Normalité asymptotique d'estimateurs à noyau de la densité et du taux de hasard pour des données censurées

Fatiha Messaci [1] & Mohamed Boukeloua [2]

Département de Mathématiques, Université des Frères Mentouri, route d'Ain El Bey, 25017 Constantine, Algérie.
E-mails : [1] *f_messaci@yahoo.fr,* [2] *boukeloua.mohamed@gmail.com*

Résumé. Dans ce travail, nous nous intéressons à l'estimation non paramétrique de la densité de probabilité et du taux de hasard d'une variable aléatoire X sujette à la censure. D'abord, nous considérons un cadre général de censure où nous supposons qu'au lieu d'observer X, nous observons une variable Z et un indicateur de censure qui détermine si l'observation est complète ou non. En utilisant une idée classique de l'estimation à noyau, nous introduisons un estimateur de la densité de X et nous établissons sa normalité asymptotique. Ensuite, nous appliquons notre résultat pour déduire la normalité asymptotique des estimateurs de la densité et du taux de hasard dans les cas de la censure à droite, la censure double et la censure mixte. Dans le premier cas $Z = \min(X, R)$ où R est une variable de censure à droite. Pour les deux autres modèles $Z = \max(\min(X, R), L)$, où L est une variable de censure à gauche et la variable X est indépendante du couple (L, R). Ce qui différencie les deux cas est que dans le premier (censure double) $L \leq R$ *p.s.* alors que dans le second (censure mixte) les variables X, R et L sont indépendantes. Finalement, nous illustrons la normalité des estimateurs étudiés par une étude de simulation complétée par des tests graphiques et numériques.

Mots-clés. Densité, taux de hasard, normalité asymptotique, censure double, censure mixte, censure à droite.

Abstract. In this work, we are concerned with the nonparametric estimation of the probability density and the failure rate functions of a random variable X which is at risk of being censored. First, we consider a general censorship setup where we assume that instead of observing X, we observe a variable Z and an indicator of censoring which determines whether the observation is complete or not. Borrowing a classical idea from the kernel-type estimation, we introduce an estimator of the density of X and we establish its asymptotic normality. Then, we apply our result in order to derive the asymptotic normality of the density and the failure rate estimators in the cases of right, doubly and twice censored data. In the first case $Z = \min(X, R)$ where R is a right censoring variable. In the two other cases $Z = \max(\min(X, R), L)$, where L is a left censoring variable and the variable X is independent of the pair (L, R), but $L \leq R$ *a.s.* in the double censorship model while the variables X, R and L are independent in the case of twice censoring.

Finally, we illustrate the normality of the studied estimators through a simulation study which is completed by graphical and numerical tests.

Keywords. Density, failure rate, asymptotic normality, double censoring, twice censoring, right censoring.

1 Résultat principal

Soit X une durée d'intérêt positive, de fonction de répartition F et de densité de probabilité f. On suppose que X peut être censurée, donc les observations disponibles consistent en un échantillon de n v.a.i.i.d. de même loi que le couple (Z, δ), où l'indicateur de censure δ prend la valeur 0 lorsque $Z = X$ (observation complète). En utilisant l'idée de l'estimation à noyau de Rosenblatt (1956), nous estimons f par

$$f_n(t) = \frac{1}{h_n} \int K\left(\frac{t-y}{h_n}\right) dF_n(y), \tag{1}$$

où F_n est un estimateur de F, continu à droite, à variation bornée et vérifiant quelques conditions supplémentaires que nous précisons dans la suite (voir l'hypothèse **H4**), K est le noyau et h_n est la fenêtre.
Dans toute la suite, pour tout ensemble A, 1_A désigne la fonction indicatrice de A.
Posons $g(t) = P(\delta = 0/X = t)$, la normalité asymptotique de f_n en un point fixé x est établie sous les hypothèses suivantes.

H1 f est dérivable en x.

H2 g est continue en x.

H3 $\exists\, a > 0$ tel que $\inf_{t\in[x-a,x+a]} g(t) > 0$.

H4 F_n est une fonction en escalier ayant des sauts éventuels uniquement en $(Z_i)_{1\leq i\leq n}$ et telle que
$\sqrt{n}\max_{i:Z_i\in[x-a,x+a]} \left| n\Delta F_n(Z_i)1_{\{\delta_i=0\}} - \frac{1_{\{\delta_i=0\}}}{g(Z_i)} \right| = O_P(1)$ et
$\sum_{i=1}^{n} \Delta F_n(Z_i)1_{\{\delta_i\neq 0\}} = o_P\left(\sqrt{h_n/n}\right)$ où $\Delta F_n(Z_i)$ est le saut de F_n au point Z_i.

