Centrum voor Wiskunde en Informatica
Centre for Mathematics and Computer Science

N. Keiding, R.D. Gill

Random truncation models and Markov processes

Department of Mathematical Statistics Report MS-R8817 November

The Centre for Mathematics and Computer Science is a research institute of the Stichting Mathematisch Centrum, which was founded on February 11, 1946, as a nonprofit institution aiming at the promotion of mathematics, computer science, and their applications. It is sponsored by the Dutch Government through the Netherlands Organization for the Advancement of Research (N.W.O.).

Random Truncation Models and Markov Processes

N. Keiding
Statistical Resaerch Unit, University of Copenhagen,
Blegdamsvej 3, DK-2200 Copenhagen N, Denmark

R.d. Gill
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

1980 Mathematics Subject Classification: Primary: 62M05, 62G05. Secondary: 60G55, 60G44.
Key Words & Phrases: counting processes, delayed entry, inference in stochastic processes, intensity function, left truncation, nonparametric maximum likelihood, product integral, survival analysis.
Note: This is a revision of report MS-R8702

Random left truncation is modelled by the conditional distribution of the random variable X of interest, given that it is larger than the truncating random variable Y; usually X and Y are assumed independent. The present paper is based on a simple reparametrization of the left truncation model as a three–state Markov process. The derivation of a nonparametric estimator of a distribution function under random truncation is then a special case of results on the statistical theory of counting processes by Aalen and Johansen. This framework also clarifies the status of the estimator as nonparametric maximum likelihood estimator, and consistency, asymptotic normality and efficiency may be derived directly as special cases of Aalen and Johansen's general theorems and later work. Besides improving the interpretability of the results and considerably shortening proofs and derivations, the present framework also allows several generalizations.

Report MS-R8817
Centre for Mathematics and Computer Science
P.O. Box 4079, 1009 AB Amsterdam, The Netherlands

1. Introduction. As has been known since Halley (1693), the construction of a life table involves following persons from an entrance age to an exit age and registering whether exit is due to death or end of observation for other reasons (censoring, in modern terminology). Kaplan and Meier (1958) initiated the modern mathematical-statistical analysis of the life table in continuous time, or equivalently, the nonparametric estimation of a distribution function from right–censored observations. Kaplan and Meier also showed that their 'product–limit' estimator was the method of choice under delayed entry, or left truncation, even though this portion of their paper has escaped the attention of many later authors. Nevertheless the practical use of life table and product–limit methods under left truncation has flourished. For recent biostatistical applications of the theory of this paper see Keiding, Bayer and Watt–Boolsen (1987) and Lagakos, Barraj and De Gruttola (1987).

A different empirical motivation for the study of nonparametric estimation under random truncation comes from astronomy, as recently summarized by Woodroofe (1985). In fact, a heuristic maximum likelihood argument for the product–limit estimator under random truncation was given by Lynden–Bell (1971).

A third apparently independent line of work on this estimator concerns estimation of the residual in truncated regression, cf. Bhattacharya, Chernoff and Yang (1983), Tsui, Jewell and Wu (1987) and Bickel and Ritov (1987).

Following Woodroofe (1985) our basic setup is that of n i.i.d. replications of the conditional distribution, given $Y<X$, of a pair of independent random variables Y and X with distribution functions G and F, of which nonparametric estimators are sought. (Obviously, this problem is ill posed unless ess. inf. $Y \leq$ ess. inf. $X <$ ess. sup. $Y \leq$ ess. sup. X, which will be assumed throughout).

The purpose of this paper is to demonstrate how an embedding of the basic nonparametric estimation problem into a simple Markov process model not only provides a considerably simpler and much more intuitive approach to a number of issues in the current literature, but also paves the way for several new results. (A practical application of a slightly more general Markov process model, containing nonparametric estimation with delayed entry, was given by Aalen, Borgan, Keiding and Thormann (1980). A rather similar Markov process model for doubly censored data was recently suggested by Samuelsen (1988).)

In the Markov process framework, *the product–limit estimators* of G and F may be derived as a direct consequence of results by Aalen and Johansen (1978), thereby placing the form and properties of the estimator in a natural perspective. (A different important perspective is that of selection bias models, cf. Vardi (1985)). It should be remarked here that another model of delayed entry, obtained without conditioning on the event that the entry time is less than the failure time, was studied by Aalen (1975, 1978), see Andersen, Borgan, Gill and Keiding (1988) for a detailed discussion of 'filtering'.

The consistency and asymptotic normality of the estimators were studied by Woodroofe (1985, cf. 1987), who did not identify the asymptotic covariance structure, and by Wang, Jewell and Tsai (1986). Our results directly derive from those of Aalen and Johansen (1978), supplemented by Gill (1983) for the edge effects. (Although the latter references as well as the general framework of Aalen (1975, 1978) explicitly account for censoring and therefore pave the way for extension of our approach to left truncated and right censored data, we do not carry through this program here. Tsai, Jewell and Wang (1987) gave some results in this direction, using a classical approach).

As an illustration of the power of the methods, we provide a simple direct derivation of the variance of the asymptotic normal distribution of $\hat{\alpha}$, where α=P{Y<X}. Such results (in slight disagreement with ours) were conjectured earlier by Chao (1987) based on a complicated influence function approach.

The maximum likelihood properties of the product–limit estimators were in particular discussed by Wang et al. (1986) and Wang (1987a, b), using Vardi's results on selection bias models as main framework. Wang's discussion of the marginal nonparametric maximum likelihood property of the product limit estimator is long and complicated. We can be brief here, because our embedding makes the general results by Johansen (1978) on nonparametric estimation in continuous–time, finite-state Markov processes directy available. Also, our discussion highlights the natural existence condition of 'no empty inner risk sets'; in the earlier literature reference was made to a general but not very intuitive condition of Vardi (1985).

Little has been said so far in the literature on efficiency of the product–limit estimator, except for some complicated algebra by Huang and Tsai (1986) in the restrictive case where ess. inf. $Y <$ ess. inf. X, ess. sup. $Y <$ ess. sup. X. We give specific directions as to which paths to follow using functional differentation to derive efficiency results directly from the efficiency of the empirical marginals of the conditional distributions of X and Y given $Y<X$ (Reeds, 1976; Gill, 1988; van der Vaart, 1988).

In a final section we show how the present framework also covers an estimation problem in steady–state renewal processes studied by Winter and Földes (1986).

2. Interpretation of random truncation models in a simple Markov process model. Woodroofe (1985) surveyed the problem of nonparametric estimation of the distributions G and F of independent, positive random variables Y and X when sampling from the conditional distribution of (X,Y) given $Y \leq X$. Define the cumulative hazard functions

$$\Gamma(y) = \int_0^y dG(y)/[1-G(y-)], \quad \Phi(x) = \int_0^x dF(x)/[1-F(x-)] \ .$$

Let $a_G < b_G$ be the essential infimum and supremum of G so that (a_G, b_G) is the interior of the convex support of G; define a_F and b_F similarly.

We assume throughout that Y and X have no common atoms; in particular $P\{Y=X\}=0$ so that

$$\alpha = P\{Y \leq X\} = P\{Y < X\}$$

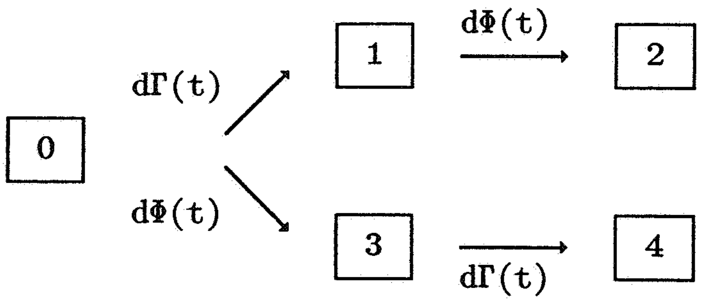
which is assumed to be positive; this is then equivalent to assuming $a_G<b_F$. Towards the end of this section we briefly mention the modifications necessary when $P\{Y=X\}>0$. To force identification we suppose also $a_G \leq a_F$ and $b_G \leq b_F$.

Define a stochastic process $U=\{U(t),\ t\in[0,\infty]\}$ by

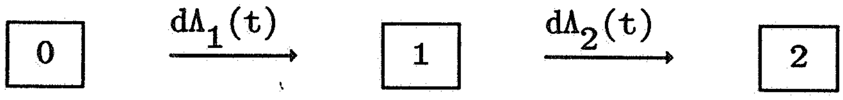
$$\begin{array}{ll} U(t) = 0 & \text{when} \quad t < X \wedge Y \\ U(t) = 1 & \text{when} \quad Y \leq t < X \\ U(t) = 2 & \text{when} \quad Y < X \leq t \\ U(t) = 3 & \text{when} \quad X \leq t < Y \\ U(t) = 4 & \text{when} \quad X \leq Y \leq t \ . \end{array}$$

It is seen that U is equivalent to (Y,X) and furthermore that the conditional distribution of (Y,X) given $(Y<X)$ is equivalent to the conditional distribution of U given $U(\infty)=2$.

PROPOSITION 2.1. U is a Markov process with $U(0)\equiv 0$ and intensities given by the diagram



In the conditional distribution given $U(\infty)=2$ (that is, $Y<X$), U is again a Markov process given by the diagram



where $\Lambda_2 = \Phi$ whereas

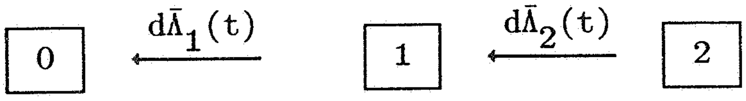
$$d\Lambda_1(t) = d\Gamma(t)\,\frac{P_{12}(t,\infty)}{P_{02}(t-,\infty)} = \frac{d\Gamma(t)}{P\{Y<X \mid X\geq t, Y\geq t\}}$$

where $P_{ij}(t,u)$ are the transition probabilities in the original Markov process.

Proof. Using product–integral formalism, Johansen (1978, 1987) defined finite–state, nonhomogeneous Markov processes from general (not necessarily continuous) intensity measures. That the conditional process given $U(\infty)=2$ is Markov with the stated intensity measures is well known and easily seen by direct calculation.

□

Consider now the time–reversed conditional Markov process $U(t)$ on $[0,\infty]$ with time running backwards and $U(\infty) \equiv 2$:



The following proposition is a standard result in Markov processes and is easily proved directly.

PROPOSITION 2.2. Consider a Markov process with states $\{0,1,2\}$ defined from intensity measures Λ_1 and Λ_2 as in Proposition 2.1. The intensities of the backwards Markov process (the "backwards intensities") are given by

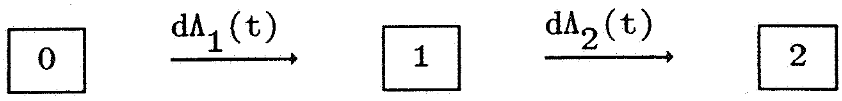
$$d\bar{\Lambda}_i(t) = d\Lambda_i(t)\,\frac{P_{2,i-1}(\infty,t-)}{P_{2,i}(\infty,t)}\ ,\quad i=1,2. \qquad \square$$

Define the (left–continuous) backwards cumulative hazard

$$\bar{\Gamma}(t) = \int_{[t,\infty)} \frac{dG(u)}{G(u)}\ ,$$

then an easy calculation from Proposition 2.2 gives $d\bar{\Lambda}_1(t) = d\bar{\Gamma}(t)$, which of course also follows directly by symmetry of time.

We now want to ask a converse question informally formulated as follows: given a Markov process



under which circumstances do there exist distribution functions G and F generating this in the above way, that is, such that

$$\frac{d\Gamma(t)}{P\{Y<X \mid X\geq t, Y\geq t\}} = d\Lambda_1(t) \ ,$$

$$dF(t)/[1-F(t-)] = d\Lambda_2(t) \ ,$$

where Y and X are independent with distribution functions G and F.

As preparation, we consider arbitrary integrated intensity measures Λ on $[0,\infty]$; define the minimal convex support Σ (which is an open, half–open or closed interval) as the smallest convex set such that $\Lambda(\Sigma^c) = 0$. Define c to be a termination point of Λ if either $\Lambda(\{c\}) = 1$ or $\Lambda(c-\epsilon,c] = \infty$ for all $\epsilon>0$, but not both.

