Estimation paramétrique dans le cadre d'un modèle d'événements concurrents

Cédric Vernier

Vernier Cédric

Estimation paramétrique dans le cadre d'un modèle dévénements concurrents

Mémoire de Statistique de survie

Sous la direction de Mme Ségolen Geffray

Année 2012 - 2013
Université de Strasbourg, UFR de Mathématiques

Ce mémoire se fonde sur la lecture de l'article de Jong-Hyeong Jeong et Jason Fine, *Direct parametric inference for the cumulative incidence function*" paru dans le journal *Applied statistics* **55** (2), 187-200

1 Introduction

Nous avons vu en cours comment traiter des données d'analyse de survie de façon non-paramétrique et semi-paramétrique. Dans ce mémoire on s'intéresse plus particulièrement au cas paramétrique avec d'une part un modèle de Weibull et d'autre part un modèle de Gompertz selon deux approches différentes afin d'analyser le cas de deux événements à risques concurrents mutuellement exclusifs dépendents et en présence de censure à droite indépendante. Nous rappellerons d'abords les principes de l'analyse de survie, les notations et définitions, et les fonctions auxquelles nous nous intéressons. Nous verrons ensuite quelques unes des différentes approches possibles afin d'estimer les fonctions d'intérêt. Dans une quatrième partie nous expliciterons les estimateurs du maximum de vraisemblance pour les paramètres des deux modèles retennus ainsi que les propriétés des estimateurs des fonctions en jeu. Enfin dans une dernière partie nous appliquerons nos résultats afin de générer des simulations avec le logiciel R.

1.1 Cadre de l'étude

Le but de cette étude est de voir comment analyser des données de survie dans le cadre d'événements concurrents dépendants et plus précisément avec un modèle paramétrique. Nos variables d'intérêt sont des durées modélisées par des variables aléatoires positives. Un problème rencontré est le problème de la censure, ce qui correspond à un type particulier de données incomplètes.
Pour illustrer le cadre d'étude, les données d'une étude sur des patientes atteintes de cancer du sein sont analysées. Chaque patiente a été traitée pour son cancer grâce à une intervention chirurgicale et un traitement par chimiothérapie et/ou par radiations. On cherche à analyser la durée qui s'écoule avant de détecter le premier événement qui survient après le traitement (rechute locale ou régionale, rechute à distance, décès). On peut donc utiliser des méthodes d'analyse de survie afin d'étudier l'évolution du risque au cours du temps de ces événements. On ne prend en compte que le premier événement qui survient, on n'a donc pour chaque patiente qu'un seul événement possible, les événements sont mutuellement exclusifs. Dans cette étude on s'intéresse plus particulièrement aux rechutes loco/régionales, ce qui sera donc notre événement d'intérêt et tous les autres types d'événements concurrents constitueront le deuxième groupe d'événements.

1.2 Notations

Nous notons C_i la variable aléatoire qui indique, si il y a censure, la durée entre le traitement et la survenue de la censure pour le patient i.

Soit T_i la variable aléatoire représentant la durée entre le traitement et la survenue d'un événement lié au cancer pour le patient i.

Soit K_i la variable aléatoire qui désigne le type d'événement survenu chez un patient i.

Soit $X_i = \min(T_i, C_i)_{i=1,..,n}$ est la variable aléatoire de la durée observée en présence de censure pour le patient i.

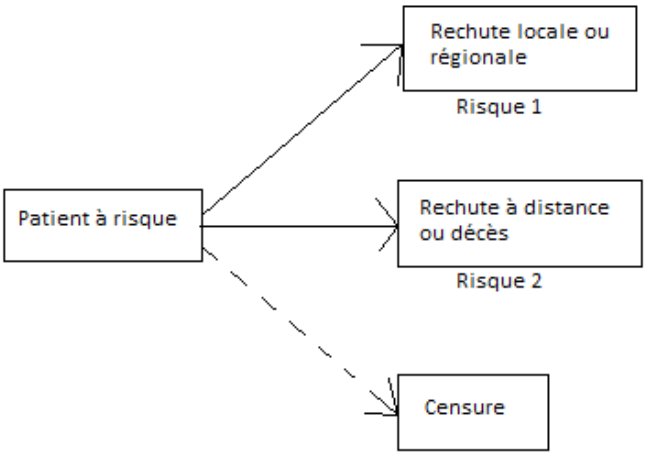


FIGURE 1 – Evènements rencontrés dans l'étude réelle illustrant l'article de De Jeong et Fine (2005)

1.3 Définitions

• Une observation est dite censurée lorsqu'on ne connait pas la durée exacte X mais qu'on dipose d'un minimum, d'un maximum ou les deux à la fois.

• On parle de censure à droite si, au lieu d'observer T on observe une durée C et si on sait que C<T

• Des événements sont dit concurrents si parmi une population la survenue de plusieurs évènements de cause différentes peut se produire. Ils sont de plus mutuellement exclusifs si la survenue d'un évènement de cause k empêche la survenue de tout autre évenement.

