Lectures on Survival Analysis

Richard D. Gill
Mathematical Institute, University Utrecht,
Budapestlaan 6, 3584 CD Utrecht, Netherlands.
gill@math.ruu.nl

To appear in: Ecole d'Eté de Probabilités de Saint Flour XXII,
ed. P. Bernard, Springer Lecture Notes in Mathematics

Preface.

These notes, though mainly written after the course was given, follow quite closely my lectures at the 22nd Saint Flour *École d'Été de Calcul des Probabilités*, 8–25 July 1992. I am really grateful to the organisers and the participants who have shaped the lecture notes in many ways.

The vague title is a cover-up for the more honest 'topics in and around survival analysis which interest me at the moment, with an audience of French probabilists in mind'. Accordingly, the main theme of the lectures—to my mind the fundamental notion in survival analysis—is product-integration, and to begin with I have tried to cover its basic theory in fair detail. Probabilistic connections are emphasized.

The next group of lectures study the Kaplan-Meier or product-limit estimator: the natural generalisation, for randomly censored survival times, of the empirical distribution function considered as nonparametric estimator of an unknown distribution. Using product-integration, the asymptotics of the Kaplan-Meier estimator are treated in two different ways: firstly, using modern empirical process theory, and secondly, using martingale methods. In both approaches a simple identity from product-integration, the Duhamel equation, does all the real work. Counting processes lurk in the background of the martingale approach though they are not treated here at length; the interested reader is urged to follow them up in the book *Statistical models based on counting processes* by P.K. Andersen, Ø. Borgan, R.D. Gill and N. Keiding (1993); the book is referred to as 'ABGK' in the sequel.

I also neglect statistical issues such as asymptotic optimality theory, partly with my audience in mind, and partly because this subject is still very fluid with, in my opinion, interesting developments ahead; the reader is referred in the meantime to Section IV.1.5 and Chapter VIII of ABGK. However beneath the surface statistical ideas, especially involving nonparametric maximum likelihood, are ever-present and give the real reason for many otherwise surprising results.

Neglected in the written notes are applications, though, in the real lectures, illustrations taken from ABGK were prominent.

Most of this part of the course covers classical material, though there are also new results. One of the most striking is the proof, using discrete time martingales, of Stute and Wang's (1993) very recent Glivenko-Cantelli theorem for the Kaplan-Meier estimator. I suspect the techniques used here could find applicability in many other problems in survival analysis of a sequential or combinatorial nature. Another striking result is the use of van der Laan's identity (on estimation of linear parameters in convex models; van der Laan, 1993a) to give a more or less two-line proof of consistency, weak convergence, and correctness of the bootstrap, of the Kaplan-Meier estimator. We also give a new bootstrap confidence band construction for the Kaplan-Meier estimator 'on the whole line' (the first 'whole line' confidence band which does not rely on any integrability conditions at all).

While the first part of the lecture notes contains an introduction to survival analysis or rather to some of the mathematical tools which can be used there, the second part goes beyond or outside survival analysis and looks at somehow related problems in multivariate time and in spatial statistics: we give an introduction to Dabrowska's multivariate product-limit estimator, to non-parametric estimation in Laslett's line-segment problem (again using van der Laan's identity), and to the estimation of inter-event distance distributions in spatial point processes. All these topics involve in some way or another variants of the Kaplan-Meier estimator. The results are taken from 'work in progress' and are sometimes provisional in nature.

Many topics central to survival anaysis (the Cox regression model; the log rank test; and so on) are missing in this course. Even when we restrict attention to product-limit type estimators, it is a pity not to have included sections on the Aalen-Johansen product-limit estimator for an inhomogenous Markov process, and to nonparametric estimation with randomly truncated data. Again, the disappointed reader is referred to ABGK to rectify such omissions.

Finally one lecture was given on something completely different: the cryptographic approach to random number generation. One section on that subject is therefore also included here 'for the record'.

Contents

1. Introduction: survival and hazard

Survival analysis is the branch of applied statistics dealing with the analysis of data on times of events in individual life-histories (human or otherwise). A more modern and broader title is *generalised event history analysis.* To begin with, the event in question was often the failure of a medical treatment designed to keep cancer patients in remission and the emergence and growth of survival analysis was directly connected to the large amount of resources put into cancer research. This area of medical statistics brought a whole new class of problems to the fore, especially the problem of how to deal with *censored data.* At first many ad hoc techniques were used to get around these problems but slowly a unifying idea emerged. This is to see such data as the result of a dynamic process in time, each further day of observation producing some new pieces of data. Tractable statistical models are based on modelling events continuously in time, conditioning on past events; and new statistical ideas such as partial likelihood are also based on this dynamic time structure.

This means that the basic notion in the mathematics of survival analysis is surely that of the *hazard rate*, and the basic mathematical tool is *product-integration*, providing the means of moving to and fro between a dynamic description in terms of hazards (or more generally, intensities) and a more static description in terms of probability densities or their tail integrals, the survival function. We start by defining these basic notions and show how the relation between hazard and survival is a general instance of a relation between *additive* and *multiplicative* interval functions.

Let T be a positive random variable, with distribution function F, representing the time of occurrence of some event. The *survival function* S is defined by

$$S(t) \;=\; \mathrm{P}(T > t),$$

the probability of surviving (not experiencing the event) up to (and including) time t. Of course $S = 1 - F$. We define the *cumulative hazard function* Λ by

$$\Lambda(t) \;=\; \int_0^t \frac{F(\mathrm{d}s)}{S(s-)}.$$

One may check (e.g., using dominated convergence for t such that $S(t-) > 0$, and a monotonicity argument for other t) that

$$\Lambda(t) \;=\; \lim \sum_i \Bigl(1 - \frac{S(t_i)}{S(t_{i-1})}\Bigr) \;=\; \sum_i \mathrm{P}(T \le t_i \mid T > t_{i-1})$$

where $0 = t_0 < t_1 < \ldots < t_n = t$ is a partition of $(0, t]$ and the limit is taken as the mesh of the partition, $\max_i |t_i - t_{i-1}|$, converges to zero.

One can also consider Λ as a measure, $\Lambda(\mathrm{d}t) = F(\mathrm{d}t)/S(t-)$. Treating $\mathrm{d}t$ not just as the length of a small time interval $[t, t + \mathrm{d}t)$ but also as the name of the interval itself, one can interpret $\Lambda(\mathrm{d}t)$ as $P(T \in \mathrm{d}t \mid T \ge t)$, hence the name *hazard*, the risk of experiencing the event (death, failure, ...) in the small time interval $\mathrm{d}t$, given survival up to the start of the interval. (It is necessary to think of the interval $\mathrm{d}t$ as left closed, right open, in contrast to ordinary time intervals which will usually be left open, right closed). For an ordinary interval $(s,t]$ we write $\Lambda(s,t) = \Lambda((s,t]) = \Lambda(t) - \Lambda(s)$ for the total hazard of the time interval. This makes Λ an *additive interval function*: for $s \le t \le u$,

$$\Lambda(s,u) \;=\; \Lambda(s,t) + \Lambda(t,u).$$

The survival function S generates another interval function, which we denote $S(s,t)$:

$$S(s,t) \;=\; \frac{S(t)}{S(s)} \;=\; \mathrm{P}(T > t \mid T > s),$$

the probability of surviving the interval $(s,t]$ given one survives its starting point s. We may call this the *conditional survival function.* This interval function is *multiplicative*: for $s \le t \le u$

$$S(s,u) \;=\; S(s,t)S(t,u).$$

From now on one must be careful: when treating S as an interval function we naturally write $S(\mathrm{d}t)$ for $S([t,t+\mathrm{d}t)) = S(t-,(t+\mathrm{d}t)-)$; informally, the probability of surviving $\mathrm{d}t = [t,t+\mathrm{d}t)$ given survival up to but not including t. This must not be confused with an infinitesimal element of the additive measure generated in the ordinary way by the function of one variable $t \mapsto S(t)$.

We now have the following facts about the interval functions S and Λ: S is multiplicative, while Λ is additive; moreover they are related by

$$\Lambda(\mathrm{d}s) \;=\; 1 - S(\mathrm{d}s)$$

or equivalently

$$S(\mathrm{d}s) \;=\; 1 - \Lambda(\mathrm{d}s).$$

Adding $\Lambda(\mathrm{d}s)$ over small time intervals $\mathrm{d}s$ forming a partition of $(0,t]$, and similarly multiplying $S(\mathrm{d}s)$ over the small intervals, these two formulas give the, for the time being informal, duality:

$$\begin{aligned}
\Lambda(t) \;&=\; \Lambda(0,t) \;=\; \int_{(0,t]} (1 - S(\mathrm{d}s)) \\
S(t) \;&=\; S(0,t) \;=\; \prod_{(0,t]} (1 - \Lambda(\mathrm{d}s)).
\end{aligned}$$

A small point of heuristics: to make the intervals match up properly one should think of $(0,t]$ as being the same as $[0+\mathrm{d}0, t+\mathrm{d}t)$. The integral $\int$ and *product-integral* $\prod$ will be defined formally as limits over partitions of $(0,t]$ with mesh converging to zero of sums and products respectively of the interval functions $1-S$ and $1-\Lambda$. We have found:

> *The hazard Λ, an additive interval function, is the additive integral of $1-S$; conversely the survival function S, seen as a multiplicative interval function, is the multiplicative integral of $1-\Lambda$.*

Since $S(s,t) = \prod_s^t(1-\mathrm{d}\Lambda)$, $\Lambda(s,t) = \int_s^t(1-\mathrm{d}S)$, it follows from this duality that the conditional distribution of T given $T > s$ has hazard function $\Lambda(s,\cdot)$ on (s,∞) or in other words hazard measure $\Lambda|_{(s,\infty)}$.

This is good motivation to study the duality both in more detail and more generality. In particular we will generalise the duality to the case when Λ and S are replaced by (square) matrix valued functions (and 1 by the identity matrix): this will produce the duality between the multiplicative transition matrix and the additive matrix intensity measure of a finite state space, time inhomogenous Markov process.

Another aspect of product-integral formalism is that it gives an effortless unification of the discrete and continuous cases. Consider two special cases of the above: that in which the distribution of T is absolutely continuous, and that in which it is discrete. In the discrete case where T has a discrete density $f(t) = \mathrm{P}(T = t)$ we define the discrete hazard function $\lambda(t) = \mathrm{P}(T = t \mid T \geq t)$ and find $\Lambda(t) = \sum_{s \leq t} \lambda(s)$, $S(t) = \prod_{s \leq t}(1 - \lambda(s))$. On the other hand, in the continous case where T has a density $f = F'$, we define the hazard rate $\lambda = f/(1 - F)$. We find $\Lambda(t) = \int_0^t \lambda(s)\mathrm{d}s$, and our product-integral representation $S(t) = \prod_0^t(1 - \Lambda(\mathrm{d}s))$ becomes $S(t) = \exp(-\Lambda(t))$, which is a much less intuitive and seemingly quite different relationship.

We will establish continuity and even differentiablity properties of the product-integral mapping which in particular gives information on how to go from discrete to continuous survival functions, and from discrete time Markov chains to continuous time Markov processes. Later (section 11) we will also take a look at product-integration over higher-dimensional (non ordered) time.

2. Product-integration.

Product-integration was introduced by Volterra (1887). An extensive survey, including some of the history of product-integration, is given by Gill and Johansen (1990). Here we take (and improve slightly) their approach, which was based on MacNerney (1963) with a key element coming from Dobrushin (1953). Another survey with many more references and applications but taking a different approach is given by Dollard and Friedman (1979).

α and μ will denote $p \times p$ matrix valued additive, respectively multiplicative, right continuous interval functions on $[0, \infty)$. The identity matrix and the zero matrix will simply be written as 1 and 0; the context will always show what is meant. The special case $p = 1$ and $\alpha \geq 0$, or $\mu \geq 1$, will be called 'the real, nonnegative case', and we will write α_0 and μ_0 instead of α and μ for emphasis. We want to establish the duality $\mu = \prod(1 + \mathrm{d}\alpha)$, $\alpha = \int(\mathrm{d}\mu - 1)$, and derive further properties of the product-integral. Intuitively the duality follows by noting that for a small interval $\mathrm{d}s$, $\mu(\mathrm{d}s) = 1 + \alpha(\mathrm{d}s)$ if and only if $\alpha(\mathrm{d}s) = \mu(\mathrm{d}s) - 1$. Now multiplying or adding over a fine partition of $(0, t]$ gives the required relations.

The approach will be to consider the real nonnegative case first, deriving the results in that case by a simple monotonicity argument. Then we show how the general case follows from this special one through a so-called domination property together with some easy algebraic identities concerning matrix sums and products. Results on hazard and survival will follow by taking $\alpha = -\Lambda$, $\mu = S$. Since $-\Lambda \leq 0$ and $S \leq 1$, the complete argument via domination is needed in this case, even though Λ and S are scalar.

This part of the theory (following MacNerney, 1963) shows that product-integrals exist as limits, under refinements of partitions of an interval, of finite products over the subintervals in the partition. Right continuity plays no role yet. Using right continuity, we strengthen this to a uniform limit as the mesh of the partition (the length of the longest subinterval) converges to zero, using an idea from Dobrushin (1953). Right continuity also allows a measure-theoretic interpretation of the main results.

To begin with we state the algebraic identities. Generalised to continuous products they will become some of the key properties of product-integrals which we will make much use of later. In fact, (1) and (2) become the Kolmogorov forward and backward equations respectively, or alternatively, Volterra integral equations; Volterra's (1887) original motivation for introducing product-integration. Equation (3) doesn't seem to have a name but is very useful all the same. Equation (4) becomes the Duhamel equation, a powerful tool expressing the difference between two product-integrals in terms of the difference of the integrands (the history of its name is not clear). Equation (5) becomes the Peano series (Peano, 1888), expressing the product-integral as a sum of repeated integrals (or as a Neumann series).

Lemma 1. *Let $a_1, \ldots, a_n$ and $b_1, \ldots, b_n$ be $p \times p$ matrices. Then (with an empty product equal to* 1*):*

$$\prod_j (1 + a_j) \ - 1 \ = \ \sum_j \Bigl(\prod_{i<j} (1 + a_i) \Bigr) a_j, \tag{1}$$

$$\prod_j (1 + a_j) \ - 1 \ = \ \sum_j a_j \Bigl(\prod_{k>j} (1 + a_k) \Bigr), \tag{2}$$

$$\prod_i (1 + a_i) \ - 1 - \sum_i a_i \ = \ \sum_{i,k\,:\,i<k} a_i \Bigl(\prod_{j\,:\,i<j<k} (1 + a_j) \Bigr) a_k, \tag{3}$$

$$\prod_j (1 + a_j) - \prod_j (1 + b_j) = \sum_j \Bigl(\prod_{i<j} (1 + a_i)(a_j - b_j) \prod_{k>j} (1 + b_k) \Bigr). \tag{4}$$

$$\prod_i (1 + a_i) \ = \ 1 + \sum_{m=1}^{n} \sum_{i_1 < i_2 < \ldots < i_m} a_{i_1} \ldots a_{i_m}. \tag{5}$$

Proof. Equation (4) is seen to be a telescoping sum if one replaces the middle term on the right, $a_j - b_j$, with $(1 + a_j) - (1 + b_j)$, and expands on this difference. Equations (1) and (2) follow by taking all b_j and all a_j respectively equal to the zero matrix 0. Equation (3) follows by taking the '-1' in (2) to the right hand side, and then substituting for $\prod(1 + a_i)$ in the right hand side of (1). Equation (5) is obvious. □

Now let α and μ respectively be additive and multiplicative interval functions which are *right continuous*:

$$\alpha(s,t) \ \to \ \alpha(s,s) \ = \ 0 \quad \text{as } t \downarrow s,$$

$$\mu(s,t) \ \to \ \mu(s,s) \ = \ 1 \quad \text{as } t \downarrow s.$$

By α_0 and μ_0 we denote respectively additive and multiplicative real right continuous interval functions with $\alpha_0 \geq 0$ and $\mu_0 - 1 \geq 0$. We suppose α is *dominated* by α_0 and

$\mu - 1$ by $\mu_0 - 1$, which means $\|\alpha\| \le \alpha_0$ and $\|\mu - 1\| \le \mu_0 - 1$. Here, $\|a\|$ is the matrix norm $\max_i \sum_j |a_{ij}|$, which means we also have

$$\|a+b\| \le \|a\| + \|b\|, \quad \|ab\| \le \|a\|\|b\|, \quad \|1\| = 1.$$

It will turn out that domination of $\mu - 1$ by $\mu_0 - 1$ for a multiplicative interval function μ_0 is equivalent to domination by an additive interval function α_0; one can then take $\mu_0 = \prod(1 + \mathrm{d}\alpha_0)$.

We say alternatively that $\mu - 1$ and α are of *bounded variation* if $\mu - 1$ and α are dominated by real, right continuous, *additive* interval functions. For the time being the right continuity of α and μ will not be important. The property will be used later when we interpret our results in terms of standard (measure theoretic) integration theory.

Let $(s,t]$ denote a fixed time interval and let $\mathcal{T}$ denote a partition of $(s,t]$ into a finite number of sub-intervals. Note the inequalities

$$1 + a + b \le (1+a)(1+b) \le \exp(a+b), \quad a, b \ge 0,$$

$$\log(xy) \le (x-1) + (y-1) \le xy - 1, \quad x, y \ge 1.$$

The first shows that $\prod_{\mathcal{T}}(1+\alpha_0)$ is bounded from above by $\exp \alpha_0(s,t)$ and *increases* under refinement of the partition $\mathcal{T}$. Similarly $\sum_{\mathcal{T}}(\mu_0 - 1)$ is bounded from below by $\log \mu_0(s,t)$ and *decreases* under refinement of the partition. This means we may define

$$\prod_{(s,t]}(1 + \mathrm{d}\alpha_0) \;=\; \lim_{\mathcal{T}} \prod_{\mathcal{T}}(1 + \alpha_0), \tag{6}$$

$$\int_{(s,t]}(\mathrm{d}\mu_0 - 1) \;=\; \lim_{\mathcal{T}} \sum_{\mathcal{T}}(\mu_0 - 1), \tag{7}$$

where the limits are taken under refinement of partitions of $(s,t]$. (Thus: for any $\varepsilon > 0$ and any partition there exists a refinement of that partition such that for all further refinements, the approximating sum or product is within ε of the limit).

Proposition 1. *For given α_0 define $\mu_0 = \prod(1 + \mathrm{d}\alpha_0)$. Then $\mu_0 \ge 1$ is a right continuous, multiplicative interval function and $\alpha_0 = \int(\mathrm{d}\mu_0 - 1)$. Conversely, for given μ_0 define $\alpha_0 = \int(\mathrm{d}\mu_0 - 1)$. Then $\alpha_0 \ge 0$ is a right continuous, additive interval function and $\mu_0 = \prod(1 + \mathrm{d}\alpha_0)$.*

Proof. The following bounds are easy to verify: for given α_0, $\mu_0 = \prod(1 + \mathrm{d}\alpha_0)$ satisfies $\exp(\alpha_0) - 1 \ge \mu_0 - 1 \ge \alpha_0 \ge 0$. Similarly, for given μ_0, $\alpha_0 = \int(\mathrm{d}\mu_0 - 1)$ satisfies $0 \le \log \mu_0 \le \alpha_0 \le \mu_0 - 1$. The right continuity is now easy to establish and additivity or multiplicativity also easily verified.

Our proof of the duality establishes the following chain of inequalities, which gives some insight into why the duality holds:

$$\begin{aligned}0 \le \sum_{\mathcal{T}}(\mu_0 - 1) \ - \alpha_0(s,t) \ &\le \ \mu_0(s,t) - \prod_{\mathcal{T}}(1+\alpha_0) \\ &\le \mu_0(s,t)\Big(\sum_{\mathcal{T}}(\mu_0 - 1) \ - \alpha_0(s,t)\Big).\end{aligned} \tag{8}$$

First, let $\alpha_0 \ge 0$ be given and define $\mu_0 = \prod(1+\mathrm{d}\alpha_0)$. Let α_j and μ_j denote the values of α_0 and μ_0 on the elements of the partition $\mathcal{T}$. Using the easy bounds on μ_0 and its multiplicativity we find

$$\begin{aligned}0 &\le \sum_{\mathcal{T}}(\mu_0 - 1) - \alpha_0(s,t) \\ &= \sum_j(\mu_j - 1 - \alpha_j) \\ &\le \sum_j \prod_{i<j}(1+\alpha_i)(\mu_j - 1 - \alpha_j)\prod_{k>j}\mu_k \\ &= \prod_j \mu_j - \prod_j(1+\alpha_j) \\ &= \mu_0(s,t) - \prod_{\mathcal{T}}(1+\alpha_0).\end{aligned}$$

Since $\prod_{\mathcal{T}}(1+\alpha_0) \to \mu_0(s,t)$ this shows that $\sum_{\mathcal{T}}(\mu_0 - 1) \to \alpha_0(s,t)$ and also gives the first half of (8).

Conversely, let $\mu_0 \ge 1$ be given and define $\alpha_0 = \int(\mathrm{d}\mu_0 - 1)$. Again using the easy bounds on α_0 and its additivity we find

$$\begin{aligned}0 &\le \mu_0(s,t) - \prod_{\mathcal{T}}(1+\alpha_0) \\ &= \prod_j \mu_j - \prod_j(1+\alpha_j) \\ &= \sum_j \prod_{i<j}\mu_i\,(\mu_j - 1 - \alpha_j)\prod_{k>j}(1+\alpha_k) \\ &\le \sum_j \prod_{i<j}\mu_i\,(\mu_j - 1 - \alpha_j)\prod_{k>j}\mu_k \\ &\le \mu_0(s,t)\Big(\sum_{\mathcal{T}}(\mu_0 - 1) \ - \alpha_0(s,t)\Big).\end{aligned}$$

Again, $\alpha_0(s,t) = \lim_{\mathcal{T}}\sum_{\mathcal{T}}(\mu_0 - 1)$ shows that $\mu_0(s,t) = \lim_{\mathcal{T}}\prod_{\mathcal{T}}(1+\alpha_0)$, and also gives the rest of (8). □

Theorem 1. *Let α be additive, right continuous, and dominated by α_0. Then*

$$\mu = \prodi(1 + \mathrm{d}\alpha) = \lim_{\mathcal{T}} \prod_{\mathcal{T}} (1 + \alpha)$$

exists, is multiplicative, right continuous, and $\mu - 1$ is dominated by $\mu_0 - 1$ where $\mu_0 = \prodi(1 + \mathrm{d}\alpha_0)$. Conversely if μ is multiplicative, right continuous, and $\mu - 1$ is dominated by $\mu_0 - 1$, then

$$\alpha = \int(\mathrm{d}\mu - 1) = \lim_{\mathcal{T}} \sum_{\mathcal{T}} (\mu - 1)$$

exists, is additive, right continuous, and is dominated by $\alpha_0 = \int(\mathrm{d}\mu_0 - 1)$. Finally, $\mu = \prodi(1 + \mathrm{d}\alpha)$ if and only if $\alpha = \int(\mathrm{d}\mu - 1)$.

Proof. Let $\mathcal{S}$ be a refinement of $\mathcal{T}$. Denote by α_i, μ_i the values of α and μ on the elements of $\mathcal{T}$; let $\mathcal{T}_i$ denote the partition of the ith element of $\mathcal{T}$ induced by $\mathcal{S}$; and let α_{ij} denote the values of α on this partition.

Let α be given. Observe that (using, in particular, (3) and (4) of Lemma 1)

$$\begin{aligned}
\prod_{\mathcal{S}}(1+\alpha) - \prod_{\mathcal{T}}(1+\alpha) &= \prod_j \Bigl(\prod_{\mathcal{T}_j}(1+\alpha)\Bigr) - \prod_j (1+\alpha_j) \\
&= \sum_j \prod_{i<j} \prod_{\mathcal{T}_i} (1+\alpha) \Bigl(\prod_{\mathcal{T}_j}(1+\alpha) - 1 - \alpha_j\Bigr) \prod_{k>j}(1+\alpha_k) \\
&= \sum_j \prod_{i<j} \prod_{\mathcal{T}_i} (1+\alpha) \Bigl(\sum_{l,n\,:\,l<n} \alpha_{jl} \prod_{m\,:\,l<m<n} (1+\alpha_{jm})\,\alpha_{jn} \Bigr) \prod_{k>j}(1+\alpha_j).
\end{aligned}$$

Now the final line of this chain of equalities is a sum of products of α_{ij} and α_i. This means that its norm is bounded by the same expression in the norms of the α_{ij} and α_i, which are bounded by α_{0ij} and α_{0i}. But the whole chain of equalities also holds for α_0 itself. Thus we have proved that

$$0 \le \Bigl\| \prod_{\mathcal{S}}(1+\alpha) - \prod_{\mathcal{T}}(1+\alpha) \Bigr\| \le \prod_{\mathcal{S}}(1+\alpha_0) - \prod_{\mathcal{T}}(1+\alpha_0).$$

Therefore existence of the product-integral of α_0 implies existence of the product-integral of α. Moreover, keeping $\mathcal{T}$ as the trivial partition with the single element $(s, t]$ but letting $\mathcal{S}$ become finer and finer, we obtain that $\prodi(1 + \mathrm{d}\alpha) - 1 - \alpha$ is dominated by $\prodi(1 + \mathrm{d}\alpha_0) - 1 - \alpha_0$.

Similarly, if μ is given and $\mu - 1$ is dominated by $\mu_0 - 1$, observe that (using (3) of

Lemma 1)

$$\begin{aligned}\sum_{\mathcal{T}}(\mu-1)-\sum_{\mathcal{S}}(\mu-1)&=\sum_i(\mu_i-1)-\sum_i\sum_{\mathcal{T}_i}(\mu-1)\\&=\sum_i\Bigl(\mu_i-1-\sum_{\mathcal{T}_i}(\mu-1)\Bigr)\\&=\sum_i\Bigl(\prod_{\mathcal{T}_i}(1+(\mu-1))-1-\sum_{\mathcal{T}_i}(\mu-1)\Bigr)\\&=\sum_i\Bigl(\sum_{j,l\,:\,j<l}(\mu_{ij}-1)\prod_{k\,:\,j<k<l}\mu_{ik}\;(\mu_{il}-1)\Bigr).\end{aligned}$$

Now $\|\mu-1\|\le\mu_0-1$ so the norm of the last line is bounded by the same expression in μ_0. Existence of the sum integral $\int(\mathrm{d}\mu_0-1)$ therefore implies existence of $\int(\mathrm{d}\mu-1)$. Again, keeping $\mathcal{T}$ as the trivial partition but letting $\mathcal{S}$ become finer, we obtain that $\mu-1-\int(\mathrm{d}\mu-1)$ is dominated by $\mu_0-1-\int(\mathrm{d}\mu_0-1)$.

For given α, domination of $\mu-1$ by μ_0-1; and for given μ, domination of α by α_0; are both easy to obtain. This implies that if $\mu=\prod\!\!\!\!\!\!\prod(1+\mathrm{d}\alpha)$ with α dominated by α_0 then $\int(\mathrm{d}\mu-1)$ exists; and similarly if we start with μ with $\mu-1$ dominated by μ_0-1.

It remains to show that $\mu=\prod\!\!\!\!\!\!\prod(1+\mathrm{d}\alpha)$ if and only if $\alpha=\int(\mathrm{d}\mu-1)$. In both directions we now have that $\mu-1-\alpha$ is dominated by $\mu_0-1-\alpha_0$. For the forwards implication, we note that $\sum_{\mathcal{T}}(\mu-1)-\alpha=\sum_{\mathcal{T}}(\mu-1-\alpha)$, which is dominated by $\sum_{\mathcal{T}}(\mu_0-1-\alpha_0)$. Taking the limit under refinements of $\mathcal{T}$ shows $\alpha=\int(\mathrm{d}\mu-1)$. Conversely, suppose $\alpha=\int(\mathrm{d}\mu-1)$. Then $\mu-\prod_{\mathcal{T}}(1+\alpha)=\sum_j\prod_{i<j}\mu_i\,(\mu_j-1-\alpha_j)\prod_{k>j}(1+\alpha_k)$. This is dominated by the same expression in μ_0 and α_0, and going to the limit gives the required result. □

Our next task is to show that the product-integral actually exists in a much stronger sense.

Theorem 2. *The product-integral exists as the limit of approximating finite products as the mesh of the partition tends to zero. The limit is uniform over all intervals $(s,t]$ contained in a fixed interval $(0,\tau]$ say.*

Proof. Let α be dominated by α_0. By right continuity and restricting attention to subintervals of the fixed interval $(0,\tau]$, α_0 can be interpreted as an ordinary finite measure. Let α_0^- denote the interval function whose value on $(s,t]$ is obtained by subtracting from $\alpha_0(s,t]$ the α_0 measure of its largest atom in $(s,t]$. For a partition $\mathcal{T}$ let $|\mathcal{T}|$ denote the mesh of the partition, i.e., the length of the largest subinterval in the partition. By a straightforward ε-δ analysis (see also section 12) one can verify that

$$|\mathcal{T}|\;\to\;0\quad\Rightarrow\quad\max_{\mathcal{T}}\alpha_0^-\;\to\;0.$$

For any chosen k, we have $\sum_{i,j\,:\,i<j}\alpha_{0i}\alpha_{0j}\le\sum_i\sum_{j\ne k}\alpha_{0i}\alpha_{0j}$. In particular taking k as the index maximising α_{0k}, and noting that $\max_{\mathcal{T}}\alpha_0$ is at least as large as the largest atom of α_0, we have $\sum_{i,j\,:\,i<j}\alpha_{0i}\alpha_{0j}\le\alpha_0\alpha_0^-$. Now applying (3) with a_j the values of

$1+\alpha$ over a partition $\mathcal{T}$, and taking norms, we obtain

$$\Bigl\|\prod_{\mathcal{T}}(1+\alpha)\ -1-\alpha\Bigr\| \ \le\ \alpha_0\mu_0\alpha_0^-.$$

Going to the limit under refinements of $\mathcal{T}$, we obtain

$$\Bigl\|\mathop{\prod\!\!\!\!\!\!\!\int}(1+\mathrm{d}\alpha)\ -1-\alpha\Bigr\| \ \le\ \alpha_0\mu_0\alpha_0^-.$$

Next we look at (4), taking for a_j the product-integral of $1+\alpha$ over the jth element of a partition $\mathcal{T}$, and for b_j the value of $1+\alpha$ itself. Taking norms and substituting the inequality we have just found for the central term $\|a_j-b_j\|$ we obtain

$$\Bigl\|\mathop{\prod\!\!\!\!\!\!\!\int}(1+\mathrm{d}\alpha)-\prod_{\mathcal{T}}(1+\alpha)\Bigr\| \ \le\ \mu_0\max_{\mathcal{T}}(\alpha_0\mu_0\alpha_0^-)$$

which gives us the required result. □

Let a_j and b_j denote the values on the jth element of a partition $\mathcal{T}$ of a given interval $(s,t]$ of two additive interval functions α and β, both dominated and right-continuous. Let $\mathcal{T}$ be one of a sequence of partitions with mesh converging to zero of this same interval. In equations (1)–(5) one can interpret the summations as integrals (or repeated integrals), with respect to the *fixed measures* α and $\alpha-\beta$, of certain step functions (depending on the partition), constant on the sub-intervals of the partition. Actually since we are looking at $p\times p$ matrices we have, componentwise, finite sums of such real integrals, but this makes no difference to the argument. By our uniformity result the integrands are uniformly close to product-integrals of α or β, taken up to or from an end-point of that sub-interval of the partition through which the variable of integration is passing. The only real complication is that (5) includes a sum of more and more terms. However the mth term of the sum is bounded uniformly by the mth element of the summable sequence $\alpha_0^m/m!$ so gives no difficulties.

All this means that we can go to the limit as $|\mathcal{T}|\to 0$ in (1)–(5) and obtain the following equations:

$$\mathop{\prod\!\!\!\!\!\!\!\int}_{(s,t]}(1+\mathrm{d}\alpha)\ -1\ =\ \int_{u\in(s,t]}\mathop{\prod\!\!\!\!\!\!\!\int}_{(s,u)}(1+\mathrm{d}\alpha)\,\alpha(\mathrm{d}u),\qquad\text{forward integral equation,}\tag{9}$$

$$\mathop{\prod\!\!\!\!\!\!\!\int}_{(s,t]}(1+\mathrm{d}\alpha)\ -1\ =\ \int_{u\in(s,t]}\alpha(\mathrm{d}u)\mathop{\prod\!\!\!\!\!\!\!\int}_{(u,t]}(1+\mathrm{d}\alpha),\qquad\text{backward integral equation,}\tag{10}$$

$$\mathop{\prod\!\!\!\!\!\!\!\int}_{(s,t]}(1+\mathrm{d}\alpha)\ -1-\alpha(s,t)\ =\ \int_{s<u<v\le t}\alpha(\mathrm{d}u)\mathop{\prod\!\!\!\!\!\!\!\int}_{(u,v)}(1+\mathrm{d}\alpha)\,\alpha(\mathrm{d}v),\quad\text{anonymous,}\tag{11}$$

$$\prod_{(s,t]}(1+\mathrm{d}\alpha) - \prod_{(s,t]}(1+\mathrm{d}\beta) \;=\; \int_{u\in(s,t]} \prod_{(s,u)}(1+\mathrm{d}\alpha)\big(\alpha(\mathrm{d}u)-\beta(\mathrm{d}u)\big)\prod_{(u,t]}(1+\mathrm{d}\beta),$$

$$\text{Duhamel, (12)}$$

$$\prod_{(s,t]}(1+\mathrm{d}\alpha) \;=\; 1+\sum_{m=1}^{\infty}\int_{s<u_1<\ldots<u_m\le t}\alpha(\mathrm{d}u_1)\ldots\alpha(\mathrm{d}u_m). \qquad \text{Peano, (13)}$$

Note how the product-integrals inside the ordinary intervals are now over intervals like (s,u), (u,v), or $(v,t]$, corresponding to the strict ordering $i<j<k$ in (1)–(5). **Exercise** to the doubtful reader: write out the proof of one of these equations in full!

It is easy to produce many more identities from (9)–(12). One equation we will come across in the next section is obtained from the Duhamel equation (12) by rewriting it as $\prod(1+\mathrm{d}\alpha+\mathrm{d}\beta) = \prod(1+\mathrm{d}\alpha) + \int\prod(1+\mathrm{d}\alpha+\mathrm{d}\beta)\mathrm{d}\beta\prod(1+\mathrm{d}\alpha)$ and then repeatedly substituting for $\prod(1+\mathrm{d}\alpha+\mathrm{d}\beta)$ in the right-hand side. One sees the terms of an infinite series appearing; the remainder term is easily shown to converge to zero, and we get a generalization of the Peano series:

$$\prod_{(s,t]}(1+\mathrm{d}\alpha+\mathrm{d}\beta) \;=\; \prod_{(s,t]}(1+\mathrm{d}\alpha) \;+$$

$$\sum_{m=1}^{\infty}\int_{s<u_1<\ldots<u_m\le t}\prod_{(s,u_1)}(1+\mathrm{d}\alpha)\beta(\mathrm{d}u_1)\prod_{(u_1,u_2)}(1+\mathrm{d}\alpha)\beta(\mathrm{d}u_2)\ldots\beta(\mathrm{d}u_m)\prod_{(u_m,t]}(1+\mathrm{d}\alpha). \tag{14}$$

This equation is actually a form of the so-called Trotter product formula from the theory of semi-groups (see Masani, 1981). If $\prod_{(s,u]}(1+\mathrm{d}\alpha)$ is nonsingular for all u one can replace each factor $\prod_{(u_i,u_{i+1})}(1+\mathrm{d}\alpha)$ on the right hand side of (14) by $(\prod_{(s,u_i]}(1+\mathrm{d}\alpha))^{-1}\prod_{(s,u_{i+1})}(1+\mathrm{d}\alpha)$. Taking out a factor (on the right) $\prod_{(s,t]}(1+\mathrm{d}\alpha)$ then produces the ordinary Peano series in the measure $\beta'(\mathrm{d}s) = \prod_{(s,u)}(1+\mathrm{d}\alpha)\beta(\mathrm{d}u)(\prod_{(s,u]}(1+\mathrm{d}\alpha))^{-1}$; thus we obtain the generalised Trotter formula:

$$\prod_{(s,t]}(1+\mathrm{d}\alpha+\mathrm{d}\beta) = \prod_{u\in(s,t]}\Big(1+\prod_{(s,u)}(1+\mathrm{d}\alpha)\beta(\mathrm{d}u)\big(\prod_{(s,u]}(1+\mathrm{d}\alpha)\big)^{-1}\Big)\prod_{(s,t]}(1+\mathrm{d}\alpha).$$

Masani (1981) points out the analogy between this formula for the multiplicative integral of a sum and the usual integration by parts formula for additive integration of a product, though he works with $\prod\exp(\mathrm{d}\alpha)$ rather than $\prod(1+\mathrm{d}\alpha)$.

One can consider (9) and (10) as Volterra integral equations by replacing the product-integrals on both sides by an unknown interval function. The solution turns out to be unique; this can be proved by the standard argument (consider the difference of two solutions, which satisfies the same equation with the '-1' removed, and repeatedly substitute left hand side in right hand side). Thus: for given s the unique solution β of

$$\beta(s,t)-1 \;=\; \int_{(s,t]}\beta(s,u-)\alpha(\mathrm{d}u) \tag{15}$$

is $\beta(s,t) = \prod_s^t (1+\mathrm{d}\alpha)$, and for given t the unique solution β of

$$\beta(s,t) - 1 \;=\; \int_{(s,t]} \alpha(\mathrm{d}u)\beta(u,t) \tag{16}$$

is the same. More generally, if ψ is a $q \times p$ matrix càdlàg function (right-continuous with left hand limits) then the unique $q \times p$ matrix càdlàg solution ϕ of

$$\phi(t) \;=\; \psi(t) + \int_{(0,t]} \phi(s-)\alpha(\mathrm{d}s) \tag{17}$$

is

$$\phi(t) \;=\; \psi(t) + \int_{(0,t]} \psi(s-)\alpha(\mathrm{d}s) \prod_{(s,t]} (1+\mathrm{d}\alpha). \tag{18}$$

The notion of *domination* has a measure-theoretic interpretation, close to the usual notion of *bounded variation.* We say that a (possibly matrix valued) interval function β is of bounded variation if and only if its variation, the interval function $|\beta|$ defined by $|\beta| = \sup_{\mathcal{T}} \sum_{\mathcal{T}} \|\beta\|$ is finite and right continuous, where the supremum runs over all partitions of a given interval. It is quite easy to check that β is of bounded variation if and only if β is bounded by an additive right continuous interval function α_0. The sufficiency is obvious, the necessity follows by defining $\alpha_0(s,t) = |\beta|(0,t) - |\beta|(0,s)$. Then trivially $|\beta|(0,t) \geq |\beta|(0,s) + \|\beta(s,t)\|$ giving us as required that $\|\beta\| \leq \alpha_0$. The following special result for multiplicative interval functions is also rather useful:

Proposition 2. *$\mu - 1$ is dominated by $\mu_0 - 1$ if and only if $\mu - 1$ is of bounded variation.*

Proof. $\mu - 1$ of bounded variation implies $\mu - 1$ is dominated by some α_0 which implies $\mu - 1$ is dominated by $\mu_0 - 1 = \prod(1+\mathrm{d}\alpha_0) - 1 \geq \alpha_0$. Conversely, $\mu - 1$ dominated by $\mu_0 - 1$ implies $\sum_{\mathcal{T}} \|\mu - 1\| \leq \sum_{\mathcal{T}} (\mu_0 - 1)$. But the latter sum decreases under refinements; hence it is finite (bounded by $\mu_0 - 1$ itself) and $\mu - 1$ is of bounded variation. □

We close with remarks on possible generalizations of the above theory. The first generalization concerns product-integration over more general time variables than the one-dimensional time t above. What if we replace t by an element of $[0,\infty)^k$ for instance? The answer is that as long as we stick to *scalar* measures α, the above theory can be pretty much reproduced. Restrict attention to subsets of $[0,\infty)^k$ which are (hyper)-rectangles (or finite unions of rectangles), and partitions which are finite sets of rectangles; all the above goes through once we fix an ordering of a finite collection of rectangles. Equations (9)–(13) need however to be carefully formulated. We return to this topic in section 12.

Another generalization is to replace α by the *random* interval function generated by a $p \times p$ matrix *semimartingale.* Now it is known that all our results hold for semimartingales when the product-integral is taken to be the Doléans-Dades exponential semimartingale, in fact defined as the solution to the stochastic integral equation (15) (see Karandikar, 1983). When $p = 1$ it turns out that no deep stochastic analysis is required to get all the results: all one needs is the fact that the (optional) quadratic variation process of the semimartingale exists (in probability) as $\lim_{\mathcal{T}} \sum_{\mathcal{T}} \alpha^2$. Hence the question:

Question. *Is there an elementary (i.e., deterministic) approach to the Doléans-Dades exponential semimartingale in the matrix case which takes as starting point just the existence (as limits of approximating finite sums of products) of the quadratic covariation processes between all components?*

Further background to this question is given by Gill and Johansen (1990). Freedman (1983) develops product-integration for continuous functions of bounded p-variation, $1 < p < 2$ (a different p from the dimension p used till now), and mentions in passing results on the case $p = 2$ and 2×2 matrices.

Most of the above theory can be further generalised to interval functions taking values in a complete normed ring. There are surely many applications making use of such generality, e.g., in the general study of Markov processes (in the next section we will only consider the case of a finite state space).

Exercise. *Find some new applications of product-integration.*

3. Markov processes and product-integrals.

The aim of this section is to put on record the main features of the application of product-integrals to Markov processes, and in preparation for that, to survival times and to the so-called *Bernoulli process.* The results we need are: the survival function is the product-integral of the (negative) hazard and the probability transition matrix is the product-integral of the matrix intensity measure. Later, in sections 7 and 10, we will introduce the connection between hazard or intensity measures and martingale theory.

Survival functions.

First we look at survival functions. Let $T > 0$ be a survival time with survival function S and upper support endpoint τ, $0 < \tau \leq \infty$, i.e., $\tau = \sup\{t : S(t) > 0\}$. Define the hazard measure $\Lambda(\mathrm{d}t) = F(\mathrm{d}t)/S(t-)$ as a measure on $[0, \infty)$; define the interval function $S(s, t) = S(t)/S(s)$ also on $[0, \infty)$ with the convention that $0/0 = 1$. We now have $\Lambda = \int(1 - \mathrm{d}S)$, $S = \prod(1 - \mathrm{d}\Lambda)$ on $[0, \infty]$ if $S(\tau-) > 0$, but otherwise only on $[0, \tau)$. Here are the distinguishing features of the two cases and terminology for them:

(i) *Termination in an atom.* $S(\tau-) > 0$, $S(\tau) = 0$: $\Lambda([0, \tau]) < \infty$, $\sup_{s<\tau} \Lambda(\{s\}) < 1$, $\Lambda(\{\tau\}) = 1$, $\Lambda((\tau, \infty)) = 0$.

(ii) *Continuous termination.* $S(\tau-) = 0$: $\Lambda([0, t]) < \infty$ and $\sup_{s<t} \Lambda(\{s\}) < 1$ for all $t < \tau$, $\Lambda([0, \tau)) = \infty$, $\Lambda([\tau, \infty))$=0.

Every nonnegative measure Λ on $[0, \infty)$ (without an atom at 0) satisfying properties (i) or (ii) corresponds to a survival function of the appropriate type (of a positive random variable). In case (ii) we define $\prod_0^\tau (1 - \mathrm{d}\Lambda) = \lim_{t\uparrow\tau} \prod_0^t (1 - \mathrm{d}\Lambda) = 0$. A defective distribution does not have a termination point. The total hazard is finite and the largest atom of the hazard measure is smaller than 1. In general, the distribution F of a random variable T with hazard measure Λ can be recovered from the hazard by the relation

$$F(\mathrm{d}t) = \prod_{[0,t)} (1 - \mathrm{d}\Lambda)\Lambda(\mathrm{d}t). \tag{1}$$

One can quite easily show that $\prod\!\!\!\!\!\!\;\prod(1-\mathrm{d}\Lambda) = \exp(-\Lambda_c)\prod(1-\Lambda_d)$ where Λ_c and Λ_d are the continuous and discrete parts of Λ respectively. Such a relation holds in general for real product-integrals. We do not emphasize it because in general it is neither intuitive nor useful. One exception is in the construction of the inhomogenous Bernoulli process, to which we now turn.

The inhomogenous Bernoulli process.

Let Λ now be a nonnegative measure on $[0,\infty)$, finite on $[0,t]$ for each $t<\infty$, and whose atoms are less than or equal to one (with no atom at 0). Let Λ_c and Λ_d denote the continuous and discrete parts of Λ and construct a point process on $[0,\infty)$ as follows: to the events of an inhomogenous Poisson process with intensity measure Λ_c add, independently over all atoms of Λ, independent events at the locations t of each atom with probabilities $\Lambda_d(\{t\})$. The probability of no event in the interval $(s,t]$ is $\exp(-\Lambda_c((s,t]))\prod_{(s,t]}(1-\Lambda_d) = \prod\!\!\!\!\!\!\;\prod_s^t(1-\mathrm{d}\Lambda)$. The expected number of events in $(s,t]$ is $\Lambda((s,t])$. Since the expected number of events in finite time intervals is finite, and all events are at distinct times with probability one, the times of the events can be ordered as say $0<T_1<T_2<\ldots$. Define $N(t)=\max\{n : T_n\leq t\}$ as the number of events in $[0,t]$. The process N has independent increments and is therefore Markov.

The distribution of the process can also be characterized through its jump times T_n as follows. Define $S(s,t)=\prod\!\!\!\!\!\!\;\prod_s^t(1-\mathrm{d}\Lambda)$; for given s this is a survival function on $t\geq s$ terminating at the first atom of Λ of size 1 after time s, if any exists; it is defective if there are no such atoms and $\Lambda((s,\infty))<\infty$. First T_1 is generated from the survival function $S(0,\cdot)$. Then, given $T_1=t_1$, $T_2>t_1$ is drawn from $S(t_1,\cdot)$; then given also $T_2=t_2$, $T_3>t_2$ is drawn from $S(t_2,\cdot)$ and so on. One proof for this goes via martingale and counting process theory (to which we return in section 7): by the independent increments property, $N-\Lambda$ is a martingale; now Jacod's (1975) representation of the compensator of a counting process shows how one can read off the conditional distributions of each T_n given its predecessors from the compensator Λ of N. See section 10 or ABGK Theorem II.7.1 for this result stated in terms of product-integration.

Markov processes.

Now we turn to (inhomogeneous) Markov processes with finite state space, continuous time. We suppose the process is defined starting at any time point $t \in [0, \infty)$ from any state, and makes a finite number of jumps in finite time intervals; its sample paths are right continuous stepfunctions. The transition probability matrices $P(s,t) = (\mathrm{P}(X(t) = j | X(s) = i)_{i,j})$ are right continuous and multiplicative. By the theory of product-integration they are product-integrals of a certain interval function or measure Q which we call the *intensity measure* if and only if they are of bounded variation (or dominated) in the sense we described in the previous section. Now

$$\begin{aligned}\|P(s,t)-1\| &= \max_i \sum_j |p_{ij}(s,t) - 1| = 2 \max_i \sum_{j \neq i} p_{ij}(s,t) \\ &\leq 2 \max_i \mathrm{P}(\exists \text{ a jump in } (s,t] | X(s) = i) \\ &\leq 2 \max_i \mathrm{E}(\# \text{ jumps in } (s,t] | X(s) = i).\end{aligned}$$

So a sufficient condition for domination is that the expected number of jumps in any interval, given any starting point at the beginning of the interval, is bounded by a finite measure. This turns out also to be a necessary condition.

If $P - 1$ is dominated then $P = \prod(1 + \mathrm{d}Q)$ where $Q = \int \mathrm{d}(P-1)$ is a dominated, additive matrix-valued measure. Since the elements of P are probabilities and the row sums are 1, the row sums of Q are zero; the diagonal elements are non-positive and the off-diagonal elements non-negative. The atoms of the diagonal elements of Q are not less than -1.

Define

$$\Lambda_i = -Q_{ii}, \qquad \pi_{ij} = \frac{\mathrm{d}Q_{ij}}{\mathrm{d}\Lambda_i}, \quad j \neq i. \tag{2}$$

The $\pi_{ij}(t)$ can be chosen to be a probability measure over $j \neq i$ for each i and t. The Λ_i are nonnegative measures, finite on finite intervals, with atoms at most 1.

Conversely, given Λ_i and π_{ij} (or equivalently given Q) with the just mentioned properties one can construct a Markov process as follows: starting at time s in state i stay there a sojourn time which has survival function $\prod_s^t(1 - \mathrm{d}\Lambda_i)$, $t > s$; on leaving state i at time t jump to a new state j with probability $\pi_{ij}(t)$. We want to show that this process has transition matrices $P = \prod(1 + \mathrm{d}Q)$ where the Q are obtained from the Λ_i and the π_{ij} by using (2) as a definition.

