GENERALISATION DE L'ESTIMATEUR KAPLAN-MEIER D'UNE LOI DE DUREE DE MAINTIEN EN PRESENCE D'OBSERVATIONS TRONQUEES A GAUCHE EXTENSION A L'ETUDE CONJOINTE DE DEUX DUREES DE MAINTIEN

Didier RULLIERE Actuaire Ecureuil Vie
Daniel SERANT I.S.F.A

Laboratoire de Sciences Actuarielle et Financière,
Université Claude Bernard Lyon 1

ETUDE SOUTENUE PAR LA SOCIETE ECUREUIL VIE

Résumé : L'estimation Kaplan-Meier de la loi d'une variable aléatoire positive est bien connue dans le cas de données pouvant être censurées à droite. Cette estimation a été prolongée aux cas de données pouvant être tronquées à gauche pour des variables discrètes. Turnbull [13].

Dans le cas univarié, cette généralisation est étendue au cas de variables aléatoires dont la loi peut avoir pour support tout sous-ensemble de $\mathbf{R}^+$. La notion d'estimateur cohérent est reprise avec une définition plus large que la définition classique permettant d'établir l'unicité et, sous certaines conditions, l'existence d'un estimateur maximum de vraisemblance généralisant l'estimateur Kaplan-Meier. Cet estimateur est obtenu explicitement sous une forme permettant de comparer les influences respectives des données censurées et tronquées sur l'estimation. L'algorithme de redistribution proposé par Efron [6] est généralisé.

Dans le cas bivarié on démontre, dans le cas particulier de censure univariée, et sous certaines conditions, l'existence (et l'unicité en absence de troncatures) de l'estimateur maximum de vraisemblance de la loi conjointe des deux durées de survie. Un exemple montre qu'il est possible que cette estimation affecte des probabilités irrationnelles à certains points montrant ainsi que toute estimation déduite de l'algorithme redistributif étendu au cas multivarié ne saurait être toujours l'estimation du maximum de vraisemblance. Des conditions de convergence d'algorithmes itératifs sont données en particulier dans le cas où il n'y a pas de données tronquées. Un algorithme itératif est proposé. L'estimation déduite de la généralisation de l'algorithme redistributif peut servir de point initial à l'algorithme itératif proposé.

Mots clés *:* Censored data ; truncated data ; Kaplan-Meier estimator ; Life table ; survival estimation ; product-limit estimator.

Summary : The Kaplan-Meier estimation of a positive random variable distribution is particularly well known for right censored data. This estimation has been extended to truncated data and to discrete random variable by Turnbull [13]. For the univariate case, this generalization has been extended to any random variables. The estimation's consistency notion is considered with a larger definition than the classic one. It provides uniqueness and, under certain conditions, the existence of a maximum likehood estimation generalizing the Kaplan-Meier one. The redistributive algorithm which was first proposed by Efron [6] is extended.

BULLETIN FRANÇAIS D'ACTUARIAT, Vol. 1, N° 2, 1997, pp. 97-114

For the estimation of bivariate survival distribution, we demonstrate in particular the case of univariate censoring and, under certain conditions, the existence (and without truncated data the unicity) of the maximum likehood estimation. An example indicates that this estimation gives non rational probabilities to some points and then shows that any Efron-like estimation, generalized to the bivariate case is not always a maximum likehood estimation.

Convergence conditions for iterative algorithms are given in presence of censoring. An algorithm is put forward. The redistributive algorithm generalization can be the initial point of this algorithm.

Keywords : Censored data ; Truncated data ; Kaplan-Meier estimator ; Life table ; Survival estimation ; Product-limit estimator.

1. INTRODUCTION

Dans ce qui suit le vocabulaire utilisé pour désigner une variable aléatoire positive sera emprunté aux études de durées de survie en un certain état et plus particulièrement aux études de survie humaine. Cependant les résultats obtenus peuvent s'appliquer à d'autres situations telles que, par exemple, l'étude de coûts de sinistres en présence de franchise et de plafond.

Dans le problème de l'estimation de lois de maintien dans un certain état, on est souvent confronté à la présence d'observations incomplètes. Ainsi, dans cette étude, nous considérons comme possibles des durées de maintien qui ne sont que partiellement observées du fait de la sortie prématurée de l'individu de l'expérience (censure droite) ou du fait que certains individus (troncatures gauches) sont observés alors qu'une partie de leur durée de maintien est déjà écoulée au début d'expérience.

Ainsi, lorsqu'on s'intéresse à l'âge au décès d'individus dans une population, on doit tenir compte du fait que certains individus ne sont observés qu'à partir d'un certain âge (ce qui est toujours le cas, si la date de naissance de ces individus est connue). On doit également intégrer le fait que certains individus sont survivants au moment de la fin de l'étude.

De très nombreuses études ont été menées sur les problèmes posés par ces données incomplètes notamment dans le cas où ces données sont incomplètes par censures à droite.

2. L'ESTIMATION DE LA LOI D'UNE DUREE DE MAINTIEN

2-a Notations et données du problème :

Soit une population observée de n individus. Pour l'individu numéro i , $i \in \{1,2,..,n\}$ on note X_i son âge au décès et on observe :

a- l'âge e_i au début de l'observation,

b- l'âge t_i auquel l'individu sort de l'observation :

- soit par décès et alors $X_i=t_i$ et on introduit un indicateur de non censure δ_i qui prend ici la valeur 1
- soit par censure et alors $X_i \geq t_i$ et δ_i est alors égal à 0.

Par convention on conviendra que, en cas d'égalité d'âges d'entrées ou de sorties observés sur l'ensemble de portefeuille, *une entrée précède une censure qui précède elle-même un décès*.

Par la suite on supposera que :

(i) Au moins un des e_i est nul. On peut toujours se ramener à cette situation en introduisant les variables $Y_i=X_i-Min\{e_1,..,e_n\}$ et en se limitant à l'étude de la loi conditionnelle de Y sachant $X \geq Min\{e_1,..,e_n\}$.

(ii) L'âge maximal de sortie correspond à un décès.

