JOURNAL DE LA SOCIÉTÉ FRANÇAISE DE STATISTIQUE

GÉRARD DERZKO
EVE LECONTE

Estimation non paramétrique d’incidences d’événements en compétition avec censure à droite

Journal de la société française de statistique, tome 145, nº 1 (2004), p. 47-69

<http://www.numdam.org/item?id=JSFS_2004__145_1_47_0>

ESTIMATION NON PARAMÉTRIQUE D'INCIDENCES D'ÉVÉNEMENTS EN COMPÉTITION AVEC CENSURE À DROITE

Gérard DERZKO * & Eve LECONTE **

RÉSUMÉ

L'estimation non paramétrique de l'incidence d'événements en compétition avec censure à droite est classiquement réalisée par deux méthodes : la méthode des « variables latentes », consistant à estimer l'incidence des événements de chaque type par l'estimateur de Kaplan-Meier, en considérant les événements d'autres types comme des censures, et la méthode des « causes spécifiques », basée sur l'estimateur de Prentice. Nous montrons que ces deux méthodes mettent en jeu deux modèles différents de sélection aboutissant à l'échantillon marqué observé, respectivement un modèle de sélection par minimum et un modèle de mélange censuré. La validité de ces modèles ne peut être testée mais leur choix peut être guidé par les conditions du dispositif expérimental. Deux algorithmes très simples permettant d'estimer non paramétriquement les incidences des événements sont proposés pour ces deux modèles, puis nous en donnons une généralisation dans des cas plus complexes. Les conséquences d'une erreur de spécification du mécanisme sont illustrées par simulation, puis un exemple d'application en cancérologie est donné.

Mots-clés : Risques compétitifs, incidence cumulée, estimation non paramétrique, censure à droite, processus de sélection.

ABSTRACT

Nonparametric estimation of the incidence of competing events with right censoring is traditionally achieved through two methods : the method of « latent variables », which uses the Kaplan-Meier estimator for the incidence of each type of events, considering the events of all other types as censoring, and the method of « specific causes », based on Prentice's estimator. We demonstrate that these two methods may be associated with two distinct selection mechanisms generating marked samples, which we will call respectively « selection by minimum » and « censored mixture ». The validity of these models cannot be tested, but choosing one of them can be driven by some features of the experimental design. Two very simple algorithms for nonparametric estimation of events' incidence are developed for these models and we extend the algorithms to more complex situations. The consequences of model misspecifications are illustrated by simulations, and an example in oncology is given.

* SANOFI-SYNTHELABO Recherche, 371, rue du Professeur Joseph Blayac, 34184 Montpellier Cedex 4.
E-mail : gerard.derzko@sanofi synthelabo.com

** Groupe de Recherche en Économie Mathématique et Quantitative, Université des Sciences Sociales, 21, allée de Brienne, 31000 Toulouse et Laboratoire de Statistique et Probabilités, Université Paul Sabatier, 31062 Toulouse Cedex 4.
E-mail : leconte@cict.fr, page web : http ://www.univ tlse1.fr/GREMAQ/Statistique/Eveweb

Keywords : Competing risks, cumulative incidence function, nonparametric estimation, right censoring, selection process.

1. Introduction

Dans les études portant sur des données de survie, on est souvent confronté au problème des risques compétitifs, où les événements affectant les sujets peuvent être de différents types. Le problème se complique fréquemment par une censure à droite. Les exemples d'application abondent, en particulier dans le domaine médical. Le problème des risques compétitifs est apparu à l'origine dans le domaine de la santé publique et des statistiques de mortalité : il s'agissait principalement d'estimer les courbes d'incidence de décès correspondant à chaque cause de décès, et prédire comment elles se modifieraient dans le cas où l'une des causes serait éradiquée. Le terme de risques compétitifs désignait au départ l'étude d'événements fatals de différents types, mais ce terme a été étendu à tout problème où l'on peut distinguer différents types de défaillances, fatals ou non. Par exemple, en cardiologie, une population à risque d'accidents cardio-vasculaires peut manifester plusieurs types d'accidents : infarctus du myocarde, accident vasculaire cérébral,... En cancérologie, on peut être intéressé par le temps écoulé jusqu'à la survenue du premier événement parmi plusieurs événements non mortels (rechute de la maladie, apparition de métastases,...). Ces événements peuvent être étudiés indistinctement en les considérant comme des réalisations d'une seule et même variable délai, mais il est encore plus intéressant d'appréhender leur comportement spécifique.

Le contexte général de cet article est celui où n unités statistiques sont expérimentées et où l'unité i $(1 \leqslant i \leqslant n)$ fournit un résultat constitué d'un couple (T_i, Y_i) ; Y_i est la réalisation d'une v.a. discrète Y qui catégorise les événements et T_i est la réalisation d'une v.a. continue positive T (en pratique le temps écoulé depuis une origine fixe jusqu'à la survenue d'un événement quelle que soit sa nature Y).

L'expérimentateur est intéressé en particulier par certaines des catégories d'événements, dont l'observation est perturbée par les autres catégories ; c'est le cas de la censure à droite, qui en se produisant avant un événement d'intérêt, vient empêcher son observation. Le problème principal est alors d'inférer sur des distributions partiellement (et aléatoirement) masquées.

Cet article se concentre sur les méthodes d'estimation non paramétrique de distributions non totalement observées, du fait de la présence d'une censure à droite. Il existe une très abondante littérature sur ce sujet, débutant par l'article de Kaplan et Meier en 1958, qui fournit un estimateur non paramétrique du maximum de vraisemblance dans la situation de base où l'étude porte sur les délais d'apparition d'un événement d'un type donné en présence d'une censure à droite indépendante de l'événement. La méthode de Kaplan-Meier est la plus connue des praticiens, et est incluse dans la plupart des logiciels du commerce ; aussi est-elle utilisée parfois sans discernement dans toutes sortes de conditions expérimentales complexes, incluant plusieurs types d'événements et de censures. Les utilisateurs considèrent alors que tous les événements en compétition avec l'événement d'intérêt se comportent comme une censure vis-à-vis de cet événement, en plus de la vraie censure : cette approche n'est valide qu'en supposant l'indépendance des événements entre eux, condition qui paraît d'évidence trop forte dans de nombreuses situations. L'emploi abusif de la procédure de Kaplan-Meier a été discutée par d'autres auteurs (voir par exemple Gooley *et al.*, 1999), et plusieurs méthodes ont été comparées (voir par exemple Arriagada *et al.*, 1992).

Bien qu'il y ait beaucoup de confusion dans la terminologie, l'approche consistant à utiliser la méthode de Kaplan-Meier en cas de risques compétitifs en considérant les autres événements comme de la censure est connue sous le nom de méthode des variables latentes. Une autre méthode d'estimation pour traiter des situations où la censure est indépendante des différents types d'événements d'intérêt, qui ne sont pas supposés indépendants entre eux, a été proposée par Prentice (Kalbfleisch et Prentice, 1980). Cette méthode est peu connue, peu utilisée en pratique, et absente de la plupart des logiciels.

Cet article a pour objet de montrer que, dans le contexte général fixé, les données observées relèvent de différents modèles de sélection de base. L'estimation de Kaplan-Meier est pertinente pour un modèle de sélection par valeur minimale avec indépendance des délais. L'estimateur de Prentice est toujours valide, et correspond à un modèle de mélange censuré. Curieusement, l'idée d'un tel modèle apparaît rarement et seulement récemment dans la littérature (voir Betensky et Schoenfeld, 2001), et la relation de ce modèle avec l'estimateur de Prentice paraît ignorée. Nous développons pour ces deux modèles de base des algorithmes simples permettant l'estimation non paramétrique des (sous-)distributions décrivant l'apparition des événements au cours du temps. Ces modèles peuvent être combinés en modèles complexes, et les algorithmes proposés se généralisent sans difficulté à ces cas.

