MODÈLES DE DURÉES

Jean-David Fermanian

Cours ENSAE 3ième année

3 *avenue Pierre Larousse, 92245 Malakoff cedex, France.*

Table des matières

1 Le cadre théorique **3**
- 1.1 Qu'est-ce qu'un modèle de durée? . . . 3
- 1.2 Les outils . . . 6
- 1.3 Censures et troncatures . . . 7
  - 1.3.1 Les censures . . . 7
  - 1.3.2 les troncatures . . . 9
- 1.4 Effet des censures et troncatures . . . 9
  - 1.4.1 Identifiabilité . . . 9
  - 1.4.2 Vraisemblance des observations . . . 11
- 1.5 Hétérogénité . . . 13

2 Les estimateurs de Kaplan-Meier (1958) et Nelson-Aalen **16**
- 2.1 De la survie empirique à Kaplan-Meier . . . 16
- 2.2 Propriétés de $\hat{S}_{KM}$ . . . 18
  - 2.2.1 Cohérence . . . 18
  - 2.2.2 Lois fortes . . . 20
  - 2.2.3 Convergence en loi . . . 21
- 2.3 $\hat{S}_{KM}$ estimateur non paramétrique du maximum de vraisemblance . . . 23
- 2.4 L'estimation du hasard intégré: Nelson-Aalen . . . 24
- 2.5 L'estimation sur données tronquées . . . 25

3 Les types de modèles **27**
- 3.1 Les modèles paramétriques . . . 27
- 3.2 Les modèles semiparamétriques . . . 31
  - 3.2.1 Les modèles à hasards proportionnels . . . 32
  - 3.2.2 Modèles à temps accéléré . . . 32
- 3.3 Estimation nonparamétrique de la densité et de la fonction de hasard . . . 34
- 3.4 Estimation nonparamétrique en présence de covariables . . . 38

4 Le modèle à hasards proportionnels **39**
- 4.1 Estimation paramétrique du modèle de Cox . . . 40
- 4.2 La vraisemblance partielle . . . 41
- 4.3 Propriétés asymptotiques du Modèle de Cox . . . 44
- 4.4 Estimation de la survie et de la fonction de hasard intégrée "de base" . . . 46
- 4.5 Le problème des ex-aequo . . . 48
- 4.6 Modèle de Cox discret . . . 49
- 4.7 L'interprétation en termes de vraisemblance marginale . . . 51

5 Autres modèles de régression **54**

6 Modèles à risques concurrents **58**
- 6.1 Représentation en termes de "fonctions spécifiques" . . . 58
- 6.2 Représentation en termes de variables latentes . . . 61
- 6.3 Identifiabilité dans les modèles à risques concurrents . . . 61

7 L'approche par processus ponctuels **64**

8 Annexes **72**
- 8.1 L'intégrale de Lebesgue-Stieltjes . . . 72
- 8.2 Convergence faible des processus à temps continu . . . 73
  - 8.2.1 Application à l'espace des fonctions continues sur [0,1] . . . 75
  - 8.2.2 Application à l'espace des fonctions cadlag . . . 76
  - 8.2.3 L'approche de D par la norme uniforme (Pollard, 1984) . . . 78

8.3 Rappels sur les processus 78
8.4 Familles de loi paramétriques utiles 80

9 Références **82**

1 Le cadre théorique

1.1 Qu'est-ce qu'un modèle de durée?

Les modèles de durées sont adaptés dès que la ou les phénomènes d'intérêt se modélisent comme des variables aléatoires positives. Il s'agit plus généralement de modéliser et d'estimer les lois décrivant le temps qui s'écoule entre deux événements: durée de vie d'un individu ou d'un système physique, durée entre le déclenchement d'une maladie et la guérison, durée d'un épisode de chômage, durée entre la demande d'un prêt et une défaillance de remboursement...

A l'origine liés aux applications en biologie, en médecine (biostatistiques, épidémiologie), et en démographie (espérance de vie aux divers âges, âge au mariage), les modèles de durées se sont révélés d'usage courant aujourd'hui en économie (analyse du marché du travail, durées de vie des entreprises), en finance (défaillances de crédit), en fiabilité (durée de vie de composants industriels) etc.

Ce domaine de recherche est très actif depuis une trentaine d'années; un grand nombre d'articles sont publiés annuellement dans la plupart des revues internationales de statistiques, qu'elles soient inspirées par les biostatistiques (Biometrika, Biometrics) ou généralistes (The Annals of Statistics, JASA, Journal of Multivariate Analysis...). A la suite des illustres initiateurs de la discipline sous sa forme moderne (Sir D.R. Cox, J. Crowley, R. Prentice...), des mathématiciens et statisticiens de premier plan y ont apporté des contributions importantes (R. Gill, P. Bickel, J. Wellner). La technicité de outils et des résultats s'est du coup fortement accrue depuis une quinzaine d'années.

Grossièrement, deux écoles sont actuellement en regard. Ces deux écoles s'attachent à résoudre souvent les mêmes problèmes mais dans des cadres conceptuels sensiblement différents: l'école scandinave (qui a largement débordé cette zone géographique, où elle a pris naissance) aborde les modèles de durées sous forme de processus ponctuels, i.e. de processus stochastiques à temps continu prenant des valeurs discrètes. Cette méthode permet l'usage de résultats puissants concernant les martingales (cf. section 7). L'école classique, elle, reste fidèle aux formulations originelles des modèles, et utilise des méthodes d'inférence statistique plus traditionnelles [1]. Certains résultats sont plus faciles à prouver, à présenter et/ou à généraliser selon le cadre dans lequel on se place. Il en découle une "saine émulation" au sein de la communauté scientifique.

Pourquoi a-t-on constitué une branche particulière de la statistique dans le but d'étudier les variables aléatoires positives? Essentiellement trois arguments sont avancés.

Tout d'abord, les donnée traitées sont rarement "complètes", c'est-à-dire qu'elles sont soumises à des problèmes d'observation qui compliquent sérieusement l'analyse: censures et troncatures sont les perturbations les plus connues. Ces perturbations nécessitent de développer des outils ad-hoc et/ou de revoir entièrement les méthodes classiques de la statistique mathématique.

De plus, les distributions paramétriques usuelles sont rarement adaptées aux phénomènes temporels étudiés; elles sont souvent centrées, ou bien leurs coefficients d'asymétrie (skewness) sont positifs, alors que les durées de vie présentent souvent des coefficients d'asymétrie négatifs. C'est pourquoi des méthodes semi et non-paramétriques ad-hoc ont été développées pour estimer de manière plus réaliste les distributions des durées (cf infra).

La figure 1 illustre ces handicaps des familles paramétriques classiques pour modéliser la distribution des durée de vie des Danois (estimation par maximum de vraisemblance; tiré de Hougaard (1999)).

1. mais pas forcément plus faibles, grâce notamment aux résultats théoriques sur les processus empiriques

Enfin, le fait que les observations se déroulent selon un processus temporel. Elles sont observées séquentiellement, introduit une dimension supplémentaire dans l'analyse. Cela permet l'introduction naturelle des statistiques d'ordre, des processus ponctuels etc.

FIG. 1 – *Estimations de la distribution des durées de vie des Danois (tiré de Hougaard (1999))*

1.2 Les outils

Soit la variable aléatoire univariée T. Usuellement, la distribution de T est décrite par la densité f ou sa fonction de répartition F. Lorsque T est positive, trois autres fonctions sont d'un usage central dans l'analyse des données de survie :

- la fonction de survie $S(t) = P(T > t)$,
- la fonction de hasard $\lambda(t) = f(t)/S(t)$,
- la fonction de hasard intégrée $\Lambda(t) = \int_0^t \lambda$.

La fonction de hasard au point t s'interprète comme la probabilité instantanée de sortir de l'état à la date t, sachant que le sujet est encore dans cet état en t, soit

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(T \in [t, t+\Delta t] | T > t). \tag{1-1}$$

Selon les domaines étudiés, les auteurs appellent λ le taux de panne, taux de décès, de sortie du chômage, de guérison etc. On s'inspirera souvent du vocabulaire démographique et médical ; on parlera alors de "survie", "décès" etc.

Par ailleurs les quantités suivantes sont parfois utilisées :

- la fonction de survie conditionnelle $S(t|t_0) = P(T > t + t_0 | T > t_0)$,
- la durée de vie moyenne restante $r(t) = E[T - t | T > t]$.

Proposition 1.1 *Les fonctions S, λ, Λ, $S(\cdot|\cdot)$ ou r caractérisent entièrement la distribution de T.*

Cette propriété implique qu'on peut raisonner indifféremment avec l'une ou l'autre de ces fonctions. En pratique, l'objet d'étude détermine la fonction adaptée au problème posé.

`Exemple.` **1.1** Supposons qu'on s'intéresse à l'espérance de vie aux divers âges. Ici, la durée T étudiée est donc la durée de l'existence, i.e. l'intervalle de temps entre la naissance et le décès. Si la durée de vie moyenne dans une population est de 80 ans et un mois, que peut tirer de cette information quelqu'un qui fête sont 80ième anniversaire? Rien. Ce qui intéresse cet individu est de savoir combien de temps en moyenne il lui reste à vivre sachant qu'il a 80 ans. La quantité pertinente est donc la distribution de T conditionnellement au fait que $T > 80$. Or, $r(t)$ et $\lambda(t)$ restent invariants au conditionnement par l'événement $T > 80$, à l'inverse de la densité (qui se trouve divisée par $P(T > 80)$). Ils constituent donc les descripteurs "naturels" du phénomène temporel qu'on cherche à appréhender.

L'exemple suivant illustre l'intérêt d'analyser une distribution avec la fonction de hasard plutôt qu'avec la densité.

`Exemple.` **1.2** Pour détecter les patients susceptibles de subir une attaque cardiaque, on leur conseille d'effectuer des visites médicales à certaines dates fixes. Pour déterminer ces dates, un examen de la fonction de hasard de la date d'apparition d'un accident cardiaque suffit en première approximation : avant chaque pic de λ, il convient d'effectuer une visite médicale.

Les fonctions précédentes sont évidemment reliées les unes aux autres par certaines relations.

Proposition 1.2 *Pour tout t tel que $P(T > t) > 0$ et tout $t_0 \leq t$, on a*

$$S(t) = \exp(-\int_0^t \lambda) = \exp(-\Lambda(t)), \tag{1-2}$$

$$S(t|t_0) = \exp(-\int_{t_0}^{t+t_0} \lambda) = \frac{S(t+t_0)}{S(t_0)}, \tag{1-3}$$

et si $\lim_{u\to+\infty} uS(u) = 0$, *alors*

$$r(t) = \frac{1}{S(t)} \int_t^{+\infty} S. \tag{1-4}$$

Preuve La seule relation non triviale est celle concernant r.

$$\begin{aligned} r(t) = E[T-t|T>t] &= \frac{E[(T-t).\mathbf{1}\{T>t\}]}{P(T>t)} \\ &= \frac{1}{S(t)} \left\{ \int_t^{+\infty} uf(u)\,du - tS(t) \right\} \\ &= \frac{1}{S(t)} \left\{ [-uS(u)]_t^{+\infty} + \int_t^{+\infty} S - tS(t) \right\} \end{aligned}$$

Si $\lim_{u\to\infty} uS(u) = 0$ alors on obtient le résultat recherché. □

`Exercice 1.1` *Montrer que la condition précédente* $\lim_{u\to\infty} uS(u) = 0$ *est nécessaire (indication : considérer une loi de Cauchy restreinte à* $\mathbb{R}^+$*).*

`Remarque` **1.1.** *Toutes les fonctions précédentes sont définies a priori sur* $\mathbb{R}$*, mais seul leur comportement sur* $\mathbb{R}^+$ *est pertinent. C'est pourquoi nous pourrions ne considérer que des fonctions définies sur* $\mathbb{R}^+$*. Ainsi, il est suffisant de considérer les variables aléatoires positives dont la densité existe sur* $\mathbb{R}_*^+$*. Notons que* S *n'est pas forcément dérivable à droite en* 0*. Le point* 0 *constitue souvent un cas particulier à traiter à part. On suppose le plus souvent que l'événement* $T = 0$ *est impossible, i.e. est de mesure nulle.*

L'exemple le plus simple nous est fourni par la distribution exponentielle. Cette dernière est définie par la relation : pour tout $t \geq 0$, $f(t) = \mu \exp(-\mu t)$. On en déduit
$\lambda(t) = \mu$, $\Lambda(t) = \mu t$, $S(t) = \exp(-\mu t) = S(t|t_0)$ pour tout $t_0 \geq 0$, $r(t) = 1/\mu$.

Les distributions exponentielles sont caractérisées par le fait que la fonction de hasard est constante. C'est ce qu'on qu'on appelle "la relation d'indépendance temporelle". Elle signifie qu'à n'importe quelle date, la probabilité de décéder est la même, sachant qu'on a vécu jusque là. On en déduit aisément

Proposition 1.3 *Une distribution possède la propriété d'indépendance temporelle si et seulement si elle est exponentielle, ou si sa fonction de survie est exponentielle, ou si* r *est une fonction constante, ou si, pour tout couple de réels positifs* (t_0,t_1), $S(t_0+t_1) = S(t_0)S(t_1)$.

Les fonctions et concepts introduits précédemment ne s'étendent pas toujours aisément dans un cadre multivarié. C'est le cas en particulier des fonctions de hasard : voir Fermanian (1997).

1.3 Censures et troncatures

Elles sont particulièrement typiques des données de survie. Elles proviennent du fait qu'on n'a pas accès à toute l'observation : au lieu d'observer des réalisations i.i.d. de durées T, on observe la réalisation de la variable aléatoire T soumise à diverses perturbations, indépendantes ou non du phénomène étudié. Ces perturbations peuvent pour la plupart être regroupées en deux grandes familles.

1.3.1 Les censures

Soit une variable aléatoire positive C. La durée T est dite censurée à droite si, au lieu de T, on observe le couple (X,δ) avec $X = \inf(T,C)$ et $\delta = \mathbf{1}\{T \leq C\}$. La durée T est dite censurée

à gauche si, au lieu de T, on observe le couple (X,δ) avec $X = \sup(T,C)$ et $\delta = \mathbf{1}\{T \leq C\}$. On notera $a \wedge b = \inf(a,b)$ et $a \vee b = \sup(a,b)$.

Remarque 1.2. *Ces deux cas peuvent être combinés ; alors, on dispose de deux censures C_1 et C_2, l'une à droite et l'autre à gauche, avec, $C_1 < C_2$. Au lieu de T, on observe le triplet (X,δ_1,δ_2) avec : $\delta_1 = \mathbf{1}\{T \leq C_1\}$, $\delta_2 = \mathbf{1}\{T \leq C_2\}$, et*

$$\begin{cases} X = C_1 & si\ T \leq C_1, \\ X = T & si\ C_1 < T \leq C_2, \\ X = C_2 & si\ C_2 < T. \end{cases}$$

Notons que les variables de censures peuvent être éventuellement dégénérées, i.e. peuvent être constantes et induire des masses de Dirac en un point fixe (on parle alors de censure fixe).

Typiquement, lorsque la durée d'intérêt T est censurée par C, on observe un échantillon i.i.d. $(X_i,\delta_i)_{i=1,\ldots,n}$. Il y a alors deux types d'observations : celles pour lesquelles $\delta_i = 1$ (d'où $X_i = T_i$), qui sont des "vraies" durées ou durées complètes, et celles pour lesquelles $\delta_i = 0$ (d'où $X_i = C_i$), qui sont des observations de censures (durées censurées ou incomplètes).

Exemple. 1.3 On cherche à étudier l'apprentissage de la marche chez les jeunes enfants. Pour cela, on décide de suivre tous les enfants d'une ville moyenne entre 10 à 16 mois. La durée d'intérêt est donc ici l'âge d'acquisition de la marche (ou l'intervalle de temps entre la naissance et les premiers pas). Sur l'échantillon examiné, trois cas se produisent : soit l'enfant sait déjà marcher avant 10 mois (censure fixe à gauche), soit il apprend à marcher entre 10 et 16 mois (données complètes, constatées), soit il ne sait toujours pas marcher à 16 mois (censure fixe à droite).

Exemple. 1.4 Il s'agit du cas de données tirées d'un panel de salariés de durée finie $T_0 = 3$ ans (par exemple, l'enquête Emploi de l'INSEE). Imaginons qu'on s'intéresse ici à la durée des épisodes de chômage. On observe le devenir de chômeurs depuis leur inscription à l'ANPE (interrogation tous les 3 mois) jusqu'à ce qu'ils disparaissent du panel. T sera alors la durée de l'épisode de chômage, qui se terminera lorsque l'individu trouvera un nouvel emploi, mais également s'il quitte le marché du travail, se met à son compte etc. Certains de ces chômeurs disparaîtront avant les trois ans de l'enquête (déménagement, décès...) ; c'est ce qu'on appelle le phénomène d'attrition, classique sur les données de panel. D'autres n'auront toujours pas trouvé d'emploi à la fin de l'enquête. On est donc en présence de données censurées par une variable aléatoire dont le support est $[0,T_0]$ (sa distribution est continue sur $[0,T_0[$, et possède une masse non nulle en T_0).

Remarque 1.3. *Sur l'exemple précédent, on constate l'importance du choix de l'origine des temps. Ainsi, supposons que l'origine des temps est la même date t^* pour tous les individus, ces derniers constituent alors le stock de chômeurs à cette date. La durée T d'intérêt ne serait alors plus la durée de l'épisode de chômage mais en fait T', durée de chômage restante, conditionnellement au fait d'être au chômage en t^*. Cette quantité est nettement moins intéressante pour nous que T. L'inférence statistique de la loi de durée de l'épisode de chômage à partir d'observations d'un échantillon tiré selon la loi de T', bien que possible, en est alors sérieusement compliquée. C'est le problème dit du "stock sampling", ou du "biais de sélection": voir, entre autres, Gill et al. (1988), Gouriéroux et Monfort (1989,1991), Manski et Lerman (1977).*

Il existe de nombreux autres types de censures, parmi lesquels

- censures progressives de type I : chaque individu i est soumis à une censure constante C_i connue. C'est le cas en particulier lorsque le processus temporel est observé entre deux dates fixes t_1^* et t_2^*, et lorsque les individus rentrent dans l'échantillon à tour de rôle. Ainsi, si la durée de l'individu i débute en $t_i \in [t_1^*,t_2^*]$, sa censure vaudra $C_i = t_2^* - t_i$.
- censures de type II, ou au r-ième décès : le processus d'observation se déroule jusqu'au r-ième décès. Précisément, si on note $T_{(1)} < T_{(2)} < \cdots < T_{(n)}$ les statistiques d'ordres, on observe en fait l'échantillon $(X_i,\delta_i)_{i=1,\ldots,n}$ avec $X_i = T_i \wedge T_{(r)}$ et $\delta_i = \mathbf{1}\{T_i \leq T_{(r)}\}$ (on suppose qu'il n'y a pas d'ex-aequo).

- censures progressives de type II : on enlève, à chaque décès, une proportion fixe (éventuellement différente à chaque étape) de sujets de l'échantillon, qui sont donc censurés.
- censures par intervalle : on observe le quadruplet $(C_1,C_2,\delta_1,\delta_2)$ avec $\delta_1 = \mathbf{1}\{T \leq C_1\}$, $\delta_2 = \mathbf{1}\{T \leq C_2\}$; on parle alors de censure par intervalle car on ne connaît que les bornes (aléatoires) entre lesquelles l'événement a lieu.

`Exemple.` **1.5** Pour détecter les composants défectueux d'un processus de production industriel, on effectue des contrôles selon des dates aléatoires. Lorsqu'on constate qu'un composant est à changer, on sait seulement qu'il a "décédé" entre les dates de deux contrôles successifs. C'est un exemple de censure par intervalles.

1.3.2 les troncatures

Les troncatures diffèrent des censures au sens où elles concernent l'échantillonnage lui-même. Ainsi, un variable T est tronquée par un sous-ensemble éventuellement aléatoire A de $\mathbb{R}^+$ si, au lieu de T, on observe T uniquement si $T \in A$. Plus précisément, les points de l'échantillon "tronqué" appartiennent tous à A, et suivent donc la loi de T conditionnée par l'appartenance à A.

Soit Z une variable aléatoire (éventuellement dégénérée en une masse de Dirac), on dit qu'il y a troncature droite lorsque T n'est observable que si elle est inférieure à Z. De même, on dit qu'il y a troncature gauche lorsque T n'est observable que si elle est supérieure à Z. En général, on observe le doublet (T,Z), avec $T \geq Z$ ou inversement (selon les cas).

`Exemple.` **1.6** *Des chercheurs travaillant pour la Société Protectrice des Animaux cherchent à étudier la durée de vie des chats. Ils ont a leur disposition des données de la SPA concernant les animaux confiés à cette institution. On dispose pour chaque chat de son âge Z à son arrivée à la SPA et de son âge T au décès. Les données ne sont donc pas forcément représentatives de la population totale des chats. En effet, on peut supposer que les chats sont confiés à la SPA lorsqu'ils sont soit très jeunes, soit très âgés. on dispose d'un échantillon de réalisations de T tronquées à gauche par une variable aléatoire Z.*

Le biais de sélection (appelé biased sampling ou stock sampling) est en fait un cas particulier de troncature. Dans le problème général de biased sampling, la probabilité de tirage d'un individu dans l'échantillon de nos données est fonction des valeurs de la variable d'intérêt T (cf Gill et al. (1988)). Or, lorsqu'il y a troncature gauche par Z, cette probabilité est $P(Z < t_i)$ lorsque la réalisation de T est t_i.

1.4 Effet des censures et troncatures

1.4.1 Identifiabilité

Notre objectif est d'estimer la loi d'une durée T à partir d'observations éventuellement incomplètes. Le premier problème qui se pose est celui de l'identifiabilité, c'est-à-dire : la loi de mes observations me permet-elle d'identifier la loi de T ? Autrement dit, si je connaissais parfaitement la loi suivie par mes observations, pourrais-je déterminer de manière unique la loi de T (par exemple, sa densité f ou sa survie S) ?

La réponse n'est pas toujours positive car avoir des données incomplètes constitue une perte d'information. En effet, considérons le cas simple d'une durée T de support $\mathbb{R}^+$ et une censure fixe droite en un point t_0. Il ne nous sera pas possible d'identifier la partie de la distribution de T qui se trouve à droite de t_0 ; nous ne pourrons identifier que la probabilité que T dépasse t_0, mais nous ne pourrons pas préciser le nombre de modes, la durée de vie restante au-delà de t, pour tout $t > t_0$, l'épaisseur de la queue de distribution...

Dans le cas d'une troncature droite fixe en t_0, le problème est similaire. J'ai en fait encore moins d'information car je ne peux identifier que la loi de T conditionnelle à $T \leq t_0$. De plus, $P(T > t_0)$ ne pourra pas même être évalué.

Dans le cas général, si durées et censures (ou troncatures) sont corrélées, il n'est pas possible d'identifier la loi de T à partir des observations. C'est pourquoi on fait très couramment l'hypothèse que durées et censures sont indépendantes. Précisons les choses. Dans ce but, notons $supp(T)$ le support de la loi de T, soit

$$supp(T) = \overline{\{t|f(t) \neq 0\}}, et$$

$$ssupp(T) = \sup\{t|t \in \ supp(T)\} \in \mathbb{R} \cup \{+\infty\}.$$

Alors,

Proposition 1.4 *Dans le cas d'une censure aléatoire droite C, si T et C sont indépendantes et si $ssupp(T) \leq ssupp(C)$, alors la loi de T est identifiable à partir de la loi des observations (X,δ).*

Preuve. Supposons qu'on connaisse la loi de (X,δ). Notons S et G les fonctions de survie respectives de T et C. Pour tout t, on a $P(X > t) = G(t)S(t)$. Donc on connaît le produit $S.G$. Par ailleurs, on connaît la fonction

$$\begin{aligned} P(X > t,\delta = 1) &= P(T > t,C > T) = E[\mathbf{1}\{T > t\}G(T)] \\ &= \int_t^{+\infty} G(u)f(u)du, \end{aligned}$$

donc on en déduit $G.f$, ou $G.S'$. En faisant le rapport de $S.G(t)$ sur $S'.G(t)$, on peut en déduire $S(t)$, tant que $G(t)$ est non nulle, i.e. tant que $t \leq ssup(C)$. Ainsi, on connaît entièrement la loi de T puisque $ssupp(T) \leq ssupp(C)$. Remarquons qu'on en déduit également la fonction G, donc la loi de C, sur l'ensemble $supp(C) \cap supp(T)$. □

Notons que l'hypothèse de la proposition précédente aurait pu s'écrire : pour tout t à l'intérieur du support de la loi de T, il existe t' tel que $S(t) < G(t')$.

L'hypothèse d'indépendance entre durée d'intérêt et perturbation est pratique et bien souvent nécessaire, mais elle n'est pas toujours réaliste. Dans le cas d'une censure causée par la fin du processus d'enquête, c'est naturel. Mais a contrario, lorsqu'un patient est soumis simultanément à plusieurs risques, il paraît légitime de considérer que ces derniers sont tous reliés à l'état de santé général. Un risque particulier a donc de fortes chances d'être corrélé avec les autres, qui le censurent éventuellement.

Cette correlation entre durées et censures constitue la trame des modèles à risques concurrents (ou compétitifs), qui seront abordés ultérieurement.

A titre d'illustration, soient T une durée et C une censure aléatoire droite. et soient les trois lois jointes (C,T) fournies par les tableaux suivants :

(C,T)	1	3
2	0	1/2
3/2	1/2	0

(C,T)	1	5
3	1/2	0
2	0	1/2

(C,T)	1	3
2	1/2	1/2
4	0	0

Par exemple, dans le premier tableau, la probabilité que $(C,T) = (3/2,1)$ vaut $1/2$. Notons que dans le dernier cas, C et T suivent des lois indépendantes. On constate que ces trois distributions du couple (C,T) fournissent la même loi concernant les observations (δ,X), en l'occurrence

(δ,X)	1	2
0	0	1/2
1	1/2	0

De même, supposons qu'on soit dans le cas d'une troncature de T par une variable aléatoire Z indépendante de T. Pour fixer les idées, supposons que ce soit une troncature gauche. On dispose d'un échantillon i.i.d. $(t_i,z_i)_{i=1,\dots,n}$ d'observations tirées dans la loi de (T,Z) conditionnellement à $T \geq Z$. La loi des observations a donc pour fonction de répartition

$$H^*(t,z) = P(T \leq t, Z \leq z | T \geq Z) = -\alpha^{-1} \int_0^t G(u \wedge z) S(du), \tag{1-5}$$

où S (respectivement G) représente la survie (resp. la fonction de répartition [2]) de T (resp. Z). On a posé $\alpha = P(T \geq Z)$. Peut-on retrouver S et G à partir de la loi des observations $(T_i,Z_i)_{i=1,\dots,n}$, i.e. à partir de la fonction de répartition H^* ? Pour simplifier, on se place dans le cas de lois continues. T et Z admettent donc des densités par rapport à la mesure de Lebesgue. On les note f et g respectivement.

Proposition 1.5 *Supposons que*

$$P(Z > t) \neq 0 \Longrightarrow P(T > t) \neq 0.$$

Alors la loi du couple (Z,T) est identifiable.

Preuve. En dérivant l'équation (1-5) par rapport à t, on obtient la quantité $\alpha^{-1} G(z) f(t)$ sur l'ensemble $\{(t,z) | t \geq z\}$. Soit un z_0 arbitraire, dans le support de la loi de C. Pour tout z dans ce support, il existe un t supérieur à z_0 et z tel que $f(t) \neq 0$; on connaît

$$G(z)/G(z_0) = \left(\alpha^{-1} G(z) f(t)\right) / \left(\alpha^{-1} G(z_0) f(t)\right).$$

En faisant tendre z vers la borne inférieure du support de la loi de C, on en déduit $G(z_0)$. On connaît donc G, donc f à une constante près. La normalisation $\int f = 1$ nous permet d'identifier f. □

1.4.2 Vraisemblance des observations

La plupart des estimateurs pertinents pour traiter un certain type de perturbations ne se transposent pas toujours facilement pour d'autres. En particulier, la théorie des estimateurs sur données censurées est nettement différente de celle sur données tronquées. Il en découle une multiplicité d'outils, dont beaucoup sont spécifiques aux type de données recueillies.

Raisonner sur l'échantillon des données "complètes" (non soumises aux biais d'observations, par exemple non-censurées) ne constitue pas une solution valable.

Illustrons cette dernière assertion avec la méthode du maximum de vraisemblance. Plaçons-nous dans le cas simple de la censure aléatoire droite C, indépendante de la durée d'intérêt T et spécifique à chaque individu. Pour simplifier, on suppose que T (respectivement C) a pour densité f (resp. g) et pour survie S (resp. G). La distribution de T est entièrement définie par la connaissance d'un paramètre θ de dimension finie. La contribution de l'individu i à la vraisemblance est alors

$$\begin{aligned}
\mathcal{L}_i &= P(X \in [x_i, x_i + dt], \delta = 1; \theta)^{\delta_i} \cdot P(X \in [x_i, x_i + dt], \delta = 0; \theta)^{1-\delta_i} \\
&= P(T \in [x_i, x_i + dt], C \geq T; \theta)^{\delta_i} \cdot P(C \in [x_i, x_i + dt], C < T; \theta)^{(1-\delta_i)} \\
&\approx [f(x_i;\theta) G(x_i-)]^{\delta_i} \cdot [g(x_i) S(x_i;\theta)]^{1-\delta_i}
\end{aligned}$$

Par hypothèse, le paramètre d'intérêt θ n'apparaît pas dans la loi de la censure; on peut donc se limiter aux termes relatifs à la loi de T, c'est-à-dire considérer que la vraisemblance est ici

$$\mathcal{L} = \prod_{i=1}^{n} f(x_i;\theta)^{\delta_i} S(x_i;\theta)^{1-\delta_i}. \tag{1-6}$$

2. cette notation est différente de celle utilisées pour les censures

En raisonnant sur le seul sous échantillon des données non censurées, on aurait obtenu

$$\tilde{\mathcal{L}} = \prod_{i=1}^{n} f(x_i;\theta)^{\delta_i}.$$

Maximiser $\mathcal{L}$ ou $\tilde{\mathcal{L}}$ ne fournit évidemment pas les mêmes estimateurs de θ. Le second est asymptotiquement biaisé.

`Exercice 1.2` *Montrer que, dans le cas d'une durée exponentielle (i.e. pour laquelle $S(t) = \exp(-\mu t)$) et d'une censure droite de densité g inconnue, indépendante de la durée, l'estimateur du maximum de vraisemblance associé à $\mathcal{L}$ est consistant pour μ. A l'opposé, l'e.m.v. associé à $\tilde{\mathcal{L}}$ fournit une estimation biaisée du paramètre μ.*

En présence de données censurées, la vraisemblance $\mathcal{L}$ précédente est en fait valable dans un cadre plus général que la censure aléatoire indépendante de la durée T, qui inclue notamment le cas de censures de type II (progressives ou non). Plus précisément, cette vraisemblance fournit des estimations des paramètres non biaisées dans le cas de mécanismes de censures indépendants (Kalbfleish et prentice (1980)). Ce sont les mécanismes tels que, pour tout t,

$$P(X \in [t,t+dt], \delta = 1 | \mathcal{H}(t)) = P(X \in [t,t+dt], \delta = 1 | T > t),$$

$\mathcal{H}(t)$ désignant l'histoire du processus lié à (T,C) jusqu'à la date t (observation des décès, censures et éventuellement covariables jusqu'au temps t). Cette identité signifie que le déroulement du processus à partir de la date t n'est pas perturbé par l'historique des censures et décès passés. De plus, (1-6) n'est valable que si le mécanisme de censure doit être non informatif, i.e. le paramètre θ ne doit pas intervenir dans le loi de C.

`Remarque` **1.4.** *On peut trouver des mécanismes de censures indépendants et informatifs : imaginer deux expériences identiques se déroulant en parallèle ; la censure serait alors la durée associée à un individu issu de l'autre échantillon.*

Reprenons le cas des données tronquées à gauche de manière indépendante, étudié dans le sous-section précédente. Il est possible d'écrire la vraisemblance des observations $(t_i,z_i)_{i=1,\dots,n}$ conditionnellement à la taille n de l'échantillon[3]. C'est

$$\mathcal{L}_1 = \prod_{i=1}^{n} \mathcal{L}(t_i,z_i | T \geq Z) = \frac{1}{P(T \geq Z)^n} \prod_{i=1}^{n} \mathcal{L}(t_i,z_i),$$

où $\mathcal{L}(t_i,z_i)$ représente la loi du vecteur (T,Z) (loi initiale non conditionnée). Rappelons que par construction, on a forcément $t_i \geq z_i$ pour tout i. Notons également que

$$P(T \geq Z) = E[G(T)] = \int G(u)\, f(u,\theta)\, du,$$

donc cette quantité fait intervenir la loi de la troncature. Ainsi, dans le cas d'un modèle paramétrique, on a

$$\mathcal{L}_1 = \left(\int G(u)\, f(u;\theta)\, du \right)^{-n} \prod_{i=1}^{n} f(t_i;\theta) g(z_i),$$

en notant $f(\cdot;\theta)$ et g les densités respectives de T et Z. Lorsque la loi de la troncature se réduit à une masse de Dirac en un point z_0, on n'observe les durées que lorsque $T \geq z_0$. La formule précédente devient alors

$$\mathcal{L}_1 = S(z_0;\theta)^{-n} \prod_{i=1}^{n} f(t_i;\theta).$$

3. cette quantité n est aléatoire, si on imagine qu'on sélectionne les individus i pour lesquels $T_i \geq Z_i$ dans une population de taille inconnue N

Une autre méthode consisterait à considérer la vraisemblance conditionnelle par rapport à n et aux valeurs de la variable de troncature observées. On obtient alors la vraisemblance conditionnelle

$$\mathcal{L}_2 = \prod_{i=1}^{n} \mathcal{L}(t_i | T \geq z_i).$$

Lorsque la troncature est indépendante de la censure, on peut estimer les paramètres de la loi de T sur l'une ou l'autre de ces quantités. Le seconde est plus commode à utiliser lorsqu'on n'a pas d'a priori sur la loi de la troncature. Dans le cas d'un modèle paramétrique, on a alors

$$\mathcal{L}_2 \propto \prod_{i=1}^{n} \frac{1}{S(z_i;\theta)} f(t_i;\theta).$$

Optimiser les deux vraisemblances en le paramètre θ nous fournira deux estimateurs convergents de θ. Néanmoins, $\mathcal{L}_1$ utilise "l'information complète" contenue dans les données, mai nécessite la connaissance de la loi de Z. Lorsque c'est le cas, optimiser $\mathcal{L}_1$ produira alors un estimateur plus précis qu'avec $\mathcal{L}_2$.

1.5 Hétérogénité

En général, tous les individus ne sont pas identiques. C'est pour cela qu'on les caractérise par l'intermédiaire de covariables (ou variables explicatives, ou variables exogènes).

Or, dans toute modélisation, le risque est grand "d'oublier" des covariables pertinentes pour décrire le phénomène étudié. De plus, les fichiers de données comportent rarement toutes le variables individuelles nécessaires en théorie. Or, le fait d'oublier ou de ne pas prendre en compte des variables pertinentes (ce qui arrive souvent) biaise les estimations des paramètres du modèle.

Nous allons formaliser le problème : soit une population totale $\mathcal{P}$, partitionnée en classes $\mathcal{P}_v$,$v \in V \subset \mathbb{R}^q$. Supposons que, vis-à-vis de la durées étudiée, les individus d'une même classe sont identiques, i.e. on peut définir $S(t,v)$, $\lambda(t,v)$ et $r(t,v)$ pour tout t et tout v.

En $t = 0$ les paramètres d'hétérogénité v sont distribués dans V selon la loi π. Comme les proportions relatives d'individus des différentes classes se modifient au cours du temps, la distribution dans V des paramètres v à l'instant t est notée π_t. La proportion théorique d'individus de la classe v présents dans la population à l'instant t est alors

$$\pi_t(v) = \frac{S(t,v)\pi(v)}{\int_V S(t,u)\pi(u)\,du}.$$

En effet, si la population est de taille N en $t = 0$, le nombre moyen d'individus de type v présents en t est alors $NS(t,v)\pi(v)$. Au niveau agrégé, on définit les fonctions de densité, de survie, de hasard et de durée de vie restante par

$$\tilde{f}(t) = E_\pi[f(t,v)] = \int f(t,v)\pi(v)\,dv,$$
$$\tilde{S}(t) = E_\pi[S(t,v)],$$
$$\tilde{\lambda}(t) = E_{\pi_t}[\lambda(t,v)],$$
$$\tilde{r}(t) = E_{\pi_t}[r(t,v)].$$

Notons que $\tilde{f}$ et $\tilde{S}$ sont calculée en considérant la distribution de l'hétérogénéité à la date 0, alors que pour $\tilde{\lambda}$ et $\tilde{r}$, elle est prise à la date t. On montre facilement que

Proposition 1.6 *Pour tout t, sous des conditions de régularité,*

$$\tilde{\lambda}(t) = -\frac{d}{dt}(\ln \tilde{S}(t)) = \frac{\tilde{f}(t)}{\tilde{S}(t)}, \qquad \tilde{r}(t) = \frac{1}{\tilde{S}(t)} \int_t^{+\infty} \tilde{S}.$$

En effet, si on peut dériver sous le signe somme, on a directement

$$-\frac{\tilde{S}'(t)}{\tilde{S}(t)} = \frac{1}{\tilde{S}(t)} \int f(t,v)\pi(v)\,dv = \frac{1}{\tilde{S}(t)} \int \lambda(t,v)S(t,v)\pi(v)\,dv = \int \lambda(t,v)\pi_t(v)\,dv.$$

`Exemple.` **1.7** Soit le problème de l'insertion des chômeurs (cf. exemple 1.4). On fait l'hypothèse que chaque individu a une probabilité constante au cours du temps de retrouver un emploi. La loi de la durée d'un épisode de chômage est donc exponentielle. Or tous les individus n'ont pas les mêmes chances de retrouver un emploi. On résume ce fait intuitivement évident par l'introduction d'un paramètre d'hétérogénéité individuelle v. On peut poser que la fonction de hasard de l'individu de paramètre v est alors $\lambda(t,v) = v$ pour tout t.

Soit alors π la loi de v. Pour simplifier, on supposera que π ne charge que deux réels de [0,1[, v_1 avec la probabilité α, et v_2, avec la probabilité $\beta = 1 - \alpha$.

