Journal of Sciences, Islamic Republic of Iran 33(1): 45 - 54 (2022) http://jsciences.ut.ac.ir
University of Tehran, ISSN 1016-1104

Estimating Cumulative Distribution Function Using Gamma Kernel

B. Mansouri*, S. A. Sayyid Al-Farttosi, H. Mombeini and R. Chinipardaz

Department of Statistics, Faculty of Mathematical Sciences and Computer, Shahid Chamran University of Ahvaz, Ahvaz, Islamic Republic of Iran

Received: 2 Jun 2021 / Revised: 23 Oct 2021 / Accepted: 21 Dec 2021

Abstract

In this article, we propose the gamma kernel estimator for the cumulative distribution functions with nonnegative support. We derive the asymptotic bias and variance of the proposed estimator in both boundary and interior regions and show that it is free of boundary bias. We also obtain the optimal smoothing parameter which minimizes the mean integrated square error (MISE). In addition to consistency, we prove the almost sure convergence of the proposed estimator and show that it follows the same approximate normal distribution as empirical distribution. We presented a simulation study to compare the performance of the proposed estimator with other estimators. We use the proposed estimator to estimate the cumulative probability distribution function of the food expenses for urban households in Iran.

Keywords: Asymmetric kernels; Cumulative distribution; Boundary problem; Almost sure convergence.

Introduction

For a given i.i.d. sample $X_1, \dots, X_n$ from an unknown continuous cumulative distribution function (CDF) $F(x)$, the empirical distribution function is defined as $F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x)$ where $I(\cdot)$ is the indicator function. Since $X_i's$ are i.i.d. from Strong Law of Large Numbers we can deduce that

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x) \to E[I(X \le x)] = F(x), \quad as\ n \to \infty, \quad w.p.\,1,$$

and since $\mathrm{Var}[I(X \le x)] = F(x)(1 - F(x))$ from Central Limit Theorem we have

$$F_n(x) \sim N\left(F(x), n^{-1}F(x)\big(1 - F(x)\big)\right), \quad as\ n \to \infty.$$

Although $F_n(x)$ is a consistent estimator for $F(x)$, the empirical distribution is not smooth. As an alternative, [1] and [2] introduced Kernel-type estimators for distribution estimation, based on symmetric kernels. The asymptotic properties of Kernel-type estimators have been investigated by [3]. Due to the asymptotical superiority of Kernel-type estimators over the empirical distribution function [4, 5], they are popular and commonly used in density and distribution estimation. However, they are not efficient for that distribution (density) functions which have bounded support due to the boundary bias. This problem is known as the boundary problem, and several approaches have so far been proposed to deal with it in regression, density estimation and cumulative distribution estimation tasks [6-13].

In an effort to solve the boundary problem in kernel density estimation, [14] introduced the beta kernel density estimator to estimate a density with support on [0, 1]. This research was the starting point for using

* Corresponding author: Tel: +989166043957; Email: b.mansouri@scu.ac.ir

asymmetric kernel functions in nonparametric density estimation. [15] developed his method by employing the gamma kernel to estimate a density with support on $[0, \infty)$. This approach has extended for estimating a density with support on $[0, \infty)$ using other asymmetric kernels [16-19]. Reference [20] showed uniform weak consistency of some asymmetric kernel density estimators including gamma, inverse Gaussian and reciprocal inverse Gaussian on each compact set in $[0, \infty)$ when the probability density function f is continuous on its support. They also showed weak convergence of these estimators in L_1. Despite there are many studies on using asymmetric kernels for density estimation, little study has been done on estimating the cumulative distribution function using asymmetric kernels. This may be because, unlike symmetric kernel estimators, the development of a density function estimator with an asymmetric kernel to an asymmetric kernel distribution function estimator is not straightforward. For example, consider the gamma kernel estimator as a probability density function [15]:

$$\widehat{f_G}(x) = n^{-1}\sum_{i=1}^{n} k_{\frac{x}{b}+1,b}(X_i),$$

where $k_{\frac{x}{b}+1,b}(t) = \frac{t^{x/b}\exp(-t/b)}{b^{x/b+1}\Gamma(x/b+1)}$, b is the smoothing parameter and x is the design point. Now, if we want, like the symmetric kernel, to estimate the distribution function by integrating $\widehat{f_G}(x)$, that is, $\tilde{F}(x) = \int_0^x \widehat{f_G}(u)\,du$ then it includes the integral $\int_0^x \frac{t^{u/b}\exp(-t/b)}{b^{u/b+1}\Gamma(u/b+1)}du$, which made studying $\tilde{F}(x)$ very hard if not possible. The same problem exists in all asymmetric kernels because in these kernels the design point x is embedded in at least one of the kernel function parameters. To overcome this difficulty, [21] have proposed a new Kernel-type estimator for the cumulative distribution function. To estimate a cumulative distribution function with non-negative support, they proposed to use estimators of the form

$$\hat{F}(x) = n^{-1}\sum_{i=1}^{n}\int_{X_i}^{\infty} k(u; x, b)du, \qquad (1)$$

where $k_{x,b}(u)$ is an asymetic kernel function on $[0, \infty)$ and x and b are the desing point and smoothing parameter, respectively. They have studied two asymmetric kernels, including the Birnbaum-Saunders kernel and the Weibull kernel and demonstrated that their proposed estimators are free from boundary bias. It is easy to verify that $\hat{F}(x)$ is a cummulative distribution function. For example, for the Birnbaum-Saunders kernel we have

$$\begin{aligned}\overline{K}_{B-S}(t; x, \sqrt{b}) &= \int_t^{\infty} k_{B-S}(u; x, \sqrt{b})du \\ &= 1 - \Phi\left(\left(\sqrt{\frac{t}{x}} - \sqrt{\frac{x}{t}}\right)\Big/\sqrt{b}\right), \\ &t > 0, b > 0, x > 0,\end{aligned}$$

where $\Phi(\cdot)$ is the standard normal distribution function. Then consider that

$$\begin{aligned}\overline{K}_{B-S}(t; x, \sqrt{b}) &= 1 - \Phi\left(\left(\sqrt{\frac{t}{x}} - \sqrt{\frac{x}{t}}\right)\Big/\sqrt{b}\right) \\ &= 1 - \Phi\left(-\left(\sqrt{\frac{x}{t}} - \sqrt{\frac{t}{x}}\right)\Big/\sqrt{b}\right) \\ &= \Phi\left(\left(\sqrt{\frac{x}{t}} - \sqrt{\frac{t}{x}}\right)\Big/\sqrt{b}\right);\end{aligned}$$

because $\Phi(-z) = 1 - \Phi(z)$. Therefore $\overline{K}_{B-S}(t; x, \sqrt{b})$ is the Birnbaum-Saunders cumulative distribution function with respect to x with parameters t and b. The novelty of the estimator (1) is that instead of integrating the nonparametric density function estimator, it directly estimates the cumulative distribution function.

In this paper, a gamma kernel estimator for estimating those distribution functions with support on $[0, \infty)$ is introduced. We derive the asymptotic bias and variance of the proposed estimator and show that it is an asymptotically consistent estimator. We investigate the convergence rate of the proposed estimator in both boundary and interior regions. We show that, unlike the gamma kernel density estimator, having a shoulder in the distribution is not a necessary condition for our proposed estimator to be unbiased. We also derive an optimal bandwidth for the proposed estimator.

We prove the almost surely convergence of the gamma kernel estimator to be the true distribution. Furthermore, we establish the asymptotic distribution of the proposed estimator. Moreover, we illustrate the performance of the proposed estimator in comparison with other commonly-used methods via a numerical study. In the numerical study, we consider various distributions including mixed distributions for which the estimation of cumulative distribution functions can be difficult. The results are promising and demonstrate the usefulness of the gamma kernel estimator. As an application, we estimate the probability distribution function of the cost of one month of food items for

urban households in Iran in 1398 AH (2019 AD).

The paper is organised as follows. Section 2 is devoted to the main theoretical results. In Section 3, a numerical study is conducted to illustrate the performance of the gamma kernel estimator. Finally, in Section 4, the cumulative distribution of a real data set is estimated via gamma kernel estimator.

Throughout the paper, it is assumed that the CDF $F(x)$ is satisfied in the following assumptions:

Assumption 1: The cumulative distribution function $F(x)$ is absolutely continuous with respect to Lebesgue measure on $(0, \infty)$ and has two first continuous and bounded derivatives.