Soit X une variable aléatoire positive représentant une durée de maintien et dont la loi est caractérisée par la fonction de survie $t \rightarrow Pr(X \geq t)$. Les variables X_i sont supposées indépendantes de loi commune la loi de X. *Phénomènes de censures et troncatures sont supposés ne pas influencer la variable durée de survie.*

On cherche, en particulier, à estimer la loi de X par l'estimation (notée $t \rightarrow S(t)$) de la fonction de survie. On notera également pour tout t positif $p(t)$ l'estimation de $Pr(X=t)$.

2-b L'estimateur Kaplan-Meier :

En l'absence de troncatures ($e_i=0$ pour tout indice i) et pour des âges de décès tous distincts l'estimateur classique de Kaplan-Meier (1958) est égal à :

$$S(t) = \prod_{\substack{i,\delta_i=1 \\ t_i<t}} (1 - \frac{1}{N(t_i)}) \quad \text{où} \quad N(t) = \sum_i I_{t_i \geq t} \ .$$

représente le nombre de survivants en t.

Pour des âges de décès pouvant être confondus (ce qui est en particulier le cas si la variable étudiée est discrète) on note : $D=\{d_1,...,d_k\}$ avec $d_1<d_2<..<d_k$ (k représente le nombre d'âges de décès distincts) l'ensemble des âges de décès $\{t_i, i \in \{1,2,..,n\}\}$.

On note $N_d(t) (= \sum_i I_{t_i=t})$ le nombre de décès observés en t, l'estimateur S se généralise par :

$$S(t) = \prod_{\substack{j=1,..,k \\ d_j<t}} (1 - \frac{N_d(d_j)}{N(d_j)}) .$$

Remarques :

(i) La fonction $N(t)$ est encore appelée effectif de la population exposée au risque de décès en t . Elle est trivialement égale à la différence entre l'effectif n de la population et le nombre d'âges de décès et de censures antérieures à t (antérieures au sens large pour les censures compte tenu de la convention choisie).

(ii) $S(t)$ n'est une fonction de survie que si la condition (ii) est réalisée.

(iii) La fonction $S(t)$ est une fonction en escalier dont les points de discontinuité correspondent aux âges de décès. La loi associée est donc une loi dont le support est égal à l'ensemble D des âges de décès qui généralise au cas des données censurées la distribution empirique affectant des probabilités égales à chaque âge de décès.

(iv) $S(t)$ est estimateur du maximum de vraisemblance.

(v) Dans le cas d'une variable discrète prenant des valeurs entières on reconnaît pour S la fonction de survie classique associée aux taux conditionnels de mortalité :

$$q_x = \frac{N_d(x)}{N(x)} \qquad x \in \{0,1,...\} .$$

(vi) En réalité, c'est l'actuaire allemand Böhmer (1912) qui le premier présenta l'estimateur de Kaplan-Meier.

2-c Les équations du Maximum de vraisemblance en présence de troncatures :

On considère ici que, par rapport à la situation précédente, certaines des données peuvent être tronquées.

Il est facile d'associer à toute loi de (fonction de survie notée S) candidate à être estimation du maximum de vraisemblance une loi de fonction de survie notée S' qui préserve la vraisemblance des observations censurées ($S'(t_i)=S(t_i)$) pour tout âge t_i correspondant à une censure), qui a pour support l'ensemble D des âges de décès et qui augmente la vraisemblance des observations non censurées.

Il suffit donc, comme pour l'estimation Kaplan-Meier, de limiter la recherche des solutions maximum de vraisemblance aux lois concentrées sur l'ensemble $D=\{d_1,...,d_k\}$ des âges de décès.

On note p_j la probabilité associée à l'âge d_j , $j \in \{1,2,..k\}$. On a alors :

$$\begin{cases} p(t) = p_j \ si \ \ il \ existe \ j \ tel \ que \ t = d_j \\ p(t) = 0 \ \ si \ t \notin D \end{cases}.$$

On a trivialement pour tout t :

$$S(t) = \sum_{s_j \geq t} p_j$$

et en particulier pour $t=0$: $\sum_{j=1,..,k} p_j = S(0) = 1$.

La vraisemblance d'une observation t_i non censurée est alors égale (sous hypothèse d'indépendance entre troncature et décès) à :

$$Pr(X_i = t_i / X_i \geq e_i) = \frac{p(t_i)}{S(e_i)}.$$

La vraisemblance d'une observation t_i censurée est égale (sous hypothèse d'indépendance entre censure, troncature et décès) à :

$$Pr(X_i \geq t_i / X_i \geq e_i) = \frac{S(t_i)}{S(e_i)}.$$

Par indépendance des observations la vraisemblance de l'ensemble des observations est donc égale à :

$$L(p_1,..., p_k) = \prod_{i=1,..n} \left[\left(\frac{S(t_i)}{S(e_i)} \right)^{1-\delta_i} \times \left(\frac{p(t_i)}{S(e_i)} \right)^{\delta_i} \right].$$

La fonction de vraisemblance étant invariante par homothétie on montre facilement que les équations du maximum de vraisemblance peuvent s'obtenir en annulant

la différentielle de L considérée comme fonction des k variables $p_1,p_2,..,p_k$ et en introduisant la contrainte $\sum_{j=1,..,k} p_j = 1$ a posteriori. Soit :

$$(1)\ \begin{cases} \dfrac{N_d(d_j)}{p_j} = \displaystyle\sum_{i=1,..,n} \frac{1}{S(e_i)} \times I_{e_i \le d_j} - \sum_{i=1,..,n} \frac{1}{S(t_i)} \times I_{e_i \le d_j} \times I_{\delta_i = 0} \quad pour\ j = 1,2..,k \\ \displaystyle\sum_{j=1,..,k} p_j \end{cases}$$

2-d Estimateur cohérent :

Définition :

Une loi estimateur sera dite cohérente si, pour toute partie A de $\mathbf{R}^+$:

L'espérance «a priori» (c'est-à-dire sous la seule connaissance des âges e_i de troncatures) du nombre aléatoire de décès intervenant dans la partie A

est égale à

la somme «a posteriori» (c'est-à-dire sous la connaissance de la totalité des informations) du nombre de décès observés dans la partie A et de l'espérance (sous la loi et à cause des censures) du nombre aléatoire de décès intervenant dans la partie A.