The new process is easily seen to be Markov, though we have not yet ruled out the possibility of it making an infinite number of jumps in finite time. Let $P^*(s,t)$ denote its transition probability matrix for going from any state to any other *with a finite number of jumps*, so that P^* may have row sums less than one. Let $P^{*(n)}$ denote the matrix of transition probabilities when exactly n jumps are made so that $P^* = \sum_{n=0}^\infty P^{*(n)}$. Now by (1), the probability, given we start in state i at time s, of having moved to state j at time t via the chain of states $i = i_0, i_1, \ldots, i_n = j$ (and so with precisely $n > 0$ jumps) is

$$\begin{aligned}\int\limits_{s<t_1<\ldots<t_n\le t} \prod_{(s,t_1)} (1-\mathrm{d}\Lambda_{i_0})\Lambda_{i_0}(\mathrm{d}t_1)\pi_{i_0i_1}(t_1) \prod_{(t_1,t_2)} (1-\mathrm{d}\Lambda_{i_1})\Lambda_{i_1}(\mathrm{d}t_2)\pi_{i_1i_2}(t_2)\ldots \\ \ldots\Lambda_{i_{n-1}}(\mathrm{d}t_n)\pi_{i_{n-1}i_n}(t_n)\prod_{(t_n,t]}(1-\mathrm{d}\Lambda_{i_n});\end{aligned} \tag{3}$$

if $n=0$ then it is just $\delta_{ij}\prod_s^t(1-\mathrm{d}\Lambda_i)$. When we add over all possible chains of length n we obtain the elements of the matrix $P^{*(n)}$. Let $\widetilde{Q}$ denote the matrix of diagonal elements of Q; note that $\widetilde{Q}_{ii}=-\Lambda_i$ and that $\mathrm{d}(Q-\widetilde{Q})_{ij}=\mathrm{d}\Lambda_i\pi_{ij}$. The result of adding over chains can be written (for $n>0$) in abbreviated form as

$$P^{*(n)}=\int\ldots\int\prod(1+\mathrm{d}\widetilde{Q})\,\mathrm{d}(Q-\widetilde{Q})\prod(1+\mathrm{d}\widetilde{Q})\,\mathrm{d}(Q-\widetilde{Q})\ldots\mathrm{d}(Q-\widetilde{Q})\prod(1+\mathrm{d}\widetilde{Q});$$

for $n=0$ we just get $P^{*(0)}=\prod(1+\mathrm{d}\widetilde{Q})$. Now adding over n to get P^* gives us an expression identical to the right hand side of (2.14) with $\alpha=\widetilde{Q}$ and $\beta=Q-\widetilde{Q}$ so $\alpha+\beta=Q$. Thus $P^*=\prod(1+\mathrm{d}Q)=P$, as we wanted to show. Note that since Q has row sums equal to zero, the multiplicands in the approximating finite products for $\prod(1+\mathrm{d}Q)$ have row sums one so P^* is a proper Markov matrix.

The Peano series (2.13) for $\prod(1+\mathrm{d}Q)$ does not have a probabilistic interpretation. What we have shown is that 'expanding about $1+\widetilde{Q}$ instead of about 1' does give a series (2.14) with an important probabilistic interpretation.

The Markov processes having nice sample paths but falling outside of this description are the processes defined probabilistically through Λ_i and π_{ij} as above but where the Λ_i have infinite mass close to some time points. There are two forms of this, according to whether this infinite mass is just before or just after the time point in question. Having infinite mass just before corresponds to an ordinary continuous termination point of the hazard measure for leaving the state, so that the process is certain to leave a certain state by a certain time point, without exit *at* any particular time being certain. (This is only an embarrassment if it is possible to re-enter the state before the termination time, leading to the possibility of infinitely many jumps in finite time. Whether or not this possibility has positive probability depends in general, i.e., when all transitions are always possible, in a complicated way on the joint behaviour of all the Q_{ij} as one approaches the termination time). Another possibility is infinite mass just after a given time point, so the sooner after the time point one enters that state the sooner one leaves it again.

Dobrushin (1954) characterizes when a Markov process is regular (has nice sample paths) as follows. Given a Markov process with transition matrices P, we say that there is infinite hazard of leaving state i just before time t if

$$\lim_{s\uparrow t}\sup_{\mathcal{T}}\sum_{\mathcal{T}}|p_{ii}-1| \;=\; \infty$$

where $\mathcal{T}$ runs through partitions of (s,t), and infinite hazard for leaving i just after t if

$$\lim_{u\downarrow t}\sup_{\mathcal{T}}\sum_{\mathcal{T}}|p_{ii}-1| \;=\; \infty$$

where $\mathcal{T}$ now runs through partitions of $(t,u]$. We say i is inaccessible just before t if

$$p_{ji}(s,u) \;\to\; 0 \quad \text{as } u\uparrow t \text{ for all } s<t \text{ and all } j$$

and inaccessible just after t if

$$p_{ji}(s,u) \;\to\; 0 \quad \text{as } u\downarrow t \text{ for all } s<t \text{ and all } j.$$

Then the process is regular if and only if infinite hazard only occurs for states which are inaccessible at the same time. If all states are always accessible, then the process is regular if and only if P is of bounded variation, and if and only if the expected numbers of jumps are of bounded variation, and if and only if the expected numbers of jumps are just finite.

4. Analytical properties of product-integration.

When we come to statistical problems we need to ask how statistical properties of an estimator of hazard or intensity measure carry over, if at all, to statistical properties of the corresponding estimators of survival functions or transition matrices. Properties of particular concern are: consistency; weak convergence; consistency of the bootstrap or other resampling schemes; asymptotic efficency; and so on. It turns out that many such results depend only on *continuity* and *differentiability* in a certain sense of the product-integral mapping taking dominated right-continuous additive interval functions (possibly matrix valued) to multiplicative ones.

We give more general theory when we come to such applications; for the time being we just show how the Duhamel equation leads naturally to certain continuity and differentiability properties. The reading of this section could be postponed till these applications first arise in section 6.

Fix an interval $[0,\tau]$ and consider the two norms on the right continuous matrix valued interval functions: the supremum norm

$$\|\beta\|_\infty = \sup_{s,t}\|\beta(s,t)\|$$

and the variation norm

$$\|\beta\|_{\mathrm{v}} = \sup_{\mathcal{T}}\sum_{\mathcal{T}}\|\beta\| = \alpha_0(0,\tau)$$

where $\mathcal{T}$ runs through all partitions of $(0,\tau]$ and α_0 is the smallest real measure dominating β (see end of section 2). Write $\overset{\infty}{\to}$ and $\overset{\mathrm{v}}{\to}$ for convergence with respect to these two norms. One easily checks that $\alpha_n \overset{\infty}{\to} \alpha$, $\limsup\|\alpha_n\|_{\mathrm{v}} = M < \infty$ implies $\|\alpha\|_{\mathrm{v}} \le M$.

Now let α and β be two additive interval functions; β will play the role of one of a sequence α_n of such functions approaching α. Let $h = \beta - \alpha$. Consider the difference

$$\begin{aligned}\prod(1+\mathrm{d}\beta) - \prod(1+\mathrm{d}\alpha) &= \int \prod(1+\mathrm{d}\beta)(\mathrm{d}\beta - \mathrm{d}\alpha)\prod(1+\mathrm{d}\alpha)\\ &= \int \prod(1+\mathrm{d}\beta)\mathrm{d}h\prod(1+\mathrm{d}\alpha).\end{aligned} \tag{1}$$

We omit the variables of integration and product-integration; the reader should be able to fill them in but if in doubt look back at the Duhamel equation (2.12) or even the discrete version (2.4). This must be shown to be small when h is small, in supremum norm. Integration by parts is the obvious thing to try: in other words, replace $\prod(1+\mathrm{d}\alpha)$ and $\prod(1+\mathrm{d}\beta)$ by integrals (the Volterra equations!) and then by Fubini reverse orders of integration.

Using the backward and forward integral equations we get

$$\begin{aligned}\int \prod(1+\mathrm{d}\beta)\mathrm{d}h\prod(1+\mathrm{d}\alpha) \;&=\; \int \mathrm{d}h + \int\int \prod(1+\mathrm{d}\beta)\mathrm{d}\beta\mathrm{d}h\\ &+ \int\int \mathrm{d}h\mathrm{d}\alpha\prod(1+\mathrm{d}\alpha) + \int\int\int \prod(1+\mathrm{d}\beta)\mathrm{d}\beta\mathrm{d}h\mathrm{d}\alpha\prod(1+\mathrm{d}\alpha).\end{aligned} \tag{2}$$

Next we can reverse the order of all integrations, carrying out the integration with respect to h *before* that with respect to α or β. One integration simply disappears and h is left as an interval function:

$$\begin{aligned}\int \prod(1+\mathrm{d}\beta)\mathrm{d}h\prod(1+\mathrm{d}\alpha) \;&=\; h + \int \prod(1+\mathrm{d}\beta)\mathrm{d}\beta h\\ &+ \int h\mathrm{d}\alpha\prod(1+\mathrm{d}\alpha)\\ &+ \int\int \prod(1+\mathrm{d}\alpha)\mathrm{d}\alpha h\mathrm{d}\beta\prod(1+\mathrm{d}\beta).\end{aligned} \tag{3}$$

Note that this identity does not depend at all on the original relationship $h = \beta - \alpha$ between α, β and h. For the reader worried about integration variables we write out the last term of (3) in full:

$$\int\int_{s<u<v\leq t} \prod_{s}^{u-}(1+\mathrm{d}\alpha)\alpha(\mathrm{d}u)h(u,v-)\beta(\mathrm{d}v)\prod_{v}^{t}(1+\mathrm{d}\beta).$$

Note also that variation norm boundedness of α and β implies supremum norm boundedness of their product-integrals. Consequently if we do have $h = \beta - \alpha$ then from (1):

$$\Big\| \prod(1+\mathrm{d}\beta) - \prod(1+\mathrm{d}\alpha) \Big\|_\infty \;\leq\; C\|h\|_\infty$$

uniformly in α and β of uniformly bounded variation norm. This is the promised continuity property of product-integration.

We strengthen this now to a *differentiability* result; to be precise, continuous Hadamard (compact) differentiability with respect to the supremum norm, but under a variation norm boundedness condition. This kind of differentiability, intermediate between the more familiar notions of Fréchet (bounded) and Gâteaux (directional) differentiability, is just what we need for various statistical applications as we will see later. Also it seems to be the best result to be hoped for under the chosen norm. We give more of the background theory in an appendix to this section and in section 6, and concentrate now on the bare analysis. The differentiablity result for the product-integral we give here is due to Gill and Johansen (1990). Statistical theory based on compact differentiability is developed in Gill (1989), Wellner (1993), and van der Vaart and Wellner (1993). More applications can be found in Gill, van der Laan and Wellner (1993).

Instead of writing $\beta = \alpha + h$ write $\beta = \alpha + th$ where t is real and close to zero. Compact or Hadamard differentiability means that $(1/t)(\prod(1 + \mathrm{d}\beta) - \prod(1 + \mathrm{d}\alpha))$ can be approximated, for t small, by a continuous linear map in h; the approximation to be uniform over compact sets of h or equivalently along sequences h_n. By continuous compact differentiability we mean that the approximation is also uniform in α (and β). The 'integration by parts' technique we have just used takes us some of the way here. We shall need just one other new technique, taken from the proof of the Helly-Bray lemma and which we will call the Helly-Bray technique.

With $\beta = \alpha + th$ the Duhamel equation gives us immediately, cf. (1),

$$\frac{1}{t}\Big(\prod(1+\mathrm{d}\beta) - \prod(1+\mathrm{d}\alpha)\Big) \;=\; \int \prod(1+\mathrm{d}\beta)\mathrm{d}h \prod(1+\mathrm{d}\alpha). \tag{4}$$

This can be rewritten as in (3), the right hand side of which, considered as a mapping from interval functions h to interval functions, both with the supremum norm, is continuous in h uniformly in α and β of uniformly bounded variation norm. In this way we can interpret $\prod(1 + \mathrm{d}\beta)\mathrm{d}h \prod(1 + \mathrm{d}\alpha)$ also for h which are not of bounded variation simply as the right hand side of (3); a definition by 'formal integration by parts'.

To establish continuous Hadamard differentiability we need to show that $\prod(1 + \mathrm{d}\beta)\mathrm{d}h \prod(1 + \mathrm{d}\alpha)$ is jointly continuous in α, β and h with respect to the supremum norm, for α and β of uniformly bounded variation norm.

Consider a sequence of triples (α_n, β_n, h_n) which converges in supremum norm to (α, β, h), and look at the diference between (3) at the nth stage and at the limit. Assume α_n and β_n (and hence also α, β) are of uniformly bounded variation norm. Since these triples will be related by $\beta_n = \alpha_n + t_n h_n$ where $t_n \to 0$ then in fact $\alpha = \beta$ but this is not important. The Helly-Bray technique is to insert two intermediate pairs (α_n, β_n, h^*) and (α, β, h^*) such that h^* is of bounded variation. Now we have the following telescoping sum:

$$\begin{aligned}\int \prod(1+\mathrm{d}\beta_n)\mathrm{d}h_n \prod(1+\mathrm{d}\alpha_n) &- \int \prod(1+\mathrm{d}\beta)\mathrm{d}h \prod(1+\mathrm{d}\alpha)\\ &= \int \prod(1+\mathrm{d}\beta_n)(\mathrm{d}h_n-\mathrm{d}h^*)\prod(1+\mathrm{d}\alpha_n)\\ &\quad + \left(\int \prod(1+\mathrm{d}\beta_n)\mathrm{d}h^* \prod(1+\mathrm{d}\alpha_n) - \int \prod(1+\mathrm{d}\beta)\mathrm{d}h^* \prod(1+\mathrm{d}\alpha)\right)\\ &\quad + \int \prod(1+\mathrm{d}\beta)(\mathrm{d}h-\mathrm{d}h^*)\prod(1+\mathrm{d}\alpha).\end{aligned}$$

On the right hand side we now have three terms. For the first and the third, integration by parts (transforming to something like (3)) and the bounded variation assumption show that these terms are bounded in supremum norm by a constant times the supremum norm of $h_n - h^*$ and $h - h^*$. The middle term converges to zero as $n \to \infty$ since the product integrals converge in supremum norm and h^* is of bounded variation. Therefore since $\|h_n - h^*\|_\infty \to \|h - h^*\|_\infty$ the 'lim sup' of the supremum norm of the left hand side is bounded by a constant times $\|h-h^*\|_\infty$, which can be made arbitrarily small by choice of h^*. This gives us the required result.

To summarize as a continuous compact differentiability result: for $\alpha'_n = \alpha_n + t_n h_n$ with $\alpha_n \overset{\infty}{\to} \alpha$, $h_n \overset{\infty}{\to} h$, $t_n \to 0$, α_n and α'_n of uniformly bounded variation, we have

$$\frac{1}{t_n}\left(\prod(1+\mathrm{d}\alpha'_n) - \prod(1+\mathrm{d}\alpha_n)\right) \overset{\infty}{\to} \int \prod(1+\mathrm{d}\alpha)\mathrm{d}h \prod(1+\mathrm{d}\alpha) \tag{5}$$

where the right hand side is a (supremum norm) continuous linear mapping in h, interpreted for h not of bounded variation by formal integration by parts (see (3)). It is also jointly continuous in α and h.

A similar but simpler mapping we have to deal with later is ordinary integration of one, say, càdlàg, function on $[0,\tau]$ with respect to another. The mapping $(x,y) \mapsto \int x\mathrm{d}y$ yields a new càdlàg function if we interpret the integration as being over the intervals $(0,t]$ for all $t \in [0,\tau]$. To investigate the continuous differentiability of this mapping, consider $(1/t_n)\big(\int x'_n \mathrm{d}y'_n - \int x_n \mathrm{d}y_n\big) = \int h_n \mathrm{d}y'_n + \int x_n \mathrm{d}k_n$ where $(x'_n, y'_n) = (x_n, y_n) + t_n(h_n, k_n)$, $(x_n, y_n) \overset{\infty}{\to} (x,y)$, $(h_n,k_n) \overset{\infty}{\to} (h,k)$, $t_n \to 0$ (the t_n are real numbers, the rest are càdlàg functions). Assume x_n, y_n, x'_n, y'_n (and consequently x, y too) are of uniformly bounded variation. By the Helly-Bray technique again one easily shows that $\int h_n \mathrm{d}y'_n + \int x_n \mathrm{d}k_n$ converges in supremum norm to $\int h\mathrm{d}y + \int x\mathrm{d}k$ where the second term is interpreted by formal integration by parts if k is not of bounded variation. The limit is a continuous linear map in (h,k), continuously in (x,y). Summarized as a continuous compact differentiabilty result: for $(x'_n, y'_n) = (x_n, y_n) + t_n(h_n, k_n)$, $(x_n, y_n) \overset{\infty}{\to} (x,y)$, $(h_n,k_n) \overset{\infty}{\to} (h,k)$, $t_n \to 0$ where x_n, y_n, x'_n, y'_n (and consequently x, y too) are of uniformly bounded variation,

$$\frac{1}{t_n}\left(\int x'_n \mathrm{d}y'_n - \int x_n \mathrm{d}y_n\right) \overset{\infty}{\to} \int h\mathrm{d}y + \int x\mathrm{d}k. \tag{6}$$

where the right hand side is a (supremum norm) continuous linear mapping in (h, k), interpreted for k not of bounded variation by formal integration by parts. It is also jointly continuous in (x, y) and (h, k). By an easier argument the integration mapping is of course also supremum norm continuous on functions of uniformly bounded variation. See Gill (1989, Lemma 3) or Gill, van der Laan and Wellner (1993) for more details.

In Dudley (1992) these techniques are related to so-called Young-integrals and it is shown that it is not possible to strengthen the results to Fréchet differentiability; at least, not with respect to the supremum norm.

Appendix on Hadamard differentiability.

Here we briefly give definitions of Hadamard differentiability and continuous Hadamard differentiability, see Gill (1989) for further background.

Let B and B' be normed vector spaces, and ϕ a mapping from $E \subseteq B$ to B'. Think for instance of spaces of interval functions under the supremum norm, and the product-integral mapping acting on bounded variation additive interval functions. First we describe a general notion of differentiability of ϕ at a point $x \in E$, then specialize to Fréchet, Hadamard and Gâteaux differentiability, and finally give some special properties of Hadamard differentiability.

Let $\mathrm{d}\phi(x)$ be a bounded linear map from B to B'. This can be considered as the derivative of ϕ if for x' close to x, $\phi(x')$ can be approximated by $\phi(x) + \mathrm{d}\phi(x) \cdot (x' - x)$. (We write $\mathrm{d}\phi(x) \cdot h$ rather than $\mathrm{d}\phi(x)(h)$ to emphasize the linearity of the mapping). Let $\mathcal{S}$ be a set of subsets of B. Then we say ϕ is $\mathcal{S}$-differentiable at x with derivative $\mathrm{d}\phi(x)$ if for each $H \in \mathcal{S}$, $(1/t)\big(\phi(x + th) - \phi(x) - t\mathrm{d}\phi(x) \cdot h\big)$ converges to $0 \in B'$ as $t \to 0 \in \mathbb{R}$ uniformly in $h \in H$ where $x + th$ is restricted to lie in the domain of ϕ.

If one takes $\mathcal{S}$ to be respectively the class of all bounded subsets of B, all compact subsets, or all singletons, then $\mathcal{S}$-differentiability is called Fréchet, Hadamard or Gâteaux differentiability, or bounded, compact or directional differentiability. Clearly Fréchet differentiability is the strongest concept (requires the most uniformity) and Gâteaux the weakest; Hadamard is intermediate. A most important property of Hadamard differentiabilty is that it supports the chain rule: the composition of differentiable mappings is differentiable with as derivative the composition of the derivatives. Hadamard differentiability is in an exact sense the weakest notion of differentiability which supports the chain rule.

Equivalent to the definition just given of Hadamard differentiability is the following: for all sequences of real numbers $t_n \to 0$ and sequences $h_n \to h \in B$,

$$\frac{1}{t_n}\big(\phi(x + t_n h_n) - \phi(x)\big) \;\to\; \mathrm{d}\phi(x) \cdot h,$$

where again $\mathrm{d}\phi(x)$ is required to be a continuous linear map from B to B'. If one strengthens this by requiring also that for all sequences $x_n \to x$

$$\frac{1}{t_n}\big(\phi(x_n + t_n h_n) - \phi(x_n)\big) \;\to\; \mathrm{d}\phi(x)\cdot h,$$

then we say ϕ is continuously Hadamard differentiable at x.

In section 6 we will give a first statistical application of this concept, a functional version of the delta method: weak convergence of an empirical process carries over to compact differentiable functionals of the empirical distribution function. Later (section 11) we will also mention applications in bootstrapping: the bootstrap of a compactly differentiable function of an empirical distribution works in probability, under continuous compact differentiabilty it works almost surely. Another application is in asymptotic optimality theory: compactly differentiable functionals of efficient estimators are also efficient (van der Vaart, 1991a).

5. Nelson-Aalen, Kaplan-Meier, Aalen-Johansen.

Censored survival data can sometimes be realistically modelled as follows. In the background are defined unobservable positive random variables

$$\begin{aligned} T_1,\dots,T_n &\sim \text{ i.i.d. } F\text{; independent of} \\ C_1,\dots,C_n &\sim \text{ i.i.d. } G. \end{aligned}$$

What one observes are, for $i = 1,\dots,n$:

$$\widetilde{T}_i = \min(T_i, C_i) \quad \text{and} \quad \Delta_i = 1\{T_i \le C_i\}.$$

This is known as the (standard or usual) *random censorship model.* We suppose F is completely unknown and the object is to estimate it, or functionals of it, using the observed data. The T_i are called survival times and the C_i censoring times; the $\widetilde{T}_i$ are censored survival times with censoring indicators Δ_i. We occasionally use the notation $\widetilde{T}_{(i)}$ for the ith censored observation in order of size, so that $\widetilde{T}_{(n)}$ is the largest observation. We let $\Delta_{(1)},\dots,\Delta_{(n)}$ be the corresponding censoring indicators; in the case of tied observations we take the uncensored ($\Delta_i = 1$) before the censored ($\Delta_i = 0$). So $\Delta_{(n)} = 0$ if and only if the largest observation, or any one of the equal largest observations if there are several, is censored.

The censoring distribution G may be known or unknown, and sometimes the C_i are observed as well as the $\widetilde{T}_i, \Delta_i$. For instance, suppose patients arrive at a hospital with arrival times A_i according to a Poisson process during the time interval $[0,\tau]$; suppose each patient on arrival is immediately treated and the treatment remains effective for a length of time T_i; suppose the patients are only observed up to time $\sigma > \tau$. From arrival this is a maximum length of time $C_i = \sigma - A_i$. If the process of arrivals has constant intensity then conditional on the number of arrivals we have a random censoring model with observable C_i drawn from the uniform distribution on $[\sigma - \tau, \sigma]$. With an inhomogenous arrival process with unknown intensity, conditional on the number of arrivals we still obtain the random censorship model with observable C_i but now with unknown G.

One may want to condition on the observed C_i, turning them into a collection of known but varying constants; or allow them to be dependent of one another or have varying distributions. Certain kinds of dependence between the censoring and survival times are possible without disturbing some of the analysis we will make. However for the most part we will work in the i.i.d. model described above. For a discussion of many censoring models occurring in practice see chapter III of ABGK.

Write $(\widetilde{T}, \Delta)$ and (T, C) for a generic observation and its unobservable forbears. Let Λ be the hazard function and S the survival function belonging to F. We do not assume F or G to be continuous. We let τ_F and τ_G be the upper support endpoints of F and G and $\tau = \tau_F \wedge \tau_G$. We define the function y by

$$y(t) = (1 - F(t-))(1 - G(t-));$$

it is the left continuous version of the survival function of $\widetilde{T}$. Obviously we can have no information from the data about F outside the time-interval $\{t : y(t) > 0\}$ unless F assigns mass zero to this interval in which case there is nothing else to know.

Intuitively the following seems a natural procedure for estimation of F: with $\mathrm{d}t = [t, t + \mathrm{d}t)$ as before, estimate $\mathrm{P}(T \in \mathrm{d}t \,|\, T \geq t) = \Lambda(\mathrm{d}t)$ by

$$\widehat{\Lambda}(\mathrm{d}t) = \frac{\#\{i : \widetilde{T}_i \in \mathrm{d}t, \Delta_i = 1\}}{\#\{i : \widetilde{T}_i \geq t\}} = \frac{\#\text{failures in } \mathrm{d}t}{\#\text{at risk at time } t-}. \tag{1}$$

Then estimate Λ by $\widehat{\Lambda}(t) = \int_0^t \widehat{\Lambda}(\mathrm{d}s)$ and S by $\widehat{S}(t) = \prod_0^t (1 - \widehat{\Lambda}(\mathrm{d}s))$; finally $\widehat{F} = 1 - \widehat{S}$. A rationale for this procedure would be: *given* $\widetilde{T}_i \geq t$, T_i and C_i are still independent; moreover the events $\{\widetilde{T}_i \in \mathrm{d}t, \Delta_i = 1\}$ and $\{T_i \in \mathrm{d}t\}$ are essentially the same event since C_i *strictly* less than T_i but both times in the same interval $\mathrm{d}t$ can hardly happen. The conditional probability of $\{T_i \in \mathrm{d}t\}$ is $\Lambda(\mathrm{d}t)$ so

$$\mathrm{P}(\widetilde{T}_i \in \mathrm{d}t, \Delta_i = 1 | \widetilde{T}_i \geq t) \approx \Lambda(\mathrm{d}t)$$

motivating (1).

The estimator is a maximum likelihood estimator in some sense, whether or not G is known. Think of the $\widetilde{T}_i$ as being random times and consider how the data grows in time: from one moment to the next there can be some failures, we can see which observations i these belong to, then there can be some censorings and again we can see which observations were involved. Correspondingly, write the likelihood of the n observations as

$$\prod_t \mathrm{P}(\#\text{failures in } \mathrm{d}t | \text{past}) \cdot$$
$$\cdot\, \mathrm{P}(\text{which failures} | \text{past and preceding}) \cdot \mathrm{P}(\text{censorings in } \mathrm{d}t | \text{past and preceding}).$$

The last pair of factors does not involve F; the first factor is a binomial probability, #failures in dt being approximately binomially distributed with parameters #at risk at time $t-$, $\Lambda(\mathrm{d}t)$, given the past. The maximum likelihood estimate of p given $X \sim \mathrm{bin}(n, p)$ is $\widehat{p} = X/n$. Now use transformation invariance to show $\widehat{F}$ is the maximum likelihood estimator of F.

The very informal argument given here is an example of the method of *partial likelihood*, invented by Cox (1975) to justify a somewhat more elaborate (and then rather controversial) estimation procedure in a similar but more complicated context (the Cox, 1972, regression model). Even if the deleted factors in the likelihood had depended on F, perhaps through some assumed relation between F and G, the idea is that one may still delete them and use what is left for valid though perhaps not optimal statistical inference.

The argument is also an example of the derivation of a non-parametric (i.e., infinite dimensional) maximum likelihood estimator. There is a formal definition of this concept, applicable for models like the present where there is no dominating measure and hence no likelihood to maximize; and the estimator we have just derived is then maximum likelihood. However statistical theory of such procedures is still at a rather primitive level and for the moment it is not too important to rigourise the definitions. It is worth pointing out though, that pursuing the likelihood idea further one can write down observed information based estimators of covariance structure which turn out to be asymptotically correct; and that the estimators turn out to have all the asymptotic optimality properties one could hope for. See ABGK section IV.1.5, Gill and van der Vaart (1993) for an attempt to connect these facts together.

The estimators we have just written down have a long history and are the basis of some of the most frequently used techniques in medical statistics. As we shall see they have elegant structure and some beautiful properties. Surprisingly it took a long time to get these properties well mapped out; for instance, the natural version of the Glivenko-Cantelli theorem for $\widehat{F}$ was only obtained in 1991, published 1993, (Stute and Wang), and this was not for want of trying.

The estimators $\widehat{S}$ of the survival function and $\widehat{F}$ of the distribution function were introduced by Kaplan and Meier (1958); apart from being named after them the estimator is also called the product-limit estimator. N. Kaplan and P. Meier actually simultaneously and independently submitted papers to the *Journal of the American Statistical Association* introducing the estimator and their published, joint paper was the result of the postal collaboration which came out of this coincidence. There are precursors in the actuarial literature, see especially Böhmer (1912). The usual formula for the estimated variance of the estimator is affectionately called Major Greenwood's formula (Greenwood, 1926).

It took till 1974 before the first rigorous large sample theory was established for the estimator (Breslow and Crowley, 1974; Meier 1975). These authors confirmed conjectures of Efron (1967), another of whose contributions was to introduce the notion of self-consistency which is important when thinking of the Kaplan-Meier estimator as a nonparametric maximum likelihood estimator (NPMLE); see section 13. We shall demonstrate Breslow and Crowley's method though streamlined through use of product-integration methods, and through using the idea of the functional delta-method (compact differentiability). Weak convergence on the whole line, for which martingale and counting process methods were needed (introduced by Aalen, 1975), was established by this author in 1980, 1983; and as just mentioned, the proper Glivenko-Cantelli theorem had to wait till Stute and Wang (1993). We give a martingale version of that theorem in section 8.

The estimator $\widehat{\Lambda}$ of the hazard function was introduced independently by Altschuler (1970) and Nelson (1969) and generalised greatly by Aalen (1972, 1975, 1976, 1978). It is now known as the Nelson-Aalen estimator. One of the generalizations is in the statistical analysis of censored observations from Markov processes. Suppose a number of particles move according to an inhomogeneous, finite state space Markov process; sometimes they are under observation, sometimes removed from observation. For each pair of states $i \neq j$ estimate the intensity measure Q_{ij} of moving from state i to state j by

$$\widehat{Q}_{ij}(\mathrm{d}t) = \frac{\#i \to j \text{ transitions observed in } \mathrm{d}t}{\#\text{observed at risk for } i \to j \text{ at time } t-};$$

the number in the denominator is the number of particles observed to be in state i at time $t-$. Put these together to form matrices $\widehat{Q}$, and product-integrate to form estimators of transition matrices $\widehat{P}$. The $\widehat{Q}_{ij}$ are 'just' generalised Nelson estimators for the hazard of the $i \to j$ transition, treating other transitions as censorings. Note that in this case the 'number at risk' can also grow through particles entering state i from other states (or elsewhere), whereas in the 'censored survival data' situation it is monotonically decreasing.

These estimation techniques were introduced by Aalen and Johansen (1978), combining Aalen's (1975) earlier developed martingale methods with tools from product-integration. The present author was able to extract from this the martingale approach to the Kaplan-Meier estimator (Gill, 1980, 1983), though neglecting the connections with product-integration.

6. Asymptotics for Kaplan Meier: empirical processes.

We give in this section a first approach to studying the large sample properties of the Kaplan-Meier estimator. This approach uses modern empirical process theory and the analytic properties (compact differentiability) of the product-integration (and ordinary integration) operations given in section 4. The idea is very simple: consider the Kaplan-Meier estimator as a functional of the empirical distribution of the data $(\widetilde{T}_i, \Delta_i)$, $i = 1, \ldots, n$, as represented by its empirical distribution function. The empirical distribution, minus the true, and multiplied by square root of n, converges in distribution to a certain Gaussian process (the celebrated Donsker theorem). The functional which maps empirical distribution function to Kaplan-Meier estimator is compactly differentiable, being the composition of a number of compactly differentiable ingredients (analysed in section 4). Now a generalised version of the delta method, which states that asymptotic normality of a standardized statistic $n^{1/2}(X_n - x)$ carries over to asymptotic normality of $n^{1/2}(\phi(X_n) - \phi(x))$ for any function ϕ differentiable at x, gives weak convergence of $n^{1/2}(\widehat{F} - F)$. Modern empirical process theory being rather elaborate, this approach does involve a lot of technical machinery. However once in working, it delivers a lot of results; in particular bootstrap and efficiency results and multivariate generalizations.

The second approach, introduced in section 7 and further developed in subsequent sections, uses modern (continuous time) martingale methods, again depending on a very elaborate theory. Once the apparatus is set up the results are got very easily and sometimes in stronger versions than by the empirical process approach. Both approaches can equally well be used to study the Aalen-Johansen estimator of the transition probabilities of an inhomogenous Markov process.

Presenting both approaches make it convenient to introduce two sets of notations, so let us first make these clear. For the empirical process approach we let F_n^1 be the empirical subdistribution function of the $\widetilde{T}_i$ with $\Delta_i = 1$; $\widetilde{F}_n$ is the empirical distribution function of all the $\widetilde{T}_i$. For the martingale approach we let N be the process counting observed failures and Y be the process giving the number at risk. So:

$$\begin{aligned} N(t) &= \#\{i : \widetilde{T}_i \le t, \Delta_i = 1\}, \\ Y(t) &= \#\{i : \widetilde{T}_i \ge t\}, \end{aligned}$$

which makes

$$\begin{aligned} F_n^1(t) &= \frac{1}{n} N(t), \\ 1 - \widetilde{F}_n(t-) &= \frac{1}{n} Y(t). \end{aligned}$$

Recall that we have defined $\widehat{\Lambda} = \int \mathrm{d}N/Y$ and $1 - \widehat{F} = \prod(1 - \mathrm{d}\widehat{\Lambda}) = \widehat{S}$. Integration and product-integration here define functions of t by integrating over all intervals $(0, t]$. We will later see that in the martingale approach a key for understanding properties of $\widehat{F} - F$ is that $M = N - \int Y \mathrm{d}\Lambda$ is a zero mean, square integrable martingale with predictable variation process $\langle M \rangle$ given by $\langle M \rangle = \int Y(1 - \Delta\Lambda)\mathrm{d}\Lambda$ where $\Delta\Lambda(t) = \Lambda(\{t\})$ denotes the atoms of Λ.

From empirical process theory we know (by the Glivenko-Cantelli theorem) that F_n^1 and $\widetilde{F}_n$ converge uniformly almost surely for $n \to \infty$ to their expectations $F^1 = \int(1 - G_-)\mathrm{d}F$ and $\widetilde{F} = 1 - (1 - F)(1 - G)$. The integral here denotes the function obtained by integrating over $[0, t]$ for each t, and the subscript minus sign denotes the left continuous version. Note that $\mathrm{d}F^1/(1 - \widetilde{F}_-) = \mathrm{d}F/(1 - F_-) = \mathrm{d}\Lambda$ on $\{t : y(t) > 0\}$; recall $y = (1 - \widetilde{F}_-)$, and $1 - F = \prod(1 - \mathrm{d}\Lambda)$.

We can write now

$$1 - \widehat{F}_n = \prod \Big(1 - \frac{\mathrm{d}F_n^1}{1 - \widetilde{F}_{n\,-}}\Big).$$

This can be thought of as the composition of three mappings:

$$\begin{aligned} (F_n^1, \widetilde{F}_n) &\mapsto \Big(F_n^1, \frac{1}{1 - \widetilde{F}_n}\Big) && (1) \\ &\mapsto \Big(\int \Big(\frac{1}{1 - \widetilde{F}_{n\,-}}\Big) \mathrm{d}F_n^1\Big) && (2) \\ &\mapsto \prod \Big(1 - \mathrm{d}\Big(\int \Big(\frac{1}{1 - \widetilde{F}_{n\,-}}\Big) \mathrm{d}F_n^1\Big)\Big). && (3) \end{aligned}$$

If we fix σ such that $y(\sigma) > 0$ and consider the mappings as applying always to functions on the interval $[0, \sigma]$ then we saw in section 4 that the third of these mappings (product-integration) is supremum norm continuous at functions of uniformly bounded variation;

we also indicated the same result for the second mapping (ordinary integration). The first mapping ('one over one minus') is trivially supremum norm continuous at pairs of functions whose second component is uniformly bounded away from zero. This is satisfied (for $n \to \infty$) by the restriction to $[0, \sigma]$ where $y(\sigma) > 0$. Monotonicity makes the bounded variation condition (asymptotically) easily true.

Applied to the 'true distribution functions' $(F^1, \widetilde{F})$ on the interval $[0, \sigma]$ the mappings yield the true survival function $1 - F$. Glivenko-Cantelli and continuity therefore give us the strong uniform consistency of the Kaplan-Meier estimator: $\widehat{F} \overset{\infty}{\to} F$ almost surely where the convergence is with respect to the supremum norm on $[0, \sigma]$ and $y(\sigma) > 0$.

With martingale methods we will later extend this (in section 8) to uniform convergence simply on $\{t : y(t) > 0\}$, the largest possible interval.

We now turn to asymptotic normality of $n^{\frac{1}{2}}(\widehat{F} - F)$. As we mentioned above we will obtain this by the delta method, in other words a first order Taylor expansion. Before going into the more formal side of this we point out that it is quite easy to work out, by an informal Taylor expansion, what the answer should be. Suppose the distributions under consideration are discrete; the integrals and product integrals in (1)–(3) are now just finite sums and products, involving multinomially distributed numbers of observations taking each possible value. Carry out a first order Taylor expansion to approximate $\widehat{F} - F$ by an expression linear in these variables. The result will again include sums and products which can be rewritten as integrals and product-integrals. This answer will also be correct for the general (non-discrete) case (this is actually a theorem!).

Quite some work is involved, especially to get the final result in a nice form; though several short cuts are possible if one knows where one is going. One obtains

$$\widehat{F}(t) - F(t) \approx \frac{1}{n} \sum_{i=1}^{n} \mathrm{IC}((\widetilde{T}_i, \Delta_i); F; t)$$

where the so-called *influence curve* for the Kaplan-Meier estimator is given by the zero-mean random variable

$$\begin{aligned} &\mathrm{IC}((\widetilde{T}, \Delta); F; t) \\ &= (1 - F(t)) \left(\frac{1\{\widetilde{T} \le t, \Delta = 1\}}{(1 - F(\widetilde{T}))(1 - G(\widetilde{T}-))} - \int_0^{\widetilde{T} \wedge t} \frac{F(\mathrm{d}s)}{(1 - F(s))(1 - F(s-))(1 - G(s-))} \right). \end{aligned}$$

Our aim is to show that the approximate equality here has an exact interpretation as asymptotic equivalence (uniformly in t in certain intervals) with a remainder term of order $o_{\mathrm{P}}(n^{-\frac{1}{2}})$. The asymptotic variance of the Kaplan-Meier estimator is just the variance of the influence curve in one observation; see (5) below.

The mappings above are also compactly differentiable and this gives us weak convergence of $n^{1/2}(\widehat{F} - F)$ in $D[0, \sigma]$ from the Donsker theorem—weak convergence of $\big(n^{1/2}(F^1_n - F^1), n^{1/2}(\widetilde{F}_n - \widetilde{F})\big)$—together with the Skorohod-Dudley almost sure convergence construction (a sequence converging in distribution can be represented by an almost surely convergent sequence on a suitably defined probability space), as we will now show. In Gill (1989) this technique is presented as a functional version of the delta-method. That paper used the weak convergence theory of Dudley (1966) as expounded in Pollard (1984), based on the open ball sigma-algebra.

Here we use weak convergence in the sense of Hoffmann-Jørgensen (see Pollard, 1990, or van der Vaart and Wellner, 1993). This notion of weak convergence is supposed to dispose with the measurability problems which plague general theories of weak convergence, but still great care is needed! We will not dwell on matters of measurability but refer to Wellner (1993) and van der Vaart and Wellner (1993) where the delta method, based on the Hoffmann-Jørgensen weak convergence and compact differentiability, is worked out in full detail. (These authors prefer to use a generalised continuous mapping theorem rather than the almost sure construction to derive the delta method. We have to agree that this is ultimately the more effective approach, but for sentimental reasons we keep to the almost sure construction).

According to the Hoffmann-Jørgensen theory, we see the empirical process $Z_n = \bigl(n^{1/2}(F^1_n - F^1), n^{1/2}(\widetilde{F}_n - \widetilde{F})\bigr)$ as an element of the space of (pairs of) cadlag functions on $[0, \sigma]$ endowed with the supremum norm and the Borel sigma-algebra. As such it is not measurable, but what must be the limiting process—a zero mean Gaussian process with as covariance structure the same structure as that of the empirical—is measurable. Weak convergence in the Donsker theorem (which is true in this context) and subsequent steps means convergence of all *outer expectations* of continuous bounded functions of the empirical process to the ordinary expectations of the same functions of the limiting process.

The Skorohod-Dudley almost sure convergence construction is also available in this set-up. We describe it here as a kind of coupling, i.e., the construction of a joint distribution on a product space with prescribed margins such that given random variables originally defined on the components of the product are as close together as possible. Let Z_n be a sequence of (possibly non-measurable) random elements converging in distribution to a measurable process Z in the sense just described. Suppose the Z_n are defined on probability spaces $(\Omega_n, \mathcal{F}_n, \mathrm{P}_n)$ and Z on $(\Omega, \mathcal{F}, \mathrm{P})$. Form the product of all these spaces together with (the unit interval, Borel sets, Lebesgue measure). Let π_n denote the coordinate projection from the product space to its n'th component (and define π similarly). Then according to the construction there exists a probability measure $\widetilde{\mathrm{P}}$ on the big product space whose projections onto the components of the product are just the original P_n and P; even more, outer expectations and probabilities computed on the product space and computed on the components coincide, or, formally: $(\widetilde{\mathrm{P}})^*\pi_n^{-1} = \mathrm{P}_n^*$. (One says that the coordinate projections π_n are perfect mappings in that they preserve outer as well as ordinary probabilities). Under $\widetilde{\mathrm{P}}$ the Z_n now converge *almost uniformly* to Z: this means that the distance from Z_n to Z, which may not be measurable, is bounded by a measurable sequence converging almost surely to zero.

Now we show how a delta-method theorem follows from combination of the Skorohod-Dudley construction and the definition of compact differentiability. Let X_n be elements of a normed vector space such that

$$Z_n = a_n(X_n - x) \xrightarrow{\mathcal{D}} Z$$

in the sense of Hoffmann-Jørgensen, where $a_n \to \infty$ is a sequence of real numbers. Let ϕ be a function from this space to another normed vector space, compactly differentiable at x in the sense that for all $t_n \to 0$ (real numbers) and all $h_n \to h$,

$$t_n^{-1}(\phi(x+t_nh_n)-\phi(x)) \to \mathrm{d}\phi(x).h$$

where $\mathrm{d}\phi(x)$ is a continuous linear map between the two spaces. By the Skorohod-Dudley almost sure convergence construction we may pretend the Z_n converge almost uniformly to Z. Now apply the definition of differentiability with x as given, $t_n = a_n^{-1}$, $h_n = Z_n$, $h = Z$, so that $x + t_nh_n = x + a_n^{-1}a_n(X_n - x) = X_n$. We obtain that $a_n(\phi(X_n) - \phi(x)) \to \mathrm{d}\phi(x) \cdot Z$ and also that the difference between $a_n(\phi(X_n) - \phi(x))$ and $\mathrm{d}\phi(x)\cdot Z_n$ converges to zero; both these convergences hold almost surely but a further short argument using measurability of the limit process and continuity of the derivative (van der Vaart and Wellner, 1993, Theorem 1.54 (ii)) shows that the convergence in fact holds almost uniformly. Almost uniform convergence implies convergence of outer expectations and hence weak convergence, giving the required results:

$$a_n\big(\phi(X_n)-\phi(x)\big) \overset{\mathcal{D}}{\to} \mathrm{d}\phi(x)\cdot Z$$

and moreover

$$a_n\big(\phi(X_n)-\phi(x)\big) - \mathrm{d}\phi(x)\cdot Z_n \overset{\mathrm{P}}{\to} 0.$$

(The last convergence of a possibly non-measurable sequence is actually 'almost uniformly').

A crucial point is that compact differentiability as defined here satisfies the chain rule (in fact it is the weakest form of differentiabilty to do so). Our three mappings above are each compactly differentiable (the first by a simple calculation, the second two by section 4).

The conclusion is therefore that $n^{1/2}(\widehat{F}-F)$ converges weakly to a certain Gaussian process, obtained by applying the derivatives of the three maps above (continuous linear maps) one after the other to the limit of the empirical process $(Z_n^1, \widetilde{Z}_n) = \big(n^{1/2}(F_n^1 - F^1), n^{1/2}(\widetilde{F}_n-\widetilde{F})\big)$. Also $n^{1/2}(\widehat{F}-F)$ is asymptotically equivalent to these maps applied to the empirical process itself. (All this, with respect to the supremum norm, on a given interval $[0,\sigma]$). Now the map $(x,y) \mapsto (x, 1/(1-y)) = (x,u)$ has derivative $(h,k) \mapsto (h, k/(1-y)^2) = (h,j)$; $(x,u) \mapsto \int(u_-\mathrm{d}x) = v$ has derivative $(h,j) \mapsto \int j_-\mathrm{d}x + \int u_-\mathrm{d}h = \ell$; and for scalar v the mapping $v \mapsto \prod(1-\mathrm{d}v)$ has derivative $\ell \mapsto -\int \prod(1-\mathrm{d}v)\mathrm{d}\ell \prod(1-\mathrm{d}v) = -\prod(1-\mathrm{d}v)\int(1-\Delta v)^{-1}\mathrm{d}\ell$ where $\Delta v = v - v_-$.

Applied to $(h,k) = (Z_n^1, \widetilde{Z}_n)$ at the point $(F^1, \widetilde{F})$ we obtain $(h,j) = (Z_n^1, \widetilde{Z}_n/(1-\widetilde{F})^2)$. The next step gives us $\ell = \int(\widetilde{Z}_{n\,-}/(1-\widetilde{F}_-)^2)\mathrm{d}F^1 + \int(1/(1-\widetilde{F}_-))\mathrm{d}Z_n^1$. Using the fact $(1-\widetilde{F}_-)^{-1}\mathrm{d}F^1 = \mathrm{d}\Lambda$ this simplifies to $\ell = \int(1-\widetilde{F}_-)^{-1}(\mathrm{d}Z_n^1 + \widetilde{Z}_{n\,-}\mathrm{d}\Lambda)$. The final stage therefore takes us to $-\prod(1-\mathrm{d}v)\int(1-\Delta v)^{-1}\mathrm{d}\ell =$

$$-(1-F)\int \frac{1}{(1-\widetilde{F}_-)(1-\Delta\Lambda)}(\mathrm{d}Z_n^1 + \widetilde{Z}_{n\,-}\mathrm{d}\Lambda). \tag{4}$$

We can get rid of the leading minus sign by taking one more step from $1-\widehat{F}$ to $\widehat{F}$. Now one can calculate the asymptotic covariance structure of $n^{1/2}(\widehat{F}-F)$, since it must be the same as that of $(1-F)\int(\mathrm{d}Z_1^1+\widetilde{Z}_{1\,-}\mathrm{d}\Lambda)/((1-\widetilde{F}_-)(1-\Delta\Lambda))$; a tedious calculation (which we leave for the reader to carry out after studying section 7, where martingale methods make it very simple) shows that the covariance of this process evaluated at the time points s and t is

$$(1-F(s))(1-F(t))\int_0^{s\wedge t}\frac{\mathrm{d}\Lambda}{(1-\Delta\Lambda)^2y} \;=\; (1-F(s))(1-F(t))\int_0^{s\wedge t}\frac{\mathrm{d}F}{(1-F)^2(1-G_-)}. \tag{5}$$

This means that the integral in (4) (i.e., dropping the factor $-(1-F)$) has uncorrelated increments. Since the limiting process is Gaussian with zero mean, we have that $n^{1/2}(\widehat{F}-F)$ is asymptotically distributed as $1-F$ times a process with independent, zero mean, increments; hence a martingale. This raises the question whether $n^{1/2}(\widehat{F}-F)/(1-F)$ has the martingale property before passing to the limit, and if so whether that can be used to give an alternative proof. To connect more closely with that approach, we rewrite (4) by noting that on $[0,\sigma]$

$$\begin{aligned} n^{1/2}(\mathrm{d}Z_n^1&+\widetilde{Z}_{n\,-}\mathrm{d}\Lambda)\\ &= n(\mathrm{d}F_n^1-(1-\widetilde{F}_{n\,-})\mathrm{d}\Lambda)-n(\mathrm{d}F^1-(1-\widetilde{F}_-)\mathrm{d}\Lambda)\\ &= \mathrm{d}N-Y\mathrm{d}\Lambda\,: \end{aligned}$$

thus $n^{1/2}(\widehat{F}-F)$ has been shown to be asymptotically equivalent to

$$n^{-\frac{1}{2}}(1-F)\int\frac{1}{(1-\Delta\Lambda)\,y}(\mathrm{d}N-Y\mathrm{d}\Lambda) \tag{6}$$

and it will turn our that the integral here is exactly a martingale. Note that this approximation is identical to the approximation in terms of the influence curve IC given earlier in this section.

To sum up: $n^{1/2}(\widehat{F}-F)/(1-F)$ is asymptotically distributed as a Gaussian martingale, and even asymptotically equivalent to a process which, for each n, is exactly a martingale in t; provided we restrict attention to an interval $[0,\sigma]$ such that $y(\sigma)>0$. Now martingale properties can be a powerful tool. It turns out that, up to a minor modification, $n^{1/2}(\widehat{F}-F)/(1-F)$ is *exactly* a martingale, for reasons intimately connected again with the Duhamel equation and with a basic martingale property connecting the hazard measure to a survival time. The martingale approach can be used, via the martingale central limit theorem, to give an alternative and in many ways more transparent derivation of asymptotic normality of $n^{1/2}(\widehat{F}-F)/(1-F)$. Moreover it yields a host of further results, in particular connected to the extension of the weak convergence result we have just obtained to weak convergence on a 'maximal interval', namely the closure of $\{t : y(t)>0\}$. This is essential if we want to establish large sample properties of statistical procedures based on the Kaplan-Meier estimator at all possible time values; e.g., a Kaplan-Meier based estimate of the mean, or confidence bands for all time-values.

