Scand J Statist 20: 51–62, 1993

Inference from Discrete Life History Data: a Counting Process Approach

DEBASIS SENGUPTA
Indian Statistical Institute, Calcutta

S. RAO JAMMALAMADAKA
University of California, Santa Barbara

ABSTRACT. This paper outlines the scope of the counting process techniques for life history data in the discrete time set-up. We give a weak convergence theorem for discrete martingales. The convergence is in the space of all real sequences endowed with the Fréchet metric, as the number of subjects go to infinity. While some results are easier to derive than in the continuous case, the possibility of multiple jumps causes complication. Subsequently the issues of estimation, testing and regression are discussed. Relationship of the present framework with the other works in this area is also pointed out.

Key words: multivariate counting process, martingale, invariance principles, regression models, non-parametric inference, parametric inference

1. Introduction

Since the first use of the counting process theory for the analysis of life history data by Aalen (1975, 1978), several researchers have advocated a discrete time formulation of the same. The two major arguments against continuous time set-up are as follows. The counting process approach demands accurate observation of time, as pointed out by Arjas (1985). Inaccuracy of measurements sometimes leads to tied event times, which is treated in an *ad hoc* manner. Secondly, grouping of data over a time interval is often necessary because of cost and feasibility considerations. This makes the continuous time results unsuitable for direct application. A third argument which can also be used is that discrete data may arise *naturally* in machines such as computers and computer-controlled devices which operate with a digital clock.

Hjort (1985) considered grouped and partially censored survival data. He obtained the exact maximum likelihood estimator of the discrete time cumulative failure rate, and suggested that its distribution should be closely related to that of the Nelson–Aalen estimator when the bin width goes to zero at a suitable rate, as the *number of subjects* go to infinity. In a subsequent paper (Hjort, 1990b) he dealt with non-parametric Bayes estimators in this context. Arjas & Haara (1987) investigated a discrete time logistic regression model. They established the asymptotic normality of the regression coefficient estimators via a martingale convergence theorem, as the *observation time* goes to infinity. The authors used this approach again for a generalized Cox regression model; see Arjas & Haara (1988).

The approach taken in this paper is similar to that of Hjort (1985), although we do not make any assumption about the origin of the discrete data. The asymptotic arguments, as the number of individuals go to infinity, follow directly from a discrete martingale invariance principle. Since the data need not be the sampled version of a continuous process, the time interval between the discrete points is not assumed to go to zero.

2. The multiplicative intensity model in discrete time

Let $(\Omega, \mathscr{F}, P)$ be a probability space and for $i = 1, 2, \ldots, n$ and $h = 1, 2, \ldots, H$, $\{\Delta L_{ih}(k)\}_{k \geqslant 1}$ be a family of stochastic processes defined on $(\Omega, \mathscr{F})$, having discrete parameter

k and state space $\{0, 1\}$. [Henceforth we shall denote all deterministic and random sequences on $\mathbb{N}$ with a dot in the argument.] Let $\mathscr{F}_{n,0} = \{\phi, \Omega\}$ and $\mathscr{F}_{n,k}$ be a sub-σ-algebra of $\mathscr{F}$ with respect to which $\{\Delta L_{ih}(l)\}_{1 \leqslant l \leqslant k}$ are measurable for each i and h. We further define $\Delta N_{nh}(k) = \Sigma_{i=1}^{n} \Delta L_{ih}(k)$ and $N_{nh}(k) = \Sigma_{l=1}^{k} \Delta N_{nh}(l)$. The process $N_{nh}(\cdot)$ can be thought of as a "counting" process with the discrete time parameter. The index n indicates the aggregate over n individuals, while h represents the type label. We impose the following restriction on $\Delta L_{ih}(\cdot)$:

$$\text{(A1)} \qquad \sum_{h=1}^{H} \Delta L_{ih}(k) \leqslant 1 \quad \text{for each } k \in \mathbb{N}.$$

This assumption says that the same individual can not have two different types of jump at the same time. We can define the predictable process $\lambda_{nh}(\cdot)$ as

$$\lambda_{nh}(k) = E[\Delta N_{nh}(k) \mid \mathscr{F}_{n,k-1}]. \tag{2.1}$$

This process is analogous to the stochastic intensity of a usual counting process. In the same manner as Aalen (1978) we postulate that

$$E[\Delta L_{ih}(k) \mid \mathscr{F}_{n,k-1}] = \alpha_h(k) X_{ih}(k) \quad \text{a.s.} \tag{2.2}$$

where for each h, $X_{1h}(\cdot), \ldots, X_{nh}(\cdot)$ are predictable binary processes which are independent and identically distributed. On the other hand, the deterministic sequence $\alpha_h(\cdot)$ is

$$\alpha_h(k) = \begin{cases} P[\Delta L_{ih}(k) = 1 \mid X_{ih}(k) = 1] & \text{if } P[X_{ih}(k) = 1] \geqslant 0 \\ 0 & \text{otherwise} \end{cases}$$

for any i. We put the following extra condition on $X_{ih}(k)$:

(A2) There is a partition $\{H_1, H_2, \ldots, H_m\}$ of the set $\{1, 2, \ldots, H\}$ such that the following holds:

$$X_{ih}(k) = 1 \text{ and } h \in H_r \quad \text{implies} \quad X_{ig}(k) = \begin{cases} 1 & \text{if } g \in H_r, \\ 0 & \text{if } g \in H_s, s \neq r. \end{cases}$$

This means the types of jumps can be grouped in such a manner that an individual can be ready for only one group of jump types at a time. This is a very reasonable assumption that helps simplify the likelihood expression in section 4.