Assumption 2: The smoothing parameter $b = b_n > 0$ satisfies the condition that $b \to 0 \; as \; n \to \infty$.

Asymptotic properties of the gamma kernel distribution estimator

In this section, it is shown that the gamma kernel estimator is asymptotically unbiased and consistent. An appropriate smoothing parameter is also obtained by minimizing the mean integrated square error. In addition, the convergence rate of the proposed estimator is examined. Almost sure convergence and asymptotic distribution of the proposed estimator is also discussed.

Suppose that $\mathrm{X}_1, \mathrm{X}_2, \ldots, \mathrm{X}_\mathrm{n}$ is a set of continuous random variables with an unknown cumulative distribution function $F(x)$. Then, the proposed estimator is

$$\hat{F}_G(x) = n^{-1} \sum_{i=1}^{n} \bar{K}_G\,(X_i; x, b), \tag{2}$$

where $\bar{K}_G(t; x, b) = \int_t^\infty k_G(u, x, b)du$, $k_G(t; x, b) = \frac{t^{x/b}\exp(-t/b)}{b^{x/b+1}\Gamma(x/b+1)}$, $t > 0, x > 0$, $b > 0$.

Theorem 1. Suppose that Assumptions 1-2 hold, then we have:

$$\mathrm{bias}(\hat{F}_G(x)) \approx b(f(x) + \frac{1}{2} x f'(x)), \tag{3}$$

$$\mathrm{var}\left(\hat{F}_G(x)\right) \approx \begin{cases} n^{-1}F(x)\big(1-F(x)\big) - n^{-1}\sqrt{xb/\pi}f(x) & if \quad x/ \\ n^{-1}F(x)\big(1-F(x)\big) + n^{-1}bf(x)(1-C_k(1+2k)) & if \end{cases} \tag{4}$$

where $C_k = \frac{\Gamma(2k+1)}{2^{1+2k}\Gamma^2(k+1)}$ as $b \to 0$ and $n^{-1}b \to \infty$.

Proof: Since $X_i's$ are identically distributed, we have

$$E_f\left(\hat{F}_G(x)\right) = E_f\big(\bar{K}_G(T; x, b)\big) = E_f\big(1 - K_G(T; x, b)\big) = E_k\big(F(T)\big),$$

where $E_k\big(F(T)\big)$ is the expectation of $F(T)$, when $T \sim k_G(t; x, b)$. Using Taylor expansion, we have:

$$E_k\big(F(T)\big) = F(x) + f(x)E(T-x) + \frac{1}{2}f'(x)E(T-x)^2 + O(b),$$

so, Equation (3) can be easily verified.

For the variance term, first consider that

$$\big(\bar{\mathrm{K}}^2{}_{\mathrm{G}}(\mathrm{T}; \mathrm{x}, \mathrm{b})\big) = \int_0^\infty \mathrm{F(t)}\big(2\mathrm{k_G(t; x, b)}\bar{\mathrm{K}}_{\mathrm{G}}\mathrm{(t; x, b)}\big)\mathrm{dt} \quad \text{(using integral by part)} \tag{5}$$

and now $T \sim g(t)$ where $g(t) = 2k_G(t; x, b)\bar{K}_G(t; x, b)$, for $t \geq 0$ (see appendix for details). Using the result of Lemma 2 in the appendix, we have

$$E(T-x) = b\left(1 - \frac{\Gamma(2\,x/b+2)}{\Gamma^2(x/b+1)2^{2x/b+1}}\right),$$
$$E(T-x)^2 = xb - \frac{3b^2\Gamma(2\,x/b+2)}{\Gamma^2(x/b+1)2^{2x/b+1}} + b^2.$$

Following [15] define, $B_b(x) = \frac{b^{-1}\Gamma(2x/b+1)}{\Gamma^2(x/b+1)2^{2x/b+1}}$. [15] proved that B_b is bounded from above by $\frac{b^{1/2}x^{-1/2}}{2\sqrt{\pi}}$ and

$$B_b(x) \approx \begin{cases} \frac{1}{2\sqrt{\pi}} b^{-1/2}x^{-1/2} & if \quad x/b \to \infty; \\ C_k b^{-1} & if \quad \frac{x}{b} \to k. \end{cases}$$

