KM 36(1991)
pag 61 - 76

The Analysis of Multivariate Censored Survival Times

Richard D. Gill

ABSTRACT. A very informal discussion is given of the problems of multivariate survival analysis. A recent proposal of Dorota Dabrowska for estimating a multivariate survival function with censored data is explained. The estimator is based on a representation of the survival function in terms of iterated odds ratio measures. The representation generalizes the one-dimensional representation of the survival function in terms of its hazard function, leading to the familiar product-limit estimator for ordinary censored data.

[This is a rough English translation of the text of my lecture at prof. G. J. Leppink's farewell symposium, given then (of course!) in fluent Dutch. However it is also a completion of that lecture so I hope that those who were present at the symposium, and especially Gerard Leppink himself, will find it interesting to read on to the end.]

I am very pleased indeed also to be speaking at this symposium to honour professor Leppink, but I must admit straight away that my relationship with him has been perhaps only a superficial one. When I arrived in Utrecht two years ago he had carefully and quietly withdrawn himself from the scene, leaving me very free to go my own way; happy to give me helpful advice but not bothered if I did not take it! However the relationship does have a prehistory: still at Cambridge but thinking of coming to the Netherlands back in 1973, I sent a deluge of letters to (I hope) all the statistical departments in the country. From the academic world I got two very hopeful replies: one from Utrecht and one from the then Mathematical Centre, Amsterdam. This resulted in a very pleasurable conversation with Gerard Leppink in his office at the Mathematical Institute (that summer was largely spent in Utrecht; I also spent much time in the maths library and in the hortus botanicus). Unfortunately there would not be an opening in Utrecht for a half year or so and nothing was certain; the Mathematical Centre had more concrete hope to offer so it was there I went in 1974. But I always preserved a very positive memory of that conversation and it made me really happy to in a sense 'come home' to Utrecht and succeed him there.

The other speakers have dwelt on many aspects of a statistican's life. In particular especially this morning's talks have concentrated on the relationship with real scientists working in applied fields, as a source of pleasure and motivation (and if not these, then at least of good stories to tell to your mathematical colleagues later). My talk will concentrate on another equally important aspect: doing real mathematics, and the opportunities to use it in an original way, for solving or understanding real life problems, is the other equally rewarding side of our profession. As de Kroon said this morning: after such an introductory lecture we tend to quickly forget the fine words and start to introduce the Banach spaces! So here they come!

Mathemathisch Instituut, Rijksuniversiteit Utrecht, Postbus 80.010, 3508 TA Utrecht.
Tel: 030-533763. Email: gill@math.ruu.nl.

The analysis of multivariate censored data is a notoriously tough but in some ways also dubious problem. I avoided getting involved in this for a long time, suspecting that the problem is not a real life problem, but an academic one, invented by mathematical statisticians looking for subjects to write papers on; and also believing that the problem is a dirty one, without mathematically nice (i.e., æsthetically pleasing) solutions. However recent work by Dorota Dabrowska (1988) has changed my mind on the second score; and I would like to tell you a bit about her contribution and my own work building on hers, together with colleagues in Amsterdam and Utrecht. She must have the credit for a really original new idea (and though many papers get written and published, truly original ideas are few and far between!). I hope I can make this idea seem natural and simple to you; and if I succeed, that will have been my own main contribution to this field. It took me a long time indeed to understand it. As for applications—I do now believe there are many applications. The nonparametric estimation of a multivariate survival function will not usually be the end-point of a statistical analysis, but it can play a crucial role in model building and testing on the way to a definitive analysis.

In one dimensional time (univariate survival data) I would make the grand claim that *the counting process approach*—modelling the conditional intensity in time of occurence of the events under study, and using the intimately connected martingale tools in the mathematical analysis of the associated statistical procedures—is *the* right mathematical approach for understanding classical methods from survival analysis: the Kaplan-Meier estimator, the log rank test, the Cox regression model, and so on. Moreover (and this is how we can see it is *right*) it leads to a far reaching and practically extremely important generalization of classical survival analysis to what one could call 'event history analysis'. But in multidimensional time—silence. The problems have been lying around for a long time now; there are published and unconvincingly analysed data-sets to play with; but so far only solutions in a one-dimensional spirit: Markov and semi-Markov models for modelling the different stages in an individual's life history; frailty models, modelling dependence between survival times of biologically related individuals by assuming that all dependence occurs through a latent variable (called frailty) summarizing the family's shared genetic make-up; etc. In these approaches the different time variables for one individual or one family unit are all firmly embedded in one-dimensional 'real-time' and a one-dimensional analysis is made.

But these problems are important, and even if the preferred analysis will be Markov, semi-Markov or frailty oriented in one-dimensional time, one would still like to judge the goodness of fit of such (rather restrictive) models against a wider multivariate background. In the biostatistical area there are applications to matched pair and litter-matched survival and carcinogenisis trials; studies of association of survial times in families, or of different organs or treatments of the same individuals; however I recently also came across a great deal of applications in astronomy, where radiation emission of certain quasars is measured in different wave lengths, there being a background radiation of varying amount per wave-length which leads to multivariate *left* censored observations rather that the usual *right* censored observations common in survival analysis. (Turning the overhead-sheet 180 degrees transforms bivariate left censored data into bivariate right censored data, so at least

for a mathematician it is clear that this could be the same problem). If you are interested in such examples, you will find sorted references at the end of the paper.