2 Fonctions d'intérêt

Comme on s'intéresse à la première occurrence d'un événement dans un modèle à deux risque concurrents dépendants car liés à la maladie, on cherche à estimer la fonction d'incidence cumulée associée au risque k (k=1 ;2), notée

$$C_k(t) = \mathbb{P}[X \leq t, K = k]$$

Cette fonction représente la probabilité d'être affecté avant l'instant t de l'événement de type k en prenant en compte le fait qu'un deuxième risque est à l'œuvre dans la population. Voir Figure 1

D'autres notations nous seront utiles pour la compréhension de l'étude, à savoir :

- la fonction de survie (ou fonction d'incidence globale) :

$$S(t) = \mathbb{P}(T \leq t)$$

qui est la probabilité qu'un événement (lié au cancer donc de type 1 ou 2) survienne avant l'instant t.
- la fonction de répartition $F(t) = 1 - S(t) = \mathbb{P}(T \geq t)$
- la fonction de hasard instantané, lorsque T admet une densité par rapport à la mesure de Lebesgue :

$$\lambda(t) = \lim_{h \to 0^+} \frac{1}{h} \mathbb{P}[t \leq T \leq t + h | T \geq t]$$

où $\mathbb{P}[t \leq T \leq t + h | T \geq t]$ représente la probabilité qu'un événement atteigne le patient dans [t,t+h[sachant qu'aucun événement ne l'avait atteint avant t.

$$\lambda(t) = dF(t)/S(t)$$

- la fonction de hasard brute instantané associée au risque k :

$$\lambda_k(t) = \lim_{h \to 0^+} \frac{1}{h} \mathbb{P}[t \leq T \leq t + h, K = k | T \geq t]$$

où $\mathbb{P}[t \leq T \leq t+h, K = k | T \geq t]$ représente la probabilité que l'événement k atteigne le patient dans [t,t+h[.

3 Modélisation

Plusieurs approches sont possibles afin d'effectuer une telle analyse, il existe des méthodes paramétrique, semi-paramétrique et non-paramétrique.

- **Approche non-paramétrique** : Lin (1997)

Avec cette approche, on ne place pas de modèle sur les fonctions qu'on cherche à estimer. Un estimateur de $S(\cdot)$ est donné par l'estimateur de Kaplan-Meier

$$\widehat{S}(t) = \prod_{i=1:t_i \leq t}^{n} \frac{n_i - d_i}{n_i}$$

où t_i sont les instants de décès (ou rechute) non-censurés ordonnés distincts.
n_i correspond au nombre de sujets à risque avant le temps t.
d_i correspond au nombre de morts (ou de personnes ayant présenté l'événement d'intérêt au temps t).

Remarque : cet estimateur n'est valide que lorsqu'on se place dans le cadre où la censure est indépendante des événements.
La fonction de hasard cumulé : $\Lambda(\cdot)$ est estimée par l'estimateur de Nelson-Aalen.

$$\widehat{\Lambda}(t) = \sum_{i=1:t_i \leq t}^{n} \frac{d_i}{n_i}$$

où t_i sont les instants de décès (ou rechute) ordonnés distincts.
n_i correspond au nombre de sujets à risque avant le temps t.
d_i correspond au nombre de morts (ou de personnes ayant présenté l'événement d'intérêt au temps t).

L'estimateur qui nous intéresse est l'estimateur de la fonction d'incidence cumulée associée à la cause k, $\widetilde{C_k}(\cdot)$ qui est donné par l'estimateur de Aalen-Johansen :

$$\widetilde{C_k}(t) = \sum_{i=1:t_i^k \leq t}^{n} \left(1 - \widehat{S}(t) - t_i^k\right) \frac{d_i^k}{n_i}$$

où t_i^k sont les instants de décès (ou rechute) de cause k, ordonnés distincts.
n_i correspond au nombre de sujets à risque avant le temps t.
d_i^k correspond au nombre de morts de cause k (ou de personnes ayant présenté l'événement d'intérêt spécifique à la cause k au temps t).

- **Approche semi-paramétrique** :Bryant et Dignam (2004)

Dans cette approche on va poser un modèle paramétrique sur $\Lambda_k(.)$ seulement et on estime S(.) de manière non-paramétrique.

On ne s'attardera pas sur cette section faute de temps.

- **Approche paramètrique** :

Il s'agit de poser un modèle paramétrique sur la fonction qu'on cherche à estimer. Dans l'article de Jeong et Fine (2006) trois approches sont proposées.

La première approche considérée dans l'article propose de poser un modèle paramétrique directement sur les fonctions $C_k(\cdot)$ pour k=1 ;2 dite *approche directe*, tandis que la deuxième propose une méthodologie standard basée sur les fonctions $\lambda_k(\cdot)$ pour k=1 ;2 dite *approche du hasard spécifique*. Ces deux approches font intervenir deux paramètres par fonctions, ce qui nous fait quatres paramètres par approche.

Notons qu'une troisième approche avait été proposée par Larson et Dinse (1985) avec un *Mixture Model* afin de représenter la distribution jointe de (T,K). Ce modèle comporte cinq paramètres, le cinquième paramètre étant un paramètre qualifié de paramètre de mise à niveau ce qui rajoute un paramètre additionnel vis à vis des deux approches précédentes, raison pour laquelle ce modèle est écarté.