Therefore when $x/b \to \infty$ we have

$$E(T-x) = b - (2xb + b^2)B_b(x) = b - \sqrt{xb/\pi} + O(b^{3/2}),$$
$$E(T-x)^2 = xb - 3b^3\left(\frac{2x}{b} + 1\right)B_b(x) + b^2 = xb(1 - 3\sqrt{xb/\pi}) + O(b^2)$$

and by using Taylor expansion, one gets

$$E\big(\bar{K}^2{}_G(T; x, b)\big) \approx F(x) - \sqrt{\frac{xb}{\pi}}f(x) + b\left(f(x) + \frac{x}{2}f'(x)\right) + O(b) \tag{6}$$

so, the variance can be simplified as follows:

$$\mathrm{var}\left(\hat{F}_G(x)\right) = \begin{cases} n^{-1}F(x)\big(1-F(x)\big) - n^{-1}\sqrt{xb/\pi}f(x) + O(bn^{-1}) & if \quad x/b \to \infty; \\ n^{-1}F(x)\big(1-F(x)\big) + n^{-1}bf(x)(1-C_k(1+2k)) + O(bn^{-1}) & if \quad x/b \to k. \end{cases}$$

It can be seen that the variance of the proposed

estimator in the interior points is smaller than the variance of the empirical distribution by the amount of $-n^{-1}\sqrt{xb/\pi}f(x)$. This is our gain from using smoothing. The story is different for the boundary points where $1-C_k(1+2k)>0$ for $k \le 2.381$. So for $x \le 2.381b$, the variance of the proposed estimator is larger than the variance of the empirical distribution (the difference is very small), however, outside this small region, the variance of the $\hat{F}(x)$ is smaller than the empirical distribution. [15] showed the same problem (increasing in variance near the boundary) in density estimation by Gamma kernel, but he also showed that this has negligible impact on the integrated variance [22] has shown that the gamma kernel probability density estimator in [15] has the boundary problem for those densities $f(x)$ which do not exhibit a shoulder, and whose derivative of $f(x)$ is not zero at $x=0$ since their MSE values in the boundary points converge to zero at the rate $O(n^{-2/3})$, instead of the usual rate of $O(n^{-4/5})$. But this is not the case for the proposed gamma kernel cumulative distribution estimator. The reason is revealed by comparing the bias and variance of the proposed estimator in Equations (3) and (4) with the corresponding equations in the gamma kernel probability density estimator in [15]. As we have just seen in the derivation of the bias and variance of the proposed estimator, here the expression $f'(x)$ appears everywhere with an x-factor and consequently, as x approaches zero, $f'(x)$ will vanish.

The mean square error (MSE) of the proposed estimator for $x/b \to \infty$ is

$$MSE\left(\hat{F}_G(x)\right) = n^{-1}F(x)\big(1-F(x)\big) - n^{-1}\sqrt{\frac{xb}{\pi}}f(x) + +b^2\left(f(x)+\frac{1}{2}xf'(x)\right)^2 + O(b^2+bn^{-1}),$$

and an estimate of the mean integrated square error (MISE) for the Gamma kernel estimator can be derived as follows (see [23] page 41):

$$\begin{aligned} MISE\left(\hat{F}_G(x)\right) &= \int_0^\infty MSE\left(\hat{F}_G(x)\right)dx \\ &= n^{-1}\int_0^\infty F(x)\big(1-F(x)\big)dx \\ &\quad - n^{-1}\frac{b^{\frac{1}{2}}}{\sqrt{\pi}}\int_0^\infty x^{\frac{1}{2}}f(x)dx \\ &\quad + b^2\int_0^\infty\left(f(x)+\frac{1}{2}xf'(x)\right)^2 dx \\ &\quad + O(b^2+bn^{-1}). \end{aligned}$$

The optimal smoothing parameter which minimizes MISE is

$$b_{opt} = \left(\int_0^\infty x^{\frac{1}{2}}f(x)dx\right)^{\frac{2}{3}}\left(4\sqrt{\pi}\int_0^\infty\left(f(x)+\frac{1}{2}xf'(x)\right)^2 dx\right)^{-2/3} n^{-2/3}. \quad (7)$$