Remarque :

Si la cohérence est vérifiée pour tout singleton $\{d_j\}$, $j \in [1,2,..,k]$ elle est vérifiée pour toute partie A dès que la loi cohérente a pour support l'ensemble D des âges de décès.

Le nombre espéré «a priori» est égal à :

$$\sum_{i=1,..n} \frac{\displaystyle\sum_{j=1,..,k} p_j \times I_{d_j \in A} \times I_{d_j \ge e_i}}{S(e_i)}.$$

Le nombre «a posteriori » est égal à :

$$\sum_{i=1,..,n} I_{t_i \in A} \times I_{\delta_i = 1} + \sum_{i=1,..n} \frac{\displaystyle\sum_{j=1,..,k} p_j \times I_{d_j \in A} \times I_{d_j \ge t_i}}{S(t_i)} \times I_{\delta_i = 0}$$

2-e Expression explicite d'un estimateur cohérent :

Proposition :

Si la fonction $t \to N(t) = \sum_{i=1,..,n} I_{e_i \le t} - \sum_{i=1,..,n} I_{t_i \le t} - \sum_{\substack{i=1,..,n \\ \delta_i=0}} I_{t_i=t}$ est strictement positive sur le segment *[0,d_k]* la fonction *S(t)* égale à $\prod_{d_j<t}\left[1-\frac{N_d(d_j)}{N(d_j)}\right]$ est une fonction de survie associée à l'<u>unique</u> loi cohérente.

Preuve :

Si on prend pour *A* l'ensemble *[t, +∞)* la relation de cohérence conduit à l'égalité : (quand *N(t)* est strictement positive) :

$$S(t)\times\left[\sum_{e_i\le t}\frac{1}{S(e_i)}-\sum_{\substack{t_i\le t \\ \delta_i=0}}\frac{1}{S(t_i)}\right]=\sum_{i=1,..,n} I_{t_i\ge t}-\sum_{i=1,..,n} I_{e_i\ge t}\ .$$

L'entier $\sum_{i=1,..,n} I_{t_i\ge t}-\sum_{i=1,..,n} I_{e_i\ge t}$ peut aussi s'écrire :

$$N(t)=\sum_{i=1,..,n} I_{e_i\le t}-\sum_{i=1,..,n} I_{t_i\le t}-\sum_{\substack{i=1,..,n \\ \delta_i=0}} I_{t_i=t}\ .$$

La relation ci-dessus montre que *S(t)* est constante sur les intervalles dont l'intersection avec l'ensemble des âges de décès est vide.

L'expression explicite de *S* se déduit immédiatement par un calcul récurrent sur les âges successifs de décès à partir de la contrainte *S(0)=1*.

Remarque :

L'entier *N(t)* représente l'effectif de la population exposée au risque de décès à l'âge *t* et la fonction *S(t)* généralise directement l'estimateur Kaplan-Meier.

2-f Relation estimateur cohérent et estimateur du maximum de vraisemblance :

Proposition 1 :

Toute loi est cohérente si et seulement si elle satisfait aux équations du maximum de vraisemblance.

Preuve :

Il suffit de prendre pour *A* successivement les singletons *{d_1} ; {d_2} ;...; {d_k}* et, pour la réciproque, utiliser la remarque ci-dessus.

Proposition 2 :

La condition $N(t)>0$ sur $[0,d_k]$ est équivalente à la propriété de prolongement de continuité en 0 de la fonction de vraisemblance sur la frontière du compact de $\mathbf{R}^k$ des valeurs possibles pour l'ensemble $\{(p_1,..,p_k)\}$ c'est-à-dire sur le compact $C=[0,1]^k \cap \{(x_1,..,x_n) / \sum x_i=1\}$.

Preuve :

Pour alléger l'écriture on suppose l'absence de censures et on suppose aussi que toutes les dates de décès sont distinctes. On alors $k=n$. Les observations sont numérotées par âges d'entrée croissants.

S'il existe t tel que $N(t)=0$ il existe un âge de décès (noté t_j) tel que :

- $N(t_j)=1$
- il n'existe aucun âge d'entrée entre t_j et t.
- j troncatures précèdent t_j

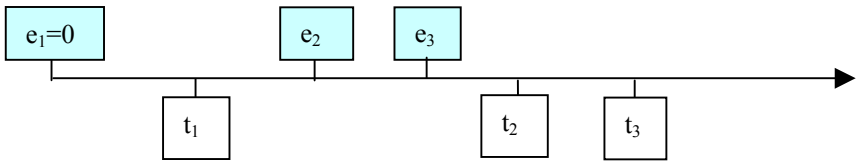
La fonction de vraisemblance s'écrit alors :

$$L(p_1,..,p_n) = \frac{p_1..p_j \times p_{j+1}...p_n}{S(e_1)...S(e_j) \times \prod_{h=j+1,..,n} S(e_h)} \ .$$

Si $S(e_{j+1})$ tend vers 0 il en est de même de $S(e_h)$ pour tout $h>j+1$ et la fonction de vraisemblance tend vers $+\infty$ ou n'a pas de limite (dans le cas le plus favorable où $\prod_{h=j+1,..,n} S(e_h) = (p_{j+1}...p_n)^{n-j}$).

Illustration :

Considérons la population (n=3) :



La fonction N(t) est nulle sur l'intervalle $]t_1 , e_2[$.

La fonction de vraisemblance est égale à

$$L(p_1,p_2,p_3) = \frac{p_1 p_2 p_3}{(p_1+p_2+p_3)\times(p_2+p_3)^2} \ .$$

Cette fonction n'a pas de limite quand p_2+p_3 tends vers 0.

Corollaire :

Si la fonction $N(t)$ reste strictement positive les équations du maximum de vraisemblance ont une solution unique qui est l'estimateur du maximum de vraisemblance.

Preuve :

La condition de continuité en 0 à la frontière du compact assure l'existence d'un estimateur maximum de vraisemblance (fonction positive, continue sur le compact et nulle sur la frontière). L'équivalence avec la relation de cohérence assure l'unicité de la solution du système des équations du maximum de vraisemblance.