- Approche directe :

Le but est de poser un modèle paramètrique sur la fonction d'incidence cumulée notée $F_k(x, \psi_k)$ pour k=1 ;2 en notant $\psi_k = (\alpha_k, \beta_k)$. La fonction d'intérêt $F_k(.)$ est impropre ce qui siginifie que $\lim_{t\to\infty} F_k(t) < 1$ dans le cadre de l'étude sur le cancer du sein. En effet, on observe une augmentation du risque de rechute pendant les 10 à 20 premières années puis la présence d'un plateau après. Pour s'ajuster sur de telles données, le modèle doit donc être tel que $F_k(x, \psi_k)$ pour k=1 ;2 soit impropre. On utilise pour cela une distribution de Gompertz :

$$F_k(x,\psi_k) = 1 - \exp\left(\frac{\beta_k}{\alpha_k}(1-\exp(\alpha_k x))\right), k = 1;2$$

avec pour fonction de hasard

$$h_k(x,\psi_k) = \left[\frac{dF_k(x,\psi_k)}{dx}\right] / (1 - F_k(x,\psi_k)) = \beta_k \exp(\alpha_k x)$$

Pour assurer la croissance de la fonction $x \to F_k(x,\psi_k)$ il faut que $\beta > 0$ et pour que la fonction soit impropre il faut que $\alpha < 0$.
On pourrait proposer d'autres distributions pourvu qu'elles s'ajustent aux données par exemple une distribution de Weibull.

• Approche du hasard spécifique :

L'événement K peut prendre deux valeurs (1 ou 2) et ces événements sont mutuellement exclusifs. La fonction de hasard instantané est alors découpée de la manière suivante :

$$\lambda(t,\psi) = \lambda_1(t,\psi_1) + \lambda_2(t,\psi_2)$$

on obtient donc pour la fonction de survie :

$$S(t\ ,\psi) = S_1(t,\psi_1)S_2(t,\psi_2)$$

avec $S_k(t,\psi_k) = \exp\left(-\int_0^t \lambda_k(u,\psi_k)du\right)$
En remplaçant ceci dans $C_k(t,\psi)$ on obtient donc :

$$C_k(t,\psi) = \int_0^t S_1(u,\psi_1)S_2(u,\psi_2)\lambda_k(u,\psi_k)du, \qquad k = 1;2$$

Un modèle de Weibull à deux paramètres est approprié pour $S_1(t,\psi_1), S_2(t,\psi_2)$ et $\lambda_k(t,\psi_k)$ ce qui donne :

$$\lambda_k(t,\kappa_k,\rho_k) = \kappa_k(\rho_k t)^{\kappa_k}/t$$

et

$$S_k(t,\kappa_k,\rho_k) = \exp(-(\rho_k t)^{\kappa_k})$$

Ceci ne nous donne pas une forme explicite de la fonction $C_k(t,\kappa_1,\rho_1,\kappa_2,\rho_2)$.
Les estimateurs sont présentés dans la section suivante.

4 Inférence paramètrique

La méthode du maximum de vraisemblance est utilisée pour faire de l'inférence.

4.1 Vraisemblance

Soit l'indicateur :

$$\delta_{ki} = \begin{cases} 1 \text{ si le ième patient est affecté par le kème événement} \\ 0 \text{ sinon.} \end{cases}$$

Les données sont donc représentées sous la forme : $(X_i, \delta_{1i}, \delta_{2i}), (i = 1, ..., n)$.

Si nos données ne comportaient pas de censures, la fonction de vraisemblance pour nos données serait :

$$L' = \prod_{i=1}^{n} f_1(x_i)^{\delta_{1i}} f_2(x_i)^{\delta_{2i}}$$

en notant $f_k(x) = dC_k(x)/dx$

Or avec la présence de données censurées indépendantes, il faut rajouter dans notre fonction de vraisemblance le cas des données censurées, on obtient donc :

$$L = \prod_{i=1}^{n} f_1(x_i)^{\delta_{1i}} f_2(x_i)^{\delta_{2i}} S(x_i)^{1-\delta_{1i}-\delta_{2i}}$$

en notant $f_k(x) = dC_k(x)/dx$ une fonction de densité potentiellement impropre.

Détaillons la vraisemblance dans les deux approches paramètriques étudiées.

• Approche du hasard spécifique :

Avec le modèle de Weibull, la vraisemblance associée à l'approche du hasard spécifique est :

$$\begin{aligned} L(\psi, x) &= \prod_{i=1}^{n} \left[S(x_i, \psi)\lambda_1(x_i, \psi_1)\right]^{\delta_{1i}} \left[S(x_i, \psi)\lambda_2(x_i, \psi_2)\right]^{\delta_{2i}} S(x_i, \psi)^{1-\delta_{1i}-\delta_{2i}} \\ &= \prod_{i=1}^{n} \lambda_1(x_i, \psi_1)^{\delta_{1i}} \lambda_2(x_i, \psi_2)^{\delta_{2i}} S_1(x_i, \psi_1) S_2(x_i, \psi_2) \end{aligned}$$

On remarque que $L(\psi, \mathbf{x})$ se factorise en deux parties , une qui dépend seulement de $\lambda_1(t, \psi_1)$ et l'autre qui dépend seulement de $\lambda_2(t, \psi_2)$. L'estimation de $C_1(t)$ et $C_2(t)$ peut donc se faire séparément.