From (2.2) it follows that

$$\lambda_{nh}(k) = \alpha_h(k) Y_{nh}(k) \quad \text{a.s.,} \tag{2.3}$$

where $Y_{nh}(k) = \Sigma_{i=1}^{n} X_{ih}(k)$, the number (out of n) of individuals ready for a type h jump at time k. We shall call $\alpha_h(\cdot)$ the deterministic intensity of type h transition. The above is similar to the celebrated multiplicative intensity model (Aalen, 1978). The crucial difference of the above model with Aalen's model is that while the time parameter in the latter is allowed to take values in an interval of the form $[0, T]$, no restriction of finiteness is necessary on the discrete time here. Further, multiple jumps of the process $N_{nh}(\cdot)$ are allowed. In this respect it is comparable to Johansen's (1983) extension of the Aalen model. Notice that a specific decomposition of the processes $\Delta N_{nh}(\cdot)$ and $Y_{nh}(\cdot)$ has been used here. We do not assume that the processes $\Delta L_{ih}(\cdot)$ and $X_{ih}(\cdot)$ for different individuals are i.i.d. This assumption will sometimes be made explicitly in order to simplify formulas.

We conclude this section with the discrete time versions of two examples from Andersen & Borgan (1985).

Example 1. Survival data with random right-censoring. Suppose that i.i.d. censored lifetimes of n individuals are denoted by $T_1, T_2, \ldots, T_n$. These are assumed to be positive and integer-valued, while any unit of time can be used. Let $D_1, D_2, \ldots, D_n$ be the corresponding censoring times. For $1 \leqslant i \leqslant n$ and each $k \in \mathbb{N}$ we define $\Delta L_i(k)$ to be the product of the indicator variables $I(T_i = k)$ and $I(D_i \geqslant k)$, and assume that the pair of indicator variables is measurable $\mathscr{F}_{n,k}$. The subscript h is dropped for simplicity. Then (2.2) holds with $X_i(k) = I(T_i \geqslant k, D_i \geqslant k)$. One can interpret $Y_n(k)$ as the number of individuals at risk at time k. If the censoring time and the notional lifetimes are independent, $\alpha(k)$ is the discrete hazard rate.

Example 2. Finite state Markov chain. Suppose we have n samples from a discrete parameter Markov chain with a finite number of states and having H possible types of transition. For $h = 1, 2, \ldots, H$, $i = 1, 2, \ldots, n$ and $k \in \mathbb{N}$ we define $\Delta L_{ih}(k)$ as the indicator of the event that a "type h" transition occurs to individual i at time k. Thus $\alpha_h(k)$ becomes a transition probability. Note that the assumption (A2) holds here with m equal to the number of non-absorbing states in the Markov chain and the set H_r represents the indices of types of transitions that are possible from the rth non-absorbing state. In the special case of competing risk data, $\Delta L_{ih}(k)$, $X_{ih}(k)$ and $\alpha_h(k)$ are the indicator of death from cause h, the indicator of being alive (irrespective of h) and the discrete cause-specific hazard rate, respectively. Similar interpretations can be given in the special case of the illness–death model.

3. Weak convergence of discrete martingales

We are interersted in the convergence of "discrete martingales" in the space of all real sequences S endowed with the Fréchet metric,

$$\rho(x(\cdot), y(\cdot)) = \sum_{k \in \mathbb{N}} 2^{-k} \frac{|x(k) - y(k)|}{1 + |x(k) - y(k)|}, \quad x(\cdot), y(\cdot) \in S.$$

Several properties of the metric space (S, ρ) are listed below. Sketches of proof may be found in Billingsley (1968, pp. 218–219).

Lemma 3.1

(a) *A sequence in (S, ρ) converges if and only if all its coordinates converge with respect to the Euclidian metric. Further, the limits of the coordinates are the coordinates of the limit.*
(b) *The space S is complete and separable with respect to the metric ρ.*
(c) *A subset T of S has a compact closure if and only if there is a positive sequence $b(\cdot)$ such that $x \in T$ implies $|x(k)| < b(k)$.*

Let $\mathscr{S}$ be the Borel σ-algebra on S. Since $(S, \mathscr{S})$ is complete and separable, tightness of a sequence of probability measures on this space is equivalent to its relative compactness, by Prohorov's theorem. The characterisation of compact sets given above allows us to relate the convergence of a sequence of probability measures to the convergence of its finite dimensional marginals. In the following, we denote convergence in distribution (in the above space) and in probability by "$\Rightarrow$" and "$\xrightarrow{P}$", respectively. Let $\{\Omega, \mathscr{F}, P\}$ be a probability space. Suppose for each i, $X_i(\cdot)$ is a mapping from Ω to S, and $Z(\cdot)$ is another such mapping. Then we have

Theorem 3.2

(a) *Suppose for all $\eta \in (0, 1)$ there is a positive real sequence $b(\cdot)$ such that*

$$\inf_{n \in \mathbb{N}} P(|X_n(k)| \leqslant b(k) \quad \textit{for all } k \in \mathbb{N}) > 1 - \eta. \tag{3.1}$$

Then $X_n(\cdot) \Rightarrow Z(\cdot)$ *if and only if for each* $n \in \mathbb{N}$ *and each subset* $\{i_1, \ldots, i_r\}$ *of* $\mathbb{N}$, *the sequence of randon vectors* $\{X_n(i_1), X_n(i_2), \ldots, X_n(i_r)\}$ *converges to* $\{Z(i_1), Z(i_2), \ldots, Z(i_r)\}$ *as* $n \to \infty$.

