

Estimation non-paramétrique du taux de hasard : application à des durées de chômage censurées à droite

Abdelhak NASSIRI, Michel DELECROIX, Michel BONNEU *

RÉSUMÉ. – Dans cet article, nous étudions plusieurs estimateurs non-paramétriques du taux de hasard dans le cas de données censurées à droite. Faute de résultats théoriques de convergence permettant une comparaison de ces estimateurs, nous procédons à une validation empirique de leurs performances. Cette approche est fondée sur une estimation empirique du MISE (*Mean Integrated Squarred Error*) à partir de données simulées. Elle met en évidence les performances d'un estimateur, dit du *k^{ème} plus proche voisin non censuré*, que l'on introduit. Nous mettons en œuvre ce nouvel estimateur sur la base de données réelles de chômage. Cette application, permet de valider les résultats obtenus pour des modèles paramétriques estimés sur ces mêmes données, et surtout de mettre en évidence la perte d'information due au manque de flexibilité des formes fonctionnelles paramétriques.

A New Nonparametric Estimator of Hazard Rate: Application to Right Censored Unemployment Duration

ABSTRACT. – Different hazard rate nonparametric estimators for randomly right-censored data are considered in this paper. Lacking sufficiently homogeneous theoretical results allowing for a comparison between their performances, an empirical procedure is adopted. This procedure is based on an empirical estimation of the MISE criterion (*Mean Integrated Squarred Error*) from simulated data. It highlights the fact that an estimator which is based on the nearest-neighbor method, and which we introduce, is more dependable thus more appropriate for unemployment data analysis. Application of this new nonparametric estimator allows us to validate results from a parametric model estimation, and particularly to show the information loss caused by inflexible parametric specification of the hazard rate.

* A. NASSIRI : ICI-ECO ; M. DELECROIX : Insee ; M. BONNEU : IUT Toulouse.

Des versions préliminaires de cet article ont été présentées aux Journées des Jeunes Economètres en avril 1991 à Toulouse, au séminaire d'Économétrie Semi- et Non-Paramétrique du GREMAQ et au séminaire d'Économétrie du Travail du GREMAQ-CEJEE, Toulouse 1992. Il figure comme chapitre III dans la thèse de Doctorat Nouveau Régime soutenue par A. NASSIRI à l'Université des Sciences Sociales-Toulouse I.

Les auteurs remercient les deux rapporteurs anonymes pour leurs remarques pertinentes et leurs critiques constructives.

1 Introduction

Nous disposons de données réelles relatives à une cohorte de demandeurs d'emploi dont nous souhaitons estimer la probabilité instantanée de sortie du chômage vers un emploi. Cependant, ces données sont censurées à droite : pour certains individus, la donnée disponible est un temps auquel le chômage était encore effectif, sans que l'on sache à quelle date un emploi a été retrouvé par la suite. Pour formaliser cette situation, on utilise le modèle d'EFRON [1967].

On note Z_i la variable aléatoire (*v.a.*) positive associée à la durée de chômage d'un individu i ($1 \leq i \leq n$), où n est la taille de l'échantillon. Les n observations disponibles T_1, \dots, T_n , sont alors définies par $T_i = \min(Z_i, C_i)$ où C_i est une *v.a.* positive, dite variable de censure. Les *v.a.* T_i et C_i sont *i.i.d.*, en outre C_i et Z_i sont supposées indépendantes entre elles. On observe, aussi, s'il y a eu censure ou pas : pour chaque individu i , on connaît la valeur de la variable indicatrice de non censure, $\delta_i = \mathbb{I}_{[Z_i < C_i]}$.

Soit f (respectivement F) la densité (respectivement la fonction de répartition) de Z_i . Son taux de hasard, noté $\lambda(t)$, est aussi appelé fonction de hasard. Il s'interprète comme la probabilité qu'un individu sorte du chômage à l'instant t , sachant qu'il a été chômeur jusqu'à cette date. Il est défini par :

$$\lambda(t) = \lim_{\Delta t \searrow 0} \frac{1}{\Delta t} \text{prob}(t \leq Z_i \leq t + \Delta t \mid Z_i \geq t), t \in \mathbb{R}^+.$$

Dans cet article, nous nous proposons de réaliser une estimation non-paramétrique de ce taux de hasard $\lambda(t)$. Grâce au développement de cette technique au cours des deux dernières décennies, on dispose aujourd'hui d'un grand nombre d'estimateurs non-paramétriques. Ils se différencient essentiellement par la méthode d'estimation fonctionnelle appliquée, et par la définition du taux de hasard retenue. Il existe, en effet, trois définitions distinctes de la fonction de hasard (voir section 2 *infra*) auxquelles s'appliquent différentes techniques : méthode du noyau, des k plus proches voisins, des fonctions orthogonales, ou des histogrammes à pas aléatoires... (cf. DROESBEKE *et alii* [1989] ; MC NICHOLS et PADGETT [1983], pour une étude bibliographique). Les résultats théoriques de convergence, établis pour ces différents estimateurs, ne permettent cependant pas de les comparer de façon satisfaisante et homogène (FÖLDES *et alii* [1981], BLUM et SUSARLA [1980], TANNERS et WONG [1983]). À notre connaissance, on ne dispose pas, par exemple, du calcul explicite d'un indicateur global de l'écart entre le taux de hasard $\lambda(t)$ et son estimation $\hat{\lambda}(t)$, qui soit valable pour tous ces estimateurs.

Dans notre étude, qui est loin d'être exhaustive, nous nous intéressons, tout particulièrement, à deux méthodes, celle du noyau et celle des k plus proches voisins. Sur la base des trois formulations du taux de hasard, nous définissons ainsi trois principaux estimateurs non-paramétriques. Afin de tenir compte de la censure, nous introduisons aussi un nouvel estimateur fondé sur une méthode dérivée des k plus proches voisins ; l'objectif étant d'en valider les

performances par rapport aux trois autres estimateurs. Pour cela, on considère comme indicateur le MISE (*Mean Integrated Square Error*) défini par :

$$\text{MISE}(\hat{\lambda}) = E\left(\int_0^{+\infty} [\hat{\lambda}(t) - \lambda(t)]^2 dt\right),$$

comme il est usuel en estimation fonctionnelle, à condition de supposer que $\lambda(t)$ est de carré intégrable.

Dans la section 3, nous procédons à une validation empirique afin d'étudier les performances des différents estimateurs considérés. Cette validation est réalisée sur la base d'une estimation empirique d'un MISE tronqué, à partir de données simulées. Elle conduit à privilégier l'estimateur non-paramétrique, dit du *k^{ème} plus proche voisin non censuré*, que nous introduisons.

Dans la section 4, nous mettons en œuvre cet estimateur sur un ensemble de données réelles relatives à une cohorte de chômeurs radiés durant l'année 1986 de l'ANPE de Toulouse. Dans cette application, nous mettons en évidence les résultats des estimations non-paramétriques par rapport à ceux obtenus dans des modèles paramétriques.

