Available online at www.sciencedirect.com
ScienceDirect
C. R. Acad. Sci. Paris, Ser. I 344 (2007) 457–460
COMPTES RENDUS MATHEMATIQUE
http://france.elsevier.com/direct/CRASS1/

Statistique/Probabilités

Estimation dans un modèle de risques concurrents éventuellement dépendants en présence de censure

Ségolen Geffray

LSTA, université Paris 6, 175, rue du Chevaleret, 75013 Paris, France

Reçu le 6 septembre 2005 ; accepté après révision le 14 février 2007

Disponible sur Internet le 21 mars 2007

Présenté par Paul Deheuvels

Résumé

Nous considérons une population soumise à $\mathcal{J}$ risques concurrents éventuellement dépendants. A chaque individu est associé un couple de variables aléatoires $(X, \mathcal{C})$. La durée de vie X est une variable positive et la cause de mort $\mathcal{C}$ prend la valeur j lorsque la mort est due à la j^{e} cause pour un j dans $\{1, \dots, \mathcal{J}\}$. Ce couple $(X, \mathcal{C})$ est censuré à droite par une v.a. positive C indépendante de $(X, \mathcal{C})$. Nous étudions les propriétés asymptotiques de l'estimateur de Aalen–Johansen $\widehat{F}_n^{(j)}$ de la fonction de répartition spécifique à la j^{e} cause $F^{(j)}(\cdot) = \mathbb{P}[X \leqslant \cdot, \mathcal{C} = j]$. ***Pour citer cet article : S. Geffray, C. R. Acad. Sci. Paris, Ser. I 344 (2007).***

Abstract

Estimation in a model of possibly dependent competing risks with censorship. We consider a population which is submitted to $\mathcal{J}$ competing causes of failure which are possibly dependent. To each individual is associated a couple of random variables $(X, \mathcal{C})$. The failure time X is non-negative and the cause of failure $\mathcal{C}$ takes a value j in $\{1, \dots, \mathcal{J}\}$ when the failure is due to the j-th cause. This couple of r.v. $(X, \mathcal{C})$ is independently right-censored by a non-negative r.v. C. We study the asymptotic properties of the Aalen–Johansen estimator $\widehat{F}_n^{(j)}$ of the subdistribution function $F^{(j)}(\cdot) = \mathbb{P}[X \leqslant \cdot, \mathcal{C} = j]$. ***To cite this article: S. Geffray, C. R. Acad. Sci. Paris, Ser. I 344 (2007).***

1. Introduction

Considérons une population exposée à $\mathcal{J}$ causes de mort mutuellement exclusives et éventuellement dépendantes. A chaque individu est associé un couple de variables $(X, \mathcal{C})$. La variable aléatoire (v.a.) X est positive, de fonction de répartition (f.r.) F et représente la durée de vie de l'individu. La v.a. $\mathcal{C}$ indique la cause du décès et ainsi prend la valeur j parmi $\{1, \dots, \mathcal{J}\}$ lorsque la mort est due à la j^{e} cause. Nous nous intéressons ici à la fonction de répartition spécifique à la j^{e} cause définie pour $t \geqslant 0$ par $F^{(j)}(t) = \mathbb{P}[X \leqslant t, \mathcal{C} = j]$. Dans la suite, nous supposons que les fonctions $F^{(j)}$ pour $j = 1, \dots, \mathcal{J}$ ont des ensembles de points de discontinuité disjoints.

Adresse e-mail : geffray@ccr.jussieu.fr.

1631-073X/$ – see front matter
doi:10.1016/j.crma.2007.02.011

Supposons que le couple $(X, \mathcal{C})$ est censuré à droite par une v.a. positive C indépendante de $(X, \mathcal{C})$ et de f.r. G. On observe alors non pas le couple $(X, \mathcal{C})$ mais le couple $(T = \min(X, C),\ J = \mathcal{C}\colon\ I(X \leqslant C))$ où $I(\cdot)$ est la fonction indicatrice. La f.r. H de la v.a. T est donnée par la relation $1 - H = (1 - F)(1 - G)$ et est à support dans l'intervalle $[0, \tau_H]$ où $\tau_H = \sup\{x\colon\ H(x) < 1\}$. Introduisons pour $j = 1, \ldots, \mathcal{J}$ et $t \geqslant 0$ la fonction $H^{(1,j)}(t) = \mathbb{P}[T \leqslant t, J = j]$.

