Available online at www.sciencedirect.com
ScienceDirect
C. R. Acad. Sci. Paris, Ser. I 346 (2008) 907–912
COMPTES RENDUS MATHEMATIQUE
http://france.elsevier.com/direct/CRASS1/

Statistique

Estimation nonparamétrique de la distribution des revenus et de l'indice de pauvreté

Galaye Dia

LERSTAD, UFR de sciences appliquées et de technologie, B.P. 234, Université Gaston-Berger-de-Saint-Louis, Saint-Louis, Sénégal

Reçu le 23 juin 2007 ; accepté après révision le 3 juillet 2008

Présenté par Paul Deheuvels

Résumé

Dans cette Note nous proposons un estimateur de l'indice de pauvreté de Foster, Greer et Thorbecke. L'estimateur est construit à l'aide du noyau de Parzen Rosenblatt. La convergence de l'estimateur est étudiée et des simulations proposées. ***Pour citer cet article : G. Dia, C. R. Acad. Sci. Paris, Ser. I 346 (2008).***

Abstract

Nonparametric estimation of income distribution and poverty index. In this Note we propose an estimator of Foster, Greer and Thorbecke class of measures $P(z,\alpha)=\int_0^z(\frac{z-x}{z})^\alpha f(x)\,\mathrm{d}x$ where z is the poverty line, $f(x)$ the density of the income distribution and α the so-called poverty aversion. The estimator is constructed with Parzen–Rosenblatt kernel. Uniform almost sure consistency and uniform mean square consistency are established. A simulation study indicates that this new estimator performs well in finite samples. ***To cite this article: G. Dia, C. R. Acad. Sci. Paris, Ser. I 346 (2008).***

Abridged English version

Let F be the cumulative distribution function of the income variable X from a population. The F.G.T. class of poverty measures is defined by $P(z,\alpha)=\int_0^z(\frac{z-x}{z})^\alpha\,\mathrm{d}F(x)$, $\alpha>0$. If $z\leqslant 0$ we set $P(z,\alpha)=0$. These poverty measures are commonly estimated by $\hat{P}_n(z,\alpha)=\frac{1}{n}\sum_{i=1}^n(1-\frac{X_i}{z})_+^\alpha$ where $x_+=\max(x,0)$ [5]. We propose in this Note a new method of kernel estimate based on the Riemann sum, namely the following estimator

$$P_n(z,\alpha)=\frac{1}{n}\sum_{j=1}^{n}\sum_{i=0}^{[z/h]}\left(1-\frac{ih}{z}\right)^\alpha K\left(\frac{X_j-ih}{h}\right)\quad\text{for } \alpha=0 \text{ or } \alpha\geqslant 1.$$

Two hypotheses are made about the density f we suppose bounded. It is either uniformly continuous or admits almost everywhere a derivative f' belonging to $L_1(\mathbf{R})$. For each case, the convergence of the above estimator will be established. Additional hypotheses are made about the kernel K, that is:

Adresses e-mail : galayedia@hotmail.com, galaye@ugb.sn.

1631-073X/$ – see front matter
doi:10.1016/j.crma.2008.07.004

$\mathbf{A_1}$) K is of bounded variation $V_{-\infty}^{u}K$ on $\mathbf{R}$. Let $V(\mathbf{R})$ be its total variation.
$\mathbf{A_2}$) $\int_{\mathbf{R}} |uK(u)|\,\mathrm{d}u < +\infty$.
$\mathbf{A_3}$) There exists a nonincreasing function λ such that for all x, x' we have $|K(x) - K(x')| \leqslant \lambda(|x - x'|)$ and $\lambda(u) \to 0$, $u \to 0$, $u \geqslant 0$. Moreover $\lambda(\frac{u}{h}) = \mathrm{O}(h)$ on bounded intervals.

We can formulate now our main results:

Theorem 0.1. *Assume that*

(1) *the hypothesis* $\mathbf{A_1}$ *holds,*
(2) f *is uniformly continuous.*

Then for all $b > 0$

$$P\Big(\lim_{n\to+\infty} \sup_{z\in[0,b]} \big|P_n(z,\alpha) - P(z,\alpha)\big| = 0\Big) = 1$$

provided $nh^2(\operatorname{Log}\operatorname{Log} n)^{-1} \to +\infty$ *as* $n \to +\infty$.

