JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

SANDRINE POIRAUD-CASANOVA
CHRISTINE THOMAS-AGNAN

Quantiles conditionnels

Journal de la société statistique de Paris, tome 139, n° 4 (1998), p. 31-44

<http://www.numdam.org/item?id=JSFS_1998__139_4_31_0>

QUANTILES CONDITIONNELS

Sandrine POIRAUD-CASANOVA,
Christine THOMAS-AGNAN*

G.R.E.M.A.Q., Université des Sciences Sociales
et Laboratoire de Statistique et Probabilités,
Université Paul Sabatier, Toulouse, France *

Résumé

Cette étude est un travail de synthèse sur les estimateurs des quantiles conditionnels dans les modèles de régression non paramétriques. Après avoir indiqué les domaines d'application, nous explicitons, pour chaque technique présentée, les motivations, le principe de calcul des estimateurs et leurs propriétés asymptotiques.

I. INTRODUCTION

On rencontre fréquemment, par exemple en médecine et en fiabilité, le problème d'estimer des quantiles d'une distribution. Rappelons que pour une proportion α fixée, le quantile d'ordre α de la loi d'une variable aléatoire X est le seuil dépassé par $(1-\alpha)\,\%$ des réalisations de X. En médecine, les praticiens aiment comparer un individu à des normes en dehors desquelles il est nécessaire de le soumettre à un bilan complet : ils se réfèrent souvent pour cela à la médiane, mais aussi au 97^{e} percentile. En fiabilité, PADGETT et LIO (1993) citent l'exemple de la détermination de la force à exercer sur des fibres de carbone pour obtenir leur point de rupture : l'ingénieur s'intéresse alors au seuil de force dépassé par la moitié des fibres, ou par 90 % des fibres.

Par ailleurs, la distribution étudiée dépend usuellement de covariables et, dans ce cas, on parlera de quantiles conditionnels ou quantiles de régression. Par exemple, les courbes de croissance des enfants présentent les quantiles de la taille ou du poids en fonction de l'âge et du sexe (COLE et GREEN (1992)). KOENKER *et al.* (1994) étudient la relation entre la vitesse de course maximale et la masse corporelle de mammifères terrestres au travers des quantiles conditionnels. SONESSON *et al.* (1993) établissent des valeurs de référence pour les limites physiologiques des cordons ombilicaux de foetus humains en fonction de leur âge. L'intérêt pour ce type de questions n'est pas récent puisqu'on trouve une application à des données de salaire dans HOGG (1975). D'autre part, l'estimation des quantiles d'ordre $\alpha/2$ et $1-\alpha/2$ d'une loi conditionnelle conduisent à la définition d'un intervalle de prédiction non paramétrique de niveau de confiance $1-\alpha$ (CSÖRGŐ et RÉVÉSZ (1984)). De plus, l'utilisation de quantiles conditionnels peut être justifiée par des préoccupations de robustesse, ce qui est le cas particulièrement pour la

* E-mail : cthomas@cict.fr

médiane conditionnelle considérée comme une alternative à l'étude de la moyenne conditionnelle lorsque la distribution n'est pas symétrique.

II. LES MODÈLES

II.1. Quantiles non conditionnels

Avant de se pencher sur les modèles de régression, parlons brièvement de l'estimation non paramétrique des quantiles non conditionnels en commençant par rappeler quelques définitions. Soit $Y_1, \ldots, Y_n$ un échantillon de taille n d'une variable aléatoire de fonction de répartition $F(.)$. On rappelle que le quantile d'ordre α de F est défini par :

$$q_\alpha = \inf \{y \,/ F(y) \geqslant \alpha\} := F^{-1}(\alpha) \tag{1}$$

pour α élément de $[0,1]$. F^{-1} est appelée inverse généralisée de F et, si F est strictement croissante, F^{-1} est égale à la fonction réciproque de F. Soit $Y_{(1)} \leqslant \cdots \leqslant Y_{(n)}$ la statistique d'ordre et soit F_n la fonction de répartition empirique des Y_i : $F_n(y) = \dfrac{1}{n} \sum_{i=1}^{n} 1(Y_i \leqslant y)$. Notons que le quantile d'ordre α de F_n appelé quantile empirique d'ordre α de l'échantillon et noté q_α^n vaut $Y_{(n\alpha)}$ si $n\alpha$ est entier et $Y_{([n\alpha]+1)}$ sinon, où $[k]$ désigne la partie entière de k. Une longue littérature existe sur les propriétés probabilistes et statistiques des quantiles empiriques (voir par exemple Csörgő et Horváth (1993)). En sus de l'estimation classique du quantile théorique par le quantile empirique, certains auteurs proposent des combinaisons linéaires plus sophistiquées des statistiques d'ordre dans le but d'améliorer l'écart quadratique moyen. Cheng et Parzen (1997) recensent les diverses propositions de telles combinaisons et proposent ensuite une approche unifiée de ces estimateurs sous la forme d'une transformation intégrale de la fonction quantile empirique :

$$\widehat{q}_\alpha = \int_0^1 q_\beta^n d_\beta K(\alpha, \beta)$$

où, pour chaque α, $K(\alpha, .)$ est une fonction de répartition. Citons quelques exemples d'estimateurs de cette famille. Harrell et Davis (1982) proposent l'estimateur

$$\widetilde{q}_\alpha = (1 - (n+1)\alpha + [(n+1)\alpha]) Y_{([(n+1)\alpha])} + ((n+1)\alpha - [(n+1)\alpha]) Y_{([(n+1)\alpha+1])}.$$