Soient n copies indépendantes $(T_i,\ J_i)_{i=1,\ldots,n}$ de (T, J). Introduisons pour $t \geqslant 0$ les fonctions de répartition empiriques $H_n(t) = \frac{1}{n}\sum_{i=1}^n I(T_i \leqslant t)$ et $H_n^{(1,j)}(t) = \frac{1}{n}\sum_{i=1}^n I(T_i \leqslant t, J_i = j)$ associées à H et $H^{(1,j)}$ respectivement. Pour toute fonction de répartition L, notons L^- la modification continue à gauche de L donnée pour $t \geqslant 0$ par $L^-(t) = \lim_{u\uparrow t} L(u)$. Pour $t \geqslant 0$, l'estimateur de Kaplan–Meier de F s'écrit $\widehat{F}_n(t) = 1 - \prod_{i=1}^n (1 - I(T_i \leqslant t,\ J_i \neq 0)/n(1 - H_n^-(T_i)))$. L'estimateur introduit par Aalen et Johansen [1] pour $F^{(j)}$ est défini pour $t \geqslant 0$ par :

$$\widehat{F}_n^{(j)}(t) = \int_0^t \frac{1 - \widehat{F}_n^-}{1 - H_n^-}\,\mathrm{d}H_n^{(1,j)}.$$

Nous obtenons la convergence faible de processus correctement centrés et normalisés, basés sur les $\widehat{F}_n^{(j)}$. Nous construisons alors des bandes de confiance asymptotiques jointes pour les $F^{(j)}$ et F. Enfin, nous énonçons deux résultats d'approximation forte.

2. Convergence faible et bandes de confiance

Comme l'estimateur de Kaplan–Meier, l'estimateur de Aalen–Johansen possède une structure de martingale en temps continu. Aalen et Johansen [1] ont obtenu la convergence faible jointe des processus $K_n^{(j)}$ pour $j = 1, \ldots, \mathcal{J}$ sur les compacts $[0, \sigma]$ où $\sigma < \tau_H$ et sous l'hypothèse de continuité des $F^{(j)}$ pour $j = 1, \ldots, \mathcal{J}$ requise pour appliquer le théorème de Rebolledo. Sur les mêmes compacts $[0, \sigma]$ où $\sigma < \tau_H$, Dauxois [3] a obtenu la convergence faible du processus $K_n^{(0)}$ en utilisant un théorème de Jakubowski et al. [6] sans effectuer d'hypothèse de continuité. Sa méthode se généralise à notre cas multidimensionnel en introduisant la martingale

$$\bigl\{\bigl(M_n^{(1,1)}, \ldots, M_n^{(1,\mathcal{J})}\bigr), \mathcal{F}_n(t)\colon\ t \in [0, \tau_H]\bigr\}$$

où pour $j = 1, \ldots, \mathcal{J}$,

$$M_n^{(1,j)}(t) = \sum_{i=1}^n \left(I(T_i \leqslant t, J_i = j) - \int_0^{t\wedge T_i} \mathrm{d}H^{(1,j)}/(1 - H^-) \right)$$

et où

$$\mathcal{F}_n(t) = \sigma\bigl\{T_i I(T_i \leqslant s),\ J_i I(T_i \leqslant s)\colon\ i = 1, \ldots, n,\ s \leqslant t\bigr\}$$

est la σ-algèbre engendrée par les événements observés avant t.

Théorème 2.1. *Posons $F^{(0)} \equiv F$, $\widehat{F}_n^{(0)} \equiv \widehat{F}_n$ et notons $\delta_{j,l} = I(j = l)$. Introduisons pour $j = 0, \ldots, \mathcal{J}$:*

$$K_n^{(j)} = \sqrt{n}\left(\widehat{F}_n^{(j)} - \frac{1 - \widehat{F}_n}{1 - F} F^{(j)}\right) \quad et \quad \widetilde{K}_n^{(j)} = \sqrt{n}\left(\widehat{F}_n^{(j)} \frac{1 - F}{1 - \widehat{F}_n} - F^{(j)}\right).$$