In the next section we will establish the martingale connections and use them in section 8 to prove one main result: the Glivenko-Cantelli theorem

$$\sup_{\{t:y(t)>0\}} |\widehat{F} - F| \to 0 \quad \text{almost surely as } n \to \infty.$$

Amazingly, this basic property of Kaplan-Meier was first established only very recently by Stute and Wang (1993). We follow their elegant proof, but replace their extensive combinatorial calculations by some structural observations involving the Duhamel equation and martingale properties.

In section 9 we will sketch weak convergence results, with statistical applications.

7. The martingale connection.

Recall the following set-up:

$$\begin{aligned} T_1, \ldots, T_n &\sim \text{ i.i.d. } F\text{; independent of} \\ C_1, \ldots, C_n &\sim \text{ i.i.d. } G. \end{aligned}$$

$$\begin{aligned} \widetilde{T}_i = \min(T_i, C_i), \quad \Delta_i &= 1\{T_i \le C_i\}. \\ N(t) &= \#\{i : \widetilde{T}_i \le t, \Delta_i = 1\}, \\ Y(t) &= \#\{i : \widetilde{T}_i \ge t\}. \\ \widehat{\Lambda}(t) &= \int_0^t \frac{N(\mathrm{d}s)}{Y(s)}, \\ 1 - \widehat{F}(t) &= \prod_0^t (1 - \widehat{\Lambda}(\mathrm{d}s)). \end{aligned}$$

We assume $F(0) = G(0) = 0$; let Λ be the hazard measure corresponding to F; and define the maximal interval on which estimation of F is possible by

$$\mathcal{T} = \{t : F(t-) < 1, G(t-) < 1\}$$

together with its upper endpoint

$$\tau = \sup \mathcal{T}.$$

So $\mathcal{T} = [0, \tau)$ or $[0, \tau]$ and $0 < \tau \le \infty$.

The source of many striking properties of the Kaplan-Meier estimator is the Duhamel equation together with the fact that the process M defined by

$$M(t) = N(t) - \int_0^t Y(s)\Lambda(\mathrm{d}s), \quad 0 \le t \le \infty,$$

is a (square integrable, zero mean) martingale on $[0, \infty]$.

Of course the definition of a martingale involves fixing a *filtration*, that is, a collection of sub σ-algebras of the basic probability space on which everything so far is defined, which is increasing and right continuous:

$$\mathcal{F}_s \subseteq \mathcal{F}_t, \quad s \leq t,$$
$$\mathcal{F}_t = \bigcap_{u>t} \mathcal{F}_u.$$

The martingale has to be adapted to the filtration, i.e.,

$$M(t) \text{ is } \mathcal{F}_t\text{-measurable for each } t.$$

The martingale property is then

$$\mathrm{E}(M(t)|\mathcal{F}_s) = M(s) \quad \forall s \leq t. \tag{1}$$

The minimal filtration which can be taken here is obviously $\mathcal{F}_t = \sigma\{M(s) : s \leq t\}$. However any larger filtration still satisfying (1) can be taken here. A rather natural choice is

$$\mathcal{F}_t = \sigma\{\widetilde{T}_i \wedge t, 1\{\widetilde{T}_i \leq t\}, \Delta_i 1\{\widetilde{T}_i \leq t\}; i = 1, \dots n\}. \tag{2}$$

Thus $\mathcal{F}_t$-measurable random variables only depend in a strict sense on 'the data available at time t': the information as to whether or not each $\widetilde{T}_i$ is less than or equal to t, and if so, its actual value and the value of Δ_i.

(We mention briefly to worried probabilists: usually one also assumes that a filtration is *complete* in the sense that $\mathcal{F}_0$ contains all P-null sets of the underlying probability space. However the assumption is not really needed: one can if necessary augment an incomplete filtration with null sets, invoke standard theorems of stochastic analysis, and then drop the null sets again, while choosing suitable versions of the processes one is working with; see Jacod and Shiryaev, 1987).

The claimed martingale property is intuitively easy to understand. It really says: given $\mathcal{F}_{t-}$ (defined as $\mathcal{F}_t$ in (2) but with '$\leq$' replaced by '$<$') there are $Y(t)$ observations still to be made. The conditional probability of an uncensored observation in $[t, t + \mathrm{d}t)$ is $\Lambda(\mathrm{d}t)$. The expected number is therefore $Y(t)\Lambda(\mathrm{d}t)$, thus $\mathrm{E}(N(\mathrm{d}t)|\mathcal{F}_{t-}) = Y(t)\Lambda(\mathrm{d}t)$ or $\mathrm{E}(M(\mathrm{d}t)|\mathcal{F}_{t-}) = 0$. Therefore $M(\mathrm{d}t)$ forms a continuous version of a sequence of martingale differences.

To actually prove the martingale property (1) is a different matter. There are many ways to do it, ranging from direct calculation to the use of general theorems on the *compensator of a counting process*, see Jacod (1975), ABGK section II.7, or section 10 below. Intermediate approaches use some calculation and some stochastic analysis. Since we need to introduce some of that anyway, here is such a hybrid proof. For an extensive introduction to the results from stochastic analysis which we need see Chapter II of ABGK.

The main tool we use is the following: the integral of a predictable process with respect to a martingale is again a martingale, under appropriate integrability conditions. Here is a suitable version of the theorem for our purposes.

Let H be a predictable process: this means that $H = H(t, \omega)$ is measurable with respect to the σ-algebra on $[0, \infty) \times \Omega$ generated by the adapted, left continuous processes. Let M be a martingale with paths of bounded variation on $[0, t]$ for each $t < \infty$.

If $\mathrm{E}\int_0^t |H(s)||M(\mathrm{d}s)| < \infty$ for each t then the process $t \mapsto \int_0^t H\mathrm{d}M$ is again a martingale on $[0, \infty)$. Intuitively, predictability means that $H(t)$ is $\mathcal{F}_{t-}$-measurable. But then $\mathrm{E}(H(t)M(\mathrm{d}t)|\mathcal{F}_{t-}) = H(t)\mathrm{E}(M(\mathrm{d}t)|\mathcal{F}_{t-}) = 0$ so $\int H\mathrm{d}M$ is the continuous time analogue of a sum of martingale differences.

This theorem can be distilled from any standard account of stochastic integration theory, as part of a rather deep and complex theory. A fairly elementary proof is given by Fleming and Harrington (1991).

Now we prove the martingale property (1). Consider the case $n = 1$ and $C = C_1 = \infty$. Thus $N(t) = 1\{T \le t\}$, $Y(t) = 1\{T \ge t\}$, where $T = T_1 \sim F$. First we show $\mathrm{E}M(\infty) = 0$. This follows from $N(\infty) = 1$ a.s. and the fact that

$$\mathrm{E}\biggl(\int_0^\infty Y(t)\Lambda(\mathrm{d}t)\biggr) = \int_0^\infty \mathrm{P}(T \ge t)\frac{F(\mathrm{d}t)}{\mathrm{P}(T \ge t)} = 1.$$

Next consider $M(\infty) - M(t)$; we show that its conditional expectation given the σ-algebra $\mathcal{F}_t = \sigma(T \wedge t, 1\{T \le t\})$ is zero. The conditional expectation can be considered separately on the event $\{T \le t\}$ and on the event $\{T > t\}$. On the former event $M(\infty) - M(t)$ is identically zero so there is nothing more to check. On $\{T > t\}$ we can compute the conditional expectation given $\mathcal{F}_t$ simply as a conditional expectation given $T > t$. Also, on this event, $M(\infty) - M(t) = 1 - \int_t^\infty 1\{T \ge s\}\Lambda(\mathrm{d}s)$. But given $T > t$, T has hazard measure $\Lambda(\mathrm{d}s)1_{(t,\infty)}$. So our previous computation for the case $t = 0$ also applies to this case: we have proved

$$\mathrm{E}\bigl(M(\infty)|\mathcal{F}_t\bigr) \;=\; M(t).$$

One can check that $\mathrm{E}M(\infty)^2 < \infty$ so M is even a square integrable martingale. This also follows from counting process theory since M is a *compensated counting process.*

Now we turn to the general case. Introduce the larger filtration

$$\mathcal{G}_t = \sigma\{T_i \wedge t, 1\{T_i \le t\}, C_i \wedge t, 1\{C_i \le t\}\}.$$

By independence of all T_i's from one another and from all the C_i, we have that the processes M_i^0 defined by

$$M_i^0(t) = 1\{T_i \le t\} - \int_0^t 1\{T_i \ge s\}\Lambda(\mathrm{d}s)$$

are all martingales with respect to $(\mathcal{G}_t)$. Let $H_i(t) = 1\{C_i \ge t\}$. The processes H_i are left continuous and adapted, hence predictable; they are also bounded. Furthermore, it is easy to check $\mathrm{E}\int_0^\infty |H_i(s)||M_i^0(\mathrm{d}s)| < \infty$ and therefore $\int H_i\mathrm{d}M_i^0$ is a martingale for each i. But $\sum_i \int H_i\mathrm{d}M_i^0 = M$ so this is also a martingale, with respect to the filtration $(\mathcal{G}_t)$. Since M is also adapted to the smaller filtration $(\mathcal{F}_t)$, it remains a martingale with respect to this filtration too.

To a square integrable martingale M one can associate its *predictable variation process* $\langle M\rangle$: this is the unique, nondecreasing, predicable process such that $M^2 - \langle M\rangle$ is again a martingale. Intuitively, $\langle M\rangle$ is characterised by

$$\langle M\rangle(\mathrm{d}t) \;=\; \mathrm{E}\bigl(M(\mathrm{d}t)^2 \bigm| \mathcal{F}_{t-}\bigr).$$

Think of $N(\mathrm{d}t)$ as being conditionally $\mathrm{bin}(Y(t), \Lambda(\mathrm{d}t))$ distributed given $\mathcal{F}_{t-}$; since $M(\mathrm{d}t)$ equals $N(\mathrm{d}t)$ minus its conditional expectation, it is plausible that $\langle M\rangle(\mathrm{d}t) = Y(t)\Lambda(\mathrm{d}t)(1-\Lambda(\mathrm{d}t))$. In fact it is true that

$$\langle M\rangle(t) \;=\; \int_0^t Y(s)(1-\Delta\Lambda(s))\Lambda(\mathrm{d}s).$$

The result can be checked by a similar procedure to the one used for the martingale property, using some further results from stochastic calculus. First one must check the result for the case $n = 1, C_1 = \infty$. This can be done by direct calculation (or by appeal to a general result on counting processes described in the next paragraph). Next we use that by independence, the *predictable covariation processes* $\langle M_i^0, M_j^0\rangle$ for $i \neq j$ are all zero; the predictable covariation process of two martingales M and M' is the unique predictable process with paths of locally bounded variation whose difference with the product MM' is a martingale. Finally we use that if H is predictable, M a square integrable martingale, and $\mathrm{E}\int_0^\infty H^2\mathrm{d}\langle M\rangle < \infty$, then $\int H\mathrm{d}M$ is also square integrable, and predictable variation and covariation may be calculated by the rules $\langle \int H\mathrm{d}M\rangle = \int H^2\mathrm{d}\langle M\rangle$, $\langle \int H\mathrm{d}M, \int H'\mathrm{d}M'\rangle = \int HH'\mathrm{d}\langle M, M'\rangle$.

Slightly less work can be done by using general properties of counting processes. Full details of the following outline of a proof can be found in ABGK section II.4. Let N be a counting process: a càdlàg process which is integer valued, zero at time zero, and with jumps of size $+1$ only; for instance the present N in the case $n = 1$. A counting process has a compensator, that is an increasing predictable process A such that $M = N - A$ is a local martingale. The word local means that there exists an increasing sequence of stopping times T_n converging almost surely to ∞ such that the stopped process M^{T_n} defined by $M^{T_n}(t) = M(T_n \wedge t)$ is a martingale for each n. Now consider $M^2 = 2\int M_-\mathrm{d}M + \int \Delta M\mathrm{d}M$. The first term is the stochastic integral of a predictable process with respect to a local martingale so again a local martingale. We further write $\int \Delta M\mathrm{d}M = N - 2\int \Delta A\mathrm{d}N + \int \Delta A\mathrm{d}A$. Since $\int \Delta A\mathrm{d}M = \int \Delta A\mathrm{d}N - \int \Delta A\mathrm{d}A$ and ΔA is again a predictable process, combining terms shows that $\int \Delta A\mathrm{d}M - A + \int \Delta A\mathrm{d}A$ is a local martingale. Thus $M^2 - \int(1-\Delta A)\mathrm{d}A$ is a local martingale or $\langle M\rangle = \int(1-\Delta A)\mathrm{d}A$.

With these tools we can now quickly derive some important martingale properties of $\widehat{F}$ and $\widehat{\Lambda}$. Define

$$\begin{aligned} J(t) \;&=\; 1\{Y(t) > 0\} \\ \frac{J(t)}{Y(t)} \;&=\; \begin{cases} 0 \text{ if } Y(t) = 0 \\ \dfrac{1}{Y(t)} \text{ otherwise.} \end{cases} \end{aligned}$$

Let $\widetilde{T}_{(n)} = \max_i \widetilde{T}_i$ and let Λ^* be the hazard measure $\Lambda^*(\mathrm{d}t) = \Lambda(\mathrm{d}t)1_{[0,\widetilde{T}_{(n)}]}$.

Now we can write

$$\begin{aligned}\widehat{\Lambda} - \Lambda^* &= \int \frac{\mathrm{d}N}{Y} - \int J \mathrm{d}\Lambda \\ &= \int \frac{J}{Y}\mathrm{d}N - \int \frac{J}{Y} Y \mathrm{d}\Lambda \\ &= \int \frac{J}{Y}\mathrm{d}M.\end{aligned}$$

Since J/Y is bounded and predictable and M is square integrable, $\widehat{\Lambda} - \Lambda^*$ is a square integrable martingale with $\langle \widehat{\Lambda} - \Lambda^* \rangle = \int (J/Y)(1 - \Delta\Lambda)\mathrm{d}\Lambda$.

Let $1 - F^* = \prod(1 - \mathrm{d}\Lambda^*)$; equivalently

$$F^*(t) = F(t \wedge \widetilde{T}_{(n)}).$$

Note that $\Lambda^*(T_{(n)}) < \infty$ almost surely and $\widehat{\Lambda}(\infty), \Lambda^*(\infty) < \infty$ almost surely.

By the Duhamel equation, for $t \in [0, \infty]$,

$$\begin{aligned}&\big(1 - \widehat{F}(t)\big) - \big(1 - F^*(t)\big) = \\ &\qquad - \int_0^t \prod_0^{s-} \big(1 - \mathrm{d}\widehat{\Lambda}(\mathrm{d}u)\big)\big(\widehat{\Lambda}(\mathrm{d}s) - \Lambda^*(\mathrm{d}s)\big) \prod_s^t \big(1 - \Lambda^*(\mathrm{d}u)\big)\end{aligned}$$

so dividing by $1 - F^*(t)$,

$$\begin{aligned}\frac{1 - \widehat{F}(t)}{1 - F^*(t)} &= 1 - \int_0^t \frac{\prod_0^{s-}(1 - \widehat{\Lambda}(\mathrm{d}u))}{\prod_0^{s}(1 - \Lambda(\mathrm{d}u))} \frac{J(s)}{Y(s)} M(\mathrm{d}s) \\ &= 1 - \int_0^t \frac{1 - \widehat{F}_-}{1 - F} \frac{J}{Y} \mathrm{d}M.\end{aligned}$$

This gives us that $(1 - \widehat{F})/(1 - F^*) - 1$ is a zero mean, square integrable martingale on $[0, t]$ for any t such that $F(t) < \infty$, with $\langle (1 - \widehat{F})/(1 - F^*) - 1 \rangle = \int ((1 - \widehat{F})/(1 - F))^2 (J(1 - \Delta\Lambda)/Y)\mathrm{d}\Lambda$.

Exercise. Compute the asymptotic variance (6.5) of the Kaplan-Meier estimator by use of stochastic analysis and the approximation (6.6).

In the next section we show how the delicate property of strong uniform consistency follows from this martingale representation and in the section after that we take another look at weak convergence properties from the martingale point of view.

8. Glivenko-Cantelli for Kaplan-Meier.

The analytic properties of the mappings 'integration' and 'product-integration' enabled us in section 6 to establish the following strong consistency result:

$$\sup_{t\in[0,\sigma]} |\widehat{F}(t) - F(t)| \ \to\ 0 \quad \text{a.s. as } n \to \infty \tag{1}$$

for any $\sigma \in \mathcal{T} = \{t : F(t-) < 1, G(t-) < 1\}$. It is now natural to ask: can we replace the interval $[0, \sigma]$ in (1) by the 'maximal interval' $\mathcal{T}$?

It has taken a surprisingly long time to resolve this basic question. Gill (1980) and Shorack and Wellner (1986) give incorrect proofs (the former even for the simpler 'in probability' result). J.-G. Wang (1987) at last gave a correct 'in probability' result and Stute and J.-L. Wang (1993) finally settled the question, in the affirmative. Their approach was completely novel though actually based on a classical technique for proving the ordinary Glivenko-Cantelli theorem. For the ordinary empirical distribution function F_n it is namely known that $F_n(t)$ is a *reverse martingale* in n (t fixed) and Doob's martingale convergence theorem is now available. Stute and Wang (1993) discovered that $\widehat{F}(t)$ (for fixed t) is a *reverse supermartingale* in n.

Here we present a simplified version of their proof, using the Duhamel equation and other martingale properties (in t; n fixed) to replace their extensive combinatorial calculations by a simple analysis of some basic structural features of the Kaplan-Meier estimator. The fact that we have a reverse supermartingale and not a martingale (in n) turns out to be really the same as the fact that in the last section, $\widehat{F} - F^*$ is a martingale in t, making $\widehat{F} - F$ (dropping the star) into a supermartingale.

First we make some general comments on the problem, to indicate why it really is a rather delicate question. If $\tau = \sup \mathcal{T}$ is such that $\tau \in \mathcal{T}$ (so $F(\tau-) < 1$, $G(\tau-) < 1$) then there is nothing more to prove. If $\tau \notin \mathcal{T}$ then either $F(\tau-) = 1$ or $G(\tau-) = 1$, or both. The case $F(\tau-) = 1$ can be handled by an easy monotonicity argument: informally, once we have proved that $\widehat{F}$ is close to F on $[0, \sigma]$ where σ is so close to τ that $F(\sigma)$ is very close to 1, then because $\widehat{F}$ is trapped between $\widehat{F}(\sigma)$ and 1 on (σ, τ), it must also be close to F there. Formally:

$$\begin{aligned}\sup_{t\in\mathcal{T}} |\widehat{F}(t) - F(t)| \ &\le\ \max\{ \sup_{t\in[0,\sigma]} |\widehat{F}(t) - F(t)|, (1 - F(\sigma)) + |\widehat{F}(\sigma) - F(\sigma)|\} \\ &\le\ \sup_{t\in[0,\sigma]} |\widehat{F}(t) - F(t)| + (1 - F(\sigma)).\end{aligned}$$

This means that the only difficult case is the case: $F(\tau-) < 1$, $G(\tau-) = 1$. With probability one in this case, all observations are strictly less than τ. The danger is that for t close to τ where the 'risk set' $\{i : \widetilde{T}_i \ge t\}$ is rather small (e.g., of size 1,2,3,...), a failure occurs, so that $\widehat{\Lambda}$ makes a large jump (of size 1, $\frac{1}{2}$, $\frac{1}{3}$, ...) causing $\widehat{F}$ to make a large jump from a value close to $F(\tau-) < 1$ some appreciable fraction of the way towards 1 (e.g., all the way, half the way, a third of the way, ...).

The in-probability result of J.-G. Wang (1987) is quite easy to obtain once we have obtained this insight. Note that by the Volterra equation

$$1-\widehat{F}(t)=1-\int_0^t(1-\widehat{F}(s-))\widehat{\Lambda}(\mathrm{d}s)$$

it follows that the increment of $\widehat{F}$ over the interval (σ,τ) is less than $\widehat{\Lambda}(\tau-)-\widehat{\Lambda}(\sigma)$ in the case of concern. But we saw that $\widehat{\Lambda}-\Lambda^*$ is a martingale, which implies in the relevant case $F(\tau-)<1, G(\tau-)=1$ that

$$\begin{aligned}\mathrm{E}\big(\widehat{\Lambda}(\tau)-\widehat{\Lambda}(\sigma)\big) &= \mathrm{E}\big(\Lambda^*(\tau)-\Lambda^*(\sigma)\big)\\ &\le \Lambda(\tau-)-\Lambda(\sigma)\end{aligned}$$

which can be made arbitrarily small by taking σ close enough to τ. Now Chebyshev's inequality shows that, uniformly in n, the nonnegative random variable $\widehat{\Lambda}(\tau)-\widehat{\Lambda}(\sigma)$ is arbitrarily small, in probability, for σ close enough to τ, hence

$$\lim_{\sigma\uparrow\tau}\sup_n \mathrm{P}\big(\widehat{F}(\tau)-\widehat{F}(\sigma)>\varepsilon\big) \;=\; 0$$

for all $\varepsilon>0$. Together with

$$\sup_{t\in[0,\sigma]}|\widehat{F}(t)-F(t)| \;\xrightarrow{\mathrm{P}}\; 0$$

for each $\sigma<\tau$, and $\lim_{\sigma\uparrow\tau}(F(\tau-)-F(\sigma))=0$, we obtain

$$\sup_{t\in[0,\tau)}|\widehat{F}(t)-F(t)| \;\xrightarrow{\mathrm{P}} 0.$$

Already a martingale property was involved here. Let us now look at the Stute-Wang strong consistency proof. We do not distinguish between the different special cases any more but give a single proof covering all cases.

The proof will in fact give much more. We will consider any measurable function $\phi\ge 0$, with support in $\mathcal{T}$, i.e., ϕ is zero outside $\mathcal{T}$, and such that $\int_0^\infty\phi\mathrm{d}F<\infty$, and show that

$$\int_0^\infty\phi\mathrm{d}\widehat{F} \;\to\; \int_0^\infty\phi\mathrm{d}F \quad \text{as } n\to\infty \text{ a.s.} \tag{2}$$

The integrals over $[0,\infty)$ can obviously everywhere be replaced by integrals over $\mathcal{T}$. Consider now ϕ equal to indicator functions $1_{[0,\sigma)}$ and $1_{[0,\sigma]}$. We can find a countable set of such indicator functions (e.g.: σ runs through all rationals and all jump points of F in $\mathcal{T}$, together with the point τ itself, though $1_{[0,\tau]}$ is not included if $\tau\notin\mathcal{T}$) such that convergence of $\int\phi\mathrm{d}\widehat{F}$ to $\int\phi\mathrm{d}F$ for all such ϕ implies uniform convergence of $\widehat{F}$ to F on $\mathcal{T}$.

So we only have to consider from now on a sequence of random variables (indexed by sample size n) $\int_{\mathcal{T}}\phi\mathrm{d}\widehat{F}$, ϕ with support in $\mathcal{T}$, $\phi\ge 0$, and $\int_{\mathcal{T}}\phi\mathrm{d}F<\infty$. We will show that this sequence is a nonnegative reverse supermartingale: inserting the variable n

and dropping the range of integration $\mathcal{T}$, this means

$$\mathrm{E}\Big(\int \phi \mathrm{d}\widehat{F}_n \,\Big|\, \int \phi \mathrm{d}\widehat{F}_{n+1}, \int \phi \mathrm{d}\widehat{F}_{n+2}, \ldots\Big) \;\leq\; \int \phi \mathrm{d}\widehat{F}_{n+1}. \tag{3}$$

We also show that $\mathrm{E}(\int \phi \mathrm{d}\widehat{F}_n) \leq \int \phi \mathrm{d}F$ for all n. Doob's supermartingale convergence theorem now implies that $\int \phi \mathrm{d}\widehat{F}_n$ converges almost surely and in expectation to some limiting random variable. However it is not difficult to see that the limit must lie in the tail σ-field generated by the the sequence of observations $(\widetilde{T}_n, \Delta_n)$; therefore by Kolmogorov's zero-one law it must be non-random and equal to the limit of the expected values of the sequence. (Or note that the limit is in the symmetric σ-field generated by the observations hence non-random by the Hewitt-Savage zero-one law). Therefore the required

$$\int \phi \mathrm{d}\widehat{F}_n \;\to\; \int \phi \mathrm{d}F \quad \text{a.s.}$$

will follow from the reverse supermartingale property (3) together with

$$\mathrm{E}\Big(\int \phi \mathrm{d}\widehat{F}_n\Big) \;\to\; \int \phi \mathrm{d}F. \tag{4}$$

We call proving (3) and (4) 'establishing the reverse supermartingale property' and 'identifying the limit' respectively. Stute and Wang (1993) used extensive and quite different looking calculations (combinatorial versus analytic) to prove these two facts. In fact it turns out that in both cases exactly the same martingale ideas can be used. We start with 'identifying the limit'.

Identifying the limit.

In the previous section we showed that the Duhamel equation for comparing $1 - \widehat{F}$ to $1 - F$ could be written in terms of an integral with respect to the basic martingale M. However we only got this martingale structure on the random time interval $[0, T]$ where

$$T = \widetilde{T}_{(n)}$$

due to problems of division by zero. In the previous section we got round this problem by modifying F and looking at F^* instead: this is got from F by forcing its hazard measure to be zero outside $[0, T]$. This technique is the usual one and has been used by many authors.

Here we propose a different trick: namely, instead of modifying F, let us modify $\widehat{F}$, or rather its hazard measure outside $[0, T]$, leaving F itself unchanged. One version of this trick has been known for a long time (Meier, 1975; Mauro, 1985): given T, add to the data one uncensored observation from the distribution with hazard measure $\Lambda(\mathrm{d}t)1_{(T,\infty)}(t)$. This is equivalent in some sense to forcing the largest observation to be uncensored. Another version (Altshuler, 1970) is to add to the data an inhomogenous Bernoulli process, started at time T, with intensity measure $\Lambda(\mathrm{d}t)1_{(T,\infty)}(t)$.

Rather than adding just one observation one could add many; in the limit, this comes down to actually knowing $\Lambda(\mathrm{d}t)1_{(T,\infty)}(t)$. Hence the following

Definition. *$1-\widetilde{F}$ is the survival function with hazard measure $\widetilde{\Lambda}$ equal to $\widehat{\Lambda}(\mathrm{d}t)$ on $[0,T]$, $\Lambda(\mathrm{d}t)$ on (T,∞).*

If $\widehat{F}(T)=1$ then $\widehat{\Lambda}$ terminates properly in an atom of size $+1$ and $\widetilde{F}=\widehat{F}$. If however $\widehat{F}(T)<1$ then $\widehat{\Lambda}$ is finite and has no atom of size $+1$. However the hazard measure $\widetilde{\Lambda}$ corresponding to $\widetilde{F}$ terminates in the same way as Λ at the same point.

We have the following properties of $\widetilde{F}$:

∗ $\widehat{F}$ and $\widetilde{F}$ coincide on $[0,T]$

∗ If $\widehat{F}(T)=1$ then $\widehat{F}$ and $\widetilde{F}$ coincide everywhere

∗ If $\widehat{F}(T)<1$ then $\widetilde{F}$ assigns mass $1-\widehat{F}(T)$ somewhere in (T,∞).

Note that T satisfies almost surely $\Lambda(T)<\infty$, $\widehat{\Lambda}(T)<\infty$.

Now consider the Duhamel equation comparing $\widetilde{F}$ to F, for t such that $\Lambda(t)<\infty$:

$$\begin{aligned}(1-\widetilde{F}(t))-(1-F(t)) &= -\int_0^t \big(1-\widetilde{F}(s-)\big)\big(\widetilde{\Lambda}(\mathrm{d}s)-\Lambda(\mathrm{d}s)\big)\prod_s^t\big(1-\Lambda(\mathrm{d}u)\big)\\ &= -\int_0^t 1\{Y(s)>0\}\big(1-\widehat{F}(s-)\big)\big(\widehat{\Lambda}(\mathrm{d}s)-\Lambda(\mathrm{d}s)\big)\prod_s^t\big(1-\Lambda(\mathrm{d}u)\big).\end{aligned}$$

If F terminates continuously, taking the limit as t tends to the termination point shows that this result actually holds for *all* t. Finally, recalling $\widehat{\Lambda}(\mathrm{d}s)=N(\mathrm{d}s)/Y(s)$ and $J(s)=1\{Y(s)>0\}$, we can rewrite the identity as

$$(1-\widetilde{F}(t))-(1-F(t)) \;=\; -\int_0^t \big(1-\widehat{F}(s-)\big)\frac{J(s)}{Y(s)}M(\mathrm{d}s)\prod_s^t\big(1-\Lambda(\mathrm{d}u)\big). \tag{5}$$

Now M is a square integrable martingale on $[0,\infty]$, $M(0)=0$, and for given t the integrand $(1-\widehat{F}(s-))(J(s)/Y(s))\prod_s^t(1-\mathrm{d}\Lambda)$ is a bounded, predictable process (in s). Therefore the right hand side of (5) is the evaluation at time t of a zero-mean martingale, giving us:

$$\mathrm{E}\widetilde{F}(t)=F(t)\quad\text{for all } t\in[0,\infty].$$

We turn now to integrals $\int_0^\infty \phi\mathrm{d}\widetilde{F}$ of measurable functions ϕ. Consider the class of functions $\phi\geq 0$ such that

$$\mathrm{E}\Big(\int_0^\infty \phi\mathrm{d}\widetilde{F}\Big) \;=\; \int_0^\infty \phi\mathrm{d}F.$$

This class (i) contains all right continuous step functions with a finite number of jumps and (ii) is closed under taking monotone limits, by an easy application (twice) of the monotone convergence theorem. Therefore by the monotone class argument (see, e.g., Protter, 1980, ch. 1, Theorem 8) the class contains *all* nonnegative measurable functions.

From now on restrict attention to $\phi\geq 0$ with support in $\mathcal{T}$ and such that $\int_{\mathcal{T}}\phi\mathrm{d}F<\infty$. We will show that for such ϕ,

$$\mathrm{E}\Big(\int \phi\mathrm{d}\widehat{F}\Big) \;\to\; \int \phi\mathrm{d}F\quad\text{as } n\to\infty.$$

In fact since for *any* ϕ, $\int_0^\infty \phi \mathrm{d}\widehat{F} = \int_0^\infty (\phi 1_{\mathcal{T}}) \mathrm{d}\widehat{F}$ almost surely, this result identifies the limit of $\mathrm{E} \int_0^\infty \phi \mathrm{d}\widehat{F}$ as $\int_{\mathcal{T}} \phi \mathrm{d}F$ for arbitrary F-integrable ϕ.

Fix $M < \infty$ and $\sigma \in \mathcal{T}$ and define $\phi_{\sigma,M} = (\phi \wedge M) 1_{[0,\sigma]}$. Note the following (remember, $\phi \geq 0$):

$$\begin{aligned}
\int \phi \mathrm{d}\widehat{F} &\leq \int \phi \mathrm{d}\widetilde{F} \\
\int \phi_{\sigma,M} \mathrm{d}\widetilde{F} &\leq \int \phi \mathrm{d}\widetilde{F} \\
\int \phi_{\sigma,M} \mathrm{d}\widetilde{F} &= \int \phi_{\sigma,M} \mathrm{d}\widehat{F} \quad \text{if } T \geq \sigma.
\end{aligned}$$

Whether or not $T = \widetilde{T}_{(n)} \geq \sigma$, both sides of the last line are bounded by M; and we have $\mathrm{P}(\widetilde{T}_{(n)} \geq \sigma) \to 1$ as $n \to \infty$. This gives us

$$\begin{aligned}
&\mathrm{E}\Big(\int \phi_{\sigma,M} \mathrm{d}\widehat{F}\Big) \leq \mathrm{E}\Big(\int \phi \mathrm{d}\widehat{F}\Big) \leq \mathrm{E}\Big(\int \phi \mathrm{d}\widetilde{F}\Big) = \int \phi \mathrm{d}F \quad \text{and} \\
&\mathrm{E}\Big(\int \phi_{\sigma,M} \mathrm{d}\widehat{F}\Big) = \mathrm{E}\Big(\int \phi_{\sigma,M} \mathrm{d}\widetilde{F}\Big) + o(1) \quad \text{as } n \to \infty \\
&\qquad = \int \phi_{\sigma,M} \mathrm{d}F + o(1).
\end{aligned}$$

But for ϕ with support in $\mathcal{T}$ and $\int \phi \mathrm{d}F < \infty$, $\int \phi_{\sigma,M} \mathrm{d}F$ can be made arbitrarily close to $\int \phi \mathrm{d}F$ by suitable choice of σ and M. Hence

$$\mathrm{E}\Big(\int \phi \mathrm{d}\widehat{F}\Big) \to \int \phi \mathrm{d}F \quad \text{as } n \to \infty.$$

The reverse supermartingale property.

Consider $n+1$ observations $\widetilde{T}_i, \Delta_i$, $i = 1, \ldots, n+1$. Write $\widetilde{T}_{i:n}$, $i = 1, \ldots n$ and $\widetilde{T}_{i:n+1}$, $i = 1, \ldots n+1$ for the ordered values of respectively $\widetilde{T}_1, \ldots, \widetilde{T}_n$ and $\widetilde{T}_1, \ldots, \widetilde{T}_{n+1}$. Let $\Delta_{i:n}$ and $\Delta_{i:n+1}$ denote the corresponding reordered $\Delta_1, \ldots, \Delta_n$ and $\Delta_1, \ldots, \Delta_{n+1}$. In case of tied values of the $\widetilde{T}_i$, we take the Δ_i with value 1 before those with value 0. From now on we write $\widehat{F}_n$, N_n, Y_n, $\widehat{\Lambda}_n$ and $\widehat{F}_{n+1}$, N_{n+1}, Y_{n+1}, $\widehat{\Lambda}_{n+1}$ to distinguish between statistics based on the first n and the first $n+1$ observations. Note that $\widehat{F}_n$ only depends on the $(\widetilde{T}_i, \Delta_i)$ through the $(\widetilde{T}_{i:n}, \Delta_{i:n})$. This means that

$$\mathcal{F}_n = \sigma\{(\widetilde{T}_{i:n}, \Delta_{i:n}), i \leq n; (\widetilde{T}_i, \Delta_i), i > n\}$$

is a decreasing sequence of σ-algebras to which the sequence $\int \phi \mathrm{d}\widehat{F}_n$ is adapted. The reverse supermartingale property (3) would follow from

$$\mathrm{E}\Big(\int \phi \mathrm{d}\widehat{F}_n \,\Big|\, \mathcal{F}_{n+1}\Big) \leq \int \phi \mathrm{d}\widehat{F}_{n+1}.$$

Since the $(\widetilde{T}_i, \Delta_i)$ for $i > n+1$ are independent of the others and not involved in $\widehat{F}_n$ or

$\widehat{F}_{n+1}$, this comes down to showing

$$\mathrm{E}\Bigl(\int \phi \mathrm{d}\widehat{F}_n \Bigm| \widetilde{T}_{i:n+1}, \Delta_{i:n+1}, i=1,\ldots,n+1\Bigr) \;\leq\; \int \phi \mathrm{d}\widehat{F}_{n+1}. \tag{6}$$

The key observation which will make this calculation really easy is the following fact: the joint distribution of all the $(\widetilde{T}_{i:n}, \Delta_{i:n})$, $(\widetilde{T}_{i:n+1}, \Delta_{i:n+1})$ can be represented by considering the first n pairs as the result of randomly deleting one of last $n+1$. By a random deletion we mean that the index i to be deleted is uniformly distributed on $\{1,\ldots,n+1\}$, independently of all the $(\widetilde{T}_{i:n+1}, \Delta_{i:n+1})$. This means that the conditional expectation in (6) can be computed, given the $(\widetilde{T}_{i:n+1}, \Delta_{i:n+1})$, by averaging over all the $n+1$ values of $\int \phi \mathrm{d}\widehat{F}_n$ obtained by basing $\widehat{F}_n$ on each possible deletion of one element from the $(\widetilde{T}_{i:n+1}, \Delta_{i:n+1})$.

A quick proof of this fact (which is actually not completely trivial, especially when F or G is not continuous) goes as follows. Replace $n+1$ by n for simplicity. The idea is to think of throwing n observations into a bag. Taking them out at random one by one does not change their joint distribution. The last one to come out is a random choice of the ones in the bag to start with. Let $X_1^*,\ldots,X_n^*$ be i.i.d. random vectors from a given distribution. Without loss of generality, assume this distribution has no atoms (otherwise replace the X_i^* by pairs (X_i^*, U_i) where the U_i are independent and uniform $(0,1)$ distributed). Let $I_1,\ldots,I_n$ be a random permutation of $1,\ldots,n$, independent of the X_i^*. Define

$$(X_1,\ldots,X_n) \;=\; (X_{I_1}^*,\ldots,X_{I_n}^*).$$

Now $(X_1,\ldots,X_n)$ is again a random sample from the same given distribution. Moreover the *set* of values $\{X_1,\ldots,X_{n-1}\}$ of the first $n-1$ observations is indeed obtained by random deletion of one element from the set $\{X_1,\ldots,X_n\} = \{X_1^*,\ldots,X_n^*\}$; namely in the second representation of this set we delete the one labelled I_n.

The next idea is to note that the random deletion of one element from the set of $(\widetilde{T}_{i:n+1}, \Delta_{i:n+1})$, which can be thought of as $n+1$ marked points along the line (some of them perhaps at the same position), can be carried out sequentially, in discrete time. Without ties this goes as follows: first decide whether or not to delete $(\widetilde{T}_{1:n+1}, \Delta_{1:n+1})$, with probabilty $1/(n+1)$. If so, stop; if not, move on to $(\widetilde{T}_{2:n+1}, \Delta_{2:n+1})$ and delete it with probabilty $1/n$; and so on. After k failed deletions, delete $(\widetilde{T}_{k+1:n+1}, \Delta_{k+1:n+1})$ with probabilty $1/(n+1-k)$.

When there are ties, the procedure is carried out in exactly the same way but according to the distinct values: after moving through k observations without deletions, delete one of the next group of m tied observations with probability $m/(n+1-k)$; the choice of which of the m to delete is done with equal probabilities.

Now we have set up a discrete time stochastic process description of how N_n and Y_n (and hence $\widehat{\Lambda}_n$ and $\widehat{F}_n$) are generated from N_{n+1} and Y_{n+1}. It will turn out that for this new set up, we have:

$$M_n(t) = N_n(t) - \int_0^t Y_n(s)\widehat{\Lambda}_{n+1}(\mathrm{d}s)$$

is a (discrete time, t) martingale. *Now exactly the same arguments* which related $\mathrm{E}(\int \phi \mathrm{d}\widehat{F})$ to $\int \phi \mathrm{d}F$ via the martingale M, will relate $\mathrm{E}(\int \phi \mathrm{d}\widehat{F}_n)$ to $\int \phi \mathrm{d}\widehat{F}_{n+1}$ via the martingale M_n, where the expectation is now taken with respect to our sequential random deletion experiment for given N_{n+1}, Y_{n+1}.

We prove the new martingale property as follows. Note the following, in which t is one of the values of the $\widetilde{T}_{i:n+1}$:

— if the random deletion has already been made, then $Y_n(t) = Y_{n+1}(t)$, $\Delta N_n(t) = \Delta N_{n+1}(t)$, hence trivially $\Delta N_n(t) = Y_n(t)\Delta\widehat{\Lambda}_{n+1}(t)$.

— if the random deletion has not already been made, then $Y_n(t) = Y_{n+1}(t) - 1$ while

$$\Delta N_n(t) = \begin{cases} \Delta N_{n+1}(t) - 1 \text{ with probability } \Delta N_{n+1}(t)/Y_{n+1}(t) \\ \Delta N_{n+1}(t) \text{ with probability } 1 - \Delta N_{n+1}(t)/Y_{n+1}(t) \end{cases}$$

hence

$$\begin{aligned} \mathrm{E}(\Delta N_n(t)|\text{past}) &= \Delta N_{n+1}(t) - \frac{\Delta N_{n+1}(t)}{Y_{n+1}(t)} \\ &= \frac{\Delta N_{n+1}(t)}{Y_{n+1}(t)}(Y_{n+1}(t) - 1) = Y_n(t)\Delta\widehat{\Lambda}_{n+1}(t). \end{aligned}$$

Combining both cases, $\mathrm{E}(\Delta N_n(t)|\text{past}) = Y_n(t)\Delta\widehat{\Lambda}_{n+1}(t)$.

Therefore $M_n(t)$ is a discrete time martingale. Exactly as in ‘identifying the limit’ introduce $\widetilde{F}_n$ defined to have hazard measure $\widehat{\Lambda}_n(\mathrm{d}t)$ on $\{t : Y_n(t) > 0\}$, $\widehat{\Lambda}_{n+1}(\mathrm{d}t)$ on $\{t : Y_n(t) = 0\}$. We find (cf. (5))

$$(1 - \widetilde{F}_n(t)) - (1 - \widehat{F}_{n+1}(t)) \;=\; -\int_0^t \big(1 - \widetilde{F}_n(s-)\big)\frac{J_n(s)}{Y_n(s)} M_n(\mathrm{d}s) \prod_s^t \big(1 - \widehat{\Lambda}_{n+1}(\mathrm{d}u)\big)$$

for all t, showing, since the integrand (in s) is a predictable process, that $\mathrm{E}\widetilde{F}_n(t) = \widehat{F}_{n+1}(t)$ for all t. Consequently $\mathrm{E}(\int \phi \mathrm{d}\widetilde{F}_n) = \int \phi \mathrm{d}\widehat{F}_{n+1}$. But for $\phi \geq 0$, $\int \phi \mathrm{d}\widehat{F}_n \leq \int \phi \mathrm{d}\widetilde{F}_n$, giving us the reverse supermartingale property: $\mathrm{E}(\int \phi \mathrm{d}\widehat{F}_n) \leq \int \phi \mathrm{d}\widehat{F}_{n+1}$.

One can get further information about $\mathrm{E}(\int \phi \mathrm{d}\widehat{F}_n)$ by considering exactly when $\int \phi \mathrm{d}\widehat{F}_n$ and $\int \phi \mathrm{d}\widetilde{F}_n$ could differ. Since the discrete support of $\widehat{\Lambda}_n$ is contained in that of $\widehat{\Lambda}_{n+1}$, a little reflection shows that the only possibility for a difference is in the mass $\widetilde{F}_n$ and $\widehat{F}_n$ give to the largest observation $t = \widetilde{T}_{n+1:n+1}$, in the case when (for that value of t) $Y_n(t) = 0$ but $Y_{n+1}(t) = 1$. If $\Delta\widehat{\Lambda}_{n+1}(t) = 0$ there is still no difference. So in order for there possibly to be a difference we must have, at sample size $n+1$, a unique largest observation which is furthermore uncensored; and the difference arises precisely when this is the observation to be deleted when stepping down to sample size n. In this case $\widetilde{F}_n$ assigns mass $1 - \widehat{F}_n(t-)$ to this observation while $\widehat{F}_n$ assigns zero mass. Therefore we have:

$$\begin{aligned} \mathrm{E}\Big(\int \phi \mathrm{d}\widehat{F}_n\Big) &= \int \phi \mathrm{d}\widehat{F}_{n+1} \\ &\quad - 1\{Y_{n+1}(t) = 1, \Delta N_{n+1}(t) = 1\} \cdot \phi(t) \cdot \mathrm{E}(1\{Y_n(t) = 0\}(1 - \widehat{F}_n(t-)) \end{aligned}$$

with $t = \widetilde{T}_{n+1:n+1}$. From this equality an interesting representation for the *unconditional* expectation of $\int \phi \mathrm{d}\widehat{F}_n$ can be derived, see Stute and Wang (1993):

$$\begin{aligned}\mathrm{E}\Big(\int \phi \mathrm{d}\widehat{F}_k\Big) &= \int_{\mathcal{T}} \phi \mathrm{d}F \\ &\quad - \sum_{n=k}^{\infty} \mathrm{E}\Big(\phi(\widetilde{T}_{n+1:n+1})(1-\widehat{F}_n(\widetilde{T}_{n:n}))1\{\widetilde{T}_{n:n} < \widetilde{T}_{n+1:n+1}, \Delta_{n+1:n+1}=1\}\Big).\end{aligned}$$

Putting $k = 0$ with the convention $\widehat{F}_0 = 1$, $\widetilde{T}_{0:0} = 0$, $\int \phi \mathrm{d}\widehat{F}_0 = 0$ also gives the curious identity

$$\int_{\mathcal{T}} \phi \mathrm{d}F = \sum_{n=0}^{\infty} \mathrm{E}\Big(\phi(\widetilde{T}_{n+1:n+1})(1-\widehat{F}_n(\widetilde{T}_{n:n}))1\{\widetilde{T}_{n:n} < \widetilde{T}_{n+1:n+1}, \Delta_{n+1:n+1}=1\}\Big).$$

Concluding remarks.

In retrospect the above proof can be made shorter by imitating the proof of weak consistency at the beginning of this section: by the Volterra equation and by the differentiability based proof of uniform consistency on $[0,\sigma]$ for any $\sigma \in \mathcal{T}$, it suffices to show in the crucial case $F(\tau-) < 1$, $G(\tau-) = 1$ that $\widehat{\Lambda}(\tau-) \rightarrow \Lambda(\tau-)$ almost surely as $n \rightarrow \infty$. But we have exactly the same martingale properties in the random deletion experiment relating $\widehat{\Lambda}_n$ to $\widehat{\Lambda}_{n+1}$ as usually hold relating $\widehat{\Lambda}$ to Λ, in particular, $\widetilde{\Lambda}_n - \widehat{\Lambda}_{n+1} = \int (J_n/Y_n)\mathrm{d}M_n$ with $M_n = N_n - \int Y_n \mathrm{d}\Lambda_{n+1}$. This makes $\widehat{\Lambda}_n(\tau-)$ also a reverse supermartingale and the same arguments as above can be used. There seems to be a lot of scope for further results here; for instance, weak convergence as a process jointly in n and t; study of sequential properties of other martingale connected estimators and rank tests; study of 'Kaplan-Meier U-statistics'; investigation of whether similar structure exists with fixed censoring or in the random truncation model (see section 10); and so on.

The discrete time martingale property we have found has parallels in many other combinatorial settings. For instance, bootstrap theory can be done by using the fact that the martingale property of $N - \int Y \mathrm{d}\Lambda$ in the real world carries over to a martingale propery of $N^* - \int Y^* \mathrm{d}\widehat{\Lambda}$ in the bootstrap world (as usual the star denotes the bootstrap version of any statistic). More comments will be made on this (in particular, why it is true) in section 11. Permutation distributions of k-sample rank tests can be studied by using the fact that the permutation distribution of $N_i - \int Y_i \mathrm{d}\widehat{\Lambda}$ is again a martingale, where the index i refers to the ith sample of k and $\widehat{\Lambda}$ is based on combining all k samples; see Andersen, Borgan, Gill and Keiding (1982) and Neuhaus (1992, 1993). In particular the latter author shows the very surprising result that permutation tests can be asymptotically validly made even with unequal censoring distributions, provided the right normalisation is used.

To return to strong consistency, and to be honest, it seems to this author that for statistical purposes, strong consistency is not worth much more than weak consistency. (Despite this comment, section 11 will give yet another approach to Glivenko-Cantelli theorems). In real life n is fixed and both theorems say that for n large, $\widehat{F}_n$ is uniformly close to F with high probability. Strong consistency just suggests a faster rate than weak consistency. In statistics it is more important to get a distributional approximation to $\widehat{F} - F$ so that we can say *how close* $\widehat{F}$ is likely to be to F. The next section will survey such results showing again how martingale methods can be a swift route to getting optimal results. Also we want to draw attention to some serious open problems which seem probabilistically interesting as well as important for applications.

Before this, we should comment on the reverse supermartingale property we have found. Is it just a (probabilistic) coincidence or does it have a deeper (statistical!) significance? The answer is that it is strongly connected to the property of $\widehat{F}$ of being a nonparametric maximum likelihood estimator. In classical statistics, the difference between a maximum likelihood estimator and the true parameter can be approximated as minus the score divided by the information. The score is a martingale in n with variance equal to the information; this makes score divided by information a reverse martingale (i.e., a sample mean is a reverse martingale). So certainly one should not be surprised to find that $\widehat{F}$ is approximately a reverse martingale in n. We have shown that it is almost exactly a reverse martingale; just the censoring of the largest observation can spoil the martingale property.

Further comments on the link to the maximum likelihood property can be found in ABGK (end of section X.2).