• Approche directe :

Cette fois on travaille avec la fonction d'incidence cumulée donnée par le modèle de Gompertz $F(t, \psi)$ et non plus avec la fonction de hasard spécifique $\lambda_k(t, \psi_k)$. Comme la fonction de survie globale S(t) intervient dans notre vraisemblance, il faut la décomposer en fonction de $F_1(., \psi)$ et $F_2(., \psi)$. Comme $S(t) = 1 - F_1(., \psi) - F_2(., \psi)$ on obtient la vraisemblance suivante :

$$L(\psi, \mathbf{x}) = \prod_{i=1}^{n} f_1(x_i, \psi_1)^{\delta_{1i}} f_2(x_i, \psi_2)^{\delta_{2i}} [1 - F_1(x_i, \psi_1) - F_2(x_i, \psi_2)]^{1-\delta_{1i}-\delta_{2i}}$$

Contrairement au cas du hasard spécifique, on ne peut pas décomposer ici la vraisemblance comme un produit de $F_1(x, \psi_1)$ et $F_2(x, \psi_2)$. Les quatre paramètres doivent être estimés simultanément. Notons que ceci a pour conséquence d'amplifier les erreurs du modèles, en effet une mauvaise paramétrisation sur $F_1(x, \psi_1)$ aura des répercussions sur $F_2(x, \psi_2)$ et vice versa.

4.2 Estimateur du maximum de vraisemblance (EMV)

Nous allons maintenant nous servir de l'EMV et de ses propriétés afin de déterminer des intervalles de confiances ponctuels pour nos estimateurs de la fonction d'incidence cumulée spécifique au risque k pour nos deux approches.

• Approche hasard spécifique :

Comme on peut faire l'estimation séparément, la log-vraisemblance partielle est donnée par :

$$\ell_k(\psi_k) = \sum_{i=1}^{n} \left[\delta_{ki} \log(\lambda_k(x_i, \psi_k)) + \log(S_k(x_i, \psi_k))\right], \quad k = 1; 2$$

Après avoir dérivé cette expression on peut déterminer le maximum de la vraisemblance à l'aide d'un algorithme d'optimisation (l'algorithme de Nelder-Mead est utilisé dans la partie simulation).
Ceci nous permet de déterminer $\widehat{\psi} = (\widehat{\psi_1}, \widehat{\psi_2})$ l'EMV de $\psi = (\psi_1, \psi_2)$. On pose ensuite :

$$\widehat{C_k}(t, \hat{\psi}) = \int_0^t S_1(u, \hat{\psi_1}) S_2(u, \hat{\psi_2}) \lambda_k(u, \hat{\psi_k}) du, \quad k = 1; 2$$

Pour que le théorème central limite pour l'EMV s'applique, il faut que le modèle soit régulier. Dans le cas présent nous avons choisi pour modèle des distributions de Gompertz et de Weibull qui sont tous deux des modèles réguliers. Comme $\widehat{C_k}(t, \hat{\psi})$ est une fonction homogène de $\hat{\psi}$ il est consistant et assymptotiquement Gaussien de variance asymptotique estimable par la delta-methode d'après le théorème central limite appliqué à l'EMV on a :

$$\sqrt{n}(\hat{\psi_n} - \psi) \overset{n\to\infty}{\longrightarrow} U \sim \mathcal{N}(0, I(\psi)^{-1})$$

avec $I(\psi)^{-1}$, l'inverse de la matrice d'information de Fisher.

Pour trouver la variance asymptotique de l'EMV il faut determiner l'inverse de la matrice Hessienne du log de la vraisemblance et l'évaluer en $\hat{\psi}$. $\widehat{\mathrm{Var}}(\hat{\psi}) = I(\hat{\psi}^{-1})$ Utilisont la delta-method pour le cas multivarié qui dit que :

$$\sqrt{n}\left(g(\widehat{\psi}_n - g(\psi)\right) \overset{n\to\infty}{\longrightarrow} Dg(\psi).U$$

avec g dérivable et non nulle. Appliquons ceci avec $g(\cdot) = C_k(\cdot)$ on a :

$$\sqrt{n}\left(C_k(\widehat{\psi}_n, t) - C_k(\psi, t)\right) \overset{n\to\infty}{\longrightarrow} DC_k(\psi, t).U$$

Or on calcul :

$$\mathbb{E}(DC_k(\psi, t).U) = DC_k(\psi, t).\mathbb{E}(U) = 0$$

et

$$\mathrm{Var}(DC_k(\psi, t).U) = DC_k(\psi, t).\mathrm{Var}(U).DC_k(\psi, t)' \quad \text{(où ' désigne la transposée)}$$

On obtient donc :

$$\widehat{\mathrm{Var}(\widehat{C_k(\hat{\psi},t)})} = DC_k(\psi,t)\big|_{\psi_k=\hat{\psi_k}}.\widehat{\mathrm{Var}(\hat{\psi})}.DC_k(\psi,t)\big|'_{\psi_k=\hat{\psi_k}}, \quad (k=1;2)$$

Avec $DC_k(\psi,t)$ la matrice de l'application linéaire tangente à C_k de la fonction d'incidence cumulée spécifique au risque k en fonction des paramètres.