(*b*) $\rho(X_n(\cdot), Z(\cdot)) \xrightarrow{P} 0$, *that is,* $X_n(\cdot) \xrightarrow{P} Z(\cdot)$ *if and only if* $X_n(k) \xrightarrow{P} Z(k)$ *for each* $k \in \mathbb{N}$.

Proof of part (b) can be found in Jammalamadaka & Sengupta (1990).

Now we take a triangular array of random variables $\{\xi_{n,k}; k = 1, \ldots, r_n; n = 1, 2, \ldots\}$ on the probability space. For each n and k, $\xi_{n,k}$ is measurable with respect to $\mathscr{F}_{n,k}$, a sub-σ-algebra of $\mathscr{F}$ satisfying $F_{n,k-1} \subset \mathscr{F}_{n,k}$. We assume that the above array is a martingale difference array, that is, $E_{k-1}\xi_{n,k} = E(\xi_{n,k} \mid \mathscr{F}_{n,k-1}) = 0$ a.s. We consider the weak convergence of the sum $X_n(k) = \Sigma_{l=1}^{r_n(k)} \xi_{n,l}$, where $r_n(\cdot)$ is an integer-valued nondecreasing sequence with $r_n(1) \geqslant 0$.

Theorem 3.3

Let the martingale difference array satisfy

(*a*) $\Sigma_{k=1}^{r_n} E_{k-1}\xi_{n,k}^2 I(|\xi_{n,k}| > \varepsilon) \xrightarrow{P}$ *for all* $\varepsilon > 0$.

(*b*) $\Sigma_{k=1}^{r_n(l)} E_{k-1}\xi_{n,k}^2 \xrightarrow{P} \Sigma_{k=1}^{l} \tau^2(k)$ *for each* $l \geqslant 0$, *where* $\tau(\cdot)$ *is a sequence in* ℓ_2.

Then $X_n(\cdot) \Rightarrow \Sigma_{k=1}^{(\cdot)} \tau(k)z_k$, *where* $z_1, z_2 \ldots$ *are i.i.d. standard normal random variables.*

Proof. The continuous time version of this theorem was given by Aalen (1977, see theorem A.1 there) as a generalization of corollary 3.8 of McLeish (1974). In fact one can easily obtain discrete equivalents of McLeish's theorem 3.2, theorem 3.6 and corollary 3.8, from which the above follows. The only difference is in the condition for tightness of the sequence $X_n(\cdot)$, which is given in (3.1). Notice that

$$\text{LHS of (3.1)} \geqslant 1 - \sup_{n \in \mathbb{N}} \sum_{k \in \mathbb{N}} P[|X_n(k)| > b(k)] \geqslant 1 - \sup_{n \in \mathbb{N}} \sum_{k \in \mathbb{N}} b^{-2}(k) E \sum_{l=1}^{r_n(k)} \xi_{n,l}^2.$$

As in McLeish (1974), we can work with a simple modification of the array which makes $E\, \Sigma_{l=1}^{r_n(k)} \xi_{n,l}^2$ converge, so that it is bounded. Hence the right hand side above can be made smaller than $1 - \Sigma_{k \in \mathbb{N}} \eta 2^{-k}$ by suitably choosing $b(k)$. Thus the tightness condition is verified rather easily in the discrete set-up. The remaining part of the proof is as in McLeish (1974) and Aalen (1977). We omit the details. □

For a multivariate generalization of theorem 3.3, we take $\{\xi_{n,k}^{(h)}; k = 1, \ldots, r_n; n = 1, 2, \ldots\}$, for $h = 1, \ldots, H$. These are martingale difference arrays with respect to the array of σ-algebras $\{\mathscr{F}_{n,k}\}$. Let $\mathbf{X}_n(k)$ and $\boldsymbol{\xi}_{n,k}$ be vectors (of size H) of the corresponding unidimensional quantities. We indicate the transpose of a vector with a prime. The product topology is used in the space of vector sequences.

Theorem 3.4

Let the martingale difference arrays satisfy

(*a*) $\Sigma_{k=1}^{r_n} E_{k-1}\boldsymbol{\xi}'_{n,k}\boldsymbol{\xi}_{n,k} I(\boldsymbol{\xi}'_{n,k}\boldsymbol{\xi}_{n,k} > \varepsilon) \xrightarrow{P} 0$ *for all* $\varepsilon > 0$.

or

(*a*′) $\Sigma_{k=1}^{r_n} E_{k-1}[\xi_{n,k}^{(h)}]^2 I(|\xi_{n,k}^{(h)}| > \varepsilon) \xrightarrow{P} 0$ *for all* $\varepsilon > 0$ *and* $h = 1, \ldots, H$.

and

(*b*) $\Sigma_{k=1}^{r_n(l)} E_{k-1}\boldsymbol{\xi}_{n,k}\boldsymbol{\xi}'_{n,k} \xrightarrow{P} \Sigma_{k=1}^{l} \mathbf{T}(k)\mathbf{T}(k)'$ *for each* $l \geqslant 0$, *where each component of* $\mathbf{T}(k)$ (*which is* $H \times H$) *is a sequence in* ℓ_2.