So, the optimal smoothing parameter is $O(n^{-2/3})$. Plugging back b_{opt} in the MISE we have

$$\begin{aligned} MISE_{opt}\left(\hat{F}_G(x)\right) \approx n^{-1}\int_0^\infty F(x)\big(1-F(x)\big)dx \\ -\frac{3}{4}\frac{\left(\int_0^\infty x^{\frac{1}{2}}f(x)dx\right)^{\frac{4}{3}}(n\sqrt{\pi})^{-\frac{4}{3}}}{\left(4\int_0^\infty\left(f(x)+\frac{1}{2}xf'(x)\right)^2 dx\right)^{\frac{1}{3}}} \\ + O\left(n^{-\frac{4}{3}}\right) \end{aligned}$$

so, it can be seen that, the optimal MISE of the proposed estimator is smaller than the MISE of the empirical distribution.

Theorem 2. *Suppose that Assumptions 1-2 holds, then as* $n \to \infty$,

a.	*b.*
$\hat{F}_G(x) \xrightarrow{a.s.} F(x)$	$\hat{F}_G(x) \sim N(F(x), n^{-1}F(x)\big(1-F(x)\big))$

Proof: The proof is the same as in [24]. For part a, consider that

$$\begin{aligned} \sup_x\left|\hat{F}_G(x)-F(x)\right| \\ \le \sup_x\left|\hat{F}_G(x)-E(\hat{F}_G(x))\right| \\ + \sup_x\left|E(\hat{F}_G(x))-F(x)\right|. \end{aligned}$$

We have $n\left|\hat{F}_G(x)-E(\hat{F}_G(x))\right| = \sum_{i=1}^n \xi_i$, where $\xi_i = \int_{X_i}^\infty k_G(y;x,b)dy - E\left\{\int_{X_i}^\infty k_G(y;x,b)dy\right\}$. Now we have $E(\xi_i)=0$, and

$$\begin{aligned} \sigma^2 = E(\xi_i^2) = F(x)(1-F(x)) \\ + b\left[f(x)\left(1-F(x)-\frac{1}{\sqrt{\pi}}\sqrt{1+\frac{x}{b}}\right) \right. \\ \left. + xf'(x)\left(\frac{1}{2}-F(x)\right)\right] + O(b^2), \end{aligned}$$

since $\{\xi_i\}_{i=1}^n$ are independent zero mean random variables and $|\xi_i| \le 1$, by using Bernstein's inequality we have

$$P\left(\left|\sum_{i=1}^{n}\xi_i\right| > \frac{\varepsilon}{n}\right) \le 2\exp\left\{-1/2\left(\frac{\varepsilon^2}{n^3\sigma^2 + n\varepsilon/3}\right)\right\}.$$

Thus $\left|\hat{F}_G(x) - E(\hat{F}_G(x))\right| \to 0$ almost completely and consequently $\left|\hat{F}_G(x) - E(\hat{F}_G(x))\right| \xrightarrow{a.s.} 0$.

This result can also be proofed by using Glivenko–Cantelli Theorem. Note that

$$\hat{F}_G(x) = n^{-1}\sum_{i=1}^{n}\bar{K}_G(X_i; x, b) = \int \bar{K}_G(y; x, b)dF_n(y) \quad \text{and} \quad E\left(\hat{F}_G(x)\right) = \int \bar{K}_G(y; x, b)dF(y)$$

$$\Rightarrow \hat{F}_G(x) - E(\hat{F}_G(x)) = \int \bar{K}_G(y; x, b)\left(dF_n(y) - dF(y)\right) = \int \left(F_n(y) - F(y)\right)k_G(y; x, b)dy$$

$$\Rightarrow \sup_x\left|\hat{F}_G(x) - E(\hat{F}_G(x))\right| \le \|F - F_n\|\sup_x \int k_G(y; x, b)dy \le \|F - F_n\| \xrightarrow{a.s.} 0.$$

Recalling from Equation (3) that $E(\hat{F}_G(x)) - F(x)$ is bounded by $b(f(x) + xf'(x)/2) + O(b)$, it can be concluded that $\hat{F}_G(x) \xrightarrow{a.s.} F(x)$.