9. Confidence bands for Kaplan-Meier.

We saw in section 7 that

$$\frac{\widehat{F} - F^*}{1 - F^*} = \int \frac{1 - \widehat{F}_-}{1 - F} \frac{J}{Y} \mathrm{d}M. \tag{1}$$

This makes $(\widehat{F} - F^*)/(1 - F^*)$ a square integrable martingale on $[0, \sigma]$ for each σ such that $F(\sigma) < 1$. By the recipe $\langle \int H \mathrm{d}M \rangle = \int H^2 \mathrm{d}\langle M \rangle$ we find

$$\langle n^{\frac{1}{2}} \frac{\widehat{F} - F^*}{1 - F^*} \rangle = \int \frac{(1 - \widehat{F}_-)^2}{(1 - F)^2} \frac{nJ}{Y} (1 - \Delta\Lambda) \mathrm{d}\Lambda. \tag{2}$$

Suppose σ also satisfies $G(\sigma-) < 1$, so in fact $y(\sigma) > 0$. For $n \to \infty$ the right hand side converges in probability (by the Glivenko-Cantelli theorem for Y/n and by uniform weak consistency of $\widehat{F}$) to the deterministic, increasing function

$$\begin{aligned} C &= \int \frac{(1 - F_-)^2}{(1 - F)^2} \frac{1}{y} (1 - \Delta\Lambda) \mathrm{d}\Lambda \\ &= \int \frac{\mathrm{d}\Lambda}{(1 - \Delta\Lambda) y}. \end{aligned} \tag{3}$$

If F is continuous we also have that the jumps of $n^{1/2}(\widehat{F} - F^*)/(1 - F^*)$ are uniformly bounded by

$$n^{-\frac{1}{2}} \frac{1}{1 - F(\sigma)} \frac{n}{Y(\sigma)} \xrightarrow{\mathrm{P}} 0 \quad \text{as } n \to \infty.$$

These two facts (when F is continuous) are all that is needed to conclude from Rebolledo's martingale central limit theorem that $n^{1/2}(\widehat{F} - F^*)/(1 - F^*)$ converges in distribution, for $n \to \infty$, to a continuous Gaussian martingale with predictable variation (or variance function) equal to the deterministic function C (see ABGK, section II.5). Here weak convergence on $D[0, \sigma]$ is in the classical sense of weak convergence with the Skorohod metric on $D[0, \sigma]$, but since the limit proces is continuous, this is equivalent to weak convergence in the modern sense with respect to the supremum norm.

When F can have jumps the martingale central limit theorem is not directly applicable. Gill (1980) shows how it can be applied after splitting up the jump of N at time t, conditionally given the past $\mathrm{bin}(Y(t), \Delta\Lambda(t))$ distributed, into $Y(t)$ separate $\mathrm{Bernoulli}(\Delta\Lambda(t))$ distributed jumps at equidistant time points in a small time interval inserted into the real line at time t. On an expanded time interval one gets weak convergence to a continuous process with as variance function a version of the function C, with its jumps linearly interpolated over inserted time intervals. The inserted time intervals can then be deleted again giving a result for the original process.

In any case one has, on $[0, \sigma]$, that with probability converging to 1 the functions F^* and F coincide. Denoting by B the standard Brownian motion, this gives us the final result

$$n^{\frac{1}{2}} \frac{\widehat{F} - F}{1 - F} \overset{\mathcal{D}}{\to} B \circ C \tag{4}$$

on $D[0, \sigma]$, supremum norm, assuming only $F(\sigma) < 1$ and $G(\sigma-) < 1$.

We showed in section 6 how this result followed from empirical process theory and quite a lot of calculations (in fact calculation of the limiting variance was even omitted there). A point we want to make is that once the martingale connections have been made, the conclusion (4), including the formula (3) for the asymptotic variance function, is a completely transparent consequence of the Duhamel equation (1) and the easy computation (2).

In statistical applications this result can be used in many ways. Here we discuss its use in *confidence band* constructions: with one aim being to draw attention to an open problem posed in Gill (1983).

From now on we restrict attention to the case when F is continuous. Let σ, satisfying $y(\sigma) > 0$ be fixed. The function C is not known but it is natural to estimate it by

$$\widehat{C} = \int \frac{n \mathrm{d}\widehat{\Lambda}}{(1 - \Delta\widehat{\Lambda})Y}.$$

This estimator is uniformly weakly consistent on $[0, \sigma]$ (also for discontinuous F). Let

q_α be the $1-\alpha$-quantile of the distribution of $\sup_{0\le s\le 1}|B(s)|$. Then we have:

$$\begin{aligned}
&\lim_{n\to\infty} \mathrm{P}\left(\sup_{[0,\sigma]}\left|n^{\frac12}\frac{\widehat{F}-F}{1-\widehat{F}}\right| > \sqrt{\widehat{C}(\sigma)}q_\alpha\right)\\
&\quad= \lim_{n\to\infty} \mathrm{P}\left(\sup_{[0,\sigma]}\left|n^{\frac12}\frac{\widehat{F}-F}{1-F}\right| > \sqrt{C(\sigma)}q_\alpha\right)\\
&\quad= \mathrm{P}\left(\sup_{[0,\sigma]}\left|\frac{B\circ C}{\sqrt{C(\sigma)}}\right| > q_\alpha\right)\\
&\quad= \mathrm{P}\left(\sup_{[0,1]}|B| > q_\alpha\right) \;=\; 1-\alpha
\end{aligned}$$

since

$$\frac{B\circ C}{\sqrt{C(\sigma)}} \;\sim\; B\circ\left(\frac{C}{C(\sigma)}\right).$$

Thus:

$$\mathrm{P}\left(F \text{ lies between } \widehat{F}\pm n^{-\frac12}\sqrt{\widehat{C}(\sigma)}(1-\widehat{F}) \text{ on } [0,\sigma]\right) \;\to\; 1-\alpha \quad \text{as } n\to\infty;$$

or in other words $\widehat{F}\pm n^{-\frac12}\sqrt{\widehat{C}(\sigma)}(1-\widehat{F})$ is an asymptotic $1-\alpha$ confidence band for F on $[0,\sigma]$. The band is called the Renyi band after its uncensored data analogue (Renyi, 1953) and was introduced independently by Gill (1980) and Nair (1981). It is actually a special case (‘$d=0$’) of a class of bands proposed by Gillespie and Fisher (1979). The similar band for the hazard function was introduced by Aalen (1976).

This band is easy to derive and use in practice but it has two drawbacks. Firstly, in order to use it we must specify σ in advance and the interpretation of the theory is that $Y(\sigma)/n$ must not be very close to zero if we want the true coverage probability of the band to be close to the nominal $1-\alpha$. Secondly, the width of the band is determined strongly by $C(\sigma)$ which suggests that the band ‘concentrates on times close to σ’—it gives a tight interval round $\widehat{F}(t)$ at $t=\sigma$ at the cost presumably of a rather wide interval for small t.

But fortunately many other bands are possible. The Brownian motion is only one of many well understood Gaussian processes, and there are simple transformations changing it into others. Two natural choices are: transformation to a Brownian bridge; and transformation to an Ornstein-Uhlenbeck process. Both transformations address our second problem; the first perhaps is also a solution to the first problem.

For the first transformation we note that the process

$$\frac{B(t)}{1+t} \text{ has covariance structure } \frac{s\wedge t}{(1+s)(1+t)} = \frac{s}{1+s}\left(1-\frac{t}{1+t}\right)$$

for $s < t$. This is a time transformation of the Brownian bridge; defining

$$K = \frac{C}{1+C}$$

and similarly $\widehat{K} = \widehat{C}/(1+\widehat{C})$ we can write, since $1/(1+C) = 1 - K$,

$$(1-K)\,B \circ C \ \sim \ B^0 \circ K$$

where B^0 denotes the Brownian bridge.

Fixing σ as before, we have immediately

$$n^{\frac{1}{2}} \frac{1-K}{1-F}(\widehat{F} - F) \ \xrightarrow{\mathcal{D}} \ B^0 \circ K \tag{6}$$

on $[0, \sigma]$. Letting $q_{\alpha,u}$ denote the $1-\alpha$ quantile of the supremum of the absolute value B^0 on $[0, u]$, $u < 1$, and making use of the uniformly consistent estimator of K on $[0, \sigma]$, we have:

$$\mathrm{P}\biggl(F \text{ lies between } \widehat{F} \pm n^{-\frac{1}{2}} \frac{1-\widehat{F}}{1-\widehat{K}} q_{\alpha, \widehat{K}(\sigma)} \text{ on } [0,\sigma] \biggr) \ \rightarrow \ 1-\alpha \quad \text{as } n \rightarrow \infty; \tag{7}$$

another confidence band for F. This is called the Hall and Wellner band after its inventors, Hall and Wellner (1980). It has the rather attractive property of reducing to a Kolmogorov-Smirnov type band (fixed width) if there is no censoring. At the end of this section we mention another band having this property. (The Hall-Wellner band is actually also a member of the earlier mentioned Gillespie-Fisher class of bands; take '$c = d$').

Now we can describe the open problem: can σ be replaced by the largest observation $\widetilde{T}_{(n)}$ in (7), eliminating the need to choose some σ and getting a confidence band on the largest possible interval?

Certainly one can carry through part of the argument: it turns out that the weak convergence in (6) is true on the maximal interval $[0, \tau]$, *without any further conditions on F or G*; see Gill (1983) and Ying (1989). If we could extend this to

$$n^{\frac{1}{2}} \frac{1-\widehat{K}}{1-\widehat{F}}(\widehat{F} - F) \ \xrightarrow{\mathcal{D}} \ B^0 \circ K \tag{8}$$

on $[0, \tau]$, without conditions on F or G, then the confidence band construction 'on the maximal interval' will be valid too.

The problem is completely open; perhaps the new techniques in the Stute-Wang theorem (section 8) could help resolve this. Possibly (8) is only true subject to some modification, e.g., of $\widehat{K}$, but still leading to something like (7) with $\sigma = \tau$. We think the problem is rather important since so far there is no theorem justifying 'common practice', which is to compute a confidence band on a large interval whose endpoint σ is such that $Y(\sigma)$ is rather small.

The previous transformation seemed canonical in some way—it is the most direct way to transform to a Brownian bridge. However one should note that the number n

enters into the computation of the band in *three places*: not just in the leading $n^{-1/2}$ but also in the weight function $1-\widehat{K}$ and in the quantile $q_{\alpha,\widehat{K}(\sigma)}$ since $\widehat{K}=\widehat{C}/(1+\widehat{C})$ and $\widehat{C} = \int(n\mathrm{d}N)/((Y-\Delta N)Y$. It is easy to check (e.g., by scaling properties of Brownian motion), that replacing n in all these locations by cn for any $0<c<\infty$ keeps (7) true. Alternatively imagine adding to the data many observations censored at zero; n increases but N and Y do not change. So n is 'an arbitrary constant' in this construction. This means that (7) is not quite as canonical as it first seems, and draws some doubt as to whether (7) can be extended to the maximal interval. Still we may pose as open problem: construct asymptotically valid confidence bands for F on the maximal interval $[0,\widetilde{T}_{(n)}]$.

The Brownian bridge process (like Brownian motion) has two nice properties: (i) it is Markov, (ii) it is Gaussian. There is, up to rescaling, exactly one *stationary* Gaussian Markov process and that is the Ornstein-Uhlenbeck process. Can we get from B or B^0 to OU by time and space transformations? Start with B^0. To achieve stationarity we must obviously at least have constant variance. Now the process

$$\frac{B^0(t)}{\sqrt{t(1-t)}} \text{ has covariance structure } \sqrt{\frac{s}{1-s}}\sqrt{\frac{1-t}{t}}$$
$$= \exp\biggl(-\biggl(\log\sqrt{\frac{t}{1-t}}-\log\sqrt{\frac{s}{1-s}}\biggr)\biggr),$$

for $s<t$. Thus letting $\phi(t)=\log\sqrt{(t/(1-t))}$ and $\iota(t)=t$ we see that

$$\frac{B^0}{\sqrt{\iota(1-\iota)}}\circ\phi^{-1} \text{ has covariance structure } \exp(-|u-v|).$$

Thus

$$\begin{aligned}\sqrt{\frac{n}{K(1-K)}}\frac{1-K}{1-F}(\widehat{F}-F) &= n^{\frac{1}{2}}\sqrt{\frac{1-K}{K}}\frac{1}{1-F}(\widehat{F}-F)\\ &= n^{\frac{1}{2}}\frac{\widehat{F}-F}{\sqrt{((1-F)^2C)}} \xrightarrow{\mathcal{D}} OU\circ\log\sqrt{C}\end{aligned}$$

and hence, using consistent estimators,

$$\mathrm{P}\biggl(F \text{ lies between } \widehat{F}\pm q_{\alpha,\log\sqrt{(\widehat{C}(\sigma_2)/\widehat{C}(\sigma_1))}}n^{-1/2}(1-\widehat{F})\sqrt{\widehat{C}} \text{ on } [\sigma_1,\sigma_2]\biggr) \rightarrow 1-\alpha$$

where $q_{\alpha,u}$ is the $1-\alpha$ quantile of the supremum of the absolute value of the Ornstein-Uhlenbeck process over an interval of length u. This band is called the EP band ('equal precision') since each *interval* forming the band has asymptotically equal probability that F passes through it. It was proposed by Nair (1981), omitting unfortunately many important details from an unpublished report of one year before. See also Nair (1984) and Hjort (1985b).

Another possibility is not to transform to a known process but to use analytic methods, simulation, or bootstrapping to obtain or estimate the quantile of the limiting law of an unfamilar process. Going back to Gill's (1983) results, this paper actually establishes, using martingale inequalities to control the right-endpoint problem, weak convergence on the whole line of $n^{\frac{1}{2}} h \cdot (\widehat{F}-F)/(1-F)$ for any nonincreasing weight function h such that $\int_0^\infty h^2 \mathrm{d}C < \infty$; the result for $h=(1-K)=1/(1+C)$ follows since $\int(1/(1+C)^2)\mathrm{d}C = [1/(1+C)] < \infty$. (More nice tail results for Kaplan-Meier using some product-integration and martingale methods are given by Yang, 1992, 1993). Many choices of h can be taken; for instance $h=(1-K)^\alpha$ or $h=y^\alpha$ for $\alpha > \frac{1}{2}$, where $y=(1-F)(1-G)$. In particular the choice $h=y$ leads to the result

$$n^{\frac{1}{2}}(1-G)(\widehat{F}-F) \overset{\mathcal{D}}{\to} y \cdot B \circ C$$

on $[0,\sigma]$, supremum norm. Moreover the techniques based on 'in probability linear bounds' in Gill (1983) show that even

$$n^{\frac{1}{2}}(1-\widehat{G})(\widehat{F}-F) \overset{\mathcal{D}}{\to} y \cdot B \circ C$$

where $1-\widehat{G} = y/(1-\widehat{F})$, the Kaplan-Meier estimator of the censoring distribution.

We will show in section 11 that the bootstrap works for this process: consequently, with stars from now on indicating bootstrap versions, the $1-\alpha$ quantile of the supremum of the absolute value of $y \cdot B \circ C$ can be consistently estimated by that of $n^{\frac{1}{2}}(1-\widehat{G}^*)(\widehat{F}^*-\widehat{F})$ (or if you prefer, $n^{\frac{1}{2}}(1-\widehat{G})(\widehat{F}^*-\widehat{F})$). Denoting this estimated quantile by q^*_α gives us the confidence band $\widehat{F} \pm q^*_\alpha n^{-1/2}/(1-\widehat{G})$ on the whole line:

$$\mathrm{P}\biggl(F \text{ lies between } \widehat{F} \pm q^*_\alpha n^{-1/2}/(1-\widehat{G}) \text{ on } [0,\tau]\biggr) \to 1-\alpha$$

as $n \to \infty$. These bounds reduce to Kolmogorov-Smirnov when there is no censoring, are valid even if F is not continuous, but require a modest simulation experiment to compute. They have a width which for t close to τ (the bigger n, the closer you must get) becomes very large (include values outside $[0,1]$ to both sides) and are therefore not quite what we are looking for. But maybe they are the best we can get.

More details and an alternative derivation of these bands are given in section 11.

10. Point processes, martingales and Markov processes.

The theory of counting processes was so far de-emphasized but it lies at the basis of the martingale connection in our study of the Kaplan-Meier estimator in sections 7 and 9. Also our study of Markov processes in section 3 is incomplete without showing how the matrix intensity measure is involved in a key martingale property of certain counting processes associated with (and equivalent to) the Markov process. The aim of this section is to put the main facts on record, emphasizing the connections with product-integration. The interested reader can follow up the tremendously rich statistical implications of this theory in ABGK.

To begin with we introduce some notation and terminology. Consider a sequence (T_n, J_n), $n=1,2,\ldots$, of random elements where the T_n take values in $(0,\infty]$ and the J_n in some measurable space $(E,\mathcal{E})$. Actually if $T_n=\infty$ then J_n is undefined, or more accurately, takes the value $\emptyset$ for some distinct point $\emptyset \notin E$. We consider the T_n as a sequence of ordered random times of certain events and the J_n as labels or marks describing the nature of the event at each time. We suppose that different events cannot occur simultaneously and that there are no accumulation points or explosions of events: thus, $T_1 > 0$, $T_{n+1} > T_n$ if T_n is finite, otherwise $T_{n+1} = \infty$ too; for all finite τ there exists an n with $T_n > \tau$. We call the process (T_n, J_n) a marked point process with marks in E.

Many stochastic processes can be described in terms of an underlying marked point process. For instance, the paths of a Markov process of the type studied in section 3 can be described, together with the state at time 0, by the times of jumps from one state to another, marked for instance by the label of the new state, or, by the pair of labels (origin state, destination state). This description preserves the time stucture of the process in the strict sense that the description of the Markov process up to time t is equivalent to the description of the marked point process up to time t (together with the intitial state), for every t.

The process (T_n, J_n) can be represented in several other ways: in terms of random measures, and in terms of counting processes. As a random measure, we consider the points (T_n, J_n) as the locations of the atoms of a counting measure μ on the product space $[0, \infty) \times E$. Thus for measurable sets $B \subseteq [0, \infty) \times E$ we define

$$\mu(B) = \#\{n : (T_n, J_n) \in B\}.$$

Another useful representation is in terms of counting processes: for measurable $A \subseteq E$ we define the process N_A by

$$N_A(t) = \mu([0, t] \times A) = \#\{n : T_n \leq t, J_n \in A\}.$$

The *counting processes* N_A are càdlàg, finite, integer valued step functions with jumps of size $+1$ only, zero at time zero, and for disjoint A and A', the processes N_A and $N_{A'}$ do not jump simultaneously. If E is finite then the collection $(N_{\{i\}} : i \in E)$ is called a *multivariate counting process*.

Our aim is to describe the distribution of the point process through a notion of conditional intensity or random intensity measure. This requires us to fix a filtration $(\mathcal{F}_t)$ specifying for each t, what is considered 'to have occurred at or before time t'. We certainly want the point process to be adapted in a proper sense to this filtration: different ways to say the same thing are to assume that all the counting processes N_A are adapted to $(\mathcal{F}_t)$ in the usual sense, or that all the T_n are $(\mathcal{F}_t)$-stopping times with J_n being $\mathcal{F}_{T_n}$ measurable. The point process is called *self-exciting* when the filtration is the minimal filtration to which the process is adapted, commonly denoted $(\mathcal{N}_t)$. Thus $\mathcal{N}_t$ is generated by all $N_A(s)$, $s \leq t$, $A \in \mathcal{E}$, or equivalently by all $1\{T_n \leq t\}, T_n 1\{T_n \leq t\}, J_n 1\{T_n \leq t\}$.

Slightly more general is the case of a filtration 'self-exciting from time 0' by which I mean $\mathcal{F}_t = \mathcal{F}_0 \vee \mathcal{N}_t$ for an arbitrary time-zero sigma algebra $\mathcal{F}_0$. In fact this is not really more general, since, at the cost of allowing the point process to have an event at time zero, one can take the larger mark space $E \cup \Omega, \mathcal{E} \oplus \mathcal{F}_0$ (supposing E and Ω disjoint), and let there be an event at time zero with mark identically equal to ω. The special structure $\mathcal{F}_t = \mathcal{F}_0 \vee \mathcal{N}_t$ allows rather nice results, as we shall soon see, as well as very nice explicit characterisations of stopping times T and such sigma algebras as $\mathcal{F}_T$, $\mathcal{F}_{T-}$ in terms of the paths of the point process; for this we refer to Courrège and Priouret (1966) and Jacobsen (1982). Also conditional expectations can be computed in the intuitively natural way.

To begin with we just suppose the point process is adapted to the filtration. By general process theory (the Doob-Meyer decomposition) the N_A have *compensators* $\widetilde{N}_A$: these are nondecreasing, predictable, càdlàg processes such that for each A

$$M_A = N_A - \widetilde{N}_A$$

is a local square integrable martingale, zero at time zero. The M_A are in fact localized by the (T_n) themselves, i.e., for each n, $M_A^{T_n}$ is a square integrable martingale. By more general process theory (stochastic integration) it turns out that the predictable variation and covaration processes of the M_A can be easily described in terms of the $\widetilde{N}_A$ themselves:

$$\langle M_A, M_{A'} \rangle = \widetilde{N}_{A\cap A'} - \int \Delta \widetilde{N}_A \mathrm{d}\widetilde{N}_{A'}.$$

In particular, $\langle M_A \rangle = \langle M_A, M_A \rangle = \int (1 - \Delta\widetilde{N}_A)\mathrm{d}N_A$ and for disjoint A and A', $\langle M_A, M_{A'} \rangle = -\int \Delta\widetilde{N}_A \mathrm{d}\widetilde{N}_{A'}$. If the compensators $\widetilde{N}_A$ are continuous, even more simplication occurs: $\langle M_A \rangle = \widetilde{N}_A$, $\langle M_A, M_{A'} \rangle = 0$ for disjoint A, A'.

In the random measure approach, one combines all the $\widetilde{N}_A$ into one compensating random measure $\widetilde{\mu}$ defined through the obvious extension procedure by

$$\widetilde{\mu}([0,t] \times A) = \widetilde{N}_A(t).$$

Now we suppose the filtration (or the process) is self-exciting from time 0. In this case it can be shown that, on the event $T_{n-1} \le t < T_n$, conditional expectations given $\mathcal{F}_t$ can be computed as conditional expectations given $\mathcal{F}_0$, given (T_k, J_k), $k = 1, \ldots, n-1$, and given $T_n > t$. Furthermore, the conditional distribution of T_n can be described as the distribution with hazard measure, restricted to (t, ∞), equal to that of T_n conditional only on $\mathcal{F}_0$ and (T_k, J_k), $k = 1, \ldots, n-1$. This is the same as conditioning on $\mathcal{F}_{T_{n-1}}$. Write $\Lambda_{T_n|\mathcal{F}_{T_{n-1}}}$ for the hazard measure of T_n conditional on $\mathcal{F}_0$ and (T_k, J_k), $k = 1, \ldots, n-1$. It turns out that the compensator of the counting process N_A can be described in terms just of these conditional hazard measures together with the conditional distributions of each J_n given $\mathcal{F}_0$, (T_k, J_k), $k = 1, \ldots, n-1$ and given $T_n = t$: on $(T_{n-1}, T_n]$

$$\widetilde{N}_A(\mathrm{d}t) = \Lambda_{T_n|\mathcal{F}_{T_{n-1}}}(\mathrm{d}t)\mathrm{P}(J_n \in A | \mathcal{F}_{T_{n-1}}, T_n = t).$$

Equivalently,

$$\widetilde{\mu}(\mathrm{d}t, \mathrm{d}x) = \sum_{n=1}^{\infty} 1_{(T_{n-1}, T_n]}(t) \Lambda_{T_n|\mathcal{F}_{T_{n-1}}}(\mathrm{d}t) \mathrm{P}_{J_n|\mathcal{F}_{T_{n-1}}, T_n = t}(\mathrm{d}x).$$

This result is due to Jacod (1975). There, the measurability problems associated with choosing proper versions of all these conditional distributions are properly treated. We do not prove the result here but just note that because of the decomposition we have just made into intervals between the jump times, the result needs to be proved for the case of a point process making just one jump. The calculation we made in section 7 proves the result in that case when, moreover, E consists of just one point. The reader might like as an exercise to extend that simple calculation to the case of a finite E.

The result has a simple intuitive content: $\widetilde{\mu}(\mathrm{d}t, \mathrm{d}x)$ is the probability, given the past up to just before time t, to have an event in the small time interval $\mathrm{d}t$ times the conditional probability, given there is an event, that its mark is in $\mathrm{d}x$. The result also shows how one can in principle extract the distribution (given $\mathcal{F}_0$) of the whole point process μ from a description of its compensator $\widetilde{\mu}$: from the trajectories of the $\widetilde{N}_A$ one can extract the conditional hazard measures of the 'next jump times' and given them, the distribution of 'the next jump mark'. In particular, Radon-Nikodym derivatives between two probability distributions can be found by simple algebraic manipulation of formal ratios of the expression

$$\mathrm{dP}|_{\mathcal{F}_0} \cdot \prod_t \Bigl((1 - \widetilde{\mu}(\mathrm{d}t, E))^{1-\mu(\mathrm{d}t,E)} \prod_x \widetilde{\mu}(\mathrm{d}t, \mathrm{d}x)^{\mu(\mathrm{d}t,\mathrm{d}x)} \Bigr).$$

Note the interpretation of this expression as a product of conditional distributions given the past for observing the point process in the infinitesimal time intervals $\mathrm{d}t$. More details are given in ABGK, section II.7.

Markov processes

Now we specialize the above results to Markov processes. For the Markov process of section 3, introduce the space E of pairs of distinct states (i, j). Let $\mathcal{F}_0$ be the sigma-algebra generated by $X(0)$, the state of the process at time 0. Let

$$\begin{aligned} N_{ij}(t) &= \#\text{direct transitions from } i \text{ to } j \text{ in } (0, t], \\ Y_i(t) &= 1\{\text{process is in state } i \text{ at time } t-\}. \end{aligned}$$

Let $\mathcal{F}_t$ be the sigma-algebra generated by $X(0)$ and all $N_{ij}(s)$, $(i, j) \in E$, $s \le t$. Observe that $(\mathcal{F}_t)$ is exactly the same as the filtration generated by the process X itself.

Comparison of our description of $\widetilde{\mu}$ above and the probabilistic construction of the process X from its intensity measure Q in section 3 then shows the following key result:

$$\widetilde{N}_{ij}(\mathrm{d}t) = Y_i(t) Q_{ij}(\mathrm{d}t);$$

or the processes M_{ij} defined by

$$M_{ij}(t) = N_{ij}(t) - \int_0^t Y_i(s) Q_{ij}(\mathrm{d}s)$$

are local square integrable martingales. From the general theory of compensators of counting processes mentioned above we then furthermore have

$$\langle M_{ij}, M_{i'j'}\rangle(t) = \delta_{i,i'} \int_0^t Y_i(s)(\delta_{j,j'} - \Delta Q_{ij}(s))Q_{ij'}(\mathrm{d}s).$$

This is the starting point for a martingale based analysis of Aalen-Johansen estimators of P (the probability transition matrix of the process) based on censored observations of the process, exactly parallel to our study of the Kaplan-Meier estimator sketched in section 9. For further details see Aalen and Johansen (1978), ABGK Section IV.1.3.

We conclude with some remarks concerning related problems. The *random truncation* problem concerns estimation of the distribution F of a positive random variable T, given i.i.d. observations of pairs (C, T) drawn from the *conditional* distribution of C, T given $T > C$, where $C > 0$ is (unconditionally) independent of T and also has a completely unknown distribution. Keiding and Gill (1990) show that the joint (conditional) distribution of C, T can be represented as a Markov process which starts at time 0 in a state 'waiting', at time C moves to a state 'at risk', and at time T to a state 'failure'. The transitions are called 'entry' and 'death' respectively. The (non-trivial) point here is that having completely unknown distributions for C and T corresponds to having completely unknown transition intensity measures $Q_{\text{waiting, at risk}}$ and $Q_{\text{at risk, failure}}$. The latter is moreover identical to the hazard measure corresponding to F. So results on nonparametric estimation of F can be extracted from results on Nelson-Aalen and Aalen-Johansen estimators without further work needed, once the identification between the random truncation model and the Markov model has been made.

Often of interest in practice are so-called semi-Markov or Markov renewal processes. These can be described here as a point process with state space the set of all pairs (i, j) (not just different pairs). Let, for each i, Q_{ij} denote a set of (defective) hazard measures such that $\sum_j Q_{ij}$ is also a hazard measure. We interpret an event with mark (i, j) as a jump from state i to state j and introduce N_{ij} and Y_i as before, and an initial state $X(0)$. Let $L(t)$ be the elapsed time since the last jump of the process strictly before time t. So L has left continuous paths, zero just after each jump time and then increasing linearly with slope $+1$ up to and including the next jump time. Then, the process is semi-Markov with intensity measures Q_{ij} means that N_{ij} has compensator $\widetilde{N}_{ij}$ given by

$$\widetilde{N}_{ij}(\mathrm{d}t) = Y_i(t)Q_{ij}(\mathrm{d}L(t)).$$

An ordinary renewal process has just one state.

In Gill (1981) it is shown how counting process methods can be used to study Nelson-Aalen and Kaplan-Meier type estimators for censored observations from a Markov renewal process, despite the occurrence of the random and non-monotone time transformation L in the compensator just given.

While writing on Markov processes we cannot resist drawing attention to an open problem concerning grouped observations of a *homogenous* Markov process, studied in Gill (1985). Consider a finite state space, homogenous, Markov process, on the time interval $[0, 1]$. Let the column vector J contain the indicators for the state of the process at time 1 and let L denote the column vector containing the total lengths of time spent in each state during $[0, 1]$. Is $\mathrm{E}(LJ^\top)$ positive semidefinite, whatever the initial distribution over the states and the intensities of transitions between the states?

11. Empirical processes revisited.

Here we look again at empirical process methods for analysing the Kaplan-Meier estimator, with particular reference to bootstrapping. There is some connection between the new approach given here and methods used by Pollard (1990) to study the Nelson-Aalen estimator. First we recall some of the modern terminology of empirical process theory; see van der Vaart and Wellner (1993) for the complete story.

Let $X_1, \ldots, X_n$ denote i.i.d. observations from a probability measure $\mathbb{P}$ on a space $\mathcal{X}$, and let $\mathbb{P}_n$ denote the empirical probability measure based on the X_i's. Let $\mathcal{F}$ denote a class of measurable functions from $\mathcal{X}$ to $\mathbb{R}$. Write $\mathbb{P}f$ and $\mathbb{P}_n f$ for true mean and empirical mean respectively of a function $f \in \mathcal{F}$, both supposed finite and even bounded. Define the empirical process

$$Z_n = (n^{\frac{1}{2}}(\mathbb{P}_n - \mathbb{P})f : f \in \mathcal{F}),$$

this is to be considered as a (possibly nonmeasurable) random element of the space $\ell^\infty(\mathcal{F})$, the class of bounded functions on $\mathcal{F}$ endowed with the supremum norm. An *envelope* for $\mathcal{F}$ is a function F such that $|f| \leq F$ for all $f \in \mathcal{F}$.

The class $\mathcal{F}$ is called a *Glivenko-Cantelli class* if $(\mathbb{P}_n f : f \in \mathcal{F})$ converges in supremum norm, almost uniformly, to $(\mathbb{P}f : f \in \mathcal{F})$. It is called a *Donsker class* if Z_n converges weakly to a tight Gaussian limit in $\ell^\infty(\mathcal{F})$. Many theorems giving useful conditions for a class to be Donsker or Glivenko-Cantelli are known. In particular we mention that if $\mathcal{X}$ is the real line, then the class of uniformly bounded monotone functions is both Glivenko-Cantelli and Donsker. This extends obviously to functions of uniformly bounded variation by writing them as differences of monotone functions.

One can show quite easily that a class of monotone functions, not necessarily uniformly bounded but having an integrable envelope, is also Glivenko-Cantelli. It is not clear whether monotone functions with *square integrable envelope* are Donsker (one approach might be to apply van der Vaart and Wellner, 1993, Lemma 2.42, to monotone functions bounded by M and then let $M \to \infty$). However van der Vaart (1993) at least shows that a $2 + \varepsilon$ finite moment of the envelope is sufficient.

Bootstrapping means estimating the distribution of Z_n by the conditional distribution given $\mathbb{P}_n$ of the bootstrap process

$$Z_n^* = (n^{\frac{1}{2}}(\mathbb{P}_n^* - \mathbb{P}_n)f : f \in \mathcal{F}),$$

where $\mathbb{P}_n^*$ is the empirical distribution based on a random sample of size n from $\mathbb{P}_n$. In principle this is a known or computable distribution: there are n^n possible samples of equal probabilily n^{-n} which just have to be enumerated. In practice one uses the Monte-Carlo method: actually take N samples of size n from $\mathbb{P}_n$, and use their empirical distribution.

A celebrated theorem of Giné and Zinn says that the bootstrap works (the conditional distribution of Z_n^* approaches that of Z_n) if and only if the Donsker theorem holds: in fact, if $\mathcal{F}$ has a square integrable envelope then almost surely, Z_n^* converges in distribution to the same limit as Z_n; without the integrability condition, the result holds in (outer) probability. This latter result is formulated properly in terms of a suitable metric metrizing convergence in distribution. It has all the desired (and expected) consequences, e.g., convergence in probability of quantiles of the distribution of real functionals of Z_n^*, in particular its own supremum norm.

These results mesh nicely with the notion of compact differentiability, since $\ell^\infty(\mathcal{F})$ is a normed vector space. Write for brevity $\mathbb{P}_n(\mathcal{F})$ for $(\mathbb{P}_n(f) : f \in \mathcal{F})$. If ϕ is a compactly differentiable functional mapping $\ell^\infty(\mathcal{F})$ to another normed vector space, then the *delta method* holds:

$$n^{\frac{1}{2}}(\phi(\mathbb{P}_n(\mathcal{F})) - \phi(\mathbb{P}(\mathcal{F}))) \xrightarrow{\mathcal{D}} \mathrm{d}\phi(\mathbb{P}(\mathcal{F})) \cdot Z.$$

Also bootstrap results carry over to differentiable functionals: if $\mathcal{F}$ is Donsker and ϕ is compactly differentiable at $\mathbb{P}(\mathcal{F})$ then the bootstrap works in probability for

$$n^{\frac{1}{2}}(\phi(\mathbb{P}_n^*(\mathcal{F})) - \phi(\mathbb{P}_n(\mathcal{F})));$$

if moreover $\mathcal{F}$ has a square integrable envelope and ϕ is *continuously* compactly differentiable at $\mathbb{P}(\mathcal{F})$ then the bootstrap works almost surely. For the very short and elegant proofs of these statements see van der Vaart and Wellner (1993, Theorems 3.24 and 3.25).

As second preparatory excursion we should mention some special aspects of bootstrapping the Kaplan-Meier estimator. In fact there is another sensible way to bootstrap censored survival data: rather than resampling from the observations $X_i = (\widetilde{T}_i, \Delta_i)$ it would seem more reasonable to resample from an estimate of the model supposed to generate them: thus, estimate F and G by Kaplan-Meier estimators $\widehat{F}$ and $\widehat{G}$, sample survival times T_i^* and censoring times C_i^* independently from $\widehat{F}$ and $\widehat{G}$, then form pairwise minima and indicators, and finally calculate a bootstrapped Kaplan-Meier estimator $\widehat{F}^*$ from them. It turns out (Efron, 1981) that this procedure is (probabilistically at least) identical to straight resampling from the X_i. The reason for this is the fact that the random censorship model in a strong sense is not a model at all: to *every* distribution of $X = (\widetilde{T}, \Delta)$ one can associate essentially one random censorship model which generates it, namely that with survival and censoring hazard measures given respectively by

$$\begin{aligned} \Lambda_F(\mathrm{d}t) &= \frac{\mathrm{P}(\widetilde{T} \in \mathrm{d}t, \Delta = 1)}{\mathrm{P}(\widetilde{T} \geq t)}, \\ \Lambda_G(\mathrm{d}t) &= \frac{\mathrm{P}(\widetilde{T} \in \mathrm{d}t, \Delta = 0)}{\mathrm{P}(\widetilde{T} > t \text{ or } \widetilde{T} = t, \Delta = 0)}. \end{aligned} \qquad (1)$$

Note the asymmetry here, corresponding to the asymmetry in the definition of Δ. The point is that if $\widetilde{T} \in \mathrm{d}t$ and $\Delta = 1$, we cannot know whether or not $C \in \mathrm{d}t$. The asymmetry means that $\widehat{G}$, the Kaplan-Meier estimator of G, is not defined simply by replacing Δ by $1 - \Delta$ in the definition. The correct definition can be inferred from (1). A useful consequence of the identity is the fact $(1 - \widehat{F})(1 - \widehat{G}) = 1 - \widetilde{F}_n$, corresponding to $(1 - F)(1 - G) = (1 - \widetilde{F})$.

These facts mean that any method used to study the Kaplan-Meier estimator under regular sampling can be used to study it under bootstrapping. For instance, the fact that $N - \int Y \mathrm{d}\Lambda$ is a martingale implies that $N^* - \int Y^* \mathrm{d}\widehat{\Lambda}$ is a $\mathbb{P}_n$-martingale (a direct proof is also easy), and all martingale proofs of weak convergence of $n^{\frac{1}{2}}(\widehat{F} - F)$ can be copied to find a proof of (conditional) weak convergence of $n^{\frac{1}{2}}(\widehat{F}^* - \widehat{F})$; the only complication is that F and G are no longer fixed but vary with n (as $\widehat{F}$ and $\widehat{G}$). See Akritas (1986) for the first proof that the bootstrap works for Kaplan-Meier along these lines.

As a final remark we point out that it is often wise to bootstrap studentized statistics; e.g., estimate the distribution of $n^{\frac{1}{2}}(\widehat{F} - F)/((1 - \widehat{F})\sqrt{\widehat{C}})$ with that of $n^{\frac{1}{2}}(\widehat{F}^* - \widehat{F})/((1 - \widehat{F}^*)\sqrt{\widehat{C}^*})$. It is not yet known if this does for Kaplan-Meier what it usually does, i.e., give second order rather than just first order correctness, especially if we are interested in distributions of nonlinear functionals of this such as a supremum norm. One should also realise (van Zwet, 1993) that to enjoy the extra accuracy one will have to take a number of bootstrap samples N which is a good deal larger than is customary.

After all these preparations some first results can at least be got very fast. The continuous differentiability of product-integration and the other maps involved, together with the classical Donsker theorem for $F_n^1, \widetilde{F}_n$, shows that the bootstrap works almost surely for the Kaplan-Meier estimator on any interval $[0, \sigma]$ such that $y(\sigma) > 0$.

We now show the great power of modern empirical process methods by looking at *van der Laan's identity*, a general identity for certain semiparametric estimation problems which we will study from that point of view in section 13.

The results of sections 4 and 6 show that $\widehat{F} - F$ can be approximated by

$$n^{-1}(1 - F) \int \frac{\mathrm{d}N - Y\mathrm{d}\Lambda}{(1 - F)(1 - G_-)}; \tag{2}$$

in fact the difference is uniformly $o_{\mathrm{P}}(n^{-\frac{1}{2}})$ on intervals $[0, \sigma]$ where $y(\sigma) > 0$. In fact there is a related identity which is so powerful that consistency, asymptotic normality, asymptotic efficiency, and correctness of the bootstrap, all follow from it in a few lines by appeal to the general theory sketched above. The identity is surprising and new; it is easy to obtain, and like all good things connected to Kaplan-Meier is really just another version of the Duhamel equation. In section 13 we show how the identity follows from the fact that (2) is the so-called *efficient influence curve* for estimating F, and $\widehat{F}$ is the nonparametric maximum likelihood estimator of F (keeping G fixed). From this point of view it is a special case of van der Laan's (1993a) identity for linear-convex models:

$$\widehat{F}(t) - F(t) = (\mathbb{P}_n - \mathbb{P})\mathrm{IC}_{\mathrm{eff}}(\cdot, \widehat{F}, t). \tag{3}$$

Here at last is the new identity:

$$\begin{aligned}\widehat{F}(t)-F(t) &= n^{-1}(1-\widehat{F}(t))\int_0^t \frac{\mathrm{d}N - Y\mathrm{d}\widehat{\Lambda}}{(1-\widehat{F})(1-G_-)}\\ &\quad - n^{-1}(1-\widehat{F}(t))\int_0^t \frac{\mathrm{d}(\mathrm{E}N)-(\mathrm{E}Y)\mathrm{d}\widehat{\Lambda}}{(1-\widehat{F})(1-G_-)}.\end{aligned} \tag{4}$$

Note how it is obtained from (2) by replacing F and Λ throughout by $\widehat{F}$ and $\widehat{\Lambda}$, then subtracting from the result the same functional of the expectation of N and Y. The distribution G remains everywhere fixed.

To prove the identity directly note first some major cancellations. Since $\mathrm{d}\widehat{\Lambda} = \mathrm{d}N/Y$ the first term disappears entirely; since $\mathrm{d}(\mathrm{E}N) = (\mathrm{E}Y)\mathrm{d}\Lambda$ we can simplify the second term, showing that (4) is equivalent to

$$\begin{aligned}\widehat{F}(t)-F(t) &= n^{-1}(1-\widehat{F}(t))\int_0^t \frac{(\mathrm{E}Y)(\mathrm{d}\widehat{\Lambda}-\mathrm{d}\Lambda)}{(1-\widehat{F})(1-G_-)}\\ &= n^{-1}(1-\widehat{F}(t))\int_0^t \frac{(1-F_-)(\mathrm{d}\widehat{\Lambda}-\mathrm{d}\Lambda)}{1-\widehat{F}}\\ &= \int_0^t (1-F_-)(\mathrm{d}\widehat{\Lambda}-\mathrm{d}\Lambda)\prod_{(\cdot)}^{t}(1-\mathrm{d}\widehat{\Lambda})\end{aligned}$$

which is simply a version of the Duhamel equation (2.12). The only condition needed here is that $G(t-)<1$.

So far it seems as if we have only complicated something rather more transparent. However, introduce the following two classes of functions of $(\widetilde{T},\Delta)$, both indexed by the pair (F,t):

$$\begin{aligned}f_{1,(F,t)}(\widetilde{T},\Delta) &= \frac{(1-F(t))1\{\widetilde{T}\le t,\Delta=1\}}{(1-F(\widetilde{T}))(1-G(\widetilde{T}-))},\\ f_{2,(F,t)}(\widetilde{T},\Delta) &= (1-F(t))\int_0^{t\wedge\widetilde{T}} \frac{\mathrm{d}F}{(1-F)(1-F_-)(1-G_-)}.\end{aligned}$$

For the time being G is kept fixed. Now fix σ such that $y(\sigma)>0$ and let $\mathcal{F}$ be the class of all $f_{1,(F,t)}$ together with all $f_{2,(F,t)}$ such that $t\in[0,\sigma]$ while F can be *any* distribution function on $[0,\infty)$ whatsoever.

Because $\int \mathrm{d}F/((1-F)(1-F_-)) = F/(1-F)$ and thanks to the indicator of $\widetilde{T}\le t$ in f_1, all $f\in\mathcal{F}$ are bounded by $1/(1-G(\sigma-))<\infty$. The functions f_2 are monotone as functions of $\widetilde{T}$; the functions f_1 are unimodal (increasing then decreasing). This means that $\mathcal{F}$ is an easy example of a Glivenko-Cantelli and a Donsker class. The reason this is useful is because we can rewrite our identity (4) as

$$\widehat{F}(t)-F(t) = (\mathbb{P}_n-\mathbb{P})(f_{1,(\widehat{F},t)} - f_{2,(\widehat{F},t)}).$$

Since the Glivenko-Cantelli theorem tells us $(\mathbb{P}_n - \mathbb{P})(\mathcal{F})$ goes, almost surely, uniformly to zero, we extract from the identity uniform consistency of $\widehat{F}$ on $[0, \sigma]$. Next, the Donsker theorem for $n^{\frac{1}{2}}(\mathbb{P}_n - \mathbb{P})(\mathcal{F})$ together with continuity of the limiting process in F allows us to conclude weak convergence of $(n^{\frac{1}{2}}(\widehat{F}(t) - F(t)) : t \in [0, \sigma])$ without further work.

A bootstrap conclusion is a little more tricky since in the identity G was fixed but in bootstrapping it must also be allowed to vary. To take care of this and also to further extend the results, multiply each of the functions in $\mathcal{F}$ by $(1 - G(t))$, and let not only t and F but also G now vary completely freely. In fact t is not restricted to any special interval $[0, \sigma]$ any more either. We now have that the functions in $\mathcal{F}$ are uniformly bounded by 1, and of course they retain their monotonicity properties. So the new, larger, $\mathcal{F}$ is still Glivenko-Cantelli and Donsker. But since in an obvious notation

$$(1 - G(t))(\widehat{F}(t) - F(t)) = (\mathbb{P}_n - \mathbb{P})(f_{1,(\widehat{F},G,t)} - f_{2,(\widehat{F},G,t)})$$

we can extract directly:

$$\begin{aligned} &\|(1 - G)(\widehat{F} - F)\|_\infty \to 0 \quad \text{almost surely,} \\ &\sqrt{n}(1 - G)(\widehat{F} - F) \xrightarrow{\mathcal{D}} (1 - G)Z \quad \text{on } \mathcal{T}, \|\cdot\|_\infty \end{aligned}$$

where $\mathcal{T} = \{t : G(t-) < 1\}$.

Finally for a bootstrap result, we appeal to the Giné-Zinn theorem, noting that

$$(1 - \widehat{G}(t))(\widehat{F}^*(t) - \widehat{F}(t)) = (\mathbb{P}^*_n - \mathbb{P}_n)(f_{1,(\widehat{F}^*,\widehat{G},t)} - f_{2,(\widehat{F}^*,\widehat{G},t)}).$$

Consequently

$$\sqrt{n}(1 - \widehat{G})(\widehat{F}^* - \widehat{F}) \xrightarrow{\mathcal{D}} (1 - G)Z \quad \text{on } \mathcal{T}, \|\cdot\|_\infty, \text{ almost surely.}$$

This is still not quite the required result (which should concern $\sqrt{n}(1-\widehat{G}^*)(\widehat{F}^* - \widehat{F})$) but good enough for practical application, and very directly obtained. To replace $(1 - \widehat{G})$ by $(1 - \widehat{G}^*)$ it is necessary to do a little more work: it is known that $(1 - \widehat{G})/(1 - G)$ is uniformly bounded in probability (Gill, 1983, by use of Doob's inequality) and the process $(1 - G)Z$ is tied down at its upper endpoint so there is not too much difficulty in making the required replacement.

These results, breathtakingly fast to obtain, allow many improvements and modifications. For instance instead of multiplying by $(1 - G)$ one could try $(1 - G)/y^{\frac{1}{2}-\varepsilon}$, for $\varepsilon > 0$; this leads to a class $\mathcal{F}$ whose envelope is unbounded but does have a finite $2 + \varepsilon/4$ moment. It seems unlikely however one can do quite as well as the optimal results in the martingale approach, since there a special relation between f_1 and f_2 was used which here, since F varies freely, is not available.

As we will see in section 13 it would have been in principle possible to derive these results about Kaplan-Meier *without having an explicit representation of the estimator* and *without an explicit form of the identity.* All that counts is the fact that it is a non-parametric maximum likelihood estimator in a model having certain general structural properties.

For further bootstrapping ideas see Doss and Gill (1992) and ABGK section IV.1.5.

Before leaving Kaplan-Meier, at least in a traditional context, we would like to make some conjectures concerning estimation of $F(\tau-)$ in the case $G(\tau-)=1$, $F(\tau-)<1$. Suppose both F and G have strictly positive densities 'just before τ', think for example of the typical case $F=\text{exponential}(\lambda)$, $G=\text{uniform}(0,\tau)$. We saw that in this case that $\widehat{F}(\tau)$ is consistent; however, from the result on weak convergence we see that the asymptotic variance of $n^{\frac{1}{2}}(\widehat{F}(\tau)-F(\tau))$ would be infinite if the usual formula $(1-F)^2\cdot C$ would be applicable. In fact the very easy calculation of formula (3.6) of Theorem 3.1 of van der Vaart (1991b) shows that root n, regular estimation of $F(\tau)$ is impossible. The question then arises: what rate is achievable? Does Kaplan-Meier achieve the best rate?

If G had been known one could have estimated $F(\tau)$ by $\int_0^\tau \mathrm{d}F_n^1/(1-G)$. In seems unlikely that using the censored observations too would tremendously improve the rate of convergence of this estimator, and also unlikely that knowing G is very crucial. So there is some similarity with the problem of estimation of $\mathrm{E}(X^{-1})$ based on i.i.d. observations of a positive X, and a little calculation shows that our case corresponds to that in which $\mathrm{E}(X^{-2+\varepsilon})<\infty$ for each $\varepsilon>0$, but $\mathrm{E}(X^{-2})=\infty$.

This problem has been studied (among many others) by Levit (1990). He shows that the truncated estimator $n^{-1}\sum X_i^{-1}1\{X_i>1/\sqrt{n}\}$ achieves the rate $\sqrt{(}n/\log n)$ and that this rate is optimal in a minimax sense. One can also obtain this result by using the van Trees inequality (Gill and Levit, 1992) and introducing the 'hardest parametric submodel' which is the exponential family with sufficient statistic $X1\{X>1/\sqrt{n}\}$.

This leads to the following conjecture:

Conjecture. *$F(\tau-)$ can be estimated at rate $\sqrt{(}n/\log n)$, the Kaplan-Meier estimator $\widehat{F}(\tau-)$ does not achieve this rate but the modification $\widehat{F}(\tau-1/\sqrt{n})$ does. Instead of the the non-random time $\tau-1/\sqrt{n}$ one can also use the random time $T_n=\sup\{t:Y(t)\geq\sqrt{n}\}$ here.*

12. Dabrowska's multivariate product-limit estimator.

One can very naturally generalise the random censorship model of the previous sections to higher dimensional time. Let $T=(T_1,\ldots,T_k)$ be a vector of positive random variables; let $C=(C_1,\ldots,C_k)$ be a vector of censoring times; and define $\widetilde{T}_i=T_i\wedge C_i$, $\Delta_i=1\{T_i\leq C_i\}$. Question: given n i.i.d. replicates of the vectors $(\widetilde{T},\Delta)$, how should one estimate the distribution of T?

