Fonction de survie bivariée de variables censurées à droite et à gauche

Philippe Saint-Pierre, Agathe Guilloux

Fonction de survie bivariée de variables censurées à droite et à gauche

Philippe Saint-Pierre & Agathe Guilloux

Université Pierre et Marie Curie - Paris 6
Laboratoire de Statistique Théorique et Appliquée (LSTA)
175 rue du Chevaleret, 8ème étage, bâtiment A
75013 PARIS, FRANCE
philippe.saint_pierre@upmc.fr

Résumé
L'objet de ce travail est l'estimation de la fonction de répartition (f.d.r.) bivariée dans le cas où une variable est censurée à gauche et l'autre censurée à droite. Cette question ne semble pas être étudiée dans la littérature alors que plusieurs travaux traitent du cas où les deux variables sont censurées à droite. La méthode proposée s'inspire d'une approche permettant, dans le cas bivarié avec deux censures à droite, d'exprimer la f.d.r. comme un produit intégral de la fonction de hasard cumulée. Un estimateur non paramétrique de la f.d.r. est obtenu en utilisant les équations intégrales de Doléans et de Volterra. La convergence presque sûre de l'estimateur ainsi que les propriétés de la somme des deux variables censurées sont étudiées. Des simulations seront présentées pour illustrer la méthodologie.

Abstract
The purpose of this work is the estimation of the bivariate cumulative distribution function (c.d.f.) when one variable is left-censored and the other one is right-censored. This issue does not seem to be studied in the literature while several studies deal with the case where the two variables are right-censored. The representation of the bivariate survival function as a product integral of the cumulative hazard function, obtained in the right censoring case, is generalized to left and right censoring case using Doléans equation. A nonparametric estimator of c.d.f. is obtained and the strong consistency of the estimator is investigated. Moreover, the properties of the sum of the two censored variables are studied. Simulations are presented to illustrate the methodology.

Mots-clés
Fonction de survie bivariée, censures à droite et à gauche, équation de Doléans, équation de Volterra, estimateur non paramétrique.

1 Introduction

Soit $T = (T_1, T_2)$ un couple de variables aléatoires (v.a.) à valeurs dans $\mathbb{R}_+^2$. On note F la fonction de répartition (f.d.r.) de T. On considère que la v.a. T est censurée à gauche

et à droite dans le sens où l'on observe les v.a. Y et δ définies par :

$$Y = (Y_1, Y_2) \text{ avec } Y_1 = T_1 \vee L \text{ et } Y_2 = T_2 \wedge C$$
$$\delta = (\delta_1, \delta_2) \text{ avec } \delta_1 = I(T_1 \geq L) \text{ et } \delta_2 = I(T_2 \leq C),$$

où (L, C) est un vecteur de v.a. positives de f.d.r. G, supposé indépendant du vecteur (T_1, T_2). La v.a. T_1 est donc censurée à gauche alors que T_2 est censurée à droite. On cherche à estimer la f.d.r. à partir d'un échantillon i.i.d. de la forme $(Y_i = (Y_{1,i}, Y_{2,i}), \delta_i = (\delta_{1,i}, \delta_{2,i}))$ pour $i = 1, \ldots, n$.
Par exemple, dans le cas d'un patient infecté par le VIH, cette fonction pourra permettre d'étudier conjointement (i) la durée T_1 écoulée entre la consultation du patient infecté et la date de l'infection par le VIH (cette durée peut être censurée à gauche dans le cas où le patient sait uniquement qu'il n'était pas infecté à une certaine date) (ii) la durée T_2 écoulée entre la consultation et le début du SIDA (cette durée peut être censurée à droite, si le patient ne connaît pas la date exacte de son infection mais celle du résultat négatif d'un test antérieur de dépistage).
Pour autant que nous sachions, ce problème n'a pas encore été considéré dans la littérature. Dabrowska (1988) a construit un estimateur de la f.d.r bivariée F en considérant que les v.a. T_1 et T_2 étaient toutes deux censurées à droite. Le problème de l'estimation de la f.d.r. univariée sous censure aléatoire à droite a été intensivement étudié (Andersen et al. (1993)). Le même problème mais sous censure à gauche a été beaucoup moins étudié. La plupart des auteurs (Cox et Oakes (1984), Andersen et al. (1993)) proposent de "renverser" le temps. Cette approche n'est par entièrement satisfaisante puisqu'elle exclut le cas où les v.a. T_1 (et Y_1) sont à support sur $\mathbb{R}_+$. Une autre approche, proposée par Gómez et al. (1992), consiste à écrire la f.d.r. F_1 de la v.a. T_1 comme la solution de l'équation intégrale de Doléans :