For part b define $Z_n = \frac{\sqrt{n}\,\bar{\xi}}{\sigma}$ in which $\bar{\xi} = n^{-1}\sum_{i=1}^{n}\xi_i$ and $\sigma^2 = E(\xi_i^2)$. Since $|\xi_i| \le 1$ so $E(\xi_i^3) = E(\xi_i^2\xi_i) \le E(\xi_i^2) = \sigma^2$, and by using Berry-Esseen Bound, we have

$$\sup_z |P(Z_n \le z) - \Phi(z)| \le \frac{33}{4}\frac{E(\xi_i^3)}{\sqrt{n}\sigma^3} \le \frac{33}{4}\frac{1}{\sqrt{n}\sigma}.$$

For sufficiently large ns, $\frac{1}{\sqrt{n}\sigma} \to 0$, $\sigma^2 \to F(x)(1 - F(x)$ and $E(\hat{F}_G(x)) \to F(x)$ and the result follows.

Results

In this section, the performance of the proposed estimators is compared with the Ordinary kernel method, the Boundary kernel method [11] and the empirical distribution method. Epanechnikov kernel is used for both the Ordinary kernel method and the Boundary kernel method. The optimal bandwidth proposed by [25] and [11] are used for the Ordinary kernel method, and the Boundary kernel method, respectively. Two hundred samples of two sizes $n = 200$ and $n = 500$ from eight various distributions including 1: Exponential (2), 2: Gamma (0.7,2), 3: Gamma (4,2), 4: Half Normal (0,1), 5: Log Normal (0,0.75), 6: Weibull (1.5,1.5), 7: 0.6 Gamma (4,0.4)+ 0.4 Normal (6,1), 8: 0.4 Half Normal (0,1)+ 0.6 Normal (4,1) are generated. Note that distributions 7 and 8 are mixed and estimating their cumulative distribution functions can be challenging. Then the integrated squared error $ISE_i = \int_0^{\infty}\left(\hat{F}_i(x) - F(x)\right)^2 dx$ is employed as an error metrics, where $\hat{F}_i(x)$, $i = 1, 2, 3, 4$ stands for the proposed estimator, Ordinary kernel method, the Boundary kernel method and the Empirical distribution method, respectively. In practice, the integral is approximated with summation.

Table 1 shows the mean and standard deviation of

Table 1. The mean and standard deviation of the ISE in 200 estimates of eight distributions for two sample size (n=200, 500) via four methods

$(\times 10^{-4})$	Method	Gamma kernel		Ordinary kernel		Boundary kernel		Empirical distribution	
Sample size	Example	Mean	Std.	Mean	Std.	Mean	Std.	Mean	Std.
200	1	4.31	3.78	5.90	5.06	4.31	4.03	4.85	3.92
	2	3.37	2.92	3.74	3.88	3.42	3.38	3.75	3.13
	3	4.97	4.36	6.02	5.03	5.46	4.60	5.98	4.60
	4	2.32	2.42	2.63	1.90	2.29	2.53	2.67	2.16
	5	4.45	3.86	4.81	4.43	4.56	3.9	4.78	3.91
	6	1.99	2.16	2.64	2.16	2.03	1.96	2.54	2.05
	7	2.46	1.70	5.69	3.39	2.41	1.32	2.47	1.38
	8	2.18	1.60	4.03	2.05	2.03	1.45	2.58	1.48
500	1	1.64	1.34	2.55	1.68	1.65	1.36	1.80	1.35
	2	1.65	1.58	2.41	1.98	1.67	1.64	1.73	1.60
	3	2.18	2.13	2.69	2.29	2.18	2.14	2.41	2.14
	4	1.25	1.38	1.43	1.45	1.23	1.44	1.45	1.45
	5	1.74	1.64	2.64	1.76	1.78	1.69	1.85	1.78
	6	0.84	0.80	1.16	0.96	0.86	0.83	0.97	0.81
	7	0.93	0.63	2.12	1.53	0.94	0.57	0.97	0.58
	8	1.25	0.72	6.69	1.52	1.18	0.74	1.37	0.75

the ISE for the ten distributions and the two sample sizes over two hundred repetitions. As can be seen from Table 1, in all cases, the mean and standard deviation of the ISE is decreased as the sample size is increased.

The simulation results show that for the sample size n=200 the performance of the proposed Gamma estimator is on par with the Boundary kernel method and better than the other two methods. For the sample size n=500, the proposed method's performance is slightly better than the Boundary kernel method.