Then $\mathbf{X}_n(\cdot) \Rightarrow \Sigma_{k=1}^{(\cdot)} \mathbf{T}(k)\mathbf{z}_k$, *where* $\mathbf{z}_1, \mathbf{z}_2 \ldots$ *are i.i.d. random vectors with distribution* $N(0, \mathbf{I})$.

Proof. Condition (a′) is easily seen to follow from condition (a). The converse is proved by induction on the order of the vector. Under the stated assumptions we observe that by virtue of theorem 3.3 $\mathbf{X}_n^{(h)}(\cdot) \Rightarrow \Sigma_{k=1}^{(\cdot)} (\mathbf{e}_h' \mathbf{T}(k)\mathbf{T}(k)'\mathbf{e}_h)^{1/2} z_k^{(h)}$, where $\mathbf{e}_h$ is the hth column of the $H \times H$ identity matrix. Thus for each h the family of distributions of $\mathbf{X}_n^{(h)}(\cdot)$ is relatively compact and hence tight. Since H is finite, $\mathbf{X}_n(\cdot)$ itself is tight (see Billingsley, 1968, p. 41). Therefore it is enough to show convergence of the finite dimensional distributions. Following Aalen (1977), we can use a Cramér–Wold argument via $U_{n,k} = \Sigma_{h=1}^{H} c_h(r_n^{-1}(k))\xi_{n,k}^{(h)}$, where $c_1(\cdot), \ldots, c_H(\cdot)$ are ℓ_1 sequences satisfying $\Sigma_{h=1}^{H} |c_h(k)| \leqslant 1$ for each k. Although orthogonality of the martingales has not been assumed, we can argue that

$$\begin{aligned}\sum_{k=1}^{r_n(l)} E_{k-1} U_{n,k}^2 &= \sum_{k=1}^{r_n(l)} \mathbf{c}'(r_n^{-1}(k)) E_{k-1}(\xi_{n,k}\xi_{n,k}')\mathbf{c}(r_n^{-1}(k)) \\ &= \sum_{k=1}^{l} \mathbf{c}'(k)[\mathbf{D}_n(k) - \mathbf{D}_n(k-1)]\mathbf{c}(k) \\ &\quad [\text{where } \mathbf{D}_n(\cdot) = \Sigma_{k=1}^{r_n(\cdot)} E_{k-1}(\xi_{n,k}\xi_{n,k}')] \\ &\xrightarrow{P} \sum_{k=1}^{l} \mathbf{c}'(k)\mathbf{T}(k)\mathbf{T}(k)'\mathbf{c}(k).\end{aligned}$$

Thus $U_{n,k}$ satisfies condition (b) of theorem 3.3. The rest of the argument is similar to that of Aalen (1977) and hence is omitted. □

Remark 1. In the above theorem, $\Sigma_{k=1}^{r_n(l)} E_{k-1}\xi_{n,k}^{(h)}\xi_{n,k}^{(g)}$ is *not* assumed to converge (in probability) to 0 for $h \neq g$. We can not afford to make this simplistic assumption in the discrete case.

Remark 2. Alternative sets of conditions similar to those of theorem 3.3 of Helland (1982) can also be proved to suffice.

4. Non-parametric maximum likelihood estimation

4.1. The maximum likelihood estimator

The likelihood function is

$$\prod_{k \in \mathbb{N}} \left[\left\{ \prod_{h=1}^{H} \alpha_h(k)^{\Delta N_{nh}(k)} \right\} \prod_{i=1}^{n} \left(1 - \sum_{h=1}^{H} \alpha_h(k) X_{ih}(k) \right)^{1 - \Sigma_{h=1}^{H} \Delta L_{ih}(k)} \right]. \tag{4.1}$$

In view of assumption (A2), the likelihood function reduces to

$$\prod_{k \in \mathbb{N}} \prod_{r=1}^{m} \left[\left\{ \prod_{h \in H_r} \alpha_h(k)^{\Delta N_{nh}(k)} \right\} \left(1 - \sum_{h \in H_r} \alpha_h(k) \right)^{Y_{nh_r}(k) - \Sigma_{h \in H_r} \Delta N_{nh}(k)} \right], \tag{4.2}$$

where h_r is any member of H_r. The likelihood (4.2) is evidently maximized by the estimator

$$\hat{\alpha}_{nh}(k) = J_{nh}(k) Y_{nh}^{-1}(k)\, \Delta N_{nh}(k), \tag{4.3}$$

where $J_{nh}(k) = I(Y_{nh}(k) > 0)$. It is interesting to note that the maximisation of the likelihood function is much more difficult in the continuous time case (see Karr, 1987) unless the model is modified in some manner (Johansen, 1983; Jacobsen, 1984). We emphasize the importance of assumption (A2) here. Since it holds in most practical situations, maximisation of (4.1) becomes unnecessary.