This simple question has turned out surprisingly hard to answer satisfactorily. One might have expected that some obvious generalization of the Kaplan-Meier estimator would do the trick. However it seems that each defining property of that estimator, when used to suggest a multivariate generalization, leads to a *different* proposal; some of which are very hard to study and some of which turn out not to be such very good ideas after all.

From a statistical point of view the property of nonparametric maximum likelihood estimator would seem the most essential. However in the multivariate case the NPMLE is only implicitly defined; in fact, it is severely non-unique and many choices are not even consistent. Sophisticated modification of the NPMLE idea is needed to make it work, and the analysis of the resulting (efficient) estimator is highly nontrivial; see van der Laan (1993c).

Another way to think of the Kaplan-Meier estimator is via the Nelson-Aalen estimator of the hazard measure. There is a natural multivariate generalization of the latter, so if one fixes the relation between hazard and survival, a plug-in estimator is possible. For instance in the one-dimensional case one can consider the survival function S as the solution, for given hazard function Λ, of the Volterra equation $S = 1 - \int S_- \mathrm{d}\Lambda$. This has a multivariate generalization leading to an estimator called the Volterra estimator; it turns out to have rather poor practical performance. Very much better is a more subtle proposal of Prentice and Cai (1992a,b) who point out that there is also a Volterra type equation, in higher dimensions, for the multivariate survival function *divided by the product of its one dimensional margins.* The integrating measure is no longer the multivariate hazard but a slightly more complicated, though still related, measure.

Finally one might expect: isn't there simply a relation, involving some kind of product-integration, between multivariate hazard and multivariate survival? The answer is that there is such a relation, but it does not involve multivariate generalisations of hazard measures but rather something new called *iterated odds ratio measures*. This relation lies at the basis of Dabrowska's (1988) generalised product-limit estimator and will be the subject of this section.

In this section we will concentrate on two issues concerning the Dabrowska estimator: firstly, the required extension of product-integration theory to higher dimensional time; and secondly, the derivation of the product-integral representation of a multivariate survival function in terms of iterated odds ratio measures (or interaction measures; Dabrowska, 1993). We will not discuss the estimation of these measures by their natural empirical counterparts, nor the statistical properties of the Dabrowska estimator which is obtained by plugging the empirical measures into the representation. The differentiability approach we took in section 6 works here very well and gives all the expected results: consistency, asymptotic normality, correctness of the bootstrap. Gill, van der Laan and Wellner (1993) give full details in the two dimensional case, together with a study of the Prentice-Cai estimator. Gill (1992) shows that the basic tools developed there suffice also for studying the general case.

Many further results and connections can be found in Dabrowska (1993).

Here is the general idea. Let T denote a k-dimensional survival time as above, and define its survival function S by $S(t) = \mathrm{P}(T \gg t)$ where the symbol $\gg$ denotes coordinatewise strict inequality $>$. In the one-dimensional case we formed an interval function from S by taking ratios. In higher dimensions we form a 'hyperrectangle function' by taking generalised or iterated ratios, just as an ordinary measure is formed by taking generalised differences. Let s, t be k-vectors; let $E = \{1, \ldots, k\}$ and for $A \subseteq E$ let t_A denote the lower-dimensional vector $(t_i : i \in A)$. Now we define, for $s \leq t$

(coordinatewise $\leq$), the rectangle-function

$$S(s,t) = \prod_{A\subseteq E} S((s_A, t_{E\setminus A}))^{(-1)^{|A|}}.$$

This is obtained by taking S at the top corner of the rectangle $(s,t] = \{u : s \ll u \leq t\}$, then dividing by the values of S at all corners one step down from the top, then multiplying by the values one further step down, and so on.

It is easy to check that S is multiplicative over partitions of a rectangle by subrectangles. Defining informally $L(\mathrm{d}t) = S(\mathrm{d}t) - 1$ then we should have

$$\begin{aligned} S(0,t) &= \mathop{\Large\pi}_{(0,t]} (1 + L(\mathrm{d}t)),\\ L(0,t) &= \int (S(\mathrm{d}t) - 1). \end{aligned}$$

Now $S(0,t) = \prod_{A\subseteq E} S_{E\setminus A}(t_{E\setminus A})^{-1^{|A|}}$, where S_A denotes the survival function of T_A. So a further step is required to recover the original survival function; in fact we have $S(t) = \prod_{A\subseteq E} S_A(t_A)$. The final result is therefore

$$S(t) = \prod_{A\subseteq E} \mathop{\Large\pi}_{(0_A,t_A]} (1 + L_A(\mathrm{d}s_A)).$$

We need estimators for L_A and theory for the analytical properties of the functionals which are now involved. The idea is to estimate $L(\mathrm{d}t)$ (and similarly $L_A(\mathrm{d}t_A)$) using the same idea which lies at the base of the Nelson-Aalen estimator: look just at those observations for which $\widetilde{T} \geq t$. For each coordinate $i \in E$ we can decide whether or not the underlying T_i lies in (t_i, ∞_i). Write

$$\begin{aligned} 1 + L(\mathrm{d}t) = S(\mathrm{d}t) &= \prod_{A\subseteq E} \mathrm{P}(T_A \geq t_A, T_{E\setminus A} \gg t_{E\setminus A})^{(-1)^{|A|}}\\ &= \prod_{A\subseteq E} \mathrm{P}(T_{E\setminus A} \gg t_{E\setminus A} \,|\, T \geq t)^{(-1)^{|A|}}. \end{aligned}$$

Restricting attention to the observations with $\widetilde{T} \geq t$ we can simply replace the conditional probabilities with numbers of observations known to satisfy $T_{E\setminus A} \gg t_{E\setminus A}$.

Multivariate product-integration.

The general theory of product-integration in section 2 goes through, *completely unchanged*, when we study ‘rectangle functions’ in $[0,\infty)^k$, provided these functions take scalar values so that the order of multiplication is not relevant. We consider only rectangular partitions (i.e., the Cartesian product of ordinary partitions of each coordinate axis). Whenever an order does have to be fixed—the key identities (2.1)–(2.5) need an order to be specified—we take the video-scanning or lexicographic ordering.

Proposition 1 and Theorem 1 give no problems. Theorem 2 is the first place where care is needed: there we used the fact that

$$|\mathcal{T}| \to 0 \Rightarrow \max_{\mathcal{T}} \alpha_0^- \to 0$$

where α_0^- is the measure α_0 less its largest atom, $\mathcal{T}$ denotes a partition of a fixed rectangle $(0,\tau]$, and $|\mathcal{T}|$ denotes its mesh: the largest length of an edge of a subrectangle in the partition. This is true in k dimensions too, as the following argument shows.

Suppose it were not true. Then one could find rectangles $A_n = (s_n, t_n]$ with diameter (maximum edge-length) converging to zero, with $\limsup \alpha_0^-(A_n) > 0$. Along a subsequence we can, by compactness, assume $s_n \to t$, $t_n \to t$. Now if $A \subseteq B$, $\alpha_0^-(A) \leq \alpha_0^-(B)$. So with $B(t,\delta)$ standing for the sphere, centre t, radius δ, we have $\alpha_0^-(B(t,\delta)) > c > 0$ for every $\delta > 0$. If t itself is an atom then for small enough δ it is the largest one in $B(t,\delta)$. If not we can remove the point t anyway and conclude $\alpha_0(B(t,\delta) \setminus \{t\}) > c > 0$ for all $\delta > 0$, which is impossible.

Finally, and essential for the later continuity and differentiability results, we need to establish versions of the equations (9) to (13) of section 2, including the Duhamel equation and the Peano series.

These equations were proved by passing to the limit in the discrete equations (1) to (5); and in those equations the order in which terms are taken does make a difference. However the discrete products which are involved can all be interpreted as approximations to product-integrals over various subregions of the rectangle $(s,t]$, and therefore the proof sketched in section 2 goes through, with the proper modifications of the limiting equations. To do this let $\prec$ denote lexicographic ordering in $[0,\infty)^k$. The Duhamel equation for instance becomes:

$$\prod_{(s,t]}(1+\mathrm{d}\alpha) - \prod_{(s,t]}(1+\mathrm{d}\beta)$$
$$= \int_{u\in(s,t]} \prod_{\{v\in(s,t]:v\prec u\}} (1+\mathrm{d}\alpha)\ (\alpha(\mathrm{d}u)-\beta(\mathrm{d}u)) \prod_{\{v\in(s,t]:v\succ u\}} (1+\mathrm{d}\beta).$$

The regions $\{v \in (s,t] : v \prec u\}$ and $\{v \in (s,t] : v \succ u\}$ are not rectangles, but are easily seen to be disjoint unions of at most k rectangles, so the product-integral can be defined for them by taking finite products over rectangles.

Further details are given in Gill, van der Laan and Wellner (1993).

Dabrowska's representation.

This material is taken from ABGK section X.3.1.

It is easy to check that the iterated odds ratios $S(s,t)$ defined at the beginning of the section are 'equal to one on the diagonal' and are 'right continuous'. In order to apply Theorem 1 of section 2 in order to conclude the existence of an additive measure L such that $S = \prod(1+\mathrm{d}L)$, $L = \int(\mathrm{d}S - 1)$ we must check the domination property for S. Before that however, we give some further discussion of the interpretation of the L-measures.

In fact $S(s,t)$ has an interpretation in terms of the $2\times 2\times\cdots\times 2$ contingency table for the events $s_i < T_i \le t_i$ versus $T_i > t_i$, with respect to the conditional distribution of T given $T \gg s$. Consider first the two-dimensional case: we have by definition

$$S\big((s_1,s_2),(t_1,t_2)\big) = \frac{S(t_1,t_2)S(s_1,s_2)}{S(s_1,t_2)S(t_1,s_2)}$$

since there are four subsets A to consider, two of them ($\emptyset$ and E) having an even number of elements, and two ($\{1\}$ and $\{2\}$) having an odd number. We can now rewrite $S(s,t)$ as

$$\begin{aligned} S\big((s_1,s_2),(t_1,t_2)\big) &= \frac{S(t_1,t_2)/S(s_1,t_2)}{S(t_1,s_2)/S(s_1,s_2)} \\ &= \frac{\mathrm{P}(T_1>t_1|T_2>t_2)/\mathrm{P}(T_1>s_1|T_2>t_2)}{\mathrm{P}(T_1>t_1|T_2>s_2)/\mathrm{P}(T_1>s_1|T_2>s_2)}. \end{aligned}$$

So $S(s,t)$ is the ratio of the conditional odds for $T_1 > t_1$ against $T_1 > s_1$, under the conditions $T_2 > t_2$ and $T_2 > s_2$ respectively. If T_1 and T_2 are independent, this *odds ratio* will equal 1. 'Positive dependence' between T_1 and T_2 will express itself in an odds ratio larger than one, since 'increasing T_2 leads to a higher odds on T_1 being large'. Negative dependence corresponds to an odds ratio smaller than 1. In fact we will see in a moment that if the odds ratio equals 1 for all $s \le t$, then T_1 and T_2 are independent. So in two dimensions $S-1$ is a measure of dependence indexed by all rectangles $(s,t]$.

In one dimension the odds 'ratio' is just the odds itself $\mathrm{P}(T_1 > t_1|T_1 > s_1)$. In higher dimensions, the k-dimensional iterated odds ratio is the ratio of two $k-1$ dimensional iterated odds ratios: i.e., the ratio of the iterated odds ratios for $(T_1,\ldots,T_{k-1})$ and the rectangle $((s_1,\ldots,s_{k-1}),(t_1,\ldots,t_{k-1})]$, conditional on $T_k > t_k$ and conditional on $T_k > s_k$. Now it measures multidimensional dependence or interaction: if the dependence between $T_1,\ldots,T_{k-1}$ increases as T_k increases one has a positive interaction (increasing interdependence) between $T_1,\ldots,T_k$ and the iterated odds ratio is larger than one.

The result of Theorem 1 (when we have verified the domination property) is that a dominated *additive* interval function (therefore, an ordinary signed measure) exists, let us call it L, such that

$$S(s,t) = \prod_{(s,t]}(1+\mathrm{d}L)$$

where the product-integral is understood as the limit of approximating finite products over rectangular partitions of the hyper-rectangle $(s,t]$ into small sub-hyper-rectangles. We call L the *iterated odds ratio measure* or *cumulant measure* and consider it as a measure of k-dimensional interaction (a measure of dependence when $k=2$ and just a description of the marginal distribution when $k=1$). Note that there is an L-measure, denoted L_C, for each subset of components T_C of T. Since the odds ratio for a small rectangle $(t,t+\mathrm{d}t]$ is $S(t,t+\mathrm{d}t) = 1+L(\mathrm{d}t)$ and a ratio of 1 corresponds to independence, we may interpret an L of zero as corresponding to zero-interaction or independence; a positive L corresponds to positive interaction or dependence, and similarly for a negative L. Of course things may be more complicated: L may take different signs in different regions of space.

By Theorem 1 of section 2 we may calculate L as $L(s,t) = \int_{(s,t]} L(\mathrm{d}u) = \int_{(s,t]} \mathrm{d}(S - 1)$; in other words, the L-measure of a rectangle $(s,t]$ is approximated by just adding together the deviations from independence (or interactions) $S(u, u + \mathrm{d}u) - 1$ of small rectangles $(u, u + \mathrm{d}u]$ forming a partition of $(s,t]$.

It remains to verify the domination property of the iterated odds ratios $S(s,t)$.

Let us first look at the two-dimensional case which will give the required insight for the general case. For $s \le t \le \tau$, $s \ne t$, we have

$$\begin{aligned}\bigl|S(s,t) - 1\bigr| &= \left|\frac{S(t_1,t_2)S(s_1,s_2)}{S(s_1,t_2)S(t_1,s_2)} - 1\right| \\ &\le S^*(\tau)^{-2}\bigl|S(t_1,t_2)S(s_1,s_2) - S(s_1,t_2)S(t_1,s_2)\bigr|\end{aligned}$$

where we write $S^*(\tau)$ as shorthand for $\mathrm{P}(T \gg \tau \text{ or } T = \tau)$; this may be different from $\mathrm{P}(T \ge \tau)$; taking account of the difference allows us to get a slightly stronger result. Now let a, b, c, d be the probabilities in the 2×2 table:

	$T_1 \in (s_1, t_1]$	$T_1 > t_1$
$T_2 \in (s_2, t_2]$	a	b
$T_2 > t_2$	c	d

Then the last inequality can be rewritten as

$$\begin{aligned}\bigl|S(s,t) - 1\bigr| &\le S^*(\tau)^{-2}\bigl|d(a+b+c+d) - (c+d)(b+d)\bigr| \\ &= S^*(\tau)^{-2}|ad - bc| \\ &\le S^*(\tau)^{-2}(a + bc) \\ &= S^*(\tau)^{-2}\Bigl(\mathrm{P}\bigl(T \in (s,t]\bigr) + \mathrm{P}\bigl(T_1 \in (s_1,t_1]\bigr)\mathrm{P}\bigl(T_2 \in (s_2,t_2]\bigr)\Bigr).\end{aligned}$$

The right hand side, a constant times the joint probability measure of T_1 and T_2 plus the product of their marginals, is a finite measure on $(0,\tau]$, hence the domination property holds.

Exactly the same kind of bound holds in general by taking account of the same magic cancellation of unwanted terms. Let τ be fixed and satisfy $S^*(\tau) = \mathrm{P}(T \gg \tau \text{ or } T = \tau) > 0$. We can write for $s \le t \le \tau$, $s \ne t$,

$$S(s,t) - 1 = \frac{\prod_{\text{even } C} S(s_C, t_{E\setminus C}) - \prod_{\text{odd } C} S(s_C, t_{E\setminus C})}{\prod_{\text{odd } C} S(s_C, t_{E\setminus C})}$$

where $\emptyset \subseteq C \subseteq E$. Now by the inclusion-exclusion principle

$$\begin{aligned}&\mathrm{P}(T \gg s, T_i > t_i \text{ for all } i \in E \setminus C) \\ &\quad = \mathrm{P}(T \gg s) - \mathrm{P}(T \gg s, T_i \le t_i \text{ for some } i \in E \setminus C) \\ &\quad = \mathrm{P}(T \gg s) - \sum_{i \in E\setminus C} \mathrm{P}(T \gg s, T_i \le t_i) + \sum_{i \ne j \in E\setminus C} \mathrm{P}(T \gg s, T_i \le t_i, T_j \le t_j) - \cdots\end{aligned}$$

or in other words

$$\begin{aligned} S(s_C, t_{E\setminus C}) &= S(s) + \sum_{\emptyset\subset B\subseteq E\setminus C} (-1)^{|B|}\mathrm{P}\big(T \in (s_B, t_B] \times (s_{E\setminus B}, \infty_{E\setminus B})\big) \\ &= \sum_{\emptyset\subseteq B\subseteq E\setminus C} (-1)^{|B|}\mathrm{P}\big(T \in (s_B, t_B] \times (s_{E\setminus B}, \infty_{E\setminus B})\big) \end{aligned}$$

Interchanging the roles of C and $E \setminus C$ in the numerator, and neglecting a possible sign change (if $|E|$ is odd) we get

$$S(s,t) - 1 = \pm \frac{\prod_{\text{even } C} \sum_{\emptyset\subseteq B\subseteq C} \phi_B - \prod_{\text{odd } C} \sum_{\emptyset\subseteq B\subseteq C} \phi_B}{\prod_{\text{odd } C} S(s_C, t_{E\setminus C})} \tag{1}$$

where

$$\phi_B = (-1)^{|B|}\mathrm{P}\big(T \in (s_B, t_B] \times (s_{E\setminus B}, \infty_{E\setminus B})\big).$$

Now when we expand the numerator of (1) an amazing cancellation occurs: products of ϕ_B where the sets B do not cover E cancel out, leaving just products of sets which do cover E. Before proving this, we illustrate it when $E = \{1, 2\}$:

$$\begin{aligned} &\prod_{\text{even } C} \sum_{B\subseteq C} \phi_B - \prod_{\text{odd } C} \sum_{B\subseteq C} \phi_B \qquad (2) \\ &\quad = (\phi_{12} + \phi_1 + \phi_2 + \phi_\emptyset)\phi_\emptyset - (\phi_1 + \phi_\emptyset)(\phi_2 + \phi_\emptyset) = (\phi_{12}\phi_\emptyset - \phi_1\phi_2) \end{aligned}$$

where $\{1, 2\} \cup \emptyset = E$, $\{1\} \cup \{2\} = E$.

In general, consider one element $i \in E$ and split the sums and products in (2) according to whether or not i is included in B, C: we get

$$\begin{aligned} &\prod_{\text{even } C, i\notin C} \sum_{B\subseteq C} \phi_B \cdot \prod_{\text{odd } C, i\notin C} \Big(\sum_{B\subseteq C} \phi_B + \sum_{B\subseteq C} \phi_{B\cup\{i\}}\Big) \\ &- \prod_{\text{odd } C, i\notin C} \sum_{B\subseteq C} \phi_B \cdot \prod_{\text{even } C, i\notin C} \Big(\sum_{B\subseteq C} \phi_B + \sum_{B\subseteq C} \phi_{B\cup\{i\}}\Big). \end{aligned}$$

The terms which nowhere include i are then

$$\prod_{\text{even } C, i\notin C} \sum_{B\subseteq C} \phi_B \cdot \prod_{\text{odd } C, i\notin C} \sum_{B\subseteq C} \phi_B - \prod_{\text{odd } C, i\notin C} \sum_{B\subseteq C} \phi_B \cdot \prod_{\text{even } C, i\notin C} \sum_{B\subseteq C} \phi_B = 0;$$

thus each term in the expansion of (2)—a sum of products of ϕ_B—includes a B containing i.

The result is that (1) can be bounded in absolute value by $S^*(\tau)^{-2^{|E|-1}}$ times a sum of products of ϕ_B, where each term has $\cup B = E$. Consider such a term $\prod \phi_{B_i}$. For each B_i choose $C_i \subseteq B_i$ such that $\cup C_i = E$ and the C_i are disjoint. Now bound $\prod \phi_{B_i}$ by $\prod \mathrm{P}\big(T_{C_i} \in (s_{C_i}, t_{C_i}]\big)$. These are finite measures so we have obtained the required result.

Showing that the multiplicative interval function S is of bounded variation also constitutes a proof of the fact that the *additive* interval function $\log S$ is of bounded variation and hence generates a bounded, signed measure. In fact Dabrowska (1988) originally introduced her representation via additive decompositions of this measure; see also Dabrowska (1993) for further exploration on these lines. Elsewhere in studying the Dabrowska estimator one needs the fact that if a function Y is of bounded variation then so also is $1/Y$; Gill (1992) and Gill, van der Laan and Wellner (1993) just take this fact for granted. However it is not trivially true and in fact needs a supplementary condition on the lower-dimensional sections of F; a proof can be given exactly on the lines above. We summarize these facts as a couple of exercises for the reader, together with a small research project:

Let E be a finite set; let $\mathcal{E}$ be the set of all subsets of E, including E itself and the empty set $\emptyset$, and $A, B, C \in \mathcal{E}$, $\mathcal{A} \subseteq \mathcal{E}$. The number of elements in C is denoted $|C|$. Consider the following two statements:

(i) If one expands $\displaystyle\sum_C (-1)^{|C|} \prod_{B \neq C} \sum_{A \subseteq B} \phi_A$ as $\displaystyle\sum_{\mathcal{A}} c_{\mathcal{A}} \prod_{A \in \mathcal{A}} \phi_A$ then $c_{\mathcal{A}} = 0$ for every $\mathcal{A}$ with $E \setminus \bigcup_{A \in \mathcal{A}} A \neq \emptyset$.

(ii) If one expands $\displaystyle\prod_{A:|A| \text{ is even}} \phi_A - \prod_{A:|A| \text{ is odd}} \phi_A$ as $\displaystyle\sum_{\mathcal{A}} c'_{\mathcal{A}} \prod_{A \in \mathcal{A}} \phi_A$ then $c'_{\mathcal{A}} = 0$ for every $\mathcal{A}$ with $E \setminus \bigcup_{A \in \mathcal{A}} A \neq \emptyset$.

Problems:

a) Prove (i) and (ii).

b) Suppose $F : \mathbb{R}^k \to \mathbb{R}_+$ and all its lower dimensional sections (fix some of the k arguments and let the others vary) are of bounded variation, and F is bounded away from zero. Use (i) to show that $1/F$ is of bounded variation and (ii) to show $\log F$ is of bounded variation.

c***) What about other functions of F (and $G \ldots$)? Is there a combinatorial background to these problems? Is there a non-combinatorial way to prove b)?

13. Laslett's line-segment problem.

This section and the next consider genuinely spatial problems. The problem of the next section doesn't look like a censored data problem at all but we will find that the Kaplan-Meier estimator is the answer all the same. The problem treated here, on the other hand, looks superficially like a case for Kaplan-Meier: however, that is very inappropiate, and we need to develop some new theory for nonparametric maximum likelihood estimators (NPMLEs) in missing data problems. The results are from Wijers (1991), van der Laan (1993a,b).

The following problem was introduced by Laslett (1982a,b). Consider a spatial line-segment process observed through a finite observation window W. Suppose the aim is to estimate the distribution of the lengths of the line-segments. Some line-segments are only partly observed: one or both endpoints are outside the window and the true length is unknown. As an example, Figure 1 shows a map of cracks in granitic rock on the surface in a region of Canada, only partially observable because of vegetation, soil, water, and so on.

Figure 1. Fracture patterns in a part of the Stone, Kamineni and Brown (1984) map of a granitic pluton near Lac du Bonnet, Manitoba, Canada. There are 1567 fractures. Of these, both ends are shown for only 256 fractures in the exposed areas whose lengths can be completely measured. The rest, namely 1311 fractures, are censored.

The data from this example was used to illustrate some methodological aspects of the Kaplan-Meier estimator (Chung, 1989a,b) but that does not seem correct. (In fact, the data was also used to decide on locations for storing nuclear waste). Formally we have censored observations, but is the 'random censorship' model applicable? And anyway, there is surely a *length bias* problem: longer line sements have a bigger chance of getting (partly) into the window and being observed, so the line segments observed are not a random sample from the distribution of interest.

The example of Figure 1 is very complex, in particular because of the shape of the window (another aspect is that the cracks in the rock are really surfaces of which only a section is observed, and so a stereological analysis is really needed). We will concentrate on the case of a convex (e.g., rectangular) window, see Figure 2. Also we will assume that the line-segment process is a homogenous Poisson line-segment process with segment lengths and orientations independent of one another, since this makes the problem concrete and analysable; and the results we obtain will be useful also when these assumptions are not true.

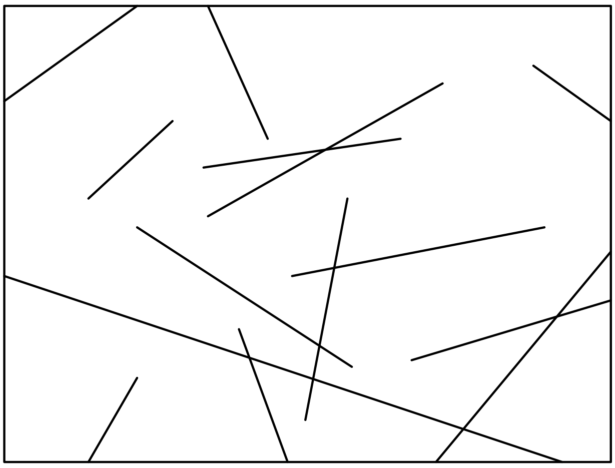

Figure 2.

To be specific, let W be a fixed compact, convex region of $\mathbb{R}^2$ with nonempty interior. Let F be a distribution, on $(0, \infty)$ of line-segment lengths; let K be a distribution on $(-\pi/2, \pi/2]$ of line-segment orientations. Let locations of 'left-hand endpoints' be generated according to a Poisson point process with constant intensity λ on $\mathbb{R}^2$. To each location of a left-hand endpoint attach a line-segment with length drawn from F and an independent orientation drawn from K, independently over locations. The line-segment process is now completely defined. (The choice of left-hand endpoints as preferred point on each line-segment is of course arbitrary; any other convention leads to the same process). The data, on the basis of which F, K and λ are to be estimated, consists of all non-empty intersections of line-segments with W. Each (at least partially) observed line-segment is either completely observed (lies inside the window), or is cut off at one or both ends by the boundary of the window. We call these possibilities: uncensored, singly censored, or doubly censored. In the singly censored case, given a preferred direction, we can further distinguish between singly-left-censored and singly-right-censored. The orientation of observed line-segments is always completely observed. How far a censored line-segment continues outside W is not known.

This problem has two non-trivial aspects. Firstly and more obviously: censoring, not all line segments are completely observed. Secondly and less obvious: length bias, the line segments hitting the window and at least partially observed have (complete) lengths which are not a random sample from F. Longer line-segments have a bigger chance to hit the window. So even if we knew the true lengths of all line-segments hitting the window, we could not estimate F by their empirical distribution.

On the other hand, the problem is not completely intractable; on the contrary, in a certain sense it is quite easy since ad hoc estimators of F are easy to construct which (conditioning on the number n of line-segments observed) are even root n consistent. For instance, consider all line-segments with left-hand endpoint in W. For such a line-segment, its length X is independent of its orientation Θ and (left-hand endpoint) location. Hence the length X is independent of the distance from the left-hand endpoint to the boundary of the window in the direction Θ. What is observed is the minimum of the two and the type (censored or uncensored), and we could use the ordinary Kaplan-Meier estimator based on just these line-segments, discarding all those whose left-hand endpoint is not observed.

Of course we could just as well take right-hand endpoints (or top, or bottom) and better still make some kind of average over these possibilities since it is not pleasant if the estimator depends strongly on an arbitrary choice of direction. Averaging uniformly over all directions avoids this arbitrariness but is rather complicated. However one could also take the average, uniformly over directions, of the N and Y processes *before* going through the steps of inverting Y, integrating with respect to N, and product-integrating the resulting $\widehat{\Lambda}$. This turns out to be the same as computing the ordinary Kaplan-Meier estimator with each line-segment included in the data set as many times as its endpoints are observed: each uncensored line-segment twice, each singly censored one once, and the doubly censored ones not at all.

The estimator just described is easy to calculate but obviously inefficent since it does not make use of the doubly censored line-segments. Its (approximate) variance cannot be calculated by the usual formula for Kaplan-Meier but one could use the bootstrap, resampling the line-segments while treating each duplicated observation as a single observation (as it really is). (**Exercise**: what is the asymptotic variance?)

We will treat the two problems (censoring, length bias) below in two different ways. (To be honest, we will in fact in these notes only solve the one-dimensional problem when the window W is an interval $[0, \tau]$ on the line $\mathbb{R}^1$). We will simply define away the length-bias problem by agreeing to estimate the length distribution of the observed line-segments. Since we will be able to establish a simple 1–1 relationship between F and its length-biased version, we can concentrate on estimating the latter and transform back later. Secondly, we will turn the censoring problem to our advantage by noting that, after reparametrization, we have a special case of a *nonparametric missing data problem.* In such models, the parameter space is convex and the distribution of one observation is linear in the unknown parameter. Now we are in a position to apply general techniques for convex-linear models developed by Wijers (1991), van der Laan (1993a).

First we prove consistency according to an elegant technique developed by Wijers (1991). Then we make use of more deep results from the theory of semiparametric models and empirical processes, and in particular van der Laan's (1993a) remarkable identity for nonparametric maximum likelihood estimators in convex-linear models, to give an alternative consistency proof for the NPMLE as well as asymptotic normality, efficiency, and a bootstrap result; this can be done even though the NPMLE is only implicitly defined and the equations which define it (the so-called self-consistency equations) are too complex to serve as the basis of a direct proof of these facts. In order to explain this approach a brief summary of the theory of asymptotically efficient estimation in semiparametric models will be given. Throughout we will only sketch the main lines of the argument, referring for the necessary computations to Wijers (1991) and van der Laan (1993b).

As we mentioned above all this will only be done in the one-dimensional case. In two dimensions there is an extra complication and the general analysis of this problem is still open, though we believe the theory for the one-dimensional case will be very useful indeed. Further remarks on this will be made later.

Here is the plan of the rest of this section. First we follow Laslett (1982a) in deriving the likelihood for F, K and λ in the general case. We make some remarks on the definition (following Kiefer and Wolfowitz, 1956) and calculation of the NPMLE and point out where the main difficulty (in going from one to two dimensions) lies. The derivation we give is heuristic but effective and serves also to introduce some useful ideas for the one-dimensional case to be studied next.

Specializing to one dimension, we follow Wijers (1991) in showing how a simple reparametrization turns the problem into a nonparametric missing data problem. The description of the problem in these terms allows one to directly write down various useful 'model identities' and to characterise the NPMLE (just of F now, or rather, a new parameter called V) through the self-consistency equation (Efron, 1967; Turnbull, 1976). The same equation used iteratively is an instance of the so-called EM algorithm (Expectation-Maximization: Dempster, Laird and Rubin, 1977) for calculating the NPMLE. We outline Wijer's consistency proof, based on simple convexity based inequalities. The inevitable hard work in actually carrying out the programme is omitted.

Then we discuss the so-called sieved NPMLE, which has the advantage that it can be computed much more quickly, while the consistency proof just given applies just as well to it as to the real NPMLE.

Next we sketch some general theory of semiparametric models and (heuristically at least) derive van der Laan's identity. We show how it can be used as an alternative route to consistency as well as many other 'higher order' properties of the sieved NPMLE. We also connect to the use of the identity in section 11 on the Kaplan-Meier estimator, this being another instance of an NPMLE in a convex-linear model. We also show how van der Vaart's (1991b) theorem tells us that certain functionals of F *cannot* be estimated (regularly) at square root of n rate.

Finally we discuss extension of the results to the general (two-dimensional) case and also what will happen on relaxation of various of our model assumptions.

Laslett's results.

For the time being then, consider the two-dimensional problem as described above, parametrized by λ (Poisson intensity), F (length distribution) and K (orientation distribution). We want to write down 'the probability density of the observed data' as a function of λ, K, and F; the joint NPMLE of these three parameters is obtained by maximizing this likelihood function over the parameter in some sense, to be made explicit later. Fix an origin O, well outside the window. We consider infinite straight lines which cross the window W, parametrized by the distance of the line to the origin r together with the orientation of the line θ. Discretize, considering small intervals $[r, r + \mathrm{d}r)$ and $[\theta, \theta + \mathrm{d}\theta)$ partitioning the ranges of r and θ. We consider now all those line-segments of the process, with left-hand endpoint lying in the strip of width $\mathrm{d}r$ between the (r, θ) and $(r + \mathrm{d}r, \theta)$ lines, and whose own orientation lies in the interval $[\theta, \theta + \mathrm{d}\theta)$, so more or less parallel to the strip. We restrict attention to strips which cross the window. As we run through the small r and θ intervals we pick up in this way, just once, every line-segment hitting the window. Moreover, what happens in different strips is independent, by familar properties of the Poisson process.

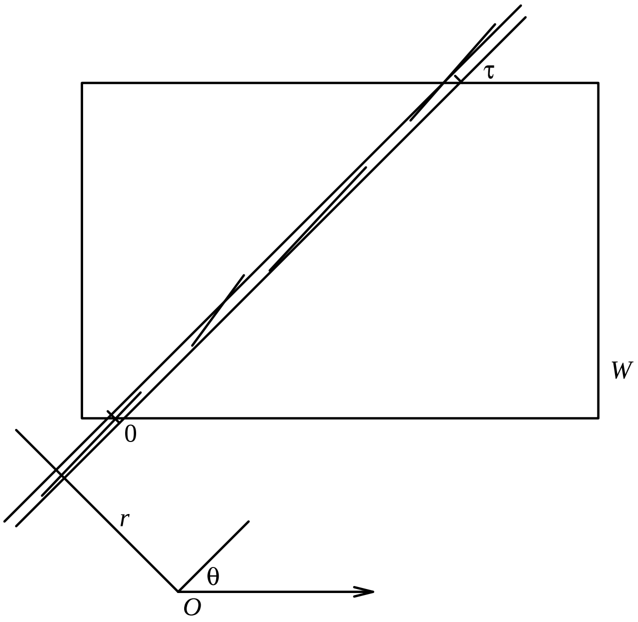

Figure 3.

Fix now one of these strips. Parametrise by the position of the left-hand endpoints of the line-segments, relative to the position where the strip enters the window. Let the length of the intersection of the window and the strip be τ, which depends of course on r and θ. Now we have a one-dimensional process of line-segments, with length distribution F and intensity $\widetilde{\lambda} = \lambda \mathrm{d}r\, K(\mathrm{d}\theta)$, observed through the interval $[0,\tau]$, since there are $\lambda \mathrm{d}r$ line-segments with left-hand endpoint in the strip per unit length, and a fraction $K(\mathrm{d}\theta)$ of them have the required orientation.

A homogenous Poisson line-segment process on the line can be considered as an inhomogenous Poisson point process in the upper half-plane, as follows: to each line-segment $[T, T{+}X]$ associate a point (T,X). The new point process has intensity measure $\widetilde{\lambda}\mathrm{d}tF(\mathrm{d}x)$; $-\infty < t < \infty$, $0 < x < \infty$. We can now calculate further using the facts that disjoint regions contain independent, Poisson distributed numbers of points with means equal to the total intensity of each region; and given the number of points n in a certain region, their locations are distributed like the set of values in an i.i.d. sample of size n from the normalized intensity (restricted to the region and normalized to have total mass one).

In the (t,x) upper half-plane draw the diagonal line with slope -1 through the origin, and draw the vertical lines $t=0$ and $t=\tau$. The two regions formed between these three lines contain all line-segments (points) which hit the window. Those in the left-hand region are 'left-censored' since they have $T<0$ but $T+X \geq 0$; they may or may not be right-censored ($T+X>\tau$). Those in the right-hand region are left-uncensord ($0 \leq T \leq \tau$) and may or may not be right-censored. Line-segments (points) outside these two regions are not observed at all since either $T+X<0$ or $T>\tau$.

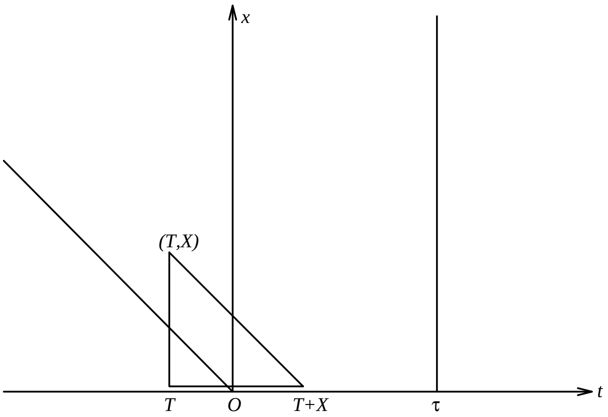

Figure 4.

Since the integrals of $\widetilde{\lambda}\mathrm{d}tF(\mathrm{d}x)$ over the two regions equal $\widetilde{\lambda}\mu_F$ and $\widetilde{\lambda}\tau$ respectively ($\mu_F = \mu$ being the mean of the length distribution F), the total numbers of observed line-segments is Poisson($\widetilde{\lambda}(\mu+\tau)$) distributed; the probabilities that an observed line-segment is left-censored or left-uncensored are $\mu/(\mu+\tau)$ and $\tau/(\mu+\tau)$ respectively. The *residual length* $Y = T + X$ of a left-censored line-segment ($T < 0$) is, by a simple calculation, continuously distributed with density $(1-F(y))/\mu$, $y > 0$ (the same formula as for the forward recurrence time in a stationary renewal process). The total length X of a left-uncensored line-segment has the original distribution F. The residual lengths of left-censored line-segments are censored (in the classical sense) at the fixed value τ. The length of left-uncensored observations are randomly censored (again in the classical sense) by $\tau - T$, independent of X and uniform$(0,\tau)$ distributed.

So the probability (density) of each of the four kinds of observations $\widetilde{X}$, up to factors not depending on F, is:

	r.u.c.	r.c.
l.u.c.	$\mathrm{d}F(\widetilde{X})$	$1-F(\widetilde{X})$
l.c.	$(1-F(\widetilde{X}))/\mu$	$\int_{\widetilde{X}}^{\infty}(1-F(y))\mathrm{d}y/\mu$

In the lower right entry (double-censored) the observed length $\widetilde{X} = \tau$ identically.

Now we recall that $\widetilde{\lambda}$ is infinitesimal so that with large probability $e^{-\widetilde{\lambda}(\tau+\mu)}$ there is no observation in the strip's transect of the window, with probability $\widetilde{\lambda}(\tau+\mu)e^{-\widetilde{\lambda}(\tau+\mu)}$ just one observation, and the probability of more than one observation may be neglected. So the probability of the observed data from one strip is proportional to a product of terms selected as follows:

	always	if obsvn.	if obsvn.	if r.u.c.	if r.c.
	$e^{-\widetilde{\lambda}(\tau+\mu)}$	$\widetilde{\lambda}(\tau+\mu)$			
if l.u.c.			$\frac{\tau}{\tau+\mu}$	$\mathrm{d}F(\widetilde{X})$	$1-F(\widetilde{X})$
if l.c.			$\frac{\mu}{\tau+\mu}$	$\frac{1-F(\widetilde{X})}{\mu}$	$\int_{\widetilde{X}}^{\infty}\frac{1-F}{\mu}$

On cancellation and substitution for $\widetilde{\lambda}$ we obtain the product of

always	if obsvn.	if u.c.	if s.c.	if d.c.
$e^{-\lambda\mathrm{d}rK(\mathrm{d}\theta)(\tau+\mu)}$	$\lambda\mathrm{d}K(\Theta)$	$\mathrm{d}F(\widetilde{X})$	$1-F(\widetilde{X})$	$\int_{\widetilde{X}}^{\infty}(1-F(y))\mathrm{d}y$

where u.c., s.c. and d.c. stand for uncensored, single-censored and double-censored respectively.

Now we multiply over all strips. The terms which only appear when a line-segment is actually observed get multiplied over the observations. The exponential term, always present, becomes an exponential of a sum over all strips crossing the window, therefore of an integral. Note that $\tau = \tau(r, \theta)$ while $\mu = \mu_F$ is constant. Splitting the integral of $(\tau + \mu)\mathrm{d}rK(\mathrm{d}\theta)$ into the sum of two integrals and integrating over r before θ, we note that in the first term, $\tau\mathrm{d}r$ is the area of the intersection of strip and window. Integrating over r gives the area of the window, denoted $|W|$. The integral of $K(\mathrm{d}\theta)$ is then equal to 1. For the second term, integrating $\mathrm{d}r$ over r gives the *diameter* of the window as seen in the θ direction, which we denote $\mathrm{diam}(W, \theta)$. Multiplied by $K(\mathrm{d}\theta)$ and integrating over θ gives the average (with respect to the distribution K) diameter, which we denote $\mathrm{E}_K\mathrm{diam}(W)$.

Let N denote the total number of observed line-segments. The result of all these computations, when we have inserted factors $\lambda(|W| + \mu_F\mathrm{E}_K\mathrm{diam}(W))$ to the powers plus and minus N, is:

$$\begin{aligned} &e^{-\lambda(|W|+\mu_F\mathrm{E}_K\mathrm{diam}(W))}(\lambda(|W| + \mu_F\mathrm{E}_K\mathrm{diam}(W)))^N \\ &\qquad\qquad\cdot\ (|W| + \mu_F\mathrm{E}_K\mathrm{diam}(W))^{-N} \cdot \prod_1^N \mathrm{d}K(\Theta_i) \\ &\qquad\qquad\cdot \prod_{\text{u.c.}} \mathrm{d}F(\widetilde{X}_i) \prod_{\text{s.c.}} (1 - F(\widetilde{X}_i)) \prod_{\text{d.c.}} \int_{\widetilde{X}_i}^{\infty} (1 - F(y))\mathrm{d}y \end{aligned} \tag{1}$$

The first line represents the Poisson distribution, with mean

$$\lambda(|W| + \mu_F\mathrm{E}_K\mathrm{diam}(W)),$$

of N; the next two lines give the joint conditional distribution, given N, of observed orientations Θ_i, censored lengths $\widetilde{X}_i$ and types u.c., s.c., and d.c. Since the range of $\widetilde{X}_i$ usually depends on Θ_i, orientations and lengths are not generally independent despite the product form. The factor $(|W| + \mu_F\mathrm{E}_K\mathrm{diam}(W))^{-N}$ depends on both K and F, and belongs just as well to the third as to the second line of (1). Together, these two lines give the conditional joint distibution of the observations given N; it is a product over $i = 1, \dots, N$ of i.i.d. observations, with a distribution depending on F and K but not λ.

Now if λ is unknown the Poisson mean of N is also completely unknown. This means that the NPMLE $(\widehat{\lambda}, \widehat{F}, \widehat{K})$ of the three parameters based on the joint likelihood (1) can be calculated by computing the NPMLEs of F and K from the conditional likelihood of the data given N, i.e., the second and third lines of (1), and then setting the observed value of N equal to its mean $\lambda(|W|+\mu_F\mathrm{E}_K\mathrm{diam}(W))$ after substitution of $\widehat{F}$ and $\widehat{K}$ for F and K respectively. In fact we will ignore λ from now on and consider only the conditional distribution of the data given $N = n$, which depends only on F and K. Asymptotics will be done 'as $n \to \infty$' which corresponds to 'as $\lambda \to \infty$'. Conventionally, asymptotics have been done for this kind of problem 'as the window W

becomes larger'. However in that case the edge effects which interest us become less and less important and in the limit maybe only turn up in some kind of second-order terms; whereas with our asymptotics, they remain equally important all the time.

So we would like to compute $\widehat{F}$ and $\widehat{K}$ by jointly maximizing the last two lines of (1). Unfortunately this does not decompose into separate maximization problems for F and K, though one can think of a natural iterative scheme: alternately determine F given K by maximizing

$$(|W| + \mu_F \mathrm{E}_K \mathrm{diam}(W))^{-n} \prod_{\text{u.c.}} \mathrm{d}F(\widetilde{X}_i) \prod_{\text{s.c.}} (1 - F(\widetilde{X}_i)) \prod_{\text{d.c.}} \int_{\widetilde{X}_i}^{\infty} (1 - F(y))\mathrm{d}y, \tag{2}$$

and K for given F by maximizing

$$(|W| + \mu_F \mathrm{E}_K \mathrm{diam}(W))^{-n} \prod_1^n \mathrm{d}K(\Theta_i). \tag{3}$$

It will be very important to have fast algorithms for the two separate maximizations then! Alternatively one could do both maximizations for a range of fixed values, e.g., of $\mathrm{E}_K\mathrm{diam}(W)$, use numerical interpolation to maximize, and then recompute at this value.

Laslett's (1982a) main contribution is to show how a version of the EM algorithm can be used to maximize (2) for given $\mathrm{E}_K\mathrm{diam}(W)$. Maximization of (3) for given μ_F is a much easier problem, left to the reader to analyse (**Exercise!**). Below we will show how (2) can be maximized in the one-dimensional case, which as far as these computations are concerned is actually not essentially easier (the likelihood looks exactly the same, only all double censored observations happen to be equal to one another).

We should explain exactly what we mean by 'maximization over F' of a expression like (2). Usually, a maximum likelihood estimator (MLE) is understood to be that value of an unknown parameter which maximizes, over possible parameter values, the density of the observations with respect to a suitable dominating measure, where the density is evaluated at the actually observed data. This function is called the likelihood function. In nonparametric problems like the present there is no dominating measure: both discrete and continuous F and K are a priori possible; if discrete, we do not know the support of the distribution; even if continuous, the distributions need not be absolutely continuous with respect to Lebesgue measure; and so on. There is therefore no likelihood to be maximized! However each *pair* of parameter values does permit calculation of a two-point likelihood function, since any two probability distributions are dominated by another measure (e.g., their sum). Thus any two parameter values can be compared to one another. The NPMLE, if it exists, is by definition (Kiefer and Wolfowitz, 1956) that value of the parameter which beats any other in all possible pairwise comparisons.

It is often easy to see that any distribution with some mass not at the observations is 'beaten' by some distribution with mass only at the observations. Computation of the NPMLE then reduces to ordinary computation of the MLE assuming a discrete distribution with known support. That happens in this problem. The NPMLE of K for given F is an implicitly weighted empirical distribution; the NPMLE of F for given K puts mass on the observed uncensored observations and also, perhaps counterintuitively, on the observed censored observations (as well as some mass to the right of the largest observation, location undetermined). We denote by the sieved NPMLE the result of maximizing only over distributions with mass on the uncensored observations (and to the right of all observations).

Consistency in the one-dimensional problem.

Now we reduce to one-dimension; the window W is the interval $[0, \tau]$ on the real line. The parameter K disappears; the (fixed) diameter of the window is just its length τ. The NPMLE of F is computed by maximizing (2), in which all doubly censored observations are identically equal to τ, which is also the value of $\mathrm{E}_K \mathrm{diam}(W)$. We have conditioned on $N = n$.

Our approach is simply by a reparametrization to absorb the difficult factor $(\mu + \tau)^{-n}$ into each of the n terms in the rest of the product, making the distribution (of one observation, $n = 1$) linear in the parameter. With now $\widetilde{\lambda} = \lambda$ let us reconsider the Poisson point process introduced above. The introduction of a second and parallel diagonal line (slope -1), intersecting the t-axis at $t = \tau$, splits the two regions of observable line segments into a total of four regions, corresponding to uncensored, singly-left-censored, singly-right-censored and doubly-censored observations. The n observed line-segments correspond to (T, X) which are distributed over the union of these four regions with distribution

$$\frac{\mathrm{d}t F(\mathrm{d}x)}{\tau + \mu} = V(\mathrm{d}x) \frac{\mathrm{d}t}{\tau + x} \tag{4}$$

where we define V, the marginal distribution of X, by

$$V(\mathrm{d}x) = \frac{\tau + x}{\tau + \mu} F(\mathrm{d}x). \tag{5}$$

This follows since, by inspection, the second factor of the right-hand side of (4) is the conditional distribution of T given $X = x$ (uniform on $[-x, \tau]$); what is left must be the marginal distribution of X.

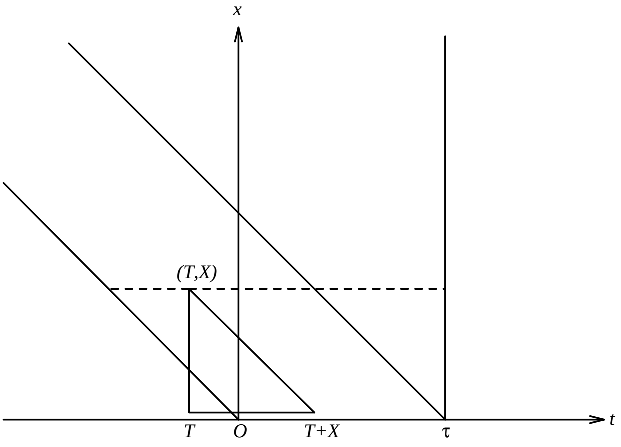

Figure 5.