Theorem 0.2. *Assume that the hypotheses* $\mathbf{A_1}$ *and* $\mathbf{A_2}$ *hold.*
Then, if f *admits almost everywhere a derivative* f' *belonging to* $L_1(\mathbf{R})$ *we have for all* $b > 0$

$$P\Big(\lim_{n\to+\infty} \sup_{z\in[0,b]} \big|P_n(z,\alpha) - P(z,\alpha)\big| = 0\Big) = 1$$

provided $nh^2(\operatorname{Log}\operatorname{Log} n)^{-1} \to +\infty$ *as* $n \to +\infty$.

Theorem 0.3. *If* f *is uniformly continuous or if hypothesis* $\mathbf{A_2}$ *holds and* f *admits almost everywhere a derivative* f' *belonging to* $L_1(\mathbf{R})$*, then we have under the hypothesis* $\mathbf{A_3}$

$$\lim_{n\to+\infty} \sup_{z\in[0,b]} E\big(P_n(z,\alpha) - P(z,\alpha)\big)^2 = 0,$$

i.e. $P_n(z,\alpha)$ *converges uniformly in quadratic mean on* $[0,b]$ *for all* $b > 0$.

We assess the performance of our estimator by giving a number of Monte Carlo experiments and compare the two estimators $P_n(z,\alpha)$ and $\hat{P}_n(z,\alpha)$ using the variance and the mean square quadratic error of the samples. Simulation studies are made in statistical package S-Plus.

1. Introduction et définition de l'estimateur

Soit $F(x)$ la distribution des revenus d'une population admettant une densité $f(x)$ continue. La famille d'indicateurs F.G.T. (Foster, Greer, Thorbecke) [1] indexés par un nombre réel $\alpha \geqslant 0$ est définie par :

$$P(z,\alpha) = \begin{cases} \displaystyle\int_0^z \left(\frac{z-x}{z}\right)^{\alpha} f(x)\,\mathrm{d}x & \text{si } z > 0, \\ 0 & \text{si } z \leqslant 0, \end{cases} \tag{1}$$

où z est appelé **seuil de pauvreté** ou **ligne de pauvreté**.

Note 1. On vérifie aisément que pour $\alpha = 0$, $P(z,\alpha) = F(x)$.

Soit $(X_1, \ldots, X_n)$ un échantillon de revenus de distribution F. L'estimateur suivant de $P(z,\alpha)$

$$\hat{P}_n(z,\alpha) = \frac{1}{n}\sum_{i=1}^{n}\left(1 - \frac{X_i}{z}\right)_+^{\alpha} \quad \text{où } x_+ = \max(0,x)$$

est l'estimateur empirique de l'indice F.G.T. de pauvreté utilisé amplement dans la pratique en économétrie et en actuariat [5]. Il est sans biais de variance égale à $n^{-1}(P(z,2\alpha)-(P(z,\alpha))^2)$. Nous nous proposons de construire un estimateur de cet indice par la méthode du noyau.

Considérons l' estimateur classique de la densité f : $\hat{f}(x)=(nh)^{-1}\sum_{j=1}^{n}K(\frac{X_j-x}{h})$ où h est positif fonction de n tendant vers zéro lorsque n tend vers l'infini et K une fonction borélienne vérifiant les hypothèses suivantes :

$$(\mathbf{H_1})\quad \sup_{-\infty<x<+\infty}\left|K(x)\right|<+\infty,\quad (\mathbf{H_2})\quad \int_{-\infty}^{+\infty}K(x)\,\mathrm{d}x=1,\quad (\mathbf{H_3})\quad \lim_{x\to\pm\infty}\left|xK(x)\right|=0. \tag{2}$$