From (4.3) one can also find the MLE of the discrete cumulative intensity $A_h(k)$ (given by $\Sigma_{l=1}^{h} \alpha_h(l)$). Its computational form coincides with that of the Nelson–Aalen estimator (Aalen, 1978). In the case of survival data with independent notional life and censoring times, the non-parametric ML or Kaplan–Meier estimator of the survival function is given by $\Pi_{i:\, k_i \leqslant k} (1 - \hat{\alpha}_n(k_i))$ where $k_1, k_2, \ldots$ are the times of observed death. When transformed to get an estimator of the hazard rate, it produces (4.3), as expected. It is also not surprising that Hjort (1985, 1990b) found (4.3) to be the MLE of $\alpha(\cdot)$ in this special case.

4.2. *Bias of the MLE*

Note that

$$\hat{\alpha}_{nh}(k) - \alpha_h(k) = J_{nh}(k) Y_{nh}^{-1}(k)\, \Delta M_{nh}(k) - (1 - J_{nh}(k))\alpha_h(k), \tag{4.4}$$

where $\Delta M_{nh}(k) = \Delta N_{nh}(k) - \lambda_{nh}(k)$ is a martingale difference with respect to the filtration $\{\mathscr{F}_{n,k}\}_{k \geqslant 1}$. It follows that

$$E[\hat{\alpha}_{nh}(k) - \alpha_h(k)] = -\alpha_h(k) P[Y_{nh}(k) = 0].$$

Thus the estimator has a finite negative bias (unless $Y_{nh}(k) > 0$ a.s.). In the i.i.d. case $P[Y_{nh}(k) = 0] = P^n[X_{1h}(k) = 0]$ and hence the bias goes to 0 at an exponential rate as $n \to \infty$.

4.3. *Consistency*

By virtue of theorem 3.2(b), a sufficient condition for the convergence (in probability) of each coordinate to the desired value will guarantee consistency. Jammalamadaka & Sengupta (1990) have used mean square convergence as a sufficient condition. In the i.i.d. case the mean squared error is

$$\alpha_h^2(k) P^n[X_{1h}(k) = 0] + \alpha_h(k)(1 - \alpha_h(k)) \sum_{i=1}^{n} \frac{1}{i} \binom{n}{i} P^i[X_{1h}(k) = 1] P^{n-i}[X_{1h}(k) = 0].$$

The above is identically 0 when $P[X_{1h}(k) = 1] = 0$. Otherwise the second term is upper-bounded by $2n^{-1}\alpha_h(k)(1 - \alpha_h(k))/P[X_{1h}(k) = 1]$. This establishes the required mean square convergence.

4.4. *Asymptotic normality*

Let $\alpha_{nh}^* = J_{nh}(k)\alpha(k)$, $A_{nh}^*(k) = \Sigma_{l=1}^{k} \alpha_{nh}^*(l)$ and $\hat{A}_{nh}(k) = \Sigma_{l=1}^{k} \hat{\alpha}_{nh}(l)$. Then

$$\hat{A}_{nh}(k) - A_{nh}^*(k) = \sum_{l=1}^{k} J_{nh}(l) Y_{nh}^{-1}(l)\, \Delta M_h(l).$$

Convergence of a properly scaled version of this martingale in the univariate and multivariate cases follow from theorems 3.3 and 3.4, respectively. However, the theorems are not suggestive of any obvious estimator of the mean squared error. This is a difficulty which is exclusive to the discrete case. The limiting covariance function will depend on the structure of dependence among individuals. In the special case of i.i.d. individuals we have the following covariance function

$$E[n(\hat{A}_{nh}(k) - A_{nh}^*(k))(\hat{A}_{ng}(k) - A_{ng}^*(k))] = \sum_{l=1}^{k} \tau_{gh}(k),$$

where

$$\tau_{hg}(k) = \begin{cases} \alpha_h(k)(1-\alpha_h(k))/E[X_{1h}(k)] & \text{if } g = h, \\ -\alpha_h(k)\alpha_g(k)/E[X_{1h}(k)] & \text{if } g \neq h, \quad h, g \in H_r \text{ for some } r, \\ 0 & \text{if } g \neq h, \quad h \in H_r, \quad g \in H_s, \quad s \neq r. \end{cases}$$

That $\tau_{hg}(k)$ is not necessarily 0 for $g \neq h$ underscores the complications arising from different kinds of jumps being allowed to occur (to different individuals) at the same time. The covariance can be consistently estimated by replacing $\alpha_h(k)$ and $\alpha_g(k)$ by their respective MLE's and $E[X_{1h}(k)]$ by $n^{-1}Y_{nh}(k)$ in the above expression. We call this estimator $\hat{\tau}_{ngh}(k)$ for future reference.

If $\Delta M_{nh}(k)$ is taken as a sum of i.i.d. random variables, one can actually establish the asymptotic normality of $\sqrt{n}(\hat{\alpha}_{nh}(\cdot) - \alpha^*_{nh}(\cdot))$. This may be done via a usual multivariate central limit theorem, which applies to the finite dimensional marginals, and making use of theorem 3.2(a). It may be possible to work in this set-up with a weaker assumption than i.i.d.