Donc $\widehat{C_k}(\hat{\psi},t)$ suit une loi normale $\mathcal{N}(0,\widehat{\mathrm{Var}(\widehat{C_k}(\hat{\psi},t))}$. On en déduit un intervalle de confiance ponctuel à 95% pour $C_k(t,\psi)$ donné par

$$\widehat{C_k}(t,\hat{\psi}) \pm 1,96\sqrt{\widehat{\mathrm{Var}(\widehat{C_k}(\hat{\psi},t))}}$$

• Approche directe :

L'approche directe est similaire à l'approche indirecte.
On a donc :

$$\ell_k(\psi_k) = \sum_{i=1}^{n} \delta_{1i}\log[f_1(x_i,\psi_1)] + \delta_{2i}\log[f_2(x_i,\psi_2)] + (1-\delta_{1i}-\delta_{2i})\log[1-F_1(x_i,\psi_1)-F_2(x_i,\psi_2)]$$

et comme précédemment pour un instant t fixé quelconque, on obtient :

$$\widehat{\mathrm{Var}(\widehat{F_k}(\hat{\psi},t))} = DF_k(\psi,t)|_{\psi_k=\hat{\psi_k}}.\widehat{\mathrm{Var}(\hat{\psi})}.DF_k(\psi,t)|'_{\psi_k=\hat{\psi_k}}, \quad (k=1;2)$$

On en conclut que $\widehat{F_k,t}(\hat{\psi})$ suit une loi normale $\mathcal{N}(0,\widehat{\mathrm{Var}(\widehat{F_k,t}(\hat{\psi}_k))}$ et qu'un intervalle de confiance à 95% pour $F_k(t,\psi_k)$ est donné par :

$$\widehat{F_k}(t,\hat{\psi}_k) \pm 1,96\sqrt{\widehat{\mathrm{Var}(\widehat{F_k}(\hat{\psi}_k))}}$$

5 Simulations

5.1 Planification des simulations

Le but est de simuler les données de risques concurrents selon les différents modèles avec un taux de hasard spécifique à une cause donnée fixé.

Scénario testés :

Le premier scénario suit l'approche directe, on simule des données selon un modèle de Gompertz avec $\alpha_1 = -0.01$, $\alpha_2 = -0.02$, $\beta_1 = 0.1$ et $\beta_2 = 0.2$.
On a

$$\lambda_1(t,\alpha_1,\beta_1) = \beta_1\exp(\alpha_1 t) = 0.1\exp(-0.01x)$$

et

$$\lambda_2(t,\alpha_2,\beta_2) = \beta_2\exp(\alpha_2 x) = 0.2\exp(-0.02x)$$

Le deuxième scénario suit l'approche indirecte, les données sont simulé selon un modèle de Weibull avec $\kappa_1 = 0.95$, $\kappa_2 = 0.7$, $\rho_1 = 0.6$ et $\rho_2 = 0.3$

$$\lambda_1(t, \kappa_1, \rho_1) = \frac{\kappa_1(\rho_1 t)^{\kappa_1}}{t} = \frac{0.95(0.6t)^{0.95}}{t}$$

et

$$\lambda_2(t, \kappa_2, \rho_2) = \frac{\kappa_2(\rho_2 t)^{\kappa_2}}{t} = \frac{0.7(0.3t)^{0.7}}{t}$$

Pour chacun de ces modèles on définit trois taux de censure : 10%, 25% et 50%.

Les variables aléatoires de censure suivent une loi exponentielle de paramètre $\lambda_{c_{j,k}}$, où j représente le scénario et k le taux de censure voulut. Par le biais de simulations d'echantillons de T_i et de C_i on détermine les $\lambda_{c_{j,k}}$. Les resultats sont les suivants :
$\lambda_{c_{1,10\%}} = 0.025$ $\lambda_{c_{1,25\%}} = 0.085$ $\lambda_{c_{1,50\%}} = 0.22$ $\lambda_{c_{2,10\%}} = 0.7$ $\lambda_{c_{2,25\%}} = 1.9$ $\lambda_{c_{2,50\%}} = 4.8$
La taille des échantillons reste fixe, avec n=100.
Enfin on utilise la méthode de Monte-Carlo pour répéter ces simulations avec M=50.

Le problème de ces simulations vient du fait de la dépendance entre les risques concurrents. Afin de générer des simulations où λ_1 et λ_2 sont interdépendants on utilise la méthode d'inversion de Bender et. al (2005).

