საქართველოს მეცნიერებათა ეროვნული აკადემიის მოამბე, ტ. 11, №3, 2017
BULLETIN OF THE GEORGIAN NATIONAL ACADEMY OF SCIENCES, vol. 11, no. 3, 2017

Mathematics

On Nonparametric Quantile Function Estimation Using Transformed Moments

Aleksandre Sborshchikovi

Faculty of Exact and Natural Sciences, I. Javakhishvili Tbilisi State University, Tbilisi, Georgia

(Presented by Academy Member Elizbar Nadaraya)

ABSTRACT. Recovery of a quantile function by means of the moments of underlying distribution is a challenging problem due to instability of approximants, when taking into account higher integer order moments of distribution. In the framework of probabilistic moment problem it is suggested to use the sequence of transformed moments. Two cases where the support of underlying distribution is bounded and unbounded from the above are considered. The uniform upper bound of the proposed approximate is established. Also the modified approximation of the quantile function is constructed. It is shown that proposed modification considerably improves the uniform approximation rate. Finally, based on the suggested approximation, new nonparametric estimate of the quantile function is constructed and corresponding error in sup-norm is investigated. The consistency in probability of corresponding estimate is derived as well. Two examples are considered, when distribution support is finite and is unbounded from the above. Two Tables with the average errors in sup-norm are recorded, and the consistency of proposed estimates are justified via simulations as well. *© 2017 Bull. Georg. Natl. Acad. Sci.*

Key words: quantile function, transformed moments, moment-based approximation

Suppose that a distribution function (*df*) F is absolutely continuous with respect to Lebesgue measure on $(0,\infty)$. In this note we only consider the case when the support of F is unbounded from the above, e.g. $\operatorname{supp}\{F\}=(0,\infty)$. There are several different approaches developed for estimating a quantile function $Q(x)=\inf\{t: F(t)\geq x\}$ of F nonparametrically. Let us mention [1-5] among others. Recall also the approximation (see [6]) that recovers Q based on the knowledge of the sequence of frequency moments $m^{-}(S)=\left\{m^{-}(j,S),\, j=1,\dots,\sim\right\}$

Here

$$m^{-}(j,S)=\int_0^\infty [S(t)]^j\,dt,\; j=1,\dots,\sim\ , \tag{1}$$

with $S=1-F$ to be the survival function of F. Namely, the following approximate was proposed:

$$Q_{\sim,S}(x) := (\bar{K}_{\sim}^{-1} m_S^{-})(x) = \sum_{k=\sim-\lfloor \sim x\rfloor}^{\sim} \sum_{j=k}^{\sim} \binom{\sim}{j}\binom{j}{k}(-1)^{j-k} m^{-}(j,S),\ x\in(0,1)\,. \tag{2}$$

In (2) and in the sequel, we will use the symbol $\lfloor \sim \rfloor$ to denote the integer part of $\sim$, while by $\lceil \sim \rceil$ we denote the rounding part of $\sim$.

To estimate the quantile function Q, one can consider the empirical counterpart $m_{\hat{s}_n}^{-} = \{m^{-}(j,\hat{s}_n),\ j=1,\ldots,\sim\}$ of m_S^{-} defined in (1) and replace the later by $m_{\hat{s}_n}^{-}$ in (2). Here $\hat{s}_n = 1-\hat{F}_n$ and $\hat{F}_n$ is the empirical *df* corresponding to the sequence of independent identically distributed (i.i.d) random variables $X_1,\ldots., X_n$ drawn from *F*. The following estimate was derived:

$$Q_{\sim,\hat{s}_n}(x) := (\bar{K}_{\sim}^{-1} m_{\hat{s}_n}^{-})(x),\ x\in(0,1)\,, \tag{3}$$

and several asymptotic properties of $Q_{\sim,\hat{s}_n}(x)$ have been investigated.

The main aim of this note is to study asymptotic behavior of another approximate and estimate of *Q* that is based on the following sequence of transformed moments $m_F^{+} = \{m^{+}(j,F),\ j=0,\ldots,\sim\}$, where

$$m^{+}(j,F) = \int_0^{\infty} x[F(x)]^j\, dF(x)\,. \tag{4}$$

In the sequel we assume that the first two moments of *X* are finite, and $\sim\in N_+$. The following notations are useful as well: by $(t,c,d), t\in(0,1)$, we denote the density function of a $Beta(\cdot,c,d)$ distribution with the shape parameters $c_x = \lceil \sim x\rceil + 1$, and $d_x = \sim - \lceil \sim x\rceil + 1$ while $\| f \|_{¢}$ denotes the sup-norm of *f* on $[\delta, 1-\delta]$, for some $0 \le ¢ \le \frac{1}{2}$. In addition let us recall the inequality (see[7]):

$$(t, \lfloor \sim x\rfloor + 1, \sim - \lfloor \sim x\rfloor + 1) \le \frac{C_1\sqrt{\sim}}{\sqrt{x(1-x)}},\ 0<x<1, \tag{5}$$

valid for some constant $C_1 > 0$.