2-g Algorithme redistributif :

Efron [6] a proposé, dans le cas de données incomplètes par censures droites un algorithme convergeant vers la solution exacte au bout d'un nombre fini d'itérations et dont le principe consiste à redistribuer à chaque étape la masse affectée à une censure à l'étape précédente, à parts égales sur les décès et les censures qui lui sont postérieures. L'algorithme est initialisé par la distribution empirique.

L'objet de ce paragraphe est d'étendre aux cas des données tronquées l'algorithme Efron et de proposer une écriture de la loi estimateur où censures et troncatures jouent des rôles symétriques.

Lemme de transformation d'un décès en censure :

Soit t_{i0} , $\delta_{i0}=1$ un âge de décès strictement inférieur à l'âge maximal de décès d_k .
On transforme l'échantillon en transformant ce décès en censure.
On note :

$$S'(t)\ ,\ D'\ ,\ \{p'_j,..\}\ ,\ N'(t)\ ,\ N'_d(t)$$

les éléments du nouvel échantillon correspondant respectivement à :

$$S(t)\ ,\ D\ ,\quad \{p_j,...\}\ ,\ N(t)\ ,\ N_d(t).$$

On a trivialement :

$D' \subset D$ avec inclusion stricte si $N_d(t_{io})=1$

$D' \cap [0,t_{io}[= D' \cap [0,t_{io}[$

$N'(t)=N(t)$ pour tout $t \neq t_{io}$; $N'_d(t)=N_d(t)$ pour tout $t \neq t_{io}$

$N'(t_{io})=N(t_{io})-1$; $N'_d(t_{io})=N_d(t_{io})-1$

De la proposition *2-e* on déduit :

$$\begin{cases} p'_j = p_j \\ \text{pour les indices } j \text{ où } d_j < t_{i_0} \end{cases}$$

$$\begin{cases} p'_{jo} = p_{jo} \times \dfrac{N'_d(t_{io})}{N'(t_{io})} \times \dfrac{N(t_{io})}{N_d(t_{io})} = p_{jo} \times \dfrac{N_d(t_{io})-1}{N(t_{io})-1} \times \dfrac{N(t_{io})}{N_d(t_{io})} \\ \text{pour l'indice } j_o \text{ tel que } d_{j_o} = t_{i_o} \end{cases}$$

$$\begin{cases} p'_j = p_j \times \dfrac{(1-\dfrac{N'_d(t_{io})}{N'(t_{io})})}{(1-\dfrac{N_d(t_{io})}{N(t_{io})})} = p_j \times \dfrac{N(t_{io})}{N(t_{io})-1} \\ \text{pour tout indice } j \text{ tel que } d_j > t_{i_0} \end{cases}$$

Corollaire 1 : Généralisation de l'algorithme d'Efron :

La probabilité p'_j (pour $d_j > t_{io}$) peut s'écrire : $p'_j = p_j + \dfrac{p_j}{N(t_{io})-1}$. Les probabilités aux âges de décès postérieurs à t_{io} sont augmentées, par la transformation du décès en censure, d'un pourcentage constant égal à l'inverse de l'effectif la population exposée au risque de décès en t_{io} après la transformation.

De plus si on suppose que $N_d(t_{io})$ est égal à *1* la probabilité p'_{jo} s'annule.

D'où l'algorithme :

Etape 0 : On part de l'échantillon où toutes les sorties par censures sont considérées comme des décès.

Etape 1 : Le premier «faux» décès rencontré (on suppose que tous les âges de sorties des «faux» décès sont distincts) est transformé en «vraie» censure et les probabilités postérieures pour toutes les sorties (décès et censures) sont augmentées du pourcentage convenable défini ci-dessus.

On procède de manière analogue à chaque étape et l'algorithme se termine à la rencontre du dernier «faux» décès.

Remarques :

En l'absence de données tronquées et si toutes les sorties postérieures à t_{jo} sont des décès (ce qui est le cas dans l'algorithme décrit précédemment les probabilités p_j sont trivialement égales à p_{jo} et l'augmentation égale à $\dfrac{p_{jo}}{N(t_{io})-1}$ des probabilités p'_j s'interprète additivement comme la répartition de la probabilité p_{jo} uniformément sur les $N(t_{io})-1$ probabilités postérieures. On retrouve ainsi l'algorithme proposé par Efron.

Dans le cas où les âges de sorties des «faux» décès sont, à une étape de l'algorithme confondus, il suffit d'introduire un ordre fictif dans les «faux» décès concernés et d'appliquer l'algorithme précédent.

Corollaire 2 : Expression analytique de l'algorithme :

Soit $p_o(t)$ l'estimation de la probabilité $Pr(X=t)$ déduite de l'échantillon $\{(e_i, t_i, \delta_i=1), i=1,2,..,n\}$ où toutes les sorties sont considérées comme des décès.

Soit C l'ensemble des âges de censures :

$C=\{c_1,c_2,..,c_h\} = \{t_i / \delta_i = 0\}$ avec $c_1<c_2<..<c_h$

Soit $N_c(t)=Card\{i / t_i=t \text{ et } \delta_i=0\}$ le nombre de censures observées à l'instant t.

On a pour tout t :

$$p(t) = p_o(t) \times \prod_{c_j \le t} \left(\frac{N(c_j) + N_c(c_j)}{N(c_j)} \right).$$

Preuve :

Il suffit en partir de l'échantillon $\{(e_i , t_i , \delta_i=1), i=1,2,..,n\}$ d'effectuer une récurrence sur la transformation de proche en proche des âges de «faux» décès en «vraies» censures en appliquant à chaque étape de la récurrence le lemme.

Remarque :

Le cas $N_c(c_j)>1$ se traite en affectant un ordre virtuel à l'ensemble $N_c(c_j)$ censures concernées.

2-h Expression de la loi cohérente en l'absence de censures droite :

Soit $G=\{g_1,..,g_h\}$ $(g_1=0<g_2<..<g_h)$ l'ensemble $\{e_i ; i=1,2,..,n\}$ des âges d'entrées.

Soit $j \rightarrow Ne(g_j)$ la suite définie par :

$Ne(0)$= Nombre de données non tronquées moins un.