Naturellement, les estimations obtenues par différentes méthodes sont différentes et induisent des conclusions et des interprétations différentes des données expérimentales. Le problème essentiel consiste alors à choisir le modèle sous-jacent de production des données. Malheureusement, ces modèles ne peuvent être distingués au moyen de tests d'hypothèses, comme cela a été remarqué très tôt par Tsiatis (1975). Cependant, bien qu'il n'y ait pas de méthode universelle pour ce choix, les conditions de l'expérimentation et la nature des événements étudiés permettent souvent de l'effectuer a priori. Nous donnons des indications et des illustrations pour ces problèmes.

La section 2 présente les deux modèles de sélection de base dans le cas de deux événements en compétition sans censure. Dans la section 3, nous généralisons ces modèles au cas d'un nombre quelconque d'événements en compétition avec censure à droite. Nous montrons que ces modèles conduisent à deux estimateurs différents de l'incidence cumulée : l'estimateur de Kaplan-Meier et l'estimateur de Prentice, et les algorithmes permettant d'obtenir ces estimateurs sont présentés. Si ces algorithmes n'ont pas grand intérêt dans les cas simples (car les estimateurs correspondants peuvent être obtenus autrement), leur utilité apparaît dans le cas de dispositifs complexes, tels ceux présentés à la fin de la section 3. Nous expliquons ensuite en section 4 comment générer des échantillons correspondant aux deux mécanismes de base dans les études de simulation, et deux échantillons simulés selon les deux mécanismes sont analysés par les deux méthodes, pour illustrer l'erreur commise dans l'estimation en supposant un modèle inapproprié. La section 5 présente un exemple d'application en cancérologie. Enfin, la section 6 s'intéresse au choix pratique du mécanisme de sélection, et la section 7 apporte des éléments complémentaires de réflexion.

2. Deux événements en compétition sans censure

2.1. Les mécanismes de sélection

Notations

Considérons deux types d'événements en compétition. Les données observées constituent un n-échantillon marqué : $(T_\imath, Y_\imath)_{i=1}^n$, où T désigne le délai d'apparition d'un événement, variable continue de fonction de répartition $U(t) = P(T < t)$ et où Y, le type de l'événement, ou encore la marque, est une variable qualitative prenant ses valeurs dans $\{0, 1\}$.

Notons $U_0(t) = P(T < t, Y = 0)$ et $U_1(t) = P(T < t, Y = 1)$ les incidences cumulées observées pour chaque type d'événement. Ces fonctions sont des sous-distributions. On a de façon évidente $U(t) = U_0(t) + U_1(t)$. Notons les distributions conditionnelles correspondantes $F_0(t) = P(T < t \mid Y = 0)$ et $F_1(t) = P(T < t \mid Y = 1)$.

Ce n-échantillon marqué peut être vu comme le résultat d'une sélection pouvant provenir de deux mécanismes différents, que nous allons détailler ci-après.

Le mécanisme de sélection par minimum (MSM)

La variable T observée peut être considérée comme le minimum de deux variables T_0 et T_1, dénommées parfois dans la littérature « variables latentes » car non complètement observables. Cela correspond aux notations d'Efron (utilisées classiquement pour modéliser le phénomène de censure à droite d'un événement) :

$$T = \min(T_0, T_1),$$
$$Y = j : T = T_j.$$

En ajoutant l'hypothèse d'indépendance des variables T_0 et T_1 (hypothèse non testable), on obtient la relation suivante :

$$P(T \geqslant t) = P(\min(T_0, T_1) \geqslant t) = P(T_0 \geqslant t)P(T_1 \geqslant t),$$

$$\text{qui peut s'écrire } 1 - U(t) = (1 - F_0(t))(1 - F_1(t)), \tag{1}$$

qui est la condition caractérisant ce mécanisme de sélection.

Une image simple pour décrire ce mécanisme est celle de deux athlètes que l'on fait courir et l'on retient le temps du gagnant, ainsi que son identité.

Ce mécanisme est le seul possible si l'une des marques désigne la censure aléatoire à droite, classiquement supposée indépendante de l'événement d'intérêt. Il correspond dans le cas de deux événements en compétition à des événements pouvant se réaliser pour chaque individu de façon certaine, mais qui ne sont éventuellement pas observés, par manque de temps. Des cas pratiques de deux événements certains et indépendants pouvant affecter des individus sont difficiles à trouver, ce qui explique que ce modèle a été souvent critiqué car correspondant à une situation irréaliste.

Si l'on ne fait pas l'hypothèse d'indépendance de T_0 et T_1, supposer que la variable observée T s'obtient par minimum n'est plus exploitable pour une estimation non paramétrique et on peut alors considérer que l'échantillon relève d'un mécanisme de mélange.

Le mécanisme de mélange (MM)

Une autre approche possible est de considérer que la variable T provient d'un mélange de distribution : deux types de délais coexistent dans l'échantillon en proportion $p_o = P(Y = 0)$ et $p_1 = 1 - p_0$, fixées au départ et donc indépendantes du temps. Ce mécanisme ne peut donc pas être supposé si une des marques désigne la censure car la censure est un processus dynamique. L'approche en termes de mélange a été proposée par d'autres auteurs (voir Betensky et Schoenfeld, 2001).

La relation $U(t) = U_0(t) + U_1(t)$ s'écrit pour ce mécanisme, les proportions p_0 et p_1 des deux types étant fixées :

$$\begin{aligned} U(t) &= F_0(t)P(Y=0) + F_1(t)P(Y=1) \\ &= F_0(t)p_0 + F_1(t)(1-p_0). \end{aligned}$$

Ce n'est pas le cas dans le MSM, où p_0 et p_1, si on voulait les écrire, seraient des fonctions de t. En effet, dans le MM, la différenciation de la formule ci-dessus donne :

$$dU(t) = P(Y=0)dF_0(t) + P(Y=1)dF_1(t),$$

alors que pour le MSM, la différenciation de la condition caractéristique (1) conduit à :

$$dU(t) = (1 - F_1(t))dF_0(t) + (1 - F_0(t))dF_1(t),$$

qui met clairement en évidence le caractère soit constant, soit fonction du temps de la décomposition linéaire de $dU(t)$ en fonction de $dF_0(t)$ et $dF_1(t)$ dans les deux modèles. De ce fait, une même paire (F_0, F_1) ne peut jamais être solution à la fois de l'une et l'autre équation.

Dans le MM, on peut toujours, comme dans le cas du MSM, associer deux variables aléatoires T_0 et T_1 aux fonctions F_0 et F_1 respectivement. Mais contrairement au cas du MSM, ces deux variables ne sont plus indépendantes, en tant que composantes d'un mélange, et elles n'ont plus d'interprétation en termes de variables latentes. La notation suivante pour la variable T issue du mélange peut alors être utile : $T = T_0 \cup_{p_0} T_1$, ce qui signifie, en posant $\tau(t) = [t, t + dt[$, que $P(T_0 \cup_{p_0} T_1 \in \tau(t)) = p_0 P(T_0 \in \tau(t)) + (1 - p_0) P(T_1 \in \tau(t))$.

L'image à retenir pour ce mécanisme est la suivante : on tire d'abord au sort l'athlète (selon une loi de Bernoulli de probabilité p_0) puis on le fait courir et on note son temps et son identité. Ce mécanisme correspond à des causes de mortalité exclusives. Il est parfois dénommé dans la littérature mécanisme des « causes spécifiques ». Il est dans la plupart des cas plus réaliste que le mécanisme de la sélection par minimum qui suppose l'indépendance des événements. Un exemple typique de ce mécanisme est le cas de l'étude de la mortalité humaine, les différents types correspondant à des causes exclusives de décès.

2.2. Estimation non paramétrique dans le cas du mécanisme de sélection par minimum

Les fonctions que l'on souhaite estimer dans ce cas sont les incidences cumulées correspondant à chaque événement, c'est-à-dire les fonctions de répartition conditionnelles F_0 et F_1. En effet, les événements étant supposés indépendants, cela a du sens de s'intéresser à la fonction de répartition d'un événement d'un type donné. Ces fonctions ne sont pas complètement observables. Par contre, les sous-distributions $U_0(t)$, $U_1(t)$ et leur somme $U(t)$ sont observables. Ces dernières peuvent donc être estimées de façon empirique sur l'échantillon.