2 Estimation non-paramétrique du taux de hasard

Le taux de hasard défini dans l'introduction admet trois formulations équivalentes, notées : $\lambda_l(t)$, $l = 1, 2, 3$. Pour tout $t \in \mathbb{R}^+$ tel que $F(t) < 1$:

$$(1) \quad \lambda_1(t) = \frac{f(t)}{S(t)},$$

où $S(t) = 1 - F(t)$ est la fonction de survie de Z_i ;

$$(2) \quad \lambda_2(t) = \frac{g(t)}{1 - F^*(t)},$$

où $g(\cdot)$ est la densité de la mesure $\nu(\cdot)$ définie par $\nu(A) = \text{prob}[(T_i \in A) \cap (\delta_i = 1)]$, pour tout borélien A de \mathbb{R}^+ . $g(\cdot)$ correspond donc à la densité des données non censurées. En revanche, $F^*(\cdot)$ est la fonction de répartition des T_i ;

$$(3) \quad \lambda_3(t) = \frac{-d[\ln S(t)]}{dt}.$$

Plusieurs estimateurs naturels non-paramétriques du taux de hasard ont été définis à partir de ces trois formulations. Pour notre validation empirique, nous en considérons trois caractérisés par la méthode du noyau appliquée aux trois formulations (1), (2) et (3).

Concernant la définition (2), il est clair que l'échantillon fournit les observations nécessaires à la construction d'un estimateur à noyau de la densité $g(\cdot)$ et d'un estimateur de $F^*(\cdot)$ (on prend la fonction de répartition empirique $F_n^*(\cdot)$ de T_i). On en déduit l'estimateur *classique* suivant :

$$(4) \quad \hat{\lambda}_2(t) = \frac{\hat{g}(t)}{1 - F_n^*(t)},$$

où $\hat{g}(t) = \frac{1}{n \cdot h} \sum_{i=1}^n \delta_i \cdot K\left(\frac{T_i - t}{h}\right)$ pour un noyau $K(\cdot)$ de PARZEN-ROSENBLATT et une fenêtre h . Cet estimateur a été proposé par BLUM et SUSARLA [1980].

En revanche, dans le cas des deux autres formulations (1) et (3), nous ne disposons pas des n observations de Z_i pour en estimer simplement la densité $f(t)$ et la fonction de survie $S(t)$. En pratique, pour estimer cette fonction de survie, on utilise l'estimateur de KAPLAN-MEIER [1958] défini par :

$$\hat{S}_{KM}(t) = \prod_{T_i < t} \left(\frac{n - r_i}{n - r_i + 1} \right)^{\delta_i},$$

où r_i représente le rang de l'observation T_i dans l'échantillon. Le procédé d'estimation de (1) et (3) consiste donc à approximer globalement la loi des Z_i par une loi discrète définie sur l'ensemble des observations des T_i , par les probabilités p_i déduites de $\hat{S}_{KM}(t)$. Soit

$$(5) \quad p_i = \frac{\delta_i}{n - r_i + 1} \prod_{j/t_j < t_i} \left(\frac{n - r_j}{n - r_j + 1} \right)^{\delta_j}.$$

On définit alors les deux estimateurs respectifs aux formulations (1) et (3) par :

$$(6) \quad \hat{\lambda}_1(t) = \frac{\hat{f}(t)}{\hat{S}_{KM}(t)},$$

$$(7) \quad \hat{\lambda}_3(t) = \frac{1}{h} \sum_{i=1}^n \frac{\delta_i}{n - r_i + 1} K\left(\frac{T_i - t}{h}\right),$$

où $\hat{f}(t) = \frac{1}{h} \sum_{i=1}^n p_i K\left(\frac{T_i - t}{h}\right)$. L'estimateur $\hat{\lambda}_1(t)$ a été notamment étudié par FÖLDES *et alii* [1981], et $\hat{\lambda}_3(t)$ par TANNERS et WONG [1983].

Les simulations effectuées dans la section 3 mettent en lumière que, comme à l'habitude, les performances pratiques des estimateurs à noyau considérés dépendent fortement du choix de la fenêtre h . Pour pallier cette difficulté, deux solutions pratiques ont été proposées. MARRON et PADGETT [1987] et

GRÉGOIRE [1991] étudient une technique usuelle de validation croisée, pour l'estimateur de la densité $f(\cdot)$. MC NICHOLS et PADGETT [1983] et MIELNICZUK [1986] définissent pour $f(\cdot)$ des estimateurs à noyau à fenêtres aléatoires, dérivés de la méthode *classique* des *k plus proches voisins*. Leur mise en œuvre pratique pose évidemment le problème de la présence d'observations censurées au voisinage du point t d'estimation. Nous avons, donc, défini un nouvel estimateur, dit du *k^{ème} plus proche voisin non censuré*. Celui-ci tient compte des observations *les plus proches de t*, pondérées par leurs poids calculés à l'aide de l'estimateur de KAPLAN-MEIER.

Ainsi, pour calculer l'estimateur $\hat{f}_P(t)$ de la densité, on utilisera les observations les plus proches de t dont la masse empirique représente un poids P fixé à l'avance. Ceci équivaut bien, dans la méthode des *k plus proches voisins* standard sur données non censurées, à utiliser les observations les plus proches de t dont la masse empirique totale vaut $\frac{k}{n}$.

En posant $d(t) = \min_y \{ \sum_{i/|T_i-t| \leq y} p_i \geq P \}$, on définit alors, pour tout $t \in \mathbb{R}^{+*}$:

$$(8) \quad \hat{f}_P(t) = \frac{P}{2.d(t)},$$

$$(9) \quad \hat{\lambda}_P(t) = \frac{\hat{f}_P(t)}{\hat{S}_{KM}(t)}.$$

Évidemment, se pose ici le problème du choix du *meilleur* poids P , mais les simulations ont révélé qu'il était moins crucial que celui du choix d'un nombre de voisins non censurés ou d'une fenêtre dans le cas des trois autres estimateurs à noyau.

Pour réaliser la validation empirique évoquée dans l'introduction, il faut d'abord comparer les performances des quatre estimateurs précédents ; le critère retenu étant le MISE. Toutefois, une comparaison théorique complète de ces différents estimateurs ($\hat{\lambda}_l(t)_{l=1,2,3}, \hat{\lambda}_P(t)$) ne s'avère pas actuellement possible : à notre connaissance, le calcul exact des erreurs quadratiques intégrées n'existe que pour l'estimateur de la densité, intégration faite sur un compact (voir, par exemple, GRÉGOIRE [1991], PONS [1986], MARRON et PADGETT [1987]). Un des problèmes sous-jacents réside évidemment dans le fait que l'on ne peut garantir, dans le cas le plus général, l'appartenance à L^2 des fonctions de hasard estimées. Nous avons donc choisi de réaliser, dans la section 3, une comparaison empirique de ces différents estimateurs sur des échantillons simulés.

3 Validation empirique des estimateurs

3.1 Simulation des données

Dans cette étude empirique, nous considérons trois lois usuelles que l'on retrouve, très souvent, dans les modélisations paramétriques des durées individuelles de chômage ou de survie :

- la loi Log-Normale centrée et réduite ;
- la loi Gamma de paramètres 1/2 et 1/2 ;
- la loi de Weibull de paramètres 3/2 et 2.

Ces trois lois induisent des spécifications différentes sur la tendance temporelle de la probabilité instantanée de sortir du chômage. Par exemple, si l'on suppose que ce taux est strictement décroissant, la loi Gamma sera une des lois adéquates. Le graphique 1 permet de visualiser l'allure de ce taux qui diffère selon la loi retenue.