5. Other estimators

5.1. Smoothing of the nonparametric estimator

The maximum likelihood estimator may need smoothing in order to appear realistic, especially when the jumps of $N_{nh}(\cdot)$ are sparse. Let us consider the unidimensional case for simplicity. One may use a kernel estimator (as in Ramlan-Hansen, 1983) of the form

$$\hat{\alpha}_K(k) = \sum_{l=k-L}^{k+L} K(k-l)\frac{\Delta N_n(l)}{Y_n(l)}, \tag{5.1}$$

where $K(\cdot)$ is suitable kernel function having support on a suitable window $\{-L, -L+1, \ldots, L\}$ and satisfying $\Sigma_{k=-L}^{L} K(k) = 1$. The above estimator falls within the purview of theorem 3.3 and hence does not need a separate discussion of asymptotic normality. Note that it is a consistent estimator of

$$\alpha^*_K(k) = \sum_{l=k-L}^{k+L} K(k-l)\alpha^*(l),$$

which will in general be different from $\alpha(k)$. Since we can not assume $L \to 0$ in the discrete case, the kernel estimator will be biased, even asymptotically. However the local averaging will reduce the variance for small sample size when the "true" $\alpha(\cdot)$ varies slowly compared to the window width $2L$.

5.2. Parametric estimator

Let $\alpha_h(k)$ be a differentiable function of the parameter vector $\boldsymbol{\theta}$. Gradient of the log-likelihood (up to time k_0) with respect to $\boldsymbol{\theta}$ is

$$\mathbf{U}(\boldsymbol{\theta}, k_0) = \sum_{k=1}^{k_0}\sum_{r=1}^{m}\left[\sum_{h\in H_r}\Delta N_{nh}(k)\frac{\dfrac{\partial\alpha_h(k)}{\partial\theta}}{\alpha_h(k)} - \sum_{h\in H_r}\frac{\partial\alpha_h(k)}{\partial\boldsymbol{\theta}}\left(\frac{Y_{nh_r} - \sum_{g\in H_r}\Delta N_{nh}(k)}{1-\sum_{g\in H_r}\alpha_g(k)}\right)\right].$$

Upon simplification, this becomes

$$\mathbf{U}(\boldsymbol{\theta}, k_0) = \sum_{k=1}^{k_0} \sum_{r=1}^{m} \sum_{h \in H_r} \frac{\partial \alpha_h(k)}{\partial \boldsymbol{\theta}} \left[\frac{\Delta M_{nh}(k)}{\alpha_h(k)} + \frac{\sum_{g \in H_r} \Delta M_{ng}(k)}{1 - \sum_{g \in H_r} \alpha_g(k)} \right.$$

$$\left. + Y_{nh(k)} \left\{ \frac{\alpha_h^0(k)}{\alpha_h(k)} - \frac{1 - \sum_{g \in H_r} \alpha_g^0(k)}{1 - \sum_{g \in H_r} \alpha_g^0(k)} \right\} \right],$$

where $\alpha_h^0(\cdot)$ is the true value of $\alpha_h(\cdot)$, corresponding to the true parameter vector $\boldsymbol{\theta}_0$. Clearly $\mathbf{U}(\boldsymbol{\theta}_0, \cdot)$ is a discrete martingale. It follows that if $\hat{\boldsymbol{\theta}}_k$ maximizes $\mathbf{U}(\boldsymbol{\theta}, k)$, then $\mathbf{U}(\hat{\boldsymbol{\theta}}, \cdot) - \mathbf{U}(\boldsymbol{\theta}_0, \cdot)$ is a martingale. Now one can do a Taylor series expansion of the score function around $\boldsymbol{\theta}_0$, as in Andersen & Gill (1982), in order to establish the asymptotic normality of the MLE of $\boldsymbol{\theta}$.

6. Tests of hypotheses

6.1. The one-sample problem

Suppose $\alpha_1^0, \ldots, \alpha_H^0$ are specified sequences and we want to test

$$\mathscr{H}_0\colon \alpha_h = \alpha_h^0, \quad h = 1, 2, \ldots, H. \tag{6.1}$$

One can use $T_{nh} = \sqrt{n}\, \Sigma_{k \geq 1} W_{nh}(k)(\hat{\alpha}_{nh}(k) - J_{nh}(k)\alpha_h^0(k))$ for this purpose, where $W_{nh}(\cdot)$ are predictable sequences converging to (possibly unknown) non-random real sequences for each h. The covariance of T_{nh} and T_{ng} can be estimated by $\Sigma_{k \geq 1} W_{ng}(k) W_{nh}(k) \hat{\tau}_{ngh}(k)$. [The consistency of $\hat{\tau}_{ngh}(k)$ is known only in the i.i.d. case.] Thus one can form an asymptotically normal statistic for $H = 1$ and asymptotically χ_H^2-distributed statistic for $H > 1$. The development here is along the lines of the continuous time case (see Hjort, 1990a).

6.2. The K-sample problem

Suppose $N_{n_1}^{(1)}, \ldots, N_{n_K}^{(K)}$ are independent H-variate counting processes and we want to test

$$\mathscr{H}_0\colon \alpha_h^{(1)}(\cdot) = \alpha_h^{(2)}(\cdot) = \ldots = \alpha_h^{(K)}(\cdot), \quad h = 1, 2, \ldots, H. \tag{6.2}$$

Following Aalen (1978) one can combine the samples and compare the pooled estimator of $\alpha_h(\cdot)$ with the estimator from individual samples. The reader is referred to Jammalamadaka & Sengupta (1990) for a detailed discussion of the resulting test statistic.