PROPOSITION 2.3. Let Λ be an intensity measure on $(0,\infty)$ with minimal convex support with endpoints $a<b$. Λ corresponds to a probability measure if and only if Λ is finite on $[a,b-\epsilon)$ for all $\epsilon>0$ and it has one and only one termination point, which is the essential supremum b.

□

PROPOSITION 2.4. Let $U=(U(s), 0\leq s\leq\infty)$ be a Markov process with state space $\{0,1,2\}$, intensity measures Λ_i: $i-1 \to i$, $i=1,2$, all other transitions having zero intensity and $P\{U(0)=0\}=P\{U(\infty)=2\}=1$. Define $\bar{\Lambda}_i$ (the backwards intensity measure from i to i–1) by

$$d\bar{\Lambda}_i(t) = d\Lambda_i(t)\,\frac{P\{U(t-)=i-1\}}{P\{U(t)=i\}}$$

and assume that Λ_1, Λ_2 as well as $\bar{\Lambda}_1, \bar{\Lambda}_2$ (with time running backwards) correspond to probability measures.

Then there exist distribution functions F and G given by

$$1-F(x) = \prod_0^x (1-d\Lambda_2)\ , \quad G(y) = \prod_{(y,\infty)} (1-d\bar{\Lambda}_1)$$

such that the Markov process corresponds to the left truncation model specified by the conditional distribution of independent random variables Y and X on $(0,\infty)$ with distribution functions G and F, given $Y<X$. These are the unique G and F subject to $a_G \le a_F$, $b_G \le b_F$.

<u>Proof</u>. The conditions directly imply that F and G are well defined distribution functions. We need to check that the construction in Proposition 2.1 of a Markov process from these F and G leads us back to the integrated intensities Λ_1 and Λ_2. Let

$$\Phi = \int dF/(1-F\_),\ \Gamma = \int dG/(1-G\_)\ .$$

Then immediately $\Phi = \Lambda_2$ and it is required only to show that

$$\frac{d\Gamma(t)}{P\{Y<X \mid Y\geq t, X>t\}} = d\Lambda_1(t)\ .$$

Starting from the left hand side we have

$$\frac{dG(t)/[1-G(t-)]}{\int_{t\leq y} \frac{1-F(y)}{1-F(t)} \frac{dG(y)}{1-G(t-)}}$$

$$= \frac{dG(t)/G(t)}{\int_{t\leq y} \frac{[1-F(y)]/[1-F(t)]}{G(t)/G(y)} \frac{dG(y)}{G(y)}}$$

$$= \frac{d\bar{\Lambda}_1(t)}{\int_{t\leq y} \frac{\Pi_{(t,y]}(1-d\Lambda_2(u))}{\Pi_{(t,y]}(1-d\bar{\Lambda}_1(u))} d\bar{\Lambda}_1(y)}$$

$$= \frac{d\bar{\Lambda}_1(t)}{\int_{t\leq y} \frac{P\{U(y+)=1|U(t+)=1\}}{P\{U(t+)=1|U(y+)=1\}} d\bar{\Lambda}_1(y)}$$

Since Λ_2 has its only termination point at its essential supremum, it has no internal increments of size 1 so that

$$P\{U(t+) = 1\}P\{U(y+) = 1\} > 0 \quad \Rightarrow \quad P\{U(t+) = 1, U(y+) = 1\} > 0$$

and we may then reduce the integrand to

$$P\{U(y+) = 1\} \;/\; P\{U(t+) = 1\}$$

and write the expression as

$$\frac{d\bar{\Lambda}_1(t)P\{U(t+)=1\}}{\int\limits_{t\leq y} P\{U(y+)=1\}d\bar{\Lambda}_1(y)}$$

$$= \frac{d\bar{\Lambda}_1(t)P\{U(t+)=1\}}{P\{U(t-)=0\}} = d\Lambda_1(t) \ . \qquad \square$$

2a. The Markov process parametrization (Λ_1, Λ_2) *and the truncation model parametrization (G, F).*

Note that while $\Lambda_2=\Phi$ corresponds to the distribution function $F_2=F$, Λ_1 corresponds to the distribution function

$$\begin{aligned} F_1(y) &= \prod_0^y (1-d\Lambda_1(u)) \\ &= \int_0^y [1-F(s)]dG(s) / \int_0^\infty [1-F(s)]dG(s) \\ &= \alpha^{-1} \int_0^y [1-F(s)]dG(s) \qquad (2.1) \end{aligned}$$

and that G may be recovered from F_1 and $F_2=F$ by the inverse relation

$$G(y) = \alpha \int_0^y [1-F_2(s)]^{-1} dF_1(s)$$

since

$$\alpha = \int_0^\infty [1 - F(s)]dG(s) = 1/\int_0^\infty [1 - F_2(s)]^{-1}dF_1(s) \ .$$

The key point of this paper is the interplay between these two alternative representations: the 'random truncation model' specified by G and F and the Markov process model specified by F_1 and F_2.

The second representation of α may be taken as starting point for a further discussion of the condition that $\bar{\Lambda}_1$ correspond to a probability measure (in the presence of the other conditions); one can prove that this happens if and only if

$$\int_0^\infty (1 - F_2(s))^{-1}dF_1(s) < \infty \ ;$$

i.e. that when we calculate "α", we find $\alpha>0$.

Note that if Λ_1 and Λ_2 are discrete and $P\{U(0)=0\}=P\{U(\infty)=2\}=1$, the condition that they correspond to probability measures implies that $\bar{\Lambda}_1$ and $\bar{\Lambda}_2$ do too.

2b. A third parametrization by the marginal conditional distributions.

Woodroofe (1985, Theorem 1) showed that the left truncation model (given by distribution function G and F) may be parameterized by the marginal conditional distributions G^* and F^* given by

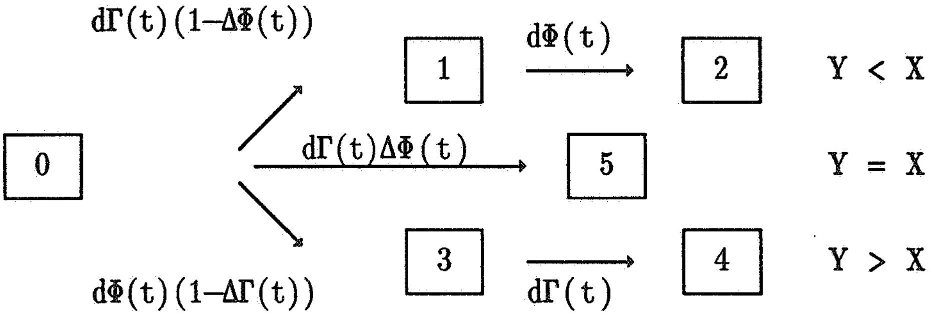
$$G^*(y) = P\{Y \le y | Y < X\} \ , \ F^*(x) = P\{X \le x | Y < X\} \ ,$$

again provided $a_G \le a_F$, $b_G \le b_F$.

In fact if $\tilde{F}$ and $\tilde{G}$ are any two distribution functions without common atoms such that $d\tilde{F}/(\tilde{G}_{\_}-\tilde{F}_{\_})$ and $d\tilde{G}/(\tilde{G}-\tilde{F})$ are respectively a forwards and backwards intensity measure (so $\tilde{F}\leq\tilde{G}$!) corresponding to probability distributions F and G, then $\tilde{F}$ and $\tilde{G}$ are the marginals of $X,Y|Y<X$ in the left truncation model with parameters F and G. This reparametrization will turn out to be convenient when discussing efficiency of the estimators. (A very general discussion of 'order conditioned independence' of random variables was recently given by Kellerer (1986)).

2c. The possibility of ties between Y *and* X.

Much of the preceding and following theory can be quite easily extended to the case when F and G have common jumps – i.e. ties between the X's and the Y's are possible. The Markov model of Proposition 2.1 has to be extended with a third route corresponding to $X=Y$ and the intensities correspondingly modified:



thus $U(t)=5$ when $0\leq X=Y\leq t$.

The theory is now also different according to whether we observe replicates of X,Y given $X<Y$ or given $X\leq Y$. The first case is much easier to handle since conditional

on $U(\infty)=2$, U remains a Markov process; in the second case, when we must condition on $U(\infty)=2$ or 5, U is no longer Markov.

3. Estimation. In this section we assume the distributions G and F to be continuous with support $(0,\infty)$; then the corresponding integrated intensities Γ and Φ are also continuous. By Y^*, X^* we denote random variables with the conditional distribution of Y, X given $Y<X$, and $G^*, F^*, \Gamma^*, \Phi^*$ denote distribution functions and integrated intensities in this distribution.

We assume that a sample of n independent identically distributed replications $(Y_1^*,X_1^*),\dots,(Y_n^*,X_n^*)$ of (Y^*,X^*) is observed. Corresponding to (Y_i^*,X_i^*), $i=1,\dots,n$, we construct (conditional) Markov processes U_i as in Proposition 2.1, which yields the following interpretation of naturally defined counting processes

$$\begin{aligned} N_1(t) &= \#\{Y_i^* \le t\} \\ &= \#\{\text{jumps by } U_1,\dots,U_n \text{ from } 0 \text{ to } 1 \text{ in } [0,t]\} \end{aligned}$$

$$\begin{aligned} N_2(t) &= \#\{Y_i^* < X_i \le t\} \\ &= \#\{\text{jumps by } U_1,\dots,U_n \text{ from } 1 \text{ to } 2 \text{ in } [0,t]\} . \end{aligned}$$

With respect to the self–exciting filtration the bivariate counting process $N(t)=(N_1(t),N_2(t))$ has compensator $A(t)= (A_1(t), A_2(t))$ given by

$$A_1(t) = \int_0^t V_1(t)d\Lambda_1(t) \ , \ A_2(t) = \int_0^t V_2(t)d\Phi(t)$$

where we have used the fact (Proposition 2.1) that $\Lambda_2=\Phi$ and where

$$V_1(t) = \#\{Y_i \ge t\} \ , \ V_2(t) = \#\{Y_i < t \le X_i\} \ ;$$

define also $J_i(t)=I\{V_i(t) > 0\}$, $i=1,2$.

3a. Estimation of the distribution of X.

According to standard methodology for statistical analysis of counting processes (Aalen, 1975, Section 5D, 1978; Aalen and Johansen, 1978) we use as estimator of the integrated intensity $\Phi(t)$ the Nelson–Aalen estimator

$$\hat{\Phi}(t) = \int_0^t \frac{J_2(u)}{V_2(u)} dN_2(u) = \sum_{i=1}^n \frac{I\{X_i^* \le t\}}{V_2(X_i)} \ .$$

It is then a basic result in the statistical analysis of counting processes that, defining

$$\tilde{\Phi}(t) = \int_0^t J_2(u) d\Phi(u) \ ,$$

the process $\hat{\Phi}(t)-\tilde{\Phi}(t)$ is a mean zero, square integrable martingale with predictable variation process given by

$$<\hat{\Phi} - \tilde{\Phi}>(t) = \int_0^t \frac{J_2(u)}{V_2(u)} d\Phi(u) \ .$$

(Note that if Φ has discrete components, the factor $d\Phi(u)$ should here be replaced by $[1-\Delta\Phi(u)]d\Phi(u)$). These properties imply the unbiasedness result

$$E(\hat{\Phi}(T)) = E(\tilde{\Phi}(T)) \tag{3.1}$$

for any stopping time T (both sides may be ∞) and suggest the estimator

$$\hat{\tau}(t) = \int_0^t \frac{J_2(u)}{[V_2(u)]^2}\, dN_2(u)$$

or a 'Greenwood–formula' – modification, acknowledging the discrete nature of $\hat{\Phi}$:

$$\hat{\tau}_G(t) = \int_0^t \frac{J_2(u)\,(V_2(u)-1)}{[V_2(u)]^3}\, dN_2(u)$$

of the mean squared error function $\tau(t)=E[<\hat{\Phi}-\tilde{\Phi}>(t)]$.

Let us take a concrete look at the process $J_2(u)=I\{V_2(u)>0\}$. Since we have assumed that ess inf X=ess inf Y=0, we will with probability one have $Y^*_{(1)}>0$, $Y^*_{(1)}$ as usual denoting the smallest Y^*, so that $V_2(u)=0$ on a proper interval $[0,Y^*_{(1)}]$. It may happen that $V_2(u)=0$ on further intervals $(U_1,Z_1],\ldots,(U_k,Z_k]$, $Z_k<X^*_{(n)}$, (it certainly becomes 0 for $u>X^*_{(n)}$). The serious problem is that in this case of 'empty inner risk sets' $\Delta\hat{\Phi}(U_i)=\Delta N_2(U_i)/V_2(U_i)=1$, "using up" the probability mass in the middle of the observation interval.