Nous simulons pour chacune de ces trois lois, 500 observations indépendantes qui représentent les durées Z_i , ($1 \leq i \leq 500$). Nous simulons aussi, dans chacun des trois cas, 500 observations indépendantes d'une loi de

Laplace de densité $q(t)$ définie sur \mathbb{R}^+ par $q(t) = \frac{e^{t-a}}{2 - e^{-a}}$, $t \in \mathbb{R}^+$. Ces observations représentent les valeurs de la v.a. de censure C_i . Avec une valeur du paramètre a fixée au voisinage de 2, nous avons observé un taux moyen de censure de 25 % environ pour les trois cas.

Les estimateurs $\hat{\lambda}_1(t)$, $\hat{\lambda}_2(t)$ et $\hat{\lambda}_3(t)$ définis respectivement en (6), (4) et (7) dépendent du choix du noyau $K(\cdot)$ et de la fenêtre h . Le choix du noyau influe peu sur les performances de ces estimateurs (cf. BOSQ et LECOUTRE [1987]), ce qui nous a conduits à privilégier la solution classique du noyau gaussien.

Le choix de la fenêtre h est, en revanche, crucial (MARRON et PADGETT [1987]). Comme fenêtre optimale, nous choisissons la valeur qui minimise le MISE. En fait, le calcul théorique du MISE étant trop compliqué dans le cas de données censurées, nous répétons 300 fois les simulations précédentes afin d'obtenir une série empirique $\{\hat{\lambda}_l^{(j)}\}_{j=1, \dots, 300}$ pour chacun des trois estimateurs $\hat{\lambda}_l(t)_{l=1,2,3}$. Il est alors possible d'estimer naturellement le MISE empirique sur un compact ¹ $[\underline{T}, \bar{T}]$:

$$\widehat{\text{MISE}}(h) = \frac{1}{300} \sum_{j=1}^{300} \int_{\underline{T}}^{\bar{T}} [\hat{\lambda}_l^{(j)}(t) - \lambda(t)]^2 dt, \quad l = 1, 2, 3, \text{ pour une fenêtre } h$$

donnée.

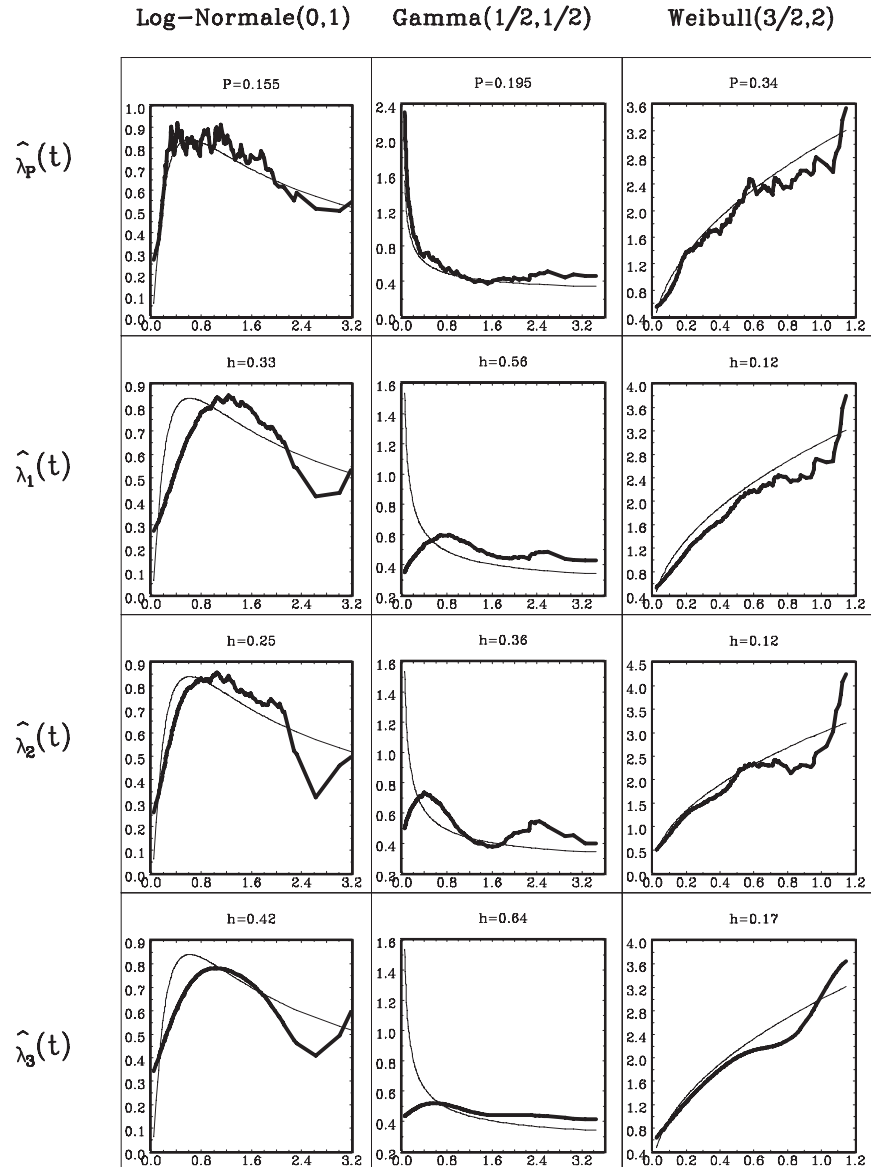
1. Dans la pratique, le compact $[\underline{T}, \bar{T}]$, inclus dans $]0, +\infty[$, est déterminé de telle sorte qu'il contienne toutes les observations.

La mise en œuvre de l'estimateur $\hat{\lambda}_P(t)$ défini en (9) repose sur le choix d'une valeur pour le poids P . Nous adoptons la même approche empirique que pour le choix de la fenêtre : simulation de 300 échantillons de taille 500 pour la variable T dans le cas de chacune des trois lois retenues. Il est alors possible d'en déduire une estimation empirique du MISE :

$$\widehat{\text{MISE}}(P) = \frac{1}{300} \sum_{j=1}^{300} \int_{\underline{T}}^{\overline{T}} [\hat{\lambda}_P^{(j)}(t) - \lambda(t)]^2 dt, \text{ pour un poids } P \text{ donné.}$$

GRAPHIQUE 1

Estimation non-paramétrique du taux de hasard des trois lois



Un balayage sur plusieurs valeurs de h et de P permet de déterminer la valeur d'une fenêtre optimale h_{opt} et d'un poids optimal P_{opt} qui minimise le MISE empirique :

$$\begin{cases} h_{opt} = \arg \min_h \widehat{\text{MISE}}(h) \\ P_{opt} = \arg \min_P \widehat{\text{MISE}}(P) \end{cases}$$

3.2 Résultats

Les résultats² des estimations du MISE sont fournis dans les tableaux 1 et 2. Les valeurs de la fenêtre optimale et du poids optimal en sont déduites et résumées dans le tableau 3.

Dans le Graphique 1, nous représentons les différents estimateurs ($\hat{\lambda}_P, \hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3$) de la fonction de hasard de chacune des trois lois. Ces estimateurs, représentés par la courbe en gras, sont calculés à l'aide de la fenêtre optimale h_{opt} ou du poids optimal P_{opt} , pour un ensemble de 500 données simulées ; l'autre courbe correspond à la fonction de hasard théorique.