In the special case of $K = 2$, the statistic $U_n = \mathbf{T}_n' \hat{\mathbf{C}}_n^{-1} \mathbf{T}_n$ can be used, where the components of $\mathbf{T}_n$ and $\hat{\mathbf{C}}_n$ are given by

$$T_{nh} = \sqrt{n} \sum_{k \geq 1} W_{nh}(k)(\hat{\alpha}_{n_1 h}^{(1)}(k) - \hat{\alpha}_{n_2 h}^{(2)}(k)),$$

$$\hat{C}_{nhg} = \sum_{k \geq 1} W_{nh}(k) W_{ng}(k)(\hat{\tau}_{n_1 hg}^{(1)}(k) + \hat{\tau}_{n_2 hg}^{(2)}(k)),$$

where $W_{nh}(\cdot)$ is a weight function as before. The asymptotic distribution of U_n is χ_H^2.

The above test was applied to the data on mating rates of "Ebony" and "Oregon" flies, which was also examined by Aalen (1978). The data consists of time measured in seconds from introduction in control chamber to the initiation of mating. Since the measurements are discrete, the method discussed here is directly applicable. The statistic U_n produced two-sided p-values of 3.1×10^{-4} and 2.5×10^{-6} for $W_n(k) = n^{-1} Y_{n_1}^{(1)}(k) Y_{n_2}^{(2)}(k)/[Y_{n_1}^{(1)}(k) + Y_{n_2}^{(2)}(k)]$ and $W_n(k) = n^{-2} Y_{n_1}^{(1)}(k) Y_{n_2}^{(2)}(k)$, respectively. Thus the hypothesis of equal (discrete) mating rates is rejected.

6.3. *Other tests*

In the same manner as above, one can formulate a test for the equality of components or groups of components of a multivariate intensity function. The asymptotic results of section 3 will again be useful. An important potential application of this test is in comparing cause-specific hazards of competing risks.

Another hypothesis of interest in the context of two samples is $\alpha^{(2)}(\cdot) = \theta \alpha^{(1)}(\cdot)$ for some unknown θ. A starting point would be to define the estimator

$$\hat{\theta}_{n,W} = \frac{\sum_{k \geq 1} W_n(k) \hat{\alpha}_{n_2}^{(2)}(k)}{\sum_{k \geq 1} W_n(k) \hat{\alpha}_{n_1}^{(1)}(k)},$$

where $W_n(\cdot)$ is a suitable weight function. The asymptotic normality of $\sqrt{n}(\hat{\theta}_{n,W} - \theta)$ is easy to establish, while a computational formula for the asymptotic variance can also be obtained routinely. This gives a test for $\alpha^{(2)} = \theta_0 \alpha^1$ for fixed θ_0. A reasonable statistic for the original problem is $\sqrt{n}\, \Sigma_{k \geq 1} M_n(k)(\hat{a}_{n_2}^{(2)}(k) - \hat{\theta}_{n,W} \hat{\alpha}_{n_1}^{(1)}(k))$ where $M_n(\cdot)$ is another weight function. Further details of these tests may be found in Jammalamadaka & Sengupta (1990).

A hypothesis of the form $(1 - \alpha^{(2)}(\cdot)) = (1 - \alpha^{(1)}(\cdot))^{\theta}$ comes naturally in the discrete proportional hazards model. A more general problem than this will be discussed in the next section.

7. Regression models

Analysis under the semiparametric regression models in the discrete set-up is difficult because of the presence of the "baseline hazard". We do not know of any partial likelihood formulation in this context that would be a function of the regression parameters only. Attempts have been made to approximate the baseline hazard by piecewise constant and other simpler forms. This essentially reduces the model to a completely parametric one. For simplicity we consider such a parametric model in the unidimensional case only, and drop the subscript h accordingly.

As in subsection 5.2, consider the likelihood up to time k_0. Let $\boldsymbol{\theta} = (\boldsymbol{\beta}'\gamma')'$ indicate the vector of all parameters, including those used to model the baseline hazard (γ) and those used for regression ($\boldsymbol{\beta}$). Further, let $\mathbf{z}_i(k)$ be the (predictable) covariate process of the ith individual. After some computations, the gradient of the log likelihood turns out to be

$$\mathbf{U}(\boldsymbol{\theta}, k_0) = \sum_{k=1}^{k_0} \sum_{i=1}^{n} \left[\frac{\partial}{\partial \boldsymbol{\theta}} \log \frac{\alpha(k, \mathbf{z}_i(k))}{1 - \alpha(k, \mathbf{z}_i(k))} \right] \{\Delta L_i(k) - \alpha(k, \mathbf{z}_i(k)) X_i(k)\}. \tag{7.1}$$

Thus $\mathbf{U}(\boldsymbol{\theta}, \cdot)$ is a martingale when evaluated at the true value of $\boldsymbol{\theta}$. This leads to the proof of asymptotic normality of the MLE's.