Dans toute la suite de cette Note, nous supposons que les hypothèses $\mathbf{H_1}$, $\mathbf{H_2}$, $\mathbf{H_3}$ sont vérifiées, que K est Riemann intégrable, et que f est bornée de support inclus dans $\mathbf{R}_+$. On désigne par x_0 la borne inférieure de ce support. Substituons dans (1) f par $\hat{f}$. On obtient

$$\tilde{J}_n(z,\alpha)=\int_0^z\left(\frac{z-x}{z}\right)^{\alpha}(nh)^{-1}\sum_{j=1}^{n}K\left(\frac{X_j-x}{h}\right)\mathrm{d}x. \tag{3}$$

En posant $\Delta_{h,i}=[hi,h(i+1)[$ et en utilisant les sommes de Riemann dans la définition de l'intégrale, on établit qu'il correspond à l'intégrale $\tilde{J}_n(z,\alpha)$ la somme suivante :

$$\tilde{P}_n(z,\alpha)=\frac{1}{n}\sum_{j=1}^{n}\sum_{i=0}^{[z/h]}\left(1-\frac{ih}{z}\right)^{\alpha}K\left(\frac{X_j-ih}{h}\right)+\vartheta_n(z) \tag{4}$$

où $[\frac{*}{h}]$ désigne la partie entière de $\frac{*}{h}$ et $\vartheta_n(z)\to 0$ presque sûrement lorsque $n\to+\infty$. Nous proposons alors comme estimateur de l'indice de pauvreté F.G.T. l'estimateur suivant :

$$P_n(z,\alpha)=\frac{1}{n}\sum_{j=1}^{n}\sum_{i=0}^{[z/h]}\left(1-\frac{ih}{z}\right)^{\alpha}K\left(\frac{X_j-ih}{h}\right)\quad \text{pour } \alpha=0 \text{ ou } \alpha\geqslant 1. \tag{5}$$

Le cas $0<\alpha<1$ reste un problème ouvert. Nous aurons besoin des hypothèses suivantes :

($\mathbf{H_4}$) K est une fonction à variation $V_{-\infty}^{u}K$ bornée sur $\mathbf{R}$. Soit $V(\mathbf{R})$ sa variation totale.
($\mathbf{H_5}$) $\int_{\mathbf{R}}|uK(u)|\,\mathrm{d}u<+\infty$.
($\mathbf{H_6}$) Il existe une fonction décroissante λ telle que pour tout x,x' on a $|K(x)-K(x')|\leqslant\lambda(|x-x'|)$ et $\lambda(u)\to 0$, $u\to 0$, $u\geqslant 0$. De plus $\lambda(\frac{u}{h})=\mathrm{O}(h)$ sur tout intervalle borné.

2. Convergence de l'estimateur

Nos résultats sont relatifs aux hypothèses additionnelles suivantes sur f :

$\mathbf{C_1}$: f est uniformément continue.
$\mathbf{C_2}$: f admet une dérivée presque partout f' $L_1(\mathbf{R})$-intégrable.

2.1. Convergence uniforme presque sûre de $P_n(z,\alpha)$

Théorème 2.1. *Supposons que les hypothèses* $\mathbf{H_4}$ *et* $\mathbf{C_1}$ *soient vérifiées. Alors pour tout $b>0$, l'estimateur $P_n(z,\alpha)$ converge uniformément presque sûrement sur $[0,b]$ vers $P(z,\alpha)$ lorsque $n\to+\infty$ i.e.*

$$P\Big(\lim_{n\to+\infty}\sup_{z\in[0,b]}\left|P_n(z,\alpha)-P(z,\alpha)\right|=0\Big)=1,$$

pourvu que $nh^2(\mathrm{Log}\,\mathrm{Log}\,n)^{-1}\to+\infty$ lorsque $n\to+\infty$.

Théorème 2.2. *Supposons que les hypothèses* $\mathbf{H_4}$, $\mathbf{H_5}$ *et* $\mathbf{C_2}$ *soient vérifiées. Alors pour tout* $b > 0$

$$P\Big(\lim_{n\to+\infty}\sup_{z\in[0,b]}\big|P_n(z,\alpha)-P(z,\alpha)\big|=0\Big)=1,$$

pourvu que $nh^2(\operatorname{Log}\operatorname{Log} n)^{-1}\to+\infty$ *lorsque* $n\to+\infty$.