Dans le cas particulier $\alpha = 1/2$, cet estimateur coïncide avec la médiane empirique classique de la distribution. Ils introduisent aussi un autre estimateur combinaison linéaire des statistiques d'ordre construit à partir d'une méthode des moments avec un argument limite. Parzen (1979) étudie le cas où le noyau $K(\alpha, \beta)$ est de type noyau de convolution

$$d_\beta K(\alpha, \beta) = h^{-1} K\left(\frac{\alpha - \beta}{h}\right) d\beta$$

où K est un noyau de probabilités. Un calcul simple montre qu'alors

$$\widehat{q}_\alpha = \sum_{i=1}^{n} \frac{1}{h_n} \left(\int_{\frac{i-1}{n}}^{\frac{i}{n}} K\left(\frac{\alpha - y}{h_n}\right) dy \right) Y_{(i)}.$$

Une approximation de ce dernier définie par

$$\overline{q}_\alpha = \sum_{i=1}^{n} \frac{1}{nh_n} K\left(\frac{\alpha - i/n}{h_n}\right) Y_{(i)}$$

est étudiée par YANG (1985). FALK (1984) établit une expression asymptotique quantifiant la performance de l'estimateur à noyaux et montrant son inefficacité relative par rapport au quantile empirique. D'autres propositions de ce type se trouvent dans KAIGH et LACHENBRUCH (1982), ZELTERMAN (1990), SHEATER et MARRON (1990), KAIGH et CHENG (1991). D'autre part, une autre approche consiste à inverser un estimateur de la fonction de répartition et c'est ainsi qu'on peut obtenir des estimateurs convergents des quantiles à partir, voir par exemple YAMATO (1973), d'estimateurs à noyau de la densité et de la fonction de répartition.

II.2. Quantiles conditionnels

Nous recensons d'abord les divers modèles de régression rencontrés dans la littérature. Nous nous limiterons au cas où les variables à expliquer et explicatives sont unidimensionnelles. On distingue en premier lieu les modèles selon la nature de la variable explicative :

- Le modèle à plan aléatoire de type (A) dont les données sont des couples $(Y_1, T_1), ..., (Y_n, T_n)$ de variables aléatoires réelles i.i.d. de même loi continue que (Y, T).
- Le modèle à plan fixe de type (F) dont les données sont des couples $(Y_1, t_1), ..., (Y_n, t_n)$ où les observations Y_i sont des variables aléatoires réelles indépendantes continues et les t_i sont des points d'observation réels non aléatoires.

$F_t(.)$ désigne la fonction de répartition de Y sachant $T = t$ dans le modèle à plan aléatoire et la fonction de répartition de Y à t fixé dans le modèle à plan fixe. Par la suite, on fera référence à la fonction F_t en utilisant le terme de fonction de répartition conditionnelle, même dans le cas du modèle à plan fixe. Le quantile conditionnel $q_\alpha(t)$ est le quantile d'ordre α de la distribution F_t défini par (1). Par ailleurs, on peut faire une hypothèse additionnelle sur F_t, à savoir : $F_t(y) = G(y - f(t))$ quels que soient y et t, où G est une fonction continue, strictement croissante et f est une fonction inconnue. Remarquons cependant que si l'on pose $G(0) = \alpha_0$ avec $\alpha_0 \in [0, 1]$, alors $f = q_{\alpha_0}$ et les fonctions quantiles pour différentes valeurs de α sont toutes parallèles. En effet, $F_t(f(t)) = G(0) = \alpha_0$ et d'autre part, pour $\alpha_0 \in [0, 1]$, $F_t(q_\alpha(t)) = \alpha \Leftrightarrow G(q_\alpha(t) - q_{\alpha_0}(t)) = \alpha \Leftrightarrow q_\alpha(t) = q_{\alpha_0}(t) + G^{-1}(\alpha)$.

Un modèle contenant cette hypothèse sera dit de type particulier et noté (P), sinon de type général et noté (G).

Les modèles que nous étudions seront désormais caractérisés par un couple de lettres selon que le modèle considéré sera à plan aléatoire ou fixe, de type (P) ou (G).

Enfin, on peut trouver des articles où les auteurs explicitent le terme d'erreur en posant dans ce cadre d'études :

$$e_i = Y_i - q_\alpha(T_i) \text{ si le type est } (A).$$
$$e_i = Y_i - q_\alpha(t_i) \text{ si le type est } (F).$$

Remarquons que les e_i sont alors i.i.d. dans un modèle à plan aléatoire mais seulement indépendantes dans un modèle à plan fixe. De plus, dans un modèle avec erreurs, on a :

$P(e_i \leqslant 0/T_i = t_i) := P_{t_i}(e_i \leqslant 0) = P_{t_i}(Y_i \leqslant q_\alpha(t_i)) = \alpha$ si le type est (A).

$P(e_i \leqslant 0) = P_{t_i}(e_i \leqslant 0) = \alpha$ si le type est (F).

Notons enfin que si l'on considère un modèle du type (FP), alors celui-ci est équivalent à la donnée d'un modèle où les e_i définies ci-dessus sont indépendantes et identiquement distribuées. En effet, un calcul simple montre l'équivalence entre $F_t(y) = G(y - q_\alpha(t))$ et $Y_i = q_\alpha(t_i) + e_i$ où les e_i sont i.i.d. de fonction de répartition G.