As an example, consider the part of the score vector that corresponds to $\boldsymbol{\beta}$. In the special case of the proportional odds ratio model (see Kalbfleisch & Prentice, 1980, p. 37) the quantity in the squared bracket simplifies to $\mathbf{z}_i(k)$. In the case of the discrete proportional hazard model $[1 - \alpha(k, \mathbf{z}_i(k))] = [1 - \alpha_0(k)]^{\exp(\beta' z_i(k))}$, it becomes $\alpha(k, \mathbf{z}_i(k))^{-1} \exp(\boldsymbol{\beta}' \mathbf{z}_i(k)) \log[1 - \alpha_0(k)] \mathbf{z}_i(k)$. This can be further simplified when the only covariate is the group indicator.

In each of these cases, one can also construct tests of the hypothesis $\boldsymbol{\beta} = 0$. The expressions for the information matrix are not given here. Presence of the nuisance parameter γ would complicate the structure of the information matrix, making it take the form of a Schur's complement.

8. Concluding remarks and scope of further work

A weakness of the formulation used here is that a consistent estimator of the mean squared error is easily found in the i.i.d. case only. If the i.i.d. assumption is dropped, some other condition involving the nature of dependence will be needed to ensure consistency. In the continuous time formulation this problem did not arise because simultaneous jumps of different individuals was ruled out. While the present formulation is more realistic, one has to look for meaningful departures from the i.i.d. assumption in order to expand the scope of the results.

The work of Hjort (1985, 1990b) should also be followed up. Since the discrete process is explicitly assumed to be the sampled version of a continuous process in his formulation, stronger results (such as uniform consistency of the kernel estimator) may be obtained. One may also try to show theoretically if the discrete process $\sqrt{n}[\hat{A}_{nh}(\cdot) - A_h(\cdot)]$ converges to the corresponding continuous process with the sampling interval going to zero at a specific rate. In this set-up, the loss of information due to sampling (say, in terms of the mean squared error) can also be studied.

The asymptotic approach used here is not a competitor of that used by Arjas (1985). His approach of letting the observation time go to infinity is motivated by a desire to process the data in real time and is suitable for regression models, where information tends to accumulate with time. We believe that both of these approaches will be useful in analyzing real data in a somewhat complementary manner.

Acknowledgements

We are indebted to an associate editor of this journal for his valuable suggestions and comments. We also thank the editor for his encouragement regarding the revision. Finally, we are thankful to Professors Petar Todorovic and Svetlozar T. Rachev of University of California, Santa Barbara for a few helpful discussions.

References

Aalen, O. O. (1975). *Statistical inference for a family of counting processes*. PhD Dissertation, Department of Statistics, University of California, Berkeley.

Aalen, O. O. (1977). Weak convergence of stochastic integrals related to counting processes. *Z. Wahr. verw. Geb.* **38,** 261–277. Correction (1979) **48,** 347.

Aalen, O. O. (1978). Nonparametric inference for a family of counting processes. *Ann. Statist.* **6,** 701–726.

Andersen, P. K. & Gill, R. D. (1982). Cox's regression model for counting processes: a large sample study. *Ann. Statist.* **10,** 1100–1120.

Andersen, P. K. & Borgan, Ø. (1985). Counting process models for life history data: a review. *Scand. J. Statist.* **12,** 97–140.

Arjas, E. (1985). Contribution to the discussion of Andersen & Borgan's paper "Counting process models for life history data: a review". *Scand. J. Statist.* **12,** 150–153.

Arjas, E. & Haara, P. (1987). A logistic regression model for hazard: asymptotic results. *Scand. J. Statist.* **14,** 1–18.

Arjas, E. & Haara, P. (1988). A note on the asymptotic normality in the Cox regression model. *Ann. Statist.* **16,** 1133–1140.

Billingsley, P. (1968). *Convergence of probability measures.* Wiley, New York.
Helland, I. S. (1982). Central limit theorems for martingales with discrete and continuous time. *Scand. J. Statist.* **9,** 79–94.
Hjort, N. L. (1985). Contribution to the discussion of Andersen & Borgan's paper "Counting process models for life history data: a review". *Scand. J. Statist.* **12,** 141–150.
Hjort, N. L. (1990a). Goodness of fit tests in models for life history data based on cumulative hazard rates. *Ann. Statist.* **18,** 1221–1258.
Hjort, N. L. (1990b). Nonparametric Bayes estimators based on beta processes in models for life history data. *Ann. Statist.* **18,** 1259–1294.
Jacobsen, M. (1984). Maximum likelihood estimation in the multiplicative intensity model: a survey. *Int. Statist. Rev.* **52,** 193–207.
Jammalamadaka, S. R. & Sengupta, D. (1990). *Inference on discrete life history data.* Technical Report No. 116, University of California, Santa Barbara.
Johansen, S. (1983). An extension of Cox's regression model. *Int. Statist. Rev.* **51,** 165–174.
Kalbfleisch, J. D. and Prentice, R. L. (1980). *The statistical analysis of failure time data.* Wiley, New York.
Karr, A. F. (1987). Maximum likelihood estimation in the multiplicative intensity model via sieves. *Ann. Statist.* **15,** 473–490.
McLeish, D. L. (1974). Dependent central limit theorems and invariance principles. *Ann. Probability* **2,** 620–628.
Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *Ann. Statist.* **11,** 453–466.

Received November 1990, in final form September 1992

Debasis Sengupta, Applied Statistics, Surveys & Computing Division, Indian Statistical Institute, 203, B.T. Road, Calcutta 700 035, India.