Results

New approximation of *Q* is defined as follows:

$$Q_{\sim}^{+}(x) := (B_{\sim}^{-1} m_F^{+})(x),\ x\in(0,1)\,, \tag{6}$$

where

$$(B_{\sim}^{-1} m_F^{+})(x) := \frac{\Gamma(\sim+2)}{\Gamma(\lceil \sim x\rceil + 1)} \sum_{j=0}^{\sim-\lceil \sim x\rceil} \frac{(-1)^j m^{+}(j+\lceil \sim x\rceil, F)}{j!(\sim - \lceil \sim x\rceil - j)!}\,.$$

Note that substitution of m_F^{+} defined by (4) in the right-hand side of (6) gives

$$Q_{\sim}^{+}(x) = \int_0^{\infty} (F(u), c_x, d_x) u\, dF(u) = \int_0^1 (t, c_x, d_x) Q(t)\, dt\,. \tag{7}$$

Hence, applying similar argument used in the proof of Theorem 1 and Corollary 1 from [8] we easily obtain the following

Theorem 1. *Let* $\tilde{Q}^{+}_{\sim} = 2Q^{+}_{\tilde{\sim}} - Q^{+}_{\sim}$ *where* $\tilde{\sim} = 2\sim$. *If* $Q^{'}$ *is bounded and* $Q^{''}$ *is bounded and continuous on* $[\text{¢}, 1-\text{¢}]$ *then for each* $x \in (0,1)$ *and* $\sim \to \infty$ *we have*

(i) $Q^{+}_{\sim} - Q(x) = \frac{1}{\sim + 2}\{(1-2x + \lceil \sim x \rceil - \sim x)Q^{'}(x) + \frac{1}{2}x(1-x)Q^{''}(x)\} + o(\frac{1}{\sim})$

(ii) $\|Q^{+}_{\sim} - Q\|_{\text{¢}} \le \frac{1}{\sim + 2}\{\frac{3}{2}\|Q^{'}\|_{\text{¢}} + \frac{1}{8}\|Q^{''}\|_{\text{¢}}\} + o(\frac{1}{\sim})$

(iii) $\tilde{Q}^{+}_{\sim}(x) - Q(x) = \frac{1}{(\sim+1)(\sim+2)}[(1-2x+(\lceil \sim x \rceil - \sim x))(\sim + 3)Q^{'}(x) + \frac{1}{2}x(1-x)Q^{''}(x)] + O(\frac{1}{\sim^{2}})$

Using the empirical df instead of F in (6) yields the estimate of Q

$$\hat{Q}^{+}_{\sim,n} := B^{-1}_{\sim} m^{+}_{\hat{F}_n} , \tag{8}$$

It is worth mentioning that given n independent copies of X, $X_1, ..., X_n$ one can rewrite the components of the sequence $m^{+}_{\hat{F}_n}$ as follows:

$$m^{+}(j, \hat{F}_n) = \int_0^{\infty} t[\hat{F}_n(t)]^j \, d\hat{F}_n(t) = \frac{1}{n}\sum_{i=1}^{n} X_{(i)}(\frac{i}{n})^j . \tag{9}$$

Here $X_{(i)}$ is the i-th order statistic of the sample $X_1, ..., X_n$

Theorem 2. *If Q is continuous on* $(0,1)$ *then* $\hat{Q}^{+}_{\sim,n} \xrightarrow{P} Q$, *uniformly on* $[\text{¢}, 1-\text{¢}]$ *as* $\sim = o(n), n \to \infty$, *for some* $0 \le \text{¢} \le \frac{1}{2}$.

Proof. Let us denote by

$$Q^{*}_{\sim,n}(x) = \int_0^{\infty} \quad (F(u), c_x, d_x) u d\hat{F}_n(u) , \tag{10}$$

and rewrite the difference between $Q^{+}_{\sim}(x)$ and $Q(x)$ as sum of the following expressions:

$$\hat{Q}^{+}_{\sim,n}(x) - Q^{*}_{\sim,n}(x) = \int_0^{\infty} [\quad (\hat{F}_n(u), c_x, d_x) - \quad (F(u), c_x, d_x) u d\hat{F}_n(u), \tag{11}$$

$$Q^{*}_{\sim,n}(x) - Q^{+}_{\sim}(x) = \int_0^{\infty} \quad (F(u), c_x, d_x) u d\left[\hat{F}_n(u) - F(u)\right] = \frac{1}{\sqrt{n}}\int_0^{\infty} \quad (F(u), c_x, d_x) u d v_{n,F}(u), \tag{12}$$

$$Q^{+}_{\sim}(x) - Q(x) = \int_0^{\infty} \quad (F(u), c_x, d_x) u dF(u) - Q(x) = \int_0^{1} \quad (t, c_x, d_x) Q(t) dt - Q(x) . \tag{13}$$