Sans vouloir faire un recensement systématique dans le cadre paramétrique, citons l'approche de KOENKER et BASSETT (1978 et 1982) pour un modèle de type (FP) où les points d'observation t_i sont, dans ce cas, des vecteurs de R^d avec $d < n$. Ils adaptent le principe des moindres carrés utilisé pour l'estimation de la moyenne conditionnelle au cas des quantiles en remplaçant la fonction carrée par la fonction ρ_α définie par :

$$\rho_\alpha(y) = \alpha y \quad \text{si} \quad y \geqslant 0 \quad \text{et} \quad \rho_\alpha(y) = (\alpha - 1)\, y \quad \text{si} \quad y \leqslant 0.$$

On retrouvera ce même principe dans les méthodes d'estimation du paragraphe 4. Les auteurs proposent alors une estimation des quantiles en cherchant une solution au problème :

$$\operatorname*{Min}_{b \in R^d} \sum_{i=1}^{n} \rho_\alpha(y_i - t_i' b).$$

L'estimateur de $q_\alpha(t)$ est alors défini par $\widehat{q}_\alpha(t) = t' b^*$ où b^* est une solution au problème de minimisation. Les auteurs démontrent un résultat de normalité asymptotique pour le vecteur des paramètres b^*.

Dans le cadre des modèles non paramétriques, on distinguera deux types de techniques d'estimation. Le premier type, décrit dans le paragraphe 3, consiste à estimer d'abord la fonction de répartition conditionnelle et à l'inverser pour obtenir le quantile conditionnel. Le second type, décrit dans le paragraphe 4, consiste en une estimation directe basée sur la propriété suivante des quantiles.

Le quantile empirique q_α^n de l'échantillon $(X_1, \ldots, X_n)$ satisfait à :

$$q_\alpha^n \in \operatorname*{Arg\,min}_{\beta \in R} \sum_{i=1}^{n} \rho_\alpha (X_i - \beta) . \tag{2}$$

Parallèlement, le quantile théorique d'une distribution de fonction de répartition F satisfait à :

$$Q_\alpha \in \operatorname*{Arg\,min}_{\beta \in R} \int \rho_\alpha (y - \beta)\, dF(y) . \tag{3}$$

III. ESTIMATION PAR INVERSION DE LA FONCTION DE RÉPARTITION

D'après (1), il est naturel de définir un estimateur $\widehat{q}_\alpha(t)$ du quantile de régression en «inversant» un estimateur $\widehat{F}_t(.)$ de la fonction de répartition conditionnelle :

$$\widehat{q}_\alpha (t) = \widehat{F}_t^{-1} (\alpha) = \inf \left\{ y / \widehat{F}_t (y) \geqslant \alpha \right\} .$$

On va donc décrire un certain nombre d'estimateurs des quantiles basés sur divers estimateurs de la fonction de répartition conditionnelle.

III.1. Estimateurs à noyaux de la fonction de répartition conditionnelle

III.1.1 *Les estimateurs du type* $\widehat{F}_t (y) = \sum_{i=1}^{n} \alpha_{ni} (t)\, 1 (Y_i \leqslant y)$

Avec la contrainte $\sum_{i=1}^{n} \alpha_{ni} (t) = 1$, les estimateurs de cette forme sont nécéssairement des fonctions de répartition et le calcul de l'inverse généralisé est explicite :

$$\widehat{q}_\alpha (t) = \inf \left\{ y / \widehat{F}_t (y) \geqslant \alpha \right\} = \inf \left\{ y / \sum_{y_i \leqslant y} \alpha_{ni} (t) \geqslant \alpha \right\} . \tag{4}$$

Pour évaluer $\widehat{q}_\alpha (t)$, il suffit donc de connaître les valeurs des poids $\alpha_{ni} (t)$ et la fonction obtenue $\widehat{q}_\alpha$ est en escalier par rapport à α. Remarquons que si $\widehat{F}_t$ est remplacée par la fonction de répartition empirique à t fixé des Y_i qui correspond aux poids $\alpha_{ni} (t) = 1/n$, alors on retrouve la définition du quantile empirique.

STONE (1977) donne des conditions sur les poids α_{ni} pour que la suite d'estimateurs ainsi définis converge ponctuellement en probabilités.

DUCHARME *et al.* (1995) et STUTE (1986) définissent dans un modèle de type (AG) les poids suivants :

$$\alpha_{ni}(t) = \frac{K\left(\frac{F_n(t) - F_n(T_i)}{h_n}\right)}{\sum_{i=1}^{n} K\left(\frac{F_n(t) - F_n(T_i)}{h_n}\right)},$$

où K est un noyau de probabilités continu et borné sur $[-1,1]$ et F_n désigne la fonction de répartition empirique des T_i à savoir : $F_n(t) = (1/n)\sum_{i=1}^{n} 1(T_i \leqslant t)$.

Ils obtiennent les résultats de convergence suivants :

i) Si $h_n \to 0$ et si $nh_n \to \infty$, alors $\underset{y}{\text{Sup}} \left|\widehat{F}_t(y) - F_t(y)\right| \underset{p.s.}{\to} 0$.

ii) Si $nh_n^3 \to \infty, nh_n^5 \to 0, R(K) = \int K^2(x)\,dx < \infty, \int xK(x)\,dx = 0$ alors

$\sqrt{nh_n}\,(\widehat{q}_\alpha(t) - q_\alpha(t)) \underset{L}{\to} \mathcal{N}\left(0, R(K)\,\sigma_\alpha^2(t)\right)$ où $\sigma_\alpha^2(t) = \alpha(1-\alpha)/f_t^2(q_\alpha(t))$

avec $f_t(y)$ densité de Y sachant $T = t$.