Le principe est le suivant, on définie une variable aléatoire U_i suivant une loi uniforme[0,1] on à le résultat suivant :

$$U_i = exp(-\Lambda(t)) \quad \Leftrightarrow \quad T_i = \Lambda^{-1}(-\ln U_i)$$

Malheuresement nous ne sommes pas en mesure de trouver Λ^{-1} de manière analytique, on utilise donc le package de R :

`Rootsolve`

et plus particulièrement la fonction

`uniroot.all`

afin de déterminer les racines de $\Lambda(T_i) + logU_i = 0$.
La cause spécifique k_i pour un individu i donné est donné selon K_i suivant une loi de Bernouilli tel que :

$$\mathbb{P}(K_i = 1) = \frac{\lambda_1(t_i)}{\lambda(t_i)}$$

$$\mathbb{P}(K_i = 2) = \frac{\lambda_2(t_i)}{\lambda(t_i)}$$

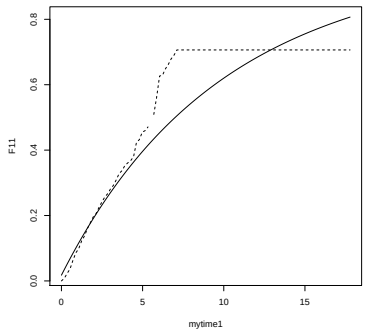
On dispose donc tous les éléments afin de créer notre jeu de données aléatoire $(X_i, \delta_{1i}, \delta_{2i})$

Afin de determiner les estimateurs des paramètres, on utilise un algorithme d'optimisation qu'on applique sur les log-vraisemblances des différentes distributions. Dans le cas présent l'algorithme de Nelder-Mead est utilisé pour sa robustesse grâce à la fonction

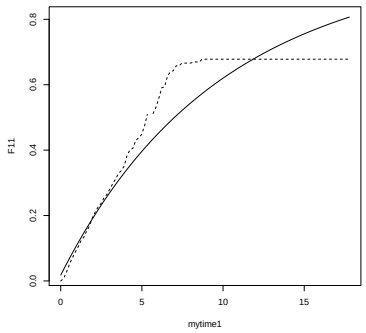
`optimix`

de la librairie du même nom.

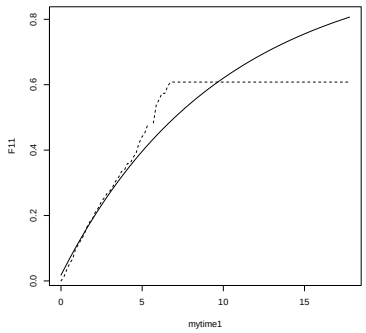
Enfin on créé deux fonctions par modèle permettant de calculer $\hat{F}_k(t)$ et $\hat{C}_k(t)$ pour k=1 ;2 et on les applique sur M jeux de données dans notre algorithme de Monte-Carlo.



(a) Estimateur de $\hat{F}_1(t)$ avec 10% de censure

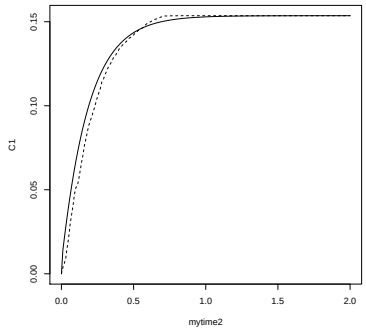


(b) [Estimateur de $\hat{F}_1(t)$ avec 25% de censure

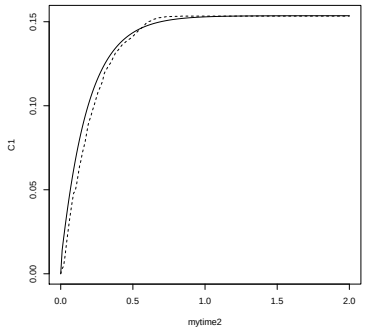


(c) [Estimateur de $\hat{F}_1(t)$ avec 50% de censure

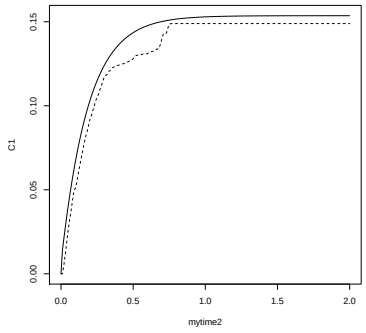
FIGURE 2 – Comparaison des estimateurs de $\hat{F}_1(t)$ avec $F_1(t)$ selon différents niveaux de censure



(a) Estimateur de $\hat{C}_1(t)$ avec 10% de censure



(b) [Estimateur de $\hat{C}_1(t)$ avec 25% de censure



(c) [Estimateur de $\hat{C}_1(t)$ avec 50% de censure

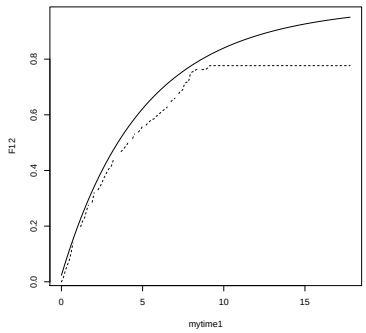
FIGURE 3 – Comparaison des estimateurs de $\hat{C}_1(t)$ avec $C_1(t)$ selon différents niveaux de censure

5.2 Résultats

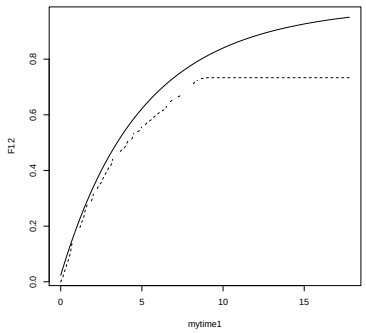
Les différents résultats sont présentés sur les figures 2, 3, 4 et 5.