T_0 et T_1 étant supposées indépendantes, on a la condition (1) :

$$1 - U(t) = (1 - F_0(t))(1 - F_1(t)).$$

En imposant de plus la condition de non-informativité suivante :

$$\left(\frac{\partial U_0}{\partial F_1}\right)(t) = \left(\frac{\partial U_1}{\partial F_0}\right)(t) \equiv 0, \tag{2}$$

c'est-à-dire qu'une variation de F_0 implique une variation de U_0 seulement et qu'une variation de F_1 implique une variation de U_1 seulement, on obtient (voir les détails de la démonstration en annexe) le système d'équations différentielles suivant en $F_0(t)$ et $F_1(t)$:

$$\begin{cases} dF_0(t) & = \dfrac{dU_0(t)}{1 - F_1(t)} \\ dF_1(t) & = \dfrac{dU_1(t)}{1 - F_0(t)} \end{cases} \tag{3}$$

soit

$$\left\{\begin{array}{ll} dF_0(t) = & (1 - F_0(t))\,\dfrac{dU_0(t)}{1 - U(t)} \\ dF_1(t) = & (1 - F_1(t))\,\dfrac{dU_1(t)}{1 - U(t)} \end{array}\right. \tag{4}$$

Remarquons que la condition de non-informativité (2) est analogue à la condition qui justifie l'usage d'une vraisemblance partielle – excluant les contributions de la censure – dans le cadre traditionnel de l'estimation par maximisation de la vraisemblance. Cette condition n'est autre qu'une condition d'exhaustivité de U_0 pour F_0 et de U_1 pour F_1 au sens de Halmos et Savage (1949).

Le système obtenu peut s'écrire sous la forme :

$$\left\{\begin{array}{l} \dfrac{dF_0(t)}{1 - F_0(t)} = \dfrac{dU_0(t)}{1 - U(t)} \\ \dfrac{dF_1(t)}{1 - F_1(t)} = \dfrac{dU_1(t)}{1 - U(t)} \end{array}\right. \tag{5}$$

qui correspond à l'égalité de deux risques, obtenue du fait de l'indépendance des deux délais d'événement : les membres de gauche sont parfois appelés dans la littérature « risques nets », alors que les membres de droite sont appelés « risques bruts » ou « risques apparents », et l'on voit qu'une résolution formelle est possible et que de plus la partie droite des équations est observable, donc estimable empiriquement. La solution de ce système est :

$$\left\{\begin{array}{lll} F_0(t) & = & 1 - \exp\left(- \displaystyle\int_0^t \frac{dU_0(x)}{1 - U(x)}\right) \\ F_1(t) & = & 1 - \exp\left(- \displaystyle\int_0^t \frac{dU_1(x)}{1 - U(x)}\right) \end{array}\right.$$

Le remplacement de U, U_0 et U_1 dans ces solutions par leurs estimateurs empiriques conduit aux estimateurs de Nelson-Aalen (Nelson, 1972, Aalen, 1978) des fonctions F_0 et F_1.

Nous choisissons plutôt d'obtenir des estimations de F_0 et de F_1 à l'aide d'un algorithme. L'échantillon total est ordonné par valeurs croissantes des temps observés. On note cet échantillon ordonné $(t_i, y_i)_{i=1}^n$. L'ensemble $\{t_1, t_2, ..., t_n\} \times \{0, 1\}$ est probabilisé en donnant la probabilité $\frac{1}{n}$ à chacun des n points observés. On cherche des estimateurs $\hat{F}_0$ et $\hat{F}_1$ de F_0 et F_1 qui vérifient le système (4) à chaque temps t. Bien entendu, ces estimateurs dépendent de n, taille de l'échantillon, mais l'indice n sera omis pour ne pas alourdir les notations.

Cela conduit à :

$$\hat{F}_j(t_i) - \hat{F}_j(t_{i-1}) = (1 - \hat{F}_j(t_{i-1}))\hat{\lambda}_{ji}, \quad j = 0,1, \tag{6}$$

$$\text{avec } \hat{\lambda}_{ji} = \frac{\hat{U}_j(t_i) - \hat{U}_j(t_{i-1})}{1 - \hat{U}(t_{i-1})},$$

avec comme conditions initiales $\hat{F}_j(0) = 0, \ j = 0,1$.

En prenant comme estimateurs de U et des $U_j, \ j = 0,1$, les estimateurs empiriques correspondants, on obtient

$$\hat{\lambda}_{ji} = \frac{\frac{1}{n}\mathbf{1}(y_i = j)}{1 - \sum_{1}^{i-1} \frac{1}{n}} = \frac{\mathbf{1}(y_i = j)}{n - i + 1}.$$

En posant $\hat{S}_j(t_i) = 1 - \hat{F}_j(t_i), \ \ j = 0,1$, on obtient immédiatement à partir de l'équation (6) :

$$\hat{S}_j(t_i) = \hat{S}_j(t_{i-1})(1 - \hat{\lambda}_{ji}),$$

avec comme conditions initiales $\hat{S}_j(0) = 1, \ j = 0,1$. Par récurrence sur i et en étendant le résultat à une fonction en escalier sur $\mathbb{R}$, on obtient alors

$$\widehat{S}_j^{KM}(t) = \prod_{i:t_i \leqslant t} \left(1 - \frac{\mathbf{1}(y_i = j)}{n - i + 1}\right), \quad j = 0,1, \tag{7}$$

qui sont les estimateurs de Kaplan et Meier (1958) des fonctions de survie $1 - F_0(t)$ et $1 - F_1(t)$, obtenus en considérant que les deux événements se censurent mutuellement. Ces estimateurs ont la propriété de vérifier l'équation (1), qui correspond à la condition d'indépendance, à chaque temps, alors que les estimateurs de Nelson-Aalen ne la vérifient qu'asymptotiquement.

Ces mêmes estimations peuvent être obtenues par l'algorithme suivant, obtenu à partir du système (3) et dont l'intérêt apparaîtra plus loin.

Algorithme

Initialisation : $\hat{F}_j(0) = 0, \ \ j = 0,1$.

Pour $i = 1$ à n, $j = 0,1$,

$$\hat{F}_j(t_i) = \hat{F}_j(t_{i-1}) + \frac{1}{n}\frac{\mathbf{1}(y_i = j)}{1 - \hat{F}_{(1-j)}(t_{i-1})}.$$

La convergence de cet algorithme est assurée car il conduit aux estimations de Kaplan-Meier. Les propriétés mathématiques de cet estimateur peuvent être trouvées au chapitre 7 de Shorack et Wellner (1986). L'intérêt de cet algorithme est qu'il s'étend à des modèles beaucoup plus complexes, qui seront vus plus loin. Les estimations des risques instantanés et des densités sont également possibles.

2.3. Estimation non paramétrique dans le cas du modèle de mélange

Dans le cas très simple du mélange de deux distributions non censurées, les proportions p_0 et p_1 peuvent être estimées par les proportions correspondantes observées dans l'échantillon. Dans le cas du mélange, les fonctions d'intérêt sont les incidences cumulées U_0 et U_1, qui correspondent aux incidences de chaque type d'événement, en présence du risque associé à l'autre type d'événement. On ne s'intéresse pas dans ce cas à F_0 et F_1, qui correspondraient à des incidences d'un événement d'un type donné, considéré comme le seul possible. Les fonctions U_0 et U_1 sont observables et peuvent être estimées de façon empirique par :

$$\widehat{U}_k^{Emp}(t) = \frac{1}{n} \sum_{i:t_i \leqslant t} \mathbf{1}(y_i = k), \quad k = 0, 1. \tag{8}$$

3. Plusieurs événements en compétition avec censure

La situation va maintenant se compliquer de la façon suivante : on est en présence de plusieurs événements compétitifs et d'un mécanisme de censure à droite. Un seul délai est observé pour chaque individu : le délai d'un des événements d'intérêt ou le délai de censure.

On est toujours en présence d'un échantillon marqué : $(T_i, Y_i)_{i=1}^n$, avec Y qui prend ses valeurs dans $\{0, 1, \ldots, p\}$. La marque 0 désigne la censure.