L'estimateur de HART (1991) est défini, dans un modèle de type (AG), par les poids suivants :

$$\alpha_{ni}(t) = \frac{1}{h_n} \int_{\frac{(i-1)}{n}}^{\frac{i}{n}} K\left(\frac{F_n(t) - u}{h_n}\right) du,$$

où K est un noyau de probabilités de support $[0,1]$.

Dans l'article de ANTOCH et JANSSEN (1989), les auteurs utilisent dans un modèle de type (FG) la caractérisation (3) du quantile. Le problème d'optimisation (3) est équivalent à : $\int \rho'_\alpha(y-t)\,dF_t(y) = 0$.

Pour une fonction de répartition G et un réel $\alpha \in [0,1]$, ces auteurs définissent

$$\lambda_{\alpha,G}(s) = -\int \rho'_\alpha(y-s)\,dG(y).$$

Mais on remarque que :

$$\lambda_{\alpha,G}(s) = -\alpha + G(s). \tag{5}$$

En effet :

$$\begin{aligned}
\lambda_{\alpha,G}(s) &= -\alpha \int_{y>s} dG(y) - (\alpha - 1)\int_{y<s} dG(y) \\
&= -\alpha\left(1 - \textstyle\int_{y<s} dG(y)\right) - (\alpha-1)\int_{y<s} dG(y) \\
&= -\alpha + \int_{y<s} dG(y) = -\alpha + G(s).
\end{aligned}$$

On a alors la caractérisation suivante : $q_\alpha(t) = \inf\{y/\lambda_{\alpha,F_t}(y) \geqslant 0\} := \lambda^{-1}_{\alpha,F_t}(0)$. En effet, d'après (5),

$$\lambda_{\alpha,F_t}(y) \geqslant 0 \Leftrightarrow F_t(y) \geqslant \alpha. \tag{6}$$

Donc $\inf\{y/\lambda_{\alpha,F_t}(y) \geqslant 0\} = \inf\{y/F_t(y) \geqslant \alpha\} = q_\alpha(t)$.
Il est donc naturel d'estimer $q_\alpha(t)$ par

$$\widehat{q}_\alpha(t) = \lambda^{-1}_{\alpha,\widehat{F}_t}(0)$$

et, d'après la relation (6), cette formulation est donc équivalente à celle de la formule (4).

Par contre, ANTOCH et JANSSEN (1989) utilisent des poids différents de ceux de DUCHARME *et al.* (1995) :

$$\begin{cases} \alpha_{ni}(t) = h_n^{-1}\displaystyle\int_{t_{i-1}}^{t_i} K\left(\frac{t-z}{h_n}\right)dz & \text{pour } 2 \leqslant i \leqslant n-1 \\ \alpha_{n1}(t) = h_n^{-1}\displaystyle\int_{-\infty}^{t_1} K\left(\frac{t-z}{h_n}\right)dz \\ \alpha_{nn}(t) = h_n^{-1}\displaystyle\int_{t_{n-1}}^{\infty} K\left(\frac{t-z}{h_n}\right)dz \end{cases}$$

où K est un noyau de probabilités.

III.1.2 *L'estimateur de* SAMANTA *de la médiane conditionnelle*

Dans un modèle de type (AG), SAMANTA (1989) et BERLINET, GANNOUN et MATZNER-LOBER (1997) effectuent préalablement à l'inversion un lissage par noyaux de type PARZEN-ROSENBLAT de la fonction de densité marginale $g(t)$ de T et de la fonction de densité conjointe $f(t,y)$ de (T,Y). En utilisant la relation classique entre la fonction de répartition conditionnelle et ces deux densités, ils posent :

$$\widehat{F}_t(y) = \int_{-\infty}^{y} f_n(u/t)\,du$$

où $f_n(y/t)) = f_n(t,y)/g_n(t)$ avec $g_n(t) = (nh_n)^{-1}\displaystyle\sum_{i=1}^{n} K\left(\frac{t-T_i}{h_n}\right)$ et $f_n(t,y) = (nh_n^2)^{-1}\displaystyle\sum_{i=1}^{n} K_1((t-T_i)/h_n)\,K_2((y-Y_i)/h_n)$, où K, K_1 et K_2 sont des noyaux de probabilités.

Sous les hypothèses énoncées par BERLINET, CADRE et GANNOUN (1998), on a le résultat de convergence suivant :

$$\sqrt{nh_n}\left(\widehat{F}_t(y) - F_t(y)\right) \underset{L}{\to} N\left(0, \sigma^2(t,y)\right)$$

où $\sigma^2(t,y) = \left(B(t,y)\left[g(t) - B(t,y)\right]/g^3(t)\right)\displaystyle\int K_1^2(u)\,du$ et $B(t,y) = \displaystyle\int_{-\infty}^{y} f(t,u)\,du$.

BERLINET *et al.* définissent ensuite un estimateur $\widehat{q}_{1/2}(t)$ de la médiane conditionnelle comme la racine de l'équation $\widehat{F}_t(y) = 1/2$.

Sous certaines conditions de régularité, $\sqrt{nh_n}\left(\widehat{q}_{1/2}(t) - q_{1/2}(t)\right) \underset{L}{\to} N\left(0, \sigma^2_{q_{1/2}}\right)$ où $\sigma^2_{q_{1/2}} = \left(4g(t) f_t^2\left(q_{1/2}(t)\right)\right)^{-1} \int K_1^2(u)\, du$ et où $f_t(y)$ désigne la densité de Y sachant $T = t$.