$$F_1(t) = F_1(\infty) - \int_t^\infty F_1(u) \frac{dH_1(u)}{H(u)} \text{ pour } t \geq 0$$
$$F_1(\infty) = 1$$

où H est la f.d.r. de la v.a. Y_1 et H_1 est définie par $H_1(t) = \mathbb{P}(Y_1 \leq t, \delta_1 = 1)$. Gómez et al. (1992) ont montré que la solution de cette équation est donnée, pour $t \geq 0$, par:

$$F_1(t) = \prod_{s>t} \left(1 - \frac{dH_1(s)}{H(s)}\right),$$

où la notation $\prod$ désigne le produit intégral.

2 Loi bivariée et fonctions de risque cumulé

Notons $\tilde{F}$ et $\tilde{G}$ les fonctions définies pour tout $(s,t) \in \mathbb{R}_+$ par :

$$\begin{aligned} \tilde{F}(s,t) &= \mathbb{P}\left(T_1 \leq s, T_2 > t\right) \text{ et} \\ \tilde{G}(s,t) &= \mathbb{P}\left(L \leq s, C > t\right). \end{aligned}$$

L'objectif est d'estimer la fonction $\tilde{F}$. Par définition des v.a. Y et δ et par indépendance entre (T_1, T_2) et (L, C), les fonctions $\tilde{H}$, K_1, K_2 et K_3, définies ci-dessous, vérifient les relations suivantes pour tout $(s,t) \in \mathbb{R}_+$:

$$\begin{aligned} &\tilde{H}(s,t) = \mathbb{P}\left(Y_1 \leq s, Y_2 > t\right) = \tilde{F}(s,t)\tilde{G}(s,t) \qquad (1) \\ &K_1(s,t) = \mathbb{P}\left(Y_1 \leq s, Y_2 > t, \delta_1 = 1, \delta_2 = 1\right) = -\int_0^s \int_t^\infty \tilde{G}(u, v-)\tilde{F}(du, dv) \\ &K_2(s,t) = \mathbb{P}\left(Y_1 \leq s, Y_2 > t, \delta_1 = 1\right) = \int_0^s \tilde{G}(u,t)\tilde{F}(du,t) \\ &K_3(s,t) = \mathbb{P}\left(Y_1 \leq s, Y_2 > t, \delta_2 = 1\right) = -\int_t^\infty \tilde{G}(s, v-)\tilde{F}(s, dv). \end{aligned}$$

Grâce aux relations précédentes, on peut définir les fonctions de risque cumulé:

$$\begin{aligned} \Lambda_{11}(s,t) &= -\int_0^s \int_0^t \frac{\tilde{F}(du,dv)}{\tilde{F}(u,v-)} = -\int_0^s \int_0^t \frac{K_1(du,dv)}{\tilde{H}(u,v-)} \qquad (2) \\ \Lambda_{10}(s,t) &= \int_0^s \frac{\tilde{F}(du,t)}{\tilde{F}(u,t)} = \int_0^s \frac{K_2(du,t)}{\tilde{H}(u,t)} \\ \Lambda_{01}(s,t) &= -\int_0^t \frac{\tilde{F}(s,dv)}{\tilde{F}(s,v-)} = -\int_0^t \frac{K_3(s,dv)}{\tilde{H}(s,v-)}. \end{aligned}$$

Les fonctions Λ_{11}, Λ_{10}, Λ_{01} sont donc estimables à l'aide des équivalents empiriques des fonctions $\tilde{H}$, K_1, K_2 et K_3. Ces versions empiriques sont calculables puisque les fonctions théoriques sont exprimées en fonction des v.a. Y et δ qui sont observables. Il reste donc à lier la fonction $\tilde{F}$ aux fonctions Λ_{11}, Λ_{10}, Λ_{01}. Définissons, pour (s,t) tels que $\tilde{F}(s,t) > 0$, la fonction A par $A(s,t) = \log \tilde{F}(s,t)$, alors on a :

$$\begin{aligned} \tilde{F}(s,t) &= \exp\{A(s,t)\} \\ &= \exp\{-\int_s^\infty \int_0^t A(du,dv) + A(\infty,t) + A(s,0) - A(\infty,0)\} \\ &= \tilde{F}(\infty,t)\tilde{F}(s,0)\exp\{-\int_s^\infty \int_0^t A(du,dv)\}. \end{aligned}$$