Dans $D^{\mathcal{J}+1}[0, \tau_H[$, l'espace des fonctions cadlag sur $[0, \tau_H[$ à valeurs dans $\mathbb{R}^{\mathcal{J}+1}$, on a :

$$\bigl(K_n^{(0)}, K_n^{(1)}, \ldots, K_n^{(\mathcal{J})}\bigr) \xrightarrow{\mathcal{D}} \bigl(K^{(0)}, K^{(1)}, \ldots, K^{(\mathcal{J})}\bigr),$$
$$\bigl(\widetilde{K}_n^{(0)}, \widetilde{K}_n^{(1)}, \ldots, \widetilde{K}_n^{(\mathcal{J})}\bigr) \xrightarrow{\mathcal{D}} \bigl(K^{(0)}, K^{(1)}, \ldots, K^{(\mathcal{J})}\bigr),$$

où les processus $K^{(j)}$ sont gaussiens de moyenne nulle et de covariance donnée pour $k, j = 0, \ldots, \mathcal{J}$ et $s, t \geqslant 0$ par :

$$\mathrm{Cov}\bigl(K^{(j)}(s), K^{(k)}(t)\bigr) = \sum_{l=1}^{\mathcal{J}} \int_0^{s\wedge t} \left(\delta_{j,l} + \frac{F^{(j)}}{1 - F}\right)\left(\delta_{k,l} + \frac{F^{(k)}}{1 - F}\right) \frac{(1 - F^-)(1 - F)}{(1 - H^-)^2}\,\mathrm{d}H^{(1,l)}.$$

Ce théorème entraîne l'égalité en distribution suivante : $K^{(j)} \stackrel{\mathcal{D}}{=} W^{(j)}(C^{(j)})$ où les $W^{(j)}$ sont des processus de Wiener corrélés et où $C^{(j)}(t) = \operatorname{Var}(K^{(j)}(t))$ représente la variance du j^e processus limite. Ces résultats permettent d'établir des bandes de confiance asymptotiques jointes pour les $F^{(j)}$ et F selon les approches de Hall–Wellner (HW) et Aalen–Nair (AN).

Pour $j = 1, \dots, \mathcal{J}$ et $t \geqslant 0$, introduisons un estimateur $\widehat{C}_n^{(j)}(t)$ de $C^{(j)}(t)$:

$$\widehat{C}_n^{(j)}(t) = \sum_{l=1}^{\mathcal{J}} \int_0^{s\wedge t} \left(\delta_{j,l} + \frac{\widehat{F}_n^{(j)}}{1-\widehat{F}_n}\right)\left(\delta_{k,l} + \frac{\widehat{F}_n^{(k)}}{1-\widehat{F}_n}\right) \frac{(1-\widehat{F}_n^-)(1-\widehat{F}_n)}{(1-H_n^-)^2}\,\mathrm{d}H_n^{(1,l)}$$

et posons

$$\mathcal{K}_n^{(j)}(t) = \begin{cases} \dfrac{\lambda_j}{\sqrt{n}}\sqrt{\widehat{C}_n^{(j)}(\sigma)} \text{ pour les bandes AN} \\ \dfrac{\lambda_j}{\sqrt{n}}\big(1+\widehat{C}_n^{(j)}(t)\big) \text{ pour les bandes HW} \end{cases} \quad \text{et} \quad \mathcal{K}_n^{(0)}(t) = \sum_{j=1}^{\mathcal{J}} \mathcal{K}_n^{(j)}(t).$$

Théorème 2.2. *Fixons $\sigma < \tau_H$. Notons W le mouvement brownien standard sur $[0,1]$ et B le pont brownien sur $[0,1]$ et posons pour des $\lambda_j > 0$:*

$$\alpha(\sigma) = \begin{cases} 1 - \displaystyle\sum_{j=1}^{\mathcal{J}} K_W(\lambda_j, 1) \textit{ pour les bandes AN avec } K_W(\lambda, a) = \mathbb{P}\Big[\sup_{0\leqslant t\leqslant a} \big|W(t)\big| \leqslant \lambda\Big], \\ 1 - \displaystyle\sum_{j=1}^{\mathcal{J}} K_B\left(\lambda_j, \frac{C_j(\sigma)}{1+C_j(\sigma)}\right) \textit{ pour les bandes HW avec } K_B(\lambda, a) = \mathbb{P}\Big[\sup_{0\leqslant s\leqslant a} \big|B(s)\big| \leqslant \lambda\Big]. \end{cases}$$