On en déduit qu'au niveau agrégé, la densité, la survie et la fonction de hasard s'écrivent respectivement

$$\begin{aligned}
\tilde{f}(t) &= E_\pi[f(t,v)] = \alpha v_1 \exp(-v_1 t) + \beta v_2 \exp(-v_2 t),\\
\tilde{S}(t) &= E_\pi[S(t,v)] = \alpha \exp(-v_1 t) + \beta \exp(-v_2 t),\\
\tilde{\lambda}(t) &= \tilde{f}(t)/\tilde{S}(t) = \frac{\alpha v_1 + \beta v_2 \exp(-(v_2 - v_1)t)}{\alpha + \beta \exp(-(v_2 - v_1)t)}.
\end{aligned}$$

Ainsi, si la population est homogène pour le phénomène de chômage, $v_1 = v_2$ et on retrouve la constance de la fonction de hasard au niveau agrégé. Par contre, si par exemple $v_2 > v_1$, les individus de type 2 quittent plus rapidement que les autres l'état de chômage ; $\tilde{\lambda}$ est alors une fonction décroissante de t alors que pour chaque individu pris séparément, elle est constante (phénomène du mover-stayer). L'interprétation de ce phénomène est simple : la proportion des chômeurs possédant le plus de chance de retrouver un emploi ($v = v_2$) diminue dans l'échantillon au fil du temps. Au bout d'un certain temps, il ne reste pratiquement dans la population que les plus difficiles à insérer.

Sans hypothèse supplémentaires, l'analyse statistique sur données agrégées (par exemple l'estimation de $\tilde{\lambda}$) ne permet pas d'inférer la distribution des durées individuelles, ni la distribution des paramètres d'hétérogénéité. Ainsi, il peut arriver qu'au niveau agrégé, la fonction de hasard soit décroissante, alors que pour tout individu, elle soit strictement croissante (cf exercice 1.3)!

`Exercice 1.3` *Supposons que la loi de T est, pour l'individu de la classe $\mathcal{P}_v$, $v > 0$, une Weibull, i.e. $S(t,v) = \exp(-vt^\beta)\mathbf{1}\{t \geq 0\}$, $\beta > 0$. De plus, on suppose que le paramètre d'hétérogénéité suit une loi gamma, i.e. $\pi(v) = v^{\rho-1}\exp(-v)/\Gamma(\rho)\mathbf{1}\{v > 0\}$, $\rho > 0$. Calculer π_t et $\tilde{\lambda}(t)$. En déduire que $\tilde{\lambda}$ est décroissante en tout point pour toute valeur des paramètres.*

En fait, le biais d'hétérogénéité nous incite toujours à sous estimer la pente moyenne de λ. En effet,

Proposition 1.7 *pour tout t, v et π,*

$$\tilde{\lambda}'(t) = E_{\pi_t}[\lambda'(t,v)] - Var_{\pi_t}[\lambda(t,v)].$$

Cette identité n'est qu'un cas particulier de la relation plus générale

Proposition 1.8 *Si $g(\cdot,v) : \mathbb{R}^+ \times V \to \mathbb{R}$ est $C^1(\mathbb{R}^+)$ pour tout v,*

$$\frac{\partial}{\partial t}E_{\pi_t}[g(t,v)] = E_{\pi_t}[\frac{\partial}{\partial t}g(t,v)] - Cov[g(t,v);\lambda(t,v)].$$

Preuve :

$$\begin{aligned}
\frac{\partial}{\partial t}E_{\pi_t}g(t,v) &= \int \frac{\partial}{\partial t}g(t,v)\pi_t(v)\,dv + \int g(t,v)\frac{\partial}{\partial t}\pi_t(v)\,dv\\
&= E_{\pi_t}[\frac{\partial}{\partial t}g(t,v)] + \int_V g(t,v)\cdot\left\{-\frac{f(t,v)\pi(v)}{\int_V S(t,v)\pi(u)\,du} + \frac{S(t,v)\pi(v)}{[\int_V S(t,u)\pi(u)\,du]^2}\cdot\int_V f(t,s)\pi(s)\,ds\right\}dv\\
&= E_{\pi_t}[\frac{\partial}{\partial t}g(t,v)] - \int g(t,v)\lambda(t,v)\pi_t(v)\,dv + E_{\pi_t}g(t,v).E_{\pi_t}\lambda(t,v).\square
\end{aligned}$$

Dans le cas particulier de l'indépendance temporelle au niveau désagrégé (lorsque, pour tout t, $\lambda(t,v) = \lambda(v)$, constante indépendante du temps), la proposition 1.7 nous apprend que la fonction de hasard agrégée est décroissante.

2 Les estimateurs de Kaplan-Meier (1958) et Nelson-Aalen

2.1 De la survie empirique à Kaplan-Meier

Soit un échantillon i.i.d. de durées $(T_i)_{i=1,\dots,n}$. La durée T a pour fonction de survie S. L'estimateur "naturel" de S est la survie empirique

$$S_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}\{T_i > x\}.$$

Comme la fonction de répartition empirique (qui vaut $1-S_n$), cette fonction possède de nombreuses qualités en termes de convergence: convergence p.s. uniforme sur $\mathbb{R}$ (th. de Glivenko-Cantelli), à la vitesse $(\ln_2 n/n)^{1/2}$ (th. de Finkelstein), convergence en loi du processus empirique associé vers un pont brownien etc (voir Shorack et Wellner (1986), par exemple).

Or, dans le cas de données censurées à droite, la variable d'intérêt n'est plus la variable observée, mais la durée sous-jacente. Plus précisément, on notera T la durée, S sa fonction de survie, C la censure aléatoire droite de survie G, $X = T \wedge C$ est la durée observée, de survie H et $\delta = \mathbf{1}\{T \leq C\}$. On supposera que la censure C est indépendante de la durée T. Notons alors que $H = S.G$.

Désormais, estimer S par la survie empirique des données observées $(X_i)_{i=1,\dots,n}$, ou même par les données observées non censurées $(X_i)_{i=1,\dots,n|\delta_i=1}$, ne fournit qu'une estimation biaisée de S. En effet, dans ce dernier cas, la survie estimée serait

$$\hat{S}(x) = \frac{\sum_{i=1}^{n} \mathbf{1}\{X_i > x, \delta_i = 1\}}{\sum_{i=1}^{n} \mathbf{1}\{\delta_i = 1\}}.$$

Alors, $\hat{S}(x)$ tend presque sûrement vers $P(T > x, \delta = 1)/P(\delta = 1)$. Cette quantité s'écrit également $-\int_x^\infty G(t-)S(dt)/P(T \leq C) \neq -\int_x^\infty S(dt) = S(x)$.

Plaçons-nous dans le cadre de données groupées, en pratique très fréquent. On observe donc l'échantillon à dates fixes $t_0 = 0, t_1, \dots, t_{K-1}$. On posera arbitrairement $t_K = +\infty$. A chaque date t_j, on connaît $m_{j,K}$, le nombre de décès dans l'intervalle $I_j =]t_{j-1}, t_j]$, $c_{j,K}$, le nombre de censures dans I_j et $n_{j,K}$ le nombre de sujets présents dans l'échantillon à la date t_{j-1} (ni décédés, ni censurés jusque là, i.e. pour lesquels $X \geq t_{j-1}$). On notera $q_{j,K}$ la probabilité de décéder dans l'intervalle I_j, sachant qu'on est vivant en t_{j-1}.

L'estimateur sur échantillon réduit se propose d'estimer chaque $q_{j,K}$ par la quantité

$$\hat{q}_{j,K}^{(1)} = \frac{m_{j,K}}{n_{j,K} - c_{j,K}}.$$

Comme $S(t_{j_0}) = P(T > t_{j_0}) = \prod_{j=1}^{j_0} P(T > t_j | T > t_{j-1}) = \prod_{j=1}^{j_0}(1 - q_{j,K})$, on peut estimer la survie de T au point t par

$$\hat{S}^{(1)}(t) = \prod_{j|t_j \leq t} (1 - \hat{q}_{j,K}^{(1)}).$$

En fait, $\hat{q}_{j,K}^{(1)}$ estime $q_{j,K}$ "comme si" il n'y avait pas de censures dans I_j. Il est découle un biais. Ainsi $\hat{q}_{j,K}^{(1)}$ surestime q_j et $\hat{S}^{(1)}(t)$ sous-estime $S(t)$.

Remarque 2.1. *On a également proposé d'estimer les quantités $q_{j,K}$ par*

$$\hat{q}_{j,K}^{(2)} = \frac{m_{j,K}}{n_{j,K} - c_{j,K}/2}.$$

Alors, $\hat{S}^{(2)}(t) = \prod_{j|t_j \leq t}(1 - \hat{q}^{(2)}_{j,K})$ *est appelé estimateur actuariel de* $S(t)$*. On peut montrer que* $\hat{S}^{(2)}(t)$ *n'estime* $S(t)$ *de manière consistante pour tout choix de points frontières* $(t_j)_{j=1,\ldots,K}$ *que si* $S(x) = 1 - (1 + C_0.G(x))^{-1/2}$*,* $C_0 > 0$ *constante arbitraire, c'est-à-dire sous des conditions particulièrement restrictives (voir Breslow et Crowley (1974)).*

Kaplan et Meier (1958) ont repris les estimateurs sur données groupées, en introduisant l'idée de faire tendre vers 0 la largeur des intervalles I_j et faire tendre vers l'infini le nombre K d'intervalles. Soit donc

$$\hat{S}_{K,n}(t) = \prod_{j|t_j \leq t} \left(1 - \frac{m_{j,K}}{(n_{j,K} - c_{j,K})}\right).$$

Si $\sup_{1 \leq j \leq K} |t_{j,K} - t_{j-1,K}| \to 0$ quand $K \to \infty$, alors

$$\hat{S}_{K,n}(t) \underset{K \to \infty}{\longrightarrow} \prod_{i=1}^{k} (1 - m_i/n_i)^{\mathbf{1}\{X^*_{(i)} \leq t\}},$$

où $X^*_{(1)} < \cdots < X^*_{(k)}$ sont les k valeurs distinctes ordonnées parmi $(X_1, \ldots, X_n)$, m_i est le nombre de décès (non censurés) à la date $X_{(i)}$, et $n_i = \sum_{j=1}^{n} \mathbf{1}\{X_j \geq X^*_{(i)}\}$ est le nombre de sujets "à risque" à la date $X^*_{(i)}$ (i.e. des individus encore ni décédés, ni censurés juste avant $X^*_{(i)}$).

Notons que si, au point $X^*_{(j)}$, on observe qu'un individu i est censuré, alors $C_i = X^*_{(j)} < T_i$. Donc l'individu i est susceptible de décéder en $t = X^*_{(j)}$. C'est pourquoi il n'y a pas lieu de le soustraire de l'ensemble à risque à cette date.

L'estimateur ainsi obtenu est l'estimateur de Kaplan et Meier de la survie S au point t, noté $\hat{S}_{KM}(t)$. On l'appelle également l'estimateur "produit-limite" (the product-limit estimator). C'est une fonction constante par intervalles, les sauts ayant lieu en chaque point $X^*_{(j)}$. Elle est cadlag (en tout point, elle est continue à droite, et possède une limite à gauche).

On note que $\hat{S}_{KM}(t) = 0$ lorsque t est supérieur ou égal à $X^*_{(k)}$, la plus grande des durées non censurées observée, et que $X^*_{(k)} = X_{(n)}$ (on n'observe pas de censures au-delà de $X^*_{(k)}$).

Il est souvent commode d'exprimer autrement $\hat{S}_{KM}(t)$. Dans ce but, on ordonne les observations $(X_i)_{i=1\ldots,n}$ de telle manière qu'en cas d'ex aequo (simultanéité temporelle) entre durées non censurées ou entre censures, on choisisse un ordre arbitraire. S'il y a ex aequo entre durées non censurées et censures, on considère que les durées ont toujours lieu avant les censures. Cela revient à ordonner les couples $(X_i, 1 - \delta_i)_{i=1,\ldots,n}$ selon l'ordre lexicographique. Soient les statistiques d'ordre associées $X_{(1)} \leq \ldots \leq X_{(n)}$, et on pose $\delta_{(i)} = \delta_k$ si $X_{(i)} = X_k$. Alors l'estimateur de Kaplan-Meier se réécrit pour tout t

$$\begin{aligned}
\hat{S}_{KM}(t) &= \prod_{i=1}^{n} \left(1 - \frac{\delta_{(i)}}{n - i + 1}\right)^{\mathbf{1}\{X_{(i)} \leq t\}} && (2\text{-}1)\\
&= \prod_{\substack{i=1,\ldots,n \\ X_{(i)} \leq t}} \left(1 - \frac{\delta_{(i)}}{n - i + 1}\right) && (2\text{-}2)\\
&= \prod_{\substack{i=1,\ldots,n \\ X_{(i)} \leq t}} \left(\frac{n - i}{n - i + 1}\right)^{\delta_{(i)}}. && (2\text{-}3)
\end{aligned}$$

Remarque 2.2. *L'identité 2-1 provient de l'identité*

$$(1 - \frac{1}{n_i})(1 - \frac{1}{n_i - 1}) \cdots (1 - \frac{1}{n_i - (m_i - 1)}) = (1 - \frac{m_i}{n_i}).$$

`Remarque` **2.3.** *La définition de $\hat{S}_{KM}(t)$ pour les "grandes" valeurs de t varie selon les auteurs. Pour nous, si $t \geq X_{(n)}$, $\hat{S}_{KM}(t) = 0$ si $\delta_{(n)} = 1$ et $\neq 0$ si $\delta_{(n)} = 0$. Certains auteurs posent $\hat{S}_{KM}(t) = 0$ pour tout $t \geq X_{(n)}$. D'autres, enfin, choisissent de ne pas définir $\hat{S}_{KM}(t)$ lorsque $t \geq X_{(n)}$ et $\delta_{(n)} = 0$.*

`Exemple.` **2.1** Calculons l'estimateur de Kaplan-Meier sur les données de Freireich (1963): "étude des durées de rémission, exprimées en semaines, de sujets atteints de leucémie, selon qu'ils ont reçu de la 6-mercaptopurnie ou un placébo". Les 21 observations se présentent de la manière suivante, un signe + indiquant une donnée censurée à droite :

$$6\ 6\ 6\ 6^+\ 7\ 9^+\ 10\ 10^+\ 11^+\ 13\ 16\ 17^+\ 19^+\ 20^+\ 22\ 23\ 25^+\ 32^+\ 32^+\ 34^+\ 35^+.$$

Alors, la fonction de survie estimée par la méthode de Kaplan et Meier, notée $\hat{S}$, fournit une fonction en escalier, non nulle sur $\mathbb{R}^+$, et telle que

$$\begin{array}{ll}
\hat{S}(t) = 1 & si\ 0 \leq t < 6 \\
\hat{S}(t) = (1 - 3/21)\hat{S}(6-) = 0{,}857 & si\ 6 \leq t < 7 \\
\hat{S}(t) = (1 - 1/17)\hat{S}(7-) = 0{,}807 & si\ 7 \leq t < 10 \\
\hat{S}(t) = (1 - 1/15)\hat{S}(10-) = 0{,}753 & si\ 10 \leq t < 13 \\
\hat{S}(t) = (1 - 1/12)\hat{S}(13-) = 0{,}690 & si\ 13 \leq t < 16 \\
\hat{S}(t) = (1 - 1/11)\hat{S}(16-) = 0{,}627 & si\ 16 \leq t < 22 \\
\hat{S}(t) = (1 - 1/7)\hat{S}(22-) = 0{,}538 & si\ 22 \leq t < 23 \\
\hat{S}(t) = (1 - 1/6)\hat{S}(23-) = 0{,}448 & si\ 23 \leq t
\end{array}$$

Son graphe est tracé sur la figure 2.

2.2 Propriétés de $\hat{S}_{KM}$

2.2.1 Cohérence

Tout d'abord, on remarque qu'en l'absence de censures, $\hat{S}_{KM}(t)$ se réduit à la survie empirique. Plus généralement,

Définition 2.1 *Un estimateur $\hat{S}$ de la fonction de survie S est dit cohérent si, pour tout t, il vérifie l'identité*

$$\hat{S}(t) = \frac{1}{n}\left[\sum_{i=1}^{n} \mathbf{1}\{X_i > t\} + \sum_{i=1}^{n} \mathbf{1}\{X_i \leq t, \delta_i = 0\} \cdot \frac{\hat{S}(t)}{\hat{S}(X_i)}\right].$$

La cohérence s'interprète de la façon suivante : la probabilité de décéder au-delà de la date t est la somme de la probabilité de n'être ni décédé, ni censuré à cette date et de la probabilité d'avoir été censuré avant la date t, et d'être toujours en vie en t, d'où le terme $\hat{S}(t)/\hat{S}(X_i)$ [4].

On montre non seulement que $\hat{S}_{KM}$ est cohérent, mais qu'en plus (voir Droesbeke et al. (1989)),

Proposition 2.1 *$\hat{S}_{KM}$ est l'unique estimateur cohérent de la survie de T.*

En particulier, lorsqu'il n'y a pas de censures, $\hat{S}_{KM}$ se réduit à l'estimateur nonparamétrique usuel de S : la fonction de répartition empirique des X_i (ou des T_i, ici).

4. ce terme est d'autant plus important que X_i est grand. En effet, une censure qui arrive "tard" est plus informative qu'une censure précoce.

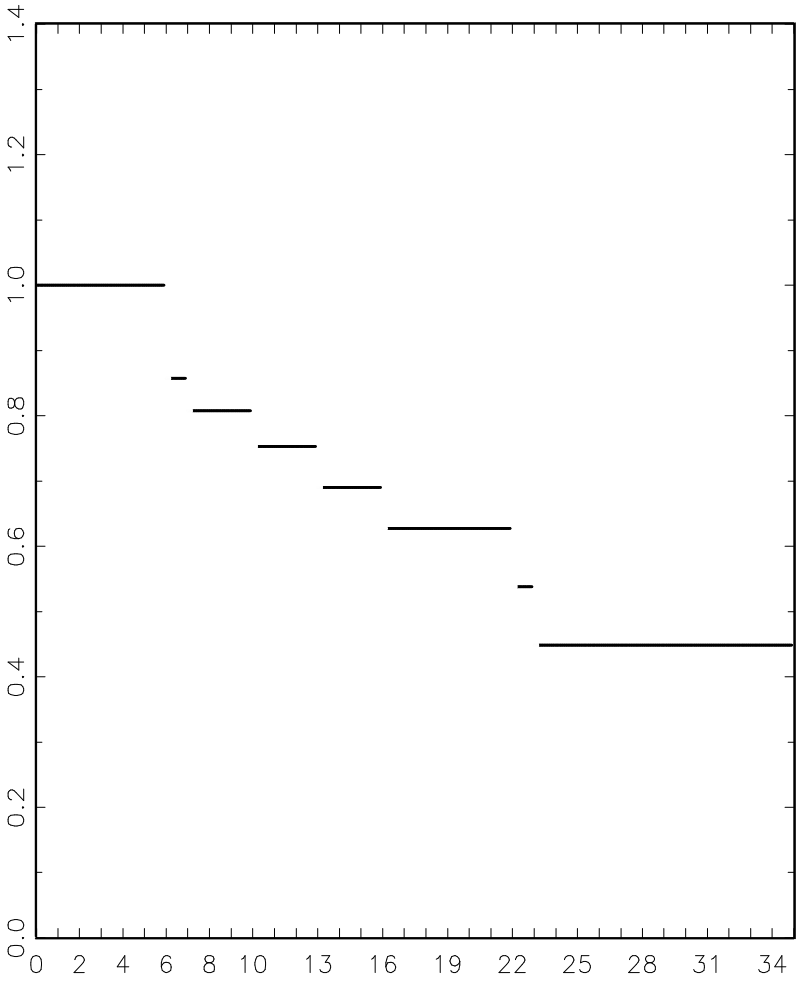

FIG. 2 – *Estimateur de Kaplan-Meier*

2.2.2 Lois fortes

Dans le cas d'un échantillon i.i.d. de variables aléatoires suivant une loi F, la fonction de répartition empirique $F_n(t) = n^{-1}\sum_{i=1}^n \mathbf{1}\{X_i \leq t\}$ (égale à 1 moins la survie empirique) est l'estimateur usuel de F. De nombreuses quantités à estimer se présentent sous la forme $E_F[\phi]$ ou $\int \phi\, dF$. Plus généralement, les quantités d'intérêt sont des fonctionnelles de la fonction de répartition empirique. Dans ce cas, une démarche usuelle consiste à remplacer F par F_n dans cette dernière formule pour obtenir un estimateur "naturel".

En analyse des données de survie, $\hat{S}_{KM}$ joue pour les données incomplètes le même rôle que la fonction de répartition empirique pour les données classiques. C'est ce qui apparaît dans les résultats de ce paragraphe.

Rappelons qu'on se place toujours dans le cadre de la censure aléatoire droite, et que H, la survie des observations X, vérifie $H = S \cdot G$ (G survie de la censure droite C). Soit A l'ensemble des atomes de X, i.e. l'ensemble des points de discontinuités de H. Soit $\tau_H = \inf\{x \geq 0 | H(x) = 0\} \in \bar{\mathbb{R}} = \mathbb{R} \cup \{+\infty\}$ la borne supérieure du support de H, et on notera $\Delta S(a) = S(a) - S(a-)$ le saut éventuel de S en a. Alors,

Théorème 2.1 (Stute et Wang (1995)) *Si S et G n'ont pas de discontinuités en commun, pour toute fonction mesurable ϕ telle que $\int |\phi| dS < \infty$, on a presque sûrement et dans $L^1(\mathbb{R})$*

$$\lim_{n\to\infty} \int \phi\, d\hat{S}_{KM} = \int_{\{x \notin A, x < \tau_H\}} \phi(x)\, S(dx) + \sum_{a \in A} \phi(a)\Delta S(a)$$

$$= \int_{\{x < \tau_H\}} \phi(x)\, S(dx) + \mathbf{1}\{\tau_H \in A\}\phi(\tau_H)\Delta S(\tau_H),$$

et le dernier terme est nul si $\tau_H = +\infty$.

Notons que lorsqu'il existe des discontinuités communes à S et G, on ne peut pas estimer ces distributions, sans hypothèses supplémentaires.

`Remarque` **2.4.** *S'il n'y a pas de censure (ou plus exactement quand la censure est portée "à l'infini", telle une masse de Dirac en un point t_0 qui "tend vers $+\infty$"), on retrouve le résultat classique*

$$\int \phi\, dS_n \underset{n\to\infty}{\longrightarrow} \int \phi\, dS,$$

presque sûrement et dans L^1. Mais il est remarquable que cette formule reste "quasiment vraie" sur données censurées ; en particulier, si S est continue en τ_H,

$$\int \phi\, \hat{S}_{KM}(dt) \underset{n\to\infty}{\longrightarrow} \int_{\{x < \tau_H\}} \phi\, S(dt) = \int_{x \leq \tau_H} \phi\, S(dt).$$

Précisons la notation $\int \phi\, d\hat{S}_{KM}$. Avec les notations précédentes,

$$\int \phi\, d\hat{S}_{KM} = \sum_{i=1}^k \phi(X^*_{(i)})\Delta\hat{S}_{KM}(X^*_{(i)}),$$

$$\Delta\hat{S}_{KM}(X^*_{(i)}) = \hat{S}_{KM}(X^*_{(i)}) - \hat{S}_{KM}(X^*_{(i)}-) = -\frac{m_i}{n_i}\hat{S}_{KM}(X^*_{(i)}-).$$

On en déduit un estimateur de la durée de vie moyenne lorsque les données sont censurées.

Corollaire 2.1 *Si S et G n'ont pas de discontinuités communes, et si $\int |x| S(dx) < \infty$, on a alors presque sûrement et dans L^1*

$$E_{\hat{S}_{KM}}[T - t | T > t] = \hat{S}^{-1}_{KM}(t) \int_t^\infty (u - t)\, \hat{S}_{KM}(du)$$

$$\underset{n\to\infty}{\longrightarrow} \begin{cases} E_S[\mathbf{1}\{T < \tau_H\}(T-t) | T > t] & \text{si } \Delta S(\tau_H) < 0 \text{ (mais alors } G(\tau_H-) = 0) \\ E_S[\mathbf{1}\{T \leq \tau_H\}(T-t) | T > t] & \text{sinon} \end{cases}$$

On a également l'équivalent du théorème de Glivenko-Cantelli lorsque les données sont censurées.

Théorème 2.2 *Si S et G n'ont pas de discontinuités en commun, on a presque sûrement*

$$\sup_{t\leq\tau_H} |\hat{S}_{KM}(t) - \tilde{S}(t)| \underset{n\to\infty}{\longrightarrow} 0.$$

où $\tilde{S}(t) = S(t)$ si $t < \tau_H$, et $\tilde{S}(t) = S(\tau_H-) + \mathbf{1}\{\tau_H \in A\}\Delta S(\tau_H)$ si $t \geq \tau_H$.

Remarquons que ce résultat nous dit que le théorème 2.1 est vrai uniformément sur la classe de fonctions $\mathcal{F} = \{\phi_t = \mathbf{1}\{]-\infty,t]\}, t \in \mathbb{R}\}$.

En pratique, il est prudent de chercher à estimer la survie sur des intervalles $[0,\tau]$, $\tau < \tau_H$, pour éviter les problèmes de biais aux bornes, et travailler ainsi avec une limite "classique" proche du cas standard i.i.d. Les limites obtenues dans les théorèmes 2.1 et 2.2 s'expriment alors très simplement. Lorsque τ_H est inconnue, un principe de prudence consiste à ne pas choisir $\tau_H >> X_{(n)}$ (car on sait que $\tau_H \geq X_{(n)}$)).

De plus, on a le résultat suivant plus précis que le précédent sur l'intervalle $[0,\tau]$, $\tau < \tau_H$.

Théorème 2.3 (Lo, Mack et Wang (1989)) *Si S est continue sur $[0,\tau]$, $\tau < \tau_H$, alors il existe une constante C_0 telle que presque sûrement*

$$\left(\frac{n}{\ln(\ln n)}\right)^{1/2} \sup_{x\in[0,\tau]} |\hat{S}_{KM}(t) - S(t)| < C_0.$$

2.2.3 Convergence en loi

Nous sommes toujours dans le cadre de la censure à droite. Soit $\phi : \mathbb{R} \longrightarrow \mathbb{R}$ une fonction mesurable telle que

$$\int \phi^2 \, dF < +\infty.$$

On peut sous des hypothèses assez faibles, énoncer un théorème central limite pour toute quantité $\int \phi \, d\hat{S}_{KM}$, appelée intégrale de Kaplan-Meier.

Théorème 2.4 (Akritas (2000)) *Si $\max(X_1,\ldots,X_n) < \tau_H$ ou $\phi(\tau_H) = 0$, et si*

$$\int_{-\infty}^{\tau_H} \frac{\phi^2(s)}{1-G(s-)} dF(s) < \infty,$$

alors

$$n^{1/2} \int_{-\infty}^{\tau_H} \phi(s) \, d\left(\hat{S}_{KM} - S\right)(s) \underset{n\to\infty}{\overset{\mathcal{L}}{\longrightarrow}} \mathcal{N}(0,\sigma^2),$$

avec

$$\sigma^2 = \int_{-\infty}^{\tau_H} \frac{S(s)}{1-H(s-)} \left[\phi(s) - \bar{\phi}(s)\right]^2 dF(s),$$

$$\bar{\phi}(s) = S^{-1}(s) \int_s^{\tau_H} \phi(u) \, dF(u).$$

Ce théorème a été prouvé en utilisant des techniques de martingales, et améliore celui établi par Stute (1995) par des techniques plus classiques de processus empiriques et de représentations en sommes de termes i.i.d.

En termes de convergence faible du processus associé à $\hat{S}_{KM}$ (cf. 8.2), on a

Théorème 2.5 (Gill (1980)) *Dans l'espace $D([0,\tau],\|\cdot\|_\infty,\mathcal{P})$ des fonctions continues à droite possédant des limites à gauche en tout point de $[0,\tau]$, muni de la norme infinie et de sa tribu de projection, si $H(\tau-) > 0$ et si S est continue sur $[0,\tau]$, alors*

$$n^{1/2}(\hat{S}_{KM} - S) \Longrightarrow Z,$$

où Z est un processus gaussien centré, de fonction de covariance

$$Cov(Z(t_1),Z(t_2)) = -S(t_1)S(t_2)\int_0^{t_1\wedge t_2}\frac{S(du)}{S^2(u)G(u)}.$$

Cela signifie que pour toute fonction ϕ de $(D([0,\tau],\|.\|_\infty,\mathcal{P})$ vers $\mathbb{R}$, mesurable continue et bornée, on a lorsque $n\to+\infty$,

$$E[\phi\left(n^{1/2}(\hat{S}_{KM}-S)\right)]\longrightarrow E[\phi(Z)].$$

En particulier (voir annexe 8.2), on obtient la normalité asymptotique de $\hat{S}_{KM}(t_0)$ en tout point $t_0\in[0,\tau]$.

Théorème 2.6 *En tout point t_0 de continuité de S, $t_0\in[0,\tau]$ et $S(\tau-)>0$,*

$$n^{1/2}(\hat{S}_{KM}(t_0)-S(t_0))\xrightarrow[n\to\infty]{\mathcal{L}}\mathcal{N}(0,V^2(t_0)),$$

avec

$$V^2(t_0) = -S^2(t_0)\int_0^{t_0}\frac{S(du)}{S^2(u)G(u)}. \tag{2-4}$$

On notera qu'on peut également écrire

$$V^2(t_0) = -S^2(t_0)\int_0^{t_0}\frac{H^{(1)}(du)}{H(u)G(u-)S(u)}. \tag{2-5}$$

Rappelons qu'on avait défini $H^{(1)}(u)=P(X>u,\delta=1)$. L'écriture 2-5 provient du fait que $H^{(1)}(du)=G(u-)S(du)$. En effet, $H^{(1)}(u)=E[\mathbf{1}\{T\le C,X>u\}]=E[G(T-)\mathbf{1}\{T>u\}]=-\int_u^{+\infty}G(t-)S(dt)$, d'où $H^{(1)}(du)=G(u-)S(du)$.

L'estimateur usuel de la variance est alors obtenu en remplaçant les quantités inconnues dans la formule 2-5 par leurs estimateurs "naturels" : remplacer S par $\hat{S}_{KM}$, $H(u)$ par $H_n(u)=n^{-1}\sum_{i=1}^n\mathbf{1}\{X_i>u\}$ et $H^{(1)}(u)$ par $H_n^{(1)}(u)=n^{-1}\sum_{i=1}^n\mathbf{1}\{X_i>u,\delta_i=1\}$. On obtient l'estimateur dit "de Greenwood"

$$\hat{V}^2(t) = -\hat{S}^2_{KM}(t)\int_0^t\frac{H_n^{(1)}(du)}{H_n(u)H_n(u-)}.$$

On peut montrer que $\hat{V}^2(t)$ converge presque sûrement, lorsque $n\to\infty$, vers la variance asymptotique de $\hat{S}_{KM}(t)$.

On peut réécrire l'estimateur de Greenwood de façon plus explicite. Notons $X^*_{(1)}<\cdots<X^*_{(k)}$ les instants de décès (non censurés) ordonnés. Comme on n'exclut pas les ex-aequo éventuels, on note m_i le nombre de décès à la date $X^*_{(i)}$, n_i le nombre d'individus à risques en $X^*_{(i)}$ (ni décédés, ni censurés à cette date, soit encore $n_i=\sum_{j=1}^n\mathbf{1}\{X_j\ge X^*_{(i)}\}$). On remarque que

$$-H_n^{(1)}(du) = \begin{cases} 0 & \text{si } u\notin\{X^*_{(i)},i=1,\ldots,k\}\\ m_i/n_i & \text{si } u=X^*_{(i)}\end{cases}$$

Alors, on estime la variance de $\hat{S}_{KM}(t)$ par $n^{-1}\hat{V}^2(t)$, ce qui s'écrit

$$\widehat{Var}_{KM}(t) = n^{-1}\hat{V}^2(t) = \hat{S}^2_{KM}(t)\sum_{i|X^*_{(i)}\le t}\frac{m_i}{(n_i-m_i)n_i}.$$

2.3 $\hat{S}_{KM}$ estimateur non paramétrique du maximum de vraisemblance

On peut voir la fonction de survie S comme un paramètre dans l'espace des fonctions de répartition (de dimension infinie), et $\hat{S}_{KM}$ comme un estimateur de S. Cet estimateur possède la propriété remarquable d'atteindre le maximum de la vraisemblance $\mathcal{L}$ sur cet espace fonctionnel. Nous allons en donner une démonstration "heuristique".

Dans ce but, notons comme d'habitude $X^*_{(0)} = 0 < X^*_{(1)} < \cdots < X^*_{(k)}$ les durées non censurées observées, m_j le nombre de décès en $X^*_{(j)}$ et c_j le nombre de censures ayant eu lieu dans l'intervalle $[X^*_{(j)}, X^*_{(j+1)}[$. Ces instants de censures seront notés $X_{j,1}, \ldots, X_{j,c_j}$.

On suppose qu'existe une incertitude sur la mesure des durées ; autrement dit, on ne connaît les instants de réalisations des décès $(X^*_{(j)})_{j=1,\ldots,k}$ qu'avec une incertitude Δ_j.

La vraisemblance "approchée" s'écrit alors, avec des notations évidentes,

$$\mathcal{L}_{app}(S) = \prod_{j=0}^{k} \left\{ \left[S(X^*_{(j)} - \Delta_j) - S(X^*_{(j)} + \Delta_j) \right]^{m_j} \cdot \prod_{l_j=1}^{c_j} S(X_{j,l_j}) \right\}.$$

On cherche donc une fonction S qui maximise $\mathcal{L}_{app}$ sur l'espace des fonctions de survie (i.e. l'espace des fonctions positives décroissantes, continues à droites et possédant en tout point une limite à gauche, égales à 1 pour tout réel négatif, et tendant vers 0 en $+\infty$).

Si une censure X_{j,l_j} est égale à une durée complète $X^*_{(j)}$, on considère que $X_{j,l_j} \geq X^*_{(j)} + \Delta_j$ en choisissant Δ_j assez petit ; en effet, la durée associée à la donnée censurée précédente est forcément strictement supérieure à $X^*_{(j)}$. Par hypothèse donc, $X_{j,l_j} \in]X^*_{(j)} + \Delta_j, X^*_{(j+1)} - \Delta_{j+1}[$, pour tout j et $l_j = 1, \ldots, c_j$. Autrement dit, on choisit les Δ_j suffisamment petits de telle manière qu'il n'y ait pas de durées censurées dans les intervalles $[X^*_{(j)} - \Delta_j, X^*_{(j)} + \Delta_j]$.

Comme S est décroissante, poser $S(X_{j,l_j}) = S(X^*_{(j)} + \Delta_j)$ donne une valeur maximale à $\mathcal{L}_{app}$, les autres valeurs de S étant supposées connues.

Par ailleurs, S peut être arbitraire entre $X^*_{(j)} - \Delta_j$ et $X^*_{(j)} + \Delta_j$ sans changer la valeur de L_{app}. Par exemple, on peut la choisir affine.

Montrons que si S maximise $\mathcal{L}_{app}$, elle doit être constante et égale à $S(X^*_{(j)} + \Delta_j)$ sur tout l'intervalle $]X^*_{(j)} + \Delta_j, X^*_{(j+1)} - \Delta_j[$. Dans ce but, on peut toujours écrire pour tout $j = 1, \ldots, k$,

$$S(X^*_{(j)} - \Delta_j) = S(X^*_{(j-1)} + \Delta_{j-1}).(1 - \lambda_j),$$

$$S(X^*_{(j)} + \Delta_j) = S(X^*_{(j)} - \Delta_j).(1 - \mu_j),$$

où les λ_j et μ_j sont des constantes comprises entre 0 et 1. On a posé arbitrairement $S(X^*_{(0)} + \Delta_0) = 1$. Un rapide calcul donne alors

$$S(X^*_{(j)} - \Delta_j) = \prod_{p=1}^{j-1} (1 - \lambda_p)(1 - \mu_p) \cdot (1 - \lambda_j) = r_j,$$

et

$$\begin{aligned} \mathcal{L}_{app} &= \prod_{j=1}^{k} \left\{ (r_j \mu_j)^{m_j} \cdot \prod_{l_j=1}^{c_j} (r_j(1 - \mu_j)) \right\} \\ &= \prod_{j=1}^{k} r_j^{m_j + c_j} \mu_j^{m_j} (1 - \mu_j)^{c_j}. \end{aligned}$$

On remarque que $r_j \leq \prod_{p=1}^{j-1}(1 - \mu_p)$, l'égalité n'ayant lieu que si $\lambda_1 = \cdots = \lambda_j$. La maximisation en les paramètres λ_j (les μ_j fixés) impose alors $\lambda_j = 0$, pour tout j. Ainsi, $S(X^*_{(j)} - \Delta_j) = S(X^*_{(j-1)} + \Delta_{j-1})$ pour tout j.

Pour définir entièrement S, il reste donc à déterminer les valeurs de S aux point $X^*_{(j)} + \Delta_j$, soit à maximiser

$$\begin{aligned}
\mathcal{L}_{app} &= \prod_{j=1}^{k} \left(\prod_{p=1}^{j-1} (1-\mu_p)^{m_j+c_j} \right) \cdot \mu_j^{m_j} (1-\mu_j)^{c_j} \\
&= \prod_{j=1}^{k} \left(\prod_{p=1}^{j} (1-\mu_p)^{m_j+c_j} \right) \cdot \frac{\mu_j^{m_j}}{(1-\mu_j)^{m_j}} \\
&= \left\{ \prod_{j=1}^{k} \mu_j^{m_j} (1-\mu_j)^{-m_j} \right\} \cdot \left\{ \prod_{p=1}^{k} \prod_{j=p}^{k} (1-\mu_p)^{m_j+c_j} \right\} \\
&= \prod_{j=1}^{k} \mu_j^{m_j} (1-\mu_j)^{n_j-m_j},
\end{aligned}$$

n_j étant le nombre de sujets à risque en $X^*_{(j)}$. On obtient alors que le maximum de $\mathcal{L}_{app}$ est obtenu avec $\mu_j = m_j/n_j$ pour tout j, ce qui donne

$$S^*(X_{(j)} + \Delta_j) = \prod_{p=1}^{j} (1 - \frac{m_j}{n_j}).$$

Comme ce résultat est indépendant du choix des constantes Δ_i, on peut les faire tendre vers 0, ce qui nous redonne l'estimateur de Kaplan-Meier.

2.4 L'estimation du hasard intégré : Nelson-Aalen

A première vue, estimer la fonction de survie S ou la fonction de hasard intégrée semble le même problème, du fait de la relation[5] $S(t) = \exp(-\Lambda(t))$. Ainsi, on peut estimer $\Lambda(t)$ par

$$\hat{\Lambda}_n^{(1)}(t) = -\ln \hat{S}_{KM}(t) = \sum_{i=1}^{n} \mathbf{1}\{X_{(i)} \leq t\} \ln(1 - \frac{\delta_{(i)}}{n-i+1}), \tag{2-6}$$

si $\hat{S}_{KM}(t) \neq 0$.