The interest in the literature has focussed on estimating not the integrated intensity of X but rather its distribution function F or (equivalently) its survivor function 1–F. The formal Aalen and Johansen (1978, Theorem 3.2) answer is to use the product–limit (or generalized Kaplan–Meier) estimator

$$1-\hat{F}(t) = \prod_{[0,t]} [1-d\hat{\Phi}(u)]$$

where the product integral reduces to the finite product

$$\prod_{i=1}^{n} \left(1 - \frac{I\{X^*_i \le t\}}{V_2(X^*_i)}\right) .$$

Unbiasedness and mean square error results derive from the fact that defining

$$1 - \tilde{F}(t) = \prod_{[0,t]} [1 - d\Phi^{*}(u)]$$

we have that

$$\frac{1-F^{*}(t)}{1-\tilde{F}(t)} - 1 = \int_0^t \frac{1-F^{*}(u-)}{1-\tilde{F}(u)} \, d[\hat{\Phi}(u) - \tilde{\Phi}(u)]$$

is a zero–mean, local square integrable martingale with predictable squared variation process given by

$$< \{1 - \hat{F}(t)\}/\{1 - \tilde{F}(t)\} - 1 >$$

$$= \int_0^t \left[\frac{1-\hat{F}(u-)}{1-\tilde{F}(u)}\right]^2 \quad d<\hat{\Phi} - \tilde{\Phi}>(u)$$

$$= \int_0^t \left[\frac{1-\hat{F}(u-)}{1-\tilde{F}(u)}\right]^2 \frac{J_2(u)}{V_2(u)} \, d\Phi(u) \ .$$

Hence for any bounded stopping time T we get

$$E\left[\frac{1-\hat{F}(T)}{1-\tilde{F}(T)}\right]^2 = 1 \ .$$

Taking into account the discrete nature of the estimator $1-\hat{F}$, the squared variation of $(1-\hat{F})/(1-\tilde{F})$ may be estimated by

$$\int_0^t J_2(u)\,(V_2(u)-1)V_2(u)^{-3}\,dN_2(u)\ ,$$

and it follows that a natural estimate of the covariance function of $1-\hat{F}$ is given by Greenwood's formula (cf. Meier (1975))

$$\text{cov}(1-\hat{F}(s),1-\hat{F}(t)) = \{1-\hat{F}(s)\}\{1-\hat{F}(t)\}\int_0^{s\wedge t} J_2(u)\,[V_2(u)\{V_2(u)-1\}]^{-1}dN_2(u)\ .$$

Note that since $d\hat{\Phi}(U_i)=1$, $i=1,\ldots,k+1$, and in particular $d\hat{\Phi}(U_1)=1$, the estimator $1-\hat{F}(t)=0$ for $t\geq U_1$. This is a serious problem if there exist values of Y_j^* (and hence X_j^*) larger that U_1 because the estimator of the distribution of X will then be supported by a proper subset consisting of the smaller observed X_j^*. Woodroofe (1985, p. 168) recognized the problem and suggested an ad hoc mending. We shall see in Section 4 below that the formal nonparametric maximum likelihood estimator does not exist in this case. Perhaps the cumulative hazard is more appropriate than the distribution function for communicating the results of the estimation since the hazard at any time x is the same for all conditional distributions of X given $X>x_0$, $x_0<x$.

3b. Estimation of the distribution of Y.

By reversing time it is immediate that the backwards integrated hazard $\bar{\Gamma}(t)$ may be estimated by a <u>backwards Nelson–Aalen estimator</u>; similarly $G(t)$ may be estimated by a generalized backwards Kaplan–Meier estimator. (Care should be taken regarding left or right continuity etc.)

The various complications are exactly as for the estimation of the distribution of X, and moreover, there are complications in estimating both distributions or not at all. In particular, there is no information in the sample on the distribution of Y on $[X_{(n)}^*, \infty)$.

Alternatively one might start from the Markov process representation of Section 3a, then estimators of the integrated intensity Λ_1 and the corresponding distribution function F_1 are immediately given as

$$\hat{\Lambda}_1(t) = \int_0^y \frac{J_1(u)}{V_1(u)} dN_1(u) = \sum_{i=1}^n \frac{I\{Y_i^* \le y\}}{\#\{Y_i^* \ge y\}}$$

and the corresponding product–limit estimator $1-\hat{F}_1$ where $\hat{F}_1$ is nothing but the empirical distribution function of the Y_i^*. Since the martingales $\hat{\Lambda}_1-\tilde{\Lambda}_1$ and $\hat{\Lambda}_2-\tilde{\Lambda}_2$ are orthogonal by the general theory of statistical analysis of counting processes, we further have the important property of approximate independence of $\hat{\Lambda}_1$ and $\hat{\Lambda}_2$; this property will be crucial for the asymptotic theory of Section 4.

Since $\hat{F}_1$ estimates

$$F_1(y) = \alpha^{-1} \int_0^y [1-F(s)]\, dG(s) \qquad (3.2)$$

one could then apply the inversion

$$\tilde{G}(y) = \tilde{\alpha} \int_0^y \{1-\hat{F}(s)\}^{-1} d\hat{F}_1(s) \ , \ \tilde{\alpha} = \int_0^\infty \{1-\hat{F}(s)\}^{-1} d\hat{F}_1(s) \ ;$$

however it is not immediate that $\tilde{G}$ equals the simple product–limit estimator $\hat{G}$ of the time–reversal approach and that $\tilde{\alpha}=\hat{\alpha}=\int \hat{G} d\hat{F}$. This is however a direct consequence of the propositions on Markov processes of Section 2 and the transformation invariance of maximum likelihood estimators which we discuss in the next section.

4. Nonparametric maximum likelihood estimation of (G,F). The purpose of this section is to show that $(\hat{G},\hat{F})$ is the nonparametric maximum likelihood estimator (NPMLE) of (G,F), and that this fact is a direct consequence of the embedding of the left truncation model into the Markov process model, for which results on NPMLE were provided by Johansen (1978). First we discuss the easier result that $\hat{F}$ is a conditional NPMLE of F given $Y_1^*,\ldots,Y_n^*$.

4a. Conditional nonparametric maximum likelihood estimation of F *given* $Y_1^*,\ldots,Y_n^*$.

As an introduction consider the factorisation of the ("full") likelihood

$$\text{lik}\,(G,F) = \alpha^{-n} \prod_i dG(Y_i)dF(X_i)$$

into the marginal likelihood of (G,F) based on $(Y_1^*,\ldots,Y_n^*)$ and the conditional likelihood of F given $(Y_1^*,\ldots,Y_n^*)$:

$$\text{lik}\,(G,F) = \text{marg.lik}_{\underset{\sim}{Y}^*}(G,F)\ \text{cond.lik}_{\underset{\sim}{X}^*|\underset{\sim}{Y}^*}(F),$$

where in particular

$$\text{cond.lik}_{\underset{\sim}{X}^*|\underset{\sim}{Y}^*}(F) = \prod_{i=1}^{n} \frac{dF(X_i^*)}{1-F(Y_i^*)} .$$

As in the derivations of the NPMLE for censored data by Kaplan and Meier (1958) and Johansen (1978), it is seen that the candidates for the maximiser $\hat{F}$ of the conditional likelihood must have support $\subseteq\{X_1^*,\ldots,X_n^*\}$, and for such F we have the simple combinatorial result

$$\text{cond.lik}(F) \quad = \quad \frac{dF(X^*_{(1)})}{[1-F(X^*_{(1)}-)]^{V_2(X^*_{(1)})}} \; \frac{dF(X^*_{(2)})}{[1-F(X^*_{(2)}-)]^{V_2(X^*_{(2)})-V_2(x^*_{(1)})+1}}$$

$$\cdots \; \frac{dF(X^*_{(n)})}{[1-F(X^*_{(n)}-)]^{V_2(X^*_{(n)})-V_2(X^*_{(n-1)})+1}}$$

$$= \quad \prod_{i=1}^{n} d\Phi(X^*_{(i)})[1-d\Phi(X_{(i)})]^{V_2(X^*_{(i)})-1}$$

using $dF(X_{(i)})=F(X_{(i)})-F(X_{(i-1)})$ and the definition $d\Phi=dF/(1-F_-)$. Recall that a discrete intensity measure Ψ with support contained in n points $a_1<\ldots<a_n$ corresponds to a probability measure if and only if $0\leq d\Psi(a_i)<1$, $i=1,\ldots,n-1$, $d\Psi(a_n)=1$. The maximisation problem is then trivial: if and only if $V_2(X_{(i)})>1$, $i=1,\ldots,n-1$ (no empty inner risk sets), the solution exists and is given by

$$d\hat{\Phi}(X_i) = 1/V_2(X_i)$$

or exactly the Nelson–Aalen estimator $\hat{\Phi}$ of Φ. By transformation invariance of maximum likelihood estimators (the relevant transformation here being the product integral) it follows that the conditional NPMLE of the survivor function $1-F=\prod(1-d\Phi)$ is the product–limit estimator

$$1-\hat{F} = \prod(1-d\hat{\Phi})$$

studied in Section 3.

This solution as well as the condition of no empty inner risk sets provide explicit examples of Theorem 2 and condition (2.10) of Vardi (1985) (as pointed out earlier by

Wang et al. (1987) and Wang (1987a)), since the left truncation model may be interpreted as a selection bias model.

4b. $\hat{F}$ is not always the unconditional MLE.

Before proceeding to the discussion of $(\hat{G},\hat{F})$ as NPMLE in the full likelihood, let us remark that if G cannot vary freely, it is easily seen that the NPMLE $\tilde{F}$ of F may differ from $\hat{F}$. Indeed Vardi (1985) showed that if G is known,

$$\tilde{F}(X^*_{(i)}) = \sum_{j=1}^{i} G(X^*_{(j)})^{-1} / \sum_{j=1}^{n} G(X^*_{(j)})^{-1}$$

(a 'weighted empirical distribution function'), and Wang (1987b) generalized this analysis by noticing that if G varies across a parametric family $G=\{G_\theta;\ \theta\in\Theta\}$, then the NPMLE of F is obtained from $\tilde{F}$ by replacing G by $G_{\hat{\theta}}$, where $\hat{\theta}$ is the MLE derived from the conditional distribution of Y^* given X^*. (It is a corollary of this analysis that when F varies freely, X^* is "M–ancillary" (Barndorff–Nielsen, 1980) w.r.t. θ.)

These results strongly suggest that the NPMLE of F in the full model may be similarly obtained by replacing G by $\hat{G}$ in $\tilde{F}$. However concrete calculations along these lines, as provided by Wang (1987a), are combinatorially involved, and we show in the next subsection that the independent parametrisation provided by the Markov process representation of the left truncation model furnishes an immediate answer.

4c. NPMLE in the left truncation model.

To the original left truncation model given by continuous distribution functions G, F varying freely over all distributions with convex support $(0,\infty)$ except that $dGdF\equiv0$ we define a conditional Markov process model as specified in Proposition 2.1 above.

NPML estimation in such Markov process models was studied by Johansen (1978) who showed that the NPMLE exists if the model is extended to allow for arbitrary increasing right–continuous integrated intensity functions with increments ≤ 1. Moreover, the likelihood is of the form

$$\prod_i \prod_t d\Lambda_i(t)^{dN_i(t)} (1 - d\Lambda_i(t))^{V_i(t) - dN_i(t)} \ .$$

It follows directly that the NPMLE of Λ_1 and Λ_2 are given by $\hat{\Lambda}_1$ and $\hat{\Lambda}_2$ as specified in Section 3.