Figure 1 shows plots of 30 estimates of the eight distributions via four methods. The true distribution is shown in boldface curve and the sample size is n = 200. The poor performance of the Ordinary kernel method especially for mixed distributions is obvious. The other estimators, even for the mixed distributions, show good agreement with the actual distribution. However, the empirical distribution estimates are not smooth.

Estimation of the probability distribution function of the monthly food cost of urban households in Iran

Household cost analysis is very important issue and many studies have been conducted in this field every year. Based on the results of these studies, the consumption pattern of households would be extracted, and by studying the trend of consumption of goods and services, justice-oriented economic policies would be evaluated, the distribution of income and facilities among households would be explained, the interrelationships of socio-economic characteristics of households would be studied, and finally the number of families below the poverty line would be extracted and the required information would be provided in national and regional accounting. Meanwhile, the cost of food is an important and unavoidable part of household consumption costs. In recent years, due to economic problems and inflation, especially in urban areas, it has become very difficult for many Iranian households to

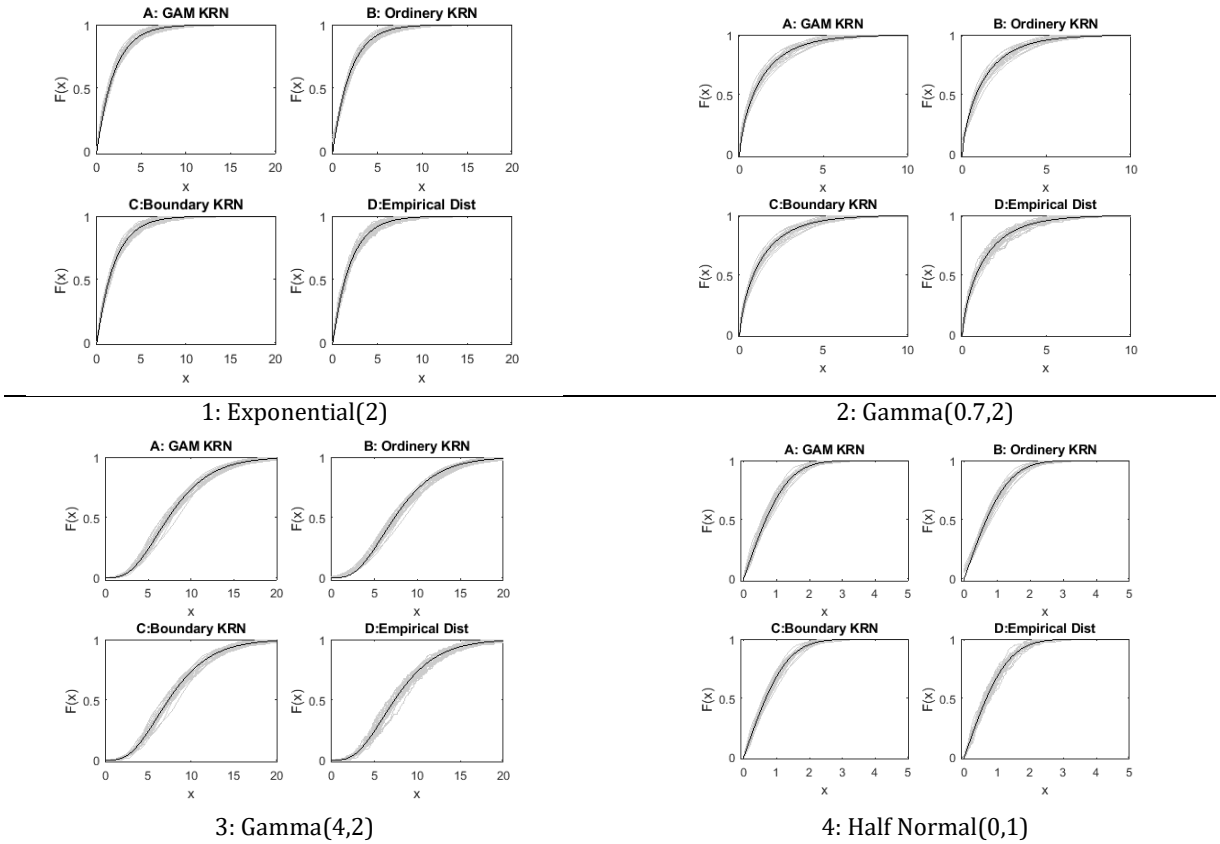

Figure 1. Plots of 30 estimates of the eight distributions via four methods. The true distribution is shown in boldface curve and the sample size is n = 200.