En fait, la plupart des auteurs préfèrent estimer directement Λ, pour fournir un estimateur à la fois plus simple et possédant de meilleures propriétés de convergence.

Lorsque T admet une densité f par rapport à la mesure de Lebesgue, on a défini la fonction de hasard intégrée par $\Lambda(t) = \int_0^t f/S$.

On peut en fait définir la fonction de hasard intégrée même si la distribution de T n'admet pas de dérivée en tout point de $\mathbb{R}^+$, par la formule

$$\Lambda(t) = -\int_0^t \frac{S(du)}{S(u-)}.$$

Plaçons nous alors dans le cadre de données censurées aléatoirement à droite. On remarque qu'on peut écrire

$$\Lambda(t) = -\int_0^t \frac{H^{(1)}(du)}{H(u-)},$$

où $H(u) = P(X > u)$ et $H^{(1)}(u) = P(X > u, \delta = 1\}$. On a utilisé le fait que $H^{(1)}(du) = G(u-)S(du)$, et que $H(u-) = G(u-)S(u-)$. Notons que s'il existe f (et donc λ) en tout point réel positif, on retrouve que le hasard intégré est simplement la primitive de λ.

5. cette relation n'est valable qu'en dimension 1 (pour les dimensions supérieures, voir infra).

En remplaçant les deux fonctions précédentes $H^{(1)}$ et H par leurs équivalents empiriques, on obtient alors, pour tout t tel que $P(T > t) > 0$, un estimateur naturel de la fonction de hasard intégrée, appelé estimateur de Nelson-Aalen :

$$\hat{\Lambda}_n^{(2)}(t) = -\int_0^t \frac{H_n^{(1)}(du)}{H_n(u-)},$$

avec $H_n(u) = n^{-1}\sum_{i=1}^n \mathbf{1}\{X_i > u\}$ et $H_n^{(1)}(u) = n^{-1}\sum_{i=1}^n \mathbf{1}\{X_i > u, \delta_i = 1\}$. Il s'écrit également

$$\hat{\Lambda}_n^{(2)}(t) = \sum_{i=1}^n \frac{\delta_i \mathbf{1}\{X_i \leq t\}}{\sum_{j=1}^n \mathbf{1}\{X_j \geq X_i\}}. \tag{2-7}$$

Remarquons que pour n grand, les deux estimateurs (2-6) et (2-7) sont équivalents.

Comme on n'utilisera exclusivement que $\hat{\Lambda}_n^{(2)}$, on le note plus simplement $\hat{\Lambda}_n$.

Pour tout t tel que $P(T > t) > 0$, cet estimateur est fortement consistant, c'est-à-dire converge presque sûrement vers $\Lambda(t)$. On peut en avoir une intuition facilement car

$$\hat{\Lambda}_n(t) = \sum_{i=1}^n \frac{\delta_i \mathbf{1}\{X_i \leq t\}}{nH(X_i-)} \left[1 + \frac{(H - H_n)(X_i-)}{H_n(X_i-)}\right].$$

Comme $\sup_{t\in\mathbb{R}} |(H_n - H)(t)|$ tend vers 0 presque sûrement, et que

$$E[\frac{\delta \mathbf{1}\{T \leq t\}}{H(X-)}] = E[G(T)\mathbf{1}\{T \leq t\}H^{-1}(T-)] = E[\mathbf{1}\{T \leq t\}S^{-1}(T-)] = \Lambda(t),$$

on en déduit que la consistance de $\hat{\Lambda}_n$ découle de la loi des grands nombres.

En fait, on a une propriété beaucoup plus forte. Soit tout d'abord un réel positif τ tel que $P(X > \tau) > 0$. Alors

Proposition 2.2 *Il existe une constante C^* telle que, avec une probabilité 1, on ait*

$$\left(\frac{n}{\ln(\ln n)}\right)^{1/2} \sup_{t\in[0,\tau]} |(\hat{\Lambda}_n - \Lambda)(t)|$$

est borné en probabilité.

De plus, on a également la convergence faible du processus $n^{1/2}(\hat{\Lambda}_n - \Lambda)$ vers un processus gaussien centré dans les espaces fonctionnels $D([0,\tau], \|\cdot\|_\infty, \mathcal{P})$ (voir Gill et al. (1993)) ou $D([0,\tau], d_0, \mathcal{D})$ (voir Fermanian (1997)) : cf annexe 8.2.

2.5 L'estimation sur données tronquées

Les estimateurs de Kaplan-Meier et de Nelson-Aalen présentés précédemment sont valables dans le cas de censures aléatoires droites, cas le plus courant en pratique. On peut en trouver des "équivalents" lorsqu'on est en présence de données tronquées aléatoirement à gauche (troncature la plus fréquente) : on observe alors T uniquement si $T \geq Z$, Z étant une variable aléatoire, qu'on suppose généralement indépendante de T.

Dans le cas le plus classique, on considère une suite i.i.d. de couples de variables aléatoires positives $(T_i, Z_i)_{i=1,\ldots,N}$, telles que T et Z sont indépendantes. Du fait de la troncature gauche, on n'observe que les couples pour lesquels $T_i \geq Z_i$. Ainsi, la taille de l'échantillon $n \leq N$ est aléatoire et la valeur N inconnue. En fait n/N tend presque sûrement vers $\alpha = P(Z \leq T)$ lorsque $N \to +\infty$, d'après la loi forte des grands nombres.

Conditionnellement à la valeur de n, les observations du "sous-échantillon" $(T_i,Z_i)_{i=1}^n$ sont i.i.d., et leur fonction de répartition jointe s'écrit

$$H^*(t,z) = P(T \leq t, Z \leq z | T \geq Z) = -\alpha^{-1} \int_0^t G(u \wedge z) S(du),$$

où S (respectivement G) représente la survie (resp. la fonction de répartition) de T (resp. Z). Le problème consiste à reconstituer S, voire G (même si la troncature offre en soi moins d'intérêt), à partir des observations $(T_i,Z_i)_{i=1,\ldots,n}$, dont on peut approcher la fonction de répartition H^*.

On notera S_n la survie empirique de $T_1,\ldots,T_n$ et on pose $C_n(u) = n^{-1} \sum_{i=1}^n \mathbf{1}\{Z_i \leq u \leq T_i\}$. Alors les équivalents des estimateurs de Kaplan-Meier et de Nelson-Aalen s'écrivent ici

$$\hat{S}_n(t) = \prod_{i|T_i \leq t} \left[\frac{nC_n(T_i) - 1}{nC_n(T_i)} \right],$$

$$\hat{\Lambda}_n(t) = -\int_0^t \frac{S_n(du)}{C_n(u)} = n^{-1} \sum_{i|T_i \leq t} C_n^{-1}(T_i).$$

On peut montrer que ces estimateurs possèdent la plupart des propriétés de convergence observées dans le cas classique censuré : convergence presque sûre de $\hat{S}_n(t)$ vers $S(t)$ et de $\hat{\Lambda}_n(t)$ vers $\Lambda(t)$ pour tout t positif à l'intérieur de $[a_T,b_T]$, compact inclu dans l'intérieur du support de S, convergence faible des processus $n^{1/2}(\hat{S}_n - S)$ et $n^{1/2}(\Lambda_n - \Lambda)$ vers des processus gaussiens centrés, dans l'espace $D([a_T,b_T],\|\cdot\|_\infty)$ etc. Pour plus de détails, voir Stute (1993), Chao et Lo (1988), Woodroofe (1985) entre autres.

`Remarque` **2.5.** *Dans le cas de données soumises simultanément à des censures droites et des troncatures gauches, la survie estimée prend toujours la forme d'un estimateur de Kaplan-Meier, mais "modifié" : voir Tsai et al. (1987), Lai et Ying (1991).*

3 Les types de modèles

On retrouve dans le cadre des modèles de durées les grandes distinction entre inférence paramétrique, semiparamétrique et nonparamétrique. On notera comme d'habitude T la durée; on supposera pour simplifier qu'elle possède en tout point $t \geq 0$ une densité $f(t)$, une survie $S(t)$, et, lorsque $S(t) > 0$, une fonction de hasard $\lambda(t)$. Il est commode d'introduire la nouvelle variable aléatoire $Y = \ln T$ car raisonner avec Y plutôt qu'avec T permet parfois de retrouver l'estimation des modèles linéaires.

3.1 Les modèles paramétriques

On suppose que la loi des durées appartient à une famille paramétrique $\{P_\theta\}_{\theta\in\Theta}$ donnée. Si nous disposions d'un échantillon $t = (t_i)_{i=1,\dots,n}$ i.i.d. tiré dans la loi de la durée T et soumis à aucun biais d'observation (en particulier les observations ne sont ni censurées, ni tronquées), on obtiendrait alors la fonction de log-vraisemblance $\ln L_n^*(t;\theta) = \sum_{i=1}^n \ln f_\theta(t_i)$. Sous des conditions usuelles de régularité (cf Gouriéroux et Monfort (1989) par exemple), la méthode classique du maximum de vraisemblance fournirait un estimateur convergent du "vrai" paramètre θ_0, i.e.

$$\hat{\theta}_n \in \arg\max_{\theta\in\Theta} L_n^*(\theta) \underset{n\to\infty}{\longrightarrow} \theta_0$$

presque sûrement ou en probabilité, selon les hypothèses vérifiées par le modèle. De plus, $\hat{\theta}_n$ serait asymptotiquement normal et asymptotiquement efficace, sa variance asymptotique $I^*(\theta_0)$ étant la borne de Cramer-Rao.

Or, comme on l'a vu précédemment, on dispose rarement d'un tel échantillon pour les modèles de durées. On peut calculer uniquement une log-vraisemblance "observée" $\ln L_n(x;\theta) = \sum_{i=1}^n l(x_i,\theta)$, c'est-à-dire obtenue à partir des observations $x = (x_i)_{i=1,\dots,n}$. Elle est généralement bien plus compliquée que la vraisemblance précédente $L_n^*(t;\theta)$ (dite "latente").

Toutefois, on peut relier les scores des modèles latents et observés car

Proposition 3.1 *le score observé est la meilleure approximation du score latent fondée sur les observations, c'est-à-dire*

$$\frac{\partial}{\partial\theta}\ln L_n(x,\theta) = E_\theta[\frac{\partial}{\partial\theta}\ln L_n^*(t;\theta)|x]. \tag{3-1}$$

Preuve

Soit $X = \Psi(T)$ l'application qui, à une donnée complète T, associe la donnée incomplète (observée) X. On a, pour tout θ et avec des notations évidentes,

$$\ln L_1^*(t;\theta) = \ln L_1(x;\theta) + \ln L_1^*(t|x;\theta).$$

En dérivant cette égalité par rapport à θ et en prenant l'espérance par rapport à la loi conditionnelle $P_\theta^{T|X=x}$, on obtient

$$E_\theta[\frac{\partial}{\partial\theta}\ln L_1^*(t;\theta)|x] = \frac{\partial}{\partial\theta}\ln L_1(x;\theta) + E_\theta[\frac{\partial}{\partial\theta}\ln L_1^*(t|x;\theta)|x].$$

Or

$$E_\theta[\frac{\partial}{\partial\theta}\ln L_1^*(t|x;\theta)|x] = \int \frac{\partial}{\partial\theta}L_1^*(t|x;\theta)\,dt = 0,$$

si on admet qu'on peut intervertir les signes d'intégration et de dérivation[6]. En sommant sur tous les points de l'échantillon, on obtient la relation demandée entre les vraisemblances.□

`Remarque 3.1.` *Cette relation est encore valable si on raisonne sur des vraisemblances conditionnées par un processus de covariables dont la loi est indépendante de θ.*

6. C'est notamment possible en un point θ_0 si la dérivée en θ de la fonction à intégrée est dominée par une fonction intégrable indépendante de tout θ dans un voisinage de θ_0.

Pour estimer θ, peut-on maximiser $L_n(x;\theta)$ en substitut de $L_n^*(t;\theta)$? Autrement dit, peut-on utiliser le score observé à la place du score latent? En général, oui, ce qui nous permet bien souvent d'utiliser la théorie du maximum de vraisemblance "classique" alors que les données sont incomplètes. Il faut pour cela que la vraisemblance des observations vérifie les conditions de régularité usuelles et surtout, que la vraie valeur θ_0 du paramètre soit identifiable à partir de la loi de X[7].

En pratique, lorsque toutes les conditions précédentes sont vérifiées, on cherche à résoudre les équations de vraisemblance. Soit on obtient des estimateurs de θ_0 explicites (cas rare). Soit on recherche numériquement des solutions au programme de maximisation de la vraisemblance (algorithmes du type Newton-Raphson, algorithme EM...).

Les méthodes de recherche de l'estimateur du maximum de vraisemblance du type Espérance-Maximisation sont souvent utiles dans le cadre de données incomplètes, du fait notamment de leurs performances au niveau numérique. Plus précisément, il s'agit d'une procédure itérative alternant une étape de recherche de la meilleure approximation de la vraisemblance latente conditionnellement aux observations (et pour la loi de paramètre courant), et une étape de maximisation d'une logvraisemblance. Ainsi, à la k-ième-étape, on dispose d'une valeur courante du paramètre notée $\theta_{(k)}$. On calcule la fonction

$$E_{\theta_{(k)}}[\ln L_n^*(t;\theta)|x] = n^{-1}\sum_{i=1}^{n} E_{\theta_{(k)}}[\ln f(T_i;\theta)|x] = Q(\theta;\theta_{(k)}), \tag{3-2}$$

puis on maximise en θ la fonction $Q(\theta;\theta_{(k)})$, pour obtenir $\theta_{(k+1)}$. Néanmoins, cet algorithme EM ne fournit pas toujours des estimateurs convergeant vers la bonne valeur du paramètre d'intérêt. Voir l'article initiateur de Dempster, Laird, Rubin (1977), Wu (1983) pour les conditions suffisantes de convergence de l'algorithme, et le survey de résultats de McLachlan et Krishnan (1997). La méthode se révèle utile pour traiter le cas de données groupées, censurées ou tronquées, de paramètres d'hétérogénéité, de mélanges de lois, d'hyperparamètres...

Notons que l'algorithme s'écrit sous une forme plus simple dans le cadre des modèles exponentiels; supposons donc qu'on puisse écrire la loi des latentes $T = (T_1,\ldots,T_n)$ sous la forme

$$f_T(t;\theta) = b(t)\exp\left(C(t)'\theta\right)/a(\theta),$$

pour tout $t = (t_1,\ldots,t_n)$. Ici, $\theta \in \mathbb{R}^q$, a,b sont des fonctions connues à valeurs réelles et $C : \mathbb{R}^n \longrightarrow \mathbb{R}^q$ est la statistique exhaustive totale du modèle. On n'observe que le vecteur x (données incomplètes), de loi

$$g_X(x) = \int_{\mathcal{T}(x)} f(t;\theta)\,dt,$$

en notant $\mathcal{T}(x)$ l'ensemble des vecteurs t qui fournissent l'observation x. Alors, en explicitant les conditions du premier ordre dans l'étape de maximisation de l'algorithme EM, on montre que ce dernier nécessite uniquement le calcul d'espérances conditionnelles de la statistique $C(t)$. Plus précisément, à la k-ième étape, on calcule la fonction

$$C_{(k)} \equiv E[C(T)|x,\theta_{(k)}],$$

et $\theta_{(k+1)}$ est défini comme racine en θ de l'équation

$$E[C(T)|\theta] = C_{(k)} = E_\theta[C(T)]. \tag{3-3}$$

En effet,

$$Q(\theta;\theta_{(k)}) = E_{\theta_{(k)}}[\ln b(T) + C'(T).\theta - \ln a(\theta)|x] = Cst + C'_{(k)}\theta - \ln a(\theta), \text{ et}$$

7. ce n'est pas toujours le cas: imaginer que T est un mélange de deux lois P_{θ_1} et P_{θ_2} de supports respectifs [0,1] et [3,4]. Si C censure toutes les observations au-delà de 2, alors θ_2 n'est pas identifiable.

$$\partial_\theta Q(\theta;\theta_{(k)}) = C_{(k)} - \partial_\theta \ln a(\theta).$$

Or

$$0 = E[\partial_\theta \ln f_T(T,\theta)] = E_\theta[C(T)] - \partial_\theta \ln a(\theta),$$

d'où la formule (3-3).

`Exemple. 3.1` Considérons durée T qui suit une loi exponentielle de paramètre $\theta = 1/\mu > 0$. Avec les notations précédentes, $a(\theta) = \mu$ et $C(T) = \sum_{i=1}^n T_i$. Cette durée est éventuellement censurée à droite par une variable aléatoire C. On observe donc un échantillon de réalisations de $X = \inf(T,C)$ et $\delta = \mathbf{1}\{T \leq C\}$. Alors, à la k-ième étape,

$$\begin{aligned} C_{(k)} &= E_{\mu_{(k)}}[\textstyle\sum T_i | x_1,\ldots,x_n;\delta_1,\ldots,\delta_n] = \sum_{i=1}^n x_i \mathbf{1}\{\delta_i = 1\} + \sum_{i=1}^n E_{\mu_{(k)}}[T_i | x_i,\delta_i = 0] \\ &= \sum_{i=1}^n x_i \mathbf{1}\{\delta_i = 1\} + \sum_{i=1}^n \mathbf{1}\{\delta_i = 0\} E_{\mu_{(k)}}[T_i | T_i > x_i] \\ &= \sum_{i=1}^n x_i \mathbf{1}\{\delta_i = 1\} + \sum_{i=1}^n \mathbf{1}\{\delta_i = 0\}(x_i + \mu_{(k)}) \\ &= \sum_{i=1}^n x_i + \mu_{(k)} \sum_{i=1}^n \mathbf{1}\{\delta_i = 0\}. \end{aligned}$$

Ainsi, $\mu_{(k+1)}$ satisfait l'équation

$$E_\mu[C(T)] = E_\mu[\sum_{i=1}^n T_i] = n\mu = C_{(k)}.$$

Une difficulté de la méthode EM consiste souvent à calculer explicitement l'espérance conditionnelle $Q(\theta;\theta_{(k)})$. Ce calcul, en dehors des modèles exponentiels, se révèle en pratique très lourd. C'est pourquoi les auteurs ont proposé des variantes de la méthode EM qui contournent cette difficulté. En particulier, les algorithmes EM simulés remplacent l'espérance conditionnelle (3-2) par une moyenne :

$$E_{\theta_{(k)}}[\ln f(T_i;\theta)|x] \# m^{-1} \sum_{j=1}^m \ln f_{\theta_{(k)}}(t_{ij}),$$

les t_{ij}, $j = 1,\ldots,m$, étant des valeurs simulées tirées suivant la loi de T_i conditionnelle à $X = x$, de paramètre $\theta_{(k)}$. Diverses variantes de cette méthode existent dans la littérature, sous le vocable "Algorithmes EM simulés" : voir le papier initial de Celeux et Diebolt (1985), Mc Fadden et Ruud (1994), une synthèse récente de Nielsen (2000). Signalons également une autre méthode par approximation stochastique de $Q(\cdot,\theta_{(k)})$ dans Delyon et al. (1999).

Par ailleurs, on peut montrer que l'information de Fisher du modèle latent s'écrit comme somme de l'information de Fisher du modèle observable et d'une matrice symétrique positive (voir Droesbeke et al. (1989)). Ainsi, perturber les données par des censures introduit une perte d'efficacité de l'estimation, visible notamment par l'accroissement de la borne d'efficacité de Cramer-Rao.

A titre d'illustration, reprenons le cas particulier important de données censurées aléatoirement à droite. On suppose le processus de censure non-informatif (c'est vrai en particulier si C est indépendante de T). La vraisemblance observable pertinente pour estimer θ est alors ramenée ici à l'équation 1-6. En supposant que C possède une densité g et une survie G, et avec des notations évidentes, on a alors

$$\begin{aligned} E_{\theta_0}[\ln L_n(X,\delta,\theta)] &= E_{\theta_0}[\delta \ln f_\theta(X) + (1-\delta)\ln S_\theta(X)] \\ &= \int G(t) \ln f_\theta(t) f_{\theta_0}(t)\, dt + \int S_{\theta_0}(c) \ln S_\theta(c) g(c)\, dc. \end{aligned}$$

Alors l'espérance du score observable est égale à

$$\frac{\partial}{\partial\theta}E_{\theta_0}[\ln L_n(x,\theta)] = \int G\frac{\partial_\theta f_\theta}{f_\theta}f_{\theta_0} + \int \frac{\partial_\theta S_\theta}{S_\theta}S_{\theta_0}g.$$

Si $\theta = \theta_0$, cette expression s'annule car

$$\frac{\partial}{\partial\theta}E_{\theta_0}[\ln L_n(x,\theta)]_{|\theta=\theta_0} = \frac{\partial}{\partial\theta}\left[\int Gf_\theta + \int S_\theta g\right] = -\frac{\partial}{\partial\theta}\int d(SG) = 0.$$

Dans les cas usuels et suffisamment réguliers, la vraie valeur du paramètre fournira donc un maximum approché (local sinon global) de la log-vraisemblance observée[8].

`Exercice 3.1` *Dans le cas d'une distribution des durées du type Weibull(μ,p) (voir Annexes) censurées à droite de manière indépendante, calculer l'estimateur du maximum de vraisemblance des paramètres d'intérêt.*

`Exercice 3.2` *Soit des durées T tronquées à gauche par une variable V indépendante de T. On note λ la fonction de hasard de T. Les données consistent en un échantillon i.i.d. $(T_i,V_i)_{i=1,\ldots,n}$, conditionné par les événements $T_i > V_i$. Montrer que la vraisemblance pertinente pour estimer la loi de T est $\prod_{i=1}^n \lambda(T_i)\exp(-\int_{V_i}^{T_i}\lambda)$. Si la loi de T appartient à une famille Weibull(μ,p), estimer ses paramètres.*

Dans le cadre des processus ponctuels, il est possible de développer une théorie du maximum de vraisemblance plus générale et plus satisfaisante: voir Andersen et al. (1993).

Application: le modèle Tobit

Le modèle Tobit est un cas particulier courant de modèle de durée. Il correspond à une durée, ou plus généralement une variable aléatoire réelle, censurée à gauche par une constante c (censure fixe, identique pour tous les individus, souvent prise égale à 0). Au lieu de la durée d'intérêt T, on n'observe donc que $X = T.\mathbf{1}\{T \geq c\} + c\mathbf{1}\{T \leq c\}$. Dans ce cas particulier, la connaissance de indicatrice δ est superflue.

Soit donc un échantillon i.i.d. $(x_1,\ldots,x_n)$. Notons f et F la densité et la survie de T. On voit facilement que

$$P(X \in [x_i - \Delta x_i, x_i])\# \begin{cases} f(x_i)\Delta x_i & \text{si } x_i > c, \\ F(c) & \text{si } x_i = c, \end{cases}$$

et donc la vraisemblance s'écrit ici

$$L_n(x) = \prod_{i=1}^n f(x_i)^{\mathbf{1}\{x_i>c\}}F(c)^{\mathbf{1}\{x_i\leq c\}}.$$

Notons m le nombre (aléatoire) de points non censurés dans l'échantillon, soit

$$m = \sum_{i=1}^n \mathbf{1}\{x_i > c\}.$$

Quitte à réindicer les observations, supposons que les durées non censurées sont les points d'indices $i = 1,\ldots,m$. Alors, la vraisemblance se réécrit

$$L_n(x) = F(c)^{n-m}\prod_{i=1}^n f(x_i).$$

A titre d'illustration, supposons que la variable latente est une gaussienne $\mathcal{N}(\mu,\sigma^2)$. Notons ϕ et Φ la densité et la fonction de répartition d'une loi gaussienne centrée réduite. Clairement, la

8. le calcul nous indique simplement que la méthode du maximum de vraisemblance peut produire un estimateur consistent

connaissance de la loi de X nous fournit celle de $\theta \equiv (\mu,\sigma^2)$. Le paramètre θ est donc identifiable dans ce modèle Tobit. La logvraisemblance s'écrit alors

$$\ln L_n(x) = (n-m)\ln\Phi(\frac{c-\mu}{\sigma}) - \frac{1}{2\sigma^2}\sum_{i=1}^{m}(x_i-\mu)^2 - \frac{m}{2}\ln(2\pi\sigma^2).$$

On constate facilement que les équations du premier ordre n'admettent pas de solutions explicites, ce qui nous incite à rechercher numériquement les racines de ce système.

A titre là aussi d'exemple, explicitons l'algorithme EM, dans le cas particulier où $\sigma = 1$. Il n'y a donc plus que le seul paramètre de position μ est estimer. Le programme de maximisation de la log-vraisemblance est ici

$$\max_{\mu}\sum_{i=1}^{n}\left(\mathbf{1}\{x_i = c\}\ln\Phi(c-\mu) - \frac{(x_i-\mu)^2}{2}\mathbf{1}\{x_i > c\}\right).$$

Soit la valeur initiale μ_0 de ce paramètre. L'espérance de la fonction critère par rapport à la loi de paramètre μ_0 est

$$\begin{aligned}
Q(\mu,\mu_0) &= \sum_{i=1}^{n} E_{\mu_0}[\ln f(T;\mu)/X = x_i] \\
&= \sum_{i=1}^{n}\left\{\frac{1}{2}E[-(T-\mu)^2|T\le c]\mathbf{1}\{x_i = c\} - \frac{(x_i-\mu)^2}{2}\mathbf{1}\{x_i > c\}\right\} - \frac{n}{2}\ln(2\pi) \\
&= -\sum_{i=1}^{n}\left\{\frac{1}{2\Phi(c-\mu_0)}\int_{-\infty}^{c}(t-\mu)^2\phi(t-\mu_0)\,dt.\mathbf{1}\{x_i = c\} + \frac{(x_i-\mu)^2}{2}.\mathbf{1}\{x_i > c\}\right\} - \frac{n}{2}\ln(2\pi) \\
&= -\frac{(n-m)}{2\Phi(c-\mu_0)}\left((\mu-\mu_0)^2\Phi(c-\mu_0) + 2(\mu_0-\mu)\phi(c-\mu_0) + (c-\mu_0)\phi(c-\mu_0) - \Phi(c-\mu_0)\right) \\
&- \sum_{i=1}^{m}\frac{(x_i-\mu)^2}{2} - \frac{n}{2}\ln(2\pi),
\end{aligned}$$

qui est un polynôme de degré deux en μ. Le coefficient dominant étant négatif, $Q(\cdot,\mu_0)$ admet un maximum noté μ_1. En général, on trouve μ_1 numériquement. Dans notre exemple, c'est la racine de la dérivée de $Q(\cdot,\mu_0)$, soit

$$\mu_1 = n^{-1}\sum_{i=1}^{m}x_i + \left(1-\frac{m}{n}\right)\left(\frac{\phi}{\Phi}(c-\mu_0) - \mu_0\right).$$

On itère ensuite le processus, en remplaçant μ_0 par la valeur courante μ_1.

`Exercice 3.3` *Soit un échantillon de n réalisations d'une variable gaussienne unidimensionnelle $\mathcal{N}(\mu,\sigma^2)$, tronquée à gauche par une constante c (modèle Tobit tronqué). Montrer que la vraisemblance des observations s'écrit*

$$L_n(x) = \prod_{i=1}^{n}\left[\sigma^{-1}\phi\left((x_i-\mu)/\sigma\right)/\left(1-\Phi((c-\mu)/\sigma)\right)\right].$$

3.2 Les modèles semiparamétriques

Si on ne veut ou ne peut pas spécifier entièrement la famille de loi à laquelle appartient T, ou bien si l'effet relatif des diverses covariables sur le phénomène étudié est pour nous le plus important à étudier, il est souvent fructueux d'utiliser des modèles semi-paramétriques. Ces derniers n'introduisent pas d'hypothèses (autres que de régularité) sur les fonctions de densité et/ou

de hasard, mais font des hypothèses sur la manière dont les diverses covariables vont influencer le déroulement du phénomène temporel.

On distingue traditionnellement deux grandes classes de modèles.

3.2.1 Les modèles à hasards proportionnels

Ils expriment un effet multiplicatif des diverses covariables sur la fonction de hasard. Dans ce but, on introduit une fonction de hasard dite "de base" λ_0, qui donne la forme générale du hasard. cette forme sera valable pour tous les individus. Les modèles à hasard proportionnels se caractérisent alors par la relation suivante, valable pour tout $t > 0$, et tout z

$$\lambda(t|z) = \lambda_0(t)c(z;\beta),$$

où z est le vecteur de covariables, β le paramètre d'intérêt et c une fonction positive. La plupart du temps, on particularise cette relation en considérant que l'effet des covariables se résume à celui d'une quantité réelle $z'\beta$ appelée index, c'est-à-dire

$$\lambda(t|z) = \lambda_0(t)c_0(z'\beta).$$

L'interprétation du modèle est la suivante : toutes choses égales par ailleurs, une covariable qui modifie c par rapport au niveau de référence, induit un effet multiplicatif de même ampleur sur le hasard à toute date t (dilatation proportionnelle de λ_0) : voir figure 3.

`Remarque` **3.2.** *Si à la fois λ_0 et c sont connues, on retombe sur un modèle paramétrique : voir annexe 8.4. Lorsque ce n'est pas le cas, le modèle est semi-paramétrique.*

Un cas particulier très important est constitué par le "modèle de Cox", qui spécifie pour c_0 la fonction exponentielle, c'est-à-dire, pour tout (t,z), on suppose que

$$\lambda(t|z) = \lambda_0(t)\exp(z'\beta).$$

Dans ce cas, lorsque $\lambda_0 > 0$, on peut réécrire le modèle sous la forme

$$\ln\Lambda_0(T) = -z'\beta + W,$$

où W suit un loi valeur-extrême : $P(W > w) = \exp(-\exp(w))$ pour tout w dans l'image de $\ln\Lambda_0(T) + z'\beta$. Cette distribution s'appelle également loi de Gumbel ou encore loi de Gompertz.

Prouvons ce dernier résultat : Pour tout $t \geq 0$ et z, $P(T > t|z) = S(t|z) = \exp(-\Lambda(t|z)) = \exp(-\Lambda_0(t)\exp(z'\beta))$. Or Λ_0 est une fonction continue strictement croissante (car $\lambda_0 > 0$), nulle en 0. Alors, en posant $u = \Lambda_0(t)$, pour tout z et tout u dans l'image de Λ_0, $P(\Lambda_0(T) > u) = \exp(-u\exp(z'\beta))$. On en déduit, pour tout z et pour tout w dans l'image de $\ln\Lambda_0 + z'\beta$, que $P(\ln\Lambda_0(t) + z'\beta > w) = \exp(-\exp(w))$.□

Classiquement, dans le modèle de Cox standard, Λ_0 est inconnue. Ce modèle revient donc à trouver β et éventuellement une fonction croissante inconnue h tels que

$$h(T) = -z'\beta + W.$$

3.2.2 Modèles à temps accéléré

Au lieu d'agir sur l'ordonnée de la fonction de hasard (dilatation verticale), il s'agit ici de modifier l'échelle des temps (l'abscisse t). Plus précisément, on dispose d'une durée T_0 de loi inconnue, dont on note λ_0 la fonction de hasard. Pour un individu de covariable z, la durée d'intérêt T sachant z, notée T_z, s'écrit

$$T_z = T_0.c(z;\beta),$$

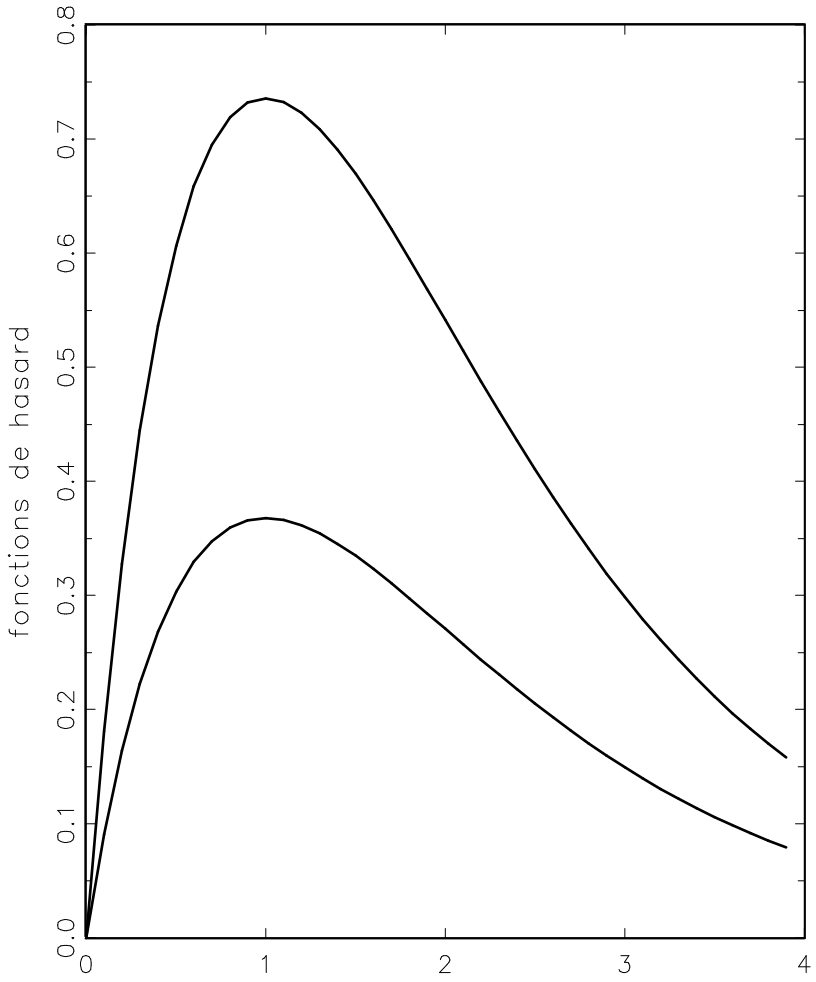

FIG. 3 – *Modèle à hasards proportionnels*

c étant une fonction positive à spécifier. Comme pour les modèles à hasards proportionnels, on choisit le plus souvent la forme exponentielle

$$T_z = T_0 . \exp(z'\beta).$$

On en déduit la relation, valable pour tout $t \geq 0$ et z,

$$\lambda(t|z) = \lambda_0(t \exp(-z'\beta)) \exp(-z'\beta), \text{ ou}$$

$$\Lambda(t|z) = \Lambda_0(t \exp(-z'\beta)).$$

En effet,

$$\begin{aligned}
\lambda(t|z) &= \lim_{\Delta \to 0} \frac{1}{\Delta t} P(T \in [t, t+\Delta t] | T > t, z) \\
&= \lim_{\Delta \to 0} \frac{1}{\Delta t} P\left(T_0 \in [\exp(-z'\beta)t, \exp(-z'\beta)t + \exp(-z'\beta)\Delta t] | T_0 > \exp(-z'\beta)t, z\right) \\
&= \exp(-z'\beta) \lim_{\tau \to 0} \frac{1}{\tau} P\left(T_0 \in [\exp(-z'\beta)t, \exp(-z'\beta)t + \tau] | T_0 > \exp(-z'\beta)t, z\right) \\
&= \exp(-z'\beta) \lambda_0\left(\exp(-z'\beta)t\right).
\end{aligned}$$

Dans le modèle à temps accéléré, les covariables modifient donc la fonction de hasard à la fois par une translation parallèle à l'axe des abscisses, et par une dilatation verticale comme dans le modèle à hasards proportionnels. En traçant les fonctions de hasard pour diverses valeurs de z, on voit rapidement si le modèle à temps accéléré et/ou le modèle à hasards proportionnels peut correspondre au phénomène étudié: comparer les allures des graphiques 3 et 4.

Ce modèle peut également s'écrire $Y = z'\beta + W$, où $Y = \ln T$ et où W est une variable de loi inconnue. En effet,

$$\begin{aligned}
P(T > t|z) &= \exp(-\Lambda(t|z)) = \exp\left(-\int_0^t \lambda(t|z)\, dt\right) \\
&= \exp\left(-\exp(-z'\beta) \int_0^t \lambda_0(u \exp(-z'\beta))\, du\right) \\
&= \exp\left(-\int_0^{t \exp(-z'\beta)} \lambda_0(v)\, dv\right) \\
&= \exp\left(-\Lambda_0(t \exp(-z'\beta))\right) = \exp\left(-\Lambda_0(\exp(\ln t - z'\beta))\right).
\end{aligned}$$

On en déduit, pour tout $w = \ln t - z'\beta$, c'est-à-dire tout w réel,

$$P(\ln T - z'\beta > w|z) = \exp\left(-\Lambda_0(\exp(w))\right).$$

Comme Λ_0 est inconnue (a priori), la loi de $\ln T - z'\beta$ est inconnue (mais indépendante de z).

`Exercice 3.4` *Montrer que l'intersection des modèles de Cox et des modèles à temps accéléré est fournie par la famille des lois de Weibull.*

3.3 Estimation nonparamétrique de la densité et de la fonction de hasard

Dans le cas d'une censure aléatoire droite, nous avons vu précédemment les estimateurs nonparamétriques naturels de la survie et de la fonction de hasard intégrée: $\hat{S}_{KM}$, estimateur de Kaplan-Meier et $\hat{\Lambda}_n$, estimateur de Nelson-Aalen. On peut déduire de ces derniers des estimateurs de la densité et de la fonction de hasard, par des techniques nonparamétriques classiques.