Where $v_{n,F}(t) = \sqrt{n}(\hat{F}_n(t) - F(t))$ represents the empirical process corresponding to the empirical *df* $\hat{F}_n$.

Note that since the function Q is continuous and the sequence of densities $\{ \quad (\cdot, \lceil \sim x \rceil + 1, \sim - \lceil \sim x \rceil + 1), \sim = 1, ...\}$ forms a ¢ –sequence at x. We conclude from (13) that $Q^{+}_{\sim}$ converges uniformly to Q as $\sim \to \infty$ (see [9], Ch. VII). From inequality (5) and the properties of $v_{n,F}$ we easily derive the upper bound for the variance of $Q^{*}_{\sim,n}(x) - Q^{+}_{\sim}(x)$ written in the form (12):

$$\frac{1}{n}\int_0^\infty {}^2(F(u),c_x,d_x)u^2dF(u) \le \frac{\sim C_1^2 E(X^2)}{nx(1-x)}. \qquad (14)$$

Applying the Lagrange formula for difference between two beta density functions under the integral in (11), one can write

$$\left|\frac{\sim+1}{\sqrt{n}}\int_0^\infty (\tilde{F}(u),c_x-1,d_x-1)\frac{\lceil \sim x\rceil - \sim \tilde{F}(u)}{\lceil \sim x\rceil(\sim - \lceil \sim x\rceil)}v_{n,F}(u)ud\hat{F}(u)\right| \le$$

$$\sup\left|v_{n,F}(u)\right|\frac{\sim+1}{\sim\sqrt{n}}\frac{C_1\sqrt{\sim}}{\sqrt{x(1-x)}}\frac{2\bar{X}}{(\frac{\lceil \sim x\rceil}{\sim})(1-\frac{\lceil \sim x\rceil}{\sim})} \le \frac{4\bar{X}C_1}{\text{¢}^3}\sqrt{\frac{\sim}{n}} \quad , \qquad (15)$$

for some $\tilde{F}(u)$ with $F(u) \le \tilde{F}(u) \le \hat{F}_n(u)$ Note that $\sup\left|v_{n,F}(u)\right| = O_P$ as $n \to \infty$

Hence, we conclude that the bounds in (14)-(15) have the order $\frac{\sim}{n}$ and $\sqrt{\frac{\sim}{n}}$ respectively, uniformly on $[\text{¢}, 1-\text{¢}]$ for some small positive ¢ .

Examples:

To conduct simulation study we consider two examples. Let us define the average errors in terms of the sup-norm for the estimate $\hat{Q}^{+}_{\sim,n}$ as follows:

$$d_{\sim,n} = \frac{1}{N}\sum_{r=1}^{N}\max_{1\le j\le\lceil \sim(1-\text{¢})\rceil}\left|\hat{Q}^{(r)}_{\sim,n}(\frac{j}{\sim}) - Q(\frac{j}{\sim})\right|.$$

Here $\hat{Q}^{(r)}_{\sim,n}$ denotes the value of $\hat{Q}^{+}_{\sim,n}$ evaluated on the *r*-th replication, and *n* is the number of replications.

In Example 1 we assume $X_i \sim Beta(\frac{1}{2};1)$ and in Example 2, let us consider the model with $X_i \sim E\text{xp}(1)$. Tables 1 and 2 display the records of average errors $d_{\sim,n}$ of $\hat{Q}^{+}_{\sim,n}$ defined in (8) when i.i.d. random variables X_i 's are drawn from $Beta(\frac{1}{2};1)$ and $E\text{xp}(1)$ distributions, respectively.

The errors recorded in Table 1 and Table 2 justify that the proposed estimate $\hat{Q}^{+}_{\sim,n}$ is consistent. In addition, one can see that the error $d_{\sim,n}$ decreases as both the parameter $\sim$ and the sample size *n* are increasing while the ratio $\frac{\sim}{n}$ is decreasing.

Table 1: The average errors in sup-norm $d_{\sim,n}$ **of** $\hat{Q}^{+}_{\sim,n}$ **are recorded. Here** X_i **'s are simulated from** $Beta(\frac{1}{2};1)$ **and** $¢ = 0$.