Les bandes de confiance de type Aalen–Nair et de type Hall–Wellner sont données par :

$$\lim_{n\to\infty} \mathbb{P}\big[r_n^{(0)}(t) \leqslant F(t) \leqslant R_n^{(0)}(t),\ r_n^{(j)}(t) \leqslant F^{(j)}(t) \leqslant R_n^{(j)}(t),\ j = 1, \dots, \mathcal{J},\ t \leqslant \sigma\big] \geqslant \alpha(\sigma),$$

avec

$$r_n^{(j)}(t) = \big(\widehat{F}_n^{(j)}(t) - \mathcal{K}_n^{(j)}(t)\big)\big(1+\mathcal{K}_n^{(0)}(t)\big)^{-1}, \qquad R_n^{(j)}(t) = \big(\widehat{F}_n^{(j)}(t) + \mathcal{K}_n^{(j)}(t)\big)\big(1-\mathcal{K}_n^{(0)}(t)\big)^{-1},$$
$$r_n^{(0)}(t) = \big(\widehat{F}_n(t) - \mathcal{K}_n^{(0)}(t)\big)\big(1-\mathcal{K}_n^{(0)}(t)\big)^{-1}, \qquad R_n^{(0)}(t) = \big(\widehat{F}_n(t) + \mathcal{K}_n^{(0)}(t)\big)\big(1+\mathcal{K}_n^{(0)}(t)\big)^{-1}.$$

Les bandes de confiance modifiées de type Aalen–Nair et de type Hall–Wellner sont données par :

$$\lim_{n\to\infty} \mathbb{P}\big[\tilde{r}_n^{(0)}(t) \leqslant F(t) \leqslant \widetilde{R}_n^{(0)}(t),\ \tilde{r}_n^{(j)}(t) \leqslant F^{(j)}(t) \leqslant \widetilde{R}_n^{(j)}(t),\ j = 1, \dots, \mathcal{J},\ t \leqslant \sigma\big] \geqslant \alpha(\sigma),$$

avec

$$\tilde{r}_n^{(j)}(t) = \widehat{F}_n^{(j)}(t)\big(1-\mathcal{K}_n^{(0)}(t)\big) - \mathcal{K}_n^{(j)}(t), \qquad \widetilde{R}_n^{(j)}(t) = \widehat{F}_n^{(j)}(t)\big(1+\mathcal{K}_n^{(0)}(t)\big) + \mathcal{K}_n^{(j)}(t),$$
$$\tilde{r}_n^{(0)}(t) = 1 - \big(1-\widehat{F}_n(t)\big)\big(1+\mathcal{K}_n^{(0)}(t)\big), \qquad \widetilde{R}_n^{(0)}(t) = 1 - \big(1-\widehat{F}_n(t)\big)\big(1-\mathcal{K}_n^{(0)}(t)\big).$$

3. Approximations fortes

Les deux résultats d'approximation forte suivants sont valables uniformément sur les intervalles aléatoires $[0, T_{n-k_n,n}]$ où $T_{n-k_n,n}$ représente la $(n-k_n)^e$ statistique d'ordre de l'échantillon $(T_1, \dots, T_n)$. La suite (k_n) désigne une suite d'entiers compris entre 1 et $n-1$. Si (k_n) est choisie négligeable devant n, alors la suite des intervalles $[0, T_{n-k_n,n}]$ est croissante pour l'inclusion à partir d'un certain rang et recouvre asymptotiquement tout compact $[0,\sigma]$ où $\sigma < \tau_H$. Dans la suite, (k_n) satisfaisait la condition suivante (H) : *pour n assez grand, $k_n \leqslant Ak_{2n}$ pour un $A > 0$, $k_n \geqslant (\log n)^2$ et la suite (k_n/n) est décroissante.*

Théorème 3.1. *Approximation forte jointe des processus $\sqrt{n}\,(\widehat{F}_n^{(j)} - F^{(j)})$ et $K_n^{(j)}$.*

Supposons que $F^{(j)}$ est continue pour $j = 1, \dots, \mathcal{J}$. Pour n assez grand, sur un espace de probabilité convenablement élargi, on peut définir $\mathcal{J}$ suites de processus gaussiens $(\tilde{L}_n^{(1)}), \dots, (\tilde{L}_n^{(\mathcal{J})})$ telles que l'on ait de façon jointe pour $j = 1, \dots, \mathcal{J}$,

$$\sup_{t \leqslant T_{n-k_n,n}} \left|\sqrt{n}\left(\widehat{F}_n^{(j)}(t) - F^{(j)}(t)\right) - \tilde{L}_n^{(j)}(t)\right| = \mathrm{O}\left(\sqrt{n}\,\frac{\log n}{k_n}\right).$$