Pour montrer ces théorèmes, on utilise le théorème 2 de Kiefer [2] et les lemmes suivants montrant que l'estimateur $P_n(z,\alpha)$ est asymptotiquement uniformément sans biais sur tout intervalle borné.

Lemme 2.1. *Si l'hypothèse* $\mathbf{C_1}$ *est vérifiée, alors pour tout* $b > 0$, *on a* :

$$\lim_{n\to+\infty}\sup_{z\in[0,b]}\big|E\big(P_n(z,\alpha)\big)-P(z,\alpha)\big|\to 0,\quad n\to+\infty.$$

Lemme 2.2. *Si les hypothèses* $\mathbf{H_5}$ *et* $\mathbf{C_2}$ *sont vérifiées, alors* :

$$\sup_{z\in\mathbf{R}}\big|E\big(P_n(z,\alpha)\big)-P(z,\alpha)\big|\leqslant h\left(\Big(\int_{\mathbf{R}}\big|f'(x)\big|\,\mathrm{d}x\Big)\Big(\int_{\mathbf{R}}\big(|u|+1\big)\big|K(u)\big|\,\mathrm{d}u\Big)+(2\alpha M+Ah)\int_{-\infty}^{+\infty}\big|K(u)\big|\,\mathrm{d}u\right)\quad(6)$$

où $M=\sup_{z\in\mathbf{R}}\frac{F(z)}{z}$ *et* $A=\sup_{x\in\mathbf{R}}f(x)$.

Remarque 1. Si K vérifie l'hypothèse $\mathbf{H_5}$, alors en utilisant l'hypothèse $\mathbf{H_1}$, le noyau $\hat{K}=\frac{K^2}{\int_{\mathbf{R}}K^2(y)\,\mathrm{d}y}$ la vérifie aussi. Des deux lemmes précédents, on tire le corollaire suivant :

Corollaire 2.1. *Sous les conditions du Lemme* 2.1 (*resp.* 2.2), *on a uniformément sur* $[0,b]$ (*resp.* $\mathbf{R}$) :

$$\lim_{n\to+\infty}E\left(\sum_{i=0}^{[z/h]}\left(1-\frac{ih}{z}\right)^{2\alpha}K^2\left(\frac{X_j-ih}{h}\right)\right)=\Big(\int_{\mathbf{R}}K^2(y)\,\mathrm{d}y\Big)P(z,2\alpha).$$

Corollaire 2.2. *Si les conditions du Théorème* 2.2 *sont vérifiées et si* $h=\mathrm{O}(n^{-1}\operatorname{Log}\operatorname{Log} n)^{1/4}$, *alors pour tout* $b>0$ *on a presque sûrement* :

$$\sup_{z\in[0,b]}\big|P_n(z,\alpha)-P(z,\alpha)\big|=\mathrm{O}\big(n^{-1}\operatorname{Log}\operatorname{Log} n\big)^{1/4}.$$

2.2. Convergence uniforme en moyenne quadratique

Théorème 2.3. *Si les hypothèses* $\mathbf{H_6}$ *et* $\mathbf{C_1}$ *sont vérifiées, alors* :

(1) $\lim_{n\to+\infty} n\operatorname{Var}(P_n(z,\alpha))=(\int_{\mathbf{R}}K^2(y)\,\mathrm{d}y)P(z,2\alpha)-(P(z,\alpha))^2$.
(2) *Pour tout* $b>0$, $\lim_{n\to+\infty}\sup_{z\in[0,b]}E(P_n(z,\alpha)-P(z,\alpha))^2=0$.

Théorème 2.4. *Supposons les hypothèses* $\mathbf{H_6}$ *et* $\mathbf{C_2}$ *vérifiées. Alors* :

(1) $\lim_{n\to+\infty} n\operatorname{Var}(P_n(z,\alpha))=(\int_{\mathbf{R}}K^2(y)\,\mathrm{d}y)P(z,2\alpha)-(P(z,\alpha))^2$.
(2) *Si de plus l'hypothèse* $\mathbf{H_5}$ *est satisfaite, on a pour tout* $b>0$,

$$\lim_{n\to+\infty}\sup_{z\in[0,b]}E\big(P_n(z,\alpha)-P(z,\alpha)\big)^2=0.$$

Pour la démonstration de ce théorème, on suppose l'hypothèse $\mathbf{C_1}$ ou $\mathbf{C_2}$ satisfaite et on démontre d'abord le Théorème 2.5 ci-dessous en utilisant le