III.1.3 *Le médianogramme*

Le médianogramme connu également sous le nom d'estimateur à fenêtre mobile s'inscrit dans cette famille d'estimateurs. Le principe du médianogramme propose, dans un modèle de type (AG), un estimateur de la médiane conditionnelle mais peut être élargi à l'estimation des quantiles.

Pour une fenêtre fixée h et un réel fixé t, on note $Q(t,h)$ la variable aléatoire représentant le nombre de points d'observations T_i appartenant à l'unique intervalle $I_{m_0} = [m_0 h, (m_0+1)h[\,, m_0 \in Z$ contenant t. $Y_{i1}, \ldots, Y_{iQ(t,h)}$ sont les Y_i tels que $T_i \in I_{m_0}$. L'estimateur $\widehat{q}_{1/2,h}(t)$ de la médiane conditionnelle $q_{1/2}(t)$ est alors défini comme la médiane empirique du sous-échantillon ordonné $Y_{(i1)} \leqslant \cdots \leqslant Y_{(iQ(t,h))}$.

Sous les hypothèses exposées dans son article, Gannoun (1991) obtient un résultat de convergence presque complète, à savoir :

$\underset{t\in C}{\text{Sup}}\left(\widehat{q}_{1/2,h}(t) - q_{1/2}(t)\right) \underset{p.co.}{\to} 0$ où C est un compact de R soumis à certaines contraintes.

III.2. Estimateurs de la fonction de répartition au sens des plus proches voisins : les estimateurs de Bhatthacharya et Gangopadhyay

Ces auteurs se placent dans un modèle de type (AG). Pour t fixé, posons $U_i = |T_i - t|$. Soit $U_{[n1]} < \cdots < U_{[nn]}$ la statistique ordonnée à partir des U_i et $Y_{[n1]}, \ldots, Y_{[nn]}$ la statistique «ordonnée» induite. Pour tout entier positif $k \leqslant n$, un estimateur au sens des «k plus proches voisins» de la fonction de répartition de Y sachant $T = t$ est : $\widehat{F}_{t,k}(y) = \frac{1}{k}\sum_{i=1}^{k} 1\left(Y_{[ni]} \leqslant y\right)$. C'est tout simplement la fonction de répartition empirique des observations aux k points t_i les plus proches de t. L'estimateur des k plus proches voisins de $q_\alpha(t)$ peut être alors défini comme le quantile d'ordre α de $\widehat{F}_{t,k}$, à savoir :

$$\widehat{q}_{\alpha,k}(t) = \inf\left\{y \,/ \widehat{F}_{t,k}(y) \geqslant \frac{[k\alpha]}{k}\right\}.$$

Une alternative à l'estimation ci-dessus est de fixer une fenêtre h et de ne considérer que les observations aux points t_i de l'intervalle $[t - h/2, t + h/2]$. Si $K_n(h) = \sum_{i=1}^{n} 1\left(U_i \leqslant h/2\right)$, on définit alors :

$$\widetilde{q}_{\alpha,h}(t) := \widehat{q}_{\alpha,K_n(h)}(t) = \inf\left\{y / \widehat{F}_{t,k}(y) \geqslant \frac{[K_n(h)\,\alpha]}{[K_n(h)]}\right\}.$$

Ces auteurs obtiennent les résultats de normalité asymptotique suivants :

- Résultat 1 :
  $n^{2/5}\left[\widehat{q}_{\alpha,[n^{4/5}s]}(t) - q_\alpha(t)\right] \underset{L}{\to} N\left(\beta s^2, \sigma^2 s^{-1}\right)$ pour tout $s \in [a,b]$ avec $0 < a < b$ et β et σ^2 constantes calculées dans l'article de BHATTHACHARYA et GANGOPADHYAY (1990).
- Résultat 2 :
  $n^{2/5}\left[\widetilde{q}_{\alpha,n^{-1/5}s}(t) - q_\alpha(t)\right] \underset{L}{\to} N\left(\gamma t^2, \tau^2 t^{-1}\right)$ pour tout $t \in [a,b]$ avec $0 < a < b$ et γ et τ^2 constantes calculées dans l'article de BHATTHACHARYA et GANGOPADHYAY (1990).

TRUONG (1989) énonce un résultat de convergence ponctuelle en probabilités pour un estimateur du même type que $\widetilde{q}$ en montrant l'optimalité de la vitesse de convergence.

Notons que les estimateurs des paragraphes 3.1.1, 3.1.3 et 3.2 sont discontinus par rapport à la variable α alors que les estimateurs du paragraphe 3.1.2 sont réguliers par rapport à cette variable.

IV. ESTIMATION DIRECTE

IV.1. Estimation par des fonctions splines de régression

L'estimation suivante a été développée par HE et SHI (1994) dans un modèle de type (AG). Il s'agit d'approximer la fonction quantile par une fonction spline. Il faut adapter le principe des moindres carrés utilisé pour l'estimation de la moyenne conditionnelle au cas des quantiles en remplaçant la fonction carrée par la fonction ρ_α.