Comme précédemment, cet échantillon peut être vu comme le résultat d'un des deux mécanismes de sélection suivants :

- Mécanisme de sélection par minimum (MSM) : c'est la généralisation du mécanisme vu dans le cas de deux événements. On considère que le temps T observé est obtenu comme minimum de $p + 1$ variables latentes que l'on supposera indépendantes, et dont l'une, T_0, est le délai de censure.
- Mécanisme de mélange censuré (MMC) : le temps observé est le minimum d'un délai de censure et d'une variable aléatoire dont la distribution est un mélange à p composantes.

Ces deux mécanismes de sélection conduisent à des estimateurs différents des fonctionnelles d'intérêt. Les deux algorithmes correspondants vont être présentés.

3.1. Mécanisme de sélection par minimum : cas général

Le couple (T, Y) est obtenu de la façon suivante :

$$\begin{aligned} T &= \min(T_0, T_1, T_2, \ldots, T_p) \\ Y &= j \ : \ T = T_j \end{aligned}$$

et les variables T_j sont supposées indépendantes deux à deux. Cette hypothèse forte n'est malheureusement pas testable à cause d'un problème d'identifiabilité (*cf.* Tsiatis, 1975).

La condition caractérisant le mécanisme de sélection s'écrit :

$$1 - U(t) = \prod_{j=0}^{p} (1 - F_j(t)) .$$

Le système d'équations (3) obtenu dans le cas de deux événements sans censure, qui découle de l'hypothèse de non-informativité, se généralise aisément au cas de $p+1$ variables :

$$dF_j(t) = \frac{dU_j(t)}{\prod_{k \neq j}(1 - F_k(t))} = (1 - F_j(t))\frac{dU_j(t)}{1 - U(t)}, \quad j = 1, \ldots, p.$$

En cherchant des estimateurs qui vérifient ce système à chaque temps, on aboutit comme précédemment aux estimateurs de Kaplan-Meier pour les $F_j(t)$, $j = 0, 1, \ldots, p$, en considérant les autres variables comme des censures, en plus de la vraie censure T_0.

L'algorithme se généralise très facilement.

Algorithme

L'échantillon total est ordonné par valeurs croissantes des temps : on le note $(t_i, y_i)_{i=1}^n$.

Initialisation : $\hat{F}_j(0) = 0$, $j = 0, \ldots, p$.

Pour $i = 1$ à n, $j = 0, \ldots, p$,

$$\hat{F}_j(t_i) = \hat{F}_j(t_{i-1}) + \frac{\mathbb{I}(y_i = j)}{n \prod_{k \neq j}(1 - \hat{F}_k(t_{i-1}))}.$$

Bien entendu, les autres fonctionnelles d'intérêt (risques instantanés, densités) peuvent également être obtenues.

3.2. Mécanisme de mélange censuré

La symétrie des $p+1$ variables latentes du mécanisme précédent ne s'applique plus ici. Nous notons donc T_0 le délai de censure et $T_{1,k}, k = 1, \ldots, p$, les p délais des événements (théoriques et sans interprétation propre).

Ces délais $T_{1,k}$ ne sont pas indépendants entre eux conditionnellement à $Y \neq 0$ (en tant que composantes d'un mélange); par contre, on fait l'hypothèse classique d'indépendance du délai de censure T_0 avec chaque $T_{1,k}$.

Le mécanisme de sélection est le suivant :

$$\begin{aligned} T &= \min(T_0, \bigcup_w T_{1,k}) \\ Y &= 0 \text{ si } T = T_0 \\ Y &= k \text{ si } T = T_{1,k}, \end{aligned} \tag{9}$$

où $T_1 = \bigcup_w T_{1,k}$ est le résultat d'un mélange, w désignant une loi de probabilité définissant la composition de ce mélange.

L'image est la suivante : on tire au sort un athlète, on le fait courir, mais il est alors en compétition avec un autre athlète (la censure) qui peut arriver avant lui : on note le temps du gagnant et son identité.

Notons $w_k = P(Y = k \mid Y \neq 0)$, $F_1(t)$ la fonction de répartition de la variable T_1 et $F_{1,k}(t)$, $k = 1, \ldots, p$, les distributions conditionnelles correspondant aux composantes du mélange. On a :

$$F_1(t) = \sum_{k=1}^{p} F_{1,k}(t)\, w_k = \sum_{k=1}^{p} I_{1,k}(t). \tag{10}$$

Il n'est pas nécessaire d'estimer explicitement les w_k, car les quantités qui nous intéressent ici sont les $I_{1,k}(t)$, incidences cumulées des événements de type k en présence des autres types, et non pas les composantes du mélange $F_{1,k}(t)$, qui correspondraient aux incidences des événements de type k, le type k étant le seul possible. Contrairement au cas de deux événements vu précédemment, ces sous-distributions $I_{1,k}(t)$ ne sont plus entièrement observables à cause de la censure.

La condition caractérisant le mécanisme de sélection s'écrit :

$$\begin{aligned} 1 - U(t) &= 1 - \sum_{k=1}^{p} U_{1,k}(t) - U_0(t) \\ &= (1 - \sum_{k=1}^{p} I_{1,k}(t)) \quad (1 - F_0(t)), \end{aligned} \tag{11}$$

où $U_0(t) = P(T < t, Y = 0)$ et $U_{1,k}(t) = P(T < t, Y = k)$, $k = 1, \ldots, p$. Ces dernières quantités sont toujours observables, donc estimables empiriquement.

En ajoutant la condition d'exhaustivité des $U_{1,k}$ pour les $I_{1,k}$ et de U_0 pour F_0, on obtient, comme précédemment :

$$\left\{ \begin{array}{lcl} dI_{1,k}(t) & = & \dfrac{dU_{1,k}(t)}{(1 - F_0(t))}, \quad k = 1, \ldots, p, \\ dF_0(t) & = & \dfrac{dU_0(t)}{(1 - \sum_{k=1}^{p} I_{1,k}(t))}. \end{array} \right. \tag{12}$$

Ce système peut se réécrire :

$$\begin{cases} dI_{1,k}(t) & = & (1 - F_1(t)) \dfrac{dU_{1,k}(t)}{(1 - U(t))}, \quad k = 1, \dots, p, \\ dF_0(t) & = & (1 - F_0(t)) \dfrac{dU_0(t)}{(1 - U(t))}, \end{cases} \tag{13}$$

et sa solution est :

$$I_{1,k}(t) = \int_0^t \frac{dU_{1,k}(x)}{1 - U(x)} \exp\Big(- \int_0^x \frac{dU_1(z)}{1 - U(z)} \Big).$$

Le remplacement de U, U_1 et $U_{1,k}$ dans la formule ci-dessus par leurs estimateurs empiriques produit pour $I_{1,k}$ l'estimateur d'Aalen-Johansen dont les propriétés mathématiques peuvent être trouvées dans Andersen *et al.* (1993, chap. IV.4.).

Estimation

L'ensemble $\{t_1, t_2, ..., t_n\} \times \{0, 1, 2, ..., p\}$ est probabilisé par :

$$\begin{aligned} P(T = t, Y = k) &= \frac{1}{n} \text{ s'il existe } i \text{ tel que } (t, k) = (t_i, y_i), \\ &= 0 \text{ sinon,} \end{aligned}$$

c'est-à-dire qu'à chaque point observé est affectée une probabilité $\frac{1}{n}$.
En utilisant (13), la discrétisation donne, pour $k = 1, \dots, p$:

$$\hat{I}_{1,k}(t_i) - \hat{I}_{1,k}(t_{i-1}) = (1 - \hat{F}_1(t_{i-1}))\, \hat{\lambda}_{1,ki} \tag{14}$$

$$\begin{aligned} \text{avec } \hat{\lambda}_{1,ki} &= \frac{\hat{U}_{1,k}(t_i) - \hat{U}_{1,k}(t_{i-1})}{1 - \hat{U}(t_{i-1})} \\ &= \frac{\mathbf{1}(y_i = k)}{n - i + 1}. \end{aligned}$$

L'estimation $\hat{F}_1(t)$ de $F_1(t) = \sum_{k=1}^p I_{1,k}(t)$ s'obtient en sommant sur k les premières équations de (12), et on aboutit alors au système suivant :

$$\begin{cases} dF_1(t) = \dfrac{dU_1(t)}{1 - F_0(t)} \\ dF_0(t) = \dfrac{dU_0(t)}{1 - F_1(t)} \end{cases}$$

qui est exactement le système (3). L'estimation qui en découle est donc l'estimation de Kaplan-Meier de la fonction de survie $S_1 = 1 - F_1$ (obtenue en pratique en regroupant tous les événements d'intérêt en un seul type).