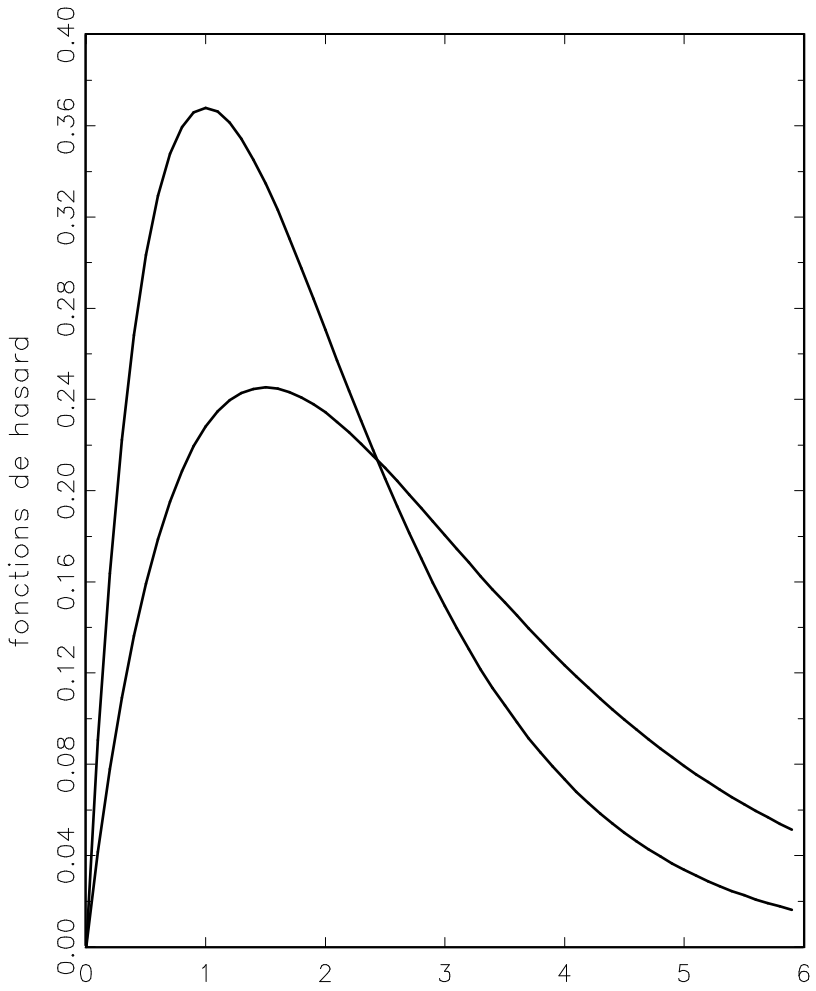

FIG. 4 – *Modèle à temps accéléré*

On appliquera la méthode du noyau de convolution (voir Bosq et Lecoutre (1987), par exemple).

Soit donc $K : \mathbb{R} \to \mathbb{R}$ un noyau réel, c'est-à-dire une fonction intégrable, d'intégrale 1. On supposera que K est continue, symétrique, à support compact, et à variations bornées. Soit, de plus, une suite de paramètres strictement positifs $(h_n)_{n\geq 1}$, dits "fenêtres". Elle vérifie $h_n \to 0$. Soit enfin $\tau < \inf\{t|P(T > t) = 0\}$.

S'il n'y a pas de censures, l'estimateur à noyau de f au point t est la convolution du noyau K avec la fonction de survie empirique S_n, i.e.

$$\tilde{f}_n(t) = -\int \frac{1}{h_n} K(\frac{t-u}{h_n}) S_n(dt) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{t-X_i}{h_n}\right).$$

En présence de données censurées, l'estimateur empirique naturel de la survie est $\hat{S}_{KM}$, ce qui nous donne

$$\hat{f}_n(t) = -\int \frac{1}{h_n} K\left(\frac{t-u}{h_n}\right) \hat{S}_{KM}(dt) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{t-X_{(i)}}{h_n}\right) \frac{\delta_{(i)}}{n-i+1} \hat{S}_{KM}(X_{(i)}-). \qquad (3\text{-}4)$$

Théorème 3.1 (Lo, Mack et Wang (1989)) *Si f (respectivement G) est continue (resp. C^2) en t, si $f(t) > 0$, si $nh_n^5 = o(\ln \ln n)$, alors presque sûrement,*

$$\limsup_n \left(\frac{nh_n}{2 \ln \ln n}\right)^{1/2} |\hat{f}_n(t) - f(t)| = \left[\frac{f(t)}{G(t)} \int K^2\right]^{1/2}.$$

Par ailleurs, on a la normalité asymptotique de cet estimateur.

Théorème 3.2 (Lo, Mack et Wang (1989)) *Si f (respectivement G) est continue (resp. C^2) en t, si $f(t) > 0$, si $nh_n^5 = o(1)$, $(\ln n)^2/(nh_n) \to 0$, alors presque sûrement,*

$$(nh_n)^{1/2}(\hat{f}_n(t) - f(t)) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0, \sigma_f^2(t)),$$

$$\sigma_f^2(t) = \frac{f(t)}{G(t)} \int K^2.$$

On remarquera que la précision des estimations se dégrade lorsque la fréquence des censures augmente (ou, de manière identique, lorsque G diminue). Des résultats de convergence uniforme sur des intervalles inclus dans $[0,\tau]$ peuvent être trouvés dans Stute et Diehl (1988).

Comme λ se définit par le rapport de f et de S, il semble naturel d'estimer la fonction de hasard par le rapport d'un estimateur de la densité sur un estimateur de la survie. Les auteurs ont étudié le plus souvent

$$\hat{\lambda}_n^{(1)} = \hat{f}_n / \hat{S}_{KM}.$$

On obtient alors les résultats classiques de convergence presque sûre, ponctuelle ou uniforme, et des résultats de convergence en loi, sous des hypothèses très similaires aux précédentes. Les vitesses de convergences sont identiques : voir, par exemple, Lo et al. (1989) (convergence presque sûre et en loi), Xiang (1994) (convergence presque sûre uniforme sur un compact)... et le survey de Padgett et McNichols (1984).

`Remarque` **3.3.** *Il est possible d'utiliser un autre rapport d'estimateurs, en remarquant qu'avec les notations habituelles, on a*

$$\lambda(t) = f(t)/S(t) = fG(t)/H(t).$$

On estime alors $H(t)$ par $H_n(t) \equiv n^{-1} \sum_{i=1}^{n} \mathbf{1}\{X_i > t\}$ et $fG(t)$ par

$$\widehat{fG}_n(t) = n^{-1}h^{-1} \sum_{i=1}^{n} K\left(\frac{t-X_i}{h_n}\right) \delta_i.$$

Cette procédure est a priori pertinente car

$$E[\widehat{fG}_n(t)] = -\int \frac{1}{h_n} K(\frac{t-u}{h_n}) G(u)\, dS(u) = \int \frac{1}{h_n} K(\frac{t-u}{h_n}) fG(u)\, du \to (fG)(t).$$

Voir Blum et Susarla (1980) pour les résultats détaillés de convergence de cette méthode.

On peut en fait estimer plus simplement la fonction de hasard, en dérivant une version lissée par un noyau de l'estimateur de Nelson-Aalen. Ainsi, on estime λ en tout point t tel que $P(T > t) > 0$ par

$$\hat{\lambda}_n^{(2)}(t) = \int \frac{1}{h_n} K\left(\frac{t-u}{h_n}\right) d\Lambda_n(du) = \frac{1}{nh_n} \sum_{i=1}^{n} \frac{\delta_i}{H_n(X_i-)} K\left(\frac{t-X_i}{h_n}\right). \tag{3-5}$$

On a le résultat de convergence uniforme presque sûre

Théorème 3.3 (Fermanian (1997)) *Si K est lipschitzien, à support compact et à variation bornée, si T et C admettent une densité continue sur $[0,\tau]$, si $nh_n^{1+\varepsilon}/\ln n \to +\infty$, pour un certain $\varepsilon > 0$, alors, pour tout $\nu > 0$, il existe une constante C^* telle que presque sûrement*

$$\left(\frac{nh_n}{\ln n}\right)^{1/2} \sup_{t\in[\nu,\tau]} |\hat{\lambda}_n^{(2)}(t) - \lambda(t)| < C^*.$$

Par ailleurs, la convergence en loi s'exprime par

Théorème 3.4 *Si K est à support compact, si λ (respectivement G) est $C^2(\mathbb{R})$ (resp. C^1) au voisinage de t, si $h_n \ln(\ln n) \to 0$, $nh_n^5 \to 0$, si $nh_n^{(1+\varepsilon)} \to \infty$, $\varepsilon > 0$, alors*

$$(nh_n)^{1/2}(\hat{\lambda}_n^{(2)}(t) - \lambda(t)) \xrightarrow[n\to\infty]{\mathcal{L}} \mathcal{N}(0,\sigma_\lambda^2(t)),$$

$$\sigma_\lambda^2(t) = \frac{\lambda(t)}{H(t)} \int K^2.$$

`Remarque` **3.4.** *Comme d'ordinaire en estimation fonctionnelle par la méthode du noyau se pose le problème du choix de la suite de fenêtres h_n. La qualité de l'estimateur en dépend cruciale-ment. Concernant les fonctions de densité ou de hasard, les auteurs ont généralement proposé des méthodes basées sur des critères de validation croisée : Patil (1993a,1993b), Grégoire (1993) etc (voir également Sarda et Vieu (1991) en présence de données non censurées). Par exemple, dans le cas d'un des estimateurs $\hat{\lambda}_n$ précédents, la méthode de validation croisée se résume de la façon suivante : on cherche $\hat{h}_0$ qui minimise le critère $ISE(h) = \int (\hat{\lambda}_n - \lambda)^2 w$ (écart quadratique intégré), w étant une fonction positive fixée à support compact. Cela revient à cherche $\hat{h}_0$ tel que*

$$\hat{h}_0 \in \arg\min_h \int \hat{\lambda}_n^2 w - \int \hat{\lambda}_n \lambda w.$$

Comme on ne connaît pas λ, il faut se contenter d'un critère ISE approché

$$\widehat{ISE}(h) = \int \hat{\lambda}_n^2 w - n^{-1} \sum_{i=1}^{n} \hat{\lambda}_{n,-i}(X_i) \frac{w}{H}(X_i)\delta_i,$$

$$\hat{\lambda}_{n,-i}(X_i) = \frac{1}{(n-1)h} \sum_{j\neq i} \frac{\delta_i}{H(X_j)} K(\frac{X_i - X_j}{h}).$$

Alors, sous des hypothèses de régularité, on montre que $\hat{h} = argmin_h \widehat{ISE}(h)$ vérifie $(\hat{h} - \hat{h}_0)/\hat{h}_0 = O_P(n^{-1/10})$, et que $n^{-1/10}(\hat{h} - \hat{h}_0)/\hat{h}_0)$ est asymptotiquement normal. Voir une autre approche dans Fermanian (1999).

3.4 Estimation nonparamétrique en présence de covariables

Si la population est fractionnée en plusieurs groupes de taille "raisonnable", il est possible d'appliquer les estimateurs nonparamétriques de S, f, Λ ou λ précédents dans chaque sous-groupe.

Si ce n'est pas le cas, on peut regrouper les individus dont les covariables ne sont pas "trop éloignées". Cette notion n'a de sens que pour des covariables continues, ou discrétisées (mais reflétant une échelle ordonnée sous-jacente).

Plus précisément, plaçons-nous dans le cadre classique de la censure aléatoire droite. Chaque individu j se voit affecté d'une covariable $z_j \in \mathbb{R}^q$, réalisation de la variable aléatoire Z continue, de densité g. Soit K un noyau de dimension q. On définit des processus empiriques conditionnels lissés par

$$\hat{H}_0(t|z) = \frac{1}{nh^q} \sum_{i=1}^{n} \mathbf{1}\{X_j > t\} K\left(\frac{z - z_j}{h}\right),$$

$$\hat{H}_1(t|z) = \frac{1}{nh^q} \sum_{i=1}^{n} \mathbf{1}\{X_j > t, \delta_j = 1\} K\left(\frac{z - z_j}{h}\right).$$

L'estimateur de Nelson-Aalen en t conditionnel à $Z = z$ est alors défini par

$$\hat{\Lambda}_n(t|z) = -\int \frac{\hat{H}_1(du|z)}{\hat{H}_0(u - |z)},$$

et l'estimateur de Kaplan-Meier conditionnel à $Z = z$ est, pour tout $t < X_{(n)}$ et z,

$$\hat{S}_{KM}(t|z) = \prod_{i=1}^{n} \left(1 - \frac{\delta_i \mathbf{1}\{X_i \leq t\} K(h^{-1}(z - Z_i))}{\sum_{j=1}^{n} \mathbf{1}\{X_j \geq X_i\} K(h^{-1}(z - Z_j))}\right).$$

Comparer cette dernière formule avec 2-1. En lissant $\hat{\Lambda}_n(\cdot|\cdot)$ et $\hat{S}_{KM}(\cdot|\cdot)$, il est possible d'obtenir les équivalents des formules 3-4 et 3-5, conditionnellement à $Z = z$ (densité et fonction de hasard conditionnels).

4 Le modèle à hasards proportionnels

Comme on l'a vu précédemment, le modèle à hasards proportionnels s'écrit sous sa forme la plus générale

$$\lambda(t|z) = \lambda_0(t)c(z;\beta),$$

en notant β le paramètre d'intérêt, z le processus de covariables et c une fonction positive fixée (et connue). λ_0 est une fonction inconnue, appelée fonction de hasard de base. On notera Λ_0 sa primitive, appelée fonction de hasard de base intégrée.

La plupart du temps, l'effet des covariables sur la durée T se résume à celui de l'index $z'\beta$. Le modèle s'écrit alors

$$\lambda(t|z) = \lambda_0(t)c_0(z'\beta).$$

La plupart des auteurs étudient en fait un cas particulier, dit "modèle de Cox", défini par

$$\lambda(t|z) = \lambda_0(t)\exp(z'\beta).$$

Lorsque la k-ième covariable est continue, le k-ième coefficient de β vérifie pour tout t

$$\beta_k = \frac{\partial}{\partial z_k}\ln\lambda(t|z).$$

Donc β_k mesure l'élasticité du taux de hasard par rapport à la k-ième covariable z_k, qui est supposée ne pas varier dans le temps.

Sauf cas contraire explicitement signalé, nous allons nous placer dans le cadre du modèle de Cox tout au long de cette partie.

`Remarque` **4.1.** *Dans la quasi totalité des cas, on se fixe la fonction c_0. Néanmoins, il est possible de l'estimer nonparamétriquement (O'Sullivan (1993), Fan et al. (1997)).*

Comme on l'a vu en 3.2.1, le modèle de Cox peut se réécrire comme un modèle linéaire car

$$\ln\Lambda_0(t|z) = -z'\beta + w, \tag{4-1}$$

où w suit une loi W de Gompertz. Evidemment, la théorie classique des MCO ne peut s'appliquer ici directement, car :

- la loi de W n'est pas centrée : $E[W] = -\gamma \sim -0{,}5772$ (γ constante d'Euler), $Var(W) = \pi^2/6$.
- on ne connaît pas en général Λ_0.
- on observe en général des données non complètes.

Notons que le premier obtacle peut être levé facilement en posant $W' = W + \gamma$, W variable centrée. On modifie le paramètre β en introduisant une composante égale à 1 pour capter la constante γ. Ainsi, si on connaît Λ_0 et si les données sont complètes, i.e. si on dispose d'un échantillon i.i.d. $(T_i,z_i)_{i=1\ldots,n}$, on peut estimer le paramètre β par la procédure classique des moindres carrés. Si $\beta \in \mathbb{R}^p$ et $z_i \in \mathbb{R}^p$, alors

$$\hat{\beta}_n = -\left[\sum_{i=1}^{n} z_i z_i'\right]^{-1} \sum_{i=1}^{n} z_i \left[\ln\Lambda_0(T_i)\right].$$

On peut montrer que le modèle de Cox s'écrit simplement en fonction de $Y' = \Lambda_0(T)$. Plus précisément, on a conditionnellement à z,

$$Y' = \exp(-z'\beta) + \tilde{W}. \tag{4-2}$$

En effet, pour tous u et z, on a

$$\begin{aligned} P(\Lambda_0(T) - \exp(-z'\beta) > u|z) &= P(T > \Lambda_0^{-1}(u + \exp(-z'\beta))|z) \\ &= \exp\left(-\Lambda\left(\Lambda_0^{-1}(u + \exp(-z'\beta))|z\right)|z\right) \\ &= \exp\left(-(u + \exp(-z'\beta)).\exp(z'\beta)\right) = \exp\left(-u\exp(z'\beta) - 1\right). \end{aligned}$$

Ainsi,

$$P(\tilde{W} + \exp(-z'\beta) > u|z) = \exp(-u\exp(z'\beta)).$$

Autrement dit, la loi de $\tilde{W}$ conditionnellement à z est centrée, a pour support $[-\exp(-z'\beta), +\infty[$, et prend la forme d'une loi exponentielle (le montrer). On en déduit que, en présence de données complètes, une procédure de type "moindres carrés non linéaires" du type

$$\arg\min_{\beta} \sum_{i=1}^{n} |\Lambda_0(X_i) - \exp(-z_i'\beta)|^2$$

fournit un convergent estimateur de β.

A partir de maintenant, on supposera que les durées sont soumises à des censures doites indépendantes de T.

4.1 Estimation paramétrique du modèle de Cox

Supposons qu'on connaisse Λ_0, éventuellement à un paramètre de dimension finie près. On est ramené à une estimation paramétrique de β sur données censurées. Comme on l'a vu en 1-6, la logvraisemblance s'écrit ici

$$\begin{aligned} \ln L_n(\beta) &= \sum_{i=1}^{n} \delta_i \ln f(X_i|z_i) + (1-\delta_i)\ln S(X_i|z_i) \\ &= \sum_{i=1}^{n} \delta_i \ln \lambda(X_i|z_i) + \ln S(X_i|z_i) \\ &= \sum_{i=1}^{n} \left\{\delta_i[z_i'\beta + \ln\lambda_0(X_i)] - \exp(z_i'\beta)\Lambda_0(X_i)\right\}, \end{aligned}$$

d'où la fonction score

$$\begin{aligned} \frac{\partial}{\partial\beta}\ln L_n(\beta) &= \sum_{i=1}^{n} z_i[\delta_i - \exp(z_i'\beta)\Lambda_0(X_i)] \\ &= -\sum_{i=1}^{n} z_i \exp(z_i'\beta) r_i^*(\beta) \end{aligned}$$

où $r_i^* = \Lambda_0(X_i) - \delta_i \exp(-z_i'\beta)$. Dans le cas d'un modèle sans censure ($\delta_i = 1$ pour tout i), r_i^* serait le résidu théorique du modèle (4-2), c'est-à-dire $r_i = \Lambda_0(X_i) - \exp(-z_i'\beta)$. On appelle les r_i^* les résidus généralisés.

`Exercice 4.1` *Montrer que dans le cas de la censure aléatoire droite, $r_i^* = E[r_i|z_i,\delta_i,X_i]$.*

L'estimateur du maximum de vraisemblance de β, noté $\hat{\beta}$, est donc racine de l'équation exprimant l'orthogonalité entre les résidus et une fonction des covariables, i.e.

$$\sum_{i=1}^{n} z_i \exp(z_i'\hat{\beta}) r_i^*(\hat{\beta}) = 0.$$

En dérivant une seconde fois la logvraisemblance $\ln L_n(\beta)$, on obtient l'information de Fisher

$$\begin{aligned} I_n(\beta) &= E_\beta[-\frac{\partial^2 \ln L_n(\beta)}{\partial\beta\partial\beta'}|z] \\ &= \sum_{i=1}^n z_i z_i' \exp(z_i'\beta) E[\Lambda_0(X_i)|z_i]. \end{aligned}$$

Pour une censure aléatoire droite C quelconque, la quantité $E[\Lambda_0(X_i)|z_i]$ est peu maniable car elle fait intervenir explicitement la loi de C. On remarque néanmoins que $E[\Lambda_0(X_i)|z_i] \leq E[\Lambda_0(T_i)|z_i]$ car Λ_0 est croissante. De plus, un calcul simple montre que, lorsque $\lim_{u\to\infty} \Lambda_0(u) = +\infty$,

$$\begin{aligned} E[\Lambda_0(X_i)|z_i] &= \int \Lambda_0(t) S(t|z_i)\, d\Lambda(t|z_i) \\ &= \exp(-z_i'\beta) \int \Lambda(t|z_i) \exp(-\Lambda(t|z_i))\, d\Lambda(t|z_i) = \exp(-z_i'\beta). \end{aligned}$$

Donc, pour tout i, il existe un facteur $\pi_i \in [0,1]$ tel que $E[\Lambda_0(X_i)|z_i] = \exp(-z_i'\beta)\pi_i$. On a alors

$$I_n(\beta) = \sum_{i=1}^n z_i z_i' \pi_i.$$

En l'absence de censure, l'information de Fisher correspondant au modèle latent $I_n^*(\beta)$ serait

$$I_n^* = \sum_{i=1}^n z_i z_i' = \sum_{i=1}^n z_i z_i' \pi_i + \sum_{i=1}^n z_i z_i'(1-\pi_i) = I_n(\beta) + \tilde{I}_n(\beta) >> I_n(\beta).$$

Donc le fait d'avoir des observations censurées introduit une perte d'information, qui s'exprime par la présence de la matrice (positive) $\tilde{I}_n(\beta)$ [9].

La théorie usuelle des tests asymptotiques peut être utilisée avec $L_n(\beta)$ et son maximisateur $\hat{\beta}$: tests de nullité de certains coefficients de β, par l'intermédiaire des tests de Wald, du score, du rapport de vraisemblance... (voir Droesbeke et al.(1989)).

La plupart des résultats de la statistique paramétrique s'expriment relativement aisément dans le cadre du modèle de Cox, lorsqu'on se donne la fonction de hasard de base λ_0. Mais la véritable originalité du modèle de Cox provient de l'introduction d'une méthode de résolution semiparamétrique particulière ; elle est fondée sur la maximisation d'une partie de la vraisemblance des observations, appelées "vraisemblance partielle".

4.2 La vraisemblance partielle

Pour tout couple de variables aléatoire (U,V), on notera f_U et $f_{U|V=v}(\cdot|v)$ les densités respectives de U et U conditionnellement à $V = v$.

Supposons qu'on observe un échantillon de réalisations d'un couple de variables (U,V), soit $(u_i,v_i)_{i=1,\ldots,m}$. Les observations ne sont a priori ni indépendantes, ni identiquement distribuées. La vraisemblance des observations s'écrit alors

$$\begin{aligned} \mathcal{L}(u_1,v_1;\ldots;u_m,v_m) &= \mathcal{L}(u_1,v_1)\mathcal{L}(u_2,v_2;\ldots;u_m,v_m|u_1,v_1) \\ &= \prod_{j=1}^m \mathcal{L}(u_j,v_j|u^{(j-1)},v^{(j-1)}), \end{aligned}$$

9. la diminution de l'information de Fisher aura pour conséquence l'accroissement de la borne de Cramer-Rao pour l'estimation de β

où $u^{(j)} = (u_1, \ldots, u_j)$ et $v^{(j)} = (v_1, \ldots, v_j)$. On peut alors écrire

$$\mathcal{L}(u_1,v_1;\ldots;u_m,v_m) = \prod_{j=1}^{m} \mathcal{L}(u_j|u^{(j-1)},v^{(j-1)}) \cdot \mathcal{L}(v_j|u^{(j)},v^{(j-1)}).$$

La vraisemblance des observations est donc le produit de deux termes. Le second, nommé vraisemblance partielle basée sur $v_1, \ldots, v_m$ dans $(u_1,v_1;\ldots;u_m,v_m)$ est par définition

$$\mathcal{L}_p(u_1,v_1;\ldots;u_m,v_m) = \prod_{j=1}^{m} \mathcal{L}(v_j|u^{(j)},v^{(j-1)}).$$

Lorsqu'il n'y a pas d'ambiguïté de notations, on l'appellera plus simplement la vraisemblance partielle de v dans (u,v). Lorsque toutes les lois conditionnelles admettent une densité par rapport à la mesure de Lebesgue, la vraisemblance partielle s'écrit

$$\mathcal{L}_p(u_1,v_1;\ldots;u_m,v_m) = \prod_{j=1}^{m} f_{V_j|U^{(j)}=u^{(j)},V^{(j-1)}=v^{(j-1)}}(v_j).$$

Si $m = 1$ la vraisemblance partielle est simplement la vraisemblance de V_1 conditionnellement à U_1. C'est le seul cas où $\mathcal{L}_p$ s'interprète de cette manière. En général, la vraisemblance partielle n'est ni une vraisemblance (i.e. une dérivé de Radon-Nikodym), ni une vraisemblance conditionnelle (vraisemblance des observations contionnellement à d'autres variables considérées alors comme fixes), ni une pseudo-vraisemblance (vraisemblance d'un modèle mal spécifié ; voir Gouriéroux et al.(1984)). Par contre, elle s'interprète parfois comme une vraisemblance marginale (cf. infra).

Dans certain cas, on peut donc utiliser $\mathcal{L}_p$ comme s'il s'agissait de la vraisemblance totale des observations. La méthode divise l'information présente dans la vraisemblance en deux parties : l'information pertinente pour estimer les paramètres du modèle, et un "bruit" qu'on peut négliger. Avec nos notations, le bruit est apporté ici par U. Pour pouvoir résumer la vraisemblance totale en la seule partie $\mathcal{L}_p$, il est nécessaire que la partie "bruitée" laissée de côté ne fasse pas intervenir de manière "informative" les paramètres qu'on cherche à estimer. Notons toutefois que formellement, ces derniers apparaissent dans cette partie de la vraisemblance via le conditionnement par $v^{(j-1)}$.

`Exemple.` **4.1** On réalise une enquête de fiabilité sur un type de machine livré dans n entreprises. Le dernier jour de chaque mois, on interroge chaque entreprise pour savoir si la machine a fonctionné correctement ou non lors du mois écoulé. On étudie alors la durée qui s'écoule entre la livraison de la machine et la première panne rencontrée. Au fil des mois, certaines entreprises ne répondent plus (disparition de l'entreprise, refus de réponse...). On suppose que ce processus d'attrition, qui engendre une censure à droite, est indépendant de la durée d'intérêt. Pour le j-ème mois, on note v_j le nombre de machines tombées en panne, u_j le nombre de machines des entreprises qui ne répondent plus (de durées censurées à droite), n_j le nombre de machines à risque au début du j-ième mois. Alors, on a évidemment $n_{j+1} = n_j - u_j - v_j$ pour tout j. Par ailleurs, on notera ϕ_j la probabilité de tomber en panne lors du j-ième mois, sachant que la machine fonctionne au début du j-ième mois . On suppose qu'on n'accède qu'à la connaissance de la suite des $(u_j,v_j)_{j=1,\ldots,m}$ (ou, ce qui revient au même, $(n_j,v_j)_j$). Alors, la vraisemblance partielle de v dans (u,v) est

$$\begin{aligned}\mathcal{L}_p &= \prod_{j=1}^{m} Prob(\ v_j \text{ pannes le mois } j|u_1,v_1;\ldots;u_{j-1},v_{j-1};\ u_j \text{ censures le mois } j) \\ &= \prod_{j=1}^{m} C^{v_j}_{n_j-u_j} \phi_j^{v_j} (1-\phi_j)^{n_j-u_j-v_j}.\end{aligned}$$

En effet, si on sait que u_j machines sont censurées le i-ième mois, seules $n_j - u_j$ machines sont désormais "à risque" ce mois-là. On a considéré qu'une machine censurée le j-ième mois ne peut

défaillir qu'à partir du $j+1$-ième, et non du j-ième mois. Il paraît raisonnable en effet de supposer que le client qui constate une panne le j-ième mois répondra à l'enquête pour se plaindre ! Notons enfin que, lors du j-ième mois, la connaissance du processus passé, i.e. des $(u_1,v_1;\ldots;u_{j-1},v_{j-1})$, se transcrit dans la vraisemblance par l'intermédiaire du seul coefficient n_j.

Il est remarquable que dans bien des cas, la vraisemblance partielle se comporte comme une véritable vraisemblance, c'est-à-dire qu'on peut développer une théorie asymptotique similaire au cas standard. Evidemment, cette "bonne" propriété dépend du choix des variables (u,v) retenu.

Ainsi, soit l'estimateur du "maximum de vraisemblance partielle"

$$\hat{\beta} \in \arg\max \mathcal{L}_p(\beta).$$

Pour obtenir la consistance de $\hat{\beta}$, il faudrait d'abord vérifier que l'argument maximum de la logvraisemblance partielle, ou de son espérance

$$E[m^{-1}\sum_{j=1}^{m} \ln \mathcal{L}(v_j|u^{(j)},v^{(j-1)};\beta)]$$

tend bien vers β lorsque $m \to \infty$, et que la logvraisemblance partielle converge uniformément par rapport au paramètre dans un voisinage de β (conditions usuelles de consistance de l'e.m.v. : voir Gourièroux et Monfort (1989)). Ces deux conditions dépendent crucialement de la forme retenue pour $\mathcal{L}_p$, et ne peuvent être précisées davantage dans le cadre de ce formalisme très général.

Par contre, si on suppose la consistance de $\hat{\beta}$, on peut comprendre pourquoi ce dernier est généralement asymptotiquement normal. En effet, cet estimateur vérifie dans les cas standards les conditions du premier ordre, i.e. est racine de $\sum_{j=1}^{m} S_j$ avec

$$S_j = \frac{\partial}{\partial\beta} \ln \mathcal{L}(v_j|u^{(j)},v^{(j-1)};\beta).$$

En supposant l'existence des densités correspondants aux distributions conditionnelles qui apparaissent dans l'expression précédente, S_j est évidemment une variable aléatoire centrée, comme tout score.

De plus, du fait de caractère séquentiel du processus observé, on a également, pour tout $j<k$,

$$E[S_jS_k] = E[S_jE[S_k|u^{(j)},v^{(j-1)}] = 0.$$

On notera I_j la variance de S_j.

Alors, le score partiel sur l'ensemble des observations est somme de variables non corrélées et centrées :

$$E_\beta[\frac{\partial}{\partial\beta}\mathcal{L}_p(\beta)] = 0, \qquad Var_\beta(\frac{\partial}{\partial\beta}\mathcal{L}_p(\beta)) = \sum_{j=1}^{m} I_j.$$

Sous des conditions de dépendance, on aura que $m^{-1/2}\partial\mathcal{L}_p/\partial\beta$ est asymptotiquement normal, de variance asymptotique approchée par $m^{-1}\sum_{j=1}^{m} I_j$, ce qui implique usuellement la normalité asymptotique de $\hat{\beta}-\beta$. Une théorie générale des propriétés de le vraisemblance partielle et de ses maximisateurs est présentée dans Wong (1986).

Nous allons appliquer cette méthode aux modèles à hasards proportionnels.

Application aux modèles à hasards proportionnels.

Soit le modèle à hasards proportionnels le plus général caractérisé par la relation, valable pour tout t et tout z,

$$\lambda(t|z) = \lambda_0(t)c(z,\beta).$$

Ici, le processus de covariables peut dépendre du temps. On suppose que la durée d'intérêt T est continue. Alors, presque sûrement, les décès auront lieu à des instants distincts $t_{(1)} < \cdots < t_{(m)}$. On le supposera désormais. On pose également $t_{(0)} = 0$. Juste avant tout instant $t_{(j)}$ [10], les indices des individus à risque (i.e. qui risquent de décéder à partir de $t_{(j)}$) forment un ensemble $\mathcal{R}_j$. Notons que nous n'avons pas fait d'hypothèses sur le type de perturbations qui pourraient affecter les durées observées (type de censures éventuelles). Dans ce cadre très général et tout en gardant les notations précédentes, u_j sera toute l'histoire du processus entre les dates $t_{(j-1)}$ et $t_{(j)}-$, plus le fait qu'un décès est observé en $t_{(j)}$; v_j sera l'indice de l'individu qui décède en $t_{(j)}$; autrement dit $v_j = (j)$ sera l'indice de la j-ième statistique d'ordre des durées observées. Alors,

$$\begin{aligned}
\mathcal{L}_p &= \prod_{j=1}^{m} Prob(\ (j) \text{ décède dans l'intervalle } [t_{(j)},t_{(j)} + \Delta t]|u_1,v_1;\ldots;u_{j-1},v_{j-1};u_j) \\
&= \prod_{j=1}^{m} Prob(T_{(j)} \in [t_{(j)},t_{(j)} + \Delta_j]|\mathcal{R}_j,\beta,u_j) \\
&= \prod_{j=1}^{m} \frac{P(T_{(j)} \in [t_{(j)},t_{(j)} + \Delta_j]|T_{(j)} \geq t_{(j)};\beta)}{\sum_{k\in\mathcal{R}_j} P(T_{(k)} \in [t_{(j)},t_{(j)} + \Delta_j]|T_{(k)} \geq t_{(j)};\beta)} \\
&\sim \prod_{j=1}^{m} \frac{\lambda(t_{(j)}|z_{(j)}(t_{(j)});\beta)}{\sum_{k\in\mathcal{R}_j} \lambda(t_{(j)}|z_{(k)}(t_{(j)});\beta)} \\
&\sim \prod_{j=1}^{m} \frac{\lambda_0(t_{(j)})c(z_{(j)}(t_{(j)});\beta)}{\sum_{k\in\mathcal{R}_j} \lambda_0(t_{(j)})c(z_{(k)}(t_{(j)});\beta)} \\
&= \prod_{j=1}^{m} \frac{c(z_{(j)}(t_{(j)});\beta)}{\sum_{k\in\mathcal{R}_j} c(z_{(k)}(t_{(j)});\beta)}.
\end{aligned}$$

Lorsqu'on est dans le cadre du modèle de Cox, la vraisemblance partielle est alors

$$\prod_{j=1}^{m} \frac{\exp\left(z'_{(j)}(t_{(j)})\beta\right)}{\sum_{k\in\mathcal{R}_j} \exp\left(z'_{(k)}(t_{(j)})\beta\right)}. \tag{4-3}$$

L'expression obtenue en 4-3 est appelée vraisemblance partielle de Cox.

`Remarque` **4.2.** *Le modèle de Cox rentre dans la classe des modèles à direction unique révélatrice (single-index models), c'est-à-dire les modèles pour lesquels l'espérance conditionnelle de T sachant $Z = z$ est fonction de la seule quantité $z'\beta$. Même si les méthodes d'estimation diffèrent très nettement de la méthode par vraisemblance partielle, on peut néanmoins estimer le paramètre β en utilisant les estimateurs single-index existants sur données incomplètes: Han (1987), Härdle et Stoker (1989), Horowitz et Härdle (1996), Ichimura (1993), Powell et al. (1989), Sherman (1993). Néanmoins, c'est déconseillé car il deviendra difficile d'évaluer la loi de T sachant z, sans hypothèse sur la loi des censures* [11]

4.3 Propriétés asymptotiques du Modèle de Cox

On se place dans le cadre de la censure aléatoire droite. Soit le modèle de Cox classique avec covariables dépendant du temps. On peut l'écrire pour tout t,

$$\lambda(t|Z) = Y(t)\lambda_0(t)\exp(Z'(t)\beta), \tag{4-4}$$

10. on dit également en $t_{(j)}-$
11. ie on ne pourra identifier λ_0 en général.

où λ_0 est le hasard de base, $Z(t)$ est le processus de covariables et $Y(t)$ est une indicatrice valant 1 si l'individu en question est à risque (pas encore décédé ni censuré) à la date t, et valant 0 sinon.

Introduire Y permet de raisonner sur un échantillon total de taille fixe n, et non de taille variable au cours du temps m précédemment utilisée, et qui correspondait aux instants de "vraies" durées.

Soit la filtration (suite croissante de σ-algèbres)

$$(\mathcal{F}_t)_{t\geq 0} = (\sigma\left((Y_i(t),Z_i(t))_{i=1,\ldots,n}\right))_{t\geq 0}\,.$$

On suppose que, pour tout i, les processus $(Y_i(\cdot))_{t\geq 0}$ et $(Z_i(\cdot))_{t\geq 0}$ sont prédictibles, i.e. pour tout t et tout i,

$$E[Y_i(t)|\mathcal{F}(t-)] = Y_i(t), \qquad E[Z_i(t)|\mathcal{F}(t-)] = Z_i(t).$$

De plus, pour tout i, $(Z_i(\cdot))_{t\geq 0}$ sera supposé localement borné (cf. annexe 8.3).

Tout d'abord, introduisons un certains nombre de notations. Pour tous vecteurs $a = (a_1,\ldots,a_p)'$ et $b = (b_1,\ldots,b_p)'$, on pose $|a| = (a'a)^{1/2}$ la norme euclidienne de a, $a\otimes b = [a_i b_j]_{i,j}$ et $a^{\otimes 2} = a\otimes a$. De plus, pour toute matrice M, $\|M\| = \sup_{i,j}|m_{i,j}|$.

Avec les notations du modèle de Cox (4-4), on posera

$$S^{(0)}(\beta,t) = n^{-1}\sum_{i=1}^{n} Y_i(t)\exp(Z_i'(t)\beta),$$

$$S^{(1)}(\beta,t) = n^{-1}\sum_{i=1}^{n} Y_i(t)Z_i(t)\exp(Z_i'(t)\beta),$$

$$S^{(2)}(\beta,t) = n^{-1}\sum_{i=1}^{n} Y_i(t)Z_i(t)^{\otimes 2}\exp(Z_i'(t)\beta),$$

$$s^{(j)}(\beta,t) = E[S^{(j)}(\beta,t)],\ j = 0,1,2,$$

$$e = s^{(1)}/s^{(0)}, \qquad v = s^{(2)}/s^{(0)} - e^{\otimes 2}.$$

Par ailleurs, on suppose que l'expérience s'arrête à une date fixe t_0 telle que $P(T > t_0) > 0$. Ainsi, les individus sont soumis à deux types de censures : l'une aléatoire de $t = 0$ à $t = t_0-$, et une autre fixe en t_0. On notera comme d'habitude $\mathcal{R}_j$ l'ensemble à risque à la date X_j (les individus k pour lesquels $X_k \geq X_j$).

Soit $\hat{\beta} \in \arg\max \mathcal{L}_p(\beta)$ où

$$\mathcal{L}_p(\beta) = \prod_{i=1}^{n}\left\{\frac{\exp(Z_i'(X_i)\beta)}{\sum_{j\in\mathcal{R}_j}\exp(Z_j'(X_i)\beta)}\right\}^{\delta_i}. \tag{4-5}$$

Notons que cette expression est identique à celle donnée en 4-3. La puissance δ_j nous permet de ne tenir compte que des individus non censurées dans la vraisemblance partielle. Du même coup, il est légitime de remplacer X_i par T_i dans l'expression 4-5.

Soient les hypothèses suivantes :

i. $\int_0^t \lambda_0 < \infty$;

ii. il existe un voisinage $\mathcal{B}$ de β tel que, pour $j = 0,1,2$ et en probabilité

$$\sup_{t\in[0,t_0],b\in\mathcal{B}} \|S^{(j)}(b,t) - s^{(j)}(b,t)\| \underset{n\to\infty}{\longrightarrow} 0;$$

iii. il existe $\delta > 0$ tel que

$$n^{-1/2} \sup_{i,t} |Z_i(t)|Y_i(t)\mathbf{1}\{Z_i'(t)\beta > -\delta|Z_i(t)|\} \underset{n\to\infty}{\longrightarrow} 0;$$

iv. pour tout b dans $\mathcal{B}$ et tout $t < t_0$, $s^{(1)}(b,t) = \partial/\partial b\, s^{(0)}(b,t)$, $s^{(2)}(b,t) = \partial^2/\partial^2 b\, s^{(0)}(b,t),$; les fonctions $s^{(j)}(b,t)$, $j = 0,1,2$ sont continues en $b \in \mathcal{B}$, uniformément en $t \in [0,t_0]$, bornées sur $\mathcal{B} \times [0,t_0]$, et $s^{(0)}$ est bornée inférieurement par une constante strictement positive sur $\mathcal{B} \times [0,t_0]$;

v. la matrice $\Sigma = \int_0^{t_0} v(\beta,t)s^{(0)}(\beta,t)\lambda_0(t)\,dt$ est définie positive.