Un espace de splines polynômiales est défini par un ordre et la donnée de nœuds. Soit $k_n - 1 \in N^*$ le nombre de nœuds notés $0 = x_0 < x_1 < \cdots < x_{k_n} = 1$ et on supposera qu'il existe α_0 tel que $\max_i (x_i - x_{i-1}) / \min_i (x_i - x_{i-1}) \leqslant \alpha_0$ uniformément en n. Soit $m+1 \in N$ l'ordre des splines. On rappelle que la dimension de l'espace des splines polynômiales d'ordre $m+1$ ayant $k_n - 1$ noeuds est $p_n = k_n + m$. On appelle $s_1, \ldots, s_{p_n}$ la partition «étendue» sur $[0,1]$ définie par :

$$0 = s_1 = \cdots = s_{m+1} \qquad s_{m+2} = x_1, \ldots, s_{m+k_n} = x_{k_n - 1}$$
$$s_{m+k_n+1} = \cdots = s_{2m+k_n+1} = 1.$$

On définit alors p_n polynômes $B_1, \ldots, B_{p_n}$ appelés B-splines normalisées associées à la partition étendue. La formule définissant les B_i a été explicitée par SCHUMAKER (1981). Si l'on pose $\pi(t) = (B_1(t), \ldots, B_{p_n}(t))'$, l'estimateur de la fonction quantile est défini par :

$$\widehat{q}_\alpha(t) = \pi(t)' \widehat{\theta}$$

où

$$\widehat{\theta} = \underset{\theta \in R^{p_n}}{\operatorname{Arg\,min}} \sum_{i=1}^{n} \rho_\alpha\left(Y_i - \pi(T_i)' \theta\right).$$

Le calcul de $\widehat{\theta}$ est facililité par le fait que p_n est très inférieur à n.

En ce qui concerne la convergence de l'estimateur, les auteurs obtiennent un résultat de type convergence en moyenne quadratique. Plus précisément :

$\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{q}_\alpha^{(j)}(T_i) - q_\alpha^{(j)}(T_i)\right)^2 = O_p\left(n^{-2(r-j)/2r+1}\right)$ pour $j = 0, \ldots, m$ où $q_\alpha^{(j)}$ représente la dérivée d'ordre j de q_α.

IV.2. Estimation par des fonctions splines de lissage

L'estimation suivante a été développée par KOENKER, NG et PORTNOY (1994) dans un modèle de type (FG). On notera $0 = t_0 < t_1 < \cdots < t_n < t_{n+1} = 1$ les abscisses des points d'observation. En adaptant de façon similaire l'estimateur de type spline de lissage de la moyenne conditionnelle, on peut d'abord penser à minimiser la quantité : $\sum_{i=1}^{n}\rho_\alpha\left(y_i - g(t_i)\right) + \lambda\int_0^1 g''^2(x)\,dx$.

Cependant, la coexistence entre la forme linéaire par morceaux du premier terme et la forme quadratique du terme de pénalité pose d'importantes difficultés de calcul de la solution. Il apparaît donc opportun de remplacer $g''^2(x)$ par $|g''(x)|$. Koenker *et al.* proposent en conséquence de minimiser la quantité

$$R_{\alpha,\lambda}(g) := \sum_{i=1}^{n}\rho_\alpha\left(y_i - g(t_i)\right) + \lambda\int_0^1 |g''(t)|\,dt \tag{7}$$

où $\lambda > 0$, $\alpha \in [0,1]$. Remarquons que dans le cas $\lambda = 0$, la solution du problème est l'interpolant des points (t_i, y_i) qui minimise la quantité $\int_0^1 |g''(t)|\,dt$. Dans le cas $\lambda = \infty$, la solution est le quantile de regression linéaire, c'est-à-dire $\widehat{q}_\alpha(t) = \widehat{\alpha} + \widehat{\beta}t$ où $\left(\widehat{\alpha}, \widehat{\beta}\right)$ est la solution de $\underset{(a,b)\in R^2}{\text{Min}}\sum_{i=1}^{n}\rho_\alpha(y_i - a - bt_i)$. On peut montrer que le problème de minimisation (7) n'a pas de solution dans l'espace de Sobolev W_1^2 des fonctions définies sur $[0,1]$ ayant une dérivée première absolument continue de carré intégrable et une dérivée seconde existant presque partout. L'espace adapté à la résolution de ce problème d'optimisation est donné par l'extension suivante de l'espace de Sobolev : soit V l'espace des distributions f au sens de Schwarz dont la dérivée f' est une mesure de variation totale finie, muni de la semi-norme $\|f\| = V(f')$ où $V(g)$ désigne la variation totale de la mesure g . Notons que dans le cas où la mesure f' est absolument continue par rapport à la mesure de Lebesgue alors f'' existe et $V(f') = \int |f''|$.

Le problème de minimisation (7) s'écrit donc :

$$\underset{g\in V}{\text{Min}}\sum_{i=1}^{n}\rho_\alpha\left(y_i - g(t_i)\right) + \lambda V(g').$$

Les auteurs montrent alors que la solution $\widehat{g}$ de ce problème est une spline linéaire ayant pour noeuds les points t_i. Pour le calcul pratique de la solution $\widehat{g}$, ce problème de minimisation s'écrit comme un problème de programmation linéaire. En effet, si l'on pose $\widehat{g}(t) = \alpha_i + \beta_i (t - t_i)$ pour $t \in [t_i, t_{i+1}[$ et $i = 1, \ldots, n$, la continuité de $\widehat{g}$ donne la relation suivante entre les coefficients : $\beta_i = (\alpha_{i+1} - \alpha_i)/h_i$ où $h_i = t_{i+1} - t_i$. En définissant le vecteur $\alpha = (\alpha_1, \ldots, \alpha_n)' = (\widehat{g}(t_1), \ldots, \widehat{g}(t_n))'$, le problème de minimisation devient :

$$\underset{\alpha \in R^n}{\text{Min}} \sum_{i=1}^{n} \rho_\alpha (y_i - \alpha_i) + \lambda \sum_{j=1}^{n} \left| d_j' \alpha \right|$$

où d_j est un vecteur à n coordonnées dépendant des h_i.