$Ne(g_j)$ = Nombre d'entrées d'âges g_j . $j>0$.

Proposition :

Si tous les âges de sortie sont des âges de décès on a pour l'estimateur cohérent :

$$p(t) = \prod_{g_j \le t} \left(1 - \frac{N_e(g_j)}{N(g_j)} \right) .$$

Preuve :

On se ramène au cas où tous les âges d'entrée sont distincts en ordonnant virtuellement toutes les entrées pour lesquelles les âges sont égaux.

On démontre alors la relation par récurrence en démontrant que si la proposition est vraie pour un échantillon elle reste vraie pour tout échantillon obtenu en déplaçant un âge d'entrée donné à n'importe quel âge supérieur.

Il suffit alors de démontrer la proposition pour le cas sans troncature (distribution empirique) et de déplacer les «fausses» observations complètes vers les «vraies» troncatures.

Corollaire :

Pour l'estimateur cohérent on a :

$$(2) \qquad p(t) = \prod_{g_j \le t} \left(1 - \frac{N_e(g_j)}{N(g_j)} \right) \times \prod_{c_j \le t} \left(1 + \frac{N_c(c_j)}{N(c_j)} \right)$$

Preuve :

Il suffit d'appliquer le corollaire 2 en remarquant que la transformation de sorties par décès en sorties par censures n'affecte pas la fonction $N(t)$ aux points g_j , $j=1,..,h$.

Remarques :

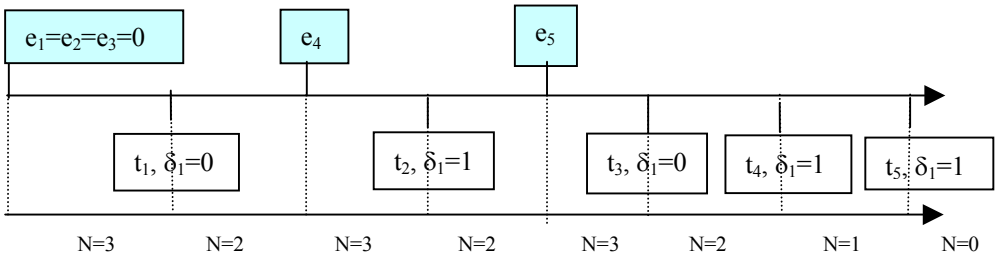
Ce dernier corollaire est intéressant dans la mesure où il permet de détecter des symétries dans l'influence sur la loi de survie du phénomène de censure droite et du phénomène de troncature gauche.

Par ailleurs on peut légitimement espérer que cette dernière relation puisse s'étendre, sous une forme peu différente, à l'estimation de la loi conjointe de plusieurs durées de survie. Nous verrons, malheureusement, qu'il n'en est rien.

Un exemple :

Considérons une population de *5* observations (graphe ci-dessous) dont *2* sont tronquées à gauche et *2* sont censurées à droite.

Le calcul de la loi cohérente de la population sans censures est d'abord donné. L'algorithme distributif est ensuite construit pour obtenir la loi cohérente finale.



1- Echantillon sans censure :

$p(t_1)=(1-2/3)=1/3$

$p(t_2)=(1-2/3)\times(1-1/3)=2/9$

$p(t_3) =p(t_4)=p(t_5)=(1-2/3)\times(1-1/3)\times(1-1/3)=4/27$

2- Algorithme de redistribution :

Première étape : le «faux» décès t_1 est transformé en censure :

Pourcentage d'augmentation $1+1/(N(t_1)-1)=1+0,5=150\%$

$\boldsymbol{p(t_1)=0}$

$p(t_2)=2/9\times1.5=1/3$

$p(t_3) =p(t_4)=p(t_5)=4/27\times1.5=2/9$

Deuxième étape : le «faux» décès t_3 est transformé en censure :

Pourcentage d'augmentation $1+1/(N(t_3)-1)=1+0.5=150\%$

$p(t_2)=1/3$

$\boldsymbol{p(t_3) =0}$

$p(t_4)=p(t_5)=2/9\times1.5=1/3$

3. ESTIMATION NON PARAMETRIQUE DE LA LOI CONJOINTE DE DEUX DUREES DE SURVIE

3-a La problématique :

L'étude simultanée de plusieurs durées de survie non indépendantes trouve de nombreuses applications dans le domaine des sciences actuarielles.

Pour en évoquer quelques-uns unes citons par exemple :

Exemple (a) : En assurance non-vie, lors de la survenance d'un sinistre, les différents coûts relatifs aux différentes garanties du contrat.

Exemple (b) : L'âge au décès et la date calendaire de décès d'un assuré.

Exemple (c) : Pour un invalide l'âge à l'entrée en invalidité et la durée de maintien en invalidité.

Exemple (d) : Pour un contrat retraite le capital acquis et l'âge au départ de la retraite.

Exemple (e) : Pour un couple les âges au décès des deux conjoints.

La généralisation au cas multivarié de l'estimation de la loi conjointe de durées de survie (c'est à dire de variables aléatoires positives) en présence de données incomplètes présente des difficultés spécifiques importantes qui ont conduit à de très nombreux travaux.

La définition des troncatures à gauche ne pose aucune difficulté particulière. Par contre les censures droites peuvent affecter indépendamment les différentes durées de survie (exemple (e)). Elles peuvent aussi n'affecter que simultanément toutes les variables observées (Exemples (b) et (d)). On parle alors de censure univariée. Elles peuvent n'affecter que certaines des variables observées (Exemples (a) , (c)).

Par ailleurs la nature des variables observées, la nature de la cause de censure (notamment dans le cas de censure simultanée) et la nature des troncatures peuvent donner des informations restrictives sur les valeurs réelles de ces variables. Dans l'exemple (b) un assuré censuré (respectivement tronqué) à l'âge x et à la date t décédera à un âge aléatoire X, et à la date $t+X$, c'est à dire sur une demi-droite du domaine $[x,\infty)\times[t,\infty)$. Par contre dans les exemples (d) et (e) une censure simultanée (respectivement une troncature gauche) ne donne aucune autre information que l'appartenance au domaine des valeurs postérieures aux valeurs de censure. (troncature).