What is so special about one-dimensional time? Think about a single survival time $T > 0$ with survival function

$$S(t) = \Pr(T > t),$$

and cumulative hazard

$$\Lambda(t) = \int_{u=0}^{u=t} \frac{-S(\mathrm{d}u)}{S(u-)},$$

from which one can reconstruct the survival function S by the *product-integral* (the notation is supposed to be suggestive, so use your imagination)

$$S(t) = \prod_{u=0}^{u=t} (1 - \Lambda(\mathrm{d}u)).$$

If you are unfamiliar with these ways of writing things, you had better start getting used to them now! Here are the intuitions behind the formulas: think of '$\mathrm{d}t$' (and similarly '$\mathrm{d}u$') not just as a (very small) length of time but also as the corresponding small time interval $[t, t+\mathrm{d}t)$. Then $-S(\mathrm{d}t)$ (the minus sign because S is decreasing) can be interpreted as the probability $\Pr(T \in \mathrm{d}t)$, $S(t-)$ as $\Pr(T \geq t)$, and $S(t)$ as $\Pr(T \geq t + \mathrm{d}t)$. Then the hazard assigned to the small interval $\mathrm{d}t$, which I write as $\Lambda(\mathrm{d}t)$, is just

$$\Lambda(\mathrm{d}t) = \frac{-S(\mathrm{d}t)}{S(t-)} = \frac{\Pr(T \in \mathrm{d}t)}{\Pr(T \geq t)} = \Pr(T \in \mathrm{d}t \,|\, T \geq t).$$

Thus $1 - \Lambda(\mathrm{d}t) = \Pr(T \geq t + \mathrm{d}t \,|\, T \geq t)$. Now we can get $S(t)$ (the probability of surviving the long time-interval from 0 to $t + \mathrm{d}t$, by multiplying together, over a sequence of small time-intervals $\mathrm{d}u$ partitioning $[0, t+\mathrm{d}t)$, the conditional probabilities $1 - \Lambda(\mathrm{d}u)$ of surviving the right endpoint of each small interval given you have survived the left endpoint. That is exactly what the product-integral $\prod$ does; see Gill and Johansen (1990) for a complete exposition of this neglected object.

Already we see: from hazards you can build survival functions. On the other hand hazards are a *natural* object to describe from an applied point of view. Engineers, doctors, biologists, and so on, have intuition, experience and theoretical knowledge of hazards (or hazard rates). Model building is conveniently done in terms of hazards. (In a moment we will

consider whether *multivariate* hazards are the right objects to use in building multivariate survival functions).

Not only are hazard rates natural from the point of view of model building but they also have at least two other characteristics, which lead to the great succes of counting process methods: they are undisturbed by censoring; and they are fundamentally connected to martingale theory.

Consider the zero-one random variable $1\{T \in \mathrm{d}t\}$ (the indicator variable that the survival time T lies in the interval $\mathrm{d}t$). Think of the random time T as being the time of an event (usually called 'failure') on the time axis drawn above; think about time slowly unrolling, so that in the beginning, the event lies somewhere still unknown in the future, till at some time point the event suddenly occurs; after that, the event lies fixed in time (its position known) behind us. Consider a given time-point t and condition on the past up to that time point (in fact, up to the beginning of the small time interval '$\mathrm{d}t$'). If the event has already happened, nothing more will happen now: the conditional expectation of $1\{T \in \mathrm{d}t\}$ is zero. Otherwise the probability of the event happening is $\Lambda(\mathrm{d}t)$. This is then also the conditional expectation of $1\{T \in \mathrm{d}t\}$. So writing $\mathcal{F}_{t-}$ for 'the past till just before time-point t' we can combine the two cases as

$$\Pr(T \in \mathrm{d}t \mid \mathcal{F}_{t-}) = \mathrm{E}(1\{T \in \mathrm{d}t\} \mid \mathcal{F}_{t-}) = \Lambda(\mathrm{d}t) \cdot 1\{T \geq t\}.$$

Add censoring into this model. We don't observe T itself but only a 'censored survival time' $\widetilde{T}$ equal to the smaller of T and a certain 'censoring time' C at which the individual in whose life the event of interest happens is lost to observation through other, independent, causes. We do at least know (at time $\widetilde{T}$) which of the two events ('failure', 'censoring') has occured. Let Δ be the 'censoring indicator' (really a 'failure indicator') equal to 1 if the time $\widetilde{T}$ is actually the failure time T, equal to 0 otherwise. Consider the indicator random variable $1\{\widetilde{T} \in \mathrm{d}t, \Delta = 1\}$ which registers 'observed failure in $\mathrm{d}t$'. Now consider again time unrolling. Again if censoring or failure has happened before time t, 'observed failure in $\mathrm{d}t$' has probablity zero. If neither has yet happened, then if failure were to happen in $\mathrm{d}t$ it would be observed (suppose $\mathrm{d}t$ is so short that if T and C are different at most one is in $\mathrm{d}t$). Moreover by independence, if failure and censoring have not yet happened then failure still has the same chances of happening as in the absence of censoring. So we can again summarize the different cases in the equation

$$\Pr(\widetilde{T} \in \mathrm{d}t, \Delta = 1 \mid \mathcal{F}_{t-}) = \mathrm{E}(1\{\widetilde{T} \in \mathrm{d}t, \Delta = 1\} \mid \mathcal{F}_{t-}) = \Lambda(\mathrm{d}t) \cdot 1\{\widetilde{T} \geq t\}.$$