IV.3. Estimation par des polynômes locaux

De même que pour les splines de régression et les splines de lissage, on peut adapter la méthode des polynômes locaux classique pour la moyenne conditionnelle aux quantiles conditionnels en remplaçant la fonction carrée par la fonction ρ_α dans le problème de minimisation qui définit le polynôme local. En toute généralité, pour un degré k, un noyau K et une fenêtre h, cela conduit simplement à résoudre en chaque point t le problème d'optimisation suivant :

$$\underset{(b_0, \ldots, b_k)}{\text{Min}} \sum_{i=1}^{n} \rho_\alpha \left(Y_i - \sum_{j=0}^{k} b_j (T_i - t)^j \right) K \left(\frac{t - T_i}{h} \right)$$

et à définir l'estimateur quantile par :

$$\widehat{q}_\alpha (t) = \widehat{b}_0.$$

De nombreux auteurs ont étudié ces estimateurs dans les modèles de type (AG). Tsybakov introduit cet estimateur dans le cas local linéaire ($k = 1$) en 1986. Lejeune et Sarda (1988) proposent un algorithme convergent pour évaluer cet estimateur dans le cas où $k = 2$. Chauduri (1991) démontre une propriété de convergence presque sûre ponctuelle dans le cas d'un noyau de type fonction indicatrice d'intervalles et pour un degré k quelconque. Fan *et al.*. (1994) établissent la normalité asymptotique de cet estimateur dans le cas où $k = 1$. Enfin, Yu et Jones (1998) implémentent l'estimateur localement linéaire et le comparent à une méthode d'estimation par inversion basée sur un estimateur à double noyau de la fonction de répartition conditionnelle.

V. CONCLUSION

Il existe donc de nombreuses techniques d'estimation des quantiles conditionnels dans le cadre non paramétrique. On peut encore généraliser le problème en considérant des données (Y_i, T_i) non indépendantes. Par exemple, ce cas est traité dans l'article de Boente et Fraiman (1995), de Berlinet, Gannoun et Matzner-Lober (1998) et dans celui de Gannoun (1991) où

l'on considère un processus $\{(Y_i, T_i)_{1 \leqslant i \leqslant n}\}$ stationnaire et α-mélangeant. Le médianogramme peut, dans ce cas, estimer la médiane conditionnelle et on a un résultat de convergence presque complète de l'estimateur vers la médiane (cf GANNOUN (1991)). Dans ce même cadre, on trouve aussi un résultat de normalité asymptotique pour l'estimateur de la médiane conditionnelle de BERLINET, GANNOUN et MATZNER-LOBER (1998). L'estimation de la médiane conditionnelle sous hypothèses ergodiques est abordée dans MARTINS ROSA et DELECROIX (1992). On peut également envisager d'ajouter dans certains cas des contraintes aux quantiles conditionnels qui nécessitent des techniques adaptées. Par exemple, POIRAUD-CASANOVA et THOMAS-AGNAN (1998) et MUKERJEE (1993) considèrent l'estimation des quantiles conditionnels sous contrainte de monotonie car, dans certains cas réels, on peut penser que la fonction quantile est une fonction monotone de la covariable (voir par exemple les données de croissance présentées dans LEJEUNE et SARDA (1988)). Enfin, BERLINET, CADRE et GANNOUN (1998) envisagent l'estimation de la médiane conditionnelle dans le cas où les variables réponses et les prédicteurs sont multidimensionnels.

Remerciements

Nous remercions les rapporteurs qui, par des remarques constructives, ont permis d'enrichir le texte ainsi que la liste des références.

BIBLIOGRAPHIE

ANTOCH J., JANSSEN P. (1989). Non parametric regression M-quantiles. Statistics & Probability Letters. 8 pp. 355-362.

BERLINET A., CADRE B., GANNOUN A. (1998). Estimation non paramétrique de la médiane conditionnelle spatiale. Preprint.

BERLINET A., GANNOU A., MATZNER-LOBER E. (1997). Asymptotic normality of the non parametric estimator of conditional median under mixing conditions. Preprint.

BERLINET A., GANNOUN A., MATZNER-LOBER E. (1998). Propriétés asymptotiques d'estimateurs convergents des quantiles conditionnels. C. R. Arcad. Sci. Paris, t. Série 1, pp. 611-614.

BHATTACHARYA P.K., GANGOPADHYAY A.K. (1990). Kernel and nearest-neighbor estimation of a conditional quantile. Ann. Math. Statist. 18 (3) pp. 1400-1415.

BOENTE G., FRAIMAN R. (1995). Asymptotic distribution of smoothers based on local means and local medians under dependance. Journal of Multivariate Analysis. 54 pp. 77-90.

CHAUDURI P. (1991). Nonparametric estimates of regression quantiles and their local Bahadur representation. Ann. Math. Statist. 19 (2) pp. 760-777.

CHENG C., PARZEN E. (1997). Unified estimators of smooth quantile and quantile density functions. Journal of Statistical and Planning Inference. 59 pp. 291-307.