Now the estimates $\hat{\Lambda}_1$ and $\hat{\Lambda}_2$ are themselves integrated intensities defining a Markov process

$$\boxed{0} \xrightarrow{\hat{\Lambda}_1} \boxed{1} \xrightarrow{\hat{\Lambda}_2} \boxed{2}$$

and by the results of Section 2 one may recover distribution functions F^0 and G^0 corresponding to a left truncation model, if and only if there is no inner jump of size one of $\hat{\Lambda}_i$ (i=1,2) – this could only happen for $\hat{\Lambda}_2$ because $\hat{\Lambda}_1$ corresponds to an ordinary empirical distribution, with jumps of size j^{-1}, $j=n,...,1$, in that order. When F^0 and G^0 exist, they coincide with $\hat{F}$ and $\hat{G}$ by Proposition 2.4 applied to $\hat{\Lambda}_i$, $i=1,2$, and it furthermore follows from the transformation invariance of maximum likelihood estimators and the definition of backwards intensities that $\tilde{G}$ of Section 3b equals $\hat{G}$.

Finally, to show that <u>an NPMLE in the left truncation model</u> does not exist if there are inner jumps of size 1 in the NPMLE for the Markov model we now only need to remark that one can then make the ("discrete") likelihood function in the left truncation model arbitrarily close to the maximum likelihood in the Markov model, without however being able to achieve this value.

As a corollary of the NPMLE property of $1-\hat{F}$ and $\hat{G}$ we remark that the NPMLE of

$$\alpha = P\{Y < X\} = P\{Y \le X\} = \int_0^\infty G(u)dF(u) = \int_0^\infty [1-F(u)]dG(u)$$

is

$$\hat{\alpha} = \int_0^\infty \hat{G}(u)d\hat{F}(u) = 1/\int_0^\infty \{1-\hat{F}(u)\}^{-1}d\hat{F}_1(u) \ .$$

5. Asymptotic results. In this section it is assumed throughout that the distributions of G and F are continuous with support $(0,\infty)$ and integrated hazards Γ and Φ.

For the asymptotic theory we shall use two alternative parametrizations: that in terms of $1-F$ and G with associated product limit estimators $1-\hat{F}$ and $\hat{G}$, and that from the counting process integrated intensities Λ_1 and Λ_2 (with corresponding d.f.s F_1 and $F_2=F$) with associated estimators $\hat{\Lambda}_1, \hat{\Lambda}_2, \hat{F}_1, \hat{F}_2$.

We need both parametrizations in the asymptotic theory. Delicate tightness problems near 0 and ∞ have been handled for product–limit estimators (cf. Gill, 1983, Ying 1987) so that here the first approach is preferable, while the second has an advantage for computation of covariances because of the orthogonality and hence asymptotic independence of the martingales $\hat{\Lambda}_1-\tilde{\Lambda}_1$ and $\hat{\Lambda}_2-\tilde{\Lambda}_2$. Thus $1-\hat{F},\hat{G}$ are used for establishing the limit theorems and the parameters of the asymptotic distribution of each estimator separately. Calculation of parameters in the joint asymptotic distribution of estimators and functionals of these is more conveniently based upon $\hat{\Lambda}_1, \hat{\Lambda}_2$; this will be illustrated by a derivation of the asymptotic distribution of $\hat{\alpha}$.

5a. Convergence on $[\epsilon, M]$, $0 < \epsilon < M < \infty$.

As we have seen, in interpreting $\hat{\Phi}$ and $1-\hat{F}$ in a practical situation, it is rather important to take account of the fact that $d\Phi$ can really only be estimated on the interval or intervals $\{t: V_2(t) > 0\}$. In demonstrating how the counting process formulation of the left–truncation problem can be used in a very direct way to derive asymptotic distribution theory for our estimators, we shall similarly take care of this problem by first only estimating

$$\Phi^{\epsilon} = \Phi - \Phi(\epsilon) \quad \text{and} \quad 1 - F^{\epsilon} = (1-F)/\{1-F(\epsilon)\}$$

on an interval $[\epsilon, M]$ whose endpoints $t=\epsilon$, M satisfy $P\{Y<t\leq X|Y<X\} > 0$.

Let $\hat{\Phi}^{\epsilon}$, $\tilde{\Phi}^{\epsilon}$, $\hat{F}^{\epsilon}$ and $\tilde{F}^{\epsilon}$ be defined similarly to Φ^{ϵ} and F^{ϵ}, and recall our notational conventions; Y and X are independent random variables with distribution functions G and F; (Y_i^*, X_i^*) for $i=1,\ldots,n$ denote independent replicates of (Y,X) conditional on $Y<X$. Thus $P\{Y_i^*<X_i^*\}=1$ while $P\{Y<X\}=\alpha<1$. Let us also write

$$v_2(t) = E[n^{-1}V_2(t)] = P\{Y_i^* < t \leq X_i^*\} = P\{Y_i^* < t\} - P\{X_i^* < t\} = C(t),$$

in Woodroofe's notation. We have

$$v_2(t) = P\{Y < t \leq X, Y < X\}/P\{Y < X\} = G(t)[1 - F(t)]/\alpha$$

$$\geq G(\epsilon)[1 - F(M)]/\alpha > 0$$

for $\epsilon \leq t \leq M$ by the assumption that Y and X have support $(0,\infty)$.

Now $n^{-1}V_2$ is the difference between two empirical distribution functions, so by the Glivenko–Cantelli theorem

$$\| n^{-1}V_2 - v_2 \|_{\epsilon}^{M} \to 0 \text{ a.s.}$$

as $n \to \infty$, where $\| \cdot \|_{\epsilon}^{M}$ denotes the supremum norm over $[\epsilon,M]$. Thus by boundedness away from zero of v_2 we also have

$$\| v_2^{-1} - nV_2^{-1} \|_{\epsilon}^{M} \to 0 \text{ a.s.}$$

as $n \to \infty$, and $J_2 \equiv 1$ on $[\epsilon,M]$ for all sufficiently large n a.s. Thus $\tilde{\Phi}^{\epsilon} \equiv \Phi^{\epsilon}$ and $\tilde{F}^{\epsilon} = F^{\epsilon}$ on $[\epsilon,M]$ for all sufficiently large n almost surely.

With these preparations made, consistency of $\hat{\Phi}^{\epsilon}$ and $\hat{F}^{\epsilon}$ as well as weak convergence of $n^{\frac{1}{2}}(\hat{\Phi}^{\epsilon}-\Phi^{\epsilon})$ and/or of $n^{\frac{1}{2}}(\hat{F}^{\epsilon}-F^{\epsilon})$ follow immediately from standard results on the Nelson–Aalen and the product–limit estimators in the counting process literature.

PROPOSITION 5.1. We have

$$\| \hat{\Phi}^{\epsilon} - \Phi^{\epsilon} \|_{\epsilon}^{M} \xrightarrow{P} 0 \text{ and } \| \hat{F}_X^{\epsilon} - F_X \|_{\epsilon}^{M} \xrightarrow{P} 0 .$$

PROOF. Apply the inequality of Lenglart (1977) exactly as Gill (1980, 1983), cf. Andersen and Borgan (1985, Appendix).

Corollary 5.1 We have

$\| \hat{\Phi} - \Phi \|_{0}^{M} \xrightarrow{P} 0$ and $\| \hat{F} - F \|_{0}^{M} \xrightarrow{P} 0.$ □

Corollary 5.1 is obtained easily from Proposition 5.1, using

$$E\{\| \hat{\Phi} - \Phi \|_0^{\epsilon}\} \leq 2\, \Phi(\epsilon)$$

by (3.1).

<u>Corollary 5.2</u> The event of the existence of $s<t<u\leq M$ with $J_2(s) = J_2(u) = 1$, $J_2(t) = 0$, ("empty inner risk sets") is asymptotically negligible.

□

It is curious that the probabilistic result of Corollary 5.2 (obtained easily from Corollary 5.1, cf. Woodroofe (1985, p. 172)) is derived via the proof of consistency of a statistical estimator!

THEOREM 5.1 Under the stated conditions,

$$n^{\frac{1}{2}}(\hat{\Phi}^{\epsilon} - \Phi^{\epsilon}) \xrightarrow{D} W^{\epsilon} \quad \text{in} \quad D[\epsilon,M] \tag{5.1}$$

as $n \to \infty$, where W^{ϵ} is a Gaussian martingale with zero mean and variance function

$$\operatorname{var} W^{\epsilon}(t) = \int_{\epsilon}^{t} \frac{1}{v_2(s)}\, d\Phi(s); \tag{5.2}$$

we also have (in fact, jointly)

$$n^{\frac{1}{2}}(\hat{F}^{\epsilon} - F^{\epsilon}) \xrightarrow{D} (1 - F^{\epsilon}) \cdot W^{\epsilon}. \tag{5.3}$$

Furthermore, $\int_{\epsilon}^{(\cdot)} n\ V_2(s)^{-2} dN_2(s)$ is a consistent estimator (in $\| \cdot \|_{\epsilon}^{M}$) of the variance function of W^{ϵ}.

□

PROOF. We use Rebolledo's (1980) version of the martingale central limit theorem and note that the verification of Rebolledo's conditions is direct in the approach used here, establishing Glivenko–Cantelli convergence for V_2/n. This approach was used earlier by Gill (1980, Section 4.2; 1983) for the product–limit estimator in a model of random censorship (though the proof is valid in our situation too (Gill, 1980, Chapter 6)), and by Andersen and Borgan (1985, Appendix) for the Nelson–Aalen estimator in a general model, including the present one.

□

Remark. Aalen (1975, Theorem 8.2), cf. Aalen (1978, Theorem 6.4) proved weak convergence of the Nelson–Aalen estimator (in a general model containing the present one) using martingale central limit theory; Aalen and Johansen (1978, Theorem 4.6) treated the product–limit estimator in a general Markov process model (containing ours). These early results relied on uniform integrability of the random variables $nJ_2(t)/V_2(t)$ over $n=1,2,...$ and $t\in[\epsilon,M]$, which is true but requires some calculation, see Aalen (1976, Proof of Lemma 4.2 in Appendix). Indeed, subsequent developments in the theory of stochastic integrals also made Aalen and Johansen's assumption (4.1) unnecessary.

□

One should note that the counting process framework allows a direct identification from the martingale central limit theorem of the asymptotic covariance structure of each of the estimators $\hat{\Phi}$ and $\hat{S}_X$ which was already suggested by the small sample arguments of Section 3a; thus no heavy calculations as used by Wang et al. (1986) are necessary. We return below to the study of $\hat{G}$ and the joint distribution of $\hat{S}_X$ and $\hat{G}$.

5b. Convergence on $[0,\infty]$.

Since $v_2(t)>0$ for all $t\in(0,\infty)$ one can ask whether or not these results can be extended to yield weak convergence in $D[0,M]$ or $D[\epsilon,\infty]$ or even $D[0,\infty]$, cf. Woodroofe (1985, Section 6; 1987). The extension of (5.3) at the righthand endpoint of the time interval was carried out by Gill (1983) for the random censorship model under natural additional conditions, see Ying (1987) for an important supplementary result. The analogous conditions in the left truncation model are automatically satisfied. We shall use the same techinques in order to study the lefthand endpoint problem, which may be of greater practical importance.

Since $\hat{S}_X$ and S_X are both close to 1 near $t=0$, one easily discovers that the extension problem for (5.3) is hardly more difficult than that for (5.1), on which we will concentrate. Also there is no hope of making an extension unless the limiting process can be extended too; for this we need to assume (cf. (5.3)) that

$$\int_0^\epsilon \frac{d\Phi(s)}{v_2(s)}\,ds < \infty\ .$$

Now

$$\int_0^\epsilon \frac{d\Phi(s)}{v_2(s)}\,ds = \alpha\int_0^\epsilon \frac{dF(s)}{G(s)(1-F(s))^2}\ .$$

Since $F(s)\to 0$ as $s\to 0$, we have finiteness if and only if

$$\int_0^\epsilon \frac{dF(s)}{G(s)} < \infty \tag{5.4}$$

for some (and then all) $\epsilon>0$. From now on we assume (5.4) holds. We will have our required result

$$\left.\begin{array}{lll} n^{\frac{1}{2}}(\hat{\Phi}-\Phi) \xrightarrow{D} W & \text{in} & D[0,\infty) \\ & & \\ n^{\frac{1}{2}}(\hat{F}-F) \xrightarrow{D} (1-F)\cdot W & \text{in} & D[0,\infty] \end{array}\right\} \qquad (5.5)$$

where W is W^{ϵ} with $\epsilon=0$ of Theorem 5.1 if for all $\delta>0$

$$\lim_{\epsilon\downarrow 0}\ \limsup_{n\to\infty} P\{n^{\frac{1}{2}} \|\tilde{\Phi}-\Phi\|_0^{\epsilon} > \delta\} = 0 \qquad (5.6)$$

and for all $\delta>0$

$$\lim_{\epsilon\downarrow 0}\ \limsup_{n\to\infty} P\{n^{\frac{1}{2}} \|\hat{\Phi}-\tilde{\Phi}\|_0^{\epsilon} > \delta\} = 0 \qquad (5.7)$$

see Billingsley (1968; Theorem 4.2) for the basic idea here and Gill (1983; Proof of Theorem 2.1) for a similar application. We look at the easier term (5.7) first.