Théorème 4.1 (Andersen et Gill (1982)) *Sous les hypothèses i-v, $\hat{\beta}$ tend en probabilité vers β quand $n \to \infty$. De plus, $\hat{\beta}$ est asymptotiquement normal, i.e.*

$$n^{1/2}(\hat{\beta} - \beta) \underset{n\to\infty}{\overset{loi}{\longrightarrow}} \mathcal{N}(0,\Sigma^{-1}).$$

Soit de plus

$$\hat{I}_n(\beta) = -\frac{\partial^2}{\partial\beta\partial\beta'} \ln \mathcal{L}_p(\beta)\cdot$$

On peut montrer que $n^{-1}\hat{I}_n(\beta)$ tend en probabilité vers Σ et fournit donc un estimateur naturel de la variance asymptotique de $\hat{\beta}$.

La vraisemblance partielle $\mathcal{L}_p$ permet de construire simplement des tests asymptotiques de $H_0 : \beta = 0$ contre $H_a : \beta \neq 0$, comme une vraisemblance classique. Ainsi, la statistique du score sous l'hypothèse nulle s'écrit

$$\xi_n = \left(\frac{\partial \ln \mathcal{L}_p}{\partial\beta}\right)'_{|\beta=0} \cdot \hat{I}_n(\beta)^{-1}_{|\beta=0} \cdot \left(\frac{\partial \ln \mathcal{L}_p}{\partial\beta}\right)_{|\beta=0}.$$

Sous H_0, ξ_n tend en loi vers un chi-deux à q degrés de liberté (q étant la dimension de β). Voir Kalbfleish et Prentice (1980) pour plus de détails.

4.4 Estimation de la survie et de la fonction de hasard intégrée "de base"

On peut, dans le cadre du modèle de Cox, chercher un estimateur de la survie de base S_0. Nous reprenons l'optique "Maximum de Vraisemblance en dimension infinie" vu en 2.3. La différence provient ici de la présence d'un processus de covariables dépendant de chaque individu et du temps. Pour alléger les notations, nous ne préciserons pas l'argument temporel dans le processus z. Plus précisément, avec les mêmes notations qu'en 2.3, en notant D_i l'ensemble des indices des individus qui décèdent en $X^*_{(i)}$, C_i les indices des individus censurés dans $[X^*_{(i)},X^*_{(i+1)}[$, et R_i l'ensemble des individus à risque à cette date, la vraisemblance approchée s'écrit

$$L_{app}(S_0) = \prod_{i=1}^{k} \left\{ \prod_{l\in D_i} \left[S_0(X^*_{(i)} - \Delta_i)^{\exp(z_l'\beta)} - S_0(X^*_{(i)} + \Delta_i)^{\exp(z_l'\beta)}\right] \cdot \prod_{l_i=1}^{c_i} S_0(X_{i,l_i})^{\exp(z'_{l_i}\beta)} \right\}.$$

Par un raisonnement identique, on montre que, pour maximiser cette quantité sur l'espace des fonctions de survie, toute solution S_0 doit être constante par morceaux, avec des sauts aux instants de "vraies" durées observées $X^*_{(i)}$. On pose alors $S_0(X^*_{(i)}) = \prod_{j=1}^{i} \alpha_j$, $\alpha_j \in]0,1]$ pour tous i et j. Il

reste donc à détermine les constantes α_j. Ces dernières doivent maximiser

$$\begin{aligned}
L_{app}(S_0) &= \prod_{i=1}^{k} \left\{ \prod_{l\in D_i} \left[(1-\alpha_i^{\exp(z_l'\beta)}) \cdot \prod_{j=1}^{i-1} \alpha_j^{\exp(z_l'\beta)} \right] \cdot \left(\prod_{j=1}^{i} \alpha_j \right)^{\sum_{l_j=1}^{c_j} \exp(z_{l_j}'\beta)} \right\} \\
&= \prod_{i=1}^{k} \left\{ \prod_{l\in D_i} (1-\alpha_i^{\exp(z_l'\beta)}) \right\} \cdot \prod_{i=1}^{k} \left\{ \prod_{j=1}^{i} \alpha_j^{\sum_{l\in D_i\cup C_i} \exp(z_l'\beta)} \cdot \alpha_i^{-\sum_{l\in D_i} \exp(z_l'\beta)} \right\} \\
&= \prod_{j=1}^{k} \prod_{l\in D_j} (1-\alpha_i^{\exp(z_l'\beta)}) \cdot \prod_{j=1}^{k} \prod_{i=j}^{k} \alpha_j^{\sum_{l\in D_i\cup C_i} \exp(z_l'\beta)} \cdot \prod_{j=1}^{k} \alpha_i^{-\sum_{l\in D_i} \exp(z_l'\beta)} \\
&= \prod_{j=1}^{k} \left\{ \prod_{l\in D_j} (1-\alpha_j^{\exp(z_l'\beta)}) \cdot \alpha_j^{\sum_{l\in R_j} \exp(z_l'\beta)} \alpha_j^{-\sum_{l\in D_j} \exp(z_l'\beta)} \right\}.
\end{aligned}$$

Si on suppose connu β, α_i est solution de l'équation $\partial/\partial\alpha_i \ln L_{app}(S_0) = 0$, soit

$$\sum_{l\in D_i} \frac{\exp(z_l'\beta)}{1-\alpha_i^{\exp(z_l'\beta)}} = \sum_{l\in R_i} \exp(z_l'\beta). \tag{4-6}$$

En remplaçant β par un estimateur $\hat{\beta}$, on peut trouver numériquement une solution $\hat{\alpha}_i$. Dans le cas particulier où il n'existe pas d'ex-aequo, D_i est réduit à un singleton et l'équation précédente donne, en notant $z_{(i)}$ la covariable relative à l'individu (i) prise à la date $X_{(i)}^*$,

$$\hat{\alpha}_i = \left(1 - \frac{\exp(z_{(i)}'\beta)}{\sum_{l\in R_i} \exp(z_l'\beta)} \right)^{\exp(-z_{(i)}'\beta)}.$$

Dans tous les cas, on estime les survies de base et conditionnelles par

$$\hat{S}_0(t) = \prod_{i|X_{(i)}^*\le t} \hat{\alpha}_i,$$

$$\hat{S}(t|z) = \prod_{i|X_{(i)}^*\le t} \hat{\alpha}_i^{\exp(z'\beta)},$$

et en remplaçant le paramètre β par un estimateur consistent. On note que si $z = 0$ pour chaque individu de l'échantillon, on retrouve l'estimateur de Kaplan-Meier. En effet, 4-6 devient

$$\sum_{l\in D_i} \frac{1}{1-\alpha_i} = \sum_{l\in R_i} 1 = n_i \Longleftrightarrow \alpha_i = (1 - m_i/n_i)$$

.

Par ailleurs, la fonction de hasard intégrée "de base" Λ_0 s'estime généralement par l'estimateur dit "de Breslow"

$$\hat{\Lambda}_0(t) = \sum_{i|X_i\le t} \frac{\delta_i}{\sum_{j\in R_i} \exp(z_j'(X_i)\hat{\beta})}.$$

On peut avoir facilement l'intuition qui a guidé à l'introduction de cet estimateur: reprendre le raisonnement que avons tenu lors de la construction de l'estimateur de Nelson-Aalen (section 2.4). On peut montrer que le processus $n^{1/2}(\hat{\Lambda}_0 - \Lambda_0)$ converge faiblement dans $D([0,\tau],d_0,\mathcal{D})$ vers un processus gaussien centré (Andersen et Gill (1982), Tsiatis (1981), et autres référence dans Andersen et al.(1993)).

`Remarque` **4.3.** *On pourrait également estimer la survie de base par la formule*

$$\hat{S}_0 = \exp(-\hat{\Lambda}_0).$$

Enfin, on peut estimer la fonction de hasard de base λ_0 en lissant un des estimateur de Λ_0 obtenu précédemment. Usuellement, en choisissant un noyau K et une fenêtre h, on utiliserait

$$\hat{\lambda}_0(t) = \int K_h(t-u)\,\hat{\Lambda}_0(dt) = \sum_{i=1}^{n} K_h(t-X_i)\frac{\delta_i}{\sum_{j\in R_i}\exp(z_j'(X_i)\hat{\beta})}.$$

Remarque 4.4. *Le modèle de Cox a fait l'objet de nombreuses extensions, tant univariées que multivariées (voir Sasieni (1993) par exemple). En particulier, il est possible d'estimer le modèle en présence d'un paramètre d'hétérogénéité. Plus précisément, on suppose que des covariables ne sont pas observables, et que leur effet peut se résumer par l'introduction dans le modèle d'une variable aléatoire supplémentaire U, de fonction de répartition F_U, de densité f_U, et indépendante du processus des covariables observables Z. Le modèle de Cox s'écrit alors*

$$\lambda(t|z,u) = \lambda_0(t)\exp(z'\beta + u).$$

Heckman et Singer (1984) indiquent comment estimer simultanément β et la distribution de U lorsque λ_0 correspond à une loi de Weibull. Néanmoins, le cadre très général du maximum de vraisemblance nonparamétrique utilisé par Heckman et Singer (1984) rend leur résultat essentiellement théorique. Par contre, Honoré (1990) a proposé une estimation plus simple de β. De manière alternative, de nombreux auteurs préfèrent spécifier la loi suivie par le paramètre d'hétérogénéité. Ainsi, on peut supposer que $\exp(u)$ *suit une loi gamma, choix effectué de manière standard. Alors, Murphy (1994,1995) a fournit des estimations de β et Λ_0 sous cette dernière hypothèse. Dans le cas de données éventuellement censurées à droite, Horowitz (1999) a fourni des estimateurs nonparamétriques des fonctions inconnues Λ_0, λ_0, F_U et f_U par application des méthodes de lissage par noyaux. Sur les modèles avec hétérogénité, on pourra consulter avec profit le survey dans le chapitre 5 d'Horowitz (1998).*

4.5 Le problème des ex-aequo

En pratique, les données ne peuvent être obtenues en temps réel, ou bien selon un pas de temps aussi petit qu'on le désire. Il est donc courant d'avoir à traiter des cas d'ex-aequo. Par exemple, les durées de chômage sont souvent repérées par mois, d'autres le sont par jours (fiabilité), ou par heures (biologie) etc. Ce phénomène d'agrégation des donnés complique généralement l'analyse. En effet, nous avons spécifié en 4.2 un modèle à hasards proportionnels en temps continu pour lequel la probabilité d'obtenir deux durées identiques est nulle. Constater des ex-aequo indique alors qu'on sort de ce cadre au sens strict, ou bien qu'utiliser directement l'analyse faite précédemment conduira à une mauvaise spécification du modèle. On peut toutefois chercher à adapter les méthodes précédentes.

Reprenons le raisonnement de la vraisemblance partielle. Ici, en gardant les notations du paragraphe 4.2, on observe m durées non censurées $t_{(j)}$, et u_j sera toute l'histoire du processus entre les dates $t_{(j-1)}$ et $t_{(j)}-$, plus le fait qu'on observe d_j décès en $t_{(j)}$. $v_j = (i_1,\ldots,i_{d_j})$ sera le vecteur des indices des d_j individus qui décèdent en $t_{(j)}$. Alors, la vraisemblance partielle de v par rapport à (u,v) est un produit de m termes $L_p^{(j)}$, $j = 1,\ldots,m$ avec

$$\begin{aligned}
L_p^{(j)}(\beta) &= \frac{P\left(i_1,\ldots,i_{d_j} \text{ décèdent en } t_{(j)} | i_l \in R_j, l=1,\ldots,d_j\right)}{\sum_{\{\mathbf{k}=(k_1,\ldots,k_{d_j})|\mathbf{k}\in R_j^{d_j}\}} P(k_1,\ldots,k_{d_j} \text{ décèdent en } t_{(j)}|\mathbf{k}\in R_j^{d_j})} \\
&= \frac{\prod_{p=1}^{d_j} P(i_p \text{décède en } t_{(j)} | i_p \in R_j)}{\sum_{\{A\subset\mathcal{P}(R_j)|\sharp A = d_j\}}\prod_{q\in A} P(q \text{ décède en } t_{(j)}|q\in R_j)} \\
&= \frac{\prod_{p=1}^{d_j}\lambda_0(t_{(j)})\exp(z_p'(t_{(j)})\beta)}{\sum_{\{A\subset\mathcal{P}(R_j)|\sharp A = d_j\}}\prod_{q\in A}\lambda_0(t_{(j)})\exp(z_q'(t_{(j)})\beta)} \\
&= \frac{\prod_{p=1}^{d_j}\exp(z_p'(t_{(j)})\beta)}{\sum_{\{A\subset\mathcal{P}(R_j)|\sharp A = d_j\}}\prod_{q\in A}\exp(z_q'(t_{(j)})\beta)}\,. \qquad (4\text{-}7)
\end{aligned}$$

Ces quantités sont complexes à calculer, particulièrement lorsque le nombre d'ex-aequo d_j est "élevé" (le nombre d'ensembles A admissibles dans le dénominateur précédent est $C_{n_j}^{d_j}$). C'est pourquoi, lorsque d_j est faible, on utilise parfois l'approximation

$$\tilde{L}_p(\beta) = \frac{\prod_{p=1}^{d_j} \exp(z_p'(t_{(j)})\beta)}{[\sum_{k \in R_j} \exp(z_k'(t_{(j)})\beta)]^{d_j}}.$$

Néanmoins, les estimations basées sur la vraisemblance partielle approchée précédente sont en général biaisées, le bais étant d'autant plus grand que les ex-aequo sont nombreux. Dans ce dernier cas, il convient d'estimer directement un modèle discret, qui s'avérera plus adapté au problème, même s'il est moins élégant (voir la section 4.6).

4.6 Modèle de Cox discret

Il est possible de donner une version discrète du modèle de Cox, adaptée notamment au cas de nombreux ex-aequo. On supposera donc que les données sont regroupées en intervalles I_j, tels que $I_j =]a_{j-1}, a_j]$, si $j = 2, \ldots, k$, et en posant arbitrairement $I_1 = [0, a_1]$ et $I_{k+1} =]a_k, +\infty[$. Les durées de décès exactes sont donc inconnues (ou ignorées). Seule est disponible (et jugée pertinente) l'information donnant l'indice de l'intervalle I_j dans lequel l'individu décède ou est censuré.

On suppose qu'un individu censuré dans l'intervalle I_j n'aurait pas pu décéder dans cet intervalle, i.e. que pour cet individu, $T > a_j$ (on aurait pu ne pas utiliser cette convention au prix de légères modifications des équations de vraisemblance, essentiellement changer $P(T > a_j)$ en $P(T > a_{j-1})$). Cette hypothèse est la transcription naturelle dans le modèle discret du principe vu précédemment selon lequel une durée complète a la priorité sur une censure ayant lieu au même moment. Enfin, on suppose que le processus $z(t)$ des covariables est constant dans chaque intervalle I_j et égal à z_j.

L'hypothèse du hasard proportionnel s'écrit ici : pour tout j,

$$P(T \in I_j | T > a_{j-1}, z) = 1 - (1 - \lambda_j)^{\exp(z_i'\beta)},$$

où $1 - \lambda_j = \exp(-\int_{a_{j-1}}^{a_j} \lambda_0)$.

`Exercice 4.2` *Montrer que, si on fait tendre la largeur des intervalles I_j vers 0, on retrouve la spécification du modèle de Cox classique en temps continu.*

En effet, le modèle à hasards proportionnels en temps continu fournit cette dernière caractérisation lorsque les observations sont groupées. En effet,

$$\begin{aligned}
P(T \in I_j | T > a_{j-1}, z) &= \frac{P(T \in I_j | z)}{P(T > a_{j-1} | z)} \\
&= \frac{S(a_{j-1}|z) - S(a_j|z)}{S(a_{j-1}|z)} \\
&= 1 - \exp\left(- \int_{a_{j-1}}^{a_j} \lambda(u|z)\, du \right) \\
&= 1 - \exp\left(- \int_{a_{j-1}}^{a_j} \exp(z'(u)\beta) \lambda_0(u)\, du \right) \\
&= 1 - (1 - \lambda_j)^{exp(z_j'\beta)},
\end{aligned}$$

en posant

$$\lambda_j = 1 - \exp(-\int_{I_j} \lambda_0).$$

On note immédiatement que $\lambda_j = P(T \in I_j | T > a_{j-1}, z = 0)$. λ_j est donc la probabilité de décéder dans l'intervalle I_j pour un individu "de référence" (par définition pour lequel $z = 0$), sachant qu'il n'a pas encore décédé. Cette quantité s'assimile donc à un taux de hasard.

Remarque 4.5. *nous adoptons ici le point de vue de données groupées, sans faire d'hypothèses sur la distribution de T, qui peut être aussi bien continue que comporter des masses non nulles en certains points.*

Le formalisme précédent englobe en fait tous ces cas. Même le cas extrême d'une distribution de T discrète peut se traiter dans un "quasi-modèle à hasards proportionnels", bien que formellement, la fonction de hasard λ n'existe pas.

En effet, on "définit"[12] *alors le hasard de T en x_i sachant z par $P(T = x_i | T \geq x_i, z)$, i.e. par $1 - (1 - \lambda_i)^{c_0(z'\beta)}$.*

Supposons que la distribution de la durée T sachant z consiste en des masses portées par un ensemble au plus dénombrable de points x_i, $i = 1,2,\cdots$, pour tout z (les x_i étant indépendants de z). Alors, la fonction de survie correspondant à $z = 0$ s'écrit

$$S_0(t) = \prod_{i | x_i \leq t} (1 - \lambda_i).$$

L'hypothèse de hasards proportionnels s'exprime alors en termes de survies

$$S(t|z) = S_0(t)^{c_0(z'\beta)} = \prod_{i | x_i \leq t} (1 - \lambda_i)^{c_0(z'\beta)}.$$

La vraisemblance du modèle discret s'écrit donc

$$\mathcal{L} = \prod_{j=1}^{k} \left\{ \prod_{l \in D_j} P(T \in I_j | z_l) \cdot \prod_{l \in C_j} P(T > a_j | z_l) \right\},$$

en notant comme précédemment D_j (resp. C_j) l'ensemble des indices des individus qui décèdent (resp. sont censurés) dans l'intervalle I_j. De plus, R_j sera l'ensemble à risque dans l'intervalle I_j (individus qui ne sont ni décédés, ni censurés dans les intervalles précédents ou, autrement dit, les individus toujours présents dans l'échantillon en a_{j-1}). On note $n_j = \sharp R_j$ et $d_j = \sharp D_j$. Par convention, les individus pour lesquels la durée observée appartient à $I_{k+1} =]a_k, +\infty[$ sont considérés comme censurés, et leur indice appartient donc à C_k. En effet, pour ces individus, on sait uniquement que leur durée T est plus grande que a_k.

Posons $\pi_j = \prod_{l \in D_j} P(T > a_j | T > a_{j-1}, z_l)$. On note que

$$\pi_j = \prod_{l \in D_j} P(T > a_j | T > a_{j-1}, z_l) = \prod_{l \in D_j} (1 - P(T \in I_j | T > a_{j-1}, z_l)) = \prod_{l \in D_j} (1 - \lambda_j)^{\exp(z_l'(a_j)\beta)}.$$

En écrivant que $\pi_j = \prod_{l \in D_j} P(T > a_j | z_l) / P(T > a_{j-1} | z_l)$ et

$$P(T > a_j | z_l) = \prod_{p=1}^{j} (1 - \lambda_p)^{\exp(z_l'\beta)},$$

12. au sens usuel, S n'est pas dérivable en x_i, donc la fonction de hasard en x_i n'existe pas. Néanmoins, dans des espaces de fonctions généralisées, dits également espaces des distributions, cette notion a un sens, et le calcul différentiel classique s'y étend. Le lecteur intéressé pourra avoir un exposé complet de cette théorie dans Schwartz (1973).

on obtient alors

$$\begin{aligned}
\mathcal{L} &= \prod_{j=1}^{k} \left\{ \prod_{l\in D_j} P(T\in I_j|T>a_{j-1},z_l)P(T>a_{j-1}|z_l)\cdot \prod_{l\in C_j} P(T>a_j|z_l) \right\} \\
&= \prod_{j=1}^{k} \left\{ \pi_j^{-1} \prod_{l\in D_j} P(T\in I_j|T>a_{j-1},z_l)\cdot \prod_{l\in C_j\cup D_j} P(T>a_j|z_l) \right\} \\
&= \left\{ \prod_{j=1}^{k} \pi_j^{-1} \prod_{l\in D_j} (1-(1-\lambda_j)^{\exp(z_l'(a_j)\beta)}) \right\} \cdot \left\{ \prod_{p=1}^{k}\prod_{j=p}^{k} \prod_{l\in C_j\cup D_j} (1-\lambda_p)^{\exp(z_l'(a_p)\beta)} \right\} \\
&= \left\{ \prod_{j=1}^{k} \pi_j^{-1} \prod_{l\in D_j} (1-(1-\lambda_j)^{\exp(z_l'(a_j)\beta)}) \right\} \cdot \left\{ \prod_{p=1}^{k} \prod_{l\in R_p} (1-\lambda_p)^{\exp(z_l'(a_p)\beta)} \right\} \\
&= \prod_{j=1}^{k} \left\{ \prod_{l\in D_j} (1-(1-\lambda_j)^{\exp(z_l'(a_j)\beta)}) \cdot \prod_{l\in R_j\setminus D_j} (1-\lambda_j)^{\exp(z_l'(a_j)\beta)} \right\}.
\end{aligned}$$

En effectuant le changement de variable $\gamma_j = \ln(-\ln(1-\lambda_j))$, $j=1,\ldots,k$, on obtient

$$\ln\mathcal{L} = \sum_{j=1}^{k} \left\{ \sum_{l\in D_j} \ln(1-\exp(-\exp(\gamma_j + z_l'(a_j)\beta))) - \sum_{l\in R_j\setminus D_j} \exp(z_l'(a_j)\beta+\gamma_j) \right\}.$$

En dérivant une fois cette identité par rapport aux γ_j, $j=1,\ldots,k$ et aux composantes du vecteur β, on obtient les equations de vraisemblance. Les estimations des paramètres sont alors les racines de ces équations. En général, il n'est pas possible d'en donner une expression explicite. Des procédures numériques dé résolution s'imposent donc. On montre que les racines $(\hat{\gamma},\hat{\beta})$ sont asymptotiquement normales, de matrice de variance-covariance asymptotique estimées par l'inverse de l'information de Fisher $-\partial^2 \ln\mathcal{L}/\partial(\gamma,\beta)\partial(\gamma,\beta)'$ (voir Kalfleish et Prentice (1980) pour plus de détails).

4.7 L'interprétation en termes de vraisemblance marginale

On va montrer que la vraisemblance partielle associée au modèle de Cox peut s'interpréter comme une vraisemblance marginale. On comprend alors mieux pourquoi la vraisemblance partielle se comporte statistiquement comme une véritable vraisemblance.

Plus précisément, plaçons-nous dans le cas le plus simple de durées non censurées. On observe donc un échantillon i.i.d. de n durées distinctes, qu'on ordonne en $t_{(1)} < \cdots < t_{(n)}$. Les covariables sont de plus supposées indépendantes du temps. La statistique des antirangs est le vecteur $r = [(1),(2),\ldots,(n)]$: r_i, la i-ième composante de r, nous donne l'indice de l'individu associé à la i-ième plus petite durée observée $t_{(i)}$. r est la réalisation de la variable aléatoire des antirangs R, assimilable à une permutation de $\{1,\ldots,n\}$.

`Exemple.` **4.2** Si $n=4$, et $t_1=5$, $t_2=17$, $t_3=12$, $t_4=15$. Alors $r=[1,3,4,2]$.

La vraisemblance de r issue des observations est dite vraisemblance marginale car on s'intéresse à la loi d'une fonction des observations et non à la loi des observations elles-mêmes. En notant $z_{(j)}$ le vecteur de covariables associé à la j-ième durée observée, la vraisemblance marginale nous

est donnée par

$$\begin{aligned}\mathcal{L}_m &= P\left(R=[(1),(2),\ldots,(n)]|z_1,\ldots,z_n\right) = \prod_{k=1}^{n} P\left(r_k=(k)|(1),\ldots,(k-1);z_1,\ldots,z_n\right)\\ &= \prod_{k=1}^{n} P\left(T_{(k)}<T_j, j\notin\{(1),(2),\ldots,(k)\}|z_1,\ldots,z_n\right)\\ &= \prod_{k=1}^{n}\int\left(\prod_{j=k+1}^{n} S(u|z_{(j)})\right) f(u|z_{(k)})\,du\\ &= \prod_{k=1}^{n}\int S_0(u)^{\sum_{j=k+1}^{n}\exp(z'_{(j)}\beta)} f(u|z_{(k)})\,du.\end{aligned}$$

Remarquons que $f(u|z_{(k)})$ est la dérivée de $-S(u|z_{(k)})$, ou de $-S_0(u)^{\exp(z'_{(k)}\beta)}$. On intègre alors facilement l'expression précédente, ce qui donne

$$\mathcal{L}_m = \prod_{k=1}^{n}\left(\frac{\sum_{j=k+1}^{n}\exp(z'_{(j)}\beta)}{\exp(z'_{(j)}\beta)}+1\right)^{-1} = \prod_{k=1}^{n}\frac{\exp(z'_{(k)}\beta)}{\sum_{j=k}^{n}\exp(z'_{(j)}\beta)},$$

c'est-à-dire la vraisemblance partielle de Cox.

Si on introduit des censures, le raisonnement se complique, même dans le cas classique (qu'on supposera) de censures indépendantes des durées. En effet, on ne peut ordonner de manière unique toutes les observations, car les données censurées correspondent à des décès pouvant avoir lieu à n'importe quel moment postérieur à la censuré constatée.

`Exemple.` **4.3** Si $n=4$, et que les données sont $114, 90^+, 63, 108^+$. Alors il y 6 vecteurs d'antirangs possibles : [3,2,4,1], [3,4,2,1], [3,2,1,4], [3,4,1,2], [3,1,2,4], [3,1,4,2].

Plus précisément, notons $t_{(1)}<\cdots<t_{(k)}$ les instants de durées complètes. Remarquons que maintenant, les indices (i), $i=1,\ldots,k$, appartiennent à $\{1,\ldots,n\}$. On observe éventuellement des censures dans l'intervalle $]t_{(i)},t_{(i+1)}]$ ($t_{(0)}=0$ et $t_{(k+1)}=+\infty$ par convention), dont les indices forment les ensembles C_i. La durée associée au j-ième individu censuré dans cet intervalle est notée $T_{i,j}$. La vraisemblance marginale de r est donc ici la probabilité que R appartienne à un ensemble de permutations compatibles avec les données, soit

$$\mathcal{L}_m = P(T_{(1)}<\cdots<T_{(k)}, T_{(i)}<T_{i,j}, \forall i=1,\ldots,k+1, j\in C_i|z_1,\ldots,z_n).$$

Notons $\mathcal{H}_i$ l'historique du processus des décès et des censures jusqu'à (i) inclu, et R_i l'ensemble à risque en $t_{(i)}$. L'événement $\mathcal{A}_i=\{T_{(i)}<T_{i,j}, j\in C_i\}$ conditionnellement à $\mathcal{H}_{i-1}$ et $T_{(i)}=u$ a pour probabilité

$$h_i(u)=\prod_{j\in C_i} S(u|z_j) = S_0(u)^{\sum_{j\in C_i}\exp(z'_{i,j}\beta)}.$$

On a alors

$$\begin{aligned}
\mathcal{L}_m &= \prod_{i=1}^{k} P(\text{l'individu } (i) \text{ fournit la } i\text{-ième durée complète}, \mathcal{A}_i | \mathcal{H}_{i-1}) \\
&= \prod_{i=1}^{k} \int P(T_{(i)} < T_l, l \in R_i, \mathcal{A}_i | \mathcal{H}_{i-1}, T_{(i)} = u) f(u|z_{(i)})\, du \\
&= \prod_{i=1}^{k} \int P(u < T_l, l \in R_i | \mathcal{H}_{i-1}, \mathcal{A}_i, T_{(i)} = u).P(\mathcal{A}_i | \mathcal{H}_{i-1}, T_{(i)} = u) f(u|z_{(i)})\, du \\
&= \prod_{i=1}^{k} \int S_0(u)^{\sum_{j \in R_i \setminus C_i} \exp(z'_{(j)}\beta)} h_i(u) f(u|z_{(i)})\, dt_{(i)} \\
&= \prod_{i=1}^{k} \int S_0(u)^{\sum_{j \in R_i} \exp(z'_{(j)}\beta)} f(u|z_{(i)})\, du \\
&= \prod_{i=1}^{k} \frac{\exp(z'_{(i)}\beta)}{\sum_{j \in R_i} \exp(z'_{(j)}\beta)},
\end{aligned}$$

et ainsi, on retrouve également la vraisemblance de Cox.

`Exercice 4.3` *Reprendre le raisonnement précédent lorsqu'il y a des ex-aequo : à l'instant $d_{(i)}$, on observe d_i décès "simultanés". Soit Q_i l'ensemble des permutations sur les indices $(1, \cdots, d_i)$, et $P = (p_1, \ldots, p_{d_i})$ un élément quelconque de Q_i. On appellera R_i l'ensemble à risque en $t_{(i)}$ et $R_i(p_r) = R_i - \{p_1, \cdots, p_{r-1}\}$. Montrer alors que la vraisemblance marginale de r s'écrit*

$$\mathcal{L}_m = \prod_{i=1}^{k} \sum_{P \in Q_i} \prod_{r=1}^{d_i} \left[\sum_{l \in R_i(p_r)} \exp(z'_l \beta) \right]^{-1} \exp(z'_{ir}\beta). \tag{4-8}$$

Retrouver la vraisemblance de Cox classique lorsque $d_i = 1$ pour tout i. On notera que la formule 4-8 diffère de la formule 4-7 obtenue à partie d'un raisonnement direct de vraisemblance partielle. Néanmoins, lorsque le nombre d'ex-aequo n'est pas trop grand, ces deux formules ne sont en général "pas trop éloignées".

Doksum (1987) a étendu l'analyse de la vraisemblance partielle en tant que vraisemblance marginale du vecteur des rangs, pour des modèles de régression du type

$$h(Y) = x'\beta + \varepsilon,$$

où h est une fonction croissante, et ε suit une loi quelconque F. Selon que h est connue, ou que ce soit F, ou bien que les deux sont inconnues (mais alors F est supposée symétrique), l'auteur propose des estimateurs de toutes les quantités inconnues des divers modèles. La méthode décrite concerne des donnés complètes.

5 Autres modèles de régression

Le modèle linéaire et ses extensions constituent la base de la modélisation. On comprend aisément l'intérêt de généraliser leur estimation à des données censurées ou tronquées. Soit donc le modèle

$$Y = Z'\beta + \varepsilon, \tag{5-1}$$

où ε est de loi inconnue, de variance finie. On suppose ε centrée, quitte à rajouter une composante égale à un dans le vecteur des paramètres β. Ici, Y peut représenter une durée, mais également toute transformation d'une durée ; on n'imposera donc pas de contrainte sur la variable expliquée, qui sera une variable aléatoire réelle quelconque. On notera S_0 et f_0 respectivement la fonction de répartition et la densité de ε [13].

En présence de données censurées à droite aléatoirement, au lieu d'un échantillon i.i.d. $(Y_i)_{i=1,\dots,n}$ de réalisations de Y, on observe un échantillon $(X_i,Z_i,\delta_i)_{i=1,\dots,n}$ où $X_i = Y_i \wedge C_i$ et $\delta_i = \mathbf{1}\{Y_i \leq C_i\}$.

Notons que le modèle à temps accéléré est équivalent au modèle de régression classique (5-1), en posant $Y = \ln T$ [14]. C'est le cas plus généralement de modèle du type $H(T) = Z'\beta + \varepsilon$, où H est une fonction que nous avons supposé connue. Cette section aurait donc pu s'intituler "estimation du modèle à temps accéléré".

Par ailleurs, si on connaît la fonction de hasard intégrée "de base" Λ_0, le modèle de Cox est un cas particulier de (5-1) (c.f. les écritures 4-1 et 4-2), pour lequel la loi de ε est connue.

La vraisemblance du triplet (X,Z,δ) est donnée par

$$f_0(X - Z'\beta)^{\delta} S_0(X - Z'\beta)^{1-\delta}.$$

Ainsi, l'équation de vraisemblance s'écrit

$$0 = \sum_{i=1}^{n} Z_i \left\{ \delta_i \frac{f_0'}{f_0}(X_i - Z_i'\beta) - (1-\delta_i)\frac{f_0}{S_0}(X_i - Z_i'\beta) \right\}. \tag{5-2}$$

Notons au passage que si on connaît la distribution suivie par ε, éventuellement à un paramètre de dimension finie près, on peut effectuer l'inférence paramétrique usuelle. Un estimateur de β sera alors une racine de l'équation (5-2).

Supposons plutôt que, dans l'équation précédente, f_0 et S_0 sont inconnues. On peut alors remplacer la quantité inconnue $-f_0'/f_0$ par une fonction "raisonnable" s, par exemple $s(t) = t$ comme dans le modèle linéaire gaussien. Comme $S_0(dt) = -f_0(t)\,dt$, l'équation (5-2) devient

$$0 = \sum_{i=1}^{n} Z_i \left\{ \delta_i s(X_i - Z_i'\beta) - \frac{(1-\delta_i)}{S_0(X_i - Z_i'\beta)} \int_{X_i - Z_i'\beta}^{\infty} s(t) S_0(dt) \right\}. \tag{5-3}$$

Dans cette dernière expression, Buckley et James (1979) ont proposé de remplacer $S_0(t)$ par l'estimateur de Kaplan-Meier $S_n(t)$ basé sur les résidus $\varepsilon_i = X_i - Z_i'\beta$, $i = 1,\dots,n$ [15].

Ainsi, un estimateur de β sera racine de l'équation

$$\Psi(\beta; s) = \sum_{i=1}^{n} Z_i \left\{ \delta_i s(X_i - Z_i'\beta) - \frac{(1-\delta_i)}{S_n(X_i - Z_i'\beta)} \int_{X_i - Z_i'\beta}^{\infty} s(t) S_n(dt) \right\}. \tag{5-4}$$

13. à ne pas confondre avec les fonctions "de base" vues dans le modèle à hasards proportionnels

14. Y est alors censurée par le logarithme de la censure initiale

15. c'est-à-dire, $S_n(t)$ est calculé comme l'estimateur de Kaplan-Meier associé aux observations $((X_i - Z_i'\beta) \wedge (C_i - Z_i'\beta); \delta_i)_{i=1,\dots,n}$.

Des procédures itératives s'avèrent en général nécessaires pour déterminer une racine de cette équation. Voir Ritov (1990) et Lai et al. (1991) pour les propriétés asymptotiques de l'estimateur de β ainsi obtenu.

L'estimateur précédent dépend crucialement du choix de la fonction s. Si cette dernière est "très proche" de la vraie fonction $-f_0'/f_0$, l'estimateur de β sera asymptotiquement proche de l'efficacité. Mais si ce n'est pas le cas...

Remarque 5.1. *Kim et Lai (1999) ont étendu la méthode de Buckley et James à des modèles de régression plus généraux du type*

$$Y = g(Z) + \varepsilon,$$

où g est inconnue ainsi que S_0. g est ici une fonction à p arguments, p étant la taille du vecteur des covariables. Il s'agit donc d'une approche de type "régression nonparamétrique". Ils proposent une méthode de résolution itérative pour estimer simultanément toutes les quantités inconnues. D'autres méthodes à partir de formules donnant directement des estimateurs de β et S_0 sont possibles, mais au prix d'hypothèses un peu plus fortes et liées au choix de paramètres de lissage (typiquement la fenêtre h) : Dabrowska (1987), Fan et Gijbels (1994), entre autres.

Pour résoudre le problème des censures aléatoires dans les modèles de régression, Koul, Susarla et Van Ryzin (1981) ont proposé un estimateur plus explicite de β. Le problème d'estimation vient du fait que l'espérance de la variable latente Y est différente de l'espérance de la variable observée X, à cause des censures. Les auteurs ont eu alors l'idée de transformer X de telle manière que la nouvelle variable observée ait la même espérance que Y. Plus précisément, considérons la variables aléatoire

$$X^* = \frac{\delta X}{G(X)},$$

en notant G la survie de C (supposée connue). Alors, on a

$$E[X^*|z] = E[E[\frac{\mathbf{1}\{Y \leq C\}Y}{G(Y)}|Y,z]|z] = E[Y|z].$$

On peut donc effectuer une régression sur X^* par moindres carrés, pour obtenir une estimation de β. Evidemment, supposer connue la loi de G est une hypothèse forte. Heureusement, cette méthode est valable même si on remplace la survie de la censure G par son estimateur naturel donné par Kaplan-Meier; en effet, C est censuré à droite par Y.

Par exemple, considérons le modèle simple à deux régresseurs

$$Y_i = \beta_0 + \beta_1 Z_i + \varepsilon_i, \ \ i = 1, \ldots, n.$$

Ici, Z_i est univarié. Avec les mêmes notations qu'au dessus, on peut estimer les deux paramètres par les formules

$$\hat{\beta}_0 = \sum_{i=1}^{n} a_{ni} \frac{\delta_i X_i}{\hat{G}(X_i)}, \ \hat{\beta}_1 = \sum_{i=1}^{n} b_{ni} \frac{\delta_i X_i}{\hat{G}(X_i)},$$

$$b_{ni} = \frac{Z_i - \bar{Z}}{\sum_{j=1}^{n}(Z_j - \bar{Z})^2}, \ a_{ni} = n^{-1} - \bar{Z} b_{ni}, \ \bar{Z} = n^{-1} \sum_{j=1}^{n} Z_j,$$

et $\hat{G}$ est l'estimateur de Kaplan-Meier de la survie des censures $(C_i)_{i=1,\ldots,n}$. Les propriétés asymptotiques ont été étudiées par Koul et al. (1981), Srinivasan et Zhou (1994).