Dans la suite on supposera la censure est univariée On supposera aussi que censures et durées de survie peuvent se situer dans la totalité du domaine des valeurs postérieures aux valeurs de troncature et de même que les durées de survie peuvent se situer dans la totalité du domaine des valeurs postérieures aux valeurs de censures.

3-b Notations et données du problème :

$X=(Y,Z)$ désigne un vecteur aléatoire de $\boldsymbol{R}^{+2}$ composantes (appelées aussi durées de survie) toutes positives.

Des ordres partiels sont définis sur $\boldsymbol{R}^{+2}$ par :

Soient $s=(s_1,s_2)$ un point de $\boldsymbol{R}^{+2}$.

On désigne par *Futur(s)* l'ensemble des points de $\boldsymbol{R}^{+2}$ dont les composantes sont supérieures ou égales aux composantes respectives de s. On note $t \geq s$ si $t \in$ Futur(s).

On désigne par *Futurstrict(s)* l'ensemble des points de $\boldsymbol{R}^{+2}$ dont les composantes sont supérieures strictement aux composantes respectives de *s*. On note $t>s$ si $t \in Futurstrict(s)$.

On désigne par *Passe(s)* l'ensemble des points de $\boldsymbol{R}^{+2}$ dont les composantes sont inférieures ou égales aux composantes respectives de s. On note $t \leq s$ si $t \in Passe(s)$.

On désigne par *Passestrict(s)* l'ensemble des points de $\boldsymbol{R}^{+2}$ dont les composantes sont inférieures strictement aux composantes respectives de *s*. On note $t<s$ si $t \in Passestrict(s)$

On dispose de l'observation d'une population de *n* couples de durées de vie. Pour l'individu numéro *i* on note $X_i=(Y_i,Z_i)$ le couple aléatoire associé.

On observe X_i sachant $X_i \in Futur(e_i)$. Si le couple e_i est non nul on dit que l'observation est tronquée et le couple e_i sera désigné par couple d'entrée.

On observe le couple de sortie t_i de l'individu et la cause de la censure ($\delta_i=0$ si l'observation est censurée - on a supposé que les deux durées étaient censurées simultanément- ; $\delta_i=1$ si les deux durées sont observées).

On reprend les mêmes hypothèses que dans les parties 1 et 2. On cherche à estimer la loi commune aux couples X_i notamment par l'estimation de la fonction de survie conjointe $Pr(X \geq t)=Pr(Y \geq y \cap Z \geq z)$ pour tout $t=(y,z)$ de $\boldsymbol{R}^{+2}$. On note $S(t)=S(y,z)$ tout estimateur de la fonction de survie.

3-c Equations du maximum de vraisemblance et estimation cohérente :

Proposition :

Une loi est une estimation cohérente si et seulement si elle est solution des équations du maximum de vraisemblance.

Preuve :

Les équations du maximum de vraisemblance sont, au détail des relations d'ordres partiels près, les mêmes que celles obtenues dans la partie 2. Ainsi le système des équations du maximum de vraisemblance est égal à :

$$(3) \quad \begin{cases} \dfrac{N_d(d_j)}{p_j} = \displaystyle\sum_{e_i \in Passé(d_j)} \frac{1}{S(e_i)} - \sum_{\substack{t_i \in Passé(d_j) \\ \delta_i=0}} \frac{1}{S(t_i)} \quad pour\ j = 1,..,k \\ \displaystyle\sum_{j=1,..,k} p_j = 1 \end{cases}$$

Remarque :

(i) Les *k* couples de sortie par «décès» sont indicés ici dans un ordre arbitraire.

(ii) De même la définition de cohérence s'étend sans difficulté et la relation de cohérence écrite pour une partie *A* de $\boldsymbol{R}^{+2}$ s'écrit :

$$\sum_{i=1,..n} \frac{\sum_{d_j \in Futur(e_i)} p_j \times I_{d_j \in A}}{S(e_i)} = \sum_{i=1,..,n} I_{t_i \in A} \times I_{\delta_i = 1} + \sum_{\substack{i=1,..n \\ \delta_i = 0}} \frac{\sum_{d_j \in Futur(t_i)} p_j \times I_{d_j \in A} \times I_{d_j \geq t_i}}{S(t_i)} .$$

Pour $A=\{dj\}$, $j=1,2,..\ ,n$ on retrouve les équations du maximum de vraisemblance.

Remarque :

Par contre aucun choix pour *A* de sous-ensemble ne conduit à la construction explicite (et donc l'unicité) d'une loi cohérente. C'est évidemment l'absence de relation d'ordre total entre les couples qui interdit cette construction «de proche en proche».

3-d Existence et unicité des équations du maximum de vraisemblance.

Proposition :

Si les données incomplètes ne sont que des censures droites il existe un estimateur maximum de vraisemblance unique, solution des équations du maximum de vraisemblance. De plus, tout algorithme itératif qui augmente strictement la vraisemblance converge vers l'estimation du maximum de vraisemblance.

Preuve :

La fonction de vraisemblance $x \rightarrow L(x)$ est continue sur le compact $C=[0,1]^k \cap \{x=(x_1, ..,x_k / \sum x_j=1\}$. Elle atteint un maximum qui n'est pas pris sur la frontière où la fonction est nulle. Ce maximum est donc atteint sur l'intérieur du compact ou, par invariance de la fonction de vraisemblance par homothétie, la différentielle (en tant que fonction de $\boldsymbol{R}^k$ est nulle).

Remarque :

- Le non-respect de la condition de continuité peut apparaître même quand les conditions de continuité sur les deux fonctions de vraisemblance marginales sont satisfaites.

Par ailleurs un développement limité du logarithme de la fonction de vraisemblance autour de tout point extrémal (c'est-à-dire de tout point de différentielle nulle) montre que tout point extrémal du compact *C* est un maximum local sur le compact ce qui assure l'unicité du point extrémal. Ce point unique est donc estimation du maximum de vraisemblance.