Now since $\hat{\Phi}-\tilde{\Phi}$ is a square integrable martingale, Lenglart's (1977) inequality gives us

$$P\{n^{\frac{1}{2}} \|\hat{\Phi}-\tilde{\Phi}\|_0^{\epsilon} > \delta\} \le \eta + P\{n <\hat{\Phi}-\tilde{\Phi}>(\epsilon) > \frac{\delta^2}{\eta}\} .$$

But

$$n < \hat{\Phi}-\tilde{\Phi} > (\epsilon) = \int_0^{\epsilon} \frac{nJ_2(s)}{V_2(s)}\, d\Phi(s) \le 2\int_0^{\epsilon} \frac{n+1}{V_2(s)+1}\, d\Phi(s) .$$

Now since $V_2(s)$ is binomially distributed,

$$E\left[\frac{n+1}{V_2(s)+1}\right] \le \frac{1}{P\{Y_i^*<s\le X_i^*\}} .$$

So

$$E(n<\hat{\Phi}-\tilde{\Phi}>(\epsilon)) \quad \le \; 2\int_0^\epsilon \frac{d\Phi(s)}{G(s)\,(1-F(s))} \;\le\; \frac{2\alpha}{(1-F(\epsilon_0))^2}\int_0^\epsilon \frac{dF(s)}{G(s)}$$

for all $\epsilon\le\epsilon_0$. Thus having assumed $\int_0^\epsilon dF(s)/G(s)<\infty$, we can prove the required result: for by Chebyshev's inequality, taking ϵ arbitrarily small, we can bound $n<\hat{\Phi}-\tilde{\Phi}>_\epsilon$ by an arbitrarily small constant with probability arbitrarily close to 1, uniformly in n; and this establishes (5.7).

As far as (5.6) is concerned, we note that

$$n^{\frac{1}{2}}\,\|\,\tilde{\Phi}-\Phi\,\|_0^\epsilon = n^{\frac{1}{2}}\int_0^\epsilon (1-J_2(s))d\Phi(s) = n^{\frac{1}{2}}\Phi(Y^*_{(1)})$$

with probability $\to 1$ as $n\to\infty$ by Corollary 5.2. It suffices therefore to show $n^{\frac{1}{2}}\Phi(Y^*_{(1)}) \xrightarrow{P} 0$ as $n\to\infty$. Now $\Lambda_1(Y^*_{(1)})$ is the minimum of n i.i.d. exponential (1) random variables, hence

$$P\{n^{\frac{1}{2}}\Phi(Y^*_{(1)})>c\} = P\{\Lambda_1(Y^*_{(1)})>\Lambda_1(\Phi^{-1}(n^{-\frac{1}{2}}c)\} = e^{-n\Lambda_1(\Phi^{-1}(n^{-\frac{1}{2}}c))} .$$

Putting $\epsilon=\Phi^{-1}(n^{-\frac{1}{2}}c)$, it suffices to prove $\Phi(\epsilon)^2/\Lambda_1(\epsilon) \rightarrow 0$ as $\epsilon\downarrow 0$. One easily verifies that

$$\int_0^\infty \frac{dF}{G} < \infty \Leftrightarrow \int_0^\epsilon \frac{d\Phi}{\Lambda_1} \downarrow 0 \text{ as } \epsilon \downarrow 0 \ .$$

But then

$$\Phi(\epsilon)^2/\Lambda_1(\epsilon) \leq \Phi(\epsilon) \int_0^\infty \frac{d\Phi}{\Lambda_1} \rightarrow 0 \text{ as } \epsilon \downarrow 0 \ ,$$

as required.

5c. Joint Weak Convergence.

By symmetry we can immediately write down weak convergence theorems for $n^{\frac{1}{2}}(\hat{\Gamma}^M - \Gamma^M)$ and/or $n^{\frac{1}{2}}(\hat{G}^M - G^M)$ and under the condition $\int_0^\infty dG/(1-F)<\infty$ drop the "M". A joint weak convergence result, e.g. for $n^{\frac{1}{2}}(\hat{F}^\epsilon - F^\epsilon)$ and $n^{\frac{1}{2}}(\hat{G}^M - G^M)$ is a little trickier. What can be argued is the following.

We certainly do have joint weak convergence of $n^{\frac{1}{2}}[n^{-1}V_1(\epsilon) - v_1(\epsilon)]$, $n^{\frac{1}{2}}[n^{-1}V_2(\epsilon) - v_2(\epsilon)]$, $n^{\frac{1}{2}}(\hat{\Lambda}_1^\epsilon - \Lambda_1^\epsilon)$ and $n^{\frac{1}{2}}(\hat{\Lambda}_2^\epsilon - \Lambda_2^\epsilon)$ in $R^2 x (D[\epsilon,M])^2$ to a bivariate normal distribution and an independent pair of independent continuous Gaussian martingales.

Consider the Markov process U starting at time $t=\epsilon$ in states 0, 1 and 2 according to the probabilities $v_1(\epsilon)$, $v_2(\epsilon)$, $1-v_1(\epsilon)-v_2(\epsilon)$ and developing in the time interval $[\epsilon,M]$ according to the finite intensity measures Λ_1^ϵ and Λ_2^ϵ. For this process we can write

$$P\{U(t)=0\} = v_1(\epsilon) \prod_{(\epsilon,t]} (1-d\Lambda_1^\epsilon)$$

$$P\{U(t)=1\} = v_1(\epsilon) \int_{(\epsilon,t]} d\Lambda_1^\epsilon(s) \prod_{(s,t]} (1-d\Lambda_2^\epsilon) + v_2(\epsilon) \prod_{(\epsilon,t]} (1-d\Lambda_2^\epsilon)$$

and from this we can calculate

$$\bar{\Lambda}_1^M(t) = \int_{[t,M)} \frac{P\{U(s-)=0\}}{P\{U(s)=1\}} d\Lambda_1^\epsilon(s) \ ,$$

$$G^M(t) = \prod_{(t,M)} (1-\bar{\Lambda}_1^M) \ .$$

Thus $1-F^\epsilon = \Pi(1-d\Lambda_2^\epsilon)$ and G^M can be constructed from $v_1(\epsilon)$, $v_2(\epsilon)$, Λ_1^ϵ and Λ_2^ϵ by the composition of a sequence of functionals involving nothing more than product integration, ordinary integration, and ordinary sums, products and ratios. By the transformation invariance of maximum likelihood estimators, $1-\hat{F}^\epsilon$ and $\hat{G}^M$ are exactly the same functionals of $V_1(\epsilon)/n$, $V_2(\epsilon)/n$, $\hat{\Lambda}_1^\epsilon$ and $\hat{\Lambda}_2^\epsilon$. Now product integration and (sum) integration of one empirical process with respect to another are compactly or Hadamard differentiable mappings from $(D[\epsilon,M])^2$ to $(D[\epsilon,M])$ with respect to the supremum norm under bounded–variation restrictions; see Gill (1989, Lemma 3 and the following Remark) [the mapping $(x,y) \rightarrow \int^{(\cdot)} x dy$] and Gill and Johansen (1987, Theorem 14) [the mapping $(x) \rightarrow \prod^{(\cdot)} (1+dx)$]. Sums, products and ratios are also compactly differentiable (in the case of ratios, as long as the denominator is bounded away from zero). So by the functional version of the δ–method (see Gill (1989, Theorem 3) or Reeds (1976)), weak convergence carries over directly.

When the extra conditions

$$\int_0^\infty (1-F)^{-1} dG < \infty, \quad \int_0^\infty G^{-1} dF < \infty \tag{5.8}$$

hold, the previously obtained extension results can be again invoked to show that we have joint weak convergence of $n^{\frac{1}{2}}(\hat{F}-F)$ and $n^{\frac{1}{2}}(\hat{G}-G)$ in $(D[0,\infty])^2$. Another compact differentiability calculation leads to asymptotic normality of $n^{\frac{1}{2}}(\hat{\alpha}-\alpha)$ where $\hat{\alpha}=\int(1-\hat{F})d\hat{G}$.

Since $\hat{G}$ and $\hat{F}$ will be dependent, the identification of the covariance structure is (as already mentioned) more conveniently based upon the orthogonal, and hence asymptotically independent martingales

$$M_i = N_i - \int_0^{(\cdot)} Y_i d\Lambda_i \ , \ i = 1,2 \ .$$

from the counting process approach.

Recall that $F_2=\Pi(1-d\Lambda_2)=F$ while $F_1=\Pi(1-d\Lambda_1)$ is given by (2.1). Also $\hat{F}_2=\hat{F}$ while $\hat{F}_1$ is the empirical distribution of the Y_i. By the simultaneous representations of $\hat{F}_i-F_i$ (i=1,2) in terms of M_i and the same martingale central limit theorem and extension results as before we can prove joint weak convergence in $(D[0,\infty])^2$ of $n^{\frac{1}{2}}(\hat{F}_i-F_i)$, u=1,2, to two independent processes $(1-F_i)\cdot W_i$ where W_i is a zero mean Gaussian martingale with var $W_i(t)=\int_0^t(v_i(s))^{-1}d\Lambda_i(s)$. In fact $(1-F_1)\cdot W_1$ has the same distribution as $B^0 \circ F_1$, where B^0 is a Brownian bridge on [0,1]. By the invariance properties of maximum likelihood estimators (see Section 4c) the simple relations quoted at the end of Section 2 between F, G, F_1, F_2 and α hold between the corresponding estimators. These rather simple expressions together with the simple form of the asymptotic covariance structure of $\hat{F}_1$ and $\hat{F}_2$ enable one to write down the asymptotic covariance structure of $\hat{F}$, $\hat{G}$ and $\hat{\alpha}$ rather easily.

5d. Asymptotic distribution of $\hat{\alpha}$.

As an example of these calculations we shall here derive the variance of the asymptotic normal distribution of $\hat{\alpha}$.

THEOREM 5.2. *Suppose* F *and* G *are continuous with common interval of support* $(0,\infty)$. *Then* $\sqrt{n}(\hat{\alpha}-\alpha) \xrightarrow{\mathfrak{D}} \cancel{\mathfrak{L}}\, N(0,\sigma^2)$ *with*

$$\sigma^2 = \alpha^3 \int_0^\infty \frac{dG}{1-F} - \alpha^2 + \alpha^3 \int_0^\infty \left[\frac{1-G}{1-F}\right]^2 \frac{dF}{G}$$

$$= \alpha^3 \int_0^\infty \frac{dF}{G} - \alpha^2 + \alpha^3 \int_0^\infty \left[\frac{F}{G}\right]^2 \frac{dG}{1-F}$$

$$< \infty$$

if and only if (5.8) holds.