			$d_{\sim,n}$			
$n/\sim$	$\sim=20$	$\sim=40$	$\sim=60$	$\sim=80$	$\sim=100$	$\sim=120$
n=100	0.0962547	0.0916660	0.0874313	0.0905589	0.119838	0.145500
n=300	0.0769729	0.0552744	0.0496481	0.0506326	0.0521042	0.0485487
n=600	0.0739440	0.0477722	0.0415074	0.0328432	0.0357198	0.0345494
n=1000	0.0714150	0.0458237	0.0342856	0.0301669	0.0298474	0.0281450

Table 2: The average errors in sup-norm $d_{\sim,n}$ **of** $\hat{Q}^{+}_{\sim,n}$ **are recorded. Here** X_i **'s are simulated from** $Exp(1)$ **and** $¢ = \frac{1}{5}$

			$d_{\sim,n}$			
$n/\sim$	$\sim=20$	$\sim=40$	$\sim=60$	$\sim=80$	$\sim=100$	$\sim=120$
n=100	0.162267	0.174211	0.186656	0.188134	0.181125	0.207305
n=300	0.100038	0.111704	0.0988033	0.109227	0.106078	0.114819
n=600	0.0930935	0.0740184	0.0787526	0.0688771	0.0755520	0.0794764
n=1000	0.0757144	0.0645206	0.0610075	0.0593421	0.0559969	0.0684779

მათემატიკა

არაპარამეტრული კვანტილის ფუნქციის შეფასება ტრანსფორმირებული მომენტებით

ა. სბორშჩიკოვი

ივანე ჯავახიშვილის სახელობის თბილისის სახელმწიფო უნივერსიტეტი, ზუსტ და საბუნებისმეტყველო მენიერებათა ფაკულტეტი

(წარმოდგენილია აკადემიის წევრის ე. ნადარაიას მიერ)

კვანტილის ფუნქციის აღდგენა იმ შემთხვევაში, როცა ცნობილია განაწილების მაღალი რიგის მომენტები, დღესდღეობით წარმოადგენს ძალზედ მნიშვნელოვან და საინტერესო პრობლემას. ალბათურ მომენტთა პრობლემის პირობებში მიღებულია ე.წ. ტრანსფორმირებული მომენტების გამოყენება. განხილულია ორი შემთხვევა, როცა განაწილების ფუნქციის განსაზღვრის არე სასრული და უსასრულოა. კვანტილის ფუნქციის მოდიფიცირებული აპროქსიმაცია იყო აგებული. ნაჩვენები იყო, რომ მოდიფიცირებული ვერსიის აპროქსიმაციის რიგი უფრო მაღალია ვიდრე თავდაპირველის.

ახალ შემოღებულ აპროქსიმაციისთვის იყო გაკეთებული არა პარამეტრული შეფასება კვანტილის ფუნქციის და ცდომილება სუპ-ნორმის პირობებში იყო ნაჩვენები.მოყვანილი იყო ამ შეფასების ძალდებულობა. ორი მაგალითი განაწილების ფუნქციის განსაზღვრის არის შესაბამისად იყო აგებული. ორი ცხრილი საშუალო ცდომილებით სუბ ნორმის პირობებში იყო დათვლილი და სიმულაციებით დამტკიცებული იყო მისი ძალდებულობა.

REFERENCES

1. Parzen E. (1979) Nonparametric statistical data modeling (with discussion). *Journal of the American Statistical Association*, 74 :105-131.
2. Harrell F.E., Davis C.E. (1982) A new distribution-free quantile estimator. *Biometrika,* 69: 635-640.
3. Jones M.C. (1992) Estimating densities, quantiles, quantile densities and density quantiles. *AISM*, 44 :721-727.
4. Cheng C., Parzen E. (1997) Unified estimators of smooth quantile and quantile density functions. *Journal of Statistical Planning and Inference*, 59: 291-307.
5. Yang S.S. (1985) A smooth nonparametric estimator of a quantile function. *Journal of the American Statistical Association,* 80 : 1004-1011.
6. Mnatsakanov R.M., Sborshchikovi A. (2017) Recovery of a quantile function from moments. *J. Comput. Appl. Math*. 315 : 354-364.
7. Chen S.X. (2000) Beta kernel for regression curves. *Statist. Sinica* 10: 73-92.
8. Mnatsakanov R.M., Garai B. (2017) On the moment-recovered approximations of regression and derivative functions with applications. *J. Comput. Appl. Math.* 315: 17-31.

9 Feller W. (1971) An Introduction to Probability Theory and Applications. *II, 2nd Edition, Wiley, New York.*

Received June, 2017