H5 K est une densité bornée et à support compact.

H6 $nh_n \longrightarrow \infty$ et $nh_n^3 \longrightarrow 0$.

Notre résultat principal est le suivant.

Théorème 1 *Sous* **H1**–**H6**, *nous avons*

$$\sqrt{nh_n}(f_n(x) - f(x)) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{f(x)}{g(x)} \int K^2(z)\, dz\right).$$

2 Applications

Nous sommes dans la situation où X est censurée à droite par une v.a.r. R et $\min(X, R)$ est censuré à gauche par une v.a.r. L. En d'autres termes $Z = \max(\min(X, R), L)$.
Dans la suite, pour toute variable aléatoire réelle V, $F_V(t) = P(V \leq t)$, $S_V(t) = 1 - F_V(t)$, $I_V = \inf\{t \in \mathbb{R}/F_V(t) > 0\}$ et $T_V = \sup\{t \in \mathbb{R}/F_V(t) < 1\}$ représentent respectivement la fonction de répartition, la fonction de survie, le point initial et le point terminal du support de V.

2.1 Le modèle de censure double

Dans le modèle de Turnbull (1974), la durée X est indépendante du couple (L, R), $P(0 \leq L \leq R) = 1$ et $\delta = 1_{\{X>R\}} + 2 \times 1_{\{X<L\}}$. Turnbull (1974) a construit des estimateurs self-consistants pour S_X, S_R et S_L notés respectivement par S_n^D, $S_R^{(n)}$ et $S_L^{(n)}$.
Comme dans Ren (1997), nous imposons les conditions suivantes.

$$S_R^{(n)}(0) = 1,\ \lim_{u\to\infty} S_L^{(n)}(u) = 0 \text{ et } S_n^D(t) = \begin{cases} 1 & \text{si } t < \min_{1\leq i\leq n} Z_i, \\ 0 & \text{si } t \geq \max_{1\leq i\leq n} Z_i. \end{cases}$$

Par application de la formule (1), nous estimons f par

$$f_n^D(t) = \frac{1}{h_n} \int K\left(\frac{t-y}{h_n}\right) dF_n^D(y),$$

où $F_n^D = 1 - S_n^D$.
C'est l'estimateur introduit par Ren (1997) qui a imposé, en particulier, les hypothèses suivantes.

D1 $\forall t \geq 0$, $S_R(t) - S_L(t) > 0$.

D2 S_X, S_R et S_L sont continues au point t pour $t \geq 0$, et $0 < S_X(t) < 1$ pour $t > 0$.

D3 $S_X(0) = S_R(0) = 1$ et $\lim_{u\to\infty} S_X(u) = \lim_{u\to\infty} S_R(u) = \lim_{u\to\infty} S_L(u) = 0$.

D4 Il existe α et β, $0 < \alpha < \beta < \infty$, tels que $P(L \in]0, \alpha[) = 0$ et $P(L \leq \beta) = 1$.

Par ailleurs, le taux de hasard de X est défini par $\lambda(x) = \frac{f(x)}{S_X(x)} 1_{\{S_X(x)\neq 0\}}$. Un estimateur naturel de $\lambda(x)$ est donc donné par

$$\lambda_n^D(x) = \frac{f_n^D(x)}{S_n^D(x)} 1_{\{S_n^D(x)\neq 0\}}. \tag{2}$$

Le résultat suivant, obtenu du Théorème 1, concerne la normalité asymptotique de f_n^D et λ_n^D.

Corollaire 1 *Soit* $x \in]I_X, T_X[$, *sous* **H1**, **H5**, **H6** *et* **D1**–**D4**, *nous avons*

i) $\sqrt{nh_n}(f_n^D(x) - f(x)) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{f(x)}{S_R(x)-S_L(x)} \int K^2(z)\,dz\right)$.

ii) $\sqrt{nh_n}(\lambda_n^D(x) - \lambda(x)) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{f(x)}{S_X^2(x)(S_R(x)-S_L(x))} \int K^2(z)\,dz\right)$.

Remarquons que Ren (1997) donne le même résultat sous les mêmes conditions sauf qu'à la place de notre hypothèse **H1** (f est dérivable au point x), il suppose que f est bornée, $f(x) > 0$ et au voisinage de x, la dérivée seconde de $f(S_R - S_L)$ existe et est bornée.