5.3 Discussion

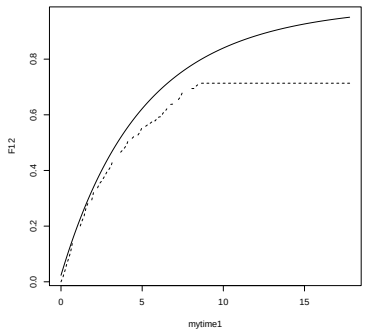
Au vu des graphiques on peut constater que pour les estimateurs $\hat{C}_k(t)$ et $\hat{F}_k(t)$, représentés en lignes discontinues sur les figures 2, 3, 4 et 5, la censure à une influence sur le biais à droite du support. En effet un constate un écart entre les courbes pleines représentants les fonctions d'incidences cumulées et les courbes discontinues représentants leurs estimateurs réspectifs. On note de plus que ce biais augmente lorsque la censure augmente.
Par ailleur on remarque, pour les estimateurs $\hat{C}_k(t)$ de $C_k(t)$,que des données présentant un grand taux de censure (50%) affecte également les estimateurs pour des durées plus faibles.
Enfin on remarque une différence notable entre les estimateurs $\hat{F}_k(t)$ par rapport à la vraie fonction d'incidence cumulée spécifique à la cause 1 et à la cause 2. Ceci est principalement dû au fait que l'algorithme d'optimisation de Nelder-Mead utilisé pour estimer les paramètres $\hat{\alpha_1}$, $\hat{\alpha_2}$, $\hat{\beta}_1$ et $\hat{\beta}_2$ du modèle de Gompertz ne permet pas de trouver de bons estimateurs avec le scénario proposé ici, d'où un biais provenant potentiellement des estimateurs $\hat{\psi}$ de ψ.



(a) Estimateur de $\hat{F}_2(t)$ avec 10% de censure

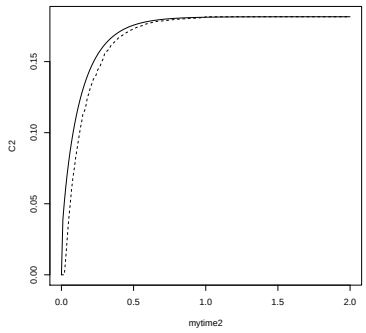


(b) [Estimateur de $\hat{F}_2(t)$ avec 25% de censure

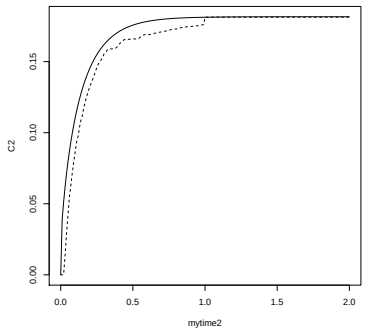


(c) [Estimateur de $\hat{F}_1(t)$ avec 50% de censure

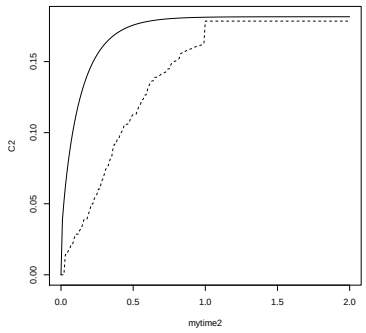
FIGURE 4 – Comparaison des estimateurs de $\hat{F}_2(t)$ avec $F_2(t)$ selon différents niveaux de censure



(a) Estimateur de $\hat{C}_2(t)$ avec 10% de censure



(b) [Estimateur de $\hat{C}_2(t)$ avec 25% de censure



(c) [Estimateur de $\hat{C}_2(t)$ avec 50% de censure

FIGURE 5 – Comparaison des estimateurs de $\hat{C}_2(t)$ avec $C_2(t)$ selon différents niveaux de censure