$\hat{\lambda}_1(t)$, $\hat{\lambda}_2(t)$ et $\hat{\lambda}_3(t)$ sont des estimateurs des trois définitions du taux de hasard fondés sur la même méthode du noyau. En les comparant, il est clair que $\hat{\lambda}_3(t)$ permet de mieux approcher le hasard théorique de chacune des trois lois : Log-Normale, Gamma et Weibull. La valeur du MISE empirique correspondant à cet estimateur est nettement plus faible (cf. Tableau 3), ce qui se traduit par un graphe de $\hat{\lambda}_3(t)$ plus proche du hasard théorique $\lambda(t)$ dans le graphique 1. En effet, bien que ces trois estimateurs partent de la même méthode non-paramétrique, l'estimateur $\hat{\lambda}_3(t)$ s'applique à une définition du taux de hasard qui, à l'encontre de celle des deux autres, permet d'estimer ce taux directement à partir de l'estimateur de KAPLAN-MEIER de la survie, sans le recours au rapport de la densité et de la survie estimées. Dans le graphique 1, on peut remarquer que, malgré l'usage d'une même méthode de lissage (noyau gaussien), les deux estimateurs $\hat{\lambda}_1(t)$ et $\hat{\lambda}_2(t)$ présentent plus de perturbations que $\hat{\lambda}_3(t)$.

Comme l'a noté un des rapporteurs, *aucun des trois estimateurs ne réussit à retracer la décroissance initiale du hasard de la loi Gamma*. Deux raisons peuvent être rappelées pour expliquer ce résultat sur cette partie du domaine de la v.a Z. D'un côté, la loi Gamma est une distribution à très forte densité au voisinage de zéro. Par conséquent, une fenêtre optimale constante sur tout le domaine, comme c'est le cas pour les trois estimateurs, ne peut que se traduire par un lissage trop fort de la densité au voisinage de zéro, d'autant plus que l'ajustement est meilleur sur le reste du domaine. D'un autre côté, il s'agit d'une variable positive. Quand on cherche à estimer la distribution de la loi au voisinage de zéro, en particulier, pour tout point t qui tombe dans la première fenêtre, peu d'observations voire aucune ne tombent dans la moitié gauche de la fenêtre. DEHEUVELS et HOMINAL [1979] ont déjà étudié ce problème en relevant l'impossibilité d'estimer correctement par des méthodes

2. Les programmes de calculs sont réalisés au moyen de procédures en GAUSS.

standards, au voisinage de zéro, la densité d'une variable aléatoire positive quand celle-ci tend vers l'infini. Ils ont appliqué deux corrections aux méthodes du noyau à l'aide des « estimateurs de la ficelle ». L'idée intuitive de ces deux corrections est, dans un cas de ne prendre que la moitié de la fenêtre, et dans l'autre cas de tenir compte de la symétrie des observations autour du point t . Nous appliquons ces corrections et d'après les résultats des simulations, nous constatons en effet qu'aucune amélioration n'a été apportée par les deux méthodes dans le cas de la loi Gamma (cf. NASSIRI [1993], pour plus de détails). Et ces deux auteurs de conclure que l'application de ces corrections nécessite des informations supplémentaires sur la densité des observations au voisinage de zéro.

TABLEAU 1

Estimation du MISE des 3 estimateurs du taux de hasard : $\hat{\lambda}_1(t)$, $\hat{\lambda}_2(t)$ et $\hat{\lambda}_3(t)$ en fonction de la fenêtre h

Loi Log-Normale				Loi Gamma				Loi Weibull			
h	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	h	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$	h	$\hat{\lambda}_1$	$\hat{\lambda}_2$	$\hat{\lambda}_3$
0.60	0.1310	0.3377	0.0734	0.80	0.4802	0.9287	0.3991	0.31	0.1909	0.2393	0.0820
0.58	0.1265	0.3088	0.0724	0.78	0.4784	0.8929	0.3984	0.30	0.1801	0.2249	0.0801
0.56	0.1221	0.2826	0.0715	0.76	0.4767	0.8598	0.3978	0.29	0.1697	0.2109	0.0782
0.54	0.1180	0.2588	0.0707	0.74	0.4750	0.8292	0.3973	0.28	0.1599	0.1972	0.0765
0.52	0.1142	0.2374	0.0701	0.72	0.4735	0.8010	0.3968	0.27	0.1505	0.1841	0.0749
0.49	0.1089	0.2092	0.0692	0.70	0.4721	0.7751	0.3964	0.26	0.1416	0.1715	0.0734
0.48	0.1073	0.2007	0.0690	0.68	0.4708	0.7512	0.3961	0.25	0.1332	0.1595	0.0720
0.47	0.1058	0.1928	0.0688	0.66	0.4696	0.7292	0.3958	0.24	0.1254	0.1482	0.0708
0.46	0.1043	0.1853	0.0687	0.64	0.4686	0.7090	0.3957	0.23	0.1181	0.1376	0.0697
0.45	0.1029	0.1783	0.0685	0.62	0.4678	0.6905	0.3958	0.22	0.1114	0.1278	0.0688
0.44	0.1016	0.1716	0.06845	0.60	0.4671	0.6735	0.3960	0.21	0.1053	0.1188	0.0680
0.43	0.1004	0.1654	0.06844	0.58	0.4666	0.6580	0.3963	0.20	0.0998	0.1107	0.0674
0.42	0.0992	0.1596	0.06843	0.56	0.46643	0.6439	0.3968	0.19	0.0948	0.1036	0.0669
0.41	0.0982	0.1541	0.06848	0.54	0.46646	0.6311	0.3976	0.18	0.0905	0.0973	0.0667
0.40	0.0972	0.1490	0.0685	0.52	0.4668	0.6196	0.3986	0.17	0.0868	0.0919	0.0666
0.37	0.0949	0.1359	0.0690	0.50	0.4674	0.6092	0.3998	0.16	0.0837	0.0874	0.0668
0.35	0.0939	0.1289	0.0696	0.48	0.4685	0.6001	0.4014	0.15	0.0813	0.0839	0.0673
0.33	0.0934	0.1232	0.0705	0.46	0.4700	0.5921	0.4034	0.14	0.0795	0.0812	0.0681
0.31	0.0935	0.1187	0.0717	0.44	0.4720	0.5853	0.4058	0.13	0.0784	0.0795	0.0692
0.29	0.0942	0.1156	0.0734	0.42	0.4748	0.5798	0.4088	0.12	0.0781	0.0787	0.0709
0.27	0.0958	0.1137	0.0756	0.40	0.4782	0.5755	0.4123	0.11	0.0786	0.0789	0.0731
0.25	0.0983	0.1134	0.0785	0.38	0.4826	0.5726	0.4166	0.10	0.0800	0.0802	0.0760
0.23	0.1020	0.1146	0.0823	0.36	0.4880	0.5713	0.4218	0.09	0.0827	0.0828	0.0799
0.21	0.1073	0.1176	0.0873	0.34	0.4948	0.5716	0.4281	0.08	0.0869	0.0870	0.0852
0.19	0.1144	0.1230	0.0937	0.32	0.5032	0.5740	0.4357	0.07	0.0934	0.0936	0.0925
0.17	0.1241	0.1311	0.1021	0.30	0.5136	0.5785	0.4448	0.06	0.1031	0.1034	0.1029
0.15	0.1372	0.1430	0.1132	0.28	0.5263	0.5857	0.4560	0.05	0.1177	0.1181	0.1180
0.13	0.1551	0.1600	0.1282	0.26	0.5421	0.5960	0.4695	0.04	0.1406	0.1411	0.1410
0.11	0.1802	0.1844	0.1488	0.24	0.5614	0.6100	0.4861	0.03	0.1789	0.1795	0.1792

* Les nombres en caractère gras correspondent au minimum du MISE.