On aboutit alors, par récurrence sur i de la formule (14), à l'estimateur de l'incidence cumulée pour la cause k ($k = 1, \ldots, p$) proposé de façon heuristique par Kalbfleisch et Prentice (1980, page 169) :

$$\hat{I}^{P_T}_{1,k}(t) = \sum_{i:t_i \leqslant t} \frac{\mathbf{I}(y_i = k)}{n - i + 1} \quad \widehat{S}^{KM}_1(t_i). \tag{15}$$

L'estimation de la fonction de survie $S_0 = 1 - F_0$ de la censure s'obtient comme celle de S_1 par Kaplan-Meier, en considérant tous les autres événements comme censurant la censure.

Algorithme

Initialisation : $\hat{I}_{1,k}(0) = 0$, $k = 1, \ldots, p$, $\hat{F}_0(0) = 0$.
Pour $i = 1$ à n, $k = 1, \ldots, p$,

$$\hat{I}_{1,k}(t_i) = \hat{I}_{1,k}(t_{i-1}) + \frac{\mathbf{I}(y_i = k)}{n\ (1 - \hat{F}_0(t_{i-1}))},$$

$$\hat{F}_0(t_i) = \hat{F}_0(t_{i-1}) + \frac{\mathbf{I}(y_i = 0)}{n \left(1 - \sum_{k=1}^{p} \hat{I}_{1,k}(t_{i-1})\right)}.$$

Les fonctions `S-Plus` qui permettent le calcul des estimateurs présentés pour deux événements en compétition en présence de censure sont disponibles sur la page web de E. Leconte.

3.3. Généralisation des mécanismes MSM et MMC

Dans le mécanisme de mélange censuré, la censure peut également être le résultat d'un mélange (plusieurs causes exclusives de censure). Par exemple, en recherche clinique, on se trouve parfois amené à comparer, dans plusieurs groupes thérapeutiques où les survies sont similaires, les différentes causes de censure à droite (mauvaise tolérance, événement intercurrent, perte de vue...) : une répartition différente de ces causes de censure entre les groupes suggérerait un biais dans la comparaison des mortalités.

Un cas encore plus général est celui où la variable délai observée est le résultat d'une sélection par minimum de plusieurs variables latentes, chacune d'entre elles pouvant être le résultat d'un mélange. C'est le cas lorsque plusieurs types de censures (mélange de censures) sont en compétition avec plusieurs types d'événements (eux-mêmes mélanges d'événements). Le cas du mécanisme de mélange censuré avec la censure résultat d'un mélange envisagé au début de cette section correspond en fait au cas de deux variables latentes seulement. Supposons de façon générale qu'on ait affaire à J variables latentes, chacune résultat d'un mélange de n_j composantes. Le couple (T, Y) est alors obtenu ainsi :

$$T = \min(\bigcup_{w_1} T_{1,k}, \ldots, \bigcup_{w_J} T_{J,k}) \tag{16}$$
$$Y = (j, k) \text{ si } T = T_{j,k}.$$

L'algorithme général pour estimer les fonctions d'incidence cumulées $I_{j,k}$, $j = 1, \ldots, J$, $k = 1, \ldots, n_j$, est le suivant : l'échantillon total, ordonné par valeurs croissantes des temps, est noté $(t_i, y_i)_{i=1}^n$ (notons qu'ici y_i désigne le couple $(j, k)_i$).

Initialisation : $\hat{I}_{j,k}(0) = 0$, $j = 1, ..., J$, $k = 1, ..., n_j$.

Pour $i = 1$ à n, $j = 1, \ldots, J$, $k = 1, \ldots, n_j$,

$$\hat{I}_{j,k}(t_i) = \hat{I}_{j,k}(t_{i-1}) + \frac{\mathbf{1}(y_i = (j,k))}{n \prod_{j' \neq j} (1 - \sum_{l=1}^{n_{j'}} \hat{I}_{j',l}(t_{i-1}))}.$$

Les estimations des incidences cumulées des variables latentes $T_1, \ldots, T_J$ (variables résultant des J mélanges) s'obtiennent par la procédure de Kaplan-Meier, en groupant tous les événements du même mélange j, et en considérant les autres événements comme des censures. L'estimateur obtenu vérifie :

$$\widehat{F}_j^{KM}(t) = \sum_k \widehat{I}_{j,k}^{Pr}(t),$$

où les $\widehat{I}_{j,k}^{Pr}$ sont les estimateurs des incidences cumulées des n_j composantes du mélange j, obtenus par l'algorithme précédent, et généralisant l'estimateur (15).

D'autre part, on peut facilement vérifier que dans le cas général les estimateurs obtenus vérifient à chaque temps t la propriété vérifiée par les quantités théoriques, à savoir :

$$1 - \sum_{j,k} \widehat{U}_{j,k}^{Emp}(t) = \prod_j \left(1 - \sum_k \widehat{I}_{j,k}^{Pr}(t)\right),$$

où les $\widehat{U}_{j,k}^{Emp}$ sont les généralisations des estimateurs définis en (8).

4. Simulation d'un échantillon marqué

4.1. Principe

Sélection par minimum

Pour générer un n-échantillon marqué provenant d'une sélection par minimum, il faut générer $p+1$ variables aléatoires indépendantes (les variables latentes) selon des lois $F_j,\ j=0,\dots,p$: est retenu pour l'individu i le minimum des $p+1$ réalisations et le numéro du type correspondant.

Sélection par minimum et mélange

On se donne une loi de probabilité w pour les types $k,\ k=1,\dots,p$, du mélange et des lois $F_{1,k}$ associées à chacun de ces types. On se donne également F_0, la loi de la censure.

Pour chaque individu i, on tire aléatoirement son type $y_i = k$ suivant la loi w, puis on tire son temps $T_{1,k}$ selon la loi $F_{1,k}$ et son délai de censure T_0 suivant la loi F_0.

Est retenu le minimum de $T_{1,k}$ et T_0 avec l'indicateur de type correspondant (0 ou k).

Les fonctions `S-Plus` qui génèrent ces deux types d'échantillon marqué dans le cas de deux événements en compétition en présence de censure sont disponibles sur la page web de E. Leconte.

4.2. Illustration

Nous avons généré un échantillon de taille 500 provenant d'un mécanisme de sélection par minimum : les deux durées d'événement suivent des lois exponentielles de paramètres respectifs 1 et 2, et elles sont censurées par un délai de censure de loi exponentielle de paramètre 1. Cet échantillon a été analysé par les deux méthodes : la méthode pertinente dans ce cas (algoMSM) et la méthode erronée (algoMMC).

De même, un échantillon de données de 500 individus provenant d'un mécanisme de mélange censuré a été généré : le mélange se compose de deux lois exponentielles de paramètres respectifs 1 et 2 en proportions égales, censurées par un délai de censure de loi exponentielle de paramètre 1. Cet échantillon a également été analysé par les deux méthodes.

La figure 1 montre les courbes d'incidences cumulées estimées dans les 4 cas. Pour visualiser l'erreur faite en postulant un mécanisme qui n'est pas le bon, nous avons tracé en traits pleins les courbes théoriques correspondant aux fonctions d'intérêt pour chaque mécanisme généré. Dans le cas des données provenant d'un MSM, il s'agit des fonctions de répartition marginales F_1 et F_2 des événements de chaque type. Pour les données provenant d'un MMC, il s'agit des incidences cumulées correspondant à chaque type d'événement $I_{1,1}$ et $I_{1,2}$, qui correspondent à des probabilités d'occurrence des événements de chaque type en présence des autres types, et dont la somme donne la fonction de répartition de la variable issue du mélange.