Une autre solution a été proposée par Stute (1993,1996), qui se rapproche de l'estimation par moindres carrés traditionnelle. L'estimateur du paramètres β est défini comme

$$\arg\min_{\beta} \sum_{i=1}^{n} w_{i:n} \left(X_{i:n} - Z'_{i:n}\beta\right)^2,$$

en définissant $X_{1:n} \leq \ldots \leq X_{n:n}$ l'échantillon ordonné des durées observées, $\delta_{i:n}$ et $Z_{i:n}$ la censure et la covariable associées à l'observation $X_{i:n}$, et le poids

$$w_{in} = \frac{\delta_{i:n}}{n-i+1} \prod_{j=1}^{i-1} \left[\frac{n-j}{n-j+1}\right]^{\delta_{j:n}}.$$

Des auteurs ont proposé des approches concurrentes : Miller (1976), Leurgans (1987),Tsiatis (1990), Fygenson et Ritov (1994)... ainsi qu'Andersen et al. (1993) (p. 581). Globalement, il semble difficile d'établir une hiérarchie claire entre ces divers estimateurs et leurs variantes.

Zheng (1984) a proposé une méthode originale dite "synthetic data method" qui étend en la généralisant la méthode de Blum-Susarla-Van Ryzin (1981). Plus précisément, supposons qu'on connaisse la survie G de la censure C. Alors, soit

$$X^* = \phi_1(X;G)\delta + \phi_2(X;G)(1-\delta),$$

où ϕ_1 et ϕ_2 sont deux fonctions telles que, presque sûrement,

$$G(Y)\phi_1(Y;G) - \int_{-\infty}^{Y} \phi_2(t;G)\, dG(t) = Y.$$

On vérifie alors que $E[X_i^*|Z_1,\ldots,Z_n] = E[Y_i|Z_1,\ldots,Z_n] = Z_i'\beta$. Alors, a partir du nouvel échantillon $(X_i^*,Z_i)_{i=1,\ldots,n}$, il est possible de résoudre le modèle linéaire par les formules usuelles (Zheng (1987), Lai et al. (1995)). Lorsqu'on ne connaît pas la loi des censures G, la méthode est encore valable à condition de remplacer G par son estimateur de Kaplan-Meier [16].

Par exemple, si on choisit $\phi_1(X;G) = X/G(X)$, et $\phi_2 = 0$, on obtient l'estimateur de Koul et al. (1981) dans le cadre du modèle linéaire sur données censurées. Cela revient alors à pondérer les observations non censurées, et à ramener à zéro les observations censurées. Si on pose

$$\phi_1(X;G) = \int \left(\frac{\mathbf{1}\{X \geq z\}}{G(s)} - \mathbf{1}\{s<0\}\right) ds,$$

on retrouve l'estimateur de Leurgans (1987).

Ces méthodes ont été étendues par Qin et Jing (2000) au cas de modèles partiellement linéaires, c'est-à-dire pour lesquels

$$Y = Z_1'\beta + g(Z_2) + \varepsilon,$$

g étant une fonction inconnue, l'erreur vérifiant $E[\varepsilon|Z_1,Z_2] = 0$. Il est alors possible, sous certaines conditions, d'estimer β et g.

Cette méthode est également valable dans le cadre de la régression nonparamétrique sur données censurées : Zheng (1988). Voir une autre approche par approximations locales dans Fan et Gijbels (1994). Les modèles non paramétriques précédents sont peu praticables avec des tailles de fichiers usuelles, dès que la dimension p de Z dépasse 3 ou 4. En effet, les vitesses de convergence des estimateurs se dégradent en fonction de la dimension de l'espace des régresseurs. C'est le problème dit de la "malédiction de la dimension" ou "curse of dimensionality", bien connu en estimation fonctionnelle (par exemple, la vitese de convergence optimale de l'estimateur à noyau de

16. on considère alors que C est censurée à droite par Y.

la régression est classiquement en $n^{-2/(4+p)}$). C'est pourquoi les modèles additifs constituent une généralisation "raisonnable" du modèle linéaire paramétrique (voir Hastie et Tibshirani (1990)), au sens où ils réalisent un compromis entre des hypothèses paramétriques toujours discutables et la régression nonparamétrique multidimentionnelle peu praticable.

Ils consistent à introduire des fonctions inconnues de $\mathbb{R}$ vers $\mathbb{R}$, qui séparent les effets de chaque composante de X de manière additive :

$$Y = g_1(Z_1) + \cdots + g_p(Z_p) + \varepsilon,$$

les fonctions univariées $g_1, \ldots, g_p$ étant à estimer. Voir une méthode générale d'estimation dans Huffer et McKeague (1991) ou Andersen et al. (1993). Dans le cadre de données censurées à droite et tronquées à gauche, Kim et Lai (1999) ont proposé une procédure itérative d'estimation de ce modèle.

Enfin, tout en restant dans la classe des modèles à direction unique révélatrice, c'est-à-dire pour lesquels $E[Y|Z = z]$ est uniquement fonction de l'index $z'\beta$, on peut atteindre un grand degré de généralité. En effet, il est possible, sur données censurées, d'estimer toutes les quantités inconnues dans le modèle

$$H(T) = Z'\beta + \varepsilon,$$

où H est une fonction strictement croissante inconnue et où ε est de loi inconnue (Gørgens et Horowitz (1999)).

Les auteurs ont cherché à généraliser les très nombreux modèles de régressions existants au cas de données incomplètes. En général, l'extension étudiée en premier est le cas de données censurées, puis le cas de données censurées à droite et tronquées à gauche simultanément. Plus rarement, les auteurs considèrent le cas de données censurées par intervalles. Par exemple, les modèles de régression par quantiles (Koenker et Bassett (1978)), plus robustes (moindre sensibilité à la présence de points aberrants par rapport au modèle linéaire) ont été généralisés au cas de données censurées par Powell (1986). Dans le même esprit, Wang (2000) a étudié dans ce cadre le modèle linéaire avec erreurs de mesure sur les variables. Et ce genre d'extensions constitue une source inépuisable de publications pour les générations présentes et futures de statisticiens.

6 Modèles à risques concurrents

Nous avons jusqu'à présent cherché à estimer la distribution d'une unique durée T. Même lorsque T était censurée par une durée C, nous n'avons pas estimé la distribution de C, considérée comme inintéresssante. Il en est autrement dans les modèles à risques concurrents. Dans cette classe de modèles, chaque individu est soumis à plusieurs risques en même temps, chacun ayant "un intérêt en soi". Le processus d'observation s'interrompt dès qu'un premier type de décès intervient. On n'observe qu'une seule durée par individu, ainsi que la nature de cette durée[17]. Notons que ce qu'on considérait jusqu'à présent comme une censure est désormais vu comme une durée d'intérêt particulière, ou une cause de décès particulière. Nous la traitons en tant que telle[18].

Plus précisément, on observe pour chaque individu une durée X et un indice J prenant les valeurs 1,2,...,p, c'est-à-dire un échantillon i.i.d. $(X_i,J_i)_{i=1,...,n}$. L'indice J indique que le décès est du J-ième type.

`Exemple.` **6.1** Un épisode de chômage peut se terminer de plusieurs manières : l'individu retrouve un nouvel emploi, ou quitte le marché du travail, ou décède, ou déménage et disparait du champ de l'enquête etc. Ces événements sont a priori corrélés. Ainsi, trouver un nouvel emploi suscite souvent un déménagement ; de même, quitter le marché du travail est plus fréquent après une (plus ou moins longue) recherche infructueuse etc.

6.1 Représentation en termes de "fonctions spécifiques"

Les observations ne nous permettent d'estimer que des fonctions des quantités $P(X > u,J = j)$, pour tous u et j. Introduisons la fonction de hasard spécifique aux décès de type j

$$\lambda_j^*(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(X \in [t,t+\Delta t],J = j|X > t).$$

Cette fonction exprime donc la probabilité instantanée de décéder selon la j-ième cause.

La fonction de hasard de X, notée comme d'habitude λ, vaut pour tout t

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(X \in [t,t+\Delta t]|X > t) = \sum_{j=1}^{p} \lambda_j^*(t).$$

La survie de X s'exprime également comme

$$S(t) = P(X > t) = \exp(-\int_0^t \lambda) = \prod_{j=1}^{p} S_j^*(t),$$

$$S_j^*(t) = \exp(-\int_0^t \lambda_j^*),\ j = 1,\ldots,p.$$

Notons que S_j^* ne s'interprète pas en général comme une fonction de survie, car λ_j^* n'est pas la fonction de hasard d'une variable aléatoire déduite "naturellement" du modèle.

Les équivalents des fonctions de densité pour des décès de type j sont

$$f_j^*(t) = \lim_{\Delta \to 0} \frac{1}{\Delta t} P(X \in [t,t+\Delta t],J = j).$$

17. c'est différent dans les modèles multivariés : on dispose alors pour chaque individu d'un vecteur de durées.

18. il pourrait néanmoins être utile de distinguer une censure des autres causes de décès, notamment lorsque certaines causes de décès sont vues comme "parasites" et ne méritent pas une estimation statistique. Cela nécessite donc de diviser les causes de décès en deux groupes : les pertinentes, a piori corrélées, sur lesquelles portera l'inférence statistique, et les autres, en général supposées indépendantes des premières, et qui en perturbent l'analyse.

On note là aussi immédiatement que f_j^* n'est pas une fonction de densité (elle n'est pas d'intégrale 1) et que, pour tout t et j,

$$f_j^*(t) = \lambda_j^*(t)S(t) = \lambda_j^*(t)\exp\left(-\sum_{k=1}^{p}\int_0^t \lambda_k^*\right).$$

Enfin, la fonction d'incidence cumulée, équivalente ici à une fonction de répartition, est définie par

$$I_j^*(t) = P(X < t,J = j) = \int_0^t f_j^* = \int_0^t \lambda_j^* S.$$

Notons que I_j^* n'est pas un vraie fonction de répartition car $\lim_{t\infty} I_j^*(t) < 1$. De plus, en général, $S_j^*(t) \neq P(X > t,J = j)$, $f_j^* \neq \lambda_j^* S_j^*$ et $dS_j^*(t)/dt \neq -f_j^*(t)$.

La connaissance des fonctions $(\lambda_j^*)_{j=1,\ldots,p}$ résume l'information disponible au vu des données. En effet, cette dernière se réduit à la connaissance des quantités I_j^*, pour tout ensemble t et tout j (ces dernières fonctions peuvent être déduites à partir de la loi des observations). Inversement, la connaissance des I_j^* permet de retrouver les fonctions S, f_j^*, et donc λ_j^*.

Il est alors évident que la vraisemblance va s'exprimer en fonction des quantités λ_j^*. Plus précisément, on a

$$\begin{aligned}
\mathcal{L} &\propto \prod_{i=1}^{n}\prod_{j=1}^{p} P(X \in [X_i,X_i + \Delta t],J_i = j)^{\mathbf{1}\{J_i=j\}} \\
&\propto \prod_{i=1}^{n}\prod_{j=1}^{p}[\lambda_j^*(X_i)S(X_i)]^{\mathbf{1}\{J_i=j\}} \\
&\propto \prod_{i=1}^{n}\lambda_{J_i}^*(X_i)S(X_i) \\
&\propto \prod_{i=1}^{n}\left\{\lambda_{J_i}^*(X_i)\prod_{j=1}^{p}\exp(-\int_0^{X_i}\lambda_j^*)\right\}.
\end{aligned}$$

Si on se concentre sur les décès de type j, i.e. sur l'estimation des fonctions f_j^* ou λ_j^*, on observe que la partie de la vraisemblance utile est alors

$$\mathcal{L}_j \propto \prod_{i=1}^{n}\left[\lambda_j^*(X_i)^{\mathbf{1}\{J_i=j\}}S_j^*(X_i)\right]. \tag{6-1}$$

Cette "vraisemblance" est identique à celle provenant de réalisations i.i.d. d'une unique durée de fonction de hasard λ_j^*, et censurée à droite par tous les décès de type k, $k \neq j$. Formellement, on peut donc reprendre l'analyse par maximum de vraisemblance vue en 2.3 par exemple. Ce fait est a priori surprenant : alors que les causes de décès sont corrélées, on peut effectuer l'analyse statistique comme si elles étaient "indépendantes les unes des autres". La contradiction n'est qu'apparente. En effet, les fonctions λ_j^* qu'on cherche à estimer ne sont pas des fonctions de hasard associées à des durées.

En n'écrivant que la partie de la vraisemblance informative sur les décès de type j, il apparaît une grande similitude avec l'équation (1-6). Un raisonnement identique à celui vu en 2.3 nous fournit donc un estimateur "de type Kaplan-Meier" spécifique au décès de type j,

$$\hat{S}_j^*(t) = \prod_{i|X_{(i)}\leq t}\left(1 - \frac{\mathbf{1}\{J_{(i)} = j\}}{n - i + 1}\right),$$

en ordonnant les observations X_i et en notant $J_{(i)}$ l'indice associé à $X_{(i)}$.

On peut aussi réécrire

$$\hat{S}_j^*(t) = \prod_{i=1,\dots,q|X_{ji}^*\leq t} \left(1 - \frac{d_{ji}}{n_{ji}}\right),$$

en notant $X_{j1}^* < \cdots < X_{jq}^*$ les instants d'observations de décès de type j, d_{ji} la nombre de décès de type j à l'instant X_{ji}^* et n_{ji} la taille de l'ensemble à risque à cette date.

$\hat{S}_j^*(t)$ estime de manière consistante $S_j^*(t)$, mais pas $P(T_j > t)$ ni $P(X > t,J = j)$. Néanmoins, la survie de X est évidemment estimée par

$$\hat{S}(t) = \prod_{j=1}^{p} \hat{S}_j^*(t),$$

et la fonction d'incidence cumulée par

$$\hat{I}_j(t) = \sum_{i|X_{ji}^*\leq t} \frac{d_{ji}}{n_{ji}} \hat{S}(X_{ji}^*).$$

Pour justifier cette dernière formule, on peut remarquer que

$$\begin{aligned}\hat{I}_j(t) &= \int_0^t S(u)\lambda_j^*(u)\,du = \int_0^t S(u)\Lambda_j^*(du) \\ &\sim \int_0^t \hat{S}(u)\hat{\Lambda}_j^*(du),\end{aligned}$$

en estimant les fonctions de hasard spécifiques Λ_j^* par les équivalents de Nelson-Aalen ie

$$\hat{\Lambda}_j^*(t) = \sum_{i=1}^{n} \frac{\mathbf{1}\{J_i = j,X_i \leq t\}}{\sum_{k=1}^{n} \mathbf{1}\{X_k \geq X_i\}}.$$

On peut obtenir des estimateurs des fonctions de hasards spécifiques λ_j^* en lissant les estimateurs précédents $\hat{\Lambda}_j^*$, de la même manière qu'en 3.3.

On peut étendre naturellement dans ce cadre les divers modèles de régressions sur données censurées à droite, en remplaçant la fonction de hasard "classique" par la ou les fonctions λ_j^* spécifiques, la survie de la durée d'intérêt classique par la ou les S_j^* etc. Néanmoins, il importe de bien comprendre que les hypothèses de type paramétriques ou semi-paramétriques doivent être faites sur les fonctions spécifiques, et non sur les distributions de variables aléatoires sous-jacentes.

Par exemple, le modèle à hasards proportionnels s'écrirait ici

$$\lambda_j^*(t) = \lambda_{0j}^*(t)\exp(z'\beta_j),$$

pour tout t et j dans J [19]. La méthode de vraisemblance partielle permet d'estimer les paramètres. En supposant qu'il n'y a pas d'ex aequo, il s'agit de maximiser la fonction de vraisemblance partielle

$$\mathcal{L}_p(\beta_1,\dots,\beta_p) = \prod_{i=1}^{n}\prod_{j=1}^{p} \left(\frac{\exp(z'(X_i)\beta_j)}{\sum_{k\in R(X_i)}\exp(z'(X_k)\beta_j)}\right)^{\mathbf{1}\{J_i=j\}},$$

où $R(X_i)$ indique l'ensemble à risque à la date X_i. On voit immédiatement que $\mathcal{L}_p$ est le produit de p vraisemblances partielles et que l'inférence sur β_j peut se mener indépendamment de la valeur des autres paramètres.

19. ou même dans un sous-ensemble d'indice de $\{1,\dots,p\}$, en considérant que les autres risques sont des censures perturbatrices indépendantes des premières

6.2 Représentation en termes de variables latentes

Certains auteurs trouvent pertinent de modéliser les modèles à risques concurrents sous forme de modèles à variables latentes. Plus précisément, ils supposent qu'on dispose de p durées sous-jacentes $T_1,\ldots,T_p$, a priori corrélées, de lois inconnues. On n'observe que le minimum X de ces p durées, et l'indice J indiquant quelle durée est observée. Le problème fondamental consiste à estimer la loi conjointe de $(T_1,\ldots,T_p)$, ou même les lois marginales de T_j, $j=1,\ldots,p$, à partir des observations (X,J).

On définit la loi jointe des p durées latentes par $Q(t_1,\ldots,t_p)=P(T_1>t_1,\ldots,T_p>t_p)$, pour tout p-uplet $(t_1,\ldots,t_p)$. Soit λ_j la fonction de hasard (marginale) de T_j. On note ∂_j la dérivée partielle par rapport à la j-ième composante, et (a_j,b_{-j}) un vecteur dont la j-ième composante vaut a_j, et dont les autres composantes valent le réel b. On a évidemment

$$\lambda_j(t)=\lim_{\Delta t\to 0}\frac{1}{\Delta t}P(T_j\in[t,t+\Delta t]|T_j>t)=\frac{(-1)}{Q(t,0_{-j})}\partial_j Q_{|t_j=t,t_{-j}=0}.$$

Notons que la survie de T_j au point t est simplement $S_j(t)=Q(t,0_{-j})$. On peut montrer que

$$\lambda_j^*(t)=\lim_{\Delta t\to 0}\frac{1}{\Delta t}P(T_j\in[t,t+\Delta t],T_k>T_j,\forall k\neq j|X>t)=\frac{(-1)}{Q(t,\ldots,t)}\partial_j Q_{|t_1=\cdots=t_p=t}. \qquad (6\text{-}2)$$

En effet,

$$\begin{aligned}
&P(T_j\in[t,t+\Delta t],T_k>T_j,\forall k\neq j|X>t)\\
&\quad=\frac{(-1)^p}{Q(t,\ldots,t)}\int \mathbf{1}\{t_k>t_j,\forall k\neq j\}.\mathbf{1}\{t_j\in[t,t+\Delta t]\}\,Q(dt_1,\ldots,dt_p)\\
&\quad=\frac{(-1)}{Q(t,\ldots,t)}\int \partial_j Q(t_j,\ldots,t_j).\mathbf{1}\{t_j\in[t,t+\Delta t]\}\,dt_j.
\end{aligned}$$

Il est important de noter que hasards marginaux λ_j et hasards spécifiques λ_j^* diffèrent en général. Par contre, ils sont identiques lorsque les durées latentes T_j sont mutuellement indépendantes. En effet, dans ce dernier cas, $Q(t_1,\ldots,t_p)=\prod_{j=1}^p S_j(t_j)$ et

$$\lambda_j(t)=\lambda_j^*(t)=-\frac{S_j'(t)}{S_j(t)},$$

$$S_j^*(t)=S_j(t).$$

Par contre, $f_j^*(t)<f_j(t)$ en général.

6.3 Identifiabilité dans les modèles à risques concurrents

Dans le cadre du modèle latent précédent, on est confronté immédiatement à un grâve problème d'identifiabilité. On a en effet le résultat suivant:

Théorème 6.1 (Tsiatis (1975)) *Soit une famille de fonctions spécifiques $(S_j^*)_{j=1,\ldots,p}$ (ou $(\lambda_j^*)_j$) fournie par un modèle à risques concurrents quelconque. Alors, il existe un modèle à risques latents (ou variables latentes) indépendants qui possède les mêmes fonctions S_j^* (ou, de manière équivalente, les mêmes λ_j^*). Il est défini par*

$$\tilde{Q}(t_1,\ldots,t_p)=\prod_{j=1}^p \tilde{S}_j(t_j),\qquad \tilde{S}_j(t)=\exp(-\int_0^t \lambda_j^*).$$

En effet, pour le modèle indépendant décrit au-dessus, la fonction de hasard spécifique aux décès de type j s'écrit

$$\tilde{\lambda}_j(t) = \frac{(-1)}{\tilde{Q}(t,\ldots,t)} \partial_j \tilde{Q}_{|t_1=\cdots=t_p=t} = -\frac{\tilde{S}_j'(t_j)}{\tilde{S}_j(t_j)} = \lambda_j^*(t).$$

Les quantités observables sont donc les mêmes dans les deux modèles, ce qui les rend indistinguables. Il existe donc toujours un modèle à risques concurrents indépendants qui génère une distribution (X,J) donnée.

Certains auteurs considèrent que la modélisation en termes de variables latentes n'est pas satisfaisante. Supposer que la loi d'une variable latente quelconque serait identique en présence ou non des autres causes de décès leur semble irréaliste. Au delà du postulat, les graves problèmes d'identifiabilité rencontrés dans le cadre latent fournissent un argument important à leur position. Mieux vaudrait alors se contenter du modèle imposé par les données, qui seul permettrait une identification de quantités pertinentes (fonctions de hasard spécifiques, fonctions d'incidence cumulée etc), plutôt que d'imposer des conditions rigoureuses et non justifiables en vue d'estimer des lois marginales d'interprétation dangereuse (voir les termes du débat dans Kalbfleish et Prentice (1980) ou Crowder (1994)).

Ce point de vue est sans doute excessif. Lorsque les conditions d'expérience ne peuvent être exactement dupliquées (par exemple, lorsqu'une cause de décès n'a plus lieu d'être), il semble légitime de chercher à prévoir le nouveau phénomène à partir d'observations tirées d'une expérience passée, même différente. C'est le cas notamment en économétrie, où les modèles à variables latentes sont d'utilisation courante, et où les expériences contrôlées sont impossibles. L'objectif principal de tels modèles est justement de déterminer la loi des variable latentes (non observées). Cette information a donc paru précieuse et pertinente aux auteurs, contredisant ceux qui ne veulent que s'en tenir aux quantités observables (ou spécifiques). Ainsi, on peut juger plus pertinent de modéliser un phénomène en supposant que les hasards marginaux λ_j vérifient un modèle de Cox, et non les fonctions de hasard spécifiques λ_j^*. Enfin et surtout, dans un cadre paramétrique, il est très souvent possible d'identifier (ie d'estimer) les λ_j comme d'ailleurs les λ_j^*. Il faut simplement le vérifier au moment de l'écriture du modèle.

Pour résumer, en pratique, deux attitudes sont possibles.

- Soit on pose un modèle sur les descripteurs des observations (X,J), c'est-à-dire sur les fonctions spécifiques. L'inconvénient est qu'on n'est pas capable de modéliser une expérience qui serait même légèrement modifiée, par exemple en supprimant une issue parmi les p possibles.
- Soit on considère un modèle à variables latentes ; sans hypothèse particulière (par exemple de nature paramétrique), on doit supposer en général que les risques sont indépendants mutuellement.

Malgré le résultat relativement décourageant du théorème 6.1, il existe néanmoins un résultat d'identifiabilité lorsque le modèle est muni de covariables dont les valeurs décrivent tout l'espace des vecteurs de durées. Nous allons détailler ce résultat.

Conditionnellement à un vecteur de covariables z, on peut souvent écrire la distribution multivariée sous la forme

$$Q(t_1,\ldots,t_p|z) \equiv P(T_1 > t_1,\ldots,T_p > t_p|z) = H\left(\exp\left(-\xi_1(z)\Phi_1(t_1)\right),\ldots,\exp\left(-\xi_p(z)\Phi_p(t_p)\right)\right), \tag{6-3}$$

où H est une fonction de distribution sur $[0,1]^p$, et $\lim_{t\to 0}\Phi_j(t) = 0$ pour tout $j = 1,\ldots,p$. Cette réécriture du modèle provient naturellement du modèle de Cox et de ses extensions multivariées.

`Exemple.` **6.2** Supposons que nous sommes en présence d'un modèle à risques concurrents de dimension 2. Chacune des durées latentes suit un modèle à hasards proportionnels. La dépendance

entre durée s'exprime uniquement par l'intermédiaire d'une variable explicative inobservable, dite composante d'hétérogénéité, notée ω et qui suit une distribution G. Il s'agit du formalisme des dit des "frailty models" (Clayton et Cuzyck (1985)). Ainsi,

$$\lambda_1(t|z_1) = \lambda_{01}(t)\exp(z'\beta_1 + c_1\omega),$$

$$\lambda_2(t|z_2) = \lambda_{02}(t)\exp(z'\beta_2 + c_2\omega).$$

Alors, la survie bivariée du modèle s'écrit

$$\begin{aligned} Q(t_1,t_2|z) &= E_\omega[P(T_1 > t_1,T_2 > t_2|\omega)] = E_\omega[\exp(-\Lambda_1(t_1|z,\omega))\exp(-\Lambda_2(t_2|z,\omega))] \\ &= \int \exp(-\Phi_1(t_1)\exp(z'\beta_1 + c_1\omega))\exp(-\Phi_2(t_2)\exp(z'\beta_2 + c_2\omega))\,dG(\omega), \end{aligned}$$

et s'exprime bien comme en (6-3), en posant

$$H(u_1,\ldots,u_p) = \int u_1^{\exp(c_1\omega)}u_2^{\exp(c_2\omega)}\,dG(\omega),$$

$$\xi_j(z) = \exp(z'\beta_j).$$

Théorème 6.2 (Heckman et Honoré (1989)) *Avec les notations de 6-3, si*

i. H est continuement différentiable, de dérivées partielles H_j, H est strictement croissante en chacun de ses arguments, et pour toute suite $(\eta_n)_{n\geq 1}$ de $[0,1]^p$ telle que $\eta_{jn} \to 1$ pour tout j, $H_j(\eta_n)$, $j = 1,\ldots,p$, possède une limite finie quand $n \to \infty$,

ii. pour tout $j = 1,\ldots,p$, $\Phi_j(1) = 1$ et $\xi_j(z_0) = 1$ pour un point fixe z_0 dans le support de Z,

iii. le support de $(\xi_1(z),\ldots,\xi_p(z))$ décrit $]0,+\infty[^p$ lorsque z varie,

iv. pour tout j, Φ_j est strictement croissante, positive, dérivable,

alors, les fonctions H, ξ_j et Φ_j sont identifiables.

Ainsi, les quantités λ_j^* (ou de manière équivalente S_j^* ou I_j^*) peuvent déterminer les fonction H, ξ_j et Φ_j, et donc déterminer $Q(\cdot|z)$ pour tout z. L'hypothèse forte est ici que les valeurs prises par les fonctions ξ_j des covariables doivent permettre de décrire tout l'espace des vecteurs de durées sous-jacentes, i.e. tout $]0,+\infty[^p$. L'estimation pratique de tels modèles sous les hypothèses du théorème 6.2 n'a toutefois pas encore été étudiée par les auteurs, à part dans Fermanian (2003).

Néanmoins, en pratique, le plupart des auteurs préfèrent postuler simplement un modèle paramétrique, lorsqu'ils souhaitent rester dans l'approche par variables latentes. Il s'agit alors de trouver une "bonne" famille paramétrique pour la loi multivariée des durées $(T_1,\ldots,T_p)$: suffisamment riche pour décrire le phénomène, mais sans trop de paramètres pour éviter les problèmes numériques. Ensuite, il suffit d'effectuer une estimation par maximum de vraisemblance.

Pour trouver des familles paramètriques de fonctions de répartitions Q, il est souvent commode de raisonner en termes de copules: exhiber des distributions marginales paramétriques F_{θ_j} pour chaque durées T_j, puis choisir une famille paramétrique de copules C_ν, $\nu \in \Theta$. Enfin, faire l'hypothèse que

$$P(T_1 \leq t_1,\ldots,T_p \leq t_p) = C_\nu(F_{\theta_1},\ldots,F_{\theta_p}),$$

pour un ν, et des $\theta_1,\ldots,\theta_p$. Ainsi, en combinant des familles de copules[20] et des familles de distributions univariées, on génère un grand nombre de distributions multivariées adaptées à des vecteurs de durées.

20. On en trouvera un grand nombre dans Joe (1997) et Nelsen (2000).

7 L'approche par processus ponctuels

La plupart des modèles de durées peuvent s'intégrer dans un formalisme plus général, développé depuis la fin des années 70 et au début des années 80. Les auteurs ont en effet remarqué que les statistiques d'intérêt et les équations de vraisemblance prennent le plus souvent la forme d'intégrales stochastiques par rapport à des processus ponctuels, c'est-à-dire des processus dont les trajectoires sont continues par morceaux et qui "comptent les événements" aux diverses dates. Or tout processus ponctuel peut s'exprimer comme somme d'un processus "quasiment" déterministe (on dit prédictible) et d'une martingale. C'est ce qu'on appelle la décomposition de Doob-Meyer. Il est alors possible d'utiliser des résultats théoriques généraux concernant la convergence des martingales, pour obtenir "directement" des résultats asymptotiques.

Cette méthode, développée initialement par l'école scandinave (Aalen, Gill, Andersen...) a rencontré un grand succès de part son esthétisme et son cadre formel très général, même si c'est au prix d'une complexité mathématique accrue. Nous allons dresser à grand traits les points importants de la théorie. Pour plus de détails, il conviendra de se référer à des ouvrages généraux (Andersen et al. (1992), Harrington et Fleming (1991)) ou aux papiers fondateurs (Aalen (1978), Gill (1980), Andersen et Gill (1982)). Certains termes utilisés dans cette section sont définis en annexe 8.3.

Définition 7.1 *Un processus ponctuel est un processus aléatoire à temps continu* $\{N(t),t \geq 0\}$ *adapté à une filtration* $\{\mathcal{F}_t,t \geq 0\}$ *avec* $N(0) = 0$, $N(t) < \infty$ *presque sûrement, et tel que presque toutes ses trajectoires soient continues à droite, constantes par morceaux et possédant des sauts de discontinuités égaux à* $+1$.

Ainsi, un processus ponctuel "compte" des événements se déroulant dans le temps. Le plus souvent, nous utiliserons la filtration "naturelle" associée au processus : pour tout t, $\mathcal{F}_t = \sigma\{N(u),0 \leq u \leq t\}$, ou cette dernière σ-algèbre en y intégrant d'autres éléments de l'histoire jusqu'en t, c'est-à-dire $\mathcal{F}_t = \sigma\{N(u),\tilde{N}(u),0 \leq u \leq t\}$.

Le processus de Poisson constitue un exemple classique de processus ponctuel. Rappelons qu'un processus de Poisson $N = (N_t)_{t\geq 0}$ compte un nombre d'événements. Il est tel que

$$P(N_t - N_s = k) = \exp(-\lambda)\frac{\lambda^k}{k!}$$

pour tout entier k, pour tous réels s,t tels que $s \leq t$, et pour une certaine constante positive λ. De plus, la variable aléatoire $N_t - N_s$, $t \geq s$, est indépendante de tout N_u avec $u \leq s$.

Si f est une fonction de $\mathbb{R}$ vers $\mathbb{R}$, éventuellement aléatoire, la quantité $\int_s^t f(u)\,dN(u)$ ou $\int_s^t f\,dN$ est une variable aléatoire, qui associe à toute réalisation ω l'intégrale de Lebesgue-Stieltjes de f par rapport à la trajectoire (à variation bornée) $N(\cdot,\omega)$.

En fait, la grande majorité des situations traitées dans les modèles de durée peuvent se formaliser de manière agréable par des processus ponctuels. Nous allons voir quelques exemples dans la suite.

`Exemple.` **7.1** Soit le cas classique d'une durée continue T censurée aléatoirement à droite par une censure C. Supposons pour simplifier que C est indépendante de T. Comme d'habitude, on observe uniquement les couples (X,δ), avec $X = \inf(T,C)$ et $\delta = \mathbf{1}\{T \leq C\}$. Soit un échantillon i.i.d. d'observations $(X_i,\delta_i)_{i=1,\ldots n}$. L'estimateur naturel de la fonction de hasard intégrée (estimateur de Nelson-Aalen) s'écrit au point t

$$\hat{\Lambda}_n(t) = \sum_{i=1}^{n} \frac{\delta_{(i)}\mathbf{1}\{X_{(i)} \leq t\}}{n-i+1}.$$

Notons, pour toute observation $i = 1,\ldots,n$,

$$N_i(t) = \mathbf{1}\{X_i \leq t,\delta_i = 1\},\ Y_i(t) = \mathbf{1}\{X_i \geq t\},\ N(t) = \sum_{i=1}^{n} N_i(t),\ Y(t) = \sum_{i=1}^{n} Y_i(t).$$

On remarque immédiatement que toutes ces quantités sont des processus ponctuels. La variable aléatoire $Y(t)$ indique le nombre d'individus à risque à la date t. $N(t)$ est le nombre de durées complètes observées jusqu'en t. De plus, l'estimateur de Nelson-Aalen s'écrit pour tout t

$$\hat{\Lambda}_n(t) = \int_0^t \frac{\mathbf{1}\{Y(u) > 0\}}{Y(u)}\, N(du).$$

Soit λ la fonction de hasard de T. Lorsque n devient grand, la limite de $\hat{\Lambda}_n$ doit être la fonction de hasard intégrée, du moins si t est plus petit que la dernière durée observée, soit

$$\Lambda^*(t) = \int_0^t \mathbf{1}\{Y(u) > 0\}\, \Lambda(du) = \int_0^{t \wedge (Max_i X_i)} \lambda.$$

Alors, on peut écrire

$$\hat{\Lambda}_n(t) - \Lambda^*(t) = \int_0^t \frac{\mathbf{1}\{Y(u) > 0\}}{Y(u)}\, M(du),$$

avec

$$M(t) = \sum_{i=1}^{n} M_i(t),\ M_i(t) = N_i(t) - \int_0^t Y_i\, d\Lambda \equiv N_i(t) - A_i(t).$$

Ce processus M joue en rôle central dans l'analyse, car c'est une martingale adaptée à la filtration naturelle $\mathcal{F}_t = \sigma\{(X_i,\delta_i)\mathbf{1}\{X_i \leq t\},i = 1,\ldots,n\}$. On peut déjà remarquer que M est d'espérance nulle car

$$\begin{aligned} E[N_i(t)] = P(X \leq t,\delta = 1) = P(T \leq t,T \leq C) = -\int_0^t G(u-)\, dS(u) \\ = \int_0^t P(X \geq u)\lambda(u)\, du = E[\int_0^t Y_i(u)\lambda(u)\, du] = E[A_i(t)]. \end{aligned}$$

Par ailleurs, de manière heuristique, on peut prouver que λ représente la probabilité instantanée de saut du processus ponctuel N. En effet, du fait de l'indépendance entre T et la censure U, on peut écrire

$$\begin{aligned} \lambda(t)\Delta t &\sim P(T \in [t,t+\Delta t])/P(T \geq t) \\ &\sim P(T \in [t,t+\Delta t]|T > t,U > t) \\ &\sim P(N(t+\Delta t) - N(t) = 1|T > t,U > t) \\ &\sim E[N(t+\Delta t) - N(t)|X > t]. \end{aligned} \tag{7-1}$$

Raisonnons sur une seule observation ($n = 1$). Notons $dN(s)$ la variable aléatoire $N(s) - N(s-ds)$, pour ds "petit". Elle suit une loi de Bernoulli, et sa probabilité instantanée de valoir 1 conditionnellement au passé (i.e. conditionnellement à $\mathcal{F}_{s-}$) est $\mathbf{1}\{X \geq s\}\lambda(s)\, ds$. Donc, d'après (7-1), $E[dN(s)|\mathcal{F}_{s-}] = \mathbf{1}\{X \geq s\}\lambda(s)\, ds = dA(s)$. Ceci est équivalent à $E[dM(s)|\mathcal{F}_{s-}] = 0$. Cette relation caractérise une martingale (cf annexe 8.3).

Donc, on espère que M sera "petit" en moyenne, et que $\int_0^t \mathbf{1}\{Y > 0\}/Y\, dM$ sera également "petite", pour que $\hat{\Lambda}_n(t) - \Lambda^*(t)$ soit également "petit". Dans ce but, il faut faire appel aux théorèmes de convergence des martingales.

`Exemple.` **7.2** Les modèles à risques concurrents peuvent être modélisés par des processus ponctuels. En effet, supposons qu'on dispose de p durées latentes $T_1,\ldots,T_p$. On n'observe que $X = \min_{1,\ldots,p} T_j$ et J l'indice de la durée correspondant à ce minimum. Alors l'information fournie par un échantillon i.i.d. de taille n se résume par le processus ponctuel multidimensionnel $N = (N_1,\ldots,N_p)$ où chaque composante de ce p-vecteur est elle-même un processus ponctuel: pour tout $j = 1,\ldots,p$,

$$N_j(t) = \sum_{i=1}^{n} \mathbf{1}\{X_i \leq t, J_i = j\}.$$

Plusieurs résultats théoriques concernent le cadre de l'exemple fondamental 7.1.

Soient T une durée et C une censure (pas forcément indépendante de T). Posons $X = T \wedge C$, $\delta = \mathbf{1}\{T \leq C\}$, $N(t) = \mathbf{1}\{X \leq t, \delta = 1\}$, $N^C(t) = \mathbf{1}\{X \leq t, \delta = 0\}$, et $\mathcal{F}_t = \sigma\{N(u), N^C(u), u \leq t\}$. Alors on a

Théorème 7.1 *Le processus M défini par*

$$M(t) = N(t) - \int_0^t \mathbf{1}\{X \geq u\}\, d\Lambda(u)$$

est une martingale si et seulement si, pour tout u tel que $P(X > u) > 0$,

$$d\Lambda(u) = -\frac{dP(T \geq u, C \geq T\}}{P(X \geq u)}. \tag{7-2}$$

Notons que si T et C sont indépendants, (7-2) est évidemment vérifiée. Par contre, l'inverse est fausse. Alors que le membre de gauche de l'équation (7-2) représente la fonction de hasard de T, le membre de droite représente la fonction de hasard "spécifique" à la durée T, en reprenant le vocabulaire des modèles à risques concurrents (voir la discussion du paragraphe 6.2).

Dans l'exemple précédent, on a décomposé le processus ponctuel N en la somme d'une martingale M et d'un processus A défini par $A(t) = \int_0^t \mathbf{1}\{X \geq u\}\, d\Lambda(u)$. On verra que la valeur de A en t est entièrement prévisible (sans aléa) à partir des valeurs passées $A(s)$, $s < t$. Plus précisément, soit un espace probabilisé muni d'une filtration $\mathcal{F}$,

Définition 7.2 *La σ-algèbre sur $[0,\infty[\times\Omega$ engendrée par les ensembles du type $\{0\} \times A_0$, $A_0 \in \mathcal{F}_0$, et $]a,b] \times A$, $0 \leq a < b < \infty$, $A \in \mathcal{F}_a$, est appelée la σ-algèbre prédictible pour la filtration $\mathcal{F}$.*

Un processus X est dit prédictible pour la filtration $\mathcal{F}$ si, en temps qu'application de $[0,\infty[\times\Omega$ vers $\mathbb{R}$, il est mesurable par rapport à la σ-algèbre prédictible pour $\mathcal{F}$. On dit également que X est un processus $\mathcal{F}$-prédictible.