Note two things, the two things I alluded to above. Censoring hasn't significantly changed the equation; the conditional probability of an observable event is of exactly the same form as without censoring, involving the same 'pre-censoring' hazard. Secondly, and more for the connoisseurs, this equation can be interpreted as stating that the difference $1\{\widetilde{T} \in \mathrm{d}t, \Delta = 1\} - \Lambda(\mathrm{d}t)1\{\widetilde{T} \geq t\}$ is a *martingale increment*—its conditional expectation given $\mathcal{F}_{t-}$ is zero—and adding it up over small intervals partitioning $[0, t + \mathrm{d}t)$ gives a *martingale*. Now martingale theory is one of the most beautiful, deep, and most powerful parts of (pure) probability theory, so one should not be surprised that this martingale

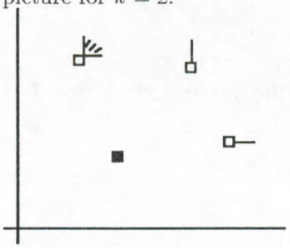
property connecting various basis ingredients of our problem—the hazard measure and the counting process registering the observed occurence of failure—can be exploited and has far reaching consequences. Moreover, if you are feeling uncomfortable about how I manipulate 'dt's then you should also be pleased to hear that martingale theory completely justifies these heuristics. (If you want to know what a martingale *really* is then consult Rabelais (1542), *Gargantua*, book I, chapter 20).

Now let's turn to multivariate time. We will use the same ideas to study why martingale techniques are not available, but also discover there is hope for some progress using other tools. I would like you to consider the natural k-variate extension of the classical random censorship model, and we will study the problem of nonparametric estimation of the multivariate survival function. In one dimension this problem is solved by the famous Kaplan-Meier estimator which makes use of the representation $S(t) = \prod_0^t (1 - \Lambda(du))$, simply plugging in the almost as well known Nelson-Aalen estimator of Λ: we estimate $\Lambda(du)$ by the number of observations *known* to lie in du, divided by the number of observations *known* to be greater than or equal to u.

Here follows the set-up and (very important) notation. Let $T = (T_1, \ldots, T_k)$ and $C = (C_1, \ldots, C_k)$ be independent k-vectors of (positive) survival times and censoring times of one individual. For each $i = 1, \ldots, k$ define $\widetilde{T}_i = T_i \wedge C_i$ and $\Delta_i = 1\{T_i \leq C_i\}$; these are the censored survival time and the (non)censoring indicator for the i'th time-dimension ($\wedge$ means 'minimum'). Put these together into two k-vectors $\widetilde{T}$ and Δ. Define $S(t) = \Pr(T \gg t)$ where t is now also a k-vector of nonnegative components, and '$\gg$' for vectors means *componentwise* '$>$'. (We use the special symbol $\gg$ to emphasize strict inequality of *all* components. The inequality sign '$\geq$' for vectors will be interpreted as componentwise $\geq$). Thus $S(t)$ is the probability that the survival time T lies *strictly inside* the 'upper, right' orthant of $\mathbb{R}_+^k$ whose corner is placed at the point t in that space.

Our data will consist of n independent and identically distributed realisations of $(\widetilde{T}, \Delta)$, and our problem is to estimate the survival function S nonparametrically.

Each of the n observations can be visualised as a point in $\mathbb{R}_+^k$ at the corresponding $\widetilde{T}$ with some mark on the point indicating the value of Δ (there are 2^k possible different values). This mark shows where the underlying T must be: at the point itself, on a half-line starting at the point, in an orthant located at the point, ..., and so on. Here is the picture for $k = 2$:

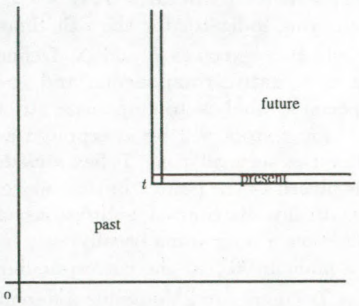


The 'points' are given as little boxes; also 'dt' is also going to be a little box rather than a little interval from now on. If the box is filled the observation lies at that point. If the box is empty then the actual observation lies on a half-line, orthant, ..., starting at that

point but not including it. When $k = 2$ then Δ can take four different values $(0,0)$, $(0,1)$, $(1,0)$, $(1,1)$. The picture shows one observation of each type.

What made life so beautifully simple in one-dimension was that there was just one way to get from time 0 to time t; and that one meets everything on the way in between. In higher dimensions there is no canonical way. In fact a *different* ad hoc estimator can be built on each 'path' from 0 to t.

Not only is there no canonical way to get from 0 to t, there is also no canonical way to define *past*, *present* and *future* at 'time' t. However motivated by our problem—estimate the probability of 'failure' inside the upper right orthant located at t—it seems sensible to take as 'future' everything inside that orthant, as 'past' everything outside, and as 'present' the border region between past and future. The present is now rather more complicated than in the one-dimensional case as the next picture shows:



However this picture is extremely important and we will study it in different ways. Firstly the picture suggests how to introduce *multivariate hazards*; in the two-dimensional case for instance we can define

$$\Lambda(\mathrm{d}t) = \Pr(T \in \mathrm{d}t \,|\, T \geq t),$$

the probability that T lies in the little box in the corner of the 'present' given that T is somewhere in the present or future; but also we can define

$$\Lambda_{1|2}(\mathrm{d}t_1 \,|\, t_2) = \Pr(T_1 \in \mathrm{d}t_1 \,|\, T \geq t),$$

$$\Lambda_{2|1}(\mathrm{d}t_2 \,|\, t_1) = \Pr(T_2 \in \mathrm{d}t_1 \,|\, T \geq t);$$

these are the conditional probabilities of the left strip and the bottom strip respectively of 'the present' (both including the corner), given T will lie in present or future (supposing the

x-axis is t_1 and the y-axis is t_2). With these three 'overlapping' conditional probabilities we can determine the conditional probabilities of each of the four disjoint parts of present and future, and conversely.