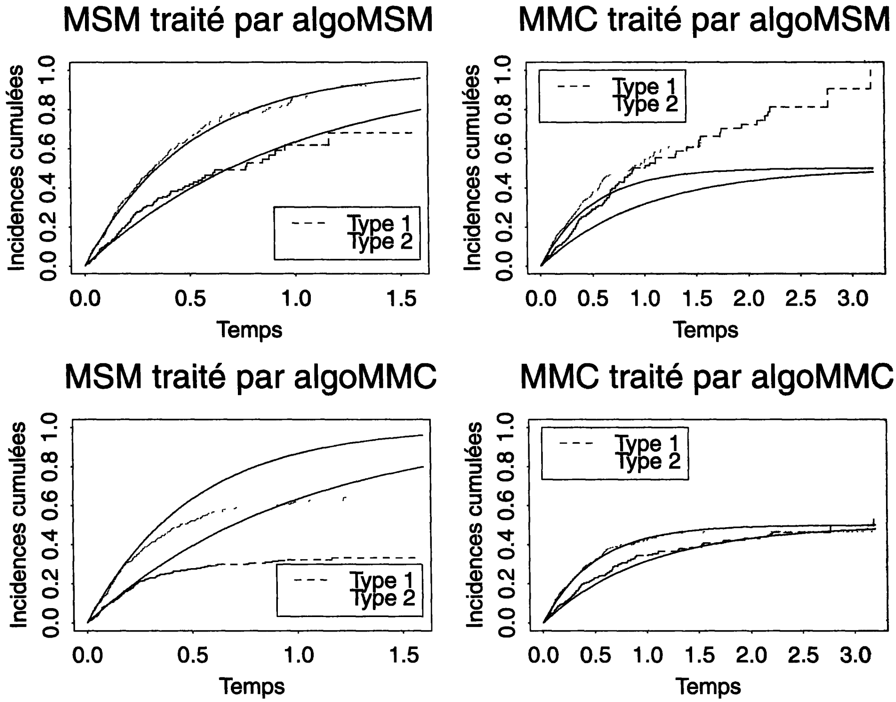

FIG 1. — Estimations des incidences cumulées pour deux échantillons MSM et MMC analysés par les deux algorithmes. Les courbes en traits pleins correspondent aux incidences cumulées théoriques.

On constate tout d'abord que, dans le cas où on applique l'algorithme pertinent, les estimations s'ajustent tout à fait aux courbes théoriques. Par contre, l'utilisation de la «mauvaise méthode» conduit à des courbes très différentes, ce qui s'explique par le fait que l'on n'estime pas les mêmes fonctions. Sur cet exemple, l'analyse de l'échantillon MSM par l'algorithme MMC (en bas à gauche) ne conduit pas à des conclusions fondamentalement différentes (les deux courbes estimées sont simplement plus basses, mais les événements de type 2 ont, pour tous les temps, une incidence cumulée plus forte que celle des événements de type 1, comme les courbes théoriques). Par contre, l'analyse de l'échantillon MMC par l'algorithe MSM (en haut à droite) est plus trompeuse : en effet, les estimations de Kaplan-Meier des fonctions de répartition des deux événements se croisent, si bien qu'on serait amené à conclure que la survie pour l'événement de type 1 est meilleure que pour le type 2 au début, mais que cette tendance s'inverse après le point d'intersection, alors qu'en réalité l'incidence cumulée de l'événement de type 2 est toujours au dessus de celle de l'événement de type 1. Il est difficile de dire dans le cas général quelle erreur est la plus grave, mais il semble malgré tout plus prudent de ne pas considérer comme indépendants des événements qui ne le sont pas, et donc d'utiliser l'algorithme MMC en cas de doute sur l'indépendance.

5. Exemple d'application

Les données de cet exemple (*cf.* tableau 1) sont des données fictives extraites de Marubini et Valsecchi, 1995 (table 10.1 page 332). Elles sont constituées des délais jusqu'à la récidive locale (événement de type 1) ou jusqu'à la métastase (événement de type 2) pour deux groupes de 35 patients chacun. On observe 10 récidives locales et 20 métastases dans chaque groupe. 5 patients par groupe n'ont pas eu d'événement et constituent donc des observations censurées.

TAB 1. — Délais jusqu'à l'apparition d'une récidive locale ou d'une métastase pour 70 patients répartis en deux groupes de traitement A et B (données fictives de Marubini, 1995).

	Groupe A	Groupe B
Récidive locale	1, 13, 17, 30, 34, 41, 78, 100, 119, 169	7, 16, 16, 20, 39, 49, 56, 73, 93, 113
Métastase	1, 6, 8, 13, 13, 15, 33, 37, 44, 45, 63, 80, 80, 89, 91, 132, 144, 171, 183, 240	1, 2, 4, 6, 8, 9, 10, 13, 17, 17, 17, 18, 18, 27, 29, 39, 50, 69, 76, 110
Données censurées	34, 60, 63, 149, 207	34, 60, 63, 78, 149

Les courbes d'incidences cumulées d'événements toutes causes confondues sont estimées par la méthode de Kaplan-Meier (voir figure 2 en haut à gauche), et la différence entre les deux groupes de traitement est statistiquement significative par le test du logrank ($p = 0,02$). L'incidence cumulée à 72 semaines est estimée à 49 % dans le groupe A et à 73 % dans le groupe B.

Suivant qu'on postule pour ces données un mécanisme MSM ou MMC, on ne va pas estimer les mêmes quantités en différentiant les deux types d'événements. Si l'on postule un mécanisme de sélection par minimum avec indépendance des deux délais, on va estimer les fonctions de répartition (ce qui revient à estimer les fonctions de survie marginales) pour le délai sans récidive et pour le délai sans métastase. Ces estimations par événement et par groupe sont présentées à la figure 2. Si l'on postule un mécanisme de mélange censuré, sans donc supposer l'indépendance des délais, on va estimer les sous-distributions $I_{1A}(t)$, $I_{2A}(t)$, $I_{1B}(t)$ et $I_{2B}(t)$ correspondant pour chaque groupe à chaque cause en présence de l'autre cause, dont la somme par groupe à chaque temps donne la fonction de répartition toutes causes par groupe. Ces différentes courbes donnent des impressions visuelles différentes, en particulier pour l'événement récidive, où les incidences cumulées estimées par la méthode MMC ne semblent pas différer entre les deux groupes, alors que les estimations faites par la méthode MSM diffèrent clairement. Quels que soient le mécanisme