L'estimateur du $k^{\text{ème}}$ plus proche voisin non censuré $\hat{\lambda}_P(t)$ est un nouvel estimateur fondé sur la première définition, à savoir $f(t)/S(t)$. Il admet une erreur minimale (au sens du MISE empirique calculé à l'aide de la valeur optimale du poids P) de l'ordre de 0.0602 pour la loi Log-Normale, de 0.2584 pour la loi Gamma et de 0.0667 pour la loi Weibull. Comparé à l'estimateur classique $\hat{\lambda}_1(t)$ appliqué à cette définition, on remarque que $\hat{\lambda}_P(t)$ apparaît nettement plus performant. En effet, pour toute valeur de P entre 0.090 et 0.36 pour la loi Log-Normale, inférieure à 0.55 pour la loi Gamma et entre 0.18 et 0.49 pour la loi Weibull, et *a fortiori* pour les valeurs du poids

TABLEAU 2

Estimation du MISE de l'estimateur $\hat{\lambda}_P(t)$ du taux de hasard en fonction du poids P

Loi Log-Normale		Loi Gamma		Loi Weibull	
P	$\hat{\lambda}_P$	P	$\hat{\lambda}_P$	P	$\hat{\lambda}_P$
1	1.2444	1	1.3931	1	1.1304
0.840	0.9459	0.850	1.1164	0.850	0.2854
0.680	0.5767	0.700	0.8245	0.700	0.1476
0.520	0.3138	0.550	0.5900	0.550	0.0903
0.360	0.1453	0.400	0.4142	0.500	0.0798
0.200	0.0646	0.305	0.3274	0.490	0.0783
0.195	0.0641	0.300	0.3234	0.480	0.0768
0.190	0.0626	0.295	0.3198	0.470	0.0754
0.185	0.0614	0.290	0.3153	0.460	0.0745
0.180	0.0606	0.285	0.3112	0.450	0.0731
0.175	0.0611	0.280	0.3071	0.440	0.0721
0.170	0.0616	0.275	0.3041	0.430	0.0714
0.165	0.0614	0.270	0.3000	0.420	0.0705
0.160	0.0610	0.265	0.2966	0.410	0.0694
0.155	0.0602	0.260	0.2923	0.400	0.0683
0.15	0.0605	0.255	0.2888	0.390	0.0679
0.145	0.0606	0.250	0.2847	0.380	0.0673
0.140	0.0616	0.245	0.2841	0.370	0.0671
0.135	0.0672	0.240	0.2806	0.360	0.0673
0.130	0.0671	0.235	0.2771	0.350	0.0668
0.125	0.0684	0.230	0.2750	0.340	0.0667
0.120	0.0714	0.225	0.2711	0.330	0.0669
0.115	0.0721	0.220	0.2684	0.320	0.0673
0.110	0.0774	0.215	0.2663	0.310	0.0676
0.105	0.0801	0.210	0.2618	0.300	0.0684
0.100	0.0859	0.205	0.2627	0.260	0.0714
0.095	0.0916	0.200	0.2598	0.220	0.0769
0.090	0.0974	0.195	0.2584	0.180	0.0858
0.085	0.1058	0.190	0.2606	0.140	0.1005
0.080	0.1084	0.185	0.2620	0.100	0.1378

* Les nombres en caractère gras correspondent au minimum du MISE.

optimal, il présente une erreur plus petite que celle de l'estimateur $\hat{\lambda}_1(t)$ calculé à l'aide la fenêtre optimale.

Enfin, il reste à préciser que $\hat{\lambda}_P(t)$ donne des résultats plus performants que le meilleur des trois estimateurs à noyau $\hat{\lambda}_3(t)$ (les cas des lois Log-Normale et Gamma), sinon équivalents (le cas de la loi Weibull). Il donne une estimation plus proche du hasard théorique. De plus, au vu du graphe, cette estimation aurait été mieux ajustée si on y avait appliqué, à l'instar des autres estimateurs, une méthode de lissage telle la méthode du noyau. Remarquons, aussi, que pour la loi de Weibull, malgré une légère supériorité du nouvel estimateur, les quatre estimateurs ajustent correctement le taux de hasard théorique avec des résultats sensiblement identiques.

Par ailleurs, ces différents estimateurs non-paramétriques s'appliquent à des échantillons *i.i.d* qui sont supposés homogènes. Ils donnent une estimation de la dépendance temporelle de la probabilité instantanée de sortir du chômage, sans que l'on sache si cette tendance estimée représente le processus temporel effectif de cette probabilité. En effet, se pose souvent le problème d'identification de ce processus par rapport à une dépendance temporelle fictive qui s'explique plutôt par une certaine hétérogénéité de la population étudiée vis-à-vis de certains facteurs plus ou moins observables³. Afin d'identifier la dépendance temporelle effective en présence de ces différents facteurs, des modèles semi-paramétriques ont été développés (BRESLOW [1974], WHITEHEAD [1980], AITKIN *et alii* [1983] et MEYER [1986 ; 1990]). Dans ces modèles, on retient une forme discrète de la fonction de hasard assez souple et simple de mise en œuvre. Qualifiée de non-paramétrique (MEYER [1990]), elle est fondée sur une *spécification exponentielle par morceaux* du taux de hasard, que l'on estime par la méthode du maximum de vraisemblance⁴.

La spécification *exponentielle constante par morceaux* repose sur une partition du domaine de la variable T en J intervalles disjoints $(\tau_{j-1}, \tau_j]$, $j = 1, \dots, J$ où $\tau_0 = 0 < \tau_1 < \dots < \tau_j < \dots < \tau_J$. La fonction de hasard

TABLEAU 3
Valeurs de la fenêtre optimale h_{opt} et du poids optimal P_{opt}

	Loi Log-Normale		Loi Gamma		Loi Weibull	
Variance*	0.7343		0.6002		0.1127	
	h_{opt}	\widehat{MISE}	h_{opt}	\widehat{MISE}	h_{opt}	\widehat{MISE}
$\hat{\lambda}_1(t)$	0.33	0.0934	0.56	0.4664	0.12	0.0781
$\hat{\lambda}_2(t)$	0.25	0.1134	0.36	0.5713	0.12	0.0787
$\hat{\lambda}_3(t)$	0.42	0.06843	0.64	0.3957	0.17	0.0666
	P_{opt}	\widehat{MISE}	P_{opt}	\widehat{MISE}	P_{opt}	\widehat{MISE}
$\hat{\lambda}_P(t)$	0.155	0.0602	0.195	0.2584	0.34	0.0667

* La variance moyenne des 300 échantillons simulés de la v.a T

3. Ce problème est connu sous le nom de paradoxe du *mover-stayer* (« *Mobile-Stable* ») (cf. ELBERS et RIDDER [1982] dans le cas des modèles paramétriques).