One can show that as F varies over all possible distributions (with finite mean μ_F),

V too varies over all possible distributions. One can recover μ_F from $\mathrm{E}_V(1/(\tau+X)) = 1/(\tau+\mu)$, and hence recover F from (5).

The i.i.d. pairs (T_i, X_i), with X_i distributed as V and the T_i conditionally uniformly distributed as above, are however not completely observed. Instead they are grouped according to a certain rule: in the region 's.l.c.' onto diagonal lines (parallel with the two others), i.e., we observe the value of $T+X$; in the region 's.r.c.' they are grouped onto vertical lines (giving the value of T); in the region 'd.c.' they are grouped together completely to a single value; and in the region 'u.c.' they are not grouped at all but remain completely observed as points.

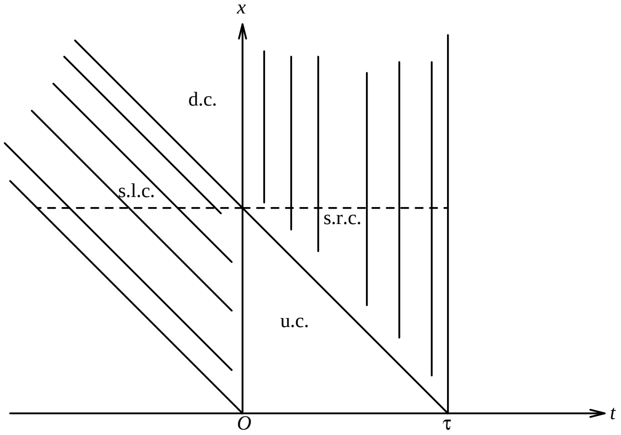

Figure 6.

This is now a more or less classical situation for computing the NPMLE of arbitrarily grouped or censored data. First we write down, by our description of the model and the grouping scheme, the self-consistency equation (Efron, 1966, Turnbull, 1976):

$$\widehat{V}(\mathrm{d}x) = \frac{1}{n}\sum_{i=1}^{n} \mathrm{P}_{\widehat{V}}(X_i \in \mathrm{d}x | \widetilde{X}_i, \Delta_i). \tag{6}$$

In words, the NPMLE of V is such that, for any region A, the estimate $\widehat{V}(A)$ should equal the average of the conditional probabilities, computed under $\widehat{V}$, that each underlying observation X_i lies in the region A given what is known about it. The right-hand side of (6) is easy to write out explicitly, thanks to our simple grouping model; we do that in a moment.

Next, one can calculate $\widehat{V}$ by the natural iteration scheme based on (6), an instance of the EM algorithm (Dempster, Laird and Rubin, 1977). To be more precise, one must first agree on the support of $\widehat{V}$ and start the iterations with a distribution which does not have a smaller support. In our case we distinguish between the NPMLE with support on all observations, and the sieved NPMLE with support only on the uncensored observations. Both satisfy (6) and both can be iteratively calculated by the EM algorithm; only the starting point must reflect our choice. (The proper NPMLE may have zero mass on *some* of the singly censored values, but this is hard to determine in advance and not important for the algorithm). We show later an alternative, faster way to calculate the sieved NPMLE.

Before proceeding we should be more specific about a difficulty which arises here: the distribution of one observation $(\widetilde{X},\Delta)$ does not determine V completely, but only depends on V through its restriction (in the sense of a defective distribution) to $[0,\tau)$ together with two 'tail integrals': $\int_{\tau-}^{\infty} \mathrm{d}V = 1-V(\tau-)$ and less trivially

$$h = h_V = \int_{\tau-}^{\infty} \frac{x-\tau}{\tau+x} V(\mathrm{d}x) = \mathrm{P}(\text{d.c.}). \tag{7}$$

This is because everything above $\tau-$ is grouped together either as a singly or as a doubly censored observation. Wijers (1991) shows that there is a 1–1 correspondence between $(F|_{[0,\tau)},\mu_F)$ and $(V|_{[0,\tau)},h_V)$. To be specific (these relations are easy to derive) use:

$$\frac{2\tau}{\tau+\mu} = 1 + \int_0^{\tau} \frac{\tau-x}{\tau+x} V(\mathrm{d}x) - h; \qquad F(\mathrm{d}x) = \frac{\tau+\mu}{\tau+x} V(\mathrm{d}x). \tag{8}$$

The first part of (8) says that the probabilities of left-uncensored plus right-uncensored equal one plus completely uncensored minus double censored. It will turn out that $V(\tau-)$ is not actually root n rate estimable, but $\int_0^{\tau}((\tau-x)/(\tau+x))V(\mathrm{d}x)$ fortunately is (or obviously, depending on how you look at it: $\tau/(\mu+\tau)$ is the probability of a left-uncensored observation so it and h and μ are trivially root n rate estimable).

From our picture of the joint distribution of (T,X) and the grouping scheme, it is easy to verify that the $\widetilde{X}$ with $\Delta=$ l.c. and those with $\Delta=$ r.c. are both continuously distributed with density $g(x)$ on $[0,\tau]$ given by

$$g(x) = \int_x^{\infty} \frac{V(\mathrm{d}y)}{\tau+y} = \int_x^{\tau-} \frac{V(\mathrm{d}y)}{\tau+y} + g(\tau-). \tag{9}$$

One can also verify (by inspection of the picture) that

$$V(\tau-) + 2\tau g(\tau-) + h = 1 \tag{10}$$

where h was defined in (7): in other words, either $X<\tau$, or $X\geq\tau$ and singly-left-censored, or $X\geq\tau$ and singly-right-censored, or X is doubly censored. These identities turn out to be useful later.

From the picture we can write out the self-consistency equation (6) explicitly as, for $x\in[0,\tau)$,

$$\begin{aligned}
\widehat{V}(\mathrm{d}x) \text{ ``} &= F_n^{\text{u.c.}}(\mathrm{d}x) + \int_{t=0}^{x-} \frac{\widehat{V}(\mathrm{d}x)\frac{\mathrm{d}t}{\tau+x}}{\int_{y=t}^{\infty}\widehat{V}(\mathrm{d}y)\frac{\mathrm{d}t}{\tau+y}} F_n^{\text{s.c.}}(\mathrm{d}t)\text{''} \\
&= F_n^{\text{u.c.}}(\mathrm{d}x) + \frac{\widehat{V}(\mathrm{d}x)}{\tau+x}\int_{t=0}^{x-} \frac{F_n^{\text{s.c.}}(\mathrm{d}t)}{\int_{y=t}^{\infty}\frac{\widehat{V}(\mathrm{d}y)}{\tau+y}} \qquad (11)\\
\widehat{h} &= h_n = F_n^{\text{d.c.}}(\{\tau\})
\end{aligned}$$

where in the denominator of (11) we note that the integral is just $\widehat{g}(y)$ which can be expressed in various ways, according to (9) and (10), in terms of $\widehat{V}|_{[0,\tau)}$ and $\widehat{h}$. The different $F_n^{*.\text{c.}}$ here denote of course empirical (sub)-distribution functions.

Equation (11) is also satisfied by the true V and F; from our picture one verifies quickly that

$$\begin{aligned} F^{\text{u.c.}}(\mathrm{d}x) &= \frac{\tau - x}{\tau + x} V(\mathrm{d}x), \quad x < \tau, \\ F^{\text{s.c.}}(\mathrm{d}x) &= 2g(x)\mathrm{d}x, \quad x < \tau, \\ F^{\text{d.c.}}(\mathrm{d}x) &= h\delta_\tau(\mathrm{d}x), \end{aligned} \tag{12}$$

where g and h were defined above and δ_τ denotes point mass at τ.

These relations all become useful when we actually work out the details of the consistency proof. The idea of the proof however is quite general. For simplicity we pretend the parameter is just V.

According to the Kiefer-Wolfowitz definition of an NPMLE, writing P_V for the distribution of a single observation and P_n for the empirical distibution of the data, we have

$$\int \log \frac{\mathrm{dP}_{\widehat{V}}}{\mathrm{d}\mu} \mathrm{dP}_n \geq \int \log \frac{\mathrm{dP}_{\widetilde{V}}}{\mathrm{d}\mu} \mathrm{d}P_n \tag{13}$$

where $\widehat{V}$ is the NPMLE of V and $\widetilde{V}$ is any other value of V, while μ is a measure dominating both $\mathrm{P}_{\widehat{V}}$ and $\mathrm{P}_{\widetilde{V}}$. A well-proven method for showing consistency in parametric models is to use this inequality with $\widetilde{V} = V$, the true parameter value. For $n \to \infty$ one hopes to be able to replace P_n by P_V, and then to obtain a contradiction with the well-known fact (from Jensen's inequality) that $\int \log(\mathrm{d}P_{\widehat{V}}/\mathrm{d}\mu)\mathrm{dP}_V \leq \int \log(\mathrm{d}P_V/\mathrm{d}\mu)\mathrm{dP}_V$ with equality if and only if $\widehat{V} = V$ (where we assume identifiability: different V have different P_V).

In our situation nothing useful comes of this since (supposing the true P_V to be continuous whereas $\mathrm{P}_{\widehat{V}}$ is discrete) the inequality (13) with $\widetilde{V} = V$ becomes a triviality. Therefore, instead of comparing, on the data, $\mathrm{P}_{\widehat{V}}$ to the true P_V, we compare it to P_{V_n} where V_n is of the same nature as $\widehat{V}$ but known (asymptotically) to be close to V. To be precise, we define the pair $(V_n|_{[0,\tau)}, h_n)$ by

$$V_n(\mathrm{d}x) = \frac{\tau + x}{\tau - x} F_n^{\text{u.c.}}(\mathrm{d}x), \quad x < \tau,$$

$$h_n = \widehat{h} = F_n^{\text{d.c.}}(\{\tau\}).$$

This estimator is consistent and moreover has a similar discrete character to the NPMLE $(\widehat{V}|_{[0,\tau)}, \widehat{h})$. We learnt this idea from Murphy (1993) where it is applied very effectively to solve a very difficult problem concerning so-called frailty models for a counting process.

Rather than proceding to study $\int \log(\mathrm{dP}_{\widehat{V}}/\mathrm{dP}_{V_n})\mathrm{dP}_n$, we exploit the *convexity* and *linearity* of our model, according to which the line-segment between $\widehat{V}$ and V_n consists also of possible parameter values, while P_V is linear in V:

$$\mathrm{P}_{(1-\varepsilon)\widehat{V}+\varepsilon V_n} = (1-\varepsilon)\mathrm{P}_{\widehat{V}} + \varepsilon \mathrm{P}_{V_n}. \tag{14}$$

The idea we will use goes back to Jewell (1982), and has been used in various contexts by Wang (1985), Pfanzagl (1988), and Groeneboom and Wellner (1992). In most of these papers the method is applied for a specific model and its general nature not emphasized. (Also, in most of these previous cases it was not necessary to introduce the ad hoc estimator V_n; one could study the line-segment between $\widehat{V}$ and the true parameter value V). Note especially that (14) holds generally in *nonparametric missing data problems* in which the data is (i.i.d. copies of) the result of applying a many-to-one mapping to a pair (X,T), where the distribution V of X is completely unknown while the conditional distribution of T given X is fixed.

Since $\widehat{V}$ beats $(1-\varepsilon)\widehat{V}+\varepsilon V_n$ on the data, and then using the linearity (14), we have

$$\begin{aligned} 0 &\geq \int \log \frac{\mathrm{dP}_{(1-\varepsilon)\widehat{V}+\varepsilon V_n}}{\mathrm{dP}_{\widehat{V}}} \mathrm{dP}_n \\ &= \int \log\biggl((1-\varepsilon)+\varepsilon\frac{\mathrm{dP}_{V_n}}{\mathrm{dP}_{\widehat{V}}}\biggr)\mathrm{dP}_n \\ &= \int \log\biggl(1+\varepsilon\biggl(\frac{\mathrm{dP}_{V_n}}{\mathrm{dP}_{\widehat{V}}}-1\biggr)\biggr)\mathrm{dP}_n \end{aligned}$$

By concavity of $\varepsilon \mapsto \log(1+\varepsilon a)$ this is also concave in ε, with a maximum at $\varepsilon = 0$ hence a derivative with respect to ε at $\varepsilon = 0$ which is nonpositive:

$$\int \biggl(\frac{\mathrm{dP}_{V_n}}{\mathrm{dP}_{\widehat{V}}}-1\biggr)\mathrm{dP}_n \leq 0,$$

or, equivalently,

$$\int \frac{\mathrm{dP}_{V_n}}{\mathrm{dP}_{\widehat{V}}}\mathrm{dP}_n \leq 1. \tag{15}$$

Our programme will be to assume, by relative compactness of the space of (possibly defective) distribution functions, that for each given ω, along some subsequence, $\widehat{V} \xrightarrow{\mathcal{D}} V_\infty$ for some possibly defective distribution V_∞. At the same time $V_n \to V$ and $\mathrm{P}_n \to \mathrm{P}_V$ in the sense of the Glivenko-Cantelli theorem. Using the self-consistency equation for $\widehat{V}$ it turns out that we know enough about $\mathrm{dP}_{V_n}/\mathrm{dP}_{\widehat{V}}$ in order to prove that the left hand side of (15) converges, along this subsequence, to its natural limit, giving the inequality

$$\int \frac{\mathrm{dP}_V}{\mathrm{dP}_{V_\infty}}\mathrm{dP}_V \leq 1. \tag{16}$$

On the other hand consider the line segment $\varepsilon \mapsto (1-\varepsilon)V_\infty+\varepsilon V$. By Jensen's inequality,

$$\varepsilon \mapsto \int \log \frac{\mathrm{dP}_{(1-\varepsilon)V_\infty+\varepsilon V}}{\mathrm{dP}_{V_\infty}}\mathrm{dP}_V$$

is maximal at $\varepsilon = 1$, and strictly maximal there unless $\mathrm{P}_{V_\infty} = \mathrm{P}_V$ (which would imply

$V_\infty = V$). But this integral equals

$$\int \log\Bigl((1-\varepsilon) + \varepsilon \frac{\mathrm{dP}_V}{\mathrm{dP}_{V_\infty}}\Bigr)\mathrm{dP}_V = \int \log\Bigl(1 + \varepsilon\Bigl(\frac{\mathrm{dP}_V}{\mathrm{dP}_{V_\infty}} - 1\Bigr)\Bigr)\mathrm{dP}_V,$$

concave in ε, being an average of concave functions. Therefore its derivative at $\varepsilon = 0$ is *nonnegative*; strictly so unless $V_\infty = V$. But this derivative equals $\int(\mathrm{dP}_V/\mathrm{dP}_{V_\infty} - 1)\mathrm{dP}_V$ so we have the reverse inequality

$$\int \frac{\mathrm{dP}_V}{\mathrm{dP}_{V_\infty}}\mathrm{dP}_V \geq 1 \tag{17}$$

with equality if and only if $V_\infty = V$. Now the usual argument (from *any* subsequence we can extract a convergent sub-subsequence, with the same limit V) shows that (for the given ω) along the original sequence $\widehat{V} \xrightarrow{\mathcal{D}} V$. This is the required strong consistency of $\widehat{V}$. (In fact a closer analysis in this specific problem shows that $\widehat{V}$ is consistent not just in the sense of weak convergence but also in the supremum norm).

We sketch the beginnings of the calculations which have to be done to carry through this argument. Recall our expressions (12) for the distribution of the data. These equations, with (9) and (10), express P_V in terms of V or rather $(V|_{[0,\tau)}, h)$, and hold also when $(V|_{[0,\tau)}, h)$ is replaced by $(\widehat{V}|_{[0,\tau)}, \widehat{h})$ or $(V_n|_{[0,\tau)}, h_n)$. Substituting into (15), and letting $\widehat{g}$ and g_n be defined as g of (9) and (10) but for the corresponding estimators, we find

$$\int \frac{\mathrm{dP}_{V_n}}{\mathrm{dP}_{\widehat{V}}}\mathrm{dP}_n = \int_0^{\tau-} \frac{\mathrm{d}V_n}{\mathrm{d}\widehat{V}}(x) F_n^{\text{u.c.}}(\mathrm{d}x) + \int_0^{\tau-} \frac{g_n(t)}{\widehat{g}(t)} F_n^{\text{s.c.}}(\mathrm{d}t) + \frac{h_n}{\widehat{h}} F_n^{\text{d.c.}}(\{\tau\}) \leq 1. \tag{18}$$

Since $V_n(\mathrm{d}x) = ((\tau+x)/(\tau-x))F_n^{\text{u.c.}}(\mathrm{d}x)$ while the self-consistency equation (11) tells us

$$\frac{\mathrm{d}F_n^{\text{u.c.}}}{\mathrm{d}\widehat{V}}(x) = 1 - \frac{1}{\tau+x}\int_0^{x-} \frac{F_n^{\text{s.c.}}(\mathrm{d}t)}{\int_t^\infty \frac{\widehat{V}(\mathrm{d}y)}{\tau+y}}$$

we find for the first part of (18) that

$$\frac{\mathrm{d}V_n}{\mathrm{d}\widehat{V}}(x) = \frac{1}{\tau - x}\left((\tau+x) - \int_0^{x-} \frac{F_n^{\text{s.c.}}(\mathrm{d}t)}{\int_t^\infty \frac{\widehat{V}(\mathrm{d}y)}{\tau+y}}\right).$$

For the second part of (18) the defining equations (9) and (10) for g allow us similarly to express $g_n/\widehat{g}$ in terms of integrals with respect to F_n and $\widehat{V}$, while the third term is trivial since h_n and $\widehat{h}$ both equal the same quantity.

The upshot of this is that (18) can be written out entirely in terms of (repeated) integrals with respect to the NPMLE $\widehat{V}$ and the empirical distribution functions $F_n^{*.\text{c.}}$. Since we assume $\widehat{V}$ converges to V_∞ and we know the empiricals converge to the true, it is now a question of pure analysis to show that (18) converges to its natural limit. Details are given in Wijers (1991). Most of the analysis is routine; the only difficulties are met 'at the endpoint' τ. Restricting the integrals over $[0,\tau)$ to integrals over $[0,\sigma]$, $\sigma<\tau$, convergence is quite easy to obtain. Since the remainder from σ to τ is nonnegative we get the inequality in the limit provided we only integrate over the smaller intervals. Finally we let σ increase up to τ and keep the inequality by monotone convergence, giving us (16). The reverse inequality (17) is true, without any further work, so $V_\infty = V$.

Exercise. Consider the classical random censorship model with G fixed and known. Take $V=F$. Prove consistency of the NPMLE (which equals the Kaplan-Meier estimator $\widehat{F}$) by Wijers' approach, using just the self-consistency equation for $\widehat{V}$ and comparing $\widehat{V}$ to V_n defined by $V_n(\mathrm{d}x)=F_n^{\text{u.c.}}(\mathrm{d}x)/(1-G(x-)$. Note that in your proof the explicit expression for $\widehat{F}$ as product-limit estimator is not made use of.

The sieved NPMLE.

We see in the above that a consistency proof for the NPMLE in a linear-convex model, in particular in a nonparametric missing data model, can be given without an explicit expression for the estimator. As we will see, much more is possible if we make use of some general theory of semiparametric estimators. First however we introduce a variant of $\widehat{V}$ which in many respects seems to be better behaved.

The above proof of Wijers used the self-consistency equation and the fact that V_n is dominated by $\widehat{V}$, but nothing more. Restricting $\widehat{V}$ to only put mass on the uncensored observations does not change these properties. (It is really rather counterintuitive that the NPMLE should place any mass on censored observations at all). We call the resulting estimator the *sieved* NPMLE and from now on consider it rather than the NPMLE itself. The sieved NPMLE is also consistent, by the above proof; it may also be calculated by iterating the self-consistency equation but in fact can nearly be calculated explicitly, by another route. Note first that by (9) and (10) we can write

$$g(t)=\int_t^\infty \frac{V(\mathrm{d}x)}{\tau+x}=\frac{1}{2\tau}\Bigl(1-h+\int_0^\tau \frac{\tau-x}{\tau+x}V(\mathrm{d}x)-\int_0^t \frac{2\tau}{\tau+x}V(\mathrm{d}x)\Bigr).$$

Since the sieved $\widehat{V}$ is actually equivalent to $F_n^{\text{u.c.}}$, one can now rewrite the self-consistency equation for the sieved estimator as

$$\widehat{V}(\mathrm{d}x)=\frac{F_n^{\text{u.c.}}(\mathrm{d}x)}{1-\dfrac{1}{\tau+x}\displaystyle\int_{t=0}^x \frac{F_n^{\text{s.c.}}(\mathrm{d}t)}{\frac{1}{2\tau}\Bigl(1-\widehat{h}+\int_0^\tau \frac{\tau-x'}{\tau+x'}\widehat{V}(\mathrm{d}x')-\int_{x'=0}^t \frac{2\tau}{\tau+x'}\widehat{V}(\mathrm{d}x')\Bigr)}}. \tag{19}$$

Fix the value of

$$\widehat{F}^{\text{u.c.}}(\tau)=\int_0^\tau \frac{\tau-x}{\tau+x}\widehat{V}(\mathrm{d}x). \tag{20}$$

Since $\widehat{V}$ puts its mass on the uncensored observations, which will not coincide with any censored observations, one can always take $x'<t<x$ in the last (lower, right) integral in (19). Thus for a chosen value of the next to last integral (20), one can recursively

calculate $\widehat{V}(\{x\})$ at each of its atoms x. Now it is not difficult to check that $\widehat{V}(\{x\})$ is a decreasing function of (trial values of) (20). So we can use (19) to form a new value of (20) given an old one; this mapping is *decreasing*; so if we start with a trial value which happens to be below the solution (the fixed point of the mapping) we come out with a value above; and vice-versa. Therefore we propose the following algorithm: given a trial value of (20) compute a new value by recursive use of (19). Take the average of the old and new values, and repeat. This algorithm converges at least as fast as 'interval halving', but close to the solution where the mapping is almost linear it is much faster; quadratic (as the Newton-Raphson algorithm) rather than linear (as interval halving or, for that matter, the EM algorithm). (By linear convergence we mean that the number of leading zero's in the error $0.0000xyz\ldots$ increases at a constant rate; by quadratic convergence we mean that it doubles at each step. This is usually called exponential and super-exponential convergence respectively).

Semiparametric models.

In order to explain van der Laan's (1993b) approach to proving not just consistency but also asymptotic normality, efficiency, and correctness of the bootstrap for the sieved estimator $\widehat{V}$ we must explain some general ideas from the theory of semiparametric models; see van der Vaart (1991b) for a brief and precise summary, or ABGK chapter VIII for an extensive introduction.

For the time being we remain within our line-segment problem. Let (X,T) and $(\widetilde{X},\Delta)$ have the same meaning as above and consider two L^2 spaces: $L^2(V)$, the space of all square integrable functions of X, and $L^2(\mathrm{P}_V)$, the space of all square integrable functions of $(\widetilde{X},\Delta)$. When we add a suffix 0 we mean the subspaces of L^2 functions of mean zero. Introduce the operator $A: L^2(V)\to L^2(\mathrm{P}_V)$ defined by

$$(Ah)(\widetilde{X},\Delta)=\mathrm{E}(h(X)\mid \widetilde{X},\Delta)$$

and its adjoint A^* which is easily checked to be

$$(A^*g)(X)=\mathrm{E}(g(\widetilde{X},\Delta)\mid X).$$

We indicate norm and inner product in these two spaces by a subscript V and P respectively. Each $h\in L^2_0(V)$ corresponds to a one-dimensional parametric submodel in our large model (all possible V), passing through the given point V, defined by

$$\mathrm{d}V_{\theta,h}\propto(1+\frac{1}{2}\theta h)^2\mathrm{d}V,\quad \theta\in\mathrm{I\!R}.$$

Write also $V_h=V_{1,h}$ and note that $V_{\theta,h}=V_{\theta h}$. With one observation of $X\sim V_{\theta,h}$ the score function for θ at $\theta=0$ (i.e., the derivative of the log likelihood), would be $h(X)$ itself. With one observation $(\widetilde{X},\Delta)$ from $\mathrm{P}_{V_{\theta,h}}$ the score function turns out to be $Ah(\widetilde{X},\Delta)$. We call A the *score operator*. (Other submodels with the same score functions could also have been considered; e.g., $\mathrm{d}V_h\propto(1+h)^+\mathrm{d}V$ or $\mathrm{d}V_h\propto\exp(h)\mathrm{d}V$).

Suppose we are interested in estimating $\kappa=\kappa(V)=V(x_0)$ for a fixed x_0. It is easy to check that the derivative of $\kappa(V_{\theta,h})$ with respect to θ at $\theta = 0$ equals

$$\int_0^{x_0} h\mathrm{d}V = \int_0^{\infty} h(x)\big(1_{[0,x_0]}(x) - V(x_0)\big)V(\mathrm{d}x).$$

The last integral is the inner product in $L_0^2(V)$ of h with

$$\dot\kappa = 1_{[0,x_0]} - V(x_0).$$

This makes $\dot\kappa$ the directional derivative of $\kappa(\mathrm{P}_{V_h})$ at $h = 0$ with respect to $h \in L_0^2(V)$.

Now in the one-dimensional submodel indexed by h, and with parameter $\theta \in \mathrm{I\!R}$, the Fisher information for θ at $\theta = 0$ based on one observation $(\widetilde{X}, \Delta)$ is the expected squared score:

$$\mathrm{E}((Ah)^2) = \|Ah\|_{\mathrm{P}}^2 = \langle Ah, Ah\rangle_{\mathrm{P}} = \langle A^*Ah, h\rangle_V. \tag{21}$$

The derivative of the parameter of interest $V_{\theta,h}(x_0)$ with respect to θ, at $\theta = 0$, is

$$\langle \dot\kappa, h\rangle_V \tag{22}$$

and hence the Cramér-Rao lower bound for n times the variance of an unbiased estimator of κ based on n i.i.d. observations is

$$\frac{\langle \dot\kappa, h\rangle_V^2}{\langle A^*Ah, h\rangle_V}. \tag{23}$$

According to general theory this quantity, also called the information bound, is also the optimal *asymptotic* variance of $\sqrt{n}(\widehat\kappa - \kappa)$ for a so-called *regular* estimator, or more precisely, sequence of estimators $\widehat\kappa = \widehat\kappa_n((\widetilde{X}_1, \Delta_1), \ldots, (\widetilde{X}_n, \Delta_n))$, for the given submodel; see van der Vaart (1991b), ABGK chapter VIII. We now can vary h, and look for the *largest* information lower bound, or in other words the *hardest parametric submodel* $\mathrm{P}_{V_{\theta,h}}$, for estimating κ at the common point $\mathrm{P}_{V_{0,h}}$. A simple calculation shows that if $\dot\kappa$ is in the range of A^*A (the so-called information operator, a map from $L^2(V)$ to itself) and therefore an inverse $(A^*A)^{-1}h \in L^2(V)$ exists, then this hardest submodel is given by

$$h = (A^*A)^{-1}\dot\kappa$$

with score function

$$g = A(A^*A)^{-1}\dot\kappa.$$

Moreover, for this submodel, the information (21), the derivative (22) and the information bound (23) all coincide and are equal to

$$\langle (A^*A)^{-1}\dot\kappa, \dot\kappa\rangle_V \tag{24}$$

Sometimes the information operator cannot be inverted at $\dot\kappa$ but still the supremum over h of the information bound (23) is finite. Van der Vaart (1991b) shows that one has a finite supremum if and only if $\dot\kappa$ is in the range of A^*. This condition is therefore a necessary condition for the existence of 'root n rate, regular' estimators of κ in our large model with V varying freely.

Define

$$\mathrm{IC}(\widetilde{X},\Delta)=\big(A(A^*A)^{-1}\dot{\kappa}\big)(\widetilde{X},\Delta),$$

supposing inverse to exist; this is the so-called *optimal influence curve* for estimating κ at V. It depends indeed both on the functional κ being estimated and the point V at which we are working. The reason for the name influence curve is that a necessary and sufficient condition for an estimator to be optimal at V is that

$$\widehat{\kappa}-\kappa=\frac{1}{n}\sum_1^n \mathrm{IC}(\widetilde{X}_i,\Delta_i)+o_{\mathrm{P}}(n^{-\frac{1}{2}}). \tag{25}$$

It is certainly easy to check that if an estimator has this stochastic expansion then it is asymptotically normal (at root n rate) with asymptotic variance equal to (24), the greatest lower bound over parametric submodels and hence called the information bound for our nonparametric model. (If an estimator is asyptotically linear in the sense of (25) for *some* fixed function of each observation then this function is called its influence curve).

The optimal influence curve has several other names and corresponding interpretations. As we saw above it is also the *score function* Ah for the hardest parametric submodel (with $h=(A^*A)^{-1}\dot{\kappa}$). It is therefore also often called the *efficient score*. Another name is *canonical gradient*. This name comes from considering κ not as a function of V but of the distribution of one observation P_V (assuming identifiability). Recall that if $\mathrm{d}V_h\approx(1+h)\mathrm{d}V$ then $\mathrm{dP}_{V_h}\approx(1+Ah)\mathrm{dP}_V$ and $\kappa(V_h)\approx\kappa(V)+\langle\dot{\kappa},h\rangle_V$. Putting $Ah=g$ or $h=(A^*A)^{-1}A^*g$ we have

$$\mathrm{dP}_{V_h}\approx(1+g)\mathrm{dP}_V$$

while

$$\begin{aligned}\kappa(V_h)&\approx\kappa(V)+\langle\dot{\kappa},(A^*A)^{-1}A^*g\rangle_V\\&=\kappa(V)+\langle(A^*A)^{-1}\dot{\kappa},A^*g\rangle_V\\&=\kappa(V)+\langle A(A^*A)^{-1}\dot{\kappa},g\rangle_{\mathrm{P}}.\end{aligned}$$

So if $\mathrm{dP}'/\mathrm{dP}\approx 1+g$ we have the corresponding

$$\begin{aligned}\kappa'\approx\kappa+\langle\mathrm{IC},g\rangle_{\mathrm{P}}&\approx\kappa+\langle\mathrm{IC},\frac{\mathrm{dP}'}{\mathrm{dP}}-1\rangle_{\mathrm{P}}\\&=\kappa+\int\mathrm{ICdP}'.\end{aligned}\tag{26}$$

The optimal influence curve IC is not the unique gradient (or derivative) since adding to it a function orthogonal to all possible $g=Ah$ does not change the linear approximation to $\kappa'-\kappa$. However the present choice is the *smallest* such derivative in terms of L^2 norm. Note that (26) suggests that if we know the distribution $\mathrm{P}'=\mathrm{P}_{V'}$ of one observation is close to P_V for a given V then we could estimate κ' with the empirical analogue $\kappa+\int\mathrm{ICdP}_n$. Choosing the version of the derivative with smallest norm corresponds to the 'local, linear estimator' with smallest variance.

So far most of what we have said has been independent of our specific model.

The form of the score operator A and its adjoint as conditional expectation operators is common to all *missing data models*. For instance, consider the classical random censorship model with unknown distribution function $F = V$ of the survival times and fixed distribution G of the censoring times, and suppose we want to estimate functionals of F such as its value at a specific point t. This is a nonparametric missing data model, and one may check that the score operator A is given by $(Ah)(\widetilde{T}, \Delta) = \mathrm{E}(h(T) \mid \widetilde{T}, \Delta)$, its adjoint is $(A^*g)(T) = \mathrm{E}(g(\widetilde{T}, \Delta) \mid T)$, and the optimal influence curve for $F(t)$ is nothing else than the influence curve of the Kaplan-Meier estimator $\widehat{F}(t)$. For an elegant proof of this, obtained by transforming from densities to hazard rates, in terms of which the inversion of the information operator is trivial (it becomes 'diagonal', corresponding to the asymptotic independent increments of the Nelson-Aalen estimator), see Ritov and Wellner (1988).

In our model (and also in the random censorship problem with fixed censoring distribution) we have another special feature: linearity. Suppose we want, as above, to estimate $\kappa = V(x_0)$. The mapping from V to κ is linear but so also is the mapping from V to P_V. This means, assuming identifiability, that the mapping from P_V to κ is *also* linear and hence our 'linear approximation' (26) is actually an *equality*, which we can write as:

$$\kappa(\mathrm{P}') = \kappa(\mathrm{P}) + \mathrm{E}_{\mathrm{P}'}(\mathrm{IC}_{\mathrm{P}}) \tag{27}$$

where we emphasize by the subscript that the influence curve or derivative is evaluated at P, not P′. Now we are ready to derive van der Laan's (1993a) identity for the NPMLE of a linear parameter in a convex-linear model. Write $\widehat{V}$, $\widehat{\kappa} = \kappa(\widehat{V})$, $\widehat{\mathrm{P}} = \mathrm{P}_{\widehat{V}}$ for the NPMLEs of V, κ and P. Since the optimal influence curve depends on the point V at which it is evaluated we can also write $\widehat{\mathrm{IC}}$ for the optimal influence curve 'at $\widehat{V}$'. As usual P_n stands for the empirical distribution of the data.

In (27) take $\widehat{P}$ for P and P for P′. This gives the equality

$$\kappa = \widehat{\kappa} + \mathrm{E}_{\mathrm{P}}(\widehat{\mathrm{IC}}).$$

Since $\widehat{\mathrm{IC}}$ is also the score function at $\widehat{V}$ of the hardest submodel *through* $\widehat{V}$, while $\widehat{V}$ is the NPMLE, this point on the curve is also the ordinary MLE within this submodel. Therefore the likelihood equation—derivative of log likelihood or sum of scores equals zero—is satisified. But this equation can be rewritten as

$$\mathrm{E}_{\mathrm{P}_n}(\widehat{\mathrm{IC}}) = 0.$$

Therefore we have the identity

$$\widehat{\kappa} = \kappa + (\mathrm{E}_{\mathrm{P}_n} - \mathrm{E}_{\mathrm{P}})(\widehat{\mathrm{IC}}).$$

Some smoothness conditions have to be checked to make sure this identity really is true. *If* it is true we are now in an excellent position to study properties of $\widehat{\kappa}$ by empirical process theory. To begin with, if $\{\mathrm{IC}_V : V \in \mathcal{V}\}$, where $\mathcal{V}$ if the space of all parameter values V, is a *Glivenko-Cantelli class*, then we have consistency of $\widehat{\kappa}$. Suppose next we can prove consistency of such a large class of functionals κ that we can prove consistency of the estimated influence curve $\widehat{\mathrm{IC}}$, in the $L^2(\mathrm{P}_V)$ sense, for the specific functional κ of interest. If then $\{\mathrm{IC}_V : V \in \mathcal{V}\}$ is also a *Donsker class*, we now have asymptotic normality and even optimality of $\widehat{\kappa}$, since to put it more informally we now have

$$\widehat{\kappa} \approx \kappa + (\mathrm{E}_{\mathrm{P}_n} - \mathrm{E}_{\mathrm{P}})(\mathrm{IC}).$$

For the *sieved* NPMLE in the line-segment model all these things are true. Above we have specified IC_V in terms of the derivative $\dot{\kappa}$ and the score operator A. It is not difficult to work out the information operator: one finds

$$\begin{aligned}(A^*Ah)(X) = 1_{[0,\tau)}(X)\frac{\tau - X}{\tau + X}h(X) \\ + \frac{2}{\tau + X}\int_0^{X\wedge\tau} \frac{\displaystyle\int_t^\infty \frac{h(x)V(\mathrm{d}x)}{\tau + x}}{\displaystyle\int_t^\infty \frac{V(\mathrm{d}x)}{\tau + x}}\mathrm{d}t \\ + 1_{(\tau,\infty)}(X)\frac{X - \tau}{\tau + X}\frac{\displaystyle\int_\tau^\infty \frac{\tau - x}{\tau + x}h(x)V(\mathrm{d}x)}{\displaystyle\int_\tau^\infty \frac{\tau - x}{\tau + x}V(\mathrm{d}x)}.\end{aligned}$$

It is not possible to explicitly invert this operator. In fact it does not even have a unique inverse—not surprisingly, since if we parametrise by V our model is not identified; different behaviours of h past τ can lead to the same parametric submodels. However it is possible to define one inverse more or less explicitly, see (29) below, in terms of an infinite series (in fact, a Neumann series, or if you prefer, a Peano series, corresponding to the inversion of a Volterra type operator). The operators involved are nice enough that one can show that $\{\mathrm{IC}_V : V \in \mathcal{V}\}$ consists of bounded functions of uniformly bounded variation, continuous in the appropriate sense in V. This means that we do have a Donsker class and the approach gives us all the information we want; see van der Laan (1993b).

In particular it is natural to consider several functionals κ simultaneously; if the now doubly indexed class of influence curves is a Donsker class, and are continuous in the appropriate sense, then from consistency we can get joint weak convergence of all the estimators. If one considers for instance the influence curves simultaneously for estimating each $V(x)$, $x \in [0,\tau)$, it turns out that we do not have a Donsker class any more; the optimal influence curves are not bounded. One can however consider all $V(x)$, $x \in [0,\sigma]$, for any chosen $\sigma < \tau$, obtain a Donsker class, and conclude weak convergence of $n^{\frac{1}{2}}(\widehat{V} - V)$ in $D[0,\sigma]$. Alternatively it is quite natural to parametrise not by $(V|_{[0,\tau)}, h)$ but by $(W|_{[0,\tau]}, h)$ where $W(\mathrm{d}x) = (\tau - x)/(\tau + x))V(\mathrm{d}x) = F^{\mathrm{u.c.}}(\mathrm{d}x)$. The $1-1$ relationship between $(W|_{[0,\tau]}, h)$ and the real parameters of interest $(F|_{[0,\tau)}, \mu)$ is very well-behaved. Moreover, since we have observations directly from $F^{\mathrm{u.c.}}$ we can find gradients for $W(x)$ which are uniformly bounded in x, and the *canonical gradients* must have the same property. The set of optimal influence functions for W, indexed now by V and x, turns out to be a Donsker class and we can prove weak convergence of $n^{\frac{1}{2}}(\widehat{W} - W)$ in $D[0,\tau]$ jointly with $n^{\frac{1}{2}}(\widehat{h} - h)$. This gives weak convergence and asymptotic optimality for $(\widehat{F}|_{[0,\sigma]}, \widehat{\mu})$ for each $\sigma < \tau$.

It seems from the above that it is not possible to estimate $V(\tau)$ or $F(\tau)$ at root n rate. We can prove this, for $V(\tau)$, by appeal to the earlier mentioned criterion of van der Vaart (1991b): is $1_{[0,\tau]} - V(\tau)$ in the range of A^* (the conditional expectation mapping from mean zero functions of $(\widetilde{X}, \Delta)$ to those of X)? We suppose V is continuous and even has a positive density at $x = \tau$. We can discard the constant, and must discover if there exists a square integrable $g(\widetilde{X}, \Delta)$ such that $1_{[0,\tau]} = \mathrm{E}(g(\widetilde{X}, \Delta) \mid X)$. The cases $x \le \tau$ and $x > \tau$ give us two equations:

$$\begin{aligned} \frac{\tau - x}{\tau + x} g^{\text{u.c.}}(x) + \frac{2}{\tau + x} \int_0^x g^{\text{s.c.}}(t) \mathrm{d}t = 1, \quad x \le \tau, \\ \frac{2}{\tau + x} \int_0^\tau g^{\text{s.c.}}(t) \mathrm{d}t + \frac{x - \tau}{\tau + x} g^{\text{d.c.}}(\tau) = 0, \quad x > \tau. \end{aligned}$$

The second equation implies that $g^{\text{d.c.}}(\tau) = 0$ and that $\int_0^\tau g^{\text{s.c.}}(t) \mathrm{d}t = 0$. So the first becomes

$$\frac{\tau - x}{\tau + x} g^{\text{u.c.}}(x) - \frac{2}{\tau + x} \int_x^\tau g^{\text{s.c.}}(t) \mathrm{d}t = 1, \quad x \le \tau,$$

which we can rewrite again as

$$g^{\text{u.c.}}(x) - \frac{2}{\tau - x} \int_x^\tau g^{\text{s.c.}}(t) \mathrm{d}t = \frac{\tau - x}{\tau - x}, \quad x \le \tau. \tag{28}$$

Here $g^{\text{u.c.}} \in L^2(F^{\text{u.c.}})$ while $g^{\text{s.c.}} \in L^2(F^{\text{s.c.}})$. If V has a density bounded away from zero then both functions are members of L^2(Lebesgue). However the right hand side of (28), the function $(\tau + x)/(\tau - x))$, is *not* square integrable on $[0, \tau]$.

Now a well-known result of Hardy (see, e.g., Ritov and Wellner, 1988) says that if

$$\widetilde{g}(x) = \frac{1}{x} \int_0^x g(t) \mathrm{d}t,$$

where g is an L^2(Lebesgue) function on the unit interval then $\|\widetilde{g}\| \le 2\|g\|$. This means that the second term on the left-hand side of (28) is also square integrable, a contradiction.

The argument can be sharpened to show that if V just has a positive density at τ, then $1_{[0,\tau]}$ is not in the range of A^*. So $V(\tau)$ cannot be root n rate regularly estimated. By consideration of the transformation from V to F it follows easily that the same applies to $F(\tau)$.

For completeness we conclude by giving the inverse of the information operator, derived in van der Laan (1993b): define first the operator $B : D[0, \tau] \to D[0, \tau]$ by

$$(Bh)(x) = \frac{2}{\tau - x} \int_x^\tau \frac{\int_y^\tau \frac{h(u)}{\tau + u} V(\mathrm{d}u)}{g(y)} \mathrm{d}y$$

and define a function α_1 and a number α_2, the latter depending on h, by

$$\begin{aligned}\alpha_1(x) &= \frac{2}{\tau - x}\frac{\int_x^\tau \frac{\mathrm{d}y}{g(y)}}{\int_0^\tau \frac{\mathrm{d}y}{g(y)}}\\ \alpha_2(h) &= \int_0^\tau \Big(\int_0^u \frac{\mathrm{d}y}{g(y)}\Big)\frac{h(u)}{\tau+u}V(\mathrm{d}u).\end{aligned}$$

Then the inverse mapping (for $\dot\kappa$ with support in $[0,\tau)$) is given on $[0,\tau)$ by

$$\begin{aligned} h = (A^*A)^{-1}(\dot\kappa) &= \phi_1 - \alpha_3\phi_2,\\ \phi_1 &= \sum_{i=0}^\infty B^i\Big(\frac{\tau+\cdot}{\tau-\cdot}\dot\kappa\Big),\\ \phi_2 &= \sum_{i=0}^\infty B^i\alpha_1,\\ \alpha_3 &= \frac{\alpha_2(\phi_1)}{1+\alpha_2(\phi_2)};\end{aligned} \tag{29}$$

on $[\tau,\infty)$ the inverse h is only determined as far as the values of the following two integrals:

$$\begin{aligned}\int_\tau^\infty \frac{h(x)}{\tau+x}V(\mathrm{d}x) &= -\frac{\int_0^\tau \frac{\int_y^\tau \frac{h(u)}{\tau+u}V(\mathrm{d}u)}{g(y)}\mathrm{d}y}{\int_0^\tau \frac{\mathrm{d}y}{g(y)}},\\ \int_\tau^\infty \frac{(x-\tau)h(x)}{\tau+x}V(\mathrm{d}x) &= 0.\end{aligned}$$

Concluding remarks.

It remains to discuss extensions and limitations of the above theory. Van der Laan's (1993a) identity is an extremely powerful tool for studying the NPMLE in linear-convex models, and many other hitherto rather difficult models can be succesfully analysed with it. When we move from the one-dimensional to the two-dimensional line-segment problem we find however that, unless K is known, we no longer have this special structure. However for a given orientation distribution K the identity is applicable for a suitable length-biased version of F, and the whole analysis of the NPMLE of F for given K should be very similar to the one-dimensional case. When K varies also, the fact that the NPMLE for F only depends on K via a single integral (the mean window diameter) while the NPMLE of K for given F is even easier to study, gives hope that one can finally make a complete analysis of the original problem using an ad hoc combination of the 'F known' results for K and vice-versa.

We assumed above a convex window, independent lengths and orientations, and a homogenous Poisson process. One can expect that the NPMLE's derived here are still reasonable estimators in other situations. Suppose the window is not convex. Each transect of the window results in data from several intervals of a single line-segment process. The joint distribution of all this data is very complex, but the marginal distribution for each interval is of the same type as above. Hence if we discard the extra information coming from single line-segments appearing or not appearing in disjoint intervals but consider the data from each interval as separate observations, the same relations hold between the means of the empirical distributions of the data $F^{*\cdot \mathrm{c}\cdot}$ and the underlying parameters as in the convex case. This is enough to give *consistency* of the NPMLE computed as if we had separate observations. The estimator will still converge at root n rate but its asymptotic variance will be of different form (since there are now many small groups of *dependent* observations). The bootstrap will probably work.

Similarly if the process is not a Poisson line-segment process but just a stationary line-segment process, one introduces the so-called Palm distribution of a typical line-segment (length, orientation) pair. We will have to assume independence in this bivariate distribution. We still will have the same relations between mean values and again at least consistency of the NPMLE computed as if the Poisson assumption holds.

If lengths and orientations are not independent, one could consider estimation of an arbitrary joint distribution. This seems a very difficult task. A sensible approach would be to partition the orientations into a small number of classes and then estimate a fixed length distribution for each class. Now the estimator above can be used for each class separately. Moreover, just to investigate whether or not lengths and orientations are independent one can compare estimators of length distributions for different classes of orientations. Bootstrap tests could be constructed.

Finally, as we said at the beginning of the section, many practical applications are really dealing with a two-dimensional section of a three-dimensional process of planar objects, producing line-segments in section. However from a model for three-dimensional planar objects one can make predictions about the two-dimensional sections. Our approach allows one to separate the edge-effects and the stereological aspects: compare the NPMLE for the length distribution with that predicted by a given model.

14. Kaplan-Meier for a spatial point process.

This section, based on Baddeley and Gill (1992), again considers a problem from spatial statistics. The problem is concerned with estimation of distance distributions when a spatial process is observed inside a finite window W. The boundary of the window prevents us completely observing all distances and there seems to be an analogy with censored data. In this case, the analogy turns out to be a useful one. Surprisingly the analogy between edge effects for point processes and censoring of survival times did not seem to have been noticed before.

We start by giving some of the background to our problem. The exploratory data analysis of observations of a spatial point process Φ often starts with the estimation of certain distance distributions: F, the distribution of the distance from an arbitrary point in space to the nearest point of the process; G, the distribution of the distance from a typical point of the process to the nearest other point of the process; and $K(r)$, the expected number of other points within distance r of a typical point of the process, divided by the intensity α. Equivalently K is proportional to the sum over all $n = 1, 2, \ldots$ of the distribution of the distance from a typical point of the process to the nth nearest point. Popular names for F, G and K are the empty space function, the nearest neighbour distance distribution, and the second moment function. For a homogeneous

Poisson process, F, G and K take known functional forms, and deviations of estimates of F, G, K from these forms are taken as indications of 'clustered' or 'inhibited' alternatives; see Diggle (1983), Ripley (1981, 1988).

However, estimation of F, G and K is hampered by edge effects when the point process is only observed within a bounded window W. Essentially the distance from a reference point x in W to the nearest point of the process is *censored* by its distance to the boundary of W. Edge effects become rapidly more severe as the dimension of space increases, or as the distance r of interest increases.

Traditionally, in spatial statistics one uses edge-corrected estimators which are weighted empirical distributions of the observed distances. The simplest approach is the 'border method' (Ripley, 1988) in which we restrict attention (when estimating F, G or K at distance r) to those reference points lying more than r units away from the boundary of W. These are the points x from which distances up to r can be observed without censoring. However, the border method throws away an appreciable number of points; in three dimensions it seems to be unacceptably wasteful, especially when estimating G. For instance, Baddeley, Moyeed, Howard, Reid and Boyde (1993) gave a case-study in which the spatial distribution of *lacunae* in the bone of the skull of a species of monkey was studied. One might like to consider the data as forty separate realisations of a stationary point process in $\mathbb{R}^3$ observed through a rather small window relative to the intensity of points; or alternatively as one realisation observed through a window consisting of forty sub-windows far apart from one another. One of these forty pieces of data is shown in figure 1. If the window is the unit cube and one considers the distance $r = 0.2$, then the border method requires one to discard almost 80% of the reference points.

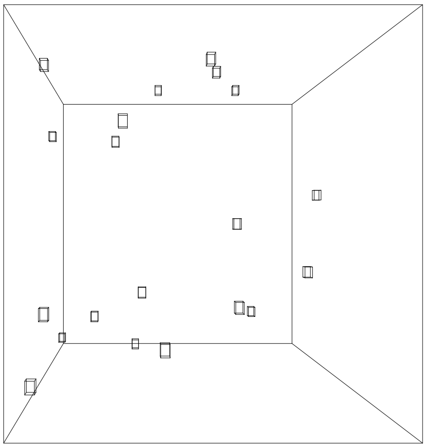

Figure 1. Spatial distribution of lacunae in skull bone, one of forty replicates.

In more sophisticated edge corrections (for estimating K), the weight $w(x, y)$ attached to the observed distance $\|x - y\|$ between two points x, y is the reciprocal of the probability that this distance will be observed under certain invariance assumptions (stationarity under translation, rotation, or both). Corrections of this type were first suggested by Miles (1974) and developed by Ripley, Lantuéjoul, Hanisch, Ohser and others; see Stoyan, Kendall and Mecke (1987), Ripley (1988), Baddeley et al. (1991) and Barendregt and Rottschäfer (1991) for recent surveys; see also Stein (1990), Doguwa and Upton (1990) and Doguwa (1990) for evidence that the last word still has not been said on the topic.