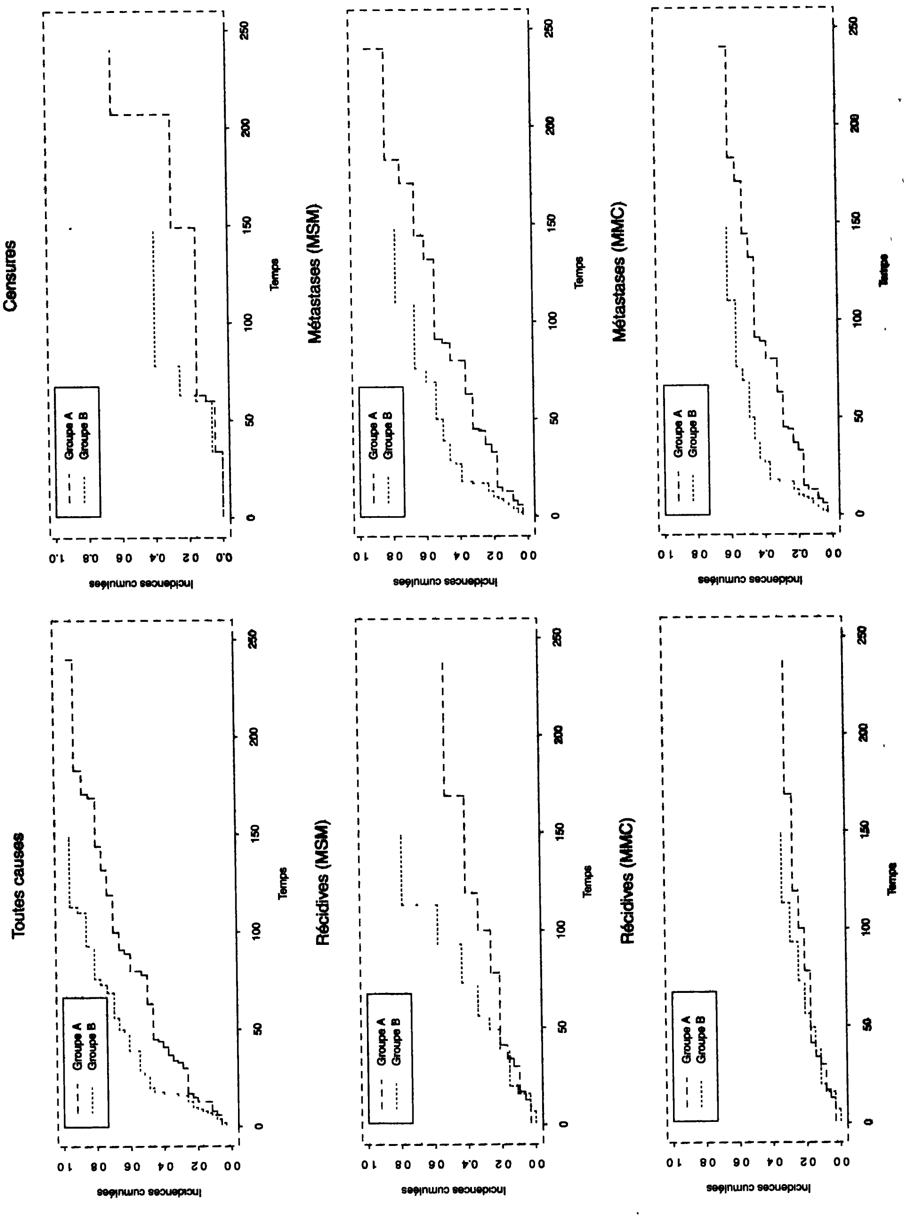

FIG 2. — Estimations des incidences cumulées par groupe pour le mécanisme MSM et le mécanisme MMC.

postulé et la méthode d'estimation utilisée, la fonction de répartition du délai de censure sera estimée de la même manière, par l'estimation de Kaplan-Meier, en considérant les délais d'événements comme des censures du délai de censure.

Les résultats des estimations à 72 semaines en utilisant les algorithmes présentés précédemment sont dans le tableau 2. Pour le MSM, on retrouve la propriété de l'estimateur de Kaplan-Meier de vérifier la condition d'indépendance à chaque temps : le produit des estimations des survies marginales pour chaque cause est égal à l'estimation de la survie globale (par exemple, à 72 semaines, $(1 - 0,206) \times (1 - 0,363) = 1 - 0,494$). Pour le MMC, on retrouve qu'à chaque temps, la somme des estimations des incidences cumulées spécifiques vaut l'estimation de l'incidence cumulée globale (par exemple, à 72 semaines, $0,173 + 0,321 = 0,494$).

TAB 2. — Estimations à 72 semaines des incidences cumulées pour les données de Marubini selon le mécanisme postulé.

	MSM		MMC	
Incidences cumulées	Groupe A	Groupe B	Groupe A	Groupe B
Toutes causes	0,494	0,733	0,494	0,733
Récidives locales	0,206	0,337	0,173	0,206
Métastases	0,363	0,598	0,321	0,528

6. Choix du mécanisme de sélection

Comme nous l'avons montré sur les exemples précédents, le choix du mécanisme de sélection conduit à des estimations différentes. La question pratique principale reste donc de spécifier correctement le modèle de sélection, sachant que sa validité ne peut être testée. Il n'y a pas de réponse universelle à cette question, bien que les arguments pour ce choix existent souvent, spécialement en faveur du modèle de mélange censuré, qui ne suppose pas l'indépendance des événements. C'est typiquement le cas lorsque la liaison entre les variables d'intérêt est un phénomène «de nature», comme dans le schéma dit «à trois états», où un patient à risque peut soit devenir malade puis décéder, soit décéder directement, lorsqu'on ne s'intéresse qu'à estimer l'incidence du premier événement : un exemple de ce schéma est celui des accidents cardiovasculaires (infarctus du myocarde, accident vasculaire cérébral, hémorragie), qui peuvent apparaître chez des patients à risque, et devenir éventuellement fatals, ou s'avérer fatals d'emblée. L'indépendance entre les occurrences d'un accident non fatal et du décès chez un patient est une supposition très peu raisonnable, car elle implique que la maladie et le décès puissent survenir chez un patient dans un ordre quelconque ! On peut difficilement concevoir une méthode d'estimation pour l'incidence de l'événement non fatal qui ne prendrait pas en compte l'existence de l'événement fatal. L'un des problèmes originels des risques compétitifs (rappelé en introduction) équivaut sur cet exemple à vouloir prévoir l'incidence d'événements non fatals, dans l'hypothèse où une nouvelle technologie médicale viendrait supprimer le risque d'événement fatal. Cette incidence ne pourrait être estimée qu'en suppposant un modèle à variables latentes avec indépendance des événements, mais ce modèle est ici a priori invalide; au contraire, la considération d'un mélange d'événements cardio-vasculaires non fatals et fatals, qui ne suppose pas l'indépendance, semble refléter correctement le contexte.

La popularité du modèle des variables latentes vient très probablement de la facilité d'estimation des fonctions de survie correspondantes par la méthode de Kaplan-Meier, présente dans tous les logiciels. D'autre part, ce modèle fournit le résultat le plus complet, à savoir la distribution multivariée des variables latentes, et l'utilisateur oublie facilement sous quelles conditions drastiques et irréalistes ce résultat est valide.

Parallèlement, c'est l'absence de logiciel pour l'estimateur de Prentice qui a certainement limité son utilisation pratique, alors que le mécanisme associé de mélange censuré est toujours valide, souvent plus réaliste, car prenant en compte les inter-relations de nature entre les variables; enfin, il fournit à l'expérimentateur les estimateurs des seules fonctions généralement estimables sans autre hypothèse, à savoir les incidences de chaque type d'événement d'intérêt en présence de tous les autres événements d'intérêt et de la censure indépendante.

Remarquons enfin que la distinction entre les deux modèles de sélection n'a pas de conséquence en terme de tests non paramétriques d'hypothèse nulle, lorsqu'on souhaite comparer différents groupes indépendants. Le test du logrank ne dépend en effet que des risques empiriques apparents, qui sont identiques dans les deux modèles de sélection. Cette remarque ne fait qu'illustrer l'impossibilité de tester le choix du modèle de sélection.

7. Quelques remarques

L'indépendance de la censure et des événements d'intérêt est une autre condition de validité des modèles à risques compétitifs; en fait la censure n'est pas un phénomène « de nature », mais plutôt une nécessité expérimentale; ce n'est pas non plus un phénomène d'intérêt (il peut l'être d'une manière accessoire seulement) ; aussi cherche-t-on à estimer, indépendamment de la censure, l'incidence du phénomène d'intérêt et, dans tout dispositif expérimental avec censure, à assurer cette indépendance, en vue d'utiliser pour l'estimation un modèle à variables latentes; ceci est souvent possible d'une manière presque parfaite : c'est le cas lorsque, dans une étude clinique comparative, les patients sont inclus à mesure qu'ils se présentent et deviennent éligibles, et l'étude se termine à la même date (fixée à l'avance) pour tous. En revanche, on ne peut garantir que les patients perdus de vue sont censurés indépendamment des événements d'intérêt, ce qui justifie le grand soin apporté généralement à limiter au maximum le nombre des perdus de vue, et à documenter la raison pour laquelle ces patients sont perdus de vue (la nature de ces raisons permet d'évaluer subjectivement le degré de dépendance).