In higher dimensions we can introduce similar things but there are now many more of them ($2^k - 1$ in dimension k). But I claim that it actually makes things better to be general! The pay-off will actually be: shorter formulas with better interpretable terms, if we take the trouble to find a convenient notation. So here is one (we are going to abuse standard notation even more than we have been doing, but I think it is going to be worthwhile): let $E = \{1, \ldots, k\}$, the set of variables (time dimensions) under consideration. We will use A, B, C for subsets of E (the new C no relation to the old one); $\emptyset$ is the empty set. These are all subsets of variables. If t is a k-vector, then by t_A I mean the subvector of length $|A|$ (the number of elements of A), obtained by collecting together the t_i with $i \in A$. Now I would like to introduce a complete set of multivariate conditional hazards (hazard measures), defined for every $\emptyset \subset A \subseteq C \subseteq E$:

$$\Lambda_{A|C}(\mathrm{d}t_A \,|\, t_C) = \Pr(T_A \in \mathrm{d}t_A \,|\, T_C \geq t_C).$$

The infinitesimal elements of these measures are the probabilities, conditional on the 'C variables' taking a value in the orthant (including faces) located at t_C, that the 'A variables' (a smaller set of variables) lie in the corresponding border between past and future; this is a hyperface ($|A| = 1$), a corner($A = C$), or something in between, depending on how many variables we are talking about. (In line with the two-dimensional notation I should have written something painful like $\Lambda_{A|C\setminus A}(\mathrm{d}t_A|t_{C\setminus A})$ but notation is meant to be abused; it's better to keep things looking easy when they indeed are!)

Now we can, keeping the last two pictures continually in mind, identify a number of key features of our problem, and then put everything together into a beautiful representation of a multivariate survival function in terms of its multivariate conditional and marginal hazard measures; and an accompanying natural estimator. On the way I will argue that rather than hazard measures, so-called *iterated odds-ratio measures*, or L-measures (Leppink measures?) are the things to concentrate on. The representation, the estimator and the L-measures were all introduced by D. Dabrowska but in a rather different way to the way done here.

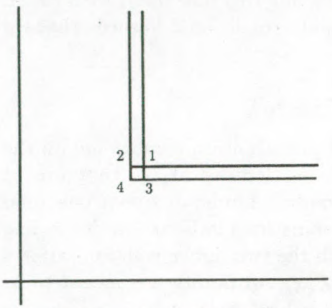
The first feature is that, looking back at our data, for each observation $(\widetilde{T}, \Delta)$, given that $\widetilde{T}$ lies in an orthant (present and future) located at t, we can decide in which of the 2^k subregions the underlying failure time T must lie. Simply move the marked boxes about, keeping the box inside 'present and future', to convince yourself of this. The crucial point is that if $\widetilde{T}$ lies somewhere on the border, we can see from the corresponding Δ_i's if the underlying T_i lie on the border too or not. (In formulas, given $\widetilde{T}_C \geq t_C$, the events $\{T_A \in \mathrm{d}t_A\}$ and $\{\widetilde{T}_A \in \mathrm{d}t_A, \Delta_A = 1_A\}$ coincide; '=' being interpreted componentwise and 1_A a vector of one's).

Next, given $\widetilde{T}_C \geq t_C$, T_C has the same distribution as the distribution of T_C given just $T_C \geq t_C$. This follows trivially from the independence of T and the censoring variables (which were also called C, I am sorry!). So the conditional hazards $\Lambda_{A|C}(\mathrm{d}t_A|t_C)$, derived

from our unknown survival function S (I won't bother you with the formula), are the same as the conditional probabilities (given $\widetilde{T}_C \geq t_C$) of observable events $\widetilde{T}_A \in \mathrm{d}t_A, \Delta_A = 1_A$.

So if we do not worry about the fact that the survival distribution is really continuous, maybe, while our data is just n discrete points, we could set about estimating these conditional hazards by just forming ratios of numbers of observations 'known to lie' in certain regions of $\mathbb{R}_+^k$.

Now I want to move away from hazards to something closely related called odds ratios. The four corners of the little box '$\mathrm{d}t$' in the previous picture are at the same time the defining corners of four overlapping orthants (all 'upper, right' orthants, got from one another by small shifts). Call the corners and the associated orthants 1,2,3,4, say (top right, top left, bottom right, bottom left).



Now look at the odds ratio

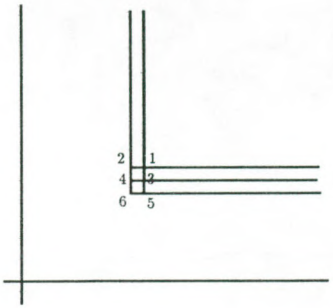
$$\frac{1/2}{3/4} = \frac{\Pr(T \in {}_1\llcorner)/\Pr(T \in {}_2\llcorner)}{\Pr(T \in {}_3\llcorner)/\Pr(T \in {}_4\llcorner)}$$

that is, the odds on 1 to 2, divided by the odds on 3 to 4. (We'll stick to $k = 2$ for the time being; but in higher dimensions we'll just recursively take ratios of odds ratios in a way you may already be able to guess). Again we note a few key features.