Pour tout n, les processus $\tilde{L}_n^{(j)}$ sont de moyenne nulle et de covariance donnée pour $k, j = 1, \dots, \mathcal{J}$ et pour $s, t \geqslant 0$ par :

$$\mathrm{Cov}\left(\tilde{L}_n^{(j)}(t), \tilde{L}_n^{(k)}(s)\right) = \sum_{l=1}^{\mathcal{J}} \int_0^{s \wedge t} \left(\delta_{j,l} + \frac{F^{(j)} - F^{(j)}(t)}{1 - F}\right)\left(\delta_{k,l} + \frac{F^{(k)} - F^{(k)}(s)}{1 - F}\right)\frac{(1 - F)^2}{(1 - H^-)^2}\,\mathrm{d}H^{(1,l)}.$$

Pour n assez grand, sur le même espace de probabilité, on peut définir $\mathcal{J}$ suites de processus gaussiens $(L_n^{(1)}), \dots, (L_n^{(\mathcal{J})})$ telles que l'on ait de façon jointe pour $j = 1, \dots, \mathcal{J}$:

$$\sup_{t \leqslant T_{n-k_n,n}} \left|K_n^{(j)}(t) - L_n^{(j)}(t)\right| = \mathrm{O}\left(\sqrt{n}\,\frac{\log n}{k_n}\right).$$

Pour tout n, les processus $L_n^{(j)}$ sont de moyenne nulle et de même covariance que les $K^{(j)}$ (cf. Theorème 2.1), à savoir, satisfont pour $k, j = 1, \dots, \mathcal{J}$ et pour $s, t \geqslant 0$:

$$\mathrm{Cov}\left(L_n^{(j)}(t), L_n^{(k)}(s)\right) = \sum_{l=1}^{\mathcal{J}} \int_0^{s \wedge t} \left(\delta_{j,l} + \frac{F^{(j)}}{1 - F}\right)\left(\delta_{k,l} + \frac{F^{(k)}}{1 - F}\right)\frac{(1 - F)^2}{(1 - H^-)^2}\,\mathrm{d}H^{(1,l)}.$$

Ce théorème s'obtient en décomposant les processus $\sqrt{n}\,(\widehat{F}_n^{(j)} - F^{(j)})$ (resp. $K_n^{(j)}$) de façon à faire apparaître les processus empiriques $\sqrt{n}\,(H_n^{(1,j)} - H^{(1,j)})$ que l'on sait approximer de façon jointe par des ponts browniens grace au résultat d'Horváth [5]. Au cours de la décomposition, il apparaît un terme de reste dont la majoration nécessite l'emploi des U-statistiques comme Stute [7,8] et des VC-classes de fonctions comme Csörgő [2] et Giné et Guillou [4].

Références

[1] O. Aalen, S. Johansen, An empirical transition matrix for nonhomogeneous Markov chains based on censored observations, Scand. J. Statist. 5 (1978) 141–150.

[2] S. Csörgő, Universal gaussian approximations under random censorship, Ann. Statist. 24 (1996) 2744–2778.

[3] J.-Y. Dauxois, A new method for proving weak convergence results applied to nonparametric estimators in survival analysis, Stochastic Process. Appl. 90 (2000) 327–334.

[4] E. Giné, A. Guillou, Laws of the iterated logarithm for censored data, Ann. Probab. 27 (1999) 2042–2067.

[5] L. Horváth, Dropping continuity and independence assumptions in random censorship models, Studia Sci. Math. Hungar. 15 (1980) 381–389.

[6] A. Jakubowski, J. Mémin, G. Pagès, Convergence en loi des suites d'intégrales stochastiques sur l'espace D^1 de Skorokhod, Probab. Theory Rel. Fields 81 (1989) 111–137.

[7] W. Stute, Strong and weak representations of cumulative hazard function and Kaplan–Meier estimators on increasing sets, J. Statist. Plann. Inference 42 (1994) 315–329.

[8] W. Stute, U-statistic processes: a martingale approach, Ann. Probab. 22 (1994) 1725–1744.