Now the estimation problem for F, G and K when observing a point process Φ through a bounded window W has some similarity with the estimation of a survival function based on a sample of randomly censored survival times. Closely following Baddeley and Gill (1992), we develop the analogy and propose Kaplan-Meier or product-limit estimators for F, G and K. Since the observed, censored distances are highly interdependent, the standard theory developed in previous sections has little to say about the statistical properties of the new estimators. In particular, classical optimality results on the Kaplan-Meier estimator with independent observations are not applicable. One may however hope that the new estimators are still better than the classical edge corrections. In fact the border method for edge correction, described above, is analogous to the so-called reduced sample estimator (discussed in Kaplan and Meier, 1958), a very inefficient competitor to the Kaplan-Meier estimator obtained by using only those observations for which the censoring time is at least t when estimating the probability of survival to time t.

The estimation of F by a Kaplan-Meier type estimator poses another new problem, since one has a *continuum* of observations: for each point in the sampling window, a censored distance to the nearest point of the process. This problem is however nicely solved using product-integration.

Together with estimates of F, G and K one would like to evaluate their accuracy. Though the estimators are based on dependent observations one may still hope that in many situations a linear approximation is possible (the delta method, section 6), leading to several proposals for variance estimators. It also leads to an evaluation of asymptotic efficiency in some simple, theoretical situations.

The next subsection recalls some definitions from spatial statistics; then we introduce our Kaplan-Meier style estimator of the empty space function F; we next discuss asymptotic properties of this estimator; and finally briefly treat the other functions G and K.

Spatial statistics.

Let Φ be a point process in $\mathbb{R}^d$, observed through a window $W \subseteq \mathbb{R}^d$. We assume W is compact and topologically regular (it is the closure of its interior), and denote its boundary by ∂W.

We may consider Φ both as a random set in $\mathbb{R}^d$ and as a random measure. The problem is, based on the data $\Phi \cap W$ (and knowledge of W itself) to estimate the functions F, G and K defined as follows.

For $x \in \mathbb{R}^d$, $A \subseteq \mathbb{R}^d$ let

$$\rho(x, A) = \inf\{\|x - a\| : a \in A\}$$

be the shortest (Euclidean) distance from x to A. Define

$$\begin{aligned} A_{\oplus r} &= \{x \in \mathbb{R}^d : \rho(x, A) \leq r\}, \\ A_{\ominus r} &= \{x \in A : \rho(x, A^c) > r\}, \end{aligned}$$

where ^{c} denotes complement. For A closed, these are respectively the dilation and erosion of A by a ball of radius r:

$$\begin{aligned} A_{\oplus r} &= \bigcup_{x \in A} B(x, r) \\ A_{\ominus r} &= \left((A^c)_{\oplus r}\right)^c \end{aligned}$$

where $B(x, r)$ is the closed ball of radius r, centre x in $\mathbb{R}^d$.

Assume now that Φ is stationary under translations and has finite positive intensity α. Thus for any bounded Borel set $A \subseteq \mathbb{R}^d$

$$\mathrm{E}\Phi(A) = \alpha |A|_d$$

where $|\cdot|_d$ denotes d-dimensional Lebesgue volume. For $r \geq 0$ define

$$\begin{aligned} F(r) &= \mathrm{P}\{\rho(0, \Phi) \leq r\} \\ &= \mathrm{P}\{\Phi(B(0, r)) > 0\}, \end{aligned}$$

$$\begin{aligned} G(r) &= \mathrm{P}\{\rho(0, \Phi \setminus \{0\}) \leq r \mid 0 \in \Phi\} \\ &= \mathrm{P}\{\Phi(B(0, r)) > 1 \mid 0 \in \Phi\}, \\ K(r) &= \alpha^{-1}\mathrm{E}\{\Phi(B(0, r) \setminus \{0\}) \mid 0 \in \Phi\}. \end{aligned}$$

By stationarity the point 0 in these expressions may be replaced by any arbitrary point x. The conditional expectations given $0 \in \Phi$, used in defining G and K above, are expectations with respect to the Palm distribution of Φ at 0. Alternative definitions using the Campbell-Mecke formula (see Stoyan, Kendall and Mecke, 1987) are

$$\begin{aligned} G(r) &= \frac{\mathrm{E}\Big(\sum_{x \in \Phi \cap A} 1\{\rho(x, \Phi \setminus \{x\}) \leq r\}\Big)}{\mathrm{E}\Phi(A)}, \\ K(r) &= \frac{\mathrm{E}\Big(\sum_{x \in \Phi \cap A} \Phi(B(x, r) \setminus \{x\})\Big)}{\mathrm{E}\Phi(A)}, \end{aligned}$$

holding for arbitrary measurable sets A with $0 < |A|_d < \infty$.

A Kaplan-Meier estimator for the empty space function.

Every reference point x in the window W contributes one possibly censored observation of the distance from an arbitrary point in space to the point process Φ; recall that $F(r) = \mathrm{P}\{\rho(x, \Phi) \leq r\}$. The analogy with survival times is to regard $\rho(x, \Phi)$ as the 'distance (time) to failure' and $\rho(x, \partial W)$ as the censoring distance. The observation is censored if $\rho(x, \partial W) < \rho(x, \Phi)$.

From the data $\Phi \cap W$ we can compute $\rho(x, \Phi \cap W)$ and $\rho(x, \partial W)$ for each $x \in W$. Note that

$$\rho(x, \Phi) \wedge \rho(x, \partial W) = \rho(x, \Phi \cap W) \wedge \rho(x, \partial W)$$

so that we can indeed observe $\rho(x, \Phi) \wedge \rho(x, \partial W)$ and $1\{\rho(x, \Phi) \leq \rho(x, \partial W)\}$ for each $x \in W$. Then the set

$$\{x \in W : \rho(x, \Phi) \wedge \rho(x, \partial W) \geq r\}$$

can be thought of as the set of points 'at risk of failure at distance r', and

$$\{x \in W : \rho(x, \Phi) = r, \ \rho(x, \Phi) \leq \rho(x, \partial W)\}$$

are the 'observed failures at distance r'. These two sets are analogous to the points counted in the empirical functions $Y(s)$, $N(\mathrm{d}s)$ respectively in the definition of the Kaplan-Meier estimator.

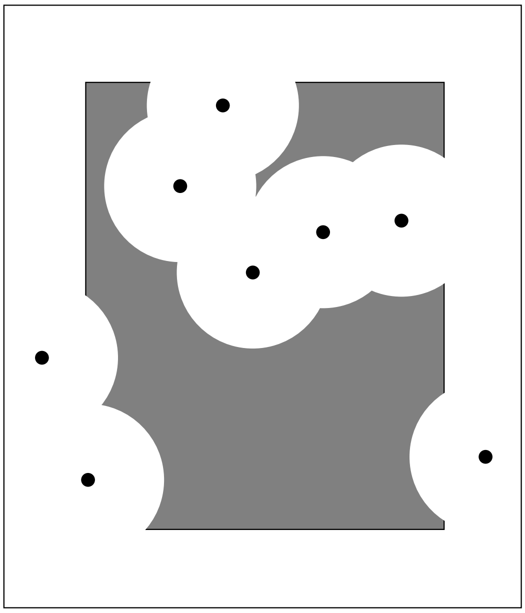

Figure 2. Geometry of the Kaplan-Meier estimator. Spatial process Φ indicated by filled dots. Points x at risk are shaded. Observed failures constitute the curved boundary of the shaded region.

Geometrically the two sets can be written as

$$W_{\ominus r} \setminus \Phi_{\oplus r}, \qquad \partial\left(\Phi_{\oplus r}\right) \cap W_{\ominus r};$$

that is, *within the eroded window* $W_{\ominus r}$, consider the region outside the union of balls of radius r centred at points of the process, and the surface of this union of balls, see Figure 2.

Definition. *Let Φ be a stationary point process and $W \subseteq \mathbb{R}^d$ a regular compact set. The Kaplan-Meier estimator $\widehat{F}$ of the empty space function F of Φ, based on data $\Phi \cap W$,* *is defined via the corresponding Nelson-Aalen esimator by*

$$\widehat{\Lambda}(r) = \int_0^r \frac{|\partial\left(\Phi_{\oplus s}\right) \cap W_{\ominus s}|_{d-1}}{|W_{\ominus s} \setminus \Phi_{\oplus s}|_d} \mathrm{d}s$$

$$1 - \widehat{F}(r) = \prod_0^r \left(1 - \widehat{\Lambda}(\mathrm{d}s)\right) = \exp(-\widehat{\Lambda}(r))$$

where $|\cdot|_{d-1}$ *denotes* $d-1$ *dimensional surface area (Hausdorff) measure.*

The reduced sample estimator (the standard border correction method) $\widehat{F}_{\mathrm{RS}}$ of F is given by

$$1 - \widehat{F}_{\mathrm{RS}}(r) = \frac{|W_{\ominus r} \setminus \Phi_{\oplus r}|_d}{|W_{\ominus r}|_d}$$

The Kaplan-Meier estimator $\widehat{F}$ is based on the continuum of observations generated by all $x \in W$. It is a proper distribution function and is even absolutely continuous, with hazard rate

$$\widehat{\lambda}(r) = \frac{|\partial\Phi_{\oplus r} \cap W_{\ominus r}|_{d-1}}{|W_{\ominus r} \setminus \Phi_{\oplus r}|_d}. \tag{1}$$

Unbiasedness and continuity.

Our first theorem will be a ‘ratio unbiasedness’ result for the hazard rate estimator $\widehat{\lambda}$.

Theorem 1. *The empty space function F is absolutely continuous with hazard rate*

$$\lambda(r) = \frac{\mathrm{E}|W \cap \partial\Phi_{\oplus r}|_{d-1}}{\mathrm{E}|W \setminus \Phi_{\oplus r}|_d}$$

for any compact regular window W*. In particular, replacing* W *by* $W_{\ominus r}$*, our estimator* $\widehat{\lambda}$ *is ‘ratio unbiased’ in the sense that the ratio of expectations of the numerator and denominator in* (1) *is equal to the true hazard rate* $\lambda(r)$ *(as long as the denominator has positive probability of being nonzero).*

Thus $\widehat{F}(r)$ respects the smoothness of the true empty space function F. The reduced-sample estimator is not even necessarily monotone.

The theorem is proved via two regularity lemmas. The first is an example of Crofton's perturbation or ‘moving manifold’ formula, see Baddeley (1977), Crofton (1869). In our case it says that the volume, within a fixed region, of a union of (possibly overlapping) balls of radius r can be determined by imagining the balls as growing at constant rate with radius s varying from 0 up to r; the finally achieved total volume equals the integral of the surface area of the intermediate objects: take $A = \Phi \cap W$, $Z = W$.

Lemma 1. *Let* $Z \subseteq \mathbb{R}^d$ *be a compact regular set and* $A \subseteq \mathbb{R}^d$ *any nonempty closed set. Then for* $r \geq 0$

$$|Z \cap A_{\oplus r}|_d = |Z \cap A|_d + \int_0^r |Z \cap \partial A_{\oplus s}|_{d-1} \mathrm{d}s.$$

The lemma is proved in Baddeley and Gill (1992) by applying the so-called co-area formula of geometric measure theory, see Federer (1969, p. 251). (It is also shown there that the integrand in the formula is measurable).

The second lemma states that the integrand $|Z \cap \partial\Phi_{\oplus s}|_{d-1}$ is uniformly bounded (over possible realisations of Φ) in such a way that dominated convergence can be used to justify interchanges of expectation and integration or differentiation (w.r.t. s).

Lemma 2 (boundedness)**.** *For any regular compact set Z*

$$|Z \cap \partial\Phi_{\oplus r}|_{d-1} \leq \frac{d}{r}|Z_{\oplus r}|_d \;\wedge\; \Phi(Z_{\oplus r})\omega_d r^{d-1}$$

where $\omega_d = |\partial B(0,1)|_{d-1} = 2\pi^{d/2}/\Gamma(d/2)$.

A formal proof of Lemma 2 is given in Baddeley and Gill (1992). Informally, note that the second term on the right is a trivial bound on the left hand side, since $\omega_d r^{d-1} = |\partial B(0,r)|_{d-1}$. For the first term, fix a realization of Φ and let y_i, $i = 1, \ldots, m$ be the distinct points of $\Phi \cap Z_{\oplus r}$. The surface whose area is taken on the left hand side is the surface of the union of (possibly overlapping) balls radius r and centres $y_1, \ldots, y_m$, intersected with Z. Note that the factor d/r equals the ratio of surface area to volume of the d-dimensional ball $B(0,r)$. Consider the segment of this ball subtended by some given subset of its surface: that is, the union of all line-segments joining a point in the given part of the surface to the centre of the ball. Again, the ratio of 'outside' surface area to volume of the *segment* is d/r. Now the surface in question, $Z \cap \partial\Phi_{\oplus r}$, can be split into m disjoint pieces, each of which is the outer surface of a (disjoint) segment of one the m balls. The total area equals d/r times the volume of the union of the segments. But the union of the segments is contained in the dilated window $Z_{\oplus r}$, so the volume of this supplies an upper bound.

Let Φ be a point process in $\mathbb{R}^d$ and $W \subset \mathbb{R}^d$ a regular compact set. The following identities follow from Lemma 1:

$$|W \cap \Phi_{\oplus r}|_d = \int_0^r |W \cap \partial\Phi_{\oplus s}|_{d-1}\,\mathrm{d}s, \tag{2}$$

$$|\{x \in W : \rho(x,\Phi) \leq \rho(x,\partial W),\ \rho(x,\Phi) \leq r\}|_d = \int_0^r |W_{\ominus s} \cap \partial\Phi_{\oplus s}|_{d-1}\,\mathrm{d}s, \tag{3}$$

$$|W_{\ominus r} \setminus \Phi_{\oplus r}|_d = |W|_d - \int_0^r |\partial(W_{\ominus s} \setminus \Phi_{\oplus s})|_{d-1}\,\mathrm{d}s. \tag{4}$$

Moreover (by standard measurability arguments from stochastic geometry) the integrands are well defined random variables for each fixed s and are almost surely measurable and integrable functions of s.

We can now prove Theorem 1. By Fubini,

$$\begin{aligned}\mathrm{E}|W \cap \Phi_{\oplus r}|_d &= \mathrm{E}\int_W 1\{x \in \Phi_{\oplus r}\}\,\mathrm{d}x \\ &= \int_W \mathrm{P}\{x \in \Phi_{\oplus r}\}\,\mathrm{d}x\end{aligned}$$

$$= F(r)\,|W|_d. \tag{5}$$

Since $|W \cap \Phi_{\oplus r}|_d$ is absolutely continuous, with derivative given in Lemma 1 and bounded as in Lemma 2, its expectation is absolutely continuous too, with derivative

$$f(r)|W|_d = \mathrm{E}|W \cap \partial\Phi_{\oplus r}|_{d-1}. \tag{6}$$

But complementarily to (5)

$$\mathrm{E}|W \setminus \Phi_{\oplus r}|_d = (1 - F(r))\,|W|_d. \tag{7}$$

Dividing (6) by (7) we obtain the first result of the theorem. The rest follows by replacing W with $W_{\ominus r}$.

Discretisation and the classical Kaplan-Meier estimator.

In practice one would not actually compute the surface areas and volumes for each $s \in [0, r]$ in order to estimate $F(r)$. Rather one would discretize W or $[0, r]$ or both.

A natural possibility is to discretize W by superimposing a regular lattice L of points, calculating for each $x_i \in W \cap L$ the censored distance $\rho(x_i, \Phi) \wedge \rho(x_i, \partial W)$ and the indicator $1\{\rho(x_i, \Phi) \leq \rho(x_i, \partial W)\}$. Then one would calculate the ordinary Kaplan-Meier estimator based on this finite dataset.

Our next result is that as the lattice becomes finer, the discrete Kaplan-Meier estimator converges to the 'theoretical' continuous estimator $\widehat{F}$.

Theorem 2. *Let $\widehat{F}_L$ be the Kaplan-Meier estimator computed from the discrete observations at the points of $W \cap L$, where $L = \varepsilon M + b$ is a rescaled, translated copy of a fixed regular lattice M. Let*

$$R = \inf\{r \geq 0 : W_{\ominus r} \cap \Phi_{\oplus r} = \emptyset\}.$$

Then as the lattice mesh ε converges to zero, $\widehat{F}_L(r) \to \widehat{F}(r)$ for any $r < R$. The convergence is uniform on any compact subinterval of $[0, R)$.

Proof. For any regular compact set $A \subseteq \mathbb{R}^d$ one has

$$\varepsilon^d \#(L \cap A) \to c|A|_d \quad \text{as } \varepsilon \to 0$$

where c is a finite positive constant. Hence the functions

$$N_L(r) = \frac{\#\,(L \cap \{x \in W : \rho(x, \Phi) \leq \rho(x, \partial W),\ \rho(x, \Phi) \leq r\})}{\#(L \cap W)}|W|_d,$$

$$Y_L(r) = \frac{\#\,(L \cap (W_{\ominus r} \setminus \Phi_{\oplus r}))}{\#(L \cap W)}|W|_d$$

converge pointwise to

$$N(r) = |\{x \in W : \rho(x, \Phi) \leq \rho(x, \partial W),\ \rho(x, \Phi) \leq r\}|_d, \tag{8}$$

$$Y(r) = |W_{\ominus r} \setminus \Phi_{\oplus r}|_d \tag{9}$$

respectively. Since $N_L(r)$ is increasing in r and the limit is continuous, $N_L \to N$ uniformly in r. Similarly, Y_L is decreasing and by (4) its limit is continuous, so it also converges uniformly.

Given (3) and by continuity of the mapping from (N,Y) to $\widehat{\Lambda} = \int \mathrm{d}N/Y$ (see sections 4 and 6) the discrete Nelson-Aalen estimator

$$\widehat{\Lambda}_L = \int \frac{\mathrm{d}N_L}{Y_L}$$

converges uniformly to $\widehat{\Lambda}$ on a closed interval where Y is strictly positive. By continuity of the product-integral mapping (section 4) $\widehat{F}_L$ converges to $\widehat{F}$. □

Further remarks on computational aspects can be found in Baddeley and Gill (1992). That paper also contains simulation results pointing to a rather satisfactory behaviour of the Kaplan-Meier estimator compared to the reduced sample estimator, though it is certainly not better in all situations.

Asymptotic properties.

A relevant 'large sample' situation is one in which the edge problem remains equally severe as in the 'small sample' case. So one would like to consider observation of the same point process through a sequence of increasingly large windows W, in such a way that (e.g.) the proportion of the window within distance r from the boundary stays appreciable. The simplest such situation is when the window W is the union of n small and distantly spread windows of fixed size and shape, so that to a good approximation one simply has n independent replicates of the situation considered in the previous section. Asymptotics as $n \to \infty$ are now easy to derive from the functional delta-method, taking as starting point a law of large numbers and a (joint) central theorem for a sum of i.i.d. replicates of the 'number of failures' and the 'number at risk' processes N and Y defined by (8) and (9) above. If the distance τ satisfies $\mathrm{E}Y(\tau) > 0$, the facts that N and Y are monotone and bounded by $|W|_d$ give the uniform LLN and CLT on $[0,\tau]$ without further restrictions (for the CLT, use the nice result of E. Giné and J. Zinn that the central limit theorem holds for i.i.d. sums of a uniformly bounded process Z satisfying $\mathrm{E}|Z(s) - Z(t)| \le c|s-t|$; see van der Vaart and Wellner, 1993). Hence $\widehat{F}$ is consistent and asymptotically normal.

We even have a bootstrap result from the Giné-Zinn equivalence theorem mentioned in Section 11 (though a jack-knife theorem would probably be more useful in practice). In practice one may well have a number of replicates but typically the number n will be small (say 5 to 10) and the windows not all of the same shape and size. Consequently the formal asymptotics cannot be expected to be very useful. We therefore only sketch them, indicating how they can be used to suggest rough variance estimators for practical use, and how theoretical efficiency calculations can be done in simple and stylized situations.

Even if we do not have i.i.d. replicates, it may still be reasonable to assume a law of large numbers and a central limit theorem for the suitably normalized processes N and Y, based now on all the data. The functional delta-method together with differentiability of the product-integration mapping tell us that if the fluctuations of the random functions

$$\frac{|W_{\ominus r} \setminus \Phi_{\oplus r}|_d}{|W|_d} \quad , \quad \frac{\int_0^r |W_{\ominus s} \cap \partial \Phi_{\oplus s}|_{d-1} \mathrm{d}s}{|W|_d}; \quad 0 \le r \le \tau$$

about their expectations are uniformly small and not too violent (in the sense that a functional central limit theorem holds as W gets larger in some way), then one may approximate $\widehat{F}(r) - F(r)$ well for $0 \le r \le \tau$ by the linear expression

$$(1 - F(r)) \int_0^r \frac{(|W_{\ominus s} \cap \partial\Phi_{\oplus s}|_{d-1} - |W_{\ominus s} \setminus \Phi_{\oplus s}|_d \lambda(s))}{y(s)} \mathrm{d}s \tag{10}$$

where $y(s) = \mathrm{E}|W_{\ominus s} \setminus \Phi_{\oplus s}|_d = (1 - F(s))|W_{\ominus s}|_d$.

If W is a union of small, distant sub-windows W_i then (10) is also a sum over the W_i of mean-zero terms, given by replacing W by W_i in (10) except in the definition of the function y. The variance of $\widehat{F}(r)$ could therefore be approximated by the sum of the squares of the summands in (10), in which one would have to replace $\lambda(\cdot)$ and F by their Kaplan-Meier estimates. This is similar to a jackknife or bootstrap analysis (which one could use if the W_i were of the *same* size and shape).

The computational problems involved in this procedure can be eased by the same sampling procedure as was used to approximate $\widehat{F}$ itself: choose points on a regular lattice intersected with W_i, or many independent random points uniformly distributed over W_i, and average the 'influence function' for one point x:

$$(1 - F(r)) \left(\frac{1\{\rho(x,\Phi) \le r,\ \rho(x,\Phi) \le \rho(x,\partial W)\}}{y(\rho(x,\Phi))} - \int_0^{r \wedge \rho(x,\Phi) \wedge \rho(x,\partial W)} \frac{\lambda(s)}{y(s)} \mathrm{d}s \right). \tag{11}$$

Expression (10) is exactly the integral over $x \in W$ of (11), with respect to Lebesgue measure, as can be seen by recognising $|\cdot|_d$ and $|\cdot|_{d-1}\mathrm{d}s$ in (10) as integrals over x and then interchanging orders of integration. In order to implement the proposal one only has to numerically tabulate an estimate of the function $\int_0^r (\lambda(s)/y(s))\mathrm{d}s$ together with the functions y and $1 - F$. After (11) has been calculated for points sampled from each subwindow W_i, one must average, square, and add over subwindows.

Alternatively one can write down the variance of the linear approximation (10), or rather, the integral over $x \in W$ of (11), in terms of the covariance structures of the random function $r(x) = \rho(x, \Phi)$ and of the window W. First of all we rewrite (10) as

$$-(1 - F(r)) \int_{x \in W} \int_{s \in (0,r]} \left(\frac{\mathrm{d}^{(s)} 1\{x \notin \Phi_{\oplus s}\} + 1\{x \notin \Phi_{\oplus s}\}\lambda(s)\mathrm{d}s}{y(s)} \right) 1\{\rho(x,\partial W) \ge s\} \mathrm{d}x.$$

After some further calculation one then arrives at

$$\begin{aligned} \mathrm{cov}(\widehat{F}(r), \widehat{F}(r')) \approx (1 - F(r))\,(1 - F(r'))\ \cdot \\ \cdot \int\limits_{x \in \mathbf{R}^d} \int\limits_{s=0}^{r} \int\limits_{s'=0}^{r'} g(\mathrm{d}s, \mathrm{d}s', x) C(W_{\ominus s}, W_{\ominus s'})(-x) \mathrm{d}x. \end{aligned} \tag{12}$$

Here, for $A, B \subseteq \mathbb{R}^d$ and $x \in \mathbb{R}^d$, $C(A,B)(x)$ is the set cross-covariance function

$$C(A,B)(x) = |A \cap (B \oplus x)|_d,$$

$B \oplus x$ being the translate of B by x, while g is given by

$$\begin{aligned} g(\mathrm{d}s, \mathrm{d}s', x) = \frac{1}{y(s)y(s')} &(\sigma(\mathrm{d}s, \mathrm{d}s')(x) + \sigma(\mathrm{d}s, s')(x)\lambda(s')\mathrm{d}s' \\ &+ \sigma(s, \mathrm{d}s')(x)\lambda(s)\mathrm{d}s + \sigma(s,s')(x)\lambda(s)\lambda(s')\mathrm{d}s\mathrm{d}s') \end{aligned} \tag{13}$$

with

$$\begin{aligned} \sigma(s,s')(x) &= \mathrm{P}\{\Phi_{\oplus s} \not\ni 0,\ \Phi_{\oplus s'} \not\ni x\} \\ &= \mathrm{P}\{y \notin \Phi_{\oplus s},\ x + y \notin \Phi_{\oplus s'}\} \\ &= \mathrm{P}\{\rho(y, \Phi) > s,\ \rho(x+y, \Phi) > s'\} \\ &= \mathrm{P}\{\Phi(B(y,s)) = 0,\ \Phi(B(x+y, s') = 0\} \end{aligned}$$

for arbitrary $y \in \mathbb{R}^d$.

One could try to estimate σ and plug the estimate into (12) using estimates of $y(s) = (1 - F(s))|W_{\ominus s}|_d$ and $\lambda(\cdot)$ also. Note that σ is actually a bivariate survival function so one could in principle use a Dabrowska-type estimator (see section 12) or just a bivariate reduced sample estimator for this purpose. However the amount of computation needed is very daunting, and the final result may be so statistically inaccurate as to be quite useless. Practical experience is badly needed here.

Finally, (10)–(12) are the starting point of a theoretical efficiency calculation, which we perform below.

The sparse Poisson limit.

Here we consider asymptotic variances of the Kaplan-Meier and reduced sample *influence functions* on a fixed window W for a Poisson process whose intensity α is sent to zero. This is the asymptotic variance of the Kaplan-Meier and reduced sample *estimators* in the large-sample case when the data consists of many independent replicates of a fixed-intensity Poisson process observed through an asymptotically small window. 'Many replicates' justifies looking at the influence function, and the case of a vanishing intensity but fixed window is the same as a vanishing window, fixed intensity. In fact if either intensity or window is small, any stationary process looks like a Poisson process.

There are just two situations to consider: (i) no random point in W, with probability $e^{-\alpha|W|_d} = 1 + \mathcal{O}(\alpha)$, and (ii), one random point in W at a position X uniformly distributed over W, occurring with probability $\alpha|W|_d e^{-\alpha|W|_d} = \alpha|W|_d + \mathcal{O}\left(\alpha^2\right)$; the remaining possibilities have probability $\mathcal{O}\left(\alpha^2\right)$.

The influence function (10) for Kaplan-Meier is the difference of two terms: a part depending on surface areas at some distances from a point of Φ, and a part depending on volumes at risk, and involving the hazard rate of the empty space function. In case (i) only the second part is present and is of order α; in case (ii) the first part is also present and is of constant order.

The empty space function for the Poisson process is

$$F(r) = 1 - \exp\bigl(-\alpha|B_r|_d\bigr)$$

and its hazard rate is

$$\lambda(r) = \frac{\mathrm{d}}{\mathrm{d}r}\left(-\log(1 - F(r))\right) = \alpha|\partial B_r|_{d-1}$$

where $B_r = B(0, r)$ is a ball of radius r, so that $|B_r|_d = r^d\omega_d/d$ and $|\partial B_r|_{d-1} = r^{d-1}\omega_d$. The 'expected number at risk' is

$$y(r) = (1 - F(r))\,|W_{\ominus r}|_d.$$

In case (i), no random points in W, the influence function (10) for Kaplan-Meier is therefore

$$\begin{aligned}
(1 - F(r)&\left\{-\int_0^r \frac{\alpha|\partial B_s|_{d-1}|W_{\ominus s}|_d}{|W_{\ominus s}|_d e^{-\alpha|B_s|_d}}\mathrm{d}s\right\}\\
&= (1 - F(r))\left\{-\int_0^r \alpha|\partial B_s|_{d-1}e^{\alpha|B_s|_d}\mathrm{d}s\right\}\\
&= e^{-\alpha|B_r|_d}\left[e^{\alpha|B_s|_d}\right]_0^r\\
&= -\left(1 - e^{-\alpha|B_r|_d}\right)\\
&= -\alpha|B_r|_d + \mathcal{O}\left(\alpha^2\right).
\end{aligned}$$

In case (ii) the influence function is

$$\begin{aligned}
(1 - F(r))&\left\{\int_0^r \frac{|\partial B(X, s) \cap W_{\ominus s}|_{d-1} - \alpha|\partial B_s|_{d-1}|W_{\ominus s} \setminus B(X, s)|_d}{|W_{\ominus s}|_d e^{-\alpha|B_s|_d}}\mathrm{d}s\right\}\\
&= e^{-\alpha|B_r|_d}\int_0^r \frac{|\partial B(X, s) \cap W_{\ominus s}|_{d-1}}{|W_{\ominus s}|_d e^{-\alpha|B_s|_d}}\mathrm{d}s + \mathcal{O}(\alpha)\\
&= \int_0^r \frac{|\partial B(x, s) \cap W_{\ominus s}|_{d-1}}{|W_{\ominus s}|_d}\mathrm{d}s + \mathcal{O}(\alpha).
\end{aligned}$$

To check this, observe that the expected influence function is therefore, to first order in α,

$$\begin{aligned}
-\alpha|B_r|_d &+ \alpha|W|_d\mathrm{E}\left(\int_0^r \frac{|\partial B(X, s) \cap W_{\ominus s}|_{d-1}}{|W_{\ominus s}|_d}\mathrm{d}s\right)\\
&= -\alpha\left(|B_r|_d - |W|_d\int_0^r \frac{\mathrm{E}|\partial B(X, s) \cap W_{\ominus s}|_{d-1}}{|W_{\ominus s}|_d}\mathrm{d}s\right).
\end{aligned}$$

By a well-known result of integral geometry (Santaló, 1976, p. 97) the expectation in the numerator is

$$\mathrm{E}|\partial B(X, s) \cap W_{\ominus s}|_{d-1} = \frac{|\partial B_s|_{d-1}\,|W_{\ominus s}|_d}{|W|_d}$$

so that the expected influence function is

$$\begin{aligned}-\alpha\left(|B_r|_d - |W|_d \int_0^r \frac{|\partial B_s|_{d-1}\,|W_{\ominus s}|_d}{|W_{\ominus s}|_d\,|W|_d}\mathrm{d}s\right)\\ &= -\alpha\left(|B_r|_d - \int_0^r |\partial B_s|_{d-1}\mathrm{d}s\right)\\ &= 0.\end{aligned}$$

What we are really looking for, the variance of the influence function, is to first order just the expectation of the square of the 'area of failures' term from case (ii) (since case (i) is now $\mathcal{O}\left(\alpha^2\right)$):

$$\alpha|W|_d\mathrm{E}\left(\int_0^r \frac{|\partial B(X,s)\cap W_{\ominus s}|_{d-1}}{|W_{\ominus s}|_d}\mathrm{d}s\right)^2.$$

For the reduced sample estimator, the calculations are similar but easier. In case (i) the estimator is identically zero; in case (ii) it is

$$\widehat{F}_{\mathrm{RS}}(r) = |B(X,r)\cap W_{\ominus r}|_d/|W_{\ominus r}|_d.$$

Since $F(r) = 1-\exp(-\alpha|B_r|_d) = \alpha|B_r|_d + \mathcal{O}\left(\alpha^2\right)$ the influence function (= estimator − estimand in this linear case) is in case (i)

$$-\alpha|B_r|_d + \mathcal{O}\left(\alpha^2\right);$$

in case (ii)

$$\frac{|B(X,r)\cap W_{\ominus r}|_d}{|W_{\ominus r}|_d} + \mathcal{O}\left(\alpha\right).$$

The expectation of the influence function is, to first order,

$$\begin{aligned}-\alpha|B_r|_d + \alpha|W|_d\mathrm{E}\left(|B(X,r)\cap W_{\ominus r}|_d\right)/|W_{\ominus r}|_d\\ &= \alpha\left\{-|B_r|_d + |W|_d\frac{|B_r|_d|W_{\ominus r}|_d/|W|_d}{|W_{\ominus r}|_d}\right\}\\ &= 0,\end{aligned}$$

as should be the case. The variance is

$$\alpha|W|_d\mathrm{E}\left(\left(\frac{|B(X,r)\cap W_{\ominus r}|_d}{|W_{\ominus r}|_d}\right)^2\right) + \mathcal{O}\left(\alpha^2\right).$$

The conclusion is that we must calculate and compare the expected squared values of

$$\int_0^r \frac{|\partial B(X,s)\cap W_{\ominus s}|_{d-1}}{|W_{\ominus s}|_d}\mathrm{d}s$$

and

$$\frac{|B(X,r) \cap W_{\ominus r}|_d}{|W_{\ominus r}|_d}$$

for $X \sim \text{uniform}(W)$.

For convenience in calculation, we will take W to be the d-dimensional unit cube centred at $(\frac{1}{2}, \ldots, \frac{1}{2})$, and replace the Euclidean metric $|| \cdot ||$ by the L_∞ metric in the definition of ρ and $A_{\oplus r}, A_{\ominus r}$. Thus F becomes the 'empty square space' function obtained by replacing $B(x,r)$ by a cube $B_\infty(x,r)$ of centre x and side length $2r$.

We need to consider all possible ways the cubes $B_\infty(X,r)$ and $W_{\ominus r}$ intersect. For given $X = x \in W$, as r increases, initially $B_\infty(x,r)$ is entirely contained in $W_{\ominus r}$, then one-by-one the faces of $B_\infty(x,r)$ pass through faces of $W_{\ominus r}$.

By symmetry we may take X uniformly distributed on the simplex $\{x : x_1 < x_2 < \ldots < x_d < \frac{1}{2}\}$. The different transitions then occur as the value $2r$ passes through x_1, then x_2, ..., then x_d; and then as $(1-2r)$ passes through $x_d, x_{d-1}, \ldots, x_1$. The latter cases are only relevant when $r > 1/4$.

After expressing the volume and surface area contributions in terms of the x_i in each case, we integrate over r (for Kaplan-Meier only) and then over x.

In one dimension the variance of $n^{1/2}(\widehat{F}(r) - F(r))$ is approximately (ignoring terms of order $O(\alpha^2)$) equal to α times the following expression:

$$\begin{cases} 2r + (1-4r)\log(1-2r) - \frac{1}{2}(\log(1-2r))^2 & \text{for } 0 \le r \le \frac{1}{4}, \\ 2r + \int_{\frac{1}{2}}^{2r} \log u \log(1-u) \mathrm{d}u - 2r \log 2r \log(1-2r) & \text{for } \frac{1}{4} \le r < \frac{1}{2}. \end{cases}$$

For the reduced sample estimator $|\Phi_{\oplus r} \cap W_{\ominus r}|_d / |W_{\ominus r}|_d$, the corresponding formula is

$$\begin{cases} 4r^2(1 - \frac{8r}{3})/(1-2r)^2 & \text{for } 0 \le r \le \frac{1}{4}, \\ (8r-1)/3 & \text{for } \frac{1}{4} \le r < \frac{1}{2}. \end{cases}$$

These functions are plotted in Figure 3 together with the corresponding curves for two and three dimensions; the latter have been calculated (by Mathematica) with a mixture of computer algebra and numerical integration (for integrals over s) and Monte-Carlo integration (for integrals over x). The new estimator is superior over a broad range of distances r, but surprisingly deteriorates at very large distances. Apparently, the kind and amount of dependence here has destroyed the optimality of Kaplan-Meier in the classical i.i.d. case.

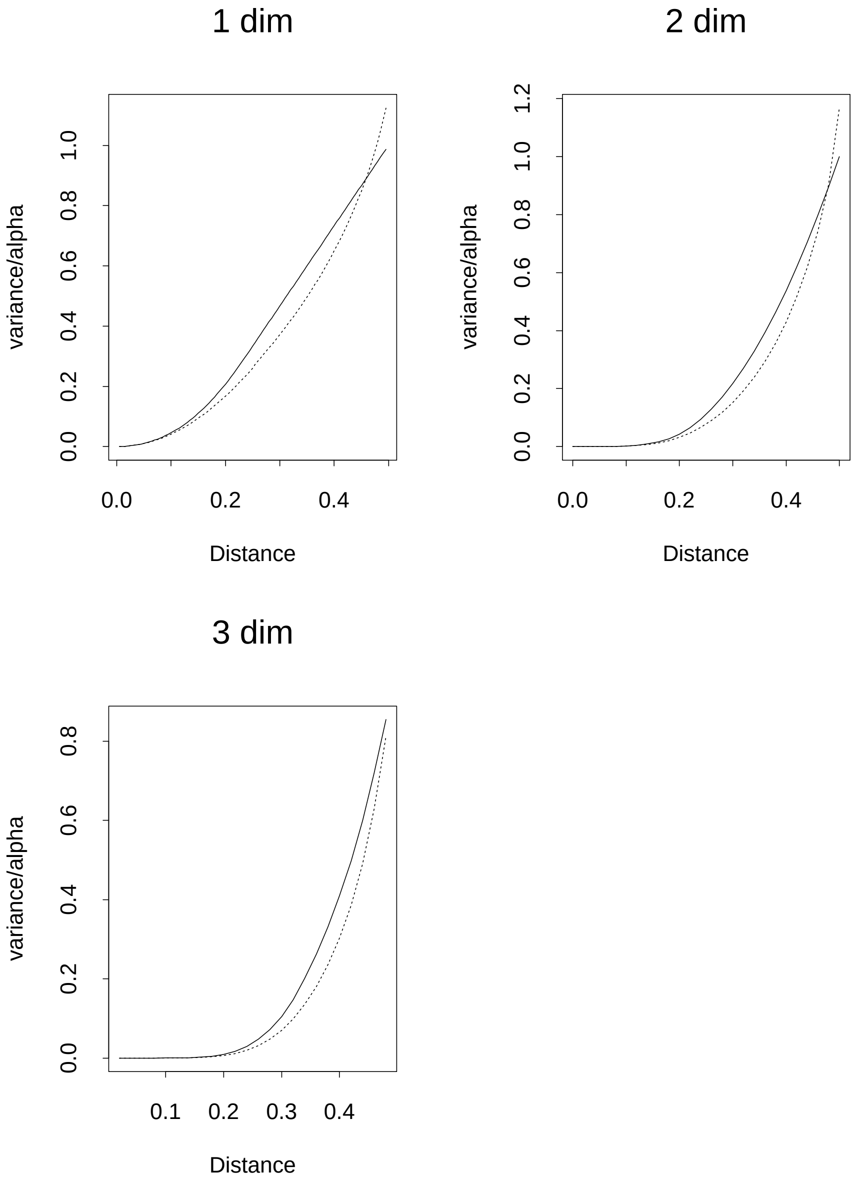

Figure 3. Sparse-limit asymptotic variances (divided by α). *Solid lines:* reduced sample estimator; *dotted lines:* Kaplan-Meier estimator.

Figure 4 shows the asymptotic relative efficiency (ratio of variances of reduced sample to Kaplan-Meier) in each dimension. The greatest gain is achieved at intermediate distances (near $\frac{1}{4}$); only for very large distances (near $\frac{1}{2}$) is there a loss in efficiency. As the dimension d increases, and hence as edge effects become more severe, Kaplan-Meier represents an ever more convincing improvement on the reduced sample estimator.

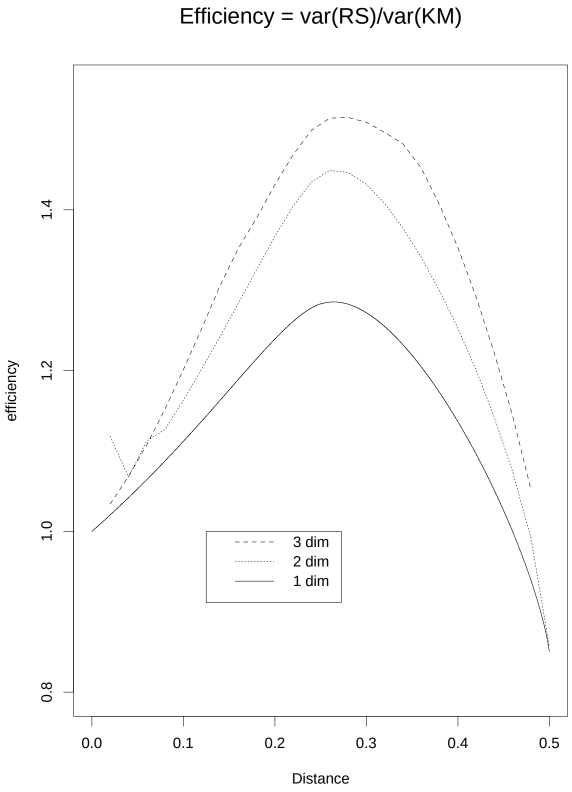

Figure 4. Asymptotic relative efficiency in 1, 2 and 3 dimensions.

The nearest neighbour distance function G.

A Kaplan-Meier estimator for G is more immediate than for F: for each point x_i of the process Φ observed in the window W, one has a censored distance from x_i to the nearest other point of Φ, censored by its distance to ∂W. Counting 'observed failures' and 'numbers at risk' as for censored data:

$$N^G(r) = \#\{x \in \Phi \cap W : \rho(x, \Phi \setminus \{x\}) \leq r,\ \rho(x, \Phi \setminus \{x\}) \leq \rho(x, \partial W)\}$$

and

$$Y^G(r) = \#\{x \in \Phi \cap W : \rho(x, \Phi \setminus \{x\}) \geq r,\ \rho(x, \partial W) \geq r\}$$

one may check that these satisfy the same mean-value relation as for ordinary randomly censored data,

$$\mathrm{E}N^G(r) = \int_0^r \mathrm{E}Y^G(s)\, \Lambda^G(\mathrm{d}s),$$

where $\Lambda^G(\mathrm{d}s) = G(\mathrm{d}s)/(1-G(s-))$, and G was defined at the beginning of this section. This motivates a Nelson-Aalen estimator

$$\widehat{\Lambda}^G(r) = \int_0^r \frac{N^G(\mathrm{d}s)}{Y^G(s)}$$

and a Kaplan-Meier estimator

$$1 - \widehat{G}(r) = \prod_{0}^{r} \left(1 - \widehat{\Lambda}^{G}(\mathrm{d}s)\right).$$

In this case there is no need for G to have any special continuity properties; in fact, G may be degenerate as in the case of a randomly translated lattice.

Linearization can be applied to $\widehat{G} - G$ just as well as for $\widehat{F} - F$ and the results used to motivate variance estimators through analogues of (10)–(12). Sparse Poisson asymptotics can also be carried out in the same way. The results show a more marked superiority of Kaplan-Meier to the reduced sample estimator than in the case of the empty-space function. Moreover, the deterioration of the Kaplan-Meier estimator at large distances is not observed any more. The situation is fundamentally different from the empty space statistic since now each point x of the process Φ contributes one observation, rather than each reference point x in the window W. The asymptotic variance is of constant order rather than of the order α. The 'leading term' in the sparse Poisson asymptotics comes from the 'number of failures' part of the influence function, for the case when exactly two points are observed in the window W.

The K function.

$K(r)$ was defined as $1/\alpha$ times the expected number of points within distance r of a typical point of the process. The possibility of defining a Kaplan-Meier estimator for $K(r)$ is not so obvious until one notices that $\alpha K(r)$ equals the sum of the distribution functions of the distance from a typical point to the nearest, second nearest, and so on. For each of the distance distributions one *can* form a Kaplan-Meier estimator, since the distance from a point $x \in \Phi$ to its kth nearest neighbour is also censored just as before by its distance to the boundary. One can check that the sequence of Kaplan-Meier estimators always satisfies the natural stochastic ordering of the distance distributions. The theory we gave for F and sketched for G can also be worked through for K.

Sparse Poisson asymptotics for K turn out to coincide exacly with those for G. The reason for this is that the cases of three or more points in the window have negligible probability compared to that for two points; so the 'leading terms' for G and K are the same. For estimating K a large number of sophisticated edge-corrections exist; see Ripley (1988), Stein (1990). It turns out that as far as the sparse Poisson asymptotics are concerned, *all* these corrections are just as good, and better than Kaplan-Meier, which itself is better than the classical border correction method (the reduced sample estimator). The sophisticated edge-corrections are in practice more complicated to compute than the Kaplan-Meier estimator, so it seems that (as is fair) the more work one does, the better the result. It is disappointing (to this author!) that Kaplan-Meier is not in the first rank, and surprising that the sophisticated edge-corrections can hardly be distinguished from one another.

More details are given in Baddeley and Gill (1992).

One might wonder whether it is possible to improve the Kaplan-Meier estimators of F, G and K by considering the observed distances as *interval-censored* rather than just right censored. This seems possible since for a point $x \in W$ which is closer to ∂W

than to other points in $\Phi \cap W$, one does know that its distance to $\Phi \setminus \{x\}$ is not greater than its distance to $(\Phi \setminus \{x\}) \cap W$; so

$$\rho(x, \partial W) \leq \rho(x, \Phi \setminus \{x\}) \leq \rho(x, (\Phi \setminus \{x\}) \cap W)$$

Similar statements can be made for the distance to the kth nearest neighbour. However, treating this data as 'randomly interval-censored data' would produce asymptotically biased estimators, since the upper limit $\rho(x, (\Phi \setminus \{x\}) \cap W)$ is strongly dependent on $\rho(x, \Phi \setminus \{x\})$, unlike the lower limit $\rho(x, \partial W)$.

15. Cryptography, statistics and the generation of randomness.

This final section is quite independent of the rest of the lecture notes. It is concerned with the subject of *random number generation* and to be more specific, with an approach to the subject developed over the last decade by computer scientists working in the area of *cryptography.* What I have to say on this subject I have learnt from the master's thesis of Brands (1991), which not only surveys the results of the cryptographic theory but also the basic ingredients in it (number theory, complexity theory including Turing machines and polynomial time computation, and so on). Another recent survey is by Luby (1993).

The traditional approach to random number generation is extensive and effective. However in my opinion it fails to explain *why* it works. It consists of a large body of useful information but somehow misses the point: in what way can a completely deterministic algorithm be said to simulate randomness? In fact 'probability theory' is notably absent in treatments of the usual approach to random number generation, which mainly discuss how to find long cycles of iteratively and deterministically determined integers which over a complete cycle have nice uniformity properties. Even 'state of the art' random number generators can turn out to be rather poor for some applications; see, e.g., Ferrenberg, Landau and Wong (1992), though an algorithm intended for use on a PC may not be the most sensible thing to use for a massive supercomputer simulation! See also Knuth (1981) for the classical theory; Marsaglia and Zaman (1991) for more recent developments, and Wang and Compagner (1993) for a nice, less orthodox, approach.

In cryptography there is a need for specially reliable random number generators. The reason for this is that the best key to a secret code is a long and completely random key (it is hardest to guess). For effective use in practice the key should however be produced deterministically, by a compact and fully automatized random number generator. However, if your adversary knows what generator you have used it may not be so difficult to guess your key after all. It is rather nice that cryptographers have not just invented their own random number generators but even developed an elaborate and elegant theory, containing nice probabilistic and even statistical ideas, which actually explains why a random number generator can simulate randomness. This theory involves the intriguing notions of one-way functions and hard-core bits; it is built on algorithmic complexity theory and in particular the distinction between polynomial and non-polynomial time algorithms as separating tractable from intractable problems; and it relies on the generally accepted (though still unproven) intractability of certain problems such as the factorization of large integers. I will argue that the theory is highly relevant to the actual use of random number generators in statistical simulation experiments, bootstrapping, randomized optimization algorithms, and so on.

A classical random number generator is an algorithm which, on given a starting number called the *seed*, produces a sequence of numbers according to a simple deterministic recursion. Usually the numbers are integers in a given, finite range, hence the numbers (eventually) follow a, usually rather long, cycle. For instance, the very well-known linear congruential generator, starting with an integer seed x_0, produces a sequence of integers x_n according to the rule

$$x_n = ax_{n-1} + b \bmod m \tag{1}$$

where the the integers a, b and m are fixed integer parameters of the method. If the parameters have been chosen appropriately the numbers x_n follow a cycle which is actually a permutation of the set of all integers modulo m, $\mathbb{Z}_m = \{0, 1, \ldots, m-1\}$. Moreover the numbers

$$u_n = x_n/m$$

behave reasonably like independent uniform $(0, 1)$ random variables and

$$y_n = \lfloor 2u_n \rfloor$$

as independent Bernoulli $(\frac{1}{2})$ variables. For good quality results m should be quite large, e.g., it should be at least a 60 bit integer (see Knuth, 1981). From uniformly distributed variables one can in principle produce numbers from any other desired distribution.

Since a random uniform$(0, 1)$ random variable is usually approximated on the computer by a number of fixed, finite precision, and since the successive bits in a uniform $(0, 1)$ random variable are independent Bernoulli $(\frac{1}{2})$ variables, a random number generator which produces Bernoulli$(\frac{1}{2})$ variables is all we really need. In fact for some choices of m the 'lower' (less significant) bits of the numbers produced by a linear congruential generator are a good deal less random than the higher bits and one may prefer to just build everything from the simulated independent Bernoulli $(\frac{1}{2})$ trials, or fair coin tosses, y_n. Note that y_n is the 'first bit' of the number u_n expressed as binary fraction.