Firstly we can estimate *conditional* odds ratios—conditional meaning, all probabilities being conditional on $T \geq t$, i.e. $T \in {}_4\llcorner$—just count observations 'known' to lie in regions 1, 2, 3, 4 and form the ratio of ratios $(1/2)/(3/4)$.

Secondly, *conditional* odds ratios coincide with *unconditional* odds ratios—get from unconditional to conditional by dividing all four probabilities by the probability of the conditioning event, $\Pr(T \in {}_4\llcorner)$. It cancels out.

Thirdly, (unconditional) odds ratios are *multiplicative*: if we multiply together the odds ratios for two adjacent small boxes, e.g., 1,2,3,4 and 3,4,5,6 in the picture below , we get the odds ratio for the union 1,2,5,6.



$$\frac{1/2}{3/4} \cdot \frac{3/4}{5/6} = \frac{1/2}{5/6}.$$

Usually the odds ratio for $\mathrm{d}t$ is going to be very close to 1: most probability lies in future, only a little in the present. Put another way, the odds ratio for 'dt' minus 1 will be infinitesimal and we will define the *iterated odds ratio measure*, denoted L, as

$$L(\mathrm{d}t) = \text{iterated odds ratio for } \mathrm{d}t \quad - 1.$$

The word 'iterated' is because I have moved back to dimension k now. In k dimensions we have an odds (ratio)$^{k-1}$. (Dabrowska's L is actually minus mine).

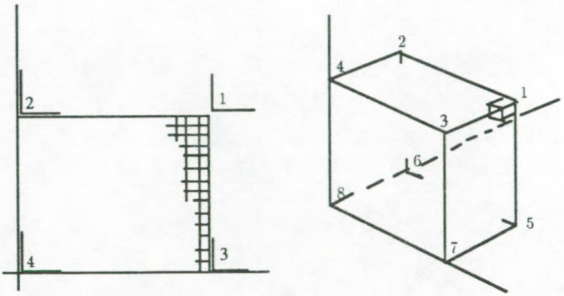
In one dimension the Leppink measure is just minus the one-dimensional hazard measure. In two dimensions one can think of it as being a marginally weighted measure of interaction: in terms of the 2×2 table with cell probabilities a, b, c, d (orthant 1, strip 2, strip 3, box 4) one can write $L(\mathrm{d}t) = (ad - bc)/((a + b)(a + c))$. It should be possible to express the higher order measures in terms of higher order interactions.

With $k = 3$, the box $\mathrm{d}t$ is a small cube with 8 corners, at the same time the corners of 8 overlapping orthants. If 1,2,3,4 are the corners on the top face and 5,6,7,8 the ones on the bottom we calculate the odds ratio *ratio*

$$\frac{1/2}{3/4} \Big/ \frac{5/6}{7/8} = \frac{1 \;\; 4\,6\,7}{2\,3\,5 \;\; 8}.$$

The second way of writing this shows that the odds ratio ratio can be calculated by starting at the very top corner (1), then dividing by all probabilities at corners one step down a rib of the cube (2,3,5), then multiplying by all probabilities one step further down (4,6,7), and finally dividing by the probability at the bottom corner (8).

The multiplicativity tells us that the iterated odds ratio for the large box $[0, t + \mathrm{d}t)$ (with bottom corner at 0 and top corner just outside t) is the product-integral, over that large box, of $1 + \mathrm{d}L$. Now the probabilities of being in the orthants located at the corners of *this* large box are the survival function at t (top corner) together with the survival functions of *subsets* C of the variables evaluated at points t_C, $C \subset E$.



These are easily seen to ocur alternatingly in numerator and denominator of the iterated odds ratio (written as a ratio of two products) as we move down (dropping more and more variables). In order to cancel out all these lower dimensional survival functions it turns out that one must just *multiply* by *every* lower dimensional odds ratio (for faces, edges, ...) of the large box $[0, t+\mathrm{d}t)$ (check this yourself!). This gives us the representation we are looking for:

$$S(t) = \prod_{C \subseteq E} \prod_{u_C \in [0_C, t_C]} (1 + L_C(\mathrm{d}u_C)).$$

The natural estimator of S associated with this representation is obtained by 'estimating' $1 + L_C(\mathrm{d}u_C)$, the iterated odds ratio for the variables with index in C and the small box $\mathrm{d}u_C$, by the corresponding empirical odds ratio based on numbers of observations with the C variables *known* to lie in the various (upper right) orthants located at the corners of $\mathrm{d}u_C$. With n observations there will be at most $n^{|C|}$ points u_C at which the empirical iterated odds ratio is unequal to 1, and where the empirical iterated odds measure (a discrete measure) is unequal to zero. One can recursively build up the values of the estimator of S on the grid of n^k points generated by the observations by simply using the multiplicativity in the representation. Dabrowska rightly calls her estimator *the* multivariate Kaplan-Meier estimator since it generalizes that famous estimator for the case $k = 1$ in a convincingly natural way (unlike previous attempts).