4. Cf. NASSIRI [1993] pour une présentation détaillée de ces modèles.

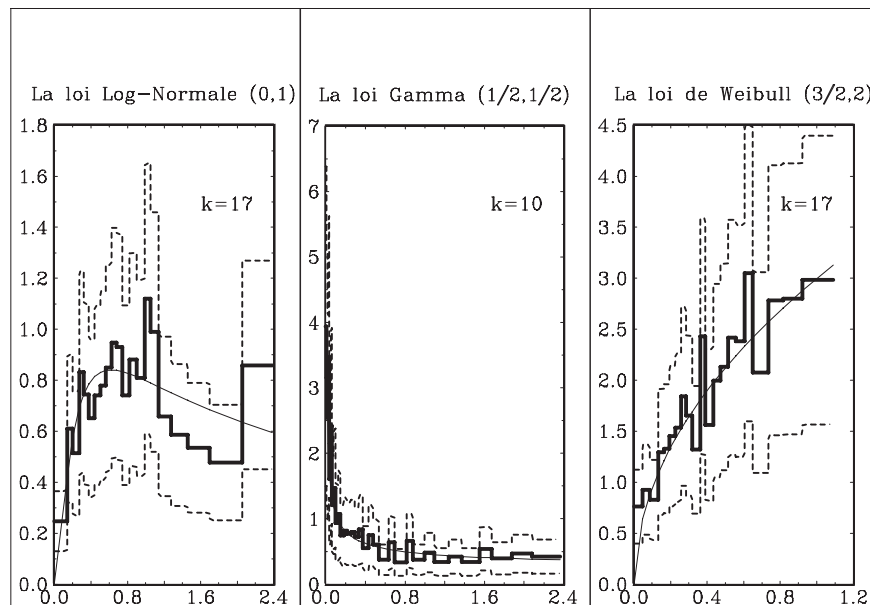
s'écrit pour tout individu i , $\lambda(t) = \lambda_j, \forall t \in (\tau_{j-1}, \tau_j], j = 1, \dots, J$, et la vraisemblance qui en découle $\prod_{j=1}^J (\lambda_j)^{\delta_j} (1 - \lambda_j)^{r_j - \delta_j}$ où δ_j représente le nombre d'individus qui sortent effectivement du chômage pendant le $j^{\text{ème}}$ intervalle : $\delta_j = \sum_{i \in R_j} \delta_{i,j}$, $\delta_{i,j}$ est une indicatrice qui prend la valeur 1 si le $i^{\text{ème}}$ individu sort effectivement du chômage au cours du $j^{\text{ème}}$ intervalle $(\tau_{j-1}, \tau_j]$, zéro sinon. $r_j = \sum_{i \in R_j} 1$ est le nombre des individus dont la durée

observée est supérieure à τ_{j-1} ; il s'agit de l'ensemble à risque dans le $j^{\text{ème}}$ intervalle.

Dans l'application de cet estimateur, le choix des J intervalles $(\tau_{j-1}, \tau_j]$, $j = 1, \dots, J$, est aussi crucial que le choix de la fenêtre dans les estimateurs non-paramétriques précédents : plus la fenêtre est large, plus la variance est importante et plus elle est petite, plus le biais est fort. En outre, dans les zones à forte densité des données censurées, il est nécessaire de fixer des intervalles de large amplitude afin de contenir au moins une observation non censurée permettant d'estimer le taux de hasard. Compte tenu de ces remarques faites lors de la mise en œuvre de cet estimateur, et dans le but de rendre ce dernier comparable à l'estimateur du $k^{\text{ème}}$ plus proche voisin non censuré, nous y adaptons la méthode des pas aléatoires. Pour cela, dans notre vraisemblance, au lieu des intervalles, nous fixons le nombre des observations non censurées, δ_j , qui doit y tomber. Dans notre vraisemblance, δ_j est égal à k et représente

GRAPHIQUE 2

Estimation du hasard constant par morceaux et de son intervalle de confiance à 0.95



l'équivalent des k plus proches voisins non censurés de l'estimateur non-paramétrique $\hat{\lambda}_P(t)$. Les intervalles sont ainsi d'amplitude aléatoire (comme c'est le cas pour $\hat{\lambda}_P(t)$), ainsi que les valeurs de r_j qui en découlent, et permettent de mieux s'adapter à la distribution des censures. Un balayage sur toute valeur possible de k , nous permet de retenir, dans le graphique 2, l'estimation qui donne le hasard constant par intervalle (escalier en gras) le plus proche du hasard théorique (courbe), ainsi que son intervalle de confiance à 95 % (escalier en tiret). Néanmoins, comme on peut le constater sur les graphes ci-dessus, les résultats des estimations ne sont pas aussi proches que $\hat{\lambda}_P(t)$ du hasard théorique, ce pour les trois lois : les valeurs du MISE se sont avérées largement supérieures à celles constatées pour les quatre autres estimateurs non-paramétriques.

Compte tenu de ces résultats, nous choisissons l'estimateur $\hat{\lambda}_P(t)$, dit du $k^{\text{ème}}$ plus proche voisin non censuré, pour l'estimation du taux de hasard sur données réelles de chômage.

4 Application aux données de chômage

Dans l'analyse du fonctionnement du marché du travail, les études portent souvent sur la durée et la fréquence des périodes de chômage. C'est ainsi qu'apparaît l'intérêt de la fonction de hasard dans la modélisation de ces durées. En effet, elle représente la probabilité qu'un individu sorte du chômage après une durée t déjà chômeé. Dans ces analyses, largement fondées sur les modèles de durées (Pour une étude bibliographique, voir FOUGÈRE [1986], BONNAL [1992] et NASSIRI [1993]), on est souvent amené à faire des hypothèses paramétriques sur la forme fonctionnelle de cette fonction⁵. Les résultats des estimations restent cependant tributaires de ces hypothèses et aboutissent à des biais qui sont d'autant plus importants que la spécification supposée est rigide et inexacte (MEYER [1986 ; 1990], HECKMAN et SINGER [1984]). Selon nous, l'intérêt majeur de ce nouvel estimateur, $\hat{\lambda}_P(t)$, est de permettre une meilleure estimation du taux de hasard qui renseigne sur sa forme fonctionnelle. Sur la base de ses résultats, il est, en outre, possible de faire le choix d'une forme paramétrique adéquate permettant de pallier le risque d'une mauvaise estimation due à une hypothèse paramétrique inexacte sur le modèle.

Les données dont on dispose sont relatives au devenir de 1394 chômeurs au sortir de leur période de chômage. Elles permettent de distinguer ceux qui ont trouvé un emploi⁶ ; les transitions vers les autres issues concurrentes donnent

5. Sauf dans le modèle de Cox [1975]. La vraisemblance, dite partielle, permet d'estimer le modèle sans avoir à spécifier la fonction de hasard qui reste, par ailleurs, indéterminée.

6. Nous ne distinguons pas entre les deux types d'emplois : emplois sur CDD ou CDI d'une part, et stages ou emplois subventionnés d'autre part. Pour une étude conjointe de ces deux types d'issues au chômage, nous renvoyons à NASSIRI [1993].

lieu à des observations censurées à droite. Dans ce cas de figure, la variable δ_i représente alors l'indication que le $i^{\text{ème}}$ chômeur a accédé à un emploi après sa période de chômage T_i . On dispose, aussi, d'un ensemble d'informations concernant les caractéristiques individuelles de ces chômeurs. Pour tenir compte de l'hétérogénéité observée dans l'échantillon, il est ainsi possible de stratifier l'échantillon selon des facteurs déterminants tels que le sexe ou l'ouverture des droits à l'indemnisation du chômage.