The new generators from cryptography theory are not much different from the classical generators. For example, the so-called *quadratic-residue* or QR-generator which we study in more detail later is defined as follows: given suitably chosen integers x_0 and m, define

$$x_n = x_{n-1}^2 \bmod m \tag{2}$$

and let

$$y_n = x_n \bmod 2$$

be the 'last bit' of x_n. Then we will show that the y_n can well approximate fair Bernoulli trials. The theorem which guarantees this (under a certain unproven but highly respectable assumption) is an asymptotic theorem, for the case that the length k of the numbers concerned, in their binary representation, $k = \lceil \log_2 m \rceil$, converges to infinity. Preliminary testing shows that a similar size of m as for the linear congruential generator produces results of similar quality (Brands, 1991). A minor difference from the classical generators is that what would be a fixed parameter m is now also considered part of the seed. The only parameter of the QR-generator is in fact the chosen length k of the numbers x_n produced inside the generator.

The idea in cryptography is that a random number generator is not a device for *creating* randomness but rather a device for *amplifying* randomness. If we consider the seed as truly random, then the output sequence y_n is also random, and we may ask how close its distribution is to the distribution of fair Bernoulli trials (the answer depending on the distribution of the seed, of course). This is very similar to the situation in chaotic dynamical systems in which a small *random* perturbation of the initial conditions produces a complete, *very* random process whose distribution is essentially unique (usually the perturbation has to be absolutely continuous with respect to Lebesgue measure but otherwise does not have to be specified).

If the seed (e.g., for the QR-generator, x_0 and m together) is chosen at random the output sequence y_n is also random but clearly its (joint) distribution is highly degenerate, especially if the output sequence is long. Suppose we generate $y_1, \ldots, y_l$ where the number l is a (low degree) polynomial in k. Specifying x_0 and m requires $2k$ binary digits; we will indeed show later how it is done using about $2k$ fair Bernoulli trials (one might conceivably use real-life fair coin tosses). Think of $l = l(k)$ as being something like k^4 and forget the factor 2. Then we are talking about using, e.g., 100 fair coin tosses to simulate 100^4: we put a hundred coin tosses in, we get a hundred *million* out. The joint distribution of $y_1, \ldots, y_l$ is highly degenerate; there are only 2^k possible, equally likely, values for the whole sequence (assuming they are all different) out of an enormous 2^l equally likely values of a true random sequence. However the degeneracy can be so well hidden that we are not aware of it. And this must hold for the classical random number generators which are routinely used by statisticians and others at exactly the kind of scale described here.

Obviously the degeneracy can be found if one looks for it: if you want a good test of whether $y_1, \ldots, y_l$ are truly random or only pseudo-random, check if the sequence you have is one of the 2^k sequences produced by the generator or one of the other 2^l sequences possible with a truly random sequence. Comparing the numbers 2^{100} with $2^{100\,000\,000}$ one sees our test constitutes a statistical test with size about zero and power about one when applied to this generator. There is a big drawback to this test however: it takes a lot of time to compute. Producing a single sequence of 100 000 000 numbers for our statistical simulation experiment is very feasible, but producing all 2^{100} possible sequences is definitely not feasible. So the just mentioned statistical test is infeasible; but there might well be tests which are feasible to compute but which just as conclusively detect pseudo-randomness from true randomness.

The aim of cryptography theory is to construct random number generators such that *no practically feasible* method can show up the difference between a generated sequence and a true random sequence. The phrase 'no practically feasible' sounds vague but can in fact be made completely precise through the notions of algorithmic complexity theory. It should be taken in an asymptotic sense, since only asymptotically (as the size of a given problem increases) can one distinguish between tractable and intractable problems. Practically feasible, or tractable, means *polynomial time*: that is, the running time of the algorithm used to compute the test is at most polynomial in the size of the problem (here, we measure size by input length k, or equivalently, by l). 'Showing up the difference' between a generated sequence and a true random sequence can also be made precise. We have a statistical testing problem with, as null hypothesis, true randomness; as alternative, the distribution inherited by the generated sequence from the distribution of the seed. A given statistical test shows up a difference if there is a difference between the size and power of the test, which are just the probabilities of 'rejecting an output sequence as looking non-random' when it is really random and when it is only pseudo-random. Again, this has to be formulated in an asymptotic sense. At the same time, 'practically feasible' is formulated in a probabilistic and asymptotic sense: the algorithm must run on average in polynomial time. We will show that the QR-generator has these properties, provided it is true (as most people believe) that factoring large integers is (on average, asymptotically) infeasible.

Factoring integers enters here because of the way we choose m: in fact we let $m = pq$ where p and q are randomly chosen primes. A statistical test which shows up the nonrandomness of this random number generator could be rebuilt into an algorithm, which doesn't take an essentially longer time to run, for factoring m. Since we believe no polynomial time algorithms exist for factorising m, there cannot be a polynomial time statistical test which the QR-generator fails. Note that if the size and power of a given test are different, one can independently repeat the test a number of times and build a new test whose power and size lie even further apart. In fact, if the power and size differ by at least one divided by a polynomial, then at most a polynomial number of replications of the test suffice to bring the size close to zero and the power close to one. Thus: 'failing a feasible statistical test' in the weak sense of power being just slightly bigger than size means that there exists a more conclusive feasible test which the generator also fails.

As we mentioned, if the seed of the QR-generator is sampled appropriately, the generator can be proven to be 'cryptographically secure' (hence statistically reliable) under a reasonable assumption (born out by all practical experience and not contradicted by any theory) about the infeasibility of factoring products of large primes. The linear congruential generator, as it is usually used, can be shown not to be secure: one can essentially recover the seed from the sequence with not too much work, and hence come up with statistical tests which overwhelmingly reject its randomness. However it is quite plausible that if not just x_0 but also (some aspects of) a, b and m are chosen at random in an appropriate way, and if not the whole x_n but just, say, y_n is output on each iteration, the generator is secure. This is an interesting open question. My feeling would be that good behaviour of a given generator in (varied and extensive) practice means that it can probably be implemented in a cryptographically secure way.

Before embarking on the theory we should pay some more attention to its relevance. In practice, does it make sense to suppose the seed of a random number generator is chosen at random? What has 'passing all feasible statistical tests' (i.e., the power and the size of any feasible test are essentially equal) got to do with how a generator is actually used in practice?

As an example, let us consider the statistical simulation experiments carried out in Nielsen, Gill, Andersen and Sørensen (1992) which aimed to show that a kind of generalised likelihood ratio test (in a certain semiparametric model from survival analysis estimated by non-parametric maximum likelihood) has the same asymptotic properties as in the parametric case. During the simulations the nominal P-value of a log likelihood ratio test, assuming an asymptotic chi-square distribution to be applicable under the null-hypothesis, was calculated for a large number of large samples from the model, under the null-hypothesis. If the conjectured asymptotic theory is true and if the chosen sample size is large enough to make it a reasonable approximation, these P-values should be approximately a sample from a uniform distribution on $(0, 1)$. Under the alternative their distribution should shift to smaller values. The results of the simulations were summarized in a number of QQ-plots of uniform quantiles set out against ordered, observed P-values; see Figure 1 for a typical case.

Figure 1. QQ-plot of uniform quantiles versus simulated nominal P-values, under the null-hypothesis.

There are a 1000 points in the graph and each point represents a test-statistic based on a sample of size 1000 from a bivariate distribution. Thus, supposing real numbers were represented by strings of 30 bits, about 60 million simulated fair coin tosses are needed to draw the graph. In fact the simulation was the completely deterministic result of repeatedly calling a random number generator, starting with an initial random seed represented as a string of about 100 bits. The random seed is the result left at the end of the previous simulation experiment; alternatively one may let the system 'reset' the seed in some mysterious way (using the system clock, perhaps) or the user can reset it: perhaps with real fair coin tosses but more likely using a coding of his or her birthday or bank account number or just with the first 'random' string of numbers which came to mind. Whichever was the case, I am completely happy to consider the initial random seed, for this simulation experiment, as truly random and perhaps even uniformly distributed on its range. Obviously if I carry out a number of simulation experiments at the same workstation using subsequent segments of the same cycle of pseudo-random numbers, different experiments are not independent of one another. However this doesn't change the interpretation of what is going on in one given experiment.

Also in a bootstrap experiment, a simulated annealing calculation, and other statistical applications, a hundred or so 'more or less' truly random, fair coin tosses, are used to generate several million up to several billion fair coin tosses.

Obviously the distribution of the output sequence does not remotely look like what it is supposed to simulate. However, we are not interested in the whole joint distribution of the output sequence but just in the distribution of a few numerical statistics, or even just of one or two zero-one valued statistics. For instance, the conclusion drawn from Figure 1 is 'this looks like a uniform sample'. One could summarize this impression by calculating some measure of distance of the observed curve from the diagonal, or one could even carry out a formal Kolmogorov-Smirnov test at the 5% level (with as conclusion 'O.K.'). The result of a bootstrap experiment is the measurement of one or two empirical quantiles, to be used in the construction of a confidence interval. The only important thing about these observed quantiles (based on several thousand replicates of a statistic computed on samples of one hundred or a thousand observations) is that they lie with large probability, under pseudo-randomness, in the same small interval (about 'the true bootstrap quantile') as under true randomness.

Conclusion: even if we produce millions of random numbers in a statistical simulation experiment, we are really only interested in the outcome of a few zero-one variables computed from all of them. In fact, our use of the simulation is based on a reliance that these variables have essentially the same distribution under pseudo-randomness as under true randomness: in other words, they should be no use as a test of randomness. If the distributions were different and known in advance, we could even use (preferably, several replicates of) our simulation experiment as a test of our generator. It would be the most sensible test to use since it tests exactly the aspect of the generator which is important for us! However, the probabilities in question are not known in advance and cannot be easily calculated, which is after all exactly the reason we were doing a simulation experiment in the first place.

Note also that even if our simulation experiment is large, we still get it finished in a reasonable length of time and if necessary could repeat it a few times. This means that the statistical test of randomness which our use of the experiment represents, is a feasible test. Consequently: a random number generator which passes all feasible tests is a random number generator which we can safely use for all practical purposes.

I would like to go into one other digression before embarking on the theory as promised. This concerns some connections between random number generators, rounding errors, and the randomness of, e.g., a classical fair coin toss.

The iterations of the linear congruential generator $x_n = ax_{n-1} + b \bmod m$ are quite easy to analyse. First of all, one can iterate a number of times without reducing modulo m and then only take the residue modulo m afterwards. This leads to the fact:

$$x_n = \left(a^n x_0 + \frac{a^n - 1}{a-1} b \right) \bmod m$$

Also, dividing the x_n and b by m, one can take the residue of real numbers modulo 1; in other words, the fractional part, denoted $\{\cdot\}$. We find

$$y_n = \{\, a^n (y_0 + \frac{b/m}{a-1}) - \frac{b/m}{a-1} \,\}.$$

This means that the pseudo-uniform random numbers produced by the linear congruential generator are nothing else than the rounding errors in a table of the mathematical function $n \mapsto \exp(\alpha n + \beta) + \gamma$, when the table values are computed to the nearest integer; take $\alpha = \log a$, $\beta = \log((b/m)/(a-1))$, $\gamma = -(b/m)/(a-1)$.

It is part of folk-lore of numerical mathematics and computer science that 'rounding errors are uniformly distributed' and much practical experience and some theory exists to support this observation. Less well established is that rounding errors in successive entries in a table of a mathematical function are approximately independent; at least, if the table entries are sufficiently far apart. This fact (which can be empirically checked) has been put to good use in the mathematical-historical study of medieval arabic astronomical tables by van Dalen (1993). Several astonomical tables known to historians of science are tabled values of known functions but with unknown parameters (some parameters have varied over the centuries, others depend on geographical location). The use of statistical techniques to determine the parameter values by non-linear regression is controversial since, apart from gross errors which are usually easy to identify, the tables have been calculated following a precise algorithm which yields exact results to the required number of (hexadecimal) digits. Thus the only error is the final rounding error; it is completely deterministic, and to consider it random or even independent, uniform, is hard for some historians to stomach.

The fair coin toss can also be considered the result of a rounding procedure. Suppose a (horizontally) spinning coin is thrown up vertically and falls back to a level surface on which it is caught and made to lie horizontally without any bouncing. The side uppermost can be computed as a function of the initial vertical speed v and rotation speed ω. In fact, we can represent the total angle through which the coin has rotated at the moment it is stopped in terms of these two parameters. We round the angle to a multiple of 2π and then look if it lies between zero and π (heads) or π and 2π (tails). The randomness of the outcome (heads or tails) is the result of the randomness, or if you like, variability, of the initial parameters v and (appropriately enough) ω. Since heads or tails is responsive to very small variations in these parameters (at least, when they are large enough to begin with), and by some symmetry properties, a smooth distribution of v, ω over a small region will make heads and tails about equally likely; see Engel (1992).

The point about this digression is that an argument about how random the seed is of a random number generator is very, very similar to the argument how random is a coin toss; in fact, we are always forced to an infinite regress in which small amounts of probability are needed to explain more; however, it is often the case that the type of randomness which we get out of the system is not critically dependent of the type of randomness we put in.

Now back to cryptography. We start with a specific example. The QR-generator, proposed by Blum and Micali (1984) and using on ideas of Rabin (1979), is specified as follows. Given a number k generate at random a prime number p and a prime number q of length at most $\lfloor k/2 \rfloor$ bits; p and q should furthermore be unequal to one another, and both should be congruent to 3 modulo 4. Subject to these restrictions p and q may be thought of as being uniformly distributed over the set of all possible pairs (in practice they will be chosen with a slightly different distribution as we will explain later;

that does not change the subsequent theory in any essential way). Define $m = pq$ (a number of at most k bits) and choose x_0 (also at most k bits) uniformly at random from $\mathbb{Z}_m^* = \{x \in \mathbb{Z}_m : \text{neither } p \text{ nor } q \text{ divides } x\}$. One may also describe $\mathbb{Z}_m^*$ as the set of elements of $\mathbb{Z}_m$ with a multiplicative inverse (modulo m); it forms a multiplicative group. Now define recursively $x_n = x_{n-1}^2 \bmod m$, $y_n = x_n \bmod 2$, $n = 1, 2, \ldots, l$ where $l = l(k)$ is at most polynomial in k. We later show how p, q and x_0 can be determined (easily: in polynomial time) from $2k$ fair coin tosses.

Define also

$$\mathrm{QR}_m = \{x^2 \bmod m : x \in \mathbb{Z}_m^*\}.$$

This is called the set of quadratic residues, modulo m. From fairly elementary number theory (the theory of the Jacobi and the Legendre symbols; the latter, as group homomorphisms from $\mathbb{Z}_m^*$ to the multiplicative group $\{-1, 0, 1\}$, have something to say about whether a number is a square or not) it follows that *exactly a quarter of the elements of* $\mathbb{Z}_m^*$ *are squares; i.e.; members of* QR_m. *Moreover, each member of* QR_m *is the square (modulo m) of exactly four different members of* $\mathbb{Z}_m^*$, *having the form* $\pm x, \pm y$. *Just one of these square roots is itself also a square. Therefore, the function* $x \mapsto x^2 \bmod m$ *is a permutation on* QR_m.

The reader is invited to calculate the table of squares of elements of $\mathbb{Z}_m^*$ in the case $p = 3$, $q = 7$, and further to investigate the sequences y_n produced by the generator.

Neglecting the factor 2 in the total length of our input string (m, x_0) we consider the QR-generator as a mapping from binary strings of length k to binary strings of length l; or rather, for a given (polynomial) dependence $l = l(k)$ as a sequence of such mappings, one for each value of k. As explained above, by the QR-generator the 2^k possible input strings are mapped into the much larger set of 2^l possible output strings. We put the uniform probability distribution on the input strings and consider the statistical problem of distinguishing the resulting probability distribution on output strings from the uniform distribution on the large set of all binary strings of length l.

Two notions are central to showing that the QR-generator (and many other generators) is reliable: the notion of a *one-way function*, and the notion of a *hard-core predicate*. A one-way function is a function which is easy to compute, while its inverse is difficult (we restrict attention here to functions which are one-to-one, with the same domain and range, hence are permutations). 'Easy' and 'difficult' mean here: on average, in polynomial time, and not in polynomial time, respectively. The notion is therefore an asymptotic notion and we are really applying it to a sequence of functions f_k, typically from a given subset of the set of binary strings of length k to another. In our description of the theory we will, for simplicity, usually suppose the function is defined on $\{0, 1\}^k$, but our examples will involve slightly more complicated domains.

An example: consider the function 'multiplication' on the set of pairs of different, ordered primes, each represented by binary integers of at most k bits. This function is easy to compute: one can easily exhibit an algorithm which runs in an most $\mathcal{O}(k^2)$ time steps, where in each time step one basic operation on just two bits is performed. However it is believed that no algorithm exists which computes the inverse of this function, 'factorization', on average in an amount of time polynomial in k. This belief is backed up by a huge amount of practical experience and much theoretical work too. A proof would in fact establish the famous conjecture '$P \subset NP$' (strict inclusion) which says, in words, that there exist problems which, though a supposed answer to them can be checked to be correct in a polynomial number of steps, no algorithm exists which solves the problem (without knowing the answer in advance) in a polynomial number of steps. In fact the existence of any one-way function at all would prove the '$P \subset NP$' conjecture. Considering the huge amount of work which has been put into this attempt, without success, it is not likely that the existence of one-way functions is going to be *proved* for quite a while.

At present the best known factorisation algorithm takes about a year on a very fast computer to factor a 100 digit product of two large, unknown primes. The same algorithm would take about a million years to factor a 200 digit number. This illustrates what it means for an algorithm to be non-polynomial time: there is in practice a rather strict limit to the size of problem which can be solved; and increasing the speed of computers has very little effect on the limit. On the other hand, and rather important for the feasibility of the QR-generator which requires one to randomly sample prime numbers, the related problem of just deciding whether a given number is prime or not, can be solved in polynomial time (using in fact a probabilistic algorithm which therefore is not guaranteed to give the right answer, but can give the right answer with a probability as close to 1 as one likes!). To decide whether or not a 100 digit number is prime takes about half a minute.

We will show in a moment that the function 'square' from QR_m to QR_m is also a one-way function, by demonstrating the equivalence of computing its inverse with the problem 'factoring' (assumed to be one-way) just described. Really we should index the set QR_m not by m but by the chosen length k, since m is not supposed to be fixed or known in advance.

The other central notion is that of a *hard-core predicate.* Though the inverse of a function f may be hard to predict, it is conceivable that a number of properties of the inverse are in fact easy to determine. For instance, it is easy to find out if a large integer is prime or not, but difficult to supply a list of its prime factors. A hard-core predicate is a property of the inverse which is essentially as difficult to determine as the inverse itself. Let such a property be described by a function B from the range of f to the set $\{0, 1\}$. Then f one-way means $x \mapsto f(x)$ is easy, but $y \mapsto f^{-1}(y)$ is difficult to compute; B hard-core for f means that $x \mapsto B(f(x))$ is easy to compute (if you knew the inverse x of $y = f(x)$, you could calculate the property easily), but $y \mapsto B(y)$ is not easy.

To be a little more precise, a one-to-one function f, say defined on $\{0, 1\}^k$ for each k, is one-way if for all polynomial time functions M and all polynomials $p(\cdot)$,

$$\mathrm{P}_k\Big(f(M(f(x))) = f(x)\Big) < \frac{1}{p(k)}$$

for all sufficiently large k, where the probability distribution P_k is (typically) the uniform distribution of x on $\{0, 1\}^k$. Also, B is hard-core for f if, on the one hand, $B(f(\cdot))$ can be computed in polynomial time, but on the other, for all polynomial time M and all polynomials $p(\cdot)$,

$$\left|\mathrm{P}_k\Big(M(x) = B(x)\Big) - \frac{1}{2}\right| < \frac{1}{p(k)}$$

for all sufficiently large k, where again x is uniformly distributed.

Note the probability of a half here: the function M guesses the value of B correctly just half of the time. This implies that B also takes the values 0 and 1 with equal probabilities, otherwise by always picking a single value one could guess right with bigger probability than a half.

As example of a hard-core predicate for the function 'square' on QR_m, we mention the so-called 'last-bit' or 'parity' function. This can be shown to be hard-core by showing that an algorithm which computes the last bit of the square root (in QR_m) of a number in QR_m (or in fact just guesses the last bit with succes probability a little better than a half), can be converted into an algorithm, not needing much more time to run, for determining the square root in its entirety. Note that the last-bit function just determines whether the square root is even or odd, or its value modulo 2. In other words, finding out if the square root of a number in QR_m is even or odd is just as difficult as finding the square root itself.

We next set up the definitions needed to discuss random number generators. A generator f is actually a polynomial time sequence of functions f_k mapping, say, $\{0,1\}^k$ to $\{0,1\}^{l(k)}$ for some polynomial function $l(\cdot)$. The domain is called the seed space and given the uniform probability distribution. Let P_n denote the uniform distribution on $\{0,1\}^n$. A feasible statistical test T of a generator f is a sequence of polynomial time functions $T_{l(k)}$ from $\{0,1\}^{l(k)}$ to $\{0,1\}$, coding for 'accept', and 'reject', where the test $T_{l(k)}$ is a test of the null hypothesis that the output sequence $y = f(x)$ is distributed as P_l against the alternative that it is distributed as $\mathrm{P}_k \circ f_k^{-1}$. We say that f passes the test T if, for any polynomial in k, the difference between the power and size of the test is eventually smaller than one divided by that polynomial. A generator is called *pseudo-random* if it passes all feasible tests.

An apparently less stringent criterion of a generator is *unpredictability.* This means that for each position from 1 up to $l-1$ in the output sequence, no feasible function exists which predicts, with better succes probability than a half, the next output bit of the sequence, given the first bits up to this position. If a generator is predictable then for some position in the output sequence one can, with some success, feasibly predict the next bit from the preceding ones. A rather nice theorem of Yao (1982) states that the property of being unpredictable is actually equivalent to being pseudo-random. In other words: passing all feasible next-bit tests implies passing *all* feasible tests. Since pseudo-randomness doesn't depend on whether the output bits are taken in their usual order or in reverse order, we have the nice corollary: forwards predictability is equivalent to backwards predictability.

Here's a sketch of the proof of the theorem. Predictable implies not pseudo-random is easy, since we can obviously construct a test of a generator from a succesful prediction of one of its output bits. For the converse, suppose the generator is not pseudo-random. This means there exists a feasible test whose size and power are 'significantly' different from one another. Denote the output sequence of the generator by $y = (y_1, \ldots, y_l)$ and let $y^* = (y_1^*, \ldots, y_l^*)$ denote a true random sequence. From these two consider all the 'cross-over' combined sequences:

$$y^{(n)} = (y_1, \dots, y_{n-1}, y_n^*, \dots, y_l^*), \quad , n = 1, \dots, l.$$

Apply the statistical test to both of $y^{(n)}$ and $y^{(n+1)}$. Since there is an appreciable difference between the probabilities of rejecting y^* and rejecting y, there has to be somewhere, at least, some difference between the probabilities of rejecting $y^{(n)}$ and $y^{(n+1)}$, since the first element of the first of these pairs is y^* and the second element of the last of the pairs is y. Here we use the fact that $l(k)$ is at most polynomial in k: a probability which is larger than one divided by some polynomial, also has this property when divided by $l(k)$.

Now if we can distinguish between $y^{(n)}$ and $y^{(n+1)}$ for some n, it seems plausible that one can predict, with some success, y_n from $(y_1, \dots, y_{n-1})$, since the only difference between $y^{(n)}$ and $y^{(n+1)}$ is whether the nth bit contains the deterministically formed y_n or the fair coin toss y_n^*. Indeed, one can show that a 'randomized' prediction algorithm can be built on comparing the results of the statistical test applied to the two sequences: $(y_1, \dots, y_{n-1}, 0, y_{n+1}^*, \dots, y_l^*)$ and $(y_1, \dots, y_{n-1}, 1, y_{n+1}^*, \dots, y_l^*)$. The algorithm is a randomised algorithm because it has to supply the fair coin tosses $(y_{n+1}^*, \dots, y_l^*)$.

To conclude the general theory, we show that given any one-way permutation, say $f : \{0,1\}^k \to \{0,1\}^k$, with a hard-core predicate B, we can construct a pseudo-random generator by iterating f, and outputting successive values of $B(f)$ (both of which are easy to do). This now famous construction is due to Blum and Micali (1984). To see this, let x be chosen at random from $\{0,1\}^k$ and let the generator output $y = g(x) = (B(f(x)), B(f(f(x))), \dots, B(f^l(x)))$. We show that g is pseudo-random by showing that it is not backwards predictable. The reason for this is that, if it were backwards predictable, we could feasibly guess, with some degree of success, the value of say $B(f^n(x))$ given the values of $B(f^{n+1}(x)), \dots, B(f^l(x))$. Knowing this latter set of values is less than knowing just $f^n(x)$, from which they may all feasibly be computed. So given $f^n(x)$ we can apparently guess $B(f^n(x))$. But this contradicts B being hard-core (here we use the fact that if f is a permutation, $f^n(x)$ is also uniformly distributed).

The mention of $\{0,1\}^k$ as domain and range of the one-way permutation f was not in any way essential here. For the QR-generator, we take as range the set of *pairs* (x, m) where $x \in \mathrm{QR}_m$ and m is the product of two different primes, congruent to 3 modulo 4, and of length $\lfloor k/2 \rfloor$ bits; and we let $f(x, m) = (x^2 \bmod m, m)$. We take $B(x, m) = \sqrt{x} \bmod 2$ where the square root is taken in QR_m. Incorporating m into both domain and range of the one-way function sets things up so that successive iterations can be done knowing m, as is needed; also it makes it clear that the pair (x, m) together form the random seed of the generator. In practice, one can sample from the seed space as follows: choose independently two random integers of length $\lfloor k/2 \rfloor$ bits and equal to 3 modulo 4 using fair coin tosses in the obvious way. Test each for primality and if it fails, increment by 4 and repeat (also demand that the second prime is different from the first). After not too many tests (by the prime-number theorem, which says that among integers of $\lfloor k/2 \rfloor$ bits, primes lie at typical distance $\mathcal{O}(k)$ apart) you will have found two primes p and q. Choose independently a random integer of length k bits by fair coin tosses; check it is not divisible by p or q (so a member of $\mathbb{Z}_m^*$; if not, repeat), and square it modulo m to obtain your initial x, element of QR_m.

Probabilists will be immediately aware that this procedure does not sample *uniformly* from the seed space. The chance a given prime p or q is selected is proportional to the distance between it and the previous prime. It seems not possible to achieve a uniform distribution on primes with a polynomial time sampling algorithm. This is not a crucial point at all, since it is just as plausible that factoring is, on average, infeasible when pairs of primes are sampled as described, as when they are sampled uniformly. In the proof above it was needed that the distribution of $f(x)$ was the same as that of x; but that is also true in our case, with x replaced by (x, m).

A more delicate point is that the sampling procedure requires more input randomness than just the fair coin tosses to start the search for p and q and to choose x, in that the primality-testing algorithm has to be probabilistic if it is to be a polynomial time algorithm. So one should also 'count the coin tosses' needed here in order to properly judge the effectiveness of the QR-algorithm as a random generator; typically $\mathcal{O}(k)$ suffice, so this is not a problem: even when these coin tosses are taken into account, we have output a much longer sequence of simulated coin tosses.

It is amusing that in a theory which depends on the notion of a *probabilistic* polynomial time algorithm (in fact, a probabilistic Turing machine) to characterise feasible and infeasible problems, one should go to so much trouble to describe how randomness can be generated, or rather expanded, in a deterministic way. A probabilistic Turing machine is supposed to be able to generate its own fair coin tosses, so looking from inside the theory, random number generators are not needed; they already exist!

We have now completed our general discussion of the theory. To apply the theory to the QR-generator, just two facts have to be verified: that squaring on QR_m is one-way, and that the parity bit of the inverse is hard-core; in other words, taking square roots is hard and just deciding if the square root is even or odd is just as hard.

We prove the one-way property; the hard-core property has a rather more elaborate (and very ingenious) proof requiring more, related, facts from number theory; see Alexi, Chor, Goldreich and Schnorr (1988) for a proof, building on earlier work of Ben-Or, Chor and Shamir (1983). Actually the highest bit is also hard-core, or even some $\log k$ bits taken together. The proof of the hard-core property goes via showing that an algorithm for determining the parity of the square root can be built into an algorithm to do factoring.

For the one-wayness of squaring, we suppose it is possible to compute square roots in QR_m, for given m, and show that this leads to a not much longer algorithm for factoring m. Our algorithm will actually be a probabilistic algorithm which leads to the right answer in polynomial time with overwhelmingly large probability. Start by picking a uniform random point in $\mathbb{Z}_m^*$. (There are $(p-1)(q-1)$ elements of $\mathbb{Z}_m^*$, so this is most of $\mathbb{Z}_m$). Square it, and we now have an element of QR_m. Take the square root in QR_m: the result is $\pm x$ or $\pm y$ for some $y \neq \pm x$. Moreover the probability is a half that the answer is *not* $\pm x$. If however we find $\pm x$, simply repeat with a new random choice of x.

After a not too large number of attempts we have found x and y with $x \neq \pm y \bmod m$, $x^2 = y^2 \bmod m$. The latter equation, rewritten as $(x-y)(x+y) = 0 \bmod m$, tells us that $x \pm y$ is divisible by p or q. Now we can use the Euclidean algorithm to determine the greatest common divisor of m and, say, $x - y$ (take the remainder of the larger number on division by the smaller; discard the larger and repeat with the remainder and the smaller number). This algorithm takes $\mathcal{O}(k^2)$ steps so is feasible. The greatest common divisor is p or q and division into m provides the other prime.

A small amount of practical experience with the QR-generator (Brands, 1991), suggests that it is just as good, in the traditional sense of passing traditional statistical tests of randomness, as a linear congruential generator of similar size and used in the same way (extraction of just one bit on each iteration). We also refer to that work for full details of the theory sketched here, including an introduction to the theory of 'polynomial time computation' based on Turing machines and Boolean circuits, and a survey of the number theoretic results needed to understand the cryptographic theory.

Appendix. Product-integrals in TEX.

For the reader interested in writing up his own research on product-integration, here are TEX macros and a postscript file (the latter to be saved as 'pi.ps') for printing a nice product-integral symbol. The files are also available by email from the author. **Exercise**: make the ultimate product-integral with METAFONT.

Bibliography.

O.O. Aalen (1972), *Estimering av Risikorater for Prevensjonsmidlet 'Spiralen'* (in Norwegian), Master's thesis, Inst. Math., Univ. Oslo.

O.O. Aalen (1975), *Statistical Inference for a Family of Counting Processes*, PhD thesis, Univ. California, Berkeley.

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

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

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

M.G. Akritas (1986), Bootstrapping the Kaplan-Meier estimator, *J. Amer. Statist. Assoc.* **81**, 1032–1038.

W. Alexi, B. Chor, O. Goldreich, and C.P. Schnorr (1988), RSA and Rabin functions: certain parts are as hard as the whole, *SIAM J. Comp.* **17**, 194–209.

B. Altshuler (1970), Theory for the measurement of competing risks in animal experiments, *Math. Biosci.* **6**, 1–11.

P.K. Andersen, Ø. Borgan, R.D. Gill, and N. Keiding, (1982), Linear nonparametric tests for comparison of counting processes, with application to censored survival data (with discussion), *Int. Statist. Rev.* **50**, 219–258; Amendment, **52**, 225 (1984).

P.K. Andersen, Ø. Borgan, R.D. Gill, and N. Keiding (1993), *Statistical Models Based on Counting Processes*, Springer, New York.

A.J. Baddeley (1987), Integrals on a moving manifold and geometrical probability, *Adv. Appl. Probab.* **9**, 588–603.

A.J. Baddeley and R.D. Gill (1992), Kaplan-Meier estimators for interpoint distance distributions of spatial point processes, Preprint 718, Dept. Math., Univ. Utrecht; revised version (1993), submitted to *Ann. Statist.*

A.J. Baddeley, R.A. Moyeed, C.V. Howard, S. Reid, and A. Boyde (1993), Analysis of a three-dimensional point pattern with replication, *Appl. Statist.* **42**, to appear.

L.G. Barendregt and M.J. Rottschäfer (1991), A statistical analysis of spatial point patterns: a case study, *Statistica Neerlandica* **45**, 345–363.

M. Ben-Or, B. Chor, and A. Shamir (1983), On the cryptographic security of single RSA bits, *Proc. 15th ACM Symp. Theor. Comp.*, 421–430.

P.J. Bickel, C.A.J. Klaassen, Y. Ritov, and J.A. Wellner (1993), *Efficient and Adaptive Inference for Semiparametric Models*, Johns Hopkins University Press, Baltimore (in press).

M. Blum and S. Micali (1984), How to generate cryptographically strong sequences of pseudo-random bits, *SIAM J. Comp.* **13**, 850–864.

P.E. Böhmer (1912), Theorie der unabhängigen Wahrscheinlichkeiten, *Rapports, Mém. et Procés-verbaux* 7^{e} *Congrès Int. Act. Amsterdam* **2**, 327–343.

S.J. Brands (1991), *The Cryptographic Approach to Pseudo-random Bit Generation*, Master's thesis, Dept. Math., Univ. Utrecht.

N.E. Breslow and J.J. Crowley (1974), A large sample study of the life table and product limit estimates under random censorship, *Ann. Statist.* **2**, 437–453.

C.F. Chung (1989a), Confidence bands for percentile residual lifetime under random censorship model, *J. Multiv. Anal.* **29**, 94–126.

C.F. Chung (1989b), Confidence bands for quantile function under random censorship, *Ann. Inst. Statist. Math.* **42**, 21–36.

P. Courrège and P. Priouret (1965), Temps d'arrêt d'un fonction aléatoire, *Publ. Inst. Stat. Univ. Paris* **14**, 245–274.

D.R. Cox (1972), Regression models and life-tables (with discussion), *J. Roy. Statist. Soc. (B)* **34**, 187–220.

D.R. Cox (1975), Partial likelihood, *Biometrika* **62**, 269–276.

M.W. Crofton (1869), Sur quelques théorèmes du calcul intégral, *Comptes Rendus de l'Académie des Sciences de Paris* **68**, 1469–1470.

D.M. Dabrowska (1988), Kaplan-Meier estimate on the plane, *Ann. Statist.* **16**, 1475–1489.

D.M. Dabrowska (1993), Product integrals and measures of dependence, Preprint, Dept. Biostatistics, Univ. Calif., Los Angeles.

B. van Dalen (1993), *Ancient and Mediaeval Astronomical Tables: Mathematical Structure and Parameter Values*, Ph.D. Thesis, Dept. Math., Univ. Utrecht.

A.P. Dempster, N.M. Laird, and D.R. Rubin (1977), Maximum likelihood estimation from incomplete data via the EM algorithm (with discussion), *J. Roy. Statist. Soc. (B)* **39**, 1–38.

P.J. Diggle (1983), *Statistical Analysis of Spatial Point Patterns*, Academic Press, London.

R.L. Dobrushin (1953), Generalization of Kolmogorov's equations for a Markov process with a finite number of possible states, *Mat. Sb. (N.S.)* **33**, 567–596 (in Russian).

R.L. Dobrushin (1954), Study of regularity of Markov processes with a finite number of possible states, *Mat. Sb. (N.S.)* **34**, 542–596 (in Russian).

S.I. Doguwa (1990), On edge-corrected kernel-based pair correlation function estimators for point processes. *Biom. J.* **32**, 95–106.

S.I. Doguwa and G.J.G. Upton (1990), On the estimation of the nearest neighbour distribution, $G(t)$, for point processes, *Biom. J.* **32**, 863–876.

J.D. Dollard and C.N. Friedman (1979), *Product Integration with Applications to Differential Equations* (with an appendix by P. R. Masani), Addison-Wesley, Reading, Massachusetts.

H. Doss and R.D. Gill (1992), A method for obtaining weak convergence results for quantile processes, with applications to censored survival data, *J. Amer. Statist. Assoc.* **87**, 869–877.

R.M. Dudley (1966), Weak convergence of probabilities on nonseparable metric spaces and empirical measures on Euclidean spaces, *Illinois J. Math.* **10**, 109–126.

R.M. Dudley (1992), Empirical processes: p-variation for $p \leq 2$ and the quantile-quantile and $\int FdG$ operators, Preprint, Dept. Math., Mass. Inst. Tech.

B. Efron (1967), The two sample problem with censored data, *Proc. 5th Berkeley Symp. Math. Statist. Probab.* **4**, 851–853.

B. Efron (1979), Bootstrap methods: Another look at the jackknife, *Ann. Statist.* **7**, 1–26.

B. Efron (1981), Censored data and the bootstrap, *J. Amer. Statist. Assoc.* **76**, 312–319.

B. Efron and I.M. Johnstone (1990), Fisher's information in terms of the hazard rate, *Ann. Statist.* **18**, 38–62.

E.M.R.A. Engel (1992), *A Road to Randomness in Physical Systems*, Springer Lecture Notes in Statistics 71.

H. Federer (1969), *Geometric Measure Theory*, Springer Verlag, Heidelberg.

A.M. Ferrenberg, D.P. Landau, and Y.J. Wong (1992), Monte Carlo simulations: hidden errors from 'good' random number generators, *Phys. Rev. Letters* **69**, 3382–3384.

T.R. Fleming and D.P. Harrington (1991), *Counting Processes and Survival Analysis*, Wiley, New York.

M.A. Freedman (1983), Operators of p-variation and the evolution representation theorem, *Trans. Amer. Math. Soc.* **279**, 95–112.

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

R.D. Gill (1980b), Nonparametric estimation based on censored observations of a Markov renewal process, *Z. Wahrsch. verw. Geb.* **53**, 97–116.

R.D. Gill (1981), Testing with replacement and the product limit estimator, *Ann. Statist.* **9**, 853–860.

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

R.D. Gill (1986), On estimating transition intensities of a Markov process with aggregated data of a certain type: 'Occurrences but no exposures', *Scand. J. Statist.* **13**, 113–134.

R.D. Gill (1989), Non- and semi-parametric maximum likelihood estimators and the von Mises method (Part 1), *Scand. J. Statist.* **16**, 97–128.

R.D. Gill (1992), Multivariate survival analysis, *Theory Prob. Appl.* **37** (English translation), 18–31 and 284–301.

R.D. Gill and S. Johansen (1990), A survey of product-integration with a view towards application in survival analysis, *Ann. Statist.* **18**, 1501–1555.

R.D. Gill and B.Ya. Levit (1992), Applications of the van Trees inequality: a Bayesian Cramér-Rao bound, Preprint 733, Dept. Math., Univ. Utrecht.

R.D. Gill, M.J. van der Laan, and J.A. Wellner (1993), Inefficient estimators of the bivariate survival function for three multivariate models, Preprint 767, Dept. Math., Univ. Utrecht.

R.D. Gill and A.W. van der Vaart (1993), Non- and semi-parametric maximum likelihood estimators and the von Mises Method (Part 2), *Scand. J. Statist.* **20**.

M.J. Gillespie and L. Fisher (1979), Confidence bands for the Kaplan-Meier survival curve estimates, *Ann. Statist.* **7**, 920–924.

M. Greenwood (1926), The natural duration of cancer, *Reports on Public Health and Medical Subjects* **33**, 1–26, His Majesty's Stationery Office, London.

P. Groeneboom and J.A. Wellner (1992), *Information Bounds and Nonparametric Maximum Likelihood Estimation*, Birkhäuser Verlag, Basel.

H.J. Hall and J.A. Wellner (1980), Confidence bands for a survival curve from censored data, *Biometrika* **67**, 133–143.

N.L. Hjort (1985a), Bootstrapping Cox's regression model, Tech. Rept. 241, Department of Statistics, Stanford University, California.

N.L. Hjort (1985b), Discussion of the paper by P.K. Andersen and Ø. Borgan, *Scand. J. Statist.* **12**, 141–150.

J. Jacod (1975), Multivariate point processes: Predictable projection, Radon-Nikodym derivatives, representation of martingales, *Z. Wahrsch. verw. Geb.* **31**, 235–253.

J. Jacod and A.N. Shiryaev (1987), *Limit Theorems for Stochastic Processes*, Springer-Verlag, Berlin.

N. Jewell (1982), Mixtures of exponential distributions, *Ann. Statist.* **10**, 479–484.

E.L. Kaplan and P. Meier (1958), Non-parametric estimation from incomplete observations, *J. Amer. Statist. Assoc.* **53**, 457–481, 562–563.

R.L. Karandikar (1983), Multiplicative stochastic integration, pp. 191–199 in: V. Mandrekar and H. Salehi (eds), *Prediction Theory and Harmonic Analysis*, North-Holland, Amsterdam.

N. Keiding and R.D. Gill (1990), Random truncation models and Markov processes, *Ann. Statist.* **18**, 582–602.

J. Kiefer and J. Wolfowitz (1956), Consistency of the maximum likelihood estimator in the presence of infinitely many nuisance parameters, *Ann. Math. Statist.* **27**, 887–906.

D.E. Knuth (1981), *The Art of Computer Programming, vol. 2: Seminumerical Algorithms*, Addison-Wesley.

M.J. van der Laan (1993a), General identity for linear parameters in convex models with applications to efficiency of the (NP)MLE, Preprint 765, Dept. Math., Univ. Utrecht.

M.J. van der Laan (1993b), Efficiency of the NPMLE in the line segment problem, Preprint 773, Math. Inst., Univ. Utrecht.

M.J. van der Laan (1993c), Repairing the NPMLE with application to the bivariate censoring model, Preprint, Dept. Math., Univ. Utrecht.

G.M. Laslett (1982a), The survival curve under monotone density constraints with application to two-dimensional line segment processes, *Biometrika* **69**, 153–160.

G.M. Laslett (1982b), Censoring and edge effects in areal and line transect sampling of rock joint traces, *Math. Geol.* **14**, 125–140.

B.Ya. Levit (1990), Approximately integrable linear statistical models in non-parametric estimation, Tech. Rep. 90–37C, Dept. Statist., Purdue Univ.

M. Luby (1993), *Pseudo-Randomness and Applications*, Princeton Univ. Press.

J.S. MacNerney (1963), Integral equations and semigroups, *Illinois J. Math.* **7**, 148–173.

P.R. Masani (1981), Multiplicative partial integration and the Trotter product formula, *Adv. Math.* **40**, 1–9.

D. Mauro (1985), A combinatoric approach to the Kaplan-Meier estimator, *Ann. Statist.* **13**, 142–149.

P. Meier (1975), Estimation of a distribution function from incomplete observations, pp. 67–87 in: J. Gani (ed.), *Perspectives in Probability and Statistics*, Applied Probability Trust, Sheffield.

R.E. Miles (1974), On the elimination of edge-effects in planar sampling, pp. 228–247 in: E.F. Harding and D.G. Kendall (eds.), *Stochastic Geometry* (a tribute to the memory of Rollo Davidson), Wiley, New York.

S.A. Murphy (1993), Consistency in a proportional hazards model incorporating a random effect, *Ann. Statist.* **21**, to appear.

V.N. Nair (1981), Plots and tests for goodness of fit with randomly censored data, *Biometrika* **68**, 99–103.

V.N. Nair (1984), Confidence bands for survival functions with censored data: a comparative study, *Technometrics* **14**, 265–275.

G. Marsaglia and A. Zaman (1991), A new class of random number generators, *Ann. Appl. Probab.* **1**, 462–480.

W. Nelson (1969), Hazard plotting for incomplete failure data, *J. Qual. Technol.* **1**, 27–52.

G. Neuhaus (1992), Conditional rank tests for the two-sample problem under random censorship: treament of ties, in: Vilaplana (ed.), *V Proceedings Statistics in the Basque Country.*

G. Neuhaus (1993), Conditional rank tests for the two-sample problem under random censorship, *Ann. Statist.* (to appear).

G.G. Nielsen, R.D. Gill, P.K. Andersen, and T.I.A. Sørensen (1992), A counting process approach to maximum likelihood estimation in frailty models, *Scand. J. Statist.* **19**, 25–43.

G. Peano (1888), Intégration par séries des équations différentielles linéaires, *Math. Ann.* **32**, 450–456.

J. Pfanzagl (1988), Consistency of maximum likelihood estimators for certain nonparametric families, in particular: mixtures, *J. Statist. Planning and Inference* **19**, 137–158.

D. Pollard (1984), *Convergence of Stochastic Processes*, Springer-Verlag, New York.

D. Pollard (1990), *Empirical processes: Theory and Applications*, Regional conference series in probability and statistics **2**, Inst. Math. Statist., Hayward, California.

R.L. Prentice and J. Cai (1992a), Covariance and survivor function estimation using censored multivariate failure time data, *Biometrika* **79**, 495–512.

R.L. Prentice and J. Cai (1992b), Marginal and conditional models for the analysis of multivariate failure time data, pp. 393–406 in: J.P. Klein and P.K. Goel (eds), *Survival Analysis: State of the Art*, Kluwer, Dordrecht.

P. Protter (1990), *Stochastic Integration and Differential Equations (a New Approach)*, Springer.

M. Rabin (1979), Digitalized signatures and public key functions as intractable as factorization, Tech. Rep. 212, Lab. Comp. Sci., Mass. Inst. Tech.

R. Rebolledo (1980), Central limit theorems for local martingales, *Z. Wahrsch. verw. Geb.* **51**, 269–286.

J.A. Reeds (1976), *On the Definition of von Mises Functionals*, PhD thesis, Research Rept. S–44, Dept. Statist., Harvard Univ.

N. Reid (1981), Influence functions for censored data, *Ann. Statist.* **9**, 78–92.

A. Rényi (1953), On the theory of order statistics, *Acta Math. Acad. Sci. Hungar.* **4**, 191–231.

B.D. Ripley (1981), *Spatial Statistics*, Wiley, New York.

B.D. Ripley (1988), *Statistical Inference for Spatial Processes*, Cambridge Univ. Press.

Y. Ritov and J.A. Wellner (1988), Censoring, martingales and the Cox model, *Contemp. Math.* **80**, 191–220.

L.A. Santaló (1976), *Integral Geometry and Geometric Probability*, Encyclopedia of Mathematics and Its Applications, vol. 1, Addison-Wesley.

G.R. Shorack and J.A. Wellner (1986), *Empirical Processes*, Wiley, New York. Corrections and changes: Tech. Rep. 167, Dept. Statist., Univ. Washington (1989).

M. Stein (1990), A new class of estimators for the reduced second moment measure of point processes, Tech. Rep. 278, Dept. Statist., Univ. Chicago.

D. Stone, D.C. Kamineni, and A. Brown (1984), Geology and fracture characteristics of the Underground Research Laboratory lease near Lac du Bonnet, Manitoba, Tech. Rep. 243, Atomic Energy of Canada Ltd. Research Co.

D. Stoyan, W.S. Kendall, and J. Mecke (1987), *Stochastic Geometry and its Applications*, Wiley, Chichester.

W. Stute and J.-L. Wang (1993), The strong law under random censorship, *Ann. Statist.*

B.W. Turnbull (1976), The empirical distribution function with arbitrarily grouped, censored and truncated data, *J. Roy. Statist. Soc. (B)* **38**, 290–295.

A.W. van der Vaart (1991a), Efficiency and Hadamard differentiability, *Scand. J. Statist.* **18**, 63–75.

A.W. van der Vaart (1991b), On differentiable functionals, *Ann. Statist.* **19**, 178–204.

A.W. van der Vaart (1993), New Donsker classes, Preprint, Dept. Math., Free Univ. Amsterdam.

A.W. van der Vaart and J.A. Wellner (1993), *Weak Convergence and Empirical Processes*, IMS Lecture Notes-Monograph Series (to appear).

V. Volterra (1887), Sulle equazioni differenziali lineari, *Rend. Acad. Lincei (Series 4)* **3**, 393–396.

J.G. Wang (1987), A note on the uniform consistency of the Kaplan-Meier estimator, *Ann. Statist.* **15**, 1313–1316.

J.-L. Wang (1985), Strong consistency of approximate maximum likelihood estimators with applications in nonparametrics, *Ann. Statist.* **13**, 932–946.

D. Wang and A. Compagner (1993), On the use of reducible polynomials as random number generators, *Math. Comp.* **60**, 363-374.

J.A. Wellner (1993), The delta-method and the bootstrap, Preprint, Dept. Statist., Univ. Washington.

B.J. Wijers (1991), Consistent non-parametric estimation for a one-dimensional line segment process observed in an interval, Preprint 683, Dept. Math., Univ. Utrecht.

S. Yang (1992), Some inequalities about the Kaplan-Meier estimator, *Ann. Statist.* **20**, 535–544.

S. Yang (1993), A central limit theorem for functionals of the Kaplan-Meier estimator, Preprint, Dept. Math., Texas Tech. Univ.

A.C. Yao (1982), Theory and applications of trapdoor functions, *Proc. 23rd IEEE Symp. Found. Comp. Sci.*, 458–463.

Z. Ying (1989), A note on the asymptotic properties of the product-limit estimator on the whole line, *Statist. Probab. Letters* **7**, 311–314.

W.R. van Zwet (1993), Wald lectures on the bootstrap, in preparation.