Soit maintenant une écriture des équations du maximum de vraisemblance (c'est à dire des équations d'annulation de la différentielle de la fonction de vraisemblance) de la forme $x=\Phi(x)$ telle que :

(i) ϕ est une fonction continue de *C* dans *C*.

(ii) Pour tout point *x* de *C* différent du point fixe $L(\Phi(x)>L(x)$.

La suite définie par $\{x_0;\ x_{n+1}=\Phi(x_n),\ n \in \boldsymbol{N}\}$ est une suite de points de *C*. Cette suite admet une suite extraite $x_{g(n)}$ qui converge vers un point λ de *C*.

Par ailleurs la suite $L(x_n)$ est croissante, majorée, donc convergente et sa limite est $L(\lambda)$ par continuité de L. La suite $L(\Phi(x(g(n))=L(x_{g(n)+1})$ converge $L(\Phi(\lambda))$. Cette suite est aussi une suite extraite de la suite $L(\lambda)$. Elle converge donc encore vers $L(\lambda)$ et $L(\Phi(\lambda)$ est égal à $L(\lambda)$. On conclut à l'égalité $\lambda=\Phi(\lambda)$ en remarquant que si λ est différent de $\phi(\lambda)$ alors $L(\phi(\lambda)$ est strictement supérieur à $L(\lambda)$!

Le point λ est donc l'unique solution des équations de vraisemblance et la suite $\{x(0)\ ;\ x(n)\ n\in \mathbf{N}\}$ du compact C a une unique valeur d'adhérence. Elle est donc convergente vers λ.

Remarques :

- En présence de données tronquées à gauche on peut montrer que comme dans le cas univarié on a une solution aux équations du maximum de vraisemblance si et seulement si la fonction de vraisemblance est prolongeable par continuité en *0* sur la frontière du compact C. Il y a donc au moins un point extrémal qui soit un maximum.

- Malheureusement la démonstration, dans ce cas, du fait que tout point extrémal est un maximum local reste à faire - si elle est vraie- ? (aucun contre exemple n'a pu être identifié).

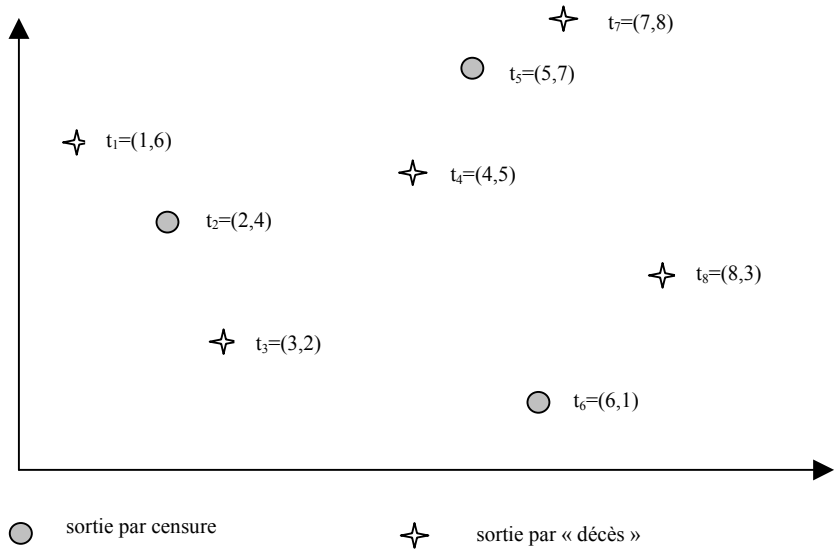
- On peut seulement alors affirmer que toute suite qui augmente la vraisemblance possède une valeur d'adhérence qui est un point extrémal (évidemment non minimal local). Il est donc indispensable de choisir comme point initial de la suite un point proche du maximum supposé de la vraisemblance.

3-e Algorithme redistributif :

Tout au moins en absence de données tronquées on pourrait imaginer généraliser l'algorithme Efron pour construire une loi cohérente.

Exemple :

L'exemple décrit ci-dessous constitue en fait un contre-exemple :



La fonction de vraisemblance est égale à :

$$L(p(t_1),\ldots,p(t_8)) = \frac{p(t_1).p(t_3)p(t_4)p(t_8)p(t_7)^2(p(t_7)+p(t_4)).(p(t_7)+p(t_8))}{(p(t_1)+p(t_3)+p(t_4)+p(t_7)+p(t_8))^5}.$$

Les équations du maximum de vraisemblance conduisent à l'unique solution :

$p(t_1) = p(t_3) = 1/8 \quad ; \quad p(t_4) = p(t_8) = (3 - \sqrt{3})/8 \quad ; \quad p(t_7) = 2\sqrt{3}/8$.

Alors que tout algorithme redistributif initialisé par la distribution empirique qui affecte des probabilités toutes égales à 1/8 aux 5 «vrais» décès et au 3 «faux» décès ne saurait en un nombre fini de pas conduire à une solution irrationnelle.

Estimation par algorithme redistributif :

On propose ici une généralisation de l'algorithme Efron en deux dimensions de la manière suivante :

Initialisation :

loi empirique qui affecte probabilité égale à *1/n* pour toutes les couples de sorties.

Etape 1 : On identifie l'ensemble des «faux» couples qui n'ont aucun «faux» couples dans leurs passés (les couples t_2 et t_6 dans l'exemple). Les probabilités (1/8 dans l'exemple de chacun ces faux couples sont réparties additivement sur les couples «vrais» et «faux» de leur futur respectif (sur l'exemple t_4,t_5,t_7 pour t_2 ; t_7 et t_8 pour t_6).

Etapes suivantes : on réitère le processus décrit ci-dessus jusqu'à disparition de tous les «faux» couples.

Sur l'exemple les probabilités finales sont respectivement :

$p(t_1) = p(t_3) = 1/8$ (égales à l'estimation Maximum de vraisemblance) ;
$p(t_4) = 1/8+1/24 \# 0.17$ (au lieu de 0.16)
$p(t_8) = 1/8+1/16 \# 0.19$ (au lieu de 0.16)
$p(t_7) = 1/8+1/24+1/16+1/8+1/24 \# 0.40$ (au lieu de 0.43)

Remarques :

- Les propriétés de cette estimation sont vraisemblablement difficiles à établir. Il semble néanmoins qu'il puisse constituer le point initial d'un algorithme itératif construit sur les équations du maximum de vraisemblance.