<u>Proof</u>. We use the representation

$$\alpha^{-1} = \int_0^\infty (1-F_2)^{-1} dF_1;$$

$\hat{\alpha}$ is given by the same relation for the estimators. We already know by the representation $\alpha = \int_0^\infty G dF$ and the generalized δ–method (Gill, 1988, Theorem 3 and Lemma 3 and Remark) that $\sqrt{n}(\hat{\alpha}-\alpha)$ is asymptotically normal with finite variance under condition (5.8). (We later check that this is equivalent to finiteness of $\sigma^2$is unbounded and $\sqrt{n}\,[(\hat{F}_2-F_2)/(1-F_2)]$ only converges in distribution on D[0,M] for each $M<\infty$, we split the integral in the new representation for α^{-1} into an integral over $D[0,M]$, to which the generalized δ–method can again be applied, and a remainder term, integrating over (M,∞). For the remainder term, we have

$$\int_M^\infty (1-F_2)^{-1}dF_1 = \int_M^\infty \frac{(1-F)^{-1}(1-F)dG}{\alpha} = \alpha^{-1}(1-G(M))$$

and similarly for the estimators. Thus

$$\begin{aligned} \sqrt{n}(\hat{\alpha}^{-1} - \alpha^{-1}) \quad &= \quad \sqrt{n}\left[\int_0^M (1-\hat{F}_2)^{-1}d\hat{F}_1 - \int_0^M (1-F_2)^{-1}dF_1\right] \\ &\quad + \sqrt{n}\,[\hat{\alpha}^{-1}(1-\hat{G}(M)) - \alpha^{-1}(1-G(M)] \\ &= \quad Z_{M,n} + R_{M,n} \;, \qquad\qquad \text{say.} \end{aligned}$$

Since $\sqrt{n}(\hat{\alpha}-\alpha)$ converges in distribution, $1-G(M)\rightarrow 0$ as $M\rightarrow\infty$, and $\sqrt{n}(\hat{G}-G)$ converges in distribution to a tied down (Gaussian) process (the limiting process converges almost surely to 0 as $M \rightarrow \infty$), we have easily

$$\lim_{M\uparrow\infty}\ \limsup_{n\rightarrow\infty} P(|R_{M,n}|>\epsilon) = 0$$

for all $\epsilon>0$. By the generalized δ–method

$$Z_{M,n} = \int_0^M (1-F_2)^{-1}d(\sqrt{n}(\hat{F}_1-F_1)) + \int_0^M \sqrt{n}\,\frac{\hat{F}_2-F_2}{1-F_2}\,\frac{dF_1}{1-F_2} + o_P(1) \tag{5.9}$$

as $n \to \infty$, for each $M<\infty$. So $Z_{M,n}$ converges in distribution to a zero mean normal variate whose variance can be calculated by replacing $\sqrt{n}(\hat{F}_1-F_1)$ and $\sqrt{n}\,[(\hat{F}_2-F_2)/(1-F_2)]$ in (5.9) by the limiting independent Gaussian processes described above. After that we can simply let $M \to \infty$ to obtain the asymptotic variance of $\sqrt{n}(\hat{\alpha}^{-1}-\alpha^{-1})$. Combining these two steps, we find that this variance is a sum of two variances σ_1^2 and σ_2^2 coming from the asymptotically independent terms in (5.9), namely

$$\sigma_1^2 = \int_0^\infty (1-F_2)^{-2} dF_1 - \left[\int_0^\infty (1-F_2)^{-1} dF_1\right]^2$$

and

$$\sigma_2^2 = \int_{s=0}^{\infty} \int_{t=0}^{\infty} \text{as cov}\left[\sqrt{n}\,\frac{\hat{F}_2(s)-F_2(s)}{1-F_2(s)}, \sqrt{n}\,\frac{\hat{F}_2(t)-F_2(t)}{1-F_2(t)}\right] \frac{F_1(ds)}{1-F_2(s)}\,\frac{F_1(dt)}{1-F_2(t)} ,$$

which must be finite under (5.8). The double integral is more conveniently evaluated as twice that over $\{0 \le t<s<\infty\}$; also use $(1-F_2)^{-1}dF_1=\alpha^{-1}dG$. Thus

$$\sigma_1^2 \;=\; \int_0^\infty (1-F)^{-1}\alpha^{-1}dG - \alpha^{-2} ,$$

$$\sigma_2^2 \;=\; 2 \iint_{0\le t<s<\infty} \int_{u=0}^{s\wedge t} \frac{F(du)}{\alpha^{-1}G(u)\,[1-F(u)]^2}\,\alpha^{-1}G(ds)\,\alpha^{-1}G(dt)$$

$$=\; 2\,\alpha^{-1} \iiint_{0\le u<t<s<\infty} \frac{G(ds)G(dt)F(du)}{G(u)\,[1-F(u)]^2}$$

$$= \alpha^{-1} \int\int_{0\leq u<t<\infty} \frac{2[1-G(t)]G(dt)F(du)}{G(u)[1-F(u)]^2}$$

$$= \alpha^{-1} \int_{u=0}^{\infty} \frac{[1-G(u)]^2F(du)}{[1-F(u)]^2G(u)} \ .$$

Hence

$$\sigma^2 = \alpha^4(\sigma_1^2 + \sigma_2^2) = \left[\alpha^3 \int_0^{\infty} \frac{dG}{1-F} - \alpha^2\right] + \alpha^3 \int_0^{\infty} \left[\frac{1-G}{1-F}\right]^2 \frac{dF}{G} \ .$$

Note that the first term is positive by e.g. Jensen's inequality applied to the G–expectation of $(1-F)^{-1}$. Also note that $\sigma^2<\infty$ implies $\int_{(0,\infty)}(1-F)^{-1}dG < \infty$, trivially, and $\int_{(0,\infty)}G^{-1}dF < \infty$ too, since $(1-G)/(1-F)$ is close to 1 near zero where $G^{-1} \to \infty$. The converse has already been established but is easy to check explicitly. The alternative expression for σ^2, under the assumed conditions, follows from symmetry or by a (rather tedious) exercise in integration by parts.

□

5e. Chao's asymptotic results.

As mentioned in the Introduction, Chao (1987), cf. Chao and Lo (1988), obtained asymptotic results for $\hat{F}$, $\hat{G}$ and $\hat{\alpha}$ using an influence function approach. By similar techniques as just demonstrated, it is easily seen that the asymptotic covariance of $\sqrt{n}(\hat{G}-G)$ and $\sqrt{n}(\hat{F}-F)$ is indeed as given by Chao (1987, formula (3.2)), except that α^{-1} should be replaced by α in that formula (twice). Chao obtained as expression for the asymptotic variance of $\sqrt{n}(\hat{\alpha}-\alpha)$

$$\sigma^2 = \int_0^\infty \left[\int_s^\infty \{1-F(t)\}dG(t)\right]^2 \frac{dF^*(s)}{C^2(s)} + \int_0^\infty \left[\int_0^s G(t)dF(t)\right]^2 \frac{dG^*(s)}{C^2(s)}$$

$$+ \quad 2\left[1-\alpha+\alpha^{-1}\int_0^\infty G(t)\log\{G(t)\}dF(t) - \int_0^\infty F^*(s)\{1-G^*(s)\}\frac{dG^*(s)}{C^2(s)}\right]$$

where C (Woodroofe's notation) was defined in Section 2 above. Chao's result differs from ours as shall be seen below.

5f. Numerical examples.

To illustrate some of the asymptotic results above, a number of Monte Carlo simulations were performed. A simple example of distributions G and F on $(0,\infty)$ and satisfying conditions (5.8):

$$\int_0^\infty (1-F)^{-1}dG < \infty \;,\quad \int_0^\infty G^{-1}dF < \infty$$

is G exponential, F gamma (3), that is $1-G(y)=e^{-y}$, $1-F(x)=[1+x+(x^2/2)]e^{-x}$ from which one may derive

$$\alpha = \int_0^\infty G(x)dF(x) = 0.875$$

and the integrals in the representation of σ^2 in Theorem 5.2 are

$$I_{11} = \int_0^\infty (1-F)^{-1}dG = \pi/2 = 1.5708 \ ,$$

$$I_{12} = \int_0^\infty [(1-G)/(1-F)]^2 G^{-1}dF = 0.0946$$

with sum 1.6654, and

$$\int_0^\infty G^{-1}dF = 1.2021 \ , \ \int_0^\infty (F/G)^2(1-F)^{-1}dG = 0.4633$$

with the same sum. It follows that the variance of the approximate distribution of $\hat{\alpha}$ is σ^2/n with $\sigma^2 = \alpha^2(1.6654\alpha - 1) = \alpha^2 \cdot 0.4572 = 0.3500$. (This result is at variance with that conjectured by Chao (1987), whose formula for this example yields σ^2=0.31).

Table 5.1 contains summary data from 10,000 Monte Carlo simulations of n independent samples from the conditional distribution of (Y,X) given Y<X for n=5,10,20,50,100 and 800. (The random number generator RAN3 of Press et al. (1986) was used on an Olivetti M24 personal computer.) Replications with empty inner risk sets were recorded but could not be included in the averages, which thus represent conditional values, given that there was no empty inner risk set.

– Table 5.1 about here –

Note first that empty inner risk sets occur also for rather large sample size n.

The approximation of Var($\hat{\alpha}$) is rather poor, indicating a very slow approach to the limiting distribution in this particular example. We show below that the problem is primarily in the (right hand) tail of the distribution. Closer scrutiny (not documented here) of the distributional form of $\hat{\alpha}$ shows that it is heavily skewed to the left, as was to be expected from the restriction $\hat{\alpha} \leq 1$. It is interesting that by calculating the estimator of σ^2 suggested by Theorem 5.2 (just replacing F,G and α by their estimates), a strong negative correlation between $\hat{\alpha}$ and $\hat{\sigma}$ is revealed: the intuitive explanation being that the closer $\hat{\alpha}$ is to 1, the closer we are to full separation between the Y_i and the X_i, in which case α=P{Y<X} becomes much easier to estimate. The estimator $\hat{\sigma}^2$ overcompensates for that feature to the extent that the distribution of $\sqrt{n}(\hat{\alpha}-\alpha)/\hat{\sigma}$ becomes skewed to the right, but now with about the correct variance. It may finally be noticed that $\hat{\sigma}^2$ is strongly (positively) correlated with max Y_i, but slightly negatively correlated with max X_i; both of these facts are again intuitively satisfactory, at least after a little reflection.

A further documentation of the above assertion that the problems are primarily in the tails derives from the following supplementary study in close accordance with the techniques of proof used here. First note that Theorem 5.2 is primarily about asymptotic distribution of the stochastic integral

$$\hat{\alpha}^{-1} = \int_0^\infty (1-\hat{F}_2)^{-1} d\hat{F}_1$$

where $\hat{F}_1$ is the empirical distribution of the Y_i and $1-\hat{F}_2=1-\hat{F}$. An investigation of the dependence of the asymptotic results on the behaviour in the tails may therefore be performed by considering the functional

$$\xi \quad = \quad \xi_{\epsilon,M} = \int_\epsilon^M \frac{1-F_2(\epsilon)}{1-F_2(x)} dF_1(x) = [1-F(\epsilon)]\alpha^{-1}[G(M)-G(\alpha)]$$

in the notation of Section 5a. This is estimated by

$$\hat{\xi} = \int_{\epsilon}^{M} \frac{1-\hat{F}(\epsilon)}{1-\hat{F}(x)}\, d\hat{F}_1(x)$$

and using the line of argument of the proof of Theorem 5.2 it is seen that

$$\sqrt{n}\,(\hat{\xi} - \xi) \xrightarrow{\mathcal{D}} \mathcal{N}(0,\sigma^2) \ ,$$

$$\sigma^2 \quad = \quad \int_{\epsilon}^{M} \left[\frac{1-F(\epsilon)}{1-F(x)}\right]^2 dF_1(x) - \xi^2 + \alpha^{-1}[1-F(\epsilon)]^2 \int_{\epsilon}^{M} \frac{[G(M)-G(u)]^2}{[1-F(u)]^2}\, \frac{dF(u)}{G(u)} .$$

Table 5.2 contains the results of a number of Monte Carlo simulations, all with sample size $n=500$, and never yielding empty inner risk sets, from $G=$exp., $F=\Gamma(3)$.

– Table 5.2 about here –

The results show that σ^2 is a good approximation to the empirical variance for $M\le5$ and many different choices of ϵ, but that it overestimates the empirical variance considerably for $M=10$, 100 or ∞. Note that $1-F(5)=.125$, $G(5)=.993$.

Note further that the theory of Section 5a did not require the integrability conditions (5.8); hence the modifications of these results for $\hat{\xi}_{\epsilon,M}$, $\epsilon>0$, $M<\infty$, also hold true without (5.8). An obvious example where (5.8) fails is $F=G$; we study in Table 5.3 below $F=G=$exponential(1). As for the results of Table 5.1, these results are conditional on no empty inner risk set.

– Table 5.3 about here –

The approximation using σ^2 is seen to be useful (actually: quite good!) for $M \leq 5$, whereas for larger M the empirical variation of $\hat{\xi}$ is much smaller than that expected from σ^2.

Finally, the extremely slow approach to normality in our example may not be typical; after all if G is exponential and F is gamma(k), k has to be at least 3 for (5.8) to hold. For the case G=exp., $F=\Gamma(5)$, one has $\alpha=.968706$, σ^2(appr.)=.03255; 8000 replications of sample size 500 gave no empty inner risk sets and an average $\hat{\alpha}$ of .968710, empirical σ^2=.03049. This indicates faster approach to the limit in Theorem 5.2 for this example.