2.2 Le modèle de censure mixte

Dans ce cas nous supposons que X, L et R sont positives et indépendantes et que $\delta = 1_{\{L<R<X\}} + 2 \times 1_{\{\min(X,R)\leq L\}}$. C'est le modèle I étudié dans Patilea et Rolin (2006) qui proposent d'estimer F par un estimateur produit-limite F_n^T. Appliquons encore une fois (1) mais en remplaçant F_n par F_n^T, nous retrouvons l' estimateur f_n^T de f introduit par Kitouni et al. (2015) qui ont montré sa convergence presque complète avec taux. Dans ce travail et afin de déduire la normalité asymptotique de f_n^T du Théorème 1, nous avons besoin des conditions suivantes déjà utilisées dans Patilea et Rolin (2006).
T1 $I_L \leq I_X$ et $T_X \leq T_R$.
T2 $\int_{\{u>I_{H_0}\}} \frac{dH_2(u)}{(F_Z(u))^2} < \infty$ où $H_k(t) = P(Z \leq t, \delta = k)$, $k \in \{0, 2\}$ et $I_{H_0} = \inf\{t \in \mathbb{R}/H_0(t) > 0\}$.
De plus, nous estimons le taux de hasard $\lambda(x)$ de X comme dans (2) par

$$\lambda_n^T(x) = \frac{f_n^T(x)}{S_n^T(x)} 1_{\{S_n^T(x)\neq 0\}}.$$

Maintenant nous sommes en mesure d'énoncer le résultat que nous visons.

Corollaire 2 *Soit* $x \in]I_X, T_X[$, *sous les hypothèses* **H1, H2, H5, H6, T1** *et* **T2**, *nous obtenons*

i) $\sqrt{nh_n}(f_n^T(x) - f(x)) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{f(x)}{F_L(x)S_R(x)} \int K^2(z)\,dz\right)$.

ii) $\sqrt{nh_n}(\lambda_n^T(x) - \lambda(x)) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{f(x)}{S_X^2(x)F_L(x)S_R(x)} \int K^2(z)\,dz\right)$.

2.3 Le modèle de censure à droite

En supposant maintenant que $L = 0$ *p.s.* dans le modèle précédent, nous nous ramenons au modèle de censure à droite. De plus, puisque dans ce cas F_n^T coïncide avec l'estimateur de Kaplan-Meier, f_n^T est l'estimateur de Blum et Susarla (1980). Nous retrouvons le résultat de normalité asymptotique donné dans Mielniczuk (1986) sous des conditions plus fortes que les nôtres.

2.4 Simulation

Une étude de simulation, consistant à comparer l'histogramme et l'estimateur de Rosenblatt (1956) basé sur un échantillon de valeurs des estimateurs étudiés avec la densité de la loi normale, a montré que le comportement de ces estimateurs est proche de celui de la loi normale pour des taux de censure avoisinant les 30 %. Ceci est confirmé par le test graphique Q-Q plot et les tests numériques de Kolmogorov-Smirnov et de Shapiro-Wilk.

Bibliographie

[1] Blum, J. R. et Susarla, V. (1980), Maximal deviation theory of density and failure rate function estimates based on censored data, *In Multivariate Analysis 5* (P.R. Krishnaiah ed.), 213–222.
[2] Kitouni, A., Boukeloua, M. et Messaci, F. (2015), Rate of strong consistency for nonparametric estimators based on twice censored data, *Statistics & Probability Letters*, 96, 255–261.
[3] Mielniczuk, J. (1986), Some asymptotic properties of kernel estimators of a density function in case of censored data, *The Annals of Statistics*, 14(2), 766–773.
[4] Patilea, V. et Rolin, J.-M. (2006), Product limit estimators of the survival function with twice censored data, *The Annals of Statistics*, 34(2), 925–938.
[5] Ren, J-J. (1997), On self-consistent estimators and kernel density estimators with doubly censored data, *Journal of Statistical Planning and Inference*, 64, 27–43.
[6] Rosenblatt, M. (1956), Remarks on some nonparametric estimates of density function, *The Annals of Mathematical Statistics*, 27, 832–837.
[7] Turnbull, B. W. (1974), Nonparametric estimation of a survivorship function with doubly censored data, *Journal of the American Statistical Association*, 69, 169–173.