Dans une étude paramétrique faite par BONNAL et FOUGÈRE [1990] sur ces mêmes données ⁷, trois spécifications ont été retenues : la loi Log-Normale, la loi de Weibull et la loi Gamma. Les résultats des vraisemblances estimées pour ces trois spécifications paramétriques permettent aux auteurs de conclure que *le meilleur modèle est le modèle Log-Normal*. Aussi, ce sont des estimations de la fonction de survie par l'estimateur non-paramétrique de KAPLAN-MEIER, qui leur permettent de préciser que *les estimations de la loi Log-Normale se superposent bien à l'estimation de KAPLAN-MEIER à l'extrémité basse de la distribution, mais moins correctement à l'extrémité haute, alors que l'observation est inverse pour les lois Weibull et Gamma (notamment pour les hommes)*.

À l'instar de cette étude, nous stratifions par rapport à la variable sexe pour constituer deux échantillons : 632 femmes et 762 hommes. À chacun des deux échantillons, nous appliquons l'estimateur non-paramétrique du $k^{\text{ème}}$ plus proche voisin non censuré à la fonction de hasard. Les résultats obtenus permettent de valider ceux des modèles paramétriques. Ils nous permettent aussi d'apporter quelques précisions supplémentaires sur l'évolution temporelle de la probabilité instantanée de trouver un emploi et, notamment, sur la différence entre les femmes et les hommes.

Pour chaque strate de chômeurs, femmes ou hommes, nous calculons l'estimateur $\hat{\lambda}_P(t)$ pour plusieurs valeurs de P entre 0 et 1. Notre premier constat est évident : la fonction de hasard estimée est d'autant plus lisse que la valeur choisie du poids est grande. De façon empirique, nous retenons alors, comme poids optimal, la valeur de P pour laquelle le graphe de $\hat{\lambda}_P(t)$ n'est ni trop lisse ni trop perturbé : 0.15 pour les femmes et 0.10 pour les hommes. Nous retenons aussi des valeurs plus fortes (0.30 pour les femmes et 0.20 pour les hommes) pour rendre compte de la tendance au lissage qui en découle. Les résultats sont représentés dans le graphique 3 où nous juxtaposons aussi les estimations paramétriques obtenues par les deux auteurs.

On remarque que plus le poids P choisi est fort, et plus l'estimation non-paramétrique de la fonction de hasard qui en découle devient lisse, si lisse qu'elle aboutit à des formes fonctionnelles assez simples qui peuvent correspondre à une loi paramétrique ou à une combinaison de lois paramétriques bien choisie. En effet, avec un poids (0.30 pour les femmes et 0.20 pour les hommes) égal au double de la valeur retenue pour le poids optimal, il apparaît que la fonction de hasard qui caractérise aussi bien les femmes que les hommes est une combinaison de deux lois paramétriques. Comme l'ont affirmé les deux auteurs BONNAL et FOUGÈRE, il s'agit d'une loi Log-Normale sur l'extrémité basse avec des valeurs modales aux alentours de cent jours pour les femmes et de cinquante jours pour les hommes. Au-delà d'un point

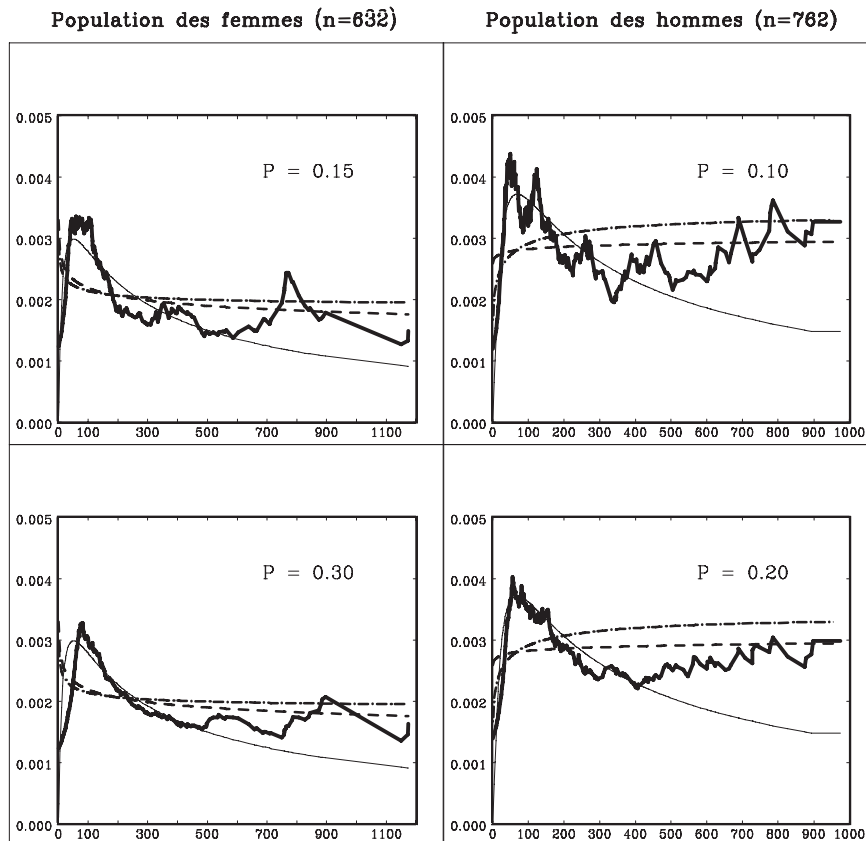
7. Pour plus de détails sur les résultats de ces deux auteurs ainsi que sur les données utilisées, nous renvoyons à leur travail.

d'inflexion, cette loi change en particulier d'un groupe à l'autre : pour le groupe des femmes, c'est une loi à hasard non monotone (type loi Log-Normale) qui admet un mode aux alentours de huit cents jours, alors que pour le groupe des hommes, c'est une loi à hasard monotone (type Weibull). Selon les résultats graphiques calculés à l'aide des valeurs retenues pour le poids optimal ⁸, le point d'inflexion se situe au niveau de six cents jours pour les femmes et de quatre cents jours pour les hommes. La combinaison de deux lois paramétriques introduirait un degré de flexibilité, mais elle fait disparaître d'importantes informations contenues dans l'échantillon, alors que ces informations sont bien présentes dans les résultats de l'estimateur non-paramétrique calculé avec le poids optimal.

En effet, la probabilité instantanée qu'un demandeur d'emploi, femme ou homme, sorte du chômage augmente sur les trois premiers mois du chômage ; celle des hommes étant plus forte que celle des femmes. Puis une chute survient. Chez les femmes, cette chute est plus forte et se stabilise à partir de

GRAPHIQUE 3

Estimations paramétriques et non-paramétriques du taux de hasard des trois lois : Log-Normale, Gamma et Weibull



8. 0.15 pour les femmes et 0.10 pour les hommes

deux cents jours avec une légère augmentation entre trois cents et quatre cents jours, et plus tard à partir de six cents jours en marquant un second pic aux environs de huit cents jours. Pour des durées de chômage supérieures à huit cents jours, la tendance est à la baisse. Chez les hommes les oscillations du taux de hasard estimé sont beaucoup plus importantes et les périodes de reprise plus nombreuses : à partir de trois mois, de six mois et particulièrement pour les durées supérieures à un an. Les particularités propres à chaque groupe, femmes ou hommes, sont encore plus nettes : on remarque qu'au-delà d'un an passé au chômage, si le taux de risque des femmes admet des oscillations qui ont une tendance à se stabiliser autour d'un niveau assez bas, celui des hommes connaît une légère tendance à la hausse.