Before going further, let me mention that that these iterated odds ratio measures could be a very important new ingredient in multivariate survival analysis. The measures are measures of dependence. Positive dependence goes with positive L-measures. At independence between two groups of variables, all Leppink measures involving variables from both groups are zero. Well known parametric multivariate survival functions have Leppink measures of particularly simple forms. Modelling dependence could be done by modelling the L-measures. This is as yet unknown territory. It is amusing, when S has a density, to write down the first few L-measures in terms of the conditional hazard *rates* (densities of the $\Lambda_{A|C}$). One finds for $k = 1$ and $k = 2$:

$$l_1 = -\lambda_1, \quad l_{12} = \lambda_{12} - \lambda_{1|2}\lambda_{2|1},$$

or schematically (just noting the indices to the left of the '|'):

$$l_1 = -(1),$$

$$l_{12} = +(12) - (1)(2),$$

and then for $k = 3$, $k = 4$:

$$l_{123} = -(123) + (12)(3) + (13)(2) + (23)(1) - 2(1)(2)(3),$$

$$\begin{aligned} l_{1234} = &+ (1234) - (123)(4) - (124)(3) - (134)(2) - (234)(1) \\ &- (12)(34) - (13)(24) - (14)(23) \\ &+ 2\{(12)(3)(4) + (13)(2)(4) + (14)(2)(3) + (34)(1)(2) + (24)(1)(3) + (23)(1)(4)\} \\ &- 6(1)(2)(3)(4). \end{aligned}$$

There is a simple algorithm for getting from one dimension to the next, but I do not know how to write down directly the coefficients in e.g. $l_{12345678910}$. An interpretation in terms of interactions in 2^k tables would be very useful.

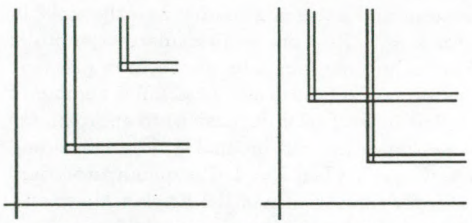
In a moment I will describe something of the mathematics behind Dabrowska's representation and the estimator, but perhaps it is more important first to say something about the results. You will be relieved to hear that the representation can be proven to be true. This is not trivial at all and involves the rather delicate problem of defining correctly the Leppink measures. The estimator does have all (or nearly all) the nice properties you could want. It is consistent, asymptotically normal at the usual $\sqrt{n}$ rate; one can in fact calculate asymptotic variances and they seem close at surprisingly small sample sizes to actual variances. It is asking for trouble to mention a specific sample size but I'll do it all the same: for $k = 2$, there is at $n = 100$ already excellent agreement between asymptotic theory and real life, for typical censoring patterns and not going too far into the tails; in fact, no worse than the situation would be for $k = 1$. How one can calculate asymptotic variances I will mention in a moment, but here is how one can *estimate* them in practice: pretend S is discrete so that the estimator is a ratio of products of a *fixed* finite number of counts of observations; use the delta method (first order Taylor expansion) to approximate with an expression *linear* in these counts; then estimate the multinomial type variances and covariances in the naive way. In case you are worried, when $k = 1$ this quaint procedure yields exactly the famous Greenwood estimator for the variance of the Kaplan-Meier estimator. The procedure sounds horrific, but one could program a computer algebra package to write the program to do the calculations. Alternatively one may bootstrap, in several ways: 'empirically', resampling from the data, or 'semiparametrically', generating artifical censored data by taking failure times from the estimated survival distribution and censoring times from the correspondingly estimated censoring distribution. These two bootstrap procedures are the same for $k = 1$ but otherwise are different.

The above description of how to estimate variances in practice serves also as description of an algorithm for getting theoretical formulas for theoretical efficiency studies:

pretend the product integral is an ordinary discrete product, so the estimator is indeed a ratio of products of numbers of observations; linearize by the delta-method; then 'de-discretise' by replacing all sums over time-points by integrals and products over time-points by product-integrals. The result will be a legal expression, a linear functional of empirical processes for the data, and its asymptotic Gaussian distribution (in particular its variance), will be the same as that of the estimator itself. It is a mathematical theorem that this formula manipulation algorithm, of which I have not given complete details here, does work. Again one could program a computer algebra package to produce and numerically evaluate expressions for asymptotic variances, given formulas for the failure and censoring distributions.

I said that the estimator had nearly all the properties one could want. What it misses is *efficiency* though the possible efficiency loss with respect to a fully efficient estimator (none is known at present) is hopefully small. A lot more work is needed here. Despite the beauty of L-measures the problem remains a nasty one and an explicit expression for an asymptotically optimal variance does not exist (if it did, we would probably know an efficient estimator too). In one special case one can do exact calculations: when all components of the failure variables and of the censoring variables are independent. Because of independence, the 'information operator' which has to be inverted in order to determine the optimal asymptotic variance maps products to products, and its inverse can be described in terms of the (known) 'one-dimensional' solution. At the same time one can simplify the asymptotic variance of the Dabrowska estimator to something manageable—and one finds the same answer! By an amazing coincidence the estimator is fully efficient 'at complete independence'. (If you knew you had independence, you wouldn't use the estimator. The conclusion to be drawn is that with *mild* dependence, you are probably *close* to efficiency).

Asymptotic variances are not pretty. There is asymptotic independence (nice) coming from 'nested orthants', but dependence from 'crossed ones', as you can easily imagine. In terms of martingale theory, there are weak martingales around (which are not of much use) but no strong ones in this problem.