COLE T.J., GREEN P.J. (1992). Smoothing reference centile curves : the LMS method and penalized likelihood. Statistics in Medecine. 11 pp. 1305-1319.

CSORGO M., HORVATH L. (1993). Weighted approximation in probability and statistics. Wiley, New-York.

CSORGO M., REVEZC P. (1984). Two approaches to constructing simultaneous confidence bounds for quantiles. Prob. and Math. Statist. 4 pp. 221-236.

DAVIS C.E., HARRELL F.E. (1982). A new distribution-free quantile estimator. Biometrika. 69 (3) pp. 635-640.

DUCHARME G.R., GANNOUN A., GUERTIN M.C., JEQUIER J.C. (1995). Reference values obtained by kernel-based estimation of quantile regression. Biometrics. 51 pp. 1105- 1116.

FALK M. (1984). Relative deficiency of Kernel type estimators of quantiles. Ann. Math. Statist. 12 (1) pp. 261-268.

FAN J., HU T.C., TRUONG Y.K. (1994). Robust nonparametric function estimation. Scand. J. Statist. 21 pp. 433-446.

GANNOUN A. (1991) . Prédiction non paramétrique : médianogramme et méthode du noyau en estimation de la médiane conditionnelle. Statistique et Analyse des données. 16 (1) pp. 23-42.

GOLDSTEIN H., PAN H. (1992). Percentile smoothing using piecewise polynomials with covariates. Biometrics. 48 pp. 1057- 1068.

HART J.D. (1991) . Comment to "Choosing a kernel regression estimator". Statistical Sciences. 6 pp. 425-427.

HE X., SHI— P. (1994). Convergence rate of B-spline estimators of non parametric conditional quantile functions. Nonparametric Statistics. 3 pp. 299-308.

HOGG R.V. (1975). Estimates of pourcentile regression lines using salary data. Journal of the American Statistical Association. 70 pp. 56-59.

KAIGH W.D., CHENG C. (1991). Subsampling quantile estimators and uniformity criteria. Comm. Statist A. 20 pp. 539-560.

KAIGH W.D., LACHENBRUCH P.A. (1982). A generalized quantile estimator. Comm. Statist A. 11 pp. 2217-2238.

KOENKER R., BASSETT G. (1978). Regression quantiles. Econometrica. 46 (1).

KOENKER R., BASSETT G. (1982). An empirical quantile function for linear models with i.i.d. errors. Journal of the American Statistical Association. 77 pp. 407-415.

KOENKER R., NG P., PORTNOY S. (1994). Quantile smoothing splines. Biometrika.81 (4) pp. 673-680.

LEJEUNE M.G., SARDA P. (1988). Quantile regression : a nonparametric approach. Computational Statistics & Data Analysis. 6 pp. 229-239.

MARTINS ROSA A.C., DELECROIX M. (1992). Ergodic processes prediction via estimation of the conditional distribution function. Pub I.S.U.P. vol XXXIX fasc 2, 95, pp. 35-56.

MUKERJEE H. (1993). An improved monotone conditional quantile estimator. Ann. Math. Statist. 21 (2) pp. 924-942.

PADGETT W.J., LIO Y.L. (1993). A smooth nonparametric quantile estimator for IFR distributions. Nonparametric Statistics. 2 pp. 195-202.

PARZEN E.. (1979). Nonparametric statistical data modeling (with comments). Journal of the American Statistical Association. 74 pp. 105-131.

POIRAUD-CASANOVA S., THOMAS-AGNAN C. (1998). Monotone nonparametric regression quantiles. Cahier technique n° 97.09.452. GREMAQ, 21 Allées de Brienne 31000 Toulouse, France.

SAMANTA M. (1989). Nonparametric estimation of conditional quantiles. Statistic and probability letters. 7 (5) pp. 407-412.

SCHUMAKER L.L. (1981). Spline functions. Wiley.

SHEATHER S.J., MARRON J.S. (1990). Kernel quantile estimators. Journal of the American Statistical Association. 85 pp. 410-416.

SONESSON S.E., FOURON J.C., DOBLIK S.P., TAWILE C., LESSARD M., SKOLL A., GUERTIN M.C., DUCHARME G.R. (1993). Reference values for Doppler velocimetric indices from the fetal and placental ends of the umbilical artery during normal pregnancy. Journal of Clinical Ultrasound. 21 pp. 317-324.

STONE C.J. (1977). Consistent nonparametric regression. Ann. Math. Statist. 5(4) pp.595-645.

STUTE W. (1986). Conditional empirical processes. Ann. Math. Statist. 14(2) pp. 638-647.

TRUONG Y.K. (1989). Asymptotic properties of kernel estimators based on local medians. Ann. Math. Statist. 17(2) pp. 606-617.

TSYBAKOV A.B. (1986). Robust reconstruction of functions by the local-approximation method. Problems of Information Transmission. 22 pp. 133-146.

YAMATO H. (1973). Uniform convergence of an estimation of a distribution function. Bull. Math. Statist. 15 pp. 69-70.

YANG S.S. (1985). A smooth nonparametric estimator of a quantile function. Journal of the American Statistical Association. 80 (392), Theory and Methods, pp. 1004-1011.

YU K., JONES M.C. (1998). Local linear quantile regression. Journal of the American Statistical Association. 93 (441) pp. 228-237.

ZELTERMAN D. (1990). Smooth nonparametric estimation of the quantile function. Journal of Statistical and Plannig Inference. 26 pp. 339-352.