Aussi bien pour les femmes que pour les hommes, on constate que la probabilité instantanée de sortie du chômage admet une tendance générale à la baisse quand le demandeur d'emploi reste au chômage plus de trois mois. On constate, aussi, qu'il est des épisodes pendant lesquels, cette tendance est soit stabilisée, soit contrecarrée à la hausse. Ces variations peuvent être le résultat d'un effet propre à la durée passée au chômage. Mais elles peuvent aussi s'expliquer par des facteurs déterminants dont on ne tient pas compte pour distinguer entre les membres de chaque groupe, ou entre les différentes issues qui leur sont offertes. Effectivement, on peut noter, par exemple, que la première reprise qui s'amorce aux alentours de deux cents jours correspond à la durée moyenne d'indemnisation (208 jours). La deuxième reprise qui commence au douzième mois de chômage, en particulier pour les hommes, se confond avec le début du chômage de longue durée (supérieure à un an) dont les victimes sont ciblées par divers dispositifs publics d'aide à l'emploi comme les stages de qualification (*cf.* NASSIRI [1993 ; 2000] pour plus de détails sur le traitement de l'hétérogénéité des individus et des issues).

5 Conclusion

Compte tenu des résultats empiriques des simulations, on peut conclure que, des différents estimateurs analysés dans cet article, *l'estimateur non-paramétrique du $k^{\text{ème}}$ plus proche voisin non censuré* est le plus performant pour approximer la fonction de hasard, en particulier des lois Log-Normale et Gamma. Dans le cas de la loi de Weibull, nous avons constaté que tous les estimateurs non-paramétriques donnent des résultats comparables. En raison de la censure à droite, il faut, toutefois, être prudent dans l'interprétation des résultats de ces estimateurs en queue de distribution, en particulier sur le dernier intervalle, par manque d'observations non censurées.

L'application de ce nouvel estimateur sur un fichier de données réelles et surtout la comparaison de ses résultats avec ceux des trois modèles paramétriques met en évidence que les formes fonctionnelles paramétriques constituent un lissage *brutal* du taux de risque. Ce lissage se trouve ainsi à l'origine d'une mauvaise estimation de la tendance temporelle de la probabilité instantanée de sortir du chômage, en particulier si celle-ci n'est pas monotone.

• Références bibliographiques

- AITKIN M., LAIRD N., FRANCIS B. (1983). – « A Reanalysis of the Stanford Heart Transplant Data », *Journal of American Statistical Association*, vol. 78, n° 382.
- BONNAL L. (1992). – « Microéconométrie de la durée de chômage et de la recherche d'emploi », *Thèse pour le Doctorat nouveau régime en Sciences Économiques*, Université de Toulouse-I.
- BONNAL L., FOUGÈRE D. (1990). – « Les déterminants individuels de la durée de chômage », *Économie et Prévision*, n° 96.
- BLUM J., SUSARLA V. (1980). – *Maximal Deviation Theory of Density and Failure Rate Function Estimates Based on Censored Data, Multivariate Analysis V*, Ed. by P. Krishnaiah, 213-222, North-Holland, Amsterdam.
- BOSQ D., LECOUTRE J.P. (1987). – *Théorie de l'estimation fonctionnelle*, *Economica*.
- BRESLOW N. (1974). – « Covariance Analysis of Censored Survival Data », *Biometrics* n° 30.
- COX D.R. (1975). – « Partial Likelihood », *Biometrika* n° 62.
- DEHEUVELS P., HOMINAL P. (1979). – « Estimation non-paramétrique de la densité compte tenu d'informations sur le support », *Revue de Statistique Appliquée*, vol. XXVII, n° 3.
- DROESBEKE J.J., FICHET B., TASSI Ph. (1989). – *Analyse statistique des durées de vie. Modélisation des données censurées*, *Economica*.
- EFRON B. (1967). – *The Two Sample Problem with Censored Data*, Proc. 5th Berkeley Symp., Vol. 4.
- ELBERS C., RIDDER A. (1982). – « True and Spurious Duration Dependence. The Identifiability of the Proportional Hazard Model », *Revue of Economic Studies* n° 49.
- FÖLDES A., REJTO L., WINTER B. (1981). – « Strong Consistency Properties of Nonparametric Estimations for Randomly Censored Data, II, Estimations of Density and Failure Rate », *Period. Math. Hungar*, 12.
- FOUGÈRE D. (1986). – « La recherche d'emploi : analyse théorique et étude économétrique », *Thèse pour le Doctorat de 3^e Cycle en Économie des Ressources Humaines*, Université de Toulouse-I.
- GREGOIRE G. (1991). – « Bandwidth Selection for the Intensity of Point. Application to Hazard Rate Models for Censored Data », *Scandinavian Journal of Statistics*.
- HASSANI S., SARDA P., VIEU Ph. (1986). – « Approche non-paramétrique en théorie de la fiabilité : revue bibliographique », *Revue de Statistique Appliquée*, Vol. XXXV, n° 4.
- HECKMAN J., SINGER B. (1984). – « Econometric Duration Analysis », *Journal of Econometrics* n° 24.
- KAPLAN S.L., MEIER P. (1958). – « Non-parametric Estimation from Incomplete Observations », *Journal of American Statistical Association*, 53.
- MARRON J.S., PADGETT W.J. (1987). – « Asymptotically Optimal Bandwidth Selection for Kernel Density from Randomly Right Censored Samples », *The Annals of Statistics* 15, n° 4.
- MC NICHOLS D.T., PADGETT W. J. (1983). – « A Modified Kernel Estimator from Randomly Right Censored Data », *Communication in Statistical, Analysis. Theory and Methods*, 13.
- MEYER B.D. (1986). – « Semiparametric Estimation of Hazard Models », *mimeo MIT*.
- MEYER B.D. (1990). – « Unemployment Insurance and Unemployment Spells ». *Econometrica*, Vol. 58, n° 4.
- MIELNICZUCK J. (1986). – « Some Asymptotic Properties of Kernel Estimators of a Density Function on Case of Censored Data », *The Annals of Statistics*, Vol. 14, n° 2.
- NASSIRI A. (1992). – « Validation empirique d'un estimateur non paramétrique du taux de hasard », *Cahier Gremaq* n° 92.29.276.
- NASSIRI A. (1993). – « Sortie du chômage et accès aux stages, une recherche économétrique de facteurs d'hétérogénéité », *Thèse pour le Doctorat Nouveau Régime en Sciences Économiques*, Université de Toulouse-I.

- PONS O. (1986). – « Vitesse de convergence des estimateurs à noyau pour l'intensité d'un processus ponctuel », *Statistics*, 17, n° 4.
- SARDA P., VIEU Ph. (1991). – « Smoothing Parameter Selection in Hazard Estimation », *Statistic and Probability Letters*, 11.
- TANNERS M.A., WONG W.H. (1983). – « The Estimation of the Hazard Rate Function from Right-Censored Data by the Kernel Method », *The Annals of Statistics*, Vol. 11, n° 3.
- WHITEHEAD J. (1980). – « Fitting Cox's Regression Model to Survival Data Using GLIM », *Applied Statistics*, 29.