The mathematical derivation of these results goes by rewriting the iterated odds ratio measures in terms of the conditional hazard measures we looked at earlier. Here is the (amazing) formula for this:

$$
\begin{aligned}
S(t) &= \prod_{C \subseteq E} \prod_{(0_C, t_C]} (1 + L_C(\mathrm{d}u_C)) \\
&= \prod_{C \subseteq E} \prod_{(0_C, t_C]} \prod_{B \subseteq C} \Pr\Big(T_B \geq u_B + \mathrm{d}u_B | T_C \geq u_C\Big)^{(-1)^{|C \setminus B|}} \\
&= \prod_{C \subseteq E} \prod_{(0_C, t_C]} \prod_{B \subseteq C} \Big(1 + \sum_{A \subseteq B} (-1)^{|A|} \Pr(T_A \ll u_A + \mathrm{d}u_A | T_C \geq u_C)\Big)^{(-1)^{|C \setminus B|}} \\
&= \prod_{C \subseteq E} \prod_{(0_C, t_C]} \prod_{B \subseteq C} \Big(1 + \sum_{A \subseteq B} (-1)^{|A|} \Lambda_{A|C}(\mathrm{d}u_A | u_C)\Big)^{(-1)^{|C \setminus B|}}.
\end{aligned}
$$

For the estimator we can write a corresponding formula. I won't explain the derivation but really there is nothing to it: write out the definition of the (conditional) iterated odds ratio for the box $\mathrm{d}u_C$, and use the inclusion-exclusion principle to express conditional probabilities of orthants in terms of the complementary events (strips along borders) $T_i < u_i + \mathrm{d}u_i$.

Now one can consider the estimator as a composition of three mappings: from the empirical distribution of the data to the empirical hazard measures, from these to the L-measures, and from these to their product-integrals. The hard part of this is from the hazard measures to the L-measures:

$$
L_C(\mathrm{d}u_C) = \prod_{B \subseteq C} \Big(1 + \sum_{A \subseteq B} (-1)^{|A|} \Lambda_{A|C}(\mathrm{d}u_A | u_C)\Big)^{(-1)^{|C \setminus B|}} - 1.
$$

This expression is easy to understand intuitively but unfortunately mathematically it does not yet make sense at all. Combine and expand the right-hand side as a ratio of two huge polynomials in '$\Lambda_{A|C}(\mathrm{d}u_A | u_C)$'; then we see that we are adding and multiplying elements of measure on all kinds of different spaces as if this was legal. But certainly expressions occur which do not have any standard meaning at all and in particular neither numerator or denominator is 'homogenous'; measures of different 'dimension' are added together.

However if one considers this 'formal' rational function in the conditional hazards, it turns out that a small number of algebraic transformations can be made which turn the formula from nonsense into something completely legal. In particular the embarrassing elements in numerator and denominator which are of less than full dimension turn out to coincide exactly, so they can be 'cancelled out' (they don't actually disappear but they can be made harmless). Just three tricks are needed (whatever k): dividing out terms of the wrong dimension, replacing 'one plus element of measure' by 'one plus atom of measure', and replacing incompatible products of transition measures by products of Radon-Nikodym derivatives and dominating measures.

Here is how it goes in dimension 2:

$$
\begin{aligned}
L(\mathrm{d}t) &= \frac{1-\Lambda_{1|2}(\mathrm{d}t_1)-\Lambda_{2|1}(\mathrm{d}t_2|t_1)+\Lambda(\mathrm{d}t)}{\big(1-\Lambda_{1|2}(\mathrm{d}t_1|t_2)\big)\big(1-\Lambda_{2|1}(\mathrm{d}t_2|t_1)\big)} - 1 \\
&= \frac{\Lambda(\mathrm{d}t)-\Lambda_{1|2}(\mathrm{d}t_1|t_2)\Lambda_{2|1}(\mathrm{d}t_2|t_1)}{\big(1-\Lambda_{1|2}(\mathrm{d}t_1|t_2)\big)\big(1-\Lambda_{2|1}(\mathrm{d}t_2|t_1)\big)} \\
&= \frac{\Lambda(\mathrm{d}t_1,\mathrm{d}t_2)-\Big(\frac{\mathrm{d}\Lambda_{1|2}(\cdot|t_2)}{\mathrm{d}\mu_1}(t_1)\frac{\mathrm{d}\Lambda_{2|1}(\cdot|t_1)}{\mathrm{d}\mu_2}(t_2)\Big)\mu_1(\mathrm{d}t_1)\mu_2(\mathrm{d}t_2)}{\big(1-\Lambda_{1|2}(\{t_1\}|t_2)\big)\big(1-\Lambda_{2|1}(\{t_2\}|t_1)\big)}.
\end{aligned}
$$

These three tricks convert a reasonably neat and moreover nicely interpretable but *illegal* formula into a huge, ugly, but *legal* one. In fact for general k one can't explicitly write out the result at all. How can one establish properties of such an object? The aim is to get statistical properties of the estimator from analytical properties of the transformations (empirical data to hazards to L-measures to survival function). We want to prove continuity and differentiablity of these mappings (considered as mappings between various Banach spaces—I told you they were coming!); this will give us to start with the validity of the representation itself (continuity gets us from discrete, where the representation is 'trivially' true, to continuous), also consistency; differentiability will give us asymptotic normality and correctness of the bootstrap (we need a kind of continuous differentiability for this).