Les processus prédictibles les plus simples sont de la forme

$$X = k_0 \mathbf{1}_{\{0\}\times A_0} + \sum_{i=1}^{n} k_i \mathbf{1}_{]a_i,b_i]\times A_i},$$

où $A_0 \in \mathcal{F}_0$, $A_i \in \mathcal{F}_{a_i}$ pour tout $i = 1,\ldots,n$ et où les k_i sont des constantes. On note que ces derniers processus possèdent des trajectoires continues à gauche. On a plus généralement

Proposition 7.1 *Tout processus X à valeurs réelles adapté à $\mathcal{F}$ et continu à gauche est $\mathcal{F}$-prédictible. Un processus est prédictible si et seulement s'il est mesurable par rapport à la plus petite σ-algèbre engendrée par les processus adaptés et continus à gauche.*

Tous les processus prédictibles ne sont pas continus à gauche. Ainsi en est-il du processus fondamental A défini dans l'exemple 7.1.

Proposition 7.2 *Soient T une durée, C une censure et $X = T \wedge C$. Alors le processus continu à droite*

$$A(t) = \int_0^t \mathbf{1}\{X \geq u\}\, d\Lambda(u)$$

est prédictible pour la filtration naturelle $\mathcal{F} = \{\mathcal{F}_t, t \geq 0\}$, $\mathcal{F}_t = \sigma\{(X_i, \delta_i)\mathbf{1}\{X_i \leq u\}, u \leq t\}$.

Par ailleurs, tous les processus ne sont pas prédictibles.

`Exercice 7.1` *Soit une martingale continue à droite possédant une probabilité non nulle de "sauter" au point t_0. Montrer alors que ce processus n'est pas prédictible.*

En fait , un processus prédictible est intuitivement un processus dont la valeur en t découle des valeurs observées avant t. Plus précisément,

Proposition 7.3 *Soit X un processus $\mathcal{F}$-prédictible. Pour tout t, $X(t)$ est $\mathcal{F}_{t-}$ mesurable.*

La notion de processus prédictible nous permet d'énoncer le résultat théorique important suivant.

Théorème 7.2 (Décomposition de Doob-Meyer) *Soit X une sous-martingale positive continue à droite adaptée à la filtration $\{\mathcal{F}_t, t \geq 0\}$. Il existe une $\mathcal{F}$-martingale continue à droite M et un processus prédictible croissant continu à droite A tel que $E[A(t)] < \infty$, presque sûrement,*

$$X(t) = M(t) + A(t),$$

pour tout $t \geq 0$. Si $A(0) = 0$ presque sûrement, et si $X = M' + A'$ est une autre décomposition avec $A'(0) = 0$, alors pour tout $t \geq 0$,

$$P\{M'(t) \neq M(t)\} = P\{A'(t) \neq A(t)\} = 0.$$

Le processus prédictible A est appelé compensateur de X. Comme un processus ponctuel est une sous-martingale positive, on déduit du théorème précédent

Corollaire 7.1 *Soit $N = \{N(t), t \geq 0\}$ un processus ponctuel adapté à une filtration $\{\mathcal{F}_t, t \geq 0\}$ avec $E[N(t)] < \infty$ pour tout t. Alors, il existe un unique processus continu à droite croissant A tel que $A(0) = 0$ presque sûrement, $E[A(t)] < \infty$ pour tout t, et $M = N - A$ est une $\mathcal{F}$-martingale continue à droite.*

L'idée de base de la méthode consiste donc à diviser l'information initiale, c'est-à-dire le processus ponctuel décrivant le déroulement de l'expérience, en une composante "régulière" A, facile à traiter car en général proche de ce qu'on cherche à obtenir, et une composante M qui s'apparente à un résidu. La structure de martingale de M permet de controller cette seconde composante, par l'emploi notamment de théorèmes asymptotiques généraux.

Il existe souvent un processus $\lambda = \{\lambda(t), t \geq 0\}$ tel que le compensateur A du processus ponctuel N puisse s'écrire pour tout t

$$A(t) = \int_0^t \lambda.$$

On appelle λ l'intensité du processus N[21].

21. l'analogie de notation avec la fonction de hasard n'est pas fortuite. Elle trouve son origine dans le cas classique traité en exemple 7.1

Les processus ponctuels rencontrés en pratique ne vérifient pas toujours la condition $E[N(t)] < \infty$ pour tout t. Cette dernière n'est parfois vérifiée que pour des processus tronqués par les valeurs prises par une suite de temps d'arrêts, suite qui tend vers $+\infty$ presque sûrement (procédé de localisation). On peut en fait généraliser la décomposition de Doob-Meyer à des sous-martingales locales (cf annexe 8.3). Ainsi, on obtient le résultat suivant :

Théorème 7.3 (Décomposition de Doob-Meyer étendue) *Soit X une sous-martingale locale positive continue à droite adaptée à la filtration $\{\mathcal{F}_t, t \geq 0\}$ et soit $(\tau_n)_n$ la suite de temps d'arrêts associée. Il existe un unique processus prédictible croissant continu à droite A tel que $A(0) = 0$ presque sûrement, $P(A(t) < \infty) = 1$ pour tout t et $X - A$ est une martingale locale continue à droite. Pour tout t, $A(t)$ est la limite presque sûre de la suite $A_n(t) \equiv X(t \wedge \tau_n)$.*

Corollaire 7.2 *Soit $N = \{N(t), t \geq 0\}$ un processus ponctuel adapté à une filtration $\{\mathcal{F}_t, t \geq 0\}$. Alors, il existe un unique processus continu à droite croissant A tel que $A(0) = 0$ presque sûrement, $A(t) < \infty$ presque sûrement pour tout t, et $M = N - A$ est une $\mathcal{F}$-martingale locale. De plus, A est localement borné et M est localement de carré intégrable. Enfin, presque sûrement pour tout t, $A(t) = \lim_{s \to t, s<t} A(s) \leq 1$.*

En fait, beaucoup de statistiques d'intérêt ne s'écrivent pas immédiatement comme des martingales, mais plutôt sous la forme $\int H\, dM$, avec M une $\mathcal{F}$-martingale et H un processus $\mathcal{F}$-prédictible (ce dernier peut être assez complexe).

`Exemple.` **7.3** Soit le modèle de Cox classique avec un processus de covariables Z dépendant du temps. La dérivée de la log-vraisemblance partielle associée au modèle de Cox classique (ou score) s'écrit avec les notations classiques

$$\begin{aligned} U(\beta) &= \sum_{i=1}^{n} \delta_i \left\{ z_i(t_i) - \frac{\sum_{j \in R_i} z_j(t_i) \exp(z_j' \beta)}{\sum_{j \in R_i} \exp(z_j' \beta)} \right\} \\ &= \sum_{i=1}^{n} \int_0^{\infty} H_i(u, \beta, z_1, \ldots, z_n)\, dN_i(u), \end{aligned}$$

en posant pour tout i, $N_i(t) = \mathbf{1}\{X_i \leq t, \delta_i = 1\}$, $Y_i(t) = \mathbf{1}\{X_i \geq t\}$ et

$$H_i(u) = z_i(u) - \frac{\sum_{j=1}^{n} z_j(u) Y_j(u) \exp(z_j'(u)\beta)}{\sum_{j=1}^{n} Y_j(u) \exp(z_j'(u)\beta)}.$$

Soit $M_i(u) = N_i(u) - \int_0^u Y_i(s) \lambda_0(s) \exp(z_i'(s)\beta)\, ds$. $(M_i(u))_{u \geq 0}$ est une martingale par rapport à la filtration $\mathcal{F} = \{\mathcal{F}_t, t \geq 0\}$, $\mathcal{F}_t = \sigma\{(X_i(u), \delta_i, z_i(u)) \mathbf{1}\{u \leq t\}, u \leq t\}$, d'après la proposition 7-2. On voit aisément (le montrer) que

$$U(\beta) = \sum_{i=1}^{n} \int_0^{\infty} H_i(u, \beta, z_1, \ldots, z_n)\, dM_i(u).$$

En fait, les quantités $\int H\, dM$ sont bien souvent elles-mêmes des martingales, en particulier de par le résultat suivant :

Théorème 7.4 *Soit N un processus ponctuel par rapport à la filtration $\mathcal{F}$ tel que $E[N(t)] < \infty$ pour tout t. Soit $M = N - A$ une $\mathcal{F}$-martingale, où A est un processus $\mathcal{F}$-prédictible tel que $A(0) = 0$, et soit H un processus $\mathcal{F}$-prédictible. Alors le processus L donné par*

$$L(t) = \int_0^t H(u)\, dM(u)$$

est une $\mathcal{F}$-martingale.

Comme pour la décomposition de Doob-Meyer, il est possible d'étendre ce résultat à des processus locaux.

Théorème 7.5 *Soit N un processus ponctuel par rapport à la filtration $\mathcal{F}$. Soit $M = N - A$ la $\mathcal{F}$-martingale locale de carré intégrable définie dans corollaire 7.2, où A est un processus $\mathcal{F}$-prédictible tel que $A(0) = 0$, et soit H un processus $\mathcal{F}$-prédictible. Alors le processus L donné par*

$$L(t) = \int_0^t H(u)\, dM(u)$$

est une $\mathcal{F}$-martingale locale de carré intégrable.

Il est légitime de se demander comment calculer les compensateurs. Dans les cas simples, il est possible de "deviner" sa forme A, et de montrer ensuite que $N - A$ est une martingale. On conclut alors que A est bien le compensateur de N par unicité de la décomposition de Doob-Meyer. C'est ainsi qu'on a procédé pour le processus apparaissant dans l'exemple 7.1.

Par ailleurs, il est souvent utile de travailler en norme L^2 pour établir des résultats théoriques (penser simplement au calcul de la variance des estimateurs, ou à l'inégalité de Bienaymé-Tchebichev pour estimer les écarts en probabilité). C'est pourquoi il est important de considérer les décomposition de Doob-Meyer associé aux carrés de martingales. Plus précisément, comme le carré d'une martingale est une sous-martingale, le théorème 7.2 donne

Théorème 7.6 *Soit M une $\mathcal{F}$-martingale (resp. locale) continue à droite, telle que $E[M^2(t)] < \infty$ pour tout t. Alors il existe un unique processus continu à droite prédictible noté $< M,M >$ tel que $< M,M > (0) = 0$ presque sûrement, $E[< M,M > (t)] < \infty$ pour tout t et $\{M^2(t) - < M,M > (t), t \geq 0\}$ est une martingale (resp. locale) continue à droite.*

On appelle $< M,M >$ le processus de variation quadratique prédictible de M. On montre sans peine que si M est une martingale (éventuellement locale) de carré intégrable telle que $M(0) = 0$ presque sûrement, alors pour tout t, $E[M^2(t)] = E[< M,M > (t)]$.

Par ailleurs, on peut relier $< M,M >$ et A lorsqu'on a la décomposition de Doob-Meyer $M = N - A$. Comme $\bar{M} = M^2 - < M,M >$ est une martingale, on sait que, de manière heuristique, $E[d\bar{M}(t)|\mathcal{F}_{t-}] = 0$. Alors $E[dM^2(t)|\mathcal{F}_{t-}] = E[d < M,M > (t)|\mathcal{F}_{t-}] = d < M,M > (t)$. Comme

$$dM^2(t) = M^2(t) - M^2(t - dt) = (M(t) + M(t - dt))dM(t) = (dM(t))^2 + 2M(t - dt)dM(t),$$

et $E[dM(t)|\mathcal{F}_{t-}] = 0$. On en déduit

$$d < M,M > (t) = E[\{dM(t)\}^2|\mathcal{F}_{t-}] = Var(dM(t)|\mathcal{F}_{t-}) = Var(dN(t) - dA(t)|\mathcal{F}_{t-}).$$

Or $E[dN(t)|\mathcal{F}_{t-}] = E[dA(t)|\mathcal{F}_{t-}] = dA(t)$. Donc, N valant 0 ou 1 dans l'intervalle infinitésimal $]t - dt, t]$, on a

$$\begin{aligned} d < M,M > (t) &= E[\{dN(t)\}^2|\mathcal{F}_{t-}] - 2E[dN(t)|\mathcal{F}_{t-}]dA(t) + \{dA(t)\}^2 \\ &= E[dN(t)|\mathcal{F}_{t-}] - \{dA(t)\}^2 = dA(t)\{1 - dA(t)\} \sim dA(t). \end{aligned}$$

On intuite donc que $< M,M >$ est égal à A. C'est bien le cas.

Théorème 7.7 *Soit N un processus ponctuel tel que $E[N(t)] < \infty$ pour tout $t \geq 0$, et soit A son compensateur. Si toutes les trajectoires de A sont continues et si $EM^2(t) < \infty$ pour tout t, avec $M = N - A$, alors $M^2 - A$ est une martingale continue à droite, i.e.*

$$< M,M >= A.$$

Le résultat reste vrai sans l'hypothèse $EM^2(t) < \infty$ si A est continu. Retenons donc la relation $EM^2(t) = EA(t) = E < M,M > (t)$, valable sous des hypothèses de régularité.

On peut également définir le processus de variation quadratique associé à deux martingales définies sur le même espace probabilisé par la relation

$$< M_1,M_2 >= \frac{1}{2}(< M_1 + M_2 > - < M_1,M_1 > - < M_2,M_2 >).$$

Des règles de calcul simples des processus de variation quadratique existent alors. Elles permettent de calculer des covariances. En particulier

Théorème 7.8 *Si pour tout $i = 1,2$, H_i est un processus prédictible borné, N_i est un processus ponctuel borné, et si la martingale $M_i = N_i - A_i$ vérfie $E[M_i^2(t)] < \infty$ pour tout t, où A_i est l'unique compensateur de N_i, alors*

$$< \int H_1\, dM_1, \int H_2\, dM_2 >= \int H_1 H_2 d < M_1,M_2 > .$$

Revenons au modèle d'un échantillon i.i.d. de durées T éventuellement censurées à droite par C. Les statistiques

$$U = \sum_{i=1}^{n} \int H_i\, d(N_i - A_i), \tag{7-3}$$

sont fréquentes (cf. exemple 7.3), avec H_i un processus $\mathcal{F}$-prédictible borné et $A(t) = \int_0^t \lambda$. On déduit du théorème précédent la propriété suivante, utile en pratique.

Proposition 7.4 *Le processus U défini en (7-3) est une $\mathcal{F}$-martingale, $E[U(t)] = 0$ et*

$$Var(U(t)) = \sum_{i=1}^{n} \int_0^t E[H_i^2(u)\mathbf{1}\{X_i \geq u\}]\lambda(u)\, du,$$

pour tout $t \geq 0$.

Par exemple, on déduit de ce résultat que le score associé à la vraisemblance partielle du modèle de Cox (cf. exemple 7.3) est une martingale. En la multipliant par le facteur $n^{1/2}$, et en invoquant un théorème central limite pour les martingales (théorème de Rebolledo : voir annexe), on en déduit alors directement la normalité asymptotique de $\hat{\beta}$, maximisateur de la vraisemblance partielle : voir Andersen et Gill (1982) ou Gill (1984).

`Exemple.` **7.4** Reprenons l'estimateur de Nelson-Aalen vu dans l'exemple 7.1. Si on cherche à prouver la convergence uniforme en probabilité de $\hat{\Lambda}_n(s)$, $0 \leq s \leq t$, il suffit de montrer que $\sup_{s\in[0,t]} \mid \int_0^s \mathbf{1}\{Y(u) > 0\}Y^{-1}(u)\, dM(u) \mid$ tend vers 0 en probabilité. Or, en utilisant l'inégalité de Lenglart (voir annexe 8.3), on obtient la propriété suivante :

Proposition 7.5 *Soit N un processus ponctuel et $M = N - A$ la martingale locale de carré intégrable associée. Si H est un processus adapté continu à gauche possédant des limites à droite (ou plus généralement prédictible et localement borné), alors, pour tout temps d'arrêt T tel que $P(T < \infty) = 1$, et tous $\varepsilon,\eta > 0$,*

$$P\left(\sup_{t\leq T}\left\{\int_0^t H(s)\, dM(s)\right\}^2 \geq \varepsilon\right) \leq \frac{\eta}{\varepsilon} + P\left(\int_0^T H^2(s) < M,M > (ds) \geq \eta\right).$$

On en déduit que

$$P\left(\sup_{s\in[0,t]}\left[\int_0^s \frac{\mathbf{1}\{Y(u)>0\}}{Y(u)}\,dM(u)\right]^2 > \varepsilon\right) \leq \frac{\eta}{\varepsilon} + P\left(\int_0^t \frac{\mathbf{1}\{Y(u)>0\}}{Y^2(u)}\,dA(u) \geq \eta\right)$$
$$\leq \quad \frac{\eta}{\varepsilon} + P\left(\frac{\Lambda(t)}{Y(t)} \geq \eta\right),$$

en majorant $Y(u)$ par $Y(t)$. Si $Y(t) \to +\infty$ en probabilité lorsque n tend vers $+\infty$, pour tout ε, on peut alors choisir η tel que le dernier membre soit arbitrairement petit, dès que n est assez grand. Ainsi est prouvée la convergence faible uniforme de $\hat{\Lambda}_n$.

Ce type de méthodologie comporte indéniablement des avantages: généralité des résultats, hypothèses de validité exprimées en termes de propriétés vérifiées par les processus sous-jacents (notamment le processus des covariables), usage d'outils mathématiques puissants... Néanmoins, l'approche par processus ponctuels perd de sa puissance lorsqu'il s'agit d'établir des résultats de convergence "fins", notamment des vitesses de convergence. Dans ce cas, des raisonnement plus classiques basés sur la théorie des processus empiriques sont souvent nécessaires.

8 Annexes

8.1 L'intégrale de Lebesgue-Stieltjes

Rappelons les principax résultats concernant l'intégrale de Lebesgue-Stieltjes.

Soit F une fonction de $\mathbb{R}$ vers $\mathbb{R}$, à variations bornées. De plus, on supposera qu'elle est continue en droite et admet une limite finie à gauche en tout point t (on dit alors qu'elle est "cadlag"). On déduit de F une mesure μ_0 sur l'ensemble des réunions finies d'intervalles de $\mathbb{R}$, par l'intermédiaire des définitions suivantes : pour tout $a \leq b$,

$$\mu_0(]a,b]) = F(b) - F(a), \qquad \mu_0(]a,b[) = F(b-) - F(a),$$

$$\mu_0([a,b]) = F(b) - F(a-), \qquad \mu_0([a,b[) = F(b-) - F(a-).$$

On peut alors étendre μ_0 en une mesure μ définie sur la σ-algèbre (ou tribu) engendrée par les intervalles, c'est-à-dire à l'ensemble des boréliens de $\mathbb{R}$ (théorème de Carathéodory : voir Shiryayev (1984), par exemple).

L'intégrale de Lebesgue-Stieltjes par rapport à F d'une fonction $\phi : \mathbb{R} \to \mathbb{R}$ mesurable est définie par $\int \phi \, d\mu$. On la note $\int \phi \, dF$, ou $\int \phi(t) \, F(dt)$.

Comme pour l'intégrale de Riemann ou de lebesgue classiques, on notera $\int_a^b \phi \, dF = \int \mathbf{1}\{t \in]a,b]\}\phi(t) \, F(dt)$. De plus, on notera

$$\int_{a-}^{b} \phi \, dF = \int \mathbf{1}\{t \in [a,b]\}\phi(t) \, F(dt) = \lim_{x \to a, x<a} \int_x^b \phi \, dF.$$

L'intégrale de Lebesgue-Stieltjes possède les propriétés usuelles des intégrales, i.e. pour toutes fonctions F et G cadlag,

$$\int_a^b F(t-) \, G(dt) + \int_a^b G(t) \, F(dt) = FG(b) - FG(a),$$

$$\int_a^b \phi \, dF + \int_b^c \phi \, dF = \int_a^c \phi \, dF,$$

$$d(F(x)G(y)) = F(dx)G(dy).$$

Si ϕ est continue sur $]a,b]$, elle est intégrable au sens de Lebesgue-Stieltjes.

Si F est une "fonction de sauts", i.e. si elle s'écrit $F(t) = \sum_{n|x_n \leq t} h_n = \sum_n h_n \mathbf{1}\{x_n \leq t\}$ pour une certaine suite au plus dénombrable de point $(x_n)_{n \leq 1}$, alors $\int \phi \, dF = \sum_n \phi(x_n) h_n$. En particulier,

$$\int \phi(t) \, d\left(\mathbf{1}\{t \leq t_0\}\right) = \phi(t_0).$$

`Exemple.` **8.1** Si S est la fonction de survie empirique d'un échantillon $(X_1, \ldots, X_n)$, on a pour toute fonction mesurable ϕ

$$n^{-1} \sum_{i=1}^{n} \phi(X_i) = -\int \phi \, dS_n = -\int \phi(t) S_n(dt) = E_{P_n}[\phi],$$

où P_n est la loi de répartition empirique.

Si F est absolument continue continue par rapport à la mesure de Lebesgue, elle possède une dérivée (dite de Radon-Nicodym) par rapport à cette mesure, qui s'identifie à la dérivée usuelle de F lorsque celle-ci existe. Alors pour toute fonction ϕ mesurable, $\int \phi\, dF = \int \phi(t) F'(t)\, dt$.

On utilise couramment ce type d'intégrales car le fonctions de répartition et de survie sont continues à droite et possèdent des limites à gauche en tout point. Les quantités qui interviennent dans les problèmes statistiques s'expriment par ailleurs très souvent comme des intégrales (de Lebesgue-Stietjes) d'une certaine fonction ϕ par rapport à une fonction de répartition ou de survie (penser aux estimateurs de Kaplan-Meier ou de Nelson-Aalen).

Remarque **8.1.** *La théorie de l'intégrale de Stieltjes en dimension supérieure ou égale à 2, bien que d'esprit similaire, est plus délicate à présenter. On se référera à Fermanian (1996).*

8.2 Convergence faible des processus à temps continu

La théorie générale de la convergence faible des processus ou, ce qui revient au même, de la convergence faible des suites de probabilités est exposée dans Billingsley (1968) et Pollard (1984). Nous allons en rappeler les principes essentiels.

Nous présenterons d'abord la théorie valable dans des espaces métriques généraux, puis ses applications aux espaces qui nous concernent :

- l'espace des fonctions continues sur $[a,b]$, noté $C([a,b])$,
- l'espace des fonctions continues à droite admettant des limites à gauche en tout point de $[a,b]$ (fonctions "cadlag"), noté $D([a,b])$.

Ces deux espaces peuvent être munis de diverses métriques et de diverses tribus. Une difficulté du sujet provient du fait qu'il existe plusieurs choix possibles pour la métrique et la tribu. De plus, il ne sera pas toujours pertinent de travailler avec la tribu borelienne, c'est-à-dire la tribu engendrée par les ouverts (approche de Pollard : cf infra).

Soit donc S un espace métrique, muni de la métrique ρ, d'une tribu $\mathcal{A}$ et d'une probabilité P. $\mathcal{A}$ sera inclue dans la tribu borélienne de (S,ρ). $(P_n)_{n\geq 1}$ désignera une suite de probabilités sur $(S,\mathcal{A})$. On notera $\mathcal{C}(S,\mathcal{A})$ l'espace des fonctions de $(S,\rho,\mathcal{A})$ vers $(\mathbb{R},\mathcal{B}(\mathcal{R}))$ continues, bornées et $\mathcal{A}/\mathcal{B}(\mathbb{R})$ **mesurables**. Si $\mathcal{A}$ est la tribu borélienne de (S,ρ), la continuité de f assure sa mesurabilité.

Définition 8.1 *La suite de probabilités $(P_n)_{n\geq 1}$ sur $\mathcal{A}$ converge faiblement vers P si pour toute fonction f de $\mathcal{C}(S,\mathcal{A})$,*

$$\int_S f\, dP_n \underset{n\to\infty}{\longrightarrow} \int_S f\, dP.$$

On note $P_n \underset{n\to\infty}{\Longrightarrow} P$.

Si $(P_n)_{n\geq 1}$ converge faiblement, sa limite est bien déterminée (il y a unicité). En effet, deux mesures de probabilité P et Q sur $\mathcal{A}$ coïncident si et seulement si, pour toute fonction f dans $\mathcal{C}(S,\mathcal{A})$, $\int_S f\, dP = \int_S f\, dQ$.

Remarque **8.2.** *Soit $(F_n)_{n\geq 1}$ et F des fonctions de répartition de variables aléatoires à valeurs dans $\mathbb{R}^m$. On définit $P_n(t) = \int_{-\infty}^t dF_n(u)$ et $P(t) = \int_{-\infty}^t dF(u)$. Alors, on peut établir la propriété suivante :*

$$\left(P_n \underset{n\to\infty}{\Longrightarrow} P\right) \Longleftrightarrow \left(\lim_{n\to\infty} F_n(x) = F(x) \text{ en tout point } x \text{ de continuité de } F\right).$$

Par la suite, une fonction $X : (\Omega,\mathcal{E},P) \longrightarrow (S,\mathcal{A})$ mesurable est dite élément aléatoire et l'on définit P_X, la distribution de X par $P_X(A) = P(X^{-1}(A))$, pour tout $A \in \mathcal{A}$. Il est équivalent

d'étudier la convergence faible des mesures et la convergence en distribution des éléments aléatoires (appelés variables aléatoires lorsque S est de dimension finie).

Définition 8.2 ***(Convergence en distribution)***. *On dit que la suite d'éléments aléatoires $(X_n)_{n\geq 1}$ converge en distribution (ou en loi) vers X lorsque $(P_{X_n})_{n\geq 1}$ converge faiblement vers P_X. Ainsi, on notera*

$$X_n \underset{n\to\infty}{\Longrightarrow} X \text{ si et seulement si } P_{X_n} \underset{n\to\infty}{\Longrightarrow} P_X.$$

Ici, S est en général un espace fonctionnel et f est le plus souvent une "fonction de fonctions". Lorsque S est de dimension finie, on peut se ramener à $S = \mathbb{R}^m$, et on retrouve le concept classique de convergence en loi : si $(X_n)_{n\geq 1}$ est une suite de variables aléatoires sur $\mathbb{R}^m$, si P_n est la probabilité induite par X_n (i.e. $P_n = P_{X_n}$) alors

$$\begin{aligned} X_n \xrightarrow[n\to\infty]{loi} X &\iff \forall f : \mathbb{R}^m \to \mathbb{R} \text{ continue bornée, } E[f(X_n)] \underset{n\to\infty}{\longrightarrow} E[f(X)] \\ &\iff \left(P_{X_n} \underset{n\to\infty}{\Longrightarrow} P_X\right) \end{aligned}$$

Soient $(X_n)_{n\geq 1}$ et X des éléments aléatoires de $(S,\rho,\mathcal{A})$. Supposons que $X_n \underset{n\to\infty}{\Longrightarrow} X$.

Soit $h : (S,\rho,\mathcal{A}) \longrightarrow (S',\rho',\mathcal{A}')$ une application mesurable, continue en tout point d'un ensemble $\mathcal{A}$-mesurable $C_0 \subset S$ tel que $P(X \in C_0) = 1$. Est-ce que la convergence faible se conserve par transformation continue? Autrement dit, si on pose $Y_n = h(X_n)$ et $Y = h(X)$, a-t-on $Y_n \underset{n\to\infty}{\Longrightarrow} Y$?

Si $\mathcal{A}$ est la tribu borélienne de (S,ρ), c'est vrai (continuous mapping theorem). Dans le cas général, c'est vrai sous des conditions un peu plus strictes : il faut que C_0 soit consitué de points dits "réguliers" (voir Pollard (1984), p.67).

Il est nécessaire de fournir des méthodes plus explicites permettant de prouver des résultats de convergence faible. Tout d'abord, on a

Théorème 8.1 *Soient $(P_n)_{n\geq 1}$ et P des probabilités sur $(S,\mathcal{A})$. $\left(P_n \underset{n\to\infty}{\Longrightarrow} P\right)$ si et seulement si toute sous-suite $(P_{n'})$ de (P_n) contient une sous-suite $(P_{n''})$ telle que $\left(P_{n''} \underset{n\to\infty}{\Longrightarrow} P\right)$.*

Définition 8.3 ***(Relative compacité)***. *Soit Π une famille de probabilités sur $(S,\mathcal{A})$. On dit que Π est relativement compact si toute suite d'éléments de Π contient une sous-suite faiblement convergente. Autrement dit si, pour toute suite $(P_n)_{n\geq 1}$ d'éléments de Π, il existe $(P_{n'})$ une sous-suite de $(P_n)_{n\geq 1}$ et une mesure Q telles que $P_{n'} \underset{n\to\infty}{\Longrightarrow} Q$.*

Notons qu'il n'y a pas équivalence entre la convergence faible de P_n vers P et le fait que $(P_n)_{n\geq 1}$ est relativement compacte ; en effet, la limite précédente Q peut ne pas être de masse un.

Soit S un espace de fonctions de $I \subset \mathbb{R}$ vers $\mathbb{R}$. Les projections de dimensions finies sont définies pour tout $k \in \mathbb{N}$ et tout k-uplets $(t_1,...,t_k) \in I^k$ par $\pi_{t_1,...,t_k} : S \longrightarrow \mathbb{R}^k$, $\pi(x) = (x(t_1),...,x(t_k))$.

En général, $\left(\forall k \in \mathbb{N}, \forall (t_1,...,t_k) \in I^k, P_n\pi^{-1}_{t_1,...,t_k} \underset{n\to\infty}{\Longrightarrow} P\pi^{-1}_{t_1,...,t_k}\right)$ n'implique pas $\left(P_n \underset{n\to\infty}{\Longrightarrow} P\right)$. Mais supposons de plus que Π est relativement compact. Alors toute sous-suite $(P_{n'})$ de $(P_n)_{n\geq 1}$ contient une sous-suite $(P_{n''})$ faiblement convergente: $P_{n''} \underset{n\to\infty}{\Longrightarrow} Q$. D'où, si les projections de dimension finies sont continues, pour tout $k \in \mathbb{N}$ et tout $(t_1,...,t_k) \in I^k$,

$$P_{n''}\pi^{-1}_{t_1,...,t_k} \underset{n\to\infty}{\Longrightarrow} Q\pi^{-1}_{t_1,...,t_k}.$$

Alors pour tout $k \in \mathbb{N}$ et tout $(t_1,...,t_k) \in I^k$, $P\pi^{-1}_{t_1,...,t_k} = Q\pi^{-1}_{t_1,...,t_k}$. Comme les lois de probabilités sur $\mathcal{A}$ sont déterminées par les lois marginales de dimensions finies, $P = Q$. Ainsi, toute sous-suite de $(P_n)_{n\geq 1}$ contient une sous-suite convergeant faiblement vers P. Donc, d'après le théorème précédent, $P_n \underset{n\to\infty}{\Longrightarrow} P$.

En résumé, la convergence de chacune des distributions de dimension finie vers une certaine probabilité et la relative compacité de la suite $(P_n)_{n\geq 1}$ assurent la convergence faible de $(P_n)_{n\geq 1}$

vers une probabilité P. La difficulté principale réside dans le fait qu'il faut trouver des critères de relative compacité.

Définition 8.4 ***(Famille de probabilités tendue)***. *Soit* Π *une famille de probabilités sur* $(S,\mathcal{A})$. Π *est dite tendue (ou "tight") si pour tout* $\varepsilon > 0$, *il existe* $K_\varepsilon \subset S$ *compact tel que, pour tout* $P \in \Pi$, $P(K_\varepsilon) > 1 - \varepsilon$.

Théorème 8.2 ***(Prohorov)***. *Si* Π *est tendue , alors elle est relativement compacte. Réciproquement, lorsque* S *est séparable et complet, si* Π *est relativement compacte, alors elle est tendue.*

Le théorème de Prohorov nous incite à raisonner dans des espaces séparables et complets. En pratique, la convergence des distributions de dimension finie est aisée à prouver. Il suffit alors, dans de tels espaces, de prouver que la suite de probabilités est tendue.

8.2.1 Application à l'espace des fonctions continues sur [0,1]

Considérons $S \equiv C([0,1]) = C$ l'espace des fonctions continues sur [0,1], muni de la métrique induite par la norme uniforme

$$\rho(x,y) = \|x - y\|_\infty = \sup_{t\in[0,1]} |x(t) - y(t)| .$$

Notons $\mathcal{C}$ l'ensemble des boréliens de $(C, \|.\|_\infty)$. Il est facile de voir que $(C, \|.\|_\infty)$ est séparable et complet.

Théorème 8.3 *Soit* $(P_n)_{n\geq 1}$ *et* P *des probabilités sur* $(C, \|.\|_\infty, \mathcal{C})$. *Si* $(P_n)_{n\geq 1}$ *est tendue et si les distributions de dimensions finies de* $(P_n)_{n\geq 1}$ *convergent faiblement vers celles de* P *(i.e. si pour tout* $k \in \mathbb{N}$, *pour tout k-uplet* $(t_1,...,t_k) \in [0,1]^k$, $P_n\pi^{-1}_{t_1,...,t_k} \underset{n\to\infty}{\Longrightarrow} P\pi^{-1}_{t_1,...,t_k}$*), alors* $P_n \underset{n\to\infty}{\Longrightarrow} P$.

Il existe des critères précis pour savoir si une famille de probabilités est tendue, liées à la régularité des trajectoires (voir Billingsley (1968), par exemple). En particulier,

Définition 8.5 ***(Module de continuité)***. *Soit* $x \in C$, *on appelle module de continuité de* x *la fonction*

$$\begin{array}{rcl} w_x :]0,+\infty[& \longrightarrow & [0,+\infty[\\ \delta & \longmapsto & \sup\limits_{\{(s,t)\in[0,1]^2, |t-s|<\delta\}} |x(t) - x(s)| . \end{array}$$

Théorème 8.4 $(P_n)_{n\geq 1}$ *est tendue si et seulement si les deux conditions suivantes sont vérifiées:*
(i) $\forall \eta > 0$, $\exists a$, $\forall n \in \mathbb{N}$, $P_n(x/\,|x(0)| > a) < \eta$,
(ii) $\forall \varepsilon, \eta > 0$, $\exists \delta \in]0,1[$, $\exists n_0 \in \mathbb{N}$, $\forall n \geq n_0$, $P_n(x/w_x(\delta) > \varepsilon) < \eta$.

Au lieu de raisonner en termes de suites de probabilités, on peut utiliser des éléments aléatoires. Soit $(X_n)_{n\geq 1}$ une suite d'éléments aléatoires de $(C,\mathcal{C})$. Posons $P_n = P_{X_n}$. Alors les conditions du théorème précédent s'écrivent

$$\begin{array}{rcl} \text{(i)} & \Longleftrightarrow & \forall \eta > 0, \exists a, \forall n \in \mathbb{N}, P(X_n \in \{x/\,|x(0)| > a\}) < \eta \\ & \Longleftrightarrow & \forall \eta > 0, \exists a, \forall n \in \mathbb{N}, P(|X_n(0)| > a) < \eta \\ & \Longleftrightarrow & \text{la suite de variables aléatoires réelles } (X_n(0))_{n\geq 1} \text{ est tendue.} \\ \text{(ii)} & \Longleftrightarrow & \forall \varepsilon, \eta > 0, \exists \delta \in]0,1[, \exists n_0 \in \mathbb{N}, \forall n \geq n_0, P(w_{X_n}(\delta) > \varepsilon) < \eta. \end{array}$$

8.2.2 Application à l'espace des fonctions cadlag

Soit $D([0,1]) = D$ l'espace des fonctions définies sur [0,1] continues à droite admettant une limite à gauche en tout point de [0,1] (fonctions "cadlag"). Cet espace est fondamental car il contient les fonctions de répartitions empiriques.

Avec la métrique $\|.\|_\infty$, l'espace D n'est pas séparable. En effet, l'ensemble non dénombrable de fonctions $\{f_t(x) = \mathbf{1}\{x \leq t\}; t \in [0,1]\}$ contient des éléments distants de 1 deux à deux pour la norme infinie. De plus, le processus empirique n'est pas mesurable pour la tribu borélienne associée à $\|.\|_\infty$. Le choix de la métrique, et plus généralement de la topologie, est important car la topologie sur D définit l'ensemble des fonctions continues sur D, qui sont utilisées dans la définition de la convergence faible. Si on souhaite travailler avec la tribu borélienne $\mathcal{D}$, un choix classique consiste en la topologie de Skorohod.

Soit $\Lambda = \{\lambda : [0,1] \longrightarrow [0,1]$, continues strictement croissantes, $\lambda(0) = \lambda(1) = 0\}$. On définit alors la distance de Skorohod :

$$\forall (x,y) \in D^2, d(x,y) = \inf \{\varepsilon > 0, \exists \lambda \in \Lambda, \forall t, |\lambda(t) - t| < \varepsilon, |x(t) - y(\lambda(t))| < \varepsilon\}$$

Cette distance induit une topologie sur D, dite topologie de Skorohod. Son interprétation est relativement simple : pour toute fonction y dans un voisinage de $x \in D$, une petite déformation uniforme de l'axe des abscisses induit une petite déformation des valeurs prises par y sur l'axe des ordonnées. Remarquons que pour définir la boule associés à $\|\cdot\|_\infty$, on ne peut pas bouger l'axe des abscisses.

On montre que (D,d) est séparable mais pas complet. De plus, toute boule $B_d(x,r)$ pour la topologie de Skorohod contient la boule $B_\infty(x,r/2)$ pour la topologie de la norme uniforme. Donc la topologie de la norme uniforme $\mathcal{U}$ contient la topologie de Skorohod $\mathcal{D}$. On peut montrer que $\mathcal{D}$ restreinte à $C([0,1])$ coïncide avec $\mathcal{U}$.

Citons un autre résultat utile : la fonction $\|\cdot\|_\infty : (D,d) \longrightarrow \mathbb{R}$, qui à f associe $\|f\|_\infty = \sup_t |f(t)|$ est continue en tout point f dans $C([0,1])$. Donc, si X_n tend faiblement vers un processus à trajectoires presque sûrement continues, on peut donc appliquer le continuous mapping theorem pour avoir la loi asymptotique de $\|X_n\|_\infty$. Ainsi, on peut retrouver les lois limites de statistiques de tests du type Kolmogorov-Smirnov...

Pour rendre D complet, on introduit alors une métrique d_0 équivalente à d. Pour $\lambda \in \Lambda$, posons

$$\begin{aligned} \|\lambda\| &= \sup_{s \neq t} \left| \log \left(\frac{\lambda(t) - \lambda(s)}{t - s} \right) \right| \text{ et} \\ d_0(x,y) &= \inf \{\varepsilon > 0, \exists \lambda \in \Lambda, \|\lambda\| < \varepsilon, \forall t, |x(t) - y(\lambda(t))| < \varepsilon\} \end{aligned}$$

On note $\mathcal{D}$ l'ensemble des boréliens de (D,d_0).

