar{\mathbf{p}}^* = \bar{\mathbf{p}}$.

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

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

$$D_\ell = \sum_{i=1}^{k} \left[x_i \frac{p_i^* - \bar{p}_i}{\bar{p}_i} + y_i \frac{q_i^* - \bar{q}_i}{\bar{q}_i} - z_i \frac{p_i^* - \bar{p}_i + q_i^* - \bar{q}_i}{1 - \bar{p}_i - \bar{q}_i} \right]. \tag{6}$$

Since $\bar{\mathbf{p}}$ maximizes the likelihood $\ell^\dagger(\mathbf{p}, \bar{\mathbf{q}})$, as a consequence of Lemma 1, the directional

derivative of $\ell^\dagger$ from $\bar{\mathbf{p}}$ toward $\mathbf{p}^*$ is

$$D_\ell^\dagger(\bar{\mathbf{p}};\mathbf{p}^*) = \sum_{i=1}^{k}\left[x_i\frac{p_i^*-\bar{p}_i}{\bar{p}_i} - z_i\frac{p_i^*-\bar{p}_i}{1-\bar{p}_i-\bar{q}_i}\right] \;\leq\; 0. \tag{7}$$

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

Thus, $\ell^\infty = \lim_{j\to\infty}\ell(\mathbf{p}^{(j)},\mathbf{q}^{(j)})$ is the maximum of ℓ over Θ. Finally, since ℓ is strictly concave, $(\mathbf{p}^{(j)},\mathbf{q}^{(j)})$ converges to $(\hat{\mathbf{p}},\hat{\mathbf{q}})$, the unique MLE (see Rockafellar, 1970, Corollary 27.2.2). $\triangleleft$

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

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

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

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

Proof: At each iteration, $(\mathbf{p}^{(j)},\mathbf{q}^{(j)}) \in \Theta(q_k^{(0)})$. The directional derivative from the limit point $(\bar{\mathbf{p}},\bar{\mathbf{q}})$ to any point $(\mathbf{p}^*,\mathbf{q}^*)$ in $\Theta(q_k^{(0)})$ is exactly as in (6) except that the upper limit of the sum is $k-1$. An argument identical to that used in Theorem 1 gives the required result. $\triangleleft$

B Proof of Theorem 2

Proof. First, suppose $\mathbf{p}$ satisfies the two properties in the statement of Theorem 2. Since ϕ is concave, $\phi(\mathbf{p}^*) - \phi(\mathbf{p}) \leq \langle \mathbf{S}(\mathbf{p}), (\mathbf{p}^* - \mathbf{p})\rangle$ for any $\mathbf{p}^*$ satisfying the same constraints as $\mathbf{p}$. (For vectors $\mathbf{x},\mathbf{y}$, $\langle \mathbf{x},\mathbf{y}\rangle = \sum_{i=1}^k x_i y_i$). By the first property, $\langle \mathbf{S}(\mathbf{p}),\mathbf{p}\rangle = 0$. Let $\mathbf{e}_1 = (0,0,\ldots,0,c_k), \mathbf{e}_2 = (0,0,\ldots,0,c_k,c_k),\ldots,\mathbf{e}_k = (c_k,c_k,\ldots,c_k)$. 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 $\mathbf{p}^* = \sum_{j=1}^k \alpha_j \mathbf{e_j}$. Then, $\langle \mathbf{S}(\mathbf{p}),\mathbf{p}^*\rangle = \sum_{i=1}^k \alpha_i \sum_{j\geq i}^k c_k S_j(\mathbf{p}) \leq 0$, by the second property. Thus, $\phi(\mathbf{p}^*) \leq \phi(\mathbf{p})$ for all such $\mathbf{p}^*$, and so $\mathbf{p}$ maximizes ϕ.