- En présence de données tronquées on peut, pour obtenir une estimation de la loi conjointe proposer une généralisation de la relation (3) sous la forme :

$$p(t) = \prod_{g_j \in Passe(t)} \left(1 - \frac{N_e(g_j)}{N(g_j)}\right) \times \prod_{c_j \in Passé(t)} \left(1 + \frac{N_c(c_j)}{N(c_j)}\right)$$

avec $N(t) = Card\{i / e_i \in Passé(t)\} - Card\{i / t_i \in Passé(t)\}$

3-f Un exemple d'algorithme itératif :

Le système d'équations du maximum de vraisemblance (3) peut s'écrire :

$$N_d(d_j) = p_j \times \Big(\sum_{e_i \in Passé(d_j)} \frac{1}{S(e_i)} - \sum_{\substack{t_i \in Passé(d_j) \\ \delta_i = 0}} \frac{1}{S(t_i)} + M - M \Big) \quad \text{j=1,2,..,k}$$

où M est un réel positif quelconque.

Soit encore en posant :

$$H_j(p_1,.,p_k) = \Big(\sum_{e_i \in Passé(d_j)} \frac{1}{S(e_i)} - \sum_{\substack{t_i \in Passé(d_j) \\ \delta_i = 0}} \frac{1}{S(t_i)} \Big),$$

$$p_j = p_j + \frac{N_d(d_j) - p_j \times H_j(p_1,..,p_k)}{M} \qquad \text{j=1,2,..,k.}$$

Soit Φ la fonction définie sur le compact C par :

$$Pour\ x = (x_1,..,x_k) \qquad \Phi(x) = \begin{bmatrix} \Phi_1(x) = x_1 + \dfrac{N_d(d_1) - x_1 H_1(x)}{M} \\ . \\ . \\ \Phi_k(x) = x_k + \dfrac{N_d(d_k) - x_k H_k(x)}{M} \end{bmatrix}.$$

- L'application Φ est trivialement continue sur le compact C.
- L'application Φ prend ses valeurs dans le compact C.

En effet on montre facilement en commutant les indices de sommation dans la double somme $\sum_{j=1,..,k} H_j(x)$ que si $\Sigma x_i = 1$ alors $\Sigma \phi_i(x) = 1$.

- L'application Φ augmente la vraisemblance.

En effet, en posant $G(x)=Ln(L(x))$ on démontre que la différence $G(\Phi(x))-G(x)$ peut s'écrire :

$$G(\Phi(x)) - G(x) = \sum_{j=1,..,k} \frac{x_j}{M} \times \left[\frac{N_d(d_j)}{x_j} - H_j(x) \right]^2 + o(\|\Phi(x) - x\|)$$

ce qui assure la croissance de la fonction de vraisemblance dès qu'on se situe dans un voisinage de l'estimation du maximum de vraisemblance.

- L'algorithme itératif $\{x_0\ ,\ x_{n+1}=\Phi(x_n)\ \ n \in N\}$ converge donc vers l'estimation maximum de vraisemblance tout au moins si le point x_0 d'initialisation de la suite est assez proche du point fixe.

REFERENCES

[1] BÖHMER P.E . (1912) Theorie der unabhängigen Wahrscheinlichkeiten Rapports. *Mémoires et Procès Verbaux du Septième Congrès International d'Actuaires, Amsterdam, Vol. 2, 327-343.*

[2] BURKE M.D. (1988) Estimation of a bivariate distribution function under random censorship. *Biometrika* ***75-2****, 379-382*

[3] CAMPBELL G. (1981) Nonparametric bivariate estimation with randomly censored data. *Biometrika* ***68, 2****, 423-426.*

[4] DABROWSKA D. (1988) Kaplan-Meier estimate on the plane. *Annals of Statistics* ***16, 4,*** *1475-1489.*

[5] DROESBEKE J.J ., FICHET B., TASSI P. (1989) Analyse statistique des durées de vie . *Economica.*

[6] EFRON B. (1967) The two sample problem with censored data. *Proceeding 5th Symposium,* ***vol. 4****, 831-853.*

[7] KAPLAN E.L., MEIER P. (1958) Non parametric estimation from incomplete observations. *Journal of American Statistical Association.* ***53****, 457-481.*

[8] LAGAKOS S.W., BARRAJ L., DE GRUTTOLA V. (1988) Non parametric analysis of truncated survival data with application to aids. *Biometrika,* ***75, 3****, 515-523.*

[9] LIN D.Y. , YING Z. (1993) A simple non parametric estimator of the bivariate survival function under univariate censoring. *Biometrika,* ***80, 3****, 573-581.*

[10] PRENTICE R.L., CAI J. (1992) Covariance and survivor function estimation using censored multivariate failure time data. *Biometrika,* ***79, 3****, 495-512.*

[11] TSAI W., LEURGANS S., CROWLEY J. (1986) Non parametric estimation of bivariate survival function in the presence of censoring. *Annals of Statistics* ***14, n°4****, 1351-1365.*

[12] TURNBULL B. (1974) Non parametric estimation of a survivor function with doubly censored and truncated data. *Journal of American Statistical Association,* ***69,*** *169-173.*

[13] TURNBULL B. (1976) The empirical distribution function with arbitarly grouped, censored and truncated data. *Journal of the Royal Statistical Society B,* ***38,*** *290-296.*

[14] WOODROOFE M. (1985) Estimating a distribution function with truncated data. *Annals of Statistics* ***13,*** *163-177.*

Didier RULLIERE
Direction Technique Ecureuil Vie
5 rue Masseran
75007 PARIS
Tél. : 01.44.49.64.31 Fax 01.45.66.93.04
Fax : 01.45.66.93.04

Daniel SERANT
I.S.F.A. Université Claude Bernard
43 Boulevard du 11 Novembre 1918
69622 Villeurbanne Cedex
Tél. :04.72.43.11.75 Fax : 04.72.43.11.76
e-mail : serant@cismsun.univ-lyon1.fr