Proposition 8.1 *Les métriques d et d_0 sont équivalentes. De plus, (D,d_0) est séparable et complet.*

Ainsi, $(D,d_0,\mathcal{D})$ est un choix raisonnable d'espace de travail car, dans cet espace, être serré équivaut à être relativement compact.

On aimerait donc obtenir la convergence faible de la suite $(X_n)_{n\geq 1}$, éléments aléatoires à valeurs dans D, en montrant que $(X_n)_{n\geq 1}$ est tendue et que les distributions de dimensions finies convergent faiblement (vers celles de X). Or, on rencontre ici un problème car, pour tout vecteur $(t_1,...,t_k)$, $\pi_{t_1,...,t_k}$ est mesurable par rapport aux boréliens de (D,d_0) mais n'est pas forcément continue. En effet, soit une function $x \in D$ discontinue en $t_0 \in]0,1[$. Posons alors, pour tout $t \in [0,1]$, $x_n(t) = x(\lambda_n(t))$ où λ_n est linéaire sur $[0,t_0]$ et $[t_0,1]$ avec $\lambda_n(t_0) = t_0 - 1/n$. Alors, $x_n \underset{n\to+\infty}{\longrightarrow} x$ dans (D,d_0), mais $x_n(t_0) \underset{n\to+\infty}{\not\longrightarrow} x(t_0)$, i.e. $\pi_{t_0}(x_n) \underset{n\to+\infty}{\not\longrightarrow} \pi_{t_0}(x)$.

En fait, pour tout $t \in]0,1[$, π_t est continue au point x si et seulement si x est continue au point t. On peut donc avoir $P_n \underset{n\to\infty}{\Longrightarrow} P$ et $P_n\pi^{-1}_{t_1,\ldots,t_k}$ ne converge pas faiblement vers $P\pi^{-1}_{t_1,\ldots,t_k}$ pour certains k-uplets $(t_1,\ldots,t_k)$.

`Exercice 8.1` *En posant P = la masse unité en* $\mathbf{1}\left([0,\frac{1}{2}[\right)$, *et P_n = la masse unité en* $\mathbf{1}\left([0,\frac{1}{2}+\frac{1}{n}[\right)$, *montrer que* $P_n \underset{n\to\infty}{\Longrightarrow} P$ *et* $P_n\pi^{-1}_{\frac{1}{2}} \underset{n\to\infty}{\not\Longrightarrow} P\pi^{-1}_{\frac{1}{2}}$.

Néanmoins, les projections de dimension finies sont "presque sûrement" continues. En effet,

Proposition 8.2 *Soit T_P l'ensemble des $t \in [0,1]$ tels que, pour tout t, π_t est continue sauf en un ensemble de P-mesure nulle de D. Alors, pour tout P, T_P contient 0,1, et son complémentaire dans [0,1] est au plus dénombrable. Donc si $(t_1,\ldots,t_k) \in T_P^k$, $\pi_{t_1,\ldots,t_k}$ est continue sauf sur un ensemble de P-mesure nulle.*

Théorème 8.5 *Si $(P_n)_{n\geq 1}$ est tendue, et si pour tout vecteur $(t_1,\ldots,t_k) \in T_P^k$, $P_n\pi^{-1}_{t_1,\ldots,t_k}$ converge faiblement vers $P\pi^{-1}_{t_1,\ldots,t_k}$, alors $P_n \underset{n\to\infty}{\Longrightarrow} P$.*

On peut étendre la notion de module de continuité aux éléments de D, et donner des critères de tightness, comme dans l'espace des fonctions continues : voir Billingsley (1968).

Théorème 8.6 ***(Théorème de Donsker).*** *Soit $(\xi_i)_{i\geq 1}$ une suite de variables aléatoires i.i.d., centrées, de variance σ^2. Soit $(X_n)_{n\geq 1}$ une suite d'éléments aléatoires de D définie par $X_n(t,\omega) = 1/(\sigma\sqrt{n})S_{[nt]}(\omega)$, $S_i(\omega) = \sum_{k=1}^{i}\xi_k(\omega)$. Alors, $X_n \underset{n\to\infty}{\overset{loi}{\longrightarrow}} W$, où W désigne un processus Brownien standard.*

Soit F la fonction de répartition de ξ_1, et F_n la fonction de répartition empirique des $(\xi_i)_{1\leq i\leq n}$, i.e. $F_n(t,\omega) = n^{-1}\sum_{i=1}^{n} 1(\xi_i(\omega) \leq t)$. Soit $Y_n(t,\omega) = \sqrt{n}\,[F_n(t,\omega) - F(t)]$. Alors, $Y_n \underset{n\to\infty}{\overset{loi}{\longrightarrow}} Y$, où Y est un pont brownien.

La convergence faible dans D du processus empirique a de multiple implications en statistiques.

Rappelons qu'un mouvement brownien W est un processus dont presque toutes les tajectoires sont continues. Il se caractérise par les propriétés suivantes.

- il est gaussien : pour tout m-uplet $(t_1,\ldots,t_m)$, la variable aléatoire $(W(t_1,\cdot),\cdots,W(t_m,\cdot))$ est un vecteur gaussien centré,
- pour tout t, $VarW(t,\cdot) = t$,
- il est à accroissements indépendants : si $t > s$, la variable $W(t,\cdot) - W(s,\cdot)$ est indépendante de toutes les variables $W(u,\cdot)$, avec $u \leq s$.

On peut générer un pont brownien à partir d'un brownien standard W par les formules

$$Y(t) = W(t) - tW(1), t \in [0,1], \text{ et réciproquement}$$

$$W(t) = (t+1)Y\left(\frac{t}{t+1}\right) \text{ pour tout } t \in [0,+\infty[.$$

Le pont brownien Y est tel que

- presque toutes ses trajectoires sont continues, pour tout t,
- $Y(t,\cdot)$ suit une loi gaussienne centrée, pour tout t,
- $E[Y(s)Y(t)] = s \wedge t - st$, pour tout $(s,t) \in [0,1]^2$.

On a remarqué que la topologie de Skorohod $\mathcal{D}$ était inclue dans la topologie $\mathcal{U}$ de la convergence uniforme. Ainsi, toute fonction $f : D \to \mathbb{R}$ $\mathcal{D}$-continue est $\mathcal{U}$-continue. Il y a moins de fonctions $\mathcal{D}$-continues que de fonctions $\mathcal{U}$-continues, donc la convergence faible dans $(D,d_0,\mathcal{D})$ est moins "riche" que celle dans $(D,\|\cdot\|_\infty,\mathcal{U})$. Comme de plus $\|.\|_\infty$ est plus intuitive et plus facile à utiliser que d_0, il est tentant d'adapter la théorie à $(D,\|.\|_\infty)$.

8.2.3 L'approche de D par la norme uniforme (Pollard, 1984)

L'idée est de raisonner dans l'espace $(D,\|\cdot\|_\infty,\mathcal{P})$, $\mathcal{P}$ étant une sous-tribu de la tribu borélienne, qui rend toutes les projections de dimensions finies continues. En fait, $\mathcal{P}$ est la σ-algèbre engendrée par les projections $\pi_{t_1,\ldots,t_k}$ pour tout k-uplet $(t_1,\ldots,t_k) \in [0,1]^k$ et tout k, soit

$$\mathcal{P} = \sigma\left\{\pi^{-1}_{t_1,\ldots,t_k}(B), k \geq 1, (t_1,\ldots,t_k) \in [0,1]^k, B \in \mathcal{B}(\mathbb{R}^k)\right\}.$$

On appelle $\mathcal{P}$ la tribu de projection de D. Ainsi $X : (\Omega,\mathcal{A},P) \to (D,\mathcal{P})$ est mesurable si et seulement si pour tout $t \in [0,1]$, $\pi_t \circ X : (\Omega,\mathcal{A},\mathcal{P}) \longrightarrow (\mathbb{R},\mathcal{B}(\mathbb{R}))$ est mesurable. Notons que $\mathcal{P}$ est engendrée par les boules ouvertes de D, alors que la tribu borélienne de $(D,\|\cdot\|_\infty)$ est engendrée par les ouverts de $\mathcal{D}$. Les deux concepts ne coïncident pas car $(D,\|\cdot\|_\infty)$ n'est pas séparable.

Ces choix ont des répercussions sur la portée d'un résultat de convergence faible. En effet, pour l'appliquer, on considère des fonctions f de $(D,\|\cdot\|_\infty,\mathcal{P})$ vers $\mathbb{R}$ continues et mesurables ; la $\mathcal{P}$-mesurabilité doit être vérifié au cas par cas. Par exemple, la fonction f qui $x \in D$ associe $f(x) = \sup_{t\in[0,1]} |x(t)|$ est $\|.\|_\infty$-continue et $\mathcal{P}$-mesurable. Plus généralement, la plupart des fonctions f utiles sont $\mathcal{P}$-mesurables.

Il convient de faire une remarque importante: $C = (C([0,1]), \|.\|_\infty)$ est un sous-espace de $(D, \|.\|_\infty)$, séparable, $\mathcal{P}$-mesurable et la tribu de projection $\mathcal{P}$ restreinte à C coïncide avec la tribu borélienne de C. Il n'y a donc plus de problème de σ-algèbre dans C.

Théorème 8.7 *Soient X et $(X_n)_{n\geq 1}$ des éléments aléatoires de $(D, \|.\|_\infty, \mathcal{P})$. Supposons que pour un sous-ensemble séparable de D, noté C_0, on a $P(X \in C_0) = 0$. Alors, $X_n \underset{n\to+\infty}{\Longrightarrow} X$ si et seulement si*

(i) pour tout S, sous-ensemble fini de $[0,1]$, $\pi_S X_n \underset{n\to+\infty}{\overset{loi}{\longrightarrow}} \pi_S X$,

(ii) pour tous $\varepsilon,\delta > 0$, il existe une grille $0 = t_0 < t_1 < \ldots < t_m = 1$, telle que

$$\limsup_{n\to+\infty} P\left(\max_i \sup_{t\in[t_i,t_{i+1}[} |X_n(t) - X_n(t_i)| > \delta\right) < \varepsilon.$$

Le théorème s'applique en particulier lorsque le processus limite a presque ses trajectoires continues. Par exemple, le processus empirique Y_n du théorème 8.6 converge en loi vers un pont brownien dans $(D,\|\cdot\|_\infty,\mathcal{P})$ comme dans $(D,d_0,\mathcal{D})$ muni de sa tribu borélienne.

En conclusion, l'approche de Pollard est à conseiller si les lois limite sont à trajectoires continues (en particulier les processus gaussiens), sinon mieux vaut l'approche par la métrique de Skorohod.

8.3 Rappels sur les processus

Soit un espace probabilisé $(\Omega,\mathcal{A},P)$.

Une filtration $\mathcal{F} = (\mathcal{F}_t)_{t\geq 0}$ est une famille croissante et continue à droite de σ-algèbres, c'est-à-dire qui vérifie les propriétés suivantes :

- pour tous $s \leq t$, $\mathcal{F}_s \subset \mathcal{F}_t$;
- $\mathcal{F}_s = \cap_{t\geq s}\mathcal{F}_t$ pour tout t;
- $P(A) = 0 \Rightarrow A \in \mathcal{F}_t$ pour tout t.

Chez certains auteurs, les deux conditions précédentes ne sont pas toujours supposées. La première exprime la continuité à droite de la filtration ; la seconde est appelée souvent propriété de "complétude" de la filtration dans l'espace Ω muni de la probabilité P. Comme ces conditions se trouveront toujours vérifiées dans les cas qui nous concernent, nous les imposons dès le départ. On notera $\mathcal{F}_{t-}$ la σ-algèbre engendrée par les $\mathcal{F}_s$, $s < t$.

Un processus $X = \{X(t), t \in \Gamma\}$ est une famille de variables aléatoires indexée par un ensemble Γ, définies sur le même espace probabilisé $(\Omega, \mathcal{A}, P)$. Si Γ est à valeurs discrètes, X est dit processus à temps discret (généralement dans ce cas, $\Gamma = \mathbb{N}$). Si Γ est un intervalle réel, X est dit processus à temps discret. Pour signifier la double dépendance en $t \in \Gamma$ et $\omega \in \Omega$ d'un processus aléatoire, on note parfois $X(t,\omega)$ la valeur de X en point t pour la réalisation ω.

Un processus $X = (X_t)_{t \geq 0}$ est adapté à $\mathcal{F}$ si pour tout t, $X(t)$ est $\mathcal{F}_t$-mesurable.

Soit un processus à temps continu X. Pour toute réalisation $\omega \in \Omega$, l'application de Γ vers $\mathbb{R}$ qui à t associe $X(t,\omega)$ est appelée la trajectoire du processus pour la réalisation ω. Le processus X est dit continu à droite (ou à gauche), à variation bornée, croissant (ou décroissant), possédant des limites à droite (ou à gauche) etc, si l'ensemble des trajectoires qui satisfont ces propriétés est de probabilité 1.

Un processus $X = \{X_t, t \geq 0\}$ est dit intégrable si $\sup_{t \geq 0} E|X(t)| < \infty$, de carré intégrable si $\sup_t E[X^2(t)] < \infty$, borné s'il existe un constante finie C telle que $P(\sup_t |X(t)| < C) = 1$.

Un processus $X = (X_t, t \geq 0)$ adapté à une filtration $\{\mathcal{F}_t, t \geq 0\}$ est une martingale si pour tout couple (s,t), $s \leq t$, $E(X(t)|\mathcal{F}_s] = X(s)$ presque sûrement. Le processus X est une surmartingale si pour tout couple (s,t), $s \leq t$, $E[X(t)|\mathcal{F}_s] \geq X(s)$ presque sûrement. X est une martingale si et seulement si on a presque sûrement, pour tout t, $E[X(t)|\mathcal{F}_{t-}] = X(t-)$. On écrit cette relation (de manière imprécise) $E[dX(t)|\mathcal{F}_{t-}] = 0$.

On remarque en effet que, pour tous s et t, $s \leq t$,

$$E[X(t)|\mathcal{F}_s] - X(s) = E[X(t) - X(s)|\mathcal{F}_s] = E[\int_s^t dX(u)|\mathcal{F}_s]$$
$$= \int_s^t E[dX(u)|\mathcal{F}_s] = \int_s^t E[E[dX(u)|\mathcal{F}_{u-}]|\mathcal{F}_s].$$

Un temps d'arrêt T est une variable aléatoire à valeurs réelles non négatives telles que $\{T \leq t\} \in \mathcal{F}_t$ pour tout t.

Une suite $(\tau_n)_{n \geq 1}$ de temps d'arrêts est dite localisante si elle est non décroissante et si pour tout t,

$$P(\tau_n \geq t) \underset{n \to \infty}{\longrightarrow} 1\,.$$

Un processus M est une martingale (resp. sousmartingale) locale pour la filtration $\mathcal{F}$ s'il existe une suite localisante (τ_n) telle que, pour tout n, $M_n \equiv \{M(t \wedge \tau_n), t \geq 0\}$ est une $\mathcal{F}$-martingale (resp. sousmartingale). Si le processus M_n précédents est de carré intégrable, M_n est dite martingale de carré intégrable. Si c'est le cas pour tout n, M est dite martingale locale de carré intégrable.

Un processus X est localement borné s'il existe une suite localisante de temps d'arrêts $(\tau_n)_n$ telle que $X_n \equiv \{X(t \wedge \tau_n, t \geq 0\}$ est un processus borné pour tout n (autrement dit : pour une certaine suite de constantes $(c_n)_{n \geq 1}$, on a pour tout n, presque sûrement $\mathbf{1}\{\tau_n > 0\} \sup_{t \leq T_n} |X(t)| \leq c_n$).

Théorème 8.8 (Inégalité de Lenglart) *Soit X un processus continu à droite et Y un processus prédictible tel que $Y(0) = 0$. Si, pour tout temps d'arrêt borné T, $E[|X(T)|] \leq E[Y(T)]$, alors, pour tout temps d'arrêt T et tout couple (ε, η) positifs,*

$$P\left(\sup_{t \leq T} |X(t)| \geq \varepsilon\right) \leq \frac{\eta}{\varepsilon} + P(Y(T) \geq \eta).$$

8.4 Familles de loi paramétriques utiles

Nous rassemblons ici quelques familles de densités utilisées habituellement en analyse des durées de vie. Evidemment, leur support est inclu dans $\mathbb{R}^+$.

i. **Lois de Weibull**. Elles constituent la famille la plus usuelle en modèles de durées. Elles sont définies par $S(t) = \exp(-(\mu t)^p)$, $\mu > 0$, $p > 0$, d'où $\lambda(t) = p\mu^p t^{p-1}$ et $f(t) = p\mu^p t^{p-1}\exp(-(\mu t)^p)$. Comme cas particulier ($p = 1$), on obtient la distribution exponentielle.
Si T suit bien une loi de weibull, en traçant la courbe $\ln t \mapsto \ln\Big(-\ln(\hat{S}_{KM}(t))\Big)$, on obtient une droite de pente p et d'ordonnées à l'origine $p\ln\mu$.
On remarque qu'on peut écrire $Y = \alpha + \sigma w$, avec $\alpha = -\ln\mu$, $\sigma = p^{-1}$ et où w est une variable aléatoire suivant une loi de valeur extrême (dite également loi de Gompertz, ou de Gumbel), définie par $P(w > t) = \exp(-\exp(t))$ pour tout $t \geq 0$ (le montrer).

ii. **Distribution Lognormale**. Elles sont utiles pour approcher des fonctions de hasard non monotones. Ici, $S(t) = 1-\Phi(p\ln(\mu t))$, où p et μ sont positif, et Φ est la fonction de répartition d'une loi normale centrée réduite. Alors, $f(t) = (2\pi)^{-1/2}p\exp(-p^2(\ln\mu t)^2/2)/t$ et

$$\lambda(t) = \frac{p}{t}\cdot\frac{\phi(p\ln(\mu t))}{1-\Phi(p\ln(\mu t))} = \frac{p}{t}\cdot \text{Mills }(p\ln(\mu t)),$$

où Mills(x) est le ratio de Mills $\phi(x)/(1-\Phi(x))$ (ϕ densité d'une normale centrée réduite).
On retombre sur le modèle bien connu $Y = \alpha + \sigma W$, $W \sim \mathcal{N}(0,1)$.

iii. **Distribution Loglogistique** Une difficulté de la famille lognormale provient du caractère non calculatoire (à la main!) de ϕ et Φ. On peut utiliser plutôt la distribution loglogistique $S(t) = (1+(\mu t)^p)^{-1}$, qui donne $\lambda(t) = \mu^p p t^{p-1}/(1+(\mu t)^p)$.
On montre que $Y = \alpha + \sigma W$, où W suit une distribution logistique: $P(W > w) = (1+\exp(w))^{-1}$.

iv. **Distribution Gamma**. Elle nous fournit une extension de la loi exponentielle différente de la famille weibull.
Ici, $f(t) = \mu^k t^{k-1}\exp(-\mu t)/\Gamma(k)$, $\mu > 0$,$k > 0$, et $\Gamma(k) = \int_0^{+\infty} x^{k-1}\exp(x)\,dx$.
De plus, on peut écrire, $Y = \alpha + \sigma W$, la denité de W étant $f_W(x) = \exp(kx - \exp(x))/\Gamma(k)$.
On peut montrer que $k^{1/2}(W - \ln k)$ est asymptotiquement normale quand $k \to \infty$.

v. **Distribution Gamma généralisée** On peut englober dans un même famille à la fois les loi Weibull et Gamma. En effet, soit

$$f(t) = \frac{\mu^{pk}pt^{pk-1}}{\Gamma(k)}\exp(-(\mu t)^p),$$

μ, k et p étant trois paramètres strictement positifs. Si $k = 1$, on retrouve la famille Weibull; si $p = 1$, on retrouve la famille Gamma.
Dans ce cas, $Y = \alpha + \sigma W$, $\alpha = -\ln\mu$, $\sigma = p^{-1}$, et W suit une loi Gamma de paramètre k.

vi. **Distributions F généralisées**. On peut en fait englober toutes les familles précédentes au sein d'une unique, en posant $Y = \alpha + \sigma W$ et

$$f_W(x) = \left(\frac{m_1}{m_2}\right)^{m_1}\frac{\exp(m_1 x)}{B(m_1,m_2)}[1+\frac{m_1\exp(x)}{m_2}]^{-(m_1+m_2)},$$

avec m_1 et m_2 deux paramètres strictement positifs, et $B(m_1,m_2) = \int_0^{+\infty} x^{m_1}(1+x)^{-m_1-m_2}\,dx = \Gamma(m_1)\Gamma(m_2)/\Gamma(m_1+m_2)$.
Si $(m_1,m_2) = (1,1)$ on retombe sur la famille loglogistique; si $m_2 = +\infty$, c'est la famille des gamma généralisées, et si $(m_1,m_2) = (1,+\infty)$, il s'agit de la famille Weibull.

On peut construire d'autres familles paramétriques

- par agrégation: $\tilde{f}_{(\theta_2,\theta_3)}(t) = \int f_{(\theta_1,\theta_2)}(t)\pi_{\theta_3}(\theta_1)\,d\theta_1$, $\{\pi_{\theta_3}\}_{\theta_3}$ étant une famille de densités sur l'espace décrit par le sous vecteur de paramètres θ_1. La nouvelle famille de loi dépend donc du nouveau vecteur de paramètres (θ_2,θ_3).

- par transformation croissante de la durée : on pose $\tilde{T} = \psi(T)$, ψ fonction strictement croissante donnée[22]. Alors la loi de $\tilde{T}$ est donnée par la survie $\tilde{S}(t) = S(\psi^{-1}(t))$ ou la fonction de hasard $\tilde{\lambda}(t) = \lambda/\psi'(\psi^{-1}(t))$. En particulier, un changement d'echelle $\tilde{T} = \alpha^{-1}T$ rentre dans ca cadre ; il fournit $\tilde{\lambda}(t) = \alpha\lambda(\alpha t)$.
- par une transformation du type "hasards proportionnels" : la nouvelle fonction de hasard est définie par $\tilde{\lambda}(t) = \alpha\lambda(t)$. Alors $\tilde{S}(t) = S(t)^{\alpha}$.

A partir des procédés précédents et des familles répertoriées, il est facile d'introduire des covariables dans les modèles. Par exemple, une transformation du type "hasards proportionnels" avec $\alpha = \exp(Z'\beta)$, Z vecteur des covariables et β paramètre à estimer, nous ramène à un modèle de Cox paramétrique si on se fixe la famille paramétrique à laquelle appartient S.

`Exemple. 8.2` Un modèle du type "Weibull à hasards proportionnels" serait défini par les relations

$$\lambda(t|z) = p\mu^p t^{p-1} \exp(z'\beta),$$

pour tout $t > 0$, μ, p et β étant trois paramètres positifs à estimer. Dans le cadre d'une censure aléatoire droite indépendante de T , cela peut être fait par la méthode du maximum de vraisemblance, i.e. en cherchant à résoudre

$$\begin{aligned}
&\arg\max_{\mu,p,\beta} \prod_{i=1}^{n} \lambda(t_i|z_i)^{\delta_i} S(t_i) \\
&= \arg\max \sum_{i=1}^{n} \delta_i \ln \lambda(t_i|z_i) - \sum_{i=1}^{n} \int_0^{t_i} \lambda(v|z_i)\, dv \\
&= \arg\max \sum_{i=1}^{n} \delta_i \ln(p\mu^p t_i^{p-1}) + \sum_{i=1}^{n} \delta_i z_i'\beta - \mu^{p-1} \sum_{i=1}^{n} t_i^p \exp(z_i'\beta).
\end{aligned}$$

Visiblement, il semble difficile de trouver des formules explicites dans ce cas particulier. Il conviendra alors de recourir à des solutions de maximisation numérique.

22. c'est ce qu'on a fait avec $Y = \ln T$ au-dessus

9 Références

Références

[1] Akritas, M.G. (2000). The central limit theorem under censoring *Bernoulli*, **6**, 1109 – 1120.

[2] Andersen, P. K. et Gill, R. D. (1982). Cox's regression model for counting processes : A large sample study. *Ann. Statist.*, **10**, 1100 – 1120.

[3] Andersen, P. K., Borgan, Ø, Gill, R. D. et Keiding, N. (1993). Statistical Models base on counting Processes, Springer Verlag.

[4] Bosq, D. and Lecoutre, J-P. (1987). *Théorie de l'estimation fonctionnelle.* Economica, Paris.

[5] Billinsley, P. (1968). *Convergence of probability measures.* Wiley, New-York.

[6] Blum, J.R. et Susarla, V. (1980). Maximal deviation theory of density and failure rate function estimates based on censored data. dans *Multivariate Analysis V*, P.R. Krishnaiah, ed., North-Holland, 213 – 222.

[7] breslow, N. et Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *Ann. Statist.*, **2**, 437 – 453.

[8] Buckley, J. et James, I. (1979). Linear regression with censored data. *Biometrika*, **66**, 429 – 436.

[9] Chao, M.-T. et Lo, S.-H. (1988). Some representations of the nonparametric maximum likelihood estimators with truncated data. *Ann. Statist.*, **16**, 661 – 668.

[10] Clayton, D.G. et Cuzik, J. (1985). Multivariate generalizations of the proportional hazards model (with discussion). -*J. Roy. Statist. Soc. A*, **148** , 82 – 117.

[11] Celeux, G. et Diebolt, J. (1985). The SEM algorithm : a probabilistic teacher algorithm derived from the EM algorithm for the mixture problem. *Computational Statistics Quaterly*, **2**, 73 – 82.

[12] Cox, D. R. (1972). Regression models and life-tables (with discussion). *J. Roy. Statist. Soc. B*, **34** , 187 – 220.

[13] Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**, 269 – 276.

[14] Cox, D.R. and Oakes, D. (1984). *Analysis of Survival Data.* Chapman and Hall, New-York.

[15] Crowder, M. (1994). Identifiability Crises in Competing Risks. *International Statistical Review* **62**, 379 – 391.

[16] Dabrowska, D. (1987). Nonparametric regression with censored data. *Scand. J. Statist.*, **14**,181 – 197.

[17] Dabrowska, D. (1989). Uniform consistency of the kernel conditional Kaplan-Meier estimate. *Ann. Statist.*, **17**, 1157 – 1167.

[18] Delyon, B., Lavielle, M. et Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.*, **27**, 94 – 128.

[19] Dempster, A.P., Laird N.M. et Rubin, D.B. (1977). Maximum likelihood estimation from incomplete date via the EM algorithm (with discussion). *J. Roy. Statist. Soc. B*, **39**, 1 – 38.

[20] Doksum, K. A. (1987). An extension of partial likelohhod methods for proportional hazard models to general transformation models. *Ann. Statist.*, **15**, 325 – 345.

[21] Droesbeke, J. J., Fichet, B. et Tassi, P. (ed.) (1989). *Analyse Statistique des durées de vie* , Collection ASU, Economica, Paris.

[22] Diehl, S. et Stute, W. (1988). Kernel density and hazard function estimation in the presence of censoring. *J. Multivariate Anal.*, **25**, 299 – 310.

[23] Fan, J. et Gijbels, I. (1994). Censored regression : local linear approximation and their applications. *J. Amer. Statist. Assoc.*, **89**, 560 – 570.

[24] Fan, J., Gijbels, I. et King, M. (1997). Local likelihood and local partial likelihood in hazard estimation. *Ann. Statist.*, **25**, 1661 – 1690.

[25] Fermanian, J.-D. (1996). Multivariate hazard rates under random censorship. Document de Travail CREST, **9603**.

[26] Fermanian, J-D. (1997). Multivariate hazard rates under random censorship. *J. Multivariate Anal.* **62**, 273 – 309.

[27] Fermanian, J.-D. (1999). A new bandwidth selector in hazard estimation. *J. Nonparametric Statist.*, **10**, 137 – 182.

[28] Fermanian, J-D (2003). Nonparametric estimation of Competing Risks models with covariatesŤ. *Journal of Multivariate Analysis* **85**, 156 – 191.

[29] Fygenson, M. et Ritov, Y. (1994). Monotone estimating equations for censored data. *Ann. Statist.*, **22**, 732 – 746.

[30] Gill, R.D. (1980). *Censoring and Stochastic Integrals*, Mathematical Centre Tracts 124, Mathematisch Centrum, Amsterdam.

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

[32] Gill, R.D. (1984). Understanding Cox(s regression model : a martingale approach. *J. Amer. Statist. Assoc.*, **79**, 441 – 447.

[33] Gill, R. D., Vardi, Y. et Wellner, J. (1988). Large sample theory of empirical distributions in biased sampling models. *Ann. Stat.*, **16**, 1069 – 1112.

[34] Gørgens, T. et Horowitz, J. L. (1999). Semiparametric estimation of a censored regression model with an unknown transformation of the dependent variable. *J. of Econometrics*, **90**, 155 – 191.

[35] Gouriéroux, C. et Monfort, A.. (1989) Econometrics based on endogenous samples. Document de Travail CREST **8903**.

[36] Gourièroux, C. et Monfort, A. (1989). Statistique et modèles économétriques. *Economica*, Paris.

[37] Gouriéroux, C. et Monfort, A.. (1991) Modèles de durées et effets de génération. Document de Travail CREST **9125**.

[38] Gouriéroux, C., Monfort, A. et Trognon, A. (1984). Pseudo Maximum Likelihood Methods : Theory. *Econometrica*, **52**, 681 – 700.

[39] Grégoire, G. (1993). Least-squares cross-validation for counting process intensities. *Scand. J. Statist.*, **20**, 343 – 360.

[40] Han, A. K. (1987). Non-parametric analysis of a generalized regression model. *J. of Econometrics*, **35**, 303 – 316.

[41] Harrington, T.R. et Fleming, D.P. (1991). *Counting processes and survival analysis*, Wiley.

[42] Härdle, W. et Stoker, T. M. (1989). Investing smooth multiple regression by the method of average derivatives. *J. Amer. Statist. Ass.*, **84**, 986 – 995.

[43] Hastie, T. J. et Tibshirani, R. J. (1990). Generalized additive models. *Monographs on Statistics and Applied Probability*, Vol. **43**, Chapman et Hall, London.

[44] Heckman, J. J. and Honoré, B. E. (1989). The identifiability of the competing risks model. *Biometrika* **76**, 325 – 330.

[45] Heckman, J. J. et Singer, B. (1984). A method for minimizing the impact of distributional assumptions in econometric models for duration data. *Econometrica*, **52**, 271 – 320.

[46] Honoré, B. E. (1990). Simple estimation of a duration model with unobserved heterogeneity. *Econometrica*, **58**, 453 – 473.

[47] Horowitz, J. L. (1998). Semiparametric methods in econometrics. *Lecture Notes in Statistics*, **131**. Spriger Verlag, New-York.

[48] Horowitz, J. L. (1999). Semiparametric estimation of a proportional hazard model with unobserved heterogeneity. *Econometrica*, **67**, 1001 – 1028.

[49] Horowitz, J. L. et Härdle, W. (1996). Direct semiparametric estimation of single-index models with discrte covariates. *J. Amer. Statist. Ass.*, **91**, 1632 – 1640.

[50] Hougaard, P. (1999). Fundamentals of Survival Data. *Biometrics* **55**, 13 – 22.

[51] Huffer, F.W. et McKeague, I.W. (1991). Weighted least squares estimation for Aalen's additive risk model. *J. Amer. Statist. Assoc.*, **86**,114 – 129.

[52] Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single index models. *J. of Econometrics*, **58**, 71 – 120.

[53] Joe, H (1997). *Multivariate models and dependence concepts.* Chapman et Hall, London.

[54] Kaplan, E. L. et Meier, P. (1958). Nonparametric estimation from incomplete observations. *J Amer. Statist. Assoc.*, **53**, 457 – 481.

[55] Kim, C.K. et Lai, T.L. (1999). Regression smoothers and additive models for censored and truncated data. *Comm. Statist. Theory Meth.*, **28**, 2717 – 2747.

[56] Koenker, R. et Bassett, G. S. (1978). Regression quantiles, *Econometrica*, **46**, 33 – 50.

[57] Koul, H., Susarla, V. et Van Ryzin, J. (1981). Regression analysis with randomly right censored data. *ann. Statist.*, **9**, 1276 – 1288.

[58] Lai, T. L. et Ying, Z. (1991). Estimating a distribution function with truncated and censored data. *Ann. Statist.*, **19**, 417 – 442.

[59] Lai, T. L. et Ying, Z. (1991). Large sample theory of a modified Buckley-James estimator for regression analysis with censored data. *Ann. Statist.*, **19**, 1370 – 1402.

[60] Lai, T.L., Ying, Z. et Zheng, Z.K. (1995). Asymptotic normality of a class of adaptive statistics with applications to synthetic data methods for censored regression. *J. Multivariate Anal.*, **52**, 259 – 279.

[61] Leurgans, S. (1987). Linear models, random censorsing and synthetic data. *Biometrika*, **74**, 301 – 309.

[62] Lin, D. Y., Wei, L. J. et Ying, Z. (1993). Checking the Cox model with cumulative sums of martingale-based residuals. *Biometrika*, **80**,557 – 572.

[63] Lo, S.H., Mack, Y.P. et Wang, J.L. (1989). Density and hazard rate estimation for censored data via strong representation of the Kaplan-Meier estimator. *Prob. Theory Rel. Fields*, **80**, 461 – 473.

[64] Mc Fadden, D.L. et Ruud, P.A. (1994). Estimation by simulation. *Review of Economics and Statistics*, **76**, 591 – 608.

[65] McLachlan, G.J. et Krishnan, T. (1997). The EM algorithm and extensions. *Wiley*, New-York.

[66] Manski, C. et Lerman, S. (1977). The estimation of choice probabilities from choice based samples. *Econometrica*, **45**, 1977 – 1988.

[67] Miller, R. (1976). Least suares regression with censored data. *Biometrika*, **63**, 449 – 464.

[68] Murphy, S. A. (1994). Consistency in a proportional hazards model incorporating a random effect. *Ann. Statist.*, **22**, 712 – 731.

[69] Murphy, S. A. (1995). Asymptotic theory for the frailty model. *Ann. Statist.*, **23**, 182 – 198.

[70] Nelsen, R. (1999). *An introduction to copulas*, Springer, New-York.

[71] Nielsen, S. F. (2000). On simulated EM algorithms. *J. of Econometrics*, **96**, 267 – 292.

[72] O'Sullivan, F. (1993). Nonparametric estimation in the Cox model. *Ann. Statist.*, **21**, 124 – 145.

[73] Padgett, W.J. et McNichols, D.T. (1984). Nonparametric density estimation from censored data. *Comm. Statist. Theory Meth.*, **13**, 1581 – 1611.

[74] Patil, P.N. (1993a). On the least squares cross-validation bandwidth for hazard estimation. *Ann. Statist.*, **21**, 1792 – 1810.

[75] Patil, P.N. (1993b). Bandwidth choice for nonparametric hazard rate estimation. *J. Statist. Plann. Inference*, **35**, 15 – 30.

[76] Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer Verlag, New-York.

[77] Powelle, J. L. (1986). Censored regression quantiles, *J. Econometrics*, **32**, 143 – 155.

[78] Powell, J. L., Stock, J. H. et Stoker, T. M. (1989). Semiparametric estimation of index coefficients. *Econometrica*, **57**, 474 – 523.

[79] PRENTICE, R.L. AND KALBFLEISH, J.D. (1978). The analysis of failure times in the presence of competing risks. *Biometrics* **34** , 541 – 554.

[80] QIN, G. ET JING, B.-Y. (2000). Asymptotic properties for estimation of partial linear models with censored data. *J. Statist. Plan. Inf.*, **84**, 95 – 110.

[81] SARDA, P. ET VIEU, P. (1991). Smoothing parameter selection in hazard estimation. *Statist. Prob. Letters*, **11**, 429 – 434.

[82] SASIENI, P. (1993). Some new estimators for Cox regression. *Ann. Statist.*, **21**, 1721 – 1759.

[83] SCHWARTZ, L. (1973). *Théorie des distributions.* Hermann, Paris.

[84] SHERMAN, R. P. (1993). The limiting distribution of the maximum rank correlation estimator. *Econometrica*, **61**, 123 – 137.

[85] SHORACK, G. ET WELLNER, J. (1986). *Empirical processes with applications to Statistics.* Wiley, New-York.

[86] SHIRYAYEV, A.N. (1984). Probability. *Springer*,New-York.

[87] SRINIVASAN, C. ET ZHOU, M. (1994). Linear regression with censoring. *J. Multivariate Anal.*, **49**, 179 – 201.

[88] STUTE, W. (1993). Almost sure representations of the product-limit estimator for truncated data. *Ann. Statist.*, **21**, 146 – 156.

[89] STUTE, W. (1993). Consistent estimation under random censorship when covariables are present. *J. Multivariate Anal.*, **45**, 89 – 103.

[90] STUTE, W. (1996). Distributional convergence under random censorship when covariables are present. *Scand. J. statist.*, **23**, 461 – 472.

[91] STUTE, W. (1995). The central limit theorem under random censorship. *Ann. Statist.*, **23**, 422 – 439.

[92] STUTE, W. ET WANG, J. L. (1993). The strong law under random censorship. *Ann. Statist.*, **21**, 1591 – 1607.

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

[94] TSIATIS, A. (1975). A nonidentifiability aspect of the problem of competing risks. *Proc. Nat. Acad. Sci. USA* **72**, 20 – 22.

[95] TSIATIS, A. (1981). A large sample study of Cox's regression model. *Ann. Statist.*, **9**, 93 – 108.

[96] TSIATIS, A. (1990). Estimating regression parameters using linear rank tests for censored data. *Ann. Statist.*, **18**, 354 – 372.

[97] WANG, Q. (2000). Estimation of linear error-in-covariables models with validation data under random censorship, *J. Multivariate Anal.*, **74**, 267 – 281.

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

[99] WONG, W. H. (1986). Theory of partial likelihood. *Ann. Statist.*, **14**, 88 – 123.

[100] WU, C.F. (1983). On the convergence properties of the EM Algorithm. *Ann. Statist.*, **11**, 95 – 103.

[101] XIANG, X. (1994). Law of the logarithm for density and hazard rate estimation for censored data. *J. Multivariate Anal.*, **49**, 278 – 286.

[102] ZHENG, Z.K. (1984). Regression analysis with censored data. Ph.D Dissertation. Comulbia Univ.

[103] ZHENG, Z.K. (1987). A class of estimator for the parameter in linear regression with censored data. *Acta Math. Appl. Sinica*, **3**, 231 – 241.

[104] ZHENG, Z.K. (1988). Strong consistency of nonparametric regression estimates with censored data. *J. Math. Res. Exposition*, **2**, 307 – 313.