Lemme 2.3. *Soient* $0 \leqslant \theta_i \leqslant 1$; $i = 1, 2$. *Alors pour tout* x, y *et* $x \neq y$ *on a* :

$$\lim_{n\to+\infty} \sup_{(\theta_1,\theta_2)\in[0,1]\times[0,1]} \left(h^{-2} \int_{-\infty}^{+\infty} \left| K\left(\frac{u-x+\theta_1 h}{h}\right) K\left(\frac{u-y+\theta_2 h}{h}\right) \right| f(u)\,\mathrm{d}u \right) = 0.$$

Théorème 2.5. *Supposons l'hypothèse* $\mathbf{H_6}$ *vérifiée. Alors pour tout* $b > 0$,

$$\lim_{n\to+\infty} \sup_{z\in[0,b]} \sum_{0\leqslant i\neq j\leqslant[\frac{z}{h}]} \left(1-\frac{ih}{z}\right)^{\alpha} \left(1-\frac{jh}{z}\right)^{\alpha} \int_{\mathbf{R}} K\left(\frac{u-ih}{h}\right) K\left(\frac{u-jh}{h}\right) f(u)\,\mathrm{d}u = 0.$$

La preuve du Théorème 2.3 (resp. 2.4) résulte alors du Lemme 2.1 (resp. 2.2) et du Corollaire 2.1.

Note 2. Le Lemme 2.3 est établi par Masry [3] avec $\theta_1 = \theta_2 = 0$ et le facteur de l'intégrale égal à h^{-1}.

Remarque 2. L'estimateur $P_n(z,\alpha)$ a pour efficacité asymptotique par rapport à $\hat{P}_n(z,\alpha)$, $e(z,\alpha) = ((\int_{\mathbf{R}} K^2(y)\,\mathrm{d}y) \times P(z,2\alpha) - (P(z,\alpha))^2)/P(z,2\alpha) - (P(z,\alpha))^2$. L'intégrale $\int_{\mathbf{R}} K^2(y)\,\mathrm{d}y$ est strictement inférieure à 1 pour les noyaux usuels [4, p. 1068]. On a alors dans ce cas $e(z,\alpha) < 1$. Dans le Théorème 2.4, la vitesse de convergence en moyenne quadratique est de l'ordre de $\mathrm{O}(\frac{1}{n})$ si h est de l'ordre de $\mathrm{O}(\frac{1}{\sqrt{n}})$.

2.3. Simulations

Nous avons fait des simulations donnant l'erreur quadratique moyenne et la variance de 50 échantillons de taille n des deux estimateurs que nous avons comparés. Le noyau de Gauss qui vérifie les hypothèses $\mathbf{H_i}$, $i = 1, \ldots, 6$, a été utilisé en prenant $h = 1/\sqrt{n \operatorname{Log} n}$. Pour une distribution de type Paréto sur $[0, 1]$ de paramètres $x_0 = 0{,}02$ et $b = 0{,}2$, nous avons calculé l'erreur quadratique moyenne $msqe1$ de $(P_{n,1}(z,\alpha), \ldots, P_{n,50}(z,\alpha))$ et $msqe2$ de $(\hat{P}_{n,1}(z,\alpha), \ldots, \hat{P}_{n,50}(z,\alpha))$ ainsi que les variances respectives σ_1 et σ_2 pour différentes valeurs de z par les formules suivantes :

$$\overline{P_n(z,\alpha)} = \frac{1}{50}\sum_{i=1}^{50} P_{n,i}(z,\alpha), \quad msqe1 = \frac{1}{50}\sum_{i=1}^{50}\big(P_{n,i}(z,\alpha) - P(z,\alpha)\big)^2,$$

$$\sigma_1 = \frac{1}{49}\sum_{i=1}^{50}\big(P_{n,i}(z,\alpha) - \overline{P_n(z,\alpha)}\big)^2.$$

$\overline{\hat{P}_n(z,\alpha)}$, $msqe_2$ et σ_2 sont calculées de façon analogue pour l'estimateur $\hat{P}_n(z,\alpha)$.