Considérons les ensembles suivant,

$$\begin{aligned}
E_1 &= \{(s,t) : A(s,t) < 0, A(\Delta s, t) = A(s, \Delta t) = 0\} \\
E_2 &= \{(s,t) : A(s,t) < 0, A(\Delta s, t) < 0, A(\Delta s, \Delta t) = 0\} \\
E_3 &= \{(s,t) : A(s,t) < 0, A(s, \Delta t) < 0, A(\Delta s, \Delta t) = 0\} \\
E_4 &= \{(s,t) : A(s,t) < 0, A(\Delta s, \Delta t) > 0\}.
\end{aligned}$$

Les ensembles E_1 et E_4 correspondent au support, respectivement, de la composante continue et de la composante discrète de A. Les ensembles E_2 et E_3 sont les supports des composantes de A qui ont des discontinuités appartenant aux lignes orthogonales aux axes des coordonnées. La fonction $A = \log \tilde{F}$ étant une fonction à variation bornée, on peut utiliser la décomposition de Jordan des fonctions à variation bornée dans le plan:

$$\begin{aligned}
-\int_s^\infty \int_0^t A(du, dv) &= \sum_{i=1}^4 A_i(s,t), \text{ où} \\
A_1(s,t) &= -\int_s^\infty \int_0^t I((u,v) \in E_1) A(du, dv) \\
A_2(s,t) &= -\sum_{u>s} \int_0^t I((u,v) \in E_2) A(\Delta u, dv) \\
A_3(s,t) &= -\sum_{v \le t} \int_s^\infty I((u,v) \in E_3) A(du, \Delta v) \\
A_4(s,t) &= -\sum_{u>s} \sum_{v \le t} I((u,v) \in E_4) A(\Delta u, \Delta v).
\end{aligned}$$

En exprimant les quantités précédentes en fonction des risques cumulés, on peut montrer le résultat suivant:

Proposition 1 *Pour tout (s,t) tel que $\tilde{F}(s,t) > 0$,*

$$\tilde{F}(s,t) = \tilde{F}(\infty, t)\tilde{F}(s,0) \prod_{i=1}^4 \exp\{A_i(s,t)\}$$

avec

$$\begin{aligned}
\exp\{A_i(s,t)\} &= \exp\left\{-\int_s^\infty \int_0^t I((u,v) \in E_i) L(du, dv)\right\}, \text{ pour } i = 1,2,3, \\
\exp\{A_4(s,t)\} &= \prod_{\substack{u>s \\ (u,v)\in E_4}} \prod_{v \le t} (1 - L(\Delta u, \Delta v)) \\
L(du, dv) &= \frac{\Lambda_{10}(du, v-)\Lambda_{01}(u, dv) - \Lambda_{11}(du, dv)}{[1 - \Lambda_{10}(\Delta u, v-)][1 - \Lambda_{01}(u, \Delta v)]}
\end{aligned}$$

et

$$\tilde{F}(s,0) = \prod_{u>s} (1 - \Lambda_{10}(du,0))$$
$$\tilde{F}(\infty,t) = \prod_{v\leq t} (1 - \Lambda_{01}(\infty,dv)).$$

3 Estimation

Dans ce paragraphe nous définissons un estimateur de $\tilde{F}$ grâce à la proposition 1 et montrons sa consistance forte. On observe un échantillon i.i.d. de la forme

$$Y_i = (Y_{1,i}, Y_{2,i}) \text{ et } \delta_i = (\delta_{1,i}, \delta_{2,i}) \text{ pour } i = 1, \ldots, n.$$