Annexes

Nous donnerons les principaux bouts de code ayant permis les simulations :

```
 #-------------------------------------------------------------
 # génération aléatoire des échantillons sous un modèle de risques concurrents indépendamm
 # selon un taux de hasard spécifique à une cause donnée fixé et paramétré par kappa1,roh1
 # n= taille de l'échantillon
 # lambdac = parametre de la loi de censure
 #-------------------------------------------------------------
mysimu2<-function(n,kappa1=k1,beta1=b1,kappa2=k2,beta2=b2,lambdac=c250)
{
ui2<-runif(n)
Ti2<-rep(0,n)
Ki2<-rep(0,n)
delta1i2<-rep(0,n)
delta2i2<-rep(0,n)
for(i in 1:n)
 {
    aux<-function(t,kappa1=k1,roh1=p1,kappa2=k2,roh2=p2,Ui=ui[i])
       {
         M2Lambda(t,kappa1,kappa2,roh1,roh2)+log(Ui) #Avec M2Lambda
# une fonction qui cacule le taux de hasard cumulé
       }
    T2<-uniroot.all(aux,lower=0,upper=1000000,kappa1=k1,kappa2=k2,roh1=p1,roh2=p2,
Ui=ui2[i])
    Ti2[i]<-T2
    #initialisation des Ki pour (Ti,Ki)
    p1i2<-M2lambda1(Ti2)/M2lambda(Ti2)
    p2i2<-M2lambda2(Ti2)/M2lambda(Ti2)
    Ki2[i]=rbinom(1,1,p2i2)+1 #1 si événement 1, 2 si événement 2
    delta2i2[i]=Ki2[i]-1
    delta1i2[i]=abs(delta2i2[i]-1)
 }
#initialisation des Ci
Ci2<-rexp(n,lambdac) #25
P<-mean((Ti2<=Ci2)==TRUE)
#matrice du couple (Ti,Ci) afin de trouver Xi
M2<-matrix(c(Ti2,Ci2),nrow=n,ncol=2)
#Xi:
Xi2<-apply(M2,1,min)
#Finalisation des donées:
Donnees<-matrix(c(Xi2,delta1i2,delta2i2),nrow=n,ncol=3)
}

#-------------------------------------------------
# calcul de l'EMV maximum de vraisemblance de  psi
#-------------------------------------------------
init1<-c(0.5,0.5) # vecteur des paramètres initiaux pour la fonction optimix
```

```
init2<-c(0.5,0.6)
t1<-optimx(init1,fn=llh1,hessian=FALSE,mydata=donnees,method=c("Nelder-Mead"))
t2<-optimx(init2,fn=llh2,hessian=FALSE,mydata=donnees,method=c("Nelder-Mead"))
psychapM2<-c(t1$par$par,t2$par$par)

#Fonction qui génère un estimateur pour C1 en fonction de données et de leur nombre
M2myestim1<-function(N,mat=donnees)
  {
 j<-1
  estim1<-matrix(c(0,0),nrow=1,ncol=2)
    for(i in 1:N)
  {
  if(mat[i,2]==1)
 {
 estim1<-rbind(estim1,mat[i,1])
 j<-j+1
 estim1[j,2]<-C1chap(mat[i,1],psychapM2)
 }
  }
  estim1<-estim1[do.call("order", as.data.frame(estim1)),]
  estim1
  }
#---------------------------------------
#Monte-Carlo
#---------------------------------------
M<-50
p<-200 #nombre de pas pour la discrétisation de l'abscisse
B1<-2.12 # ce nombre correspond au max des Ti sur 10000 données (pour T1....10000)
mytime2<-seq(0,B2,B2/p)
C2110mat<-matrix(rep(NA,(p+1)*M),nrow=p+1,ncol=M) #Exemple avec C1 pour 10\% de censure
 for(m in 1:M)
 {
mydatac210<-mysimu2(n,k1,p1,k2,p2,c210)
myestimation2110<-M2myestim1(n,mydatac210)
for (t in seq(0,B2,B2/p))
{
for(i in 1:(length(myestimation2110[,1])-1))
{
if((t>=myestimation2110[i,1]) && (t<myestimation2110[i+1,1]))
C2110mat[t*p/B2+1,m]<-myestimation2110[i,2]
}
for(j in (length(myestimation2110[,1])-1):p+1)
{
if(j<p+2)
{
C2110mat[j,m]<-myestimation2110[length(myestimation2110[,1]),2]
}
}
```

```
}
}
C2110<-apply(C2110mat,1,mean)
plot(mytime2,sort(C2110),type="l")
```

A noter que la définition des fonctions de hasards instanés, hasards instanés de cause spécifique, hasards cumulés et les fonctions de log-vraisemblances, ainsi que le paramètres diffèrent entre les modèles mais l'implémentation reste la même.

Bibliographie

Jong-Hyeong Jeong et Jason Fine (2005),Direct parametric inference for the cumulative incidence function,*Applied statistics* **55**(2), 187-200

J. Bryant et J.J. Dignam (2004), Semiparametric models for cumulative incidence functions, *Biometrics*,**60**, 182-190

R. Bender,T. Augustin et M. Blettner (2005), Generating survival times to simulate cox proportional hazards models, *Statistics in medicine*,**24**, 1713-1727

J. Beyersmann, A. Latouche, A. Buchholz et M. Schumacher (2009), Simulating competing risks data in survival analysis, *Statistics in medicine*, **28**, 956-971

D.Y. Lin (1997),non-parametric inference for cumulative incidence functions in competing risks studies, *Staistics in medicine*, **16**, 901-910