Il est intéressant de remarquer que les techniques développées ici conduisent à des solutions de nature empirique, mais ne procèdent pas par maximisation de la vraisemblance empirique; indirectement néanmoins, les estimations $\widehat{F}_j^{KM}(t)$ et $\widehat{I}_{1,k}^{Pr}(t)$ sont obtenues à partir des estimations empiriques des distributions et sous-distributions $U(t)$ et $U_{1,k}(t)$, qui maximisent la vraisemblance empirique (Tapia et Thompson, 1978), par résolution numérique d'un système différentiel, et sont également les solutions du maximum de vraisemblance empirique pour des données censurées (Kalbfleisch et Prentice, 1980).

Outre la condition d'indépendance de la censure et des événements, c'est une condition de non-informativité mutuelle qui assure l'unicité de la solution non paramétrique dans nos deux modèles. Cette condition est automatiquement vérifiée par des estimateurs empiriques, c'est-à-dire des estimateurs qui donnent uniquement du poids aux points observés.

L'exposé a été limité à l'estimation d'incidences cumulées, par souci de concision, mais les algorithmes fournissent de fait des estimations non paramétriques pour les (sous-)densités et les (sous-)risques instantanés. Ces estimateurs se prêtent comme toujours au lissage. On peut pour cela utiliser des lisseurs à noyau (par exemple lisseur à noyau discret, Derzko, 1998) pour l'estimation non paramétrique des (sous)-densités $\dfrac{dU_k(t)}{dt}$ et $\dfrac{dU_{1,k}(t)}{dt}$: des estimations lissées pour $\dfrac{dF_j(t)}{dt}$ et $\dfrac{dI_{1,k}(t)}{dt}$ s'obtiennent alors par application des algorithmes proposés. De plus, ils permettent sans difficulté l'estimation non paramétrique en présence de censure par intervalle.

Enfin, il est aisé de s'assurer que les procédures d'estimation non paramétriques développées dans le cas d'un modèle à variables latentes (*cf.* section 3.1) exigent que l'on puisse ranger sans ambiguïté les délais d'événements correspondant à des valeurs différentes de Y. Cela peut ne pas être le cas lorsque l'échantillon contient des délais d'événements ex æquo. Il est souvent recommandé de « lever l'incertitude » de façon conventionnelle (par exemple en plaçant les délais de censures ex æquo avec des délais d'événements après ces derniers) mais il est préférable de prendre cette incertitude en considération, au moyen des estimations obtenues soit avec tous les ordres de transitions compatibles avec les ex æquo, soit en échantillonnant parmi ces ordres (si les possibilités sont trop nombreuses). En revanche, les procédures d'estimation sont indifférentes aux ex æquo survenant entre des délais d'événements correspondant à différentes valeurs de la variable K, caractérisant les composantes d'un mélange à J fixé (*cf.* section 3.3). La formule d'estimation (15) est simplement à adapter pour prendre en compte la possibilité d'ex æquo, en remplaçant l'indicatrice par une fréquence.

Annexe

Obtention du système (3) d'équations différentielles en F_0 et F_1 :
En posant $\tau(t) = [t, t + dt[$, $dU(t) = P(T \in \tau(t))$, $dF_0(t) = P(T \in \tau(t) \mid Y = 0)$, et $dF_1(t) = P(T \in \tau(t) \mid Y = 1)$, on obtient par dérivation :

$$dU(t) = (1 - F_1(t))\, dF_0(t) + (1 - F_0(t))\, dF_1(t). \tag{17}$$

Posons aussi $dU_0(t) = P(T \in \tau(t)\ ,\ Y = 0)$ et $dU_1(t) = P(T \in \tau(t)\ ,\ Y = 1)$. On a alors, comme $U(t) = U_0(t) + U_1(t)$,

$$dU(t) = dU_0(t) + dU_1(t) \tag{18}$$

mais aussi :

$$\begin{cases} dU_0(t) = \left(\dfrac{\partial U_0}{\partial F_0}\right)(t)\, dF_0(t) + \left(\dfrac{\partial U_0}{\partial F_1}\right)(t)\, dF_1(t) \\ dU_1(t) = \left(\dfrac{\partial U_1}{\partial F_0}\right)(t)\, dF_0(t) + \left(\dfrac{\partial U_1}{\partial F_1}\right)(t)\, dF_1(t) \end{cases} \tag{19}$$

Si l'on impose la condition de non-informativité suivante :

$$\left(\frac{\partial U_0}{\partial F_1}\right)(t) = \left(\frac{\partial U_1}{\partial F_0}\right)(t) \equiv 0 \tag{20}$$

alors

$$\begin{cases} dU_0(t) = \left(\dfrac{\partial U_0}{\partial F_0}\right)(t)\, dF_0(t) \\ dU_1(t) = \left(\dfrac{\partial U_1}{\partial F_1}\right)(t)\, dF_1(t) \end{cases} \tag{21}$$

À partir de (18) et (21), on obtient :

$$dU(t) = \left(\frac{\partial U_0}{\partial F_0}\right)(t)\, dF_0(t) + \left(\frac{\partial U_1}{\partial F_1}\right)(t)\, dF_1(t) \tag{22}$$

Du fait de l'unicité d'une telle décomposition, on obtient de (17) et (22) :

$$\begin{cases} \left(\dfrac{\partial U_0}{\partial F_0}\right)(t) & = & 1 - F_1(t) \\ \left(\dfrac{\partial U_1}{\partial F_1}\right)(t) & = & 1 - F_0(t) \end{cases} \tag{23}$$

soit finalement, par (21) et (23) :

$$\begin{cases} dU_0(t) = (1 - F_1(t))\, dF_0(t) \\ dU_1(t) = (1 - F_0(t))\, dF_1(t) \end{cases} \tag{24}$$

d'où découle le système d'équations différentielles suivant en $F_0(t)$ et $F_1(t)$:

$$\begin{cases} dF_0(t) & = \dfrac{dU_0(t)}{1 - F_1(t)} \\ dF_1(t) & = \dfrac{dU_1(t)}{1 - F_0(t)} \end{cases}$$

avec les conditions initiales $F_0(0) = F_1(0) = 0$.

Références

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

ANDERSEN P. K., BORGAN O., GILL R. D. et KEIDING N. (1993), *Statistical Models Based on Counting Processes.* New-York : Springer-Verlag.

ARRIAGADA R., RUTQVIST L. E., KRAMAR A. et JOHANSSON H. (1992), Competing risks determining event-free survival in early breast cancer, *Br. J. Cancer* 66, 951-957.

BETENSKY R. A. et SCHOENFELD D. A. (2001), Nonparametric estimation in a cure model with random cure times, *Biometrics* 57, 282-286.

DERZKO G. (1998), Une approche intrinsèque de l'estimation non paramétrique de la densité, *C. R. Acad. Sci. Paris, Série I* 327, 985-988.

DERZKO G. (2000), Estimation non paramétrique de lois de survie, *Actes des XXXII^{èmes} Journées de Statistique*, Fès (Maroc).

GOOLEY T. A., LEISENRING W., CROWLEY J. et STORER B. E. (1999), Estimation of failure probabilities in the presence fo competing risks : new representations of old estimators, *Statistics in Medicine* 18, 695-706.

HALMOS P. R. et SAVAGE L. J. (1949), Application of the Radon-Nikodym theorem to the theory of sufficient statistics, *Ann. Math. Statist.* 20, 225-241.

KALBFLEISCH J. D. et PRENTICE R. L. (1980), *The statistical analysis of failure time data.* John Wiley & Sons, New York.

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

MARUBINI E. et VALSECCHI M. G. (1995), *Analysing data from clinical trials and observationals studies.* Statistics in Practice. Wiley.

NELSON W. (1972), Theory and applications of hazard plotting for censored failure data, *Technometrics* 14, 945-965.

SHORACK G. R. et WELLNER J. A. (1986), *Empirical processes with applications to statistics.* John Wiley & Sons, New York.

TAPIA R. A. et THOMPSON J. R. (1978), *Nonparametric probability density estimation.* John Hopkings University Press.

TSIATIS A. (1975), A non-identifiability aspect of the problem of competing risks, *Proceedings of the National Academy of Science* 72, 20-22.