Considérons les estimateurs empiriques $\widehat{\tilde{H}}$, $\widehat{K_1}$, $\widehat{K_2}$ et $\widehat{K_3}$ des fonctions $\tilde{H}$, K_1, K_2 et K_3. On peut alors définir les estimateurs des fonctions de risque cumulé pour tout $(s,t) \in \mathbb{R}_+^2$ par :

$$\widehat{\Lambda_{11}}(s,t) = -\int_0^s \int_0^t \frac{\widehat{K_1}(du,dv) I\left(\widehat{\tilde{H}}(u,v-) > 0\right)}{\widehat{\tilde{H}}(u,v-)},$$
$$\widehat{\Lambda_{10}}(s,t) = \int_0^s \frac{\widehat{K_2}(du,t) I\left(\widehat{\tilde{H}}(u,t) > 0\right)}{\widehat{\tilde{H}}(u,t)} \text{ et}$$
$$\widehat{\Lambda_{01}}(s,t) = -\int_0^t \frac{\widehat{K_3}(s,dv) I\left(\widehat{\tilde{H}}(s,v-) > 0\right)}{\widehat{\tilde{H}}(s,v-)},$$

avec la convention $0/0 = 0$. Nous sommes alors en mesure de définir l'estimateur de la fonction L par

$$\widehat{L}(\Delta u, \Delta v) = \frac{\widehat{\Lambda_{10}}(\Delta u, v-)\widehat{\Lambda_{01}}(u,\Delta v) - \widehat{\Lambda_{11}}(\Delta u, \Delta v)}{[1 - \widehat{\Lambda_{10}}(\Delta u, v-)][1 - \widehat{\Lambda_{01}}(u,\Delta v)]}.$$

Comme la fonction $\widehat{L}$ est purement discrète par construction, on obtient finalement l'estimateur $\widehat{\tilde{F}}$ de la fonction $\tilde{F}$ en posant :

$$\widehat{\tilde{F}}(s,t) = \widehat{\tilde{F}}(\infty,t)\widehat{\tilde{F}}(s,0) \prod_{u>s, 0<v\leq t} (1 - \widehat{L}(du,dv)), \tag{3}$$

où $t \mapsto \widehat{\tilde{F}}(\infty,t)$ est l'estimateur de Kaplan-Meier associé à l'observation de $(Y_{2,i}, \delta_{2,i})$ pour $i = 1, \ldots, n$, enfin $t \mapsto \widehat{\tilde{F}}(s,0)$ est l'estimateur de Kaplan-Meier dans le cas de censure à gauche (Gómez et al. (1992)) associé à l'observation de $(Y_{1,i}, \delta_{1,i})$ pour $i = 1, \ldots, n$.

On peut ensuite établir la consistance forte de l'estimateur $\widehat{\tilde{F}}$. Pour $\tau = (\tau_1, \tau_2)$, notons $\|\cdot\|_\tau$ dénote la norme sup sur $[\tau_1, \infty] \times [0, \tau_2]$.

Proposition 2 *Supposons que le vecteur de v.a.* (L, C) *est indépendant du vecteur* (T_1, T_2) *et que* $\tau = (\tau_1, \tau_2)$ *vérifie* $\tilde{H}(\tau_1, \tau_2) > 0$*, alors*

$$\| \widehat{\tilde{F}} - \tilde{F} \|_\tau \stackrel{p.s.}{\to} 0.$$

Cette proposition découle de résultats obtenus en utilisant des arguments similaires à ceux de Dabrowska (1988) et de la consistance forte des estimateurs de Kaplan-Meier pour la censure à droite et à gauche.

Les détails des calculs et des preuves peuvent être trouvés dans Guilloux et Saint-Pierre (2008).

Les propriétés de la somme des deux variables censurées, l'une à droite et l'autre à gauche, ainsi que des simulations seront également présentées.

Bibliographie

[1] Andersen, P. K., Borgan, Ø., Gill, R. D., et Keiding, N. (1993) *Statistical models based on counting processes*, Springer-Verlag, New York.
[2] Cox D. R. et Oakes, D. (1984) *Analysis of survival data*, Monographs on Statistics and Applied Probability, Chapman & Hall, London.
[3] Dabrowska D. M. (1988) Kaplan-Meier estimate on the plane. *Ann. Statist.*, 16(4) :1475-1489.
[4] Gómez, G., Julià, O. et Utzet, F. (1992) Survival analysis for left censored data. *Survival analysis : state of the art*, volume 211 of NATO Adv. Sci. Inst. Ser. E Appl. Sci., pages 269-288, Kluwer Acad. Publ., Dordrecht.
[5] Hougaard, P. (2000) *Analysis of multivariate survival data*, Statistics for Biology and Health, Springer-Verlag, New York.
[6] Guilloux A. et Saint-Pierre P. (2008) Estimateur de la fonction de répartition bivariée avec censures à droite et à gauche. *Annales de l'ISUP*, 19(2) :157-164.