I can tell you even less about this part of the project, but just assure you again that one can get further; see the references. The key is not to worry what the exact expression for L is (after legalisation), but just to be able to describe the kinds of terms which can be met with after further expansion. It turns out that the terms occuring share a great deal of structure and can be dealt with 'abstractly' all in the same way. Apart from integration by parts, only three tricks (another three) are needed: a 'Helly-Bray' trick to deal with convergence of integrals when you have no right to expect it, a 'fraction-splitting trick' for dealing with unpleasant denominators, and 'd-Delta interchange' for switching between integrating measures and atoms of measures.

I hope that you can begin to taste the flavour: a curious mixture of analysis, algebra and probability. The mathematics is algorithmic; a big part of it consists of rewriting rules for formula manipulation. There seems to be scope for a systematic approach using ideas from computer algebra. Alternatively perhaps non-standard analysis could be used to switch more easily between discrete and continuous. But this would be a rather different research project.

References

The multivariate product-limit estimator:

D. M. Dabrowska (1988), Kaplan Meier estimate on the plane, *Ann. Statist.* **16** 1475–1489.

D. M. Dabrowska (1989), Kaplan Meier estimate on the plane: weak convergence, LIL and the bootstrap, *J. Multivar. Anal.* **29** 308–325.

M. van der Laan (1990), *Dabrowska's Multivariate Product Limit Estimator and the Delta-Method*, Master's Thesis, Dept. of Maths., Utrecht Univ.

D. M. Bakker (1990), *Two Nonparametric Estimators of the Survival Function of Bivariate Right Censored Observations*, Report BS-R9035, Centre for Mathematics and Computer Science, Amsterdam.

R. D. Gill (1990), *Multivariate Survival Analysis*, to appear in *Proceedings of the 2nd World Congress of the Bernoulli Society and the 53rd Annual IMS Meeting, Uppsala*; preprint 621, Dept. of Maths, Utrecht Univ.

P. J. Bickel, R. D. Gill, and J. A. Wellner (1990), *Inefficient Estimators for Three Multivariate Models*, in preparation.

The functional delta-method, product-integration:

R. D. Gill (1989), Non and semi-parametric maximum likelihood estimators and the von-Mises method (Part I), *Scand. J. Statist.* **16** 97–128.

A. Sheehy and J. A. Wellner (1990a), *Uniform Donsker Classes of Functions*, Report 189, Dept. of Statistics, Univ. of Washington.

A. Sheehy and J. A. Wellner (1990b), *The Delta Method and the Bootstrap for Nonlinear Functions of Empirical Distribution Functions via Hadamard Derivatives*, Report, Dept. of Statistics, Univ. of Washington.

R. D. Gill and S. Johansen (1990), A survey of product-integration with a view toward application in survival analysis, *Ann. Statist.* **18** (in print).

Multivariate hazards, weak martingales:

O. Pons (1996), A test of independence for two censored survival times, *Scand. J. Statist.* **13** 173–185.

Examples from survival analysis:

A. Muñoz (1980), *Nonparametric Estimation from Censored Bivariate Observations*, Tech. Rep. 60, Div. of Biostatistics, Stanford Univ.

L. J. Wei and J. M. Lachin (1984), Two-sample asymptotically distribution-free tests for incomplete multivariate observations, *J. Amer. Statisti. Assoc.* **79** 653–661.

L. Horvath and B. S. Yandell (1988), Bootstrapped multi-dimensional product limit processes, *Austr. J. Statist.* **30** 342–358.

Matched pairs and litter-matched designs:

J. D. Holt and R. L. Prentice (1974), Survival analysis in twin studies and matched pair experiments, *Biometrika* **61** 17–30.

N. Mantel, N. R. Bohidar, and J. L. Ciminera (1977), Mantel-Haenszel analyses of litter-matched time-to-response data, with modification for recovery of interlitter information, *Cancer Res.* **37** 3863–3868.

N. Mantel and J. L. Ciminera (1979), Use of logrank scores in the analysis of litter-matched data on time to tumour appearance, *Cancer Res.* **39** 4308–4315.

Examples from astronomy:

T. Isobe, E. D. Feigelson and P. I. Nelson (1986), Statistical methods for astronomical data with upper limits. II. Correlation and regression, *Astrophys. J.* **306** 490–507.

H. A. Thronson, J. Bally, and P. Hacking (1989), The components of mid- and far-infrared emission from S0 and early-type shell galaxies, *Astron. J.* **97** 363–374.

Martingale methods:

P. K. Andersen and Ø. Borgan (1985), Counting process models for life history data: a review, *Scand. J. Statist.* **12** 97–158.

P. K. Andersen, Ø. Borgan, R. D. Gill, and N. Keiding (1991), *Statistical Models Based on Counting Processes*, Springer, in preparation.

Frailty models:

R. D. Gill (1985), Discussion (pp. 108–109) of: Multivariate generalizations of the proportional hazards model by D. R. Clayton and J. Cuzick, *J. Roy. Statist. Soc. (A)* **148** 82–117.

S. G. Self and R. L. Prentice (1986), Incorporating random effects into multivariate relative risk regression models, pp. 167–177 in: *Modern Statistical Methods in Chronic Disease Epidemiology*, eds. S. H. Moolgavkar and R. L. Prentice, Wiley.

J. P. Klein (1990), *Semiparametric Estimation of Random Effects Using the Cox Model Based on the EM Algorithm*, Dept. of Stat., Ohio State Univ.

G. G. Nielsen, P. K. A. Andersen, R. D. Gill, and T. I. A. Sørensen (1990), *A Counting Process Approach to Maximum Likelihood Estimation in Frailty Models*, in preparation.