6. Remarks on Efficiency. In this section we discuss some general theory of asymptotically efficient estimation of possibly infinite dimensional parameters in i.i.d. models due to van der Vaart (1988a,b,c), in order to indicate how efficiency results for the NPMLE in the random truncation model can now be quite easily obtained. The idea is first to show that the empirical marginals $\hat{G}^*$, $\hat{F}^*$ are jointly asymptotically efficient for G^*, F^*, and then transfer this property to $\hat{G}^M$, $\hat{F}^\epsilon$ on $[\epsilon, M]$ by the compact differentiability of the corresponding mapping. What follows is only a sketch. Van der Vaart's framework, deriving from Koshevnik & Levit (1976) and Pfanzagl (1982) is based on the notion of a *tangent cone* and the estimation of *differentiable functionals* of the probability distribution of one observation. A tangent cone at a particular point in a model (in our case: at a particular probability distribution $P=P_{F,G}$ of X,Y given X>Y generated by a particular F and G) can be thought of as a collection of score functions of one–dimensional, one–sided submodels starting at that point, and evaluated at that point. Thus with $P=P_{F,G}$ fixed, $\mathcal{T}(P) \subset L^2(P)$ is called a tangent cone if for every $g \in \mathcal{T}(P)$ there exists a submodel

$\{P_t = P_{F_t, G_t} : t \in [0,1]\}$, with $P_0 = P$, such that

$$\int [t^{-1}(dP_t^{\frac{1}{2}} - dP^{\frac{1}{2}}) - \tfrac{1}{2}g\, dP^{\frac{1}{2}}] \to 0 \text{ as } t \downarrow 0 \ . \qquad (6.1)$$

Under further regularity conditions the function g appearing here is exactly the score function $\partial/\partial t \log(dP_t/dP_0)|_{t=0}$. A parameter κ taking values in a topological vector space B is considered as a functional of P rather than of F and G (which is possible, provided it is identifiable) and one only considers the estimation of such quantities which are smooth enough that they are differentiable (at $t=0$) with respect to the parameter t of each submodel considered in (6.1):

$$t^{-1}(\kappa(P_t) - \kappa(P)) \to \dot{\kappa}_P(g) \qquad (6.2)$$

where $\dot{\kappa}_P$ is a continuous linear map from the closed linear span $\overline{\mathcal{T}(P)}$ of $\mathcal{T}(P)$ to B. It turns out that under some standard conditions on B which are met in many applications, (see van der Vaart, 1988a, Section 4.2.1) and if $\mathcal{T}(P)$ is *convex*, a nice asymptotic efficiency theory can be worked out for the estimation of *differentiable* parameters κ. Thus one needs to establish regularity properties of both the model under consideration and of the parameter to be estimated before one can consider estimation of the parameter.

Van der Vaart's definition (Definition 4.5) of an asymptotically efficient estimator is that it is regular in Hajek's sense and converges in distribution to the best limiting distribution indicated by (his version of) the Hajek convolution theorem. An efficient estimator is then also locally asymptotic minimax in a certain sense (and a converse exists). Some important theorems characterize efficiency in terms of tightness plus component–wise or co–ordinatewise efficiency, and in terms of asymptotic linearity with a particular ('optimal') influence function. In particular a regular, asymptotically linear estimator ($n^{\frac{1}{2}}$ times estimation error is asymptotically equivalent to $n^{-\frac{1}{2}}$times a sum of a function – the influence function– of each observation) is efficient if and only if its influence function lies in $\overline{\mathcal{T}(P)}$. The (optimal) influence function is then the projection of the derivative of the functional, $\dot{\kappa}_P$, into $\mathcal{T}(P)$. Finally, efficiency of an estimator and differentiability of the estimand are preserved under compactly differentiable transformations.

Before we describe the application of these theorems in our situation, we make one further remark on the technical aspects of this theory: when the parameter κ we are estimating is F or G (or both) we will naturally consider these objects as elements of the space $D[0,\infty]$. Equipped with the usual Skorohod metric and topology, this is *not* a topological vector space (addition is not continuous). Both for the efficiency theory and for compact differentiability we give $D[0,\infty]$ the supremum norm and the σ–algebra generated by the open balls (smaller than the Borel σ–algebra) and use Dudley's (1966) weak convergence theory as expounded in Gaenssler (1983) and Pollard (1984); see van der Vaart, 1988a, Section 4.1.1). The weak convergence results we already have are equivalent to weak convergence in this alternative set–up by continuity of the sample paths of the limiting processes.

To start with we consider estimation of the marginals G^* and F^* by the marginal empiricals $\hat{G}^*$ and $\hat{F}^*$. By Donsker's theorem $(\sqrt{n}(\hat{G}^*-G^*), \sqrt{n}(\hat{F}^*-F^*))$ converges in distribution in $(D[0,\infty])^2$. One easily verifies that the parameters $G^*(x)$ and $F^*(x)$, separately, for any fixed x, are differentiable functions of $P_{G,F}$ with derivatives (6.2) which can be represented as the elements of $L^2(P)$ $1\{Y\leq y\}-G^*(y)$ and $1\{X\leq x\}-F^*(x)$ respectively. More generally, $(G^*,F^*)\in(D[0,\infty])^2$ is a differentiable function of $P_{G,F}$, cf. van der Vaart (1988a, Section 3.6.1). Since G^* and F^* (separately) each vary freely as G and F vary, we find that $\mathcal{T}(P)$ contains all square integrable, zero mean functions of X and similarly of Y. In fact a careful analysis of the score function at t=0 for a submodel $\{P_{G_t,F_t}: t\in[0,1)\}$ shows that $\mathcal{T}(P)$ is convex and contains precisely all sums of such a function of X and another of Y. Now by the discussion after Lemma 4.6 of van der Vaart (1988a), it follows that $\hat{G}^*(x)$ and $\hat{F}^*(x)$ are each asymptotically efficient for each x. By tightness of $(\sqrt{n}(\hat{G}^*-G^*), \sqrt{n}(\hat{F}^*-F^*))$, Theorem 4.9, and Example

3.6.1 of van der Vaart (1988a), we have efficiency of $(\hat{G}^*,\hat{F}^*)$ as an element of $(D[0,\infty])^2$. Finally, for any given $[\epsilon,M]$, $\hat{G}^M$ and $\hat{F}^\epsilon$ are compactly differentiable functions of $\hat{G}^*,\hat{F}^*$ (cf. Section 5c). Therefore by Theorem 4.11 of van der Vaart (1988a) they are efficient estimators of $(G^M,F^\epsilon)\in(D[\epsilon,M])^2$.

A final extension procedure was ~~be~~ used by van der Vaart (1988c) to derive from this efficiency of $(\hat{G},\hat{F})$ in $(D[0,\infty])^2$ under the previously introduced integrability conditions. The only serious complication here is that it is not clear now that G,F is a differentiable parameter in the sense of (6.2), and this has to be established first. Van der Vaart (1988b) has shown that the answer to this question is yes, without any further conditions, and that furthermore this question is intimately connected to the question of whether or not the tangent cone $\mathcal{T}(P)$ described earlier is closed: in fact $\mathcal{T}(P)=\overline{\mathcal{T}(P)}$ if and only if $\int G^{-1}dF<\infty$ and $\int (1-F)^{-1}dG<\infty$ and then $(\hat{G},\hat{F})$ is efficient for (G,F).

7. An estimation problem of Winter and Földes. Recently Winter and Földes (1986) studied the following estimation problem. Consider n independent renewal processes in equilibrium with underlying distribution function H, which we shall assume absolutely continuous with density h and support $(0,\infty)$. Corresponding to a fixed time, say 0, the forward and backward recurrence times S_i and R_i are observed; then $Q_i=R_i+S_i$ is a length–biased observation corresponding to the distribution function H. We quote the following distributional results: let χ be the expectation of H,

$$\chi=\int_0^\infty [1-H(u)]du \quad ,$$

then the joint distribution of (R,S) has density $\chi^{-1}h(r+s)$, the marginal distributions of R and S are equal with density $\chi^{-1}[1-H(r)]$, and the marginal distribution of $Q=R+S$ has density $\chi^{-1}qh(q)$, the length–biased density corresponding to h.

Winter and Földes considered (a slight modification of) the ordinary product–limit estimator based on the forward recurrence times $S_1,\dots,S_n$ and showed that it is strongly consistent for the underlying survivor function $1-H$. We shall demonstrate how the derivation of this estimator follows immediately from the Markov process framework considered here. First notice that the conditional distribution of $Q=R+S$ given that $R=r$ has density

$$\frac{\chi^{-1}h(q)}{\chi^{-1}[1-H(r)]}\ , \quad r \le q < \infty$$

that is, intensity (hazard) $h(q)/[1-H(q)]$, which is just the hazard corresponding to the underlying distribution H. Now define for each i (the i is suppressed in the notation) a stochastic process U on $[0,\infty]$ with state space $\{0,1,2\}$ by

$$U(t) = \begin{cases} 0, & 0 \le t < R \\ 1, & R \le t < R+S \\ 2, & R+S \le t \ . \end{cases}$$

We have

$$P\{U(t+h) = 2|U(u), 0 \le u \le t\} = o(h)$$

for $U(t)=0$, and for $U(t)=1$ (that is, $R\le t<R+S$) this is

$$P\{R+S \le t+h|R, R+S > t\} = \frac{h(t)}{1-H(t)}h + o(h)$$

by the above result on the hazard of $R+S|R$. That this depends only on t but not on R proves that U is a Markov process

$$\boxed{0} \quad \xrightarrow{\alpha} \quad \boxed{1} \quad \xrightarrow{\beta} \quad \boxed{2}$$

with intensities

$$\alpha(t) = [1-H(t)]/\int_t^\infty [1-H(r)]dr$$

(the marginal hazard of R, equal to the residual mean lifetime function of the underlying distribution H) and

$$\beta(t) = h(t)/[1-H(t)] \ .$$

The Markov process framework of Section 2 indicates that the Nelson–Aalen and product limit estimators based on $S_1,\dots,S_n$ are natural estimators of the integrated intensity $B(t)=\int_0^t \beta(q)dq$ respectively the survivor function $1-H$ of the underlying distribution, and consistency and asymptotic normality may be obtained as shown in Section 5.

Note that the backwards intensity

$$\begin{aligned}\bar{\alpha}(t) &= \alpha(t)\,\frac{P\{U(t)=0\}}{P\{U(t)=1\}} \\ &= \alpha(t)\,\frac{P\{R>t\}}{P\{R\leq t<R+S\}}\end{aligned}$$

$$= \alpha(t) \frac{\chi^{-1} \int_t^{\infty} [1-H(r)]dr}{\chi^{-1} \int_0^t \int_{t-r}^{\infty} h(r+s)dsdr}$$

$$= \frac{1-H(t)}{\int_t^{\infty} [1-H(r)]dr} \frac{\int_t^{\infty} [1-H(r)]dr}{\int_0^t [1-H(t)]dr} = \frac{1}{t} ,$$

the intensity of a uniform distribution on some interval $[0,A]$. Since it has been assumed that R has support $(0,\infty)$, this shows that the present model <u>may not</u> be interpreted as a left truncation model, which would require that $\bar{\alpha}(t)$ corresponded to a probability distribution on $(0,\infty)$.

The fact that $\bar{\alpha}(t)$ is uniform corresponds to Winter and Földes' statement that (R,S) contain no more information than $R+S$ about H. This might already have been gleaned from the likelihood function based on observation of $(R_1,S_1),\ldots,(R_n,S_n)$, which is

$$\chi^{-n} \prod_{i=1}^{n} h(r_i + s_i)$$

from which the NPMLE of H is readily derived as

$$\hat{H}(t) = \sum_{i=1}^{n} \frac{I\{R_i+S_i \le t\}}{R_i+S_i} \Big/ \sum_{i=1}^{n} \frac{1}{R_i+S_i} ,$$

that is the Cox–Vardi estimator in the terminology of Winter and Földes (Cox 1969, Vardi 1985).

It follows that the estimators based on the forward recurrence times $S_1,\ldots,S_n$ are not NPMLE. The difference between the situation here and that of Section 3 is that not only the intensity $\beta(t)$, but also $\alpha(t)$ depends only on the estimand H. In Section 3 λ_1 depended on both parameters γ and Φ in such a way that even when Φ was fixed, λ_1 could vary freely by varying γ.