Tableau 1

z	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8
$\alpha = 0, n = 5000$								
$msqe1$	0,00005502	0,00005216	0,00003863	0,00003361	0,00002962	0,00001934	0,00001481	6,88e–006
$msqe2$	0,00005574	0,00005092	0,00004104	0,00003374	0,00002943	0,00001981	0,00001516	6,62e–006
σ_1	0,00005495	0,00004786	0,00003941	0,00003429	0,00002959	0,0000196896	0,00001370	6,32e–006
σ_2	0,00005574	0,00005080	0,00003983	0,00003443	0,00002996	0,0000196898	0,00001412	6,59e–006
$\alpha = 1, n = 5000$								
$msqe1$	0,00002045	0,00002754	0,000029039	0,00002777	0,0000258544	0,000023532	0,000020740	0,000018020
$msqe2$	0,00002074	0,00002757	0,000029044	0,00002776	0,0000258545	0,000023534	0,000020741	0,000018024
σ_1	0,00002082	0,000028051	0,00002962	0,000028260	0,000026316	0,000023980	0,000021150	0,000018384
σ_2	0,00002085	0,000028052	0,00002963	0,000028266	0,000026319	0,000023984	0,000021152	0,000018388
$\alpha = 2, n = 5000$								
$msqe1$	9,7601e–006	0,00001848	0,00002350	0,00002622	0,00002738	0,00002718	0,00002639	0,00002520
$msqe2$	9,7653e–006	0,00001901	0,00002338	0,00002644	0,00002743	0,00002728	0,00002647	0,00002527
σ_1	9,6411e–006	0,00001849	0,00002324	0,00002595	0,000027008	0,00002687	0,000026041	0,000024823
σ_2	9,5640e–006	0,00001846	0,00002323	0,00002594	0,000027004	0,00002686	0,000026040	0,000024822

Remarque 3. Notons $MSQE_1$ (resp. $MSQE_2$) l'erreur quadratique moyenne de $P_n(z,\alpha)$ (resp. $\hat{P}_n(z,\alpha)$). En utilisant la décomposition classique de l'erreur quadratique moyenne et comme $nh^2 \to 0$ quand $n \to +\infty$ on a, d'après le Lemme 2.2 et le Théorème 2.4, $\frac{MSQE_1}{MSQE_2} \to e(z,\alpha)$ quand $n \to +\infty$.

Les calculs sont faits sur les trois indices classiques $P(z,0)$, $P(z,1)$, $P(z,2)$ communément appelés respectivement **le taux de pauvreté** (headcount ratio), **la profondeur de la pauvreté** (poverty gap index or depth of poverty) et **la sévérité de la pauvreté** (severity of poverty index) [1].

Un tableau comparatif des résultats des simulations montre que pour de petits échantillons de taille inférieure ou égale à 1000, en tout point z, notre estimateur a une plus petite variance pour les trois valeurs de α considérées. Pour de grands échantillons, le Tableau 1 illustre la **Remarque 3**. Nous pouvons conclure que notre estimateur est recommandable pour les petits échantillons et aussi bon que $\hat{P}_n(z,\alpha)$ pour les grands échantilllons.

Références

[1] J.E. Foster, J. Greer, E. Thorbecke, A class of decomposable poverty measures, Econometrica 52 (1984) 761–776.

[2] J. Kiefer, On large deviations of the empirical distribution of vector chance variable and law of iterated logarithm, Pacific J. Math. 11 (1961) 143–154.

[3] E. Masry, Recursive probability density estimation for weakly dependent stationary process, IEEE Trans. Inform. Theory IT-32 (2) (1986) 254–267.

[4] E. Parzen, On estimation of probability density function and mode, Ann. Math. Statist. 33 (1962) 1065–1076.

[5] C. Seidl, Poverty measurement: a survey, in: D. Bös, M. Rose, C. Seidl (Eds.), Welfare and Efficiency in Public Economics, Springer-Verlag, Heidelberg, 1988, pp. 71–147.