On the other hand, suppose $\mathbf{p}$ maximizes ϕ with $0 \leq p_i \leq c_i$ and $p_1 \leq p_2 \leq \cdots \leq p_k$. Since $p_1 > 0$ and $p_k < c_k$, $(1+\epsilon)\mathbf{p}$ satisfies the same constraints as $\mathbf{p}$ for sufficiently small

ϵ. Thus, $\lim_{\epsilon \to 0} \frac{\phi((1+\epsilon)\mathbf{p}) - \phi(\mathbf{p})}{\epsilon} = \sum_{i=1}^{k} p_i S_i(\mathbf{p}) = 0$. Also, for $0 < \epsilon < \frac{c_k - p_k}{c_k}$, $\mathbf{p} + \epsilon \mathbf{e}_{k-i+1}$ satisfies the same constraints as $\mathbf{p}$ for $1 \leq i \leq k$, so that

$$\lim_{\epsilon \searrow 0} \frac{\phi(\mathbf{p} + \epsilon \mathbf{e}_{k-i+1}) - \phi(\mathbf{p})}{\epsilon} = \sum_{j \geq i}^{k} c_k S_j(\mathbf{p}) \leq 0,$$

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, ϕ is strictly concave.

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

C Proof that (5) Describes the MLE

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

$$\sum_{i=l}^{k} S_i(\hat{\mathbf{p}}) = \sum_{i=l}^{k_{j+1}^*-1} S_i(\hat{\mathbf{p}}) \leq 0$$

which establishes the second condition in Theorem 2. $\triangleleft$

References

Ayer, M, Brunk, H.D., Ewing, G.M., Reid, W.T., Silverman, E. (1955). An empirical distribution function for sampling with incomplete information. *The Annals of Mathematical Statistics* **26**, 641-647.

Banerjee, M., Wellner J.A. (2001). Likelihood ratio tests for monotone functions. *Annals of Statistics* **29**, 1699-1731.

Banerjee, M., Wellner J.A. (2003a). Likelihood ratio, score and Wald statistics in models with monotone functions: Some comparisons. Submitted for publication.

Banerjee, M., Wellner J.A. (2003b). Confidence intervals for current status data. Submitted for publication.

Barlow, R.E., Bartholomew, D.J., Bremner, J.M., Brunk, H.D. (1972) *Statistical Inference under Order Restrictions*. New York: Wiley.

Groeneboom, P., Wellner, J.A. (1980) *Nonparametric Maximum Likelihood Estimators for Interval Censoring and Denconvolution*. Boston: Birkhäuser-Boston.

Hudgens, M.G., Satten, G.A., Longini, I.M.,Jr. (2001). Nonparametric maximum likelihood estimation for competing risks survival data subject to interval censoring and truncation. *Biometrics* **57**, 74-80.

Jewell, N.P., Shiboski, S. (1990). Statistical analysis of HIV infectivity based on partner studies. *Biometrics* **46**, 1133-1150.

Jewell, N.P., van der Laan, M. (2003). Current status data: Review, recent developments and open problems. *Handbook of Statistics*, eds. N. Balakrishnan and C.R. Rao, New York: North-Holland, to appear.

Jewell, N.P., van der Laan, M., Henneman, T. (2003). Nonparametric estimation from current status data with competing risks. *Biometrika*, **90**, 183-197.

Jongbloed, G. (1995) *Three Statistical Inverse Problems*. Ph.D. dissertation, Delft: Delft University of Technology.

Kalbfleisch, J.D., Prentice, R.L. (2002) *The Statistical Analysis of Failure Time Data*. Second Edition. New York: Wiley.

Krailo, M.D., Pike, M.C. (1983). Estimation of the distribution of age at natural menopause from prevalence data. *American Journal of Epidemiology* **117**, 356-361.

Kruskal, J.B. (1964) Nonmetric multidimensional scaling: a numerical method. *Psychometrika* **29**, 115-129.

Lange, K. (1999) *Numerical Analysis for Statisticians.* New York: Springer.

MacMahon, B., Worcester, J. (1966). Age at menopause, United States 1960–1962. *National Center for Health Statistics; Vital and Health Statistics, Series 11: Data from the National Health Survey, no. 19* Washington, DC: DHEW Publication no. (HSM) 66-1000.

Rockafellar, R.T. (1970) *Convex Analysis.* Princeton: Princeton University Press.

Sasabuchi, S., Inutska, M., Kulatunga, D.D.S. (1983). A multivariate version of isotonic regression. *Biometrika* **70**, 465-472.

Wu, Chien-Fu (1982). Some algorithms for concave and isotonic regression. *TIMS/Studies in the Management Sciences* **19**, 105-116.