Weak convergence of the Winter–Földes estimator is immediate from our results in Section 5. In particular, in order to achieve the extension to convergence on $[0,M]$ it should be required that

$$\int_0^{\epsilon} d\Phi(s)/v_2(s) < \infty$$

in the terminology of Section 5c, and using $d\Phi(t)=\beta(t)dt$ and

$$v_2(t) = P\{U(t) = 1\} = \int_0^t \frac{1-H(s)}{\chi} \frac{1-H(t)}{1-H(s)} ds = \frac{1}{\chi}[1 - H(t)] \ ,$$

the integrability condition translates into

$$\int_0^{\epsilon} t^{-1}h(t)dt < \infty \ ,$$

or finiteness of $E(X^{-1})$ where X has the underlying ("length–unbiased") interarrival time distribution H. It may easily be seen from Gill, Vardi and Wellner (1988) that the same condition is needed to ensure weak convergence of the Cox–Vardi estimator.

Acknowledgements. The authors are grateful to W.–M. Huang, W.–Y. Tsai, M.–C. Wang and B.B. Winter for supplying copies of unpublished work, to Per Kragh Andersen, Søren Johansen, Aad van der Vaart and Jon Wellner for important contributions, and to Lars Sommer Hansen and Claus Holst for help with some of the calculations.

Table 5.1. Results from 10,000 Monte Carlo replications of samples of size n from the conditional distribution (Y,X|Y<X) with Y exponential, X gamma (3) and Y and X independent.

Sample size n	Frequency of replications with empty inner risk set	Mean $\hat{\alpha}$	n Var($\hat{\alpha}$) (obs.)
5	0.0364	0.9014	0.0988
10	0.0091	0.8770	0.1646
20	0.0030	0.8743	0.1869
50	0.0003	0.8744	0.1985
100	0.0003	0.8742	0.2127
800	0.0000	0.8748	0.2400
∞ (Theoretical value)	0	0.875	0.3500

Table 5.2. Results of Monte Carlo simulations of finite integrals $\xi_{\epsilon,M}$. Sample size n=500, 500 replications except ϵ=2, M=5 (2500 repl.) G=exp., F=Γ(3).

ϵ	M	ξ	average $\hat{\xi}$	σ^2	emp.σ^2
.001	.2	.2060	.2050	.170	.131
.5	2	.5308	.5297	.330	.306
.05	5	1.0794	1.0806	.272	.306
.2	5	.9269	.9256	.444	.432
.5	5	.6756	.6772	.521	.588
2	5	.09945	.09969	.154	.166
.01	10	1.1314	1.1306	1.901	.361
.001	100	1.1417	1.1447	3.37	1.59
0	∞	1.1429	1.1448	.597	.375

Table 5.3. Results of Monte Carlo simulation of finite integrals $\xi_{\epsilon,M}$. Sample size n=500, 500 replications except ϵ=.2, M=5 (10,000 repl.). G=F=exp.

ϵ	M	freq. of empty inner risk sets	ξ	avg. $\hat{\xi}$	σ^2	emp. σ^2
.01	1	0	1.232	1.238	2.963	3.053
.5	2	.002	.572	.574	1.009	1.034
.2	5	.0017	1.330	1.340	6.84	7.83
2	5	.002	.0348	.0357	.1232	.1306
.1	10	.008	1.637	1.644	17.38	12.15
.5	10	.002	.736	.749	7.13	5.88
.01	100	.004	1.960	1.980	201.2	23.2
.001	1000	.002	1.996	2.027	2006	34

REFERENCES

AALEN, O.O. (1975). Statistical inference for a family of counting processes. Ph.D. dissertation, Department of Statistics, University of California, Berkeley.

AALEN, O.O. (1976). Nonparametric inference in connection with multiple decrement models. *Scand. J. Statist.* **3** 15-27.

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

AALEN, O.O., BORGAN, Ø., KEIDING, N., and THORMANN, J. (1980). Interaction between life history events: Nonparametric analysis of prospective and retrospective data in the presence of censoring. *Scand. J. Statist.* **7** 161-171.

AALEN, O.O. and JOHANSEN, S. (1978). An empirical transition matrix for nonhomogeneous Markov chains based on censored observations. *Scand. J. Statist.* **5** 141-150.

ANDERSEN, P.K. and BORGAN, Ø. (1985). Counting process models for life history data: A review (with discussion). *Scand. J. Statist.* **12** 97-158.

ANDERSEN, P.K., BORGAN, Ø., GILL, R. and KEIDING, N. (1988). Censoring, truncation and filtering in statistical models based on counting processes. *Contemporary Mathematics* (to appear).

BARNDORFF-NIELSEN, O. (1978). *Information and Exponential Families in Statistical Theory*. Wiley, Chichester.

BEGUN, J.M., HALL, W.J., HUANG, W.M. and WELLNER, J.A. (1983). Information and asymptotic efficiency in parametric-nonparametric models. *Ann. Statist.* **11** 432-452.

BHATTACHARYA, P.K., CHERNOFF, H. and YANG, S.S. (1983). Nonparametric estimation of the slope of a truncated regression. *Ann. Statist.* **11** 505-514.

BICKEL, P.J. and RITOV, J. (1987). Large sample theory of estimation in biased sampling regression models I. Techn. Rep. 115, Dept. Statistics, Univ. California, Berkeley.

BILLINGSLEY, P. (1968). *Convergence of Probability Measures*. Wiley, New York.

CHAO, M.-T. (1987). Influence curves for randomly truncated data. *Biometrika* **74** 426-429.

CHAO, M.-T. and LO, S.-H. (1988). Some representations of the nonparametric maximum likelihood estimators with truncated data. *Ann. Statist.* **16**, 661-668.

COX, D.R. (1969). Some sampling problems in technology. In *New Developments in Survey Sampling* (N.L. Johnson and H. Smith, Jr., eds.) 506-527. Wiley, New York.

DUDLEY, R.M. (1966). Weak convergence of measures on nonseparable metric spaces and empirical measures on Euclidean spaces. *Ill. J. Math.* **10** 109-126.

GAENSSLER, P. (1983). *Empirical Processes*. Lecture Notes-Monograph Series *3*, Institute of Mathematical Statistics, Hayward, Ca.

GILL, R.D. (1980). *Censoring and Stochastic Integrals*. Mathematical Centre Tracts 124. Mathematical Centre, Amsterdam.

GILL, R.D. (1983). Large sample behavior of the product-limit estimator on the whole line. *Ann. Statist.* **11** 49-58.

GILL, R.D. (1989). Non- and semi-parametric maximum likelihood estimators and the von Mises method (part I). *Scand. J. Statist.* (to appear).

GILL, R.D. and JOHANSEN, S. (1987). Product-integrals and counting processes. Preprint 4, Institute of Mathematical Statistics, University of Copenhagen.

GILL, R.D., VARDI, Y. and WELLNER, J.A. (1988). Large sample theory of empirical distributions in biased sampling models. *Ann. Statist.* (to appear).

HALLEY, E. (1693). An estimate of the degrees of the mortality of mankind drawn from curious tables of the births and funerals at the city of Breslaw. *Phil. Trans. Roy. Soc. London* **17** 596-610. Reprinted in *J. Inst. Act.* **112** 278-301 (1985).

HUANG, W.-M. and TSAI, W.-Y. (1986). Asymptotic optimality of the nonparametric conditional maximum likelihood estimator under random truncation. Technical Report, Lehigh University and Brookhaven National Laboratory.

JOHANSEN, S. (1978). The product limit estimator as maximum likelihood estimator. *Scand. J. Statist.* **5** 195-199.

JOHANSEN, S., (1987). Product integrals and Markov processes. *CWI Newsletter* **12** 3-13; originally appeared (1977) as: Preprint no. 3, Institute of Mathematical Statistics, Univ. of Copenhagen.

KAPLAN, E.L. and MEIER, P. (1958). Non-parametric estimation from incomplete observations. *J. Amer. Statist. Assoc.* **53** 457-481.

KEIDING, N., BAYER, T. and WATT- BOOLSEN, S. (1987). Confirmatory analysis of survival data using left truncation of the life times of primary survivors. *Statist. in Medicine* **6** 939-944.

KELLERER, H.G. (1986). Order conditioned independence of real random variables. *Math. Ann.* **273** 507-528.

KOSHEVNIK, YU.A. and LEVIT, B.YA. (1976). On a nonparametric analogue of the information matrix. *Theor. Prob. Appl.* **21** 738-753.

LAGAKOS, S.W., BARRAJ, L.M. and De GRUTTOLA, V. (1987). Nonparametric analysis of truncated survival data, with application to AIDS. *Technical Report No. 544z*, Division of Biostatistics, Dana-Farber Cancer Institute.

LENGLART, E. (1977). Relation de domination entre deux processus. *Ann. inst. Henri Poincaré* **13** 171-179.

LYNDEN-BELL, D. (1971). A method of allowing for known observational selection in small samples applied to 3CR quasars. *Monthly Notices Roy. Astron. Soc.* **155** 95-118.

MEIER, P. (1975). Estimation of a distribution function from incomplete observations. *Perspectives in Probability and Statistics* 67-87. Ed. J. Gani. Applied Probability Trust, Sheffield, England.

PFANZAGL, J. (1982). (with Wefelmeyer, W.) *Contributions to a general asymptotic statistical theory*. Lecture notes in statistics **13**, Springer.

POLLARD, D. (1984). *Convergence of stochastic processes*. Springer, New York.

PRESS, H.W., FLANNERY, P.B., TEUKOLSKY, A.S. and VETTERLING, T.W. (1986). *Numerical Recipes*. Cambridge University Press.

REBOLLEDO, R. (1980). Central limit theorems for local martingales. *Z. Wahrscheinlichkeitsth. verw. Geb.* **51** 269-286.

REEDS, J.A., III (1976). On the definition of von Mises functionals. Ph.D. Thesis, Department of Statistics, Harvard University.

SAMUELSEN, S.O. (1988). Nonparametric estimation of the cumulative intensity from doubly censored data. Asymptotic theory. *Scand. J. Statist.* (to appear).

TSAI, W-Y., JEWELL, N.P. and WANG, M-C. (1987). A note on the product-limit estimator under right censoring and left truncation. *Biometrika* **74** 883-886.

TSUI, K-L., JEWELL, N.P. and WU, C.F.J. (1987). A nonparametric approach to the truncated regression problem. Manuscript.

VAN DER VAART, A. (1988a). *Statistical Estimation in Large Parameter Spaces*. CWI Tract **44**, Centre for Mathematics and Computer Science, Amsterdam.

VAART, A. VAN DER (1988b). On differentiable functionals. Techn. Rep., Free Univ. Amsterdam.

VAART, A. VAN DER (1988c). Efficiency and Hadamard differentiability. Techn. Rep., Free Univ. Amsterdam.

VARDI, Y. (1985). Empirical distributions in selection bias models. *Ann. Statist.* **13** 178-203.

WANG, M-C. (1987a). Product-limit estimates: a generalized maximum likelihood study. *Commun. Statist. - Theory Meth.* **16**, 3117-3132.

WANG, M-C., (1987b). A semi-parametric model for randomly truncated data. Paper 635, Department of Biostatistics, Johns Hopkins University.

WANG, M.-C., JEWELL, N.P. and TSAI, W.-Y. (1986). Asymptotic properties of the product limit estimate under random truncation. *Ann. Statist.* **14** 1597-1605.

WINTER, B.B. and FÖLDES, A. (1986). Product-limit estimators for use with length-biased data. Technical Report, Department of Mathematics, University of Ottawa.

WOODROOFE, M. (1985). Estimating a distribution function with truncated data. *Ann. Statist.* **13** 163-177.

WOODROOFE, M. (1987). Correction to "Estimating a distribution function with truncated data." *Ann. Statist.* **15**, 883.

YING, Z. (1987). A note on the asymptotic properties of the product-limit estimator on the whole line. Manuscript, Department of Statistics, University of Illinois at Urbana-Champaign.

Statistical Research Unit
University of Copenhagen
Blegdamsvej 3
DK-2200 Copenhagen N Denmark

Centre for Mathematics and Computer Science
Kruislaan 413
1098 SJ Amsterdam
The Netherlands