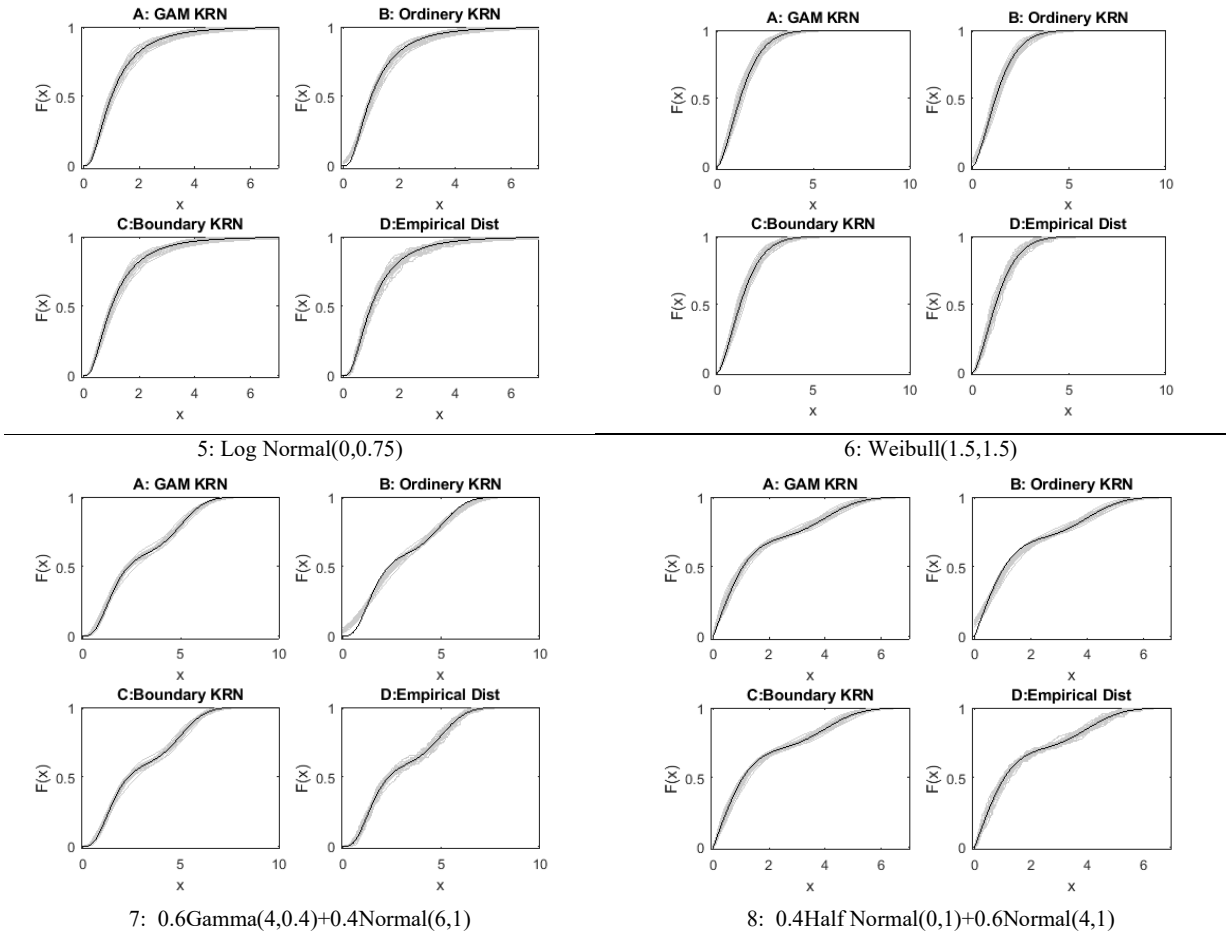

Figure 1. Plots of 30 estimates of the eight distributions via four methods. The true distribution is shown in boldface curve and the sample size is n = 200.

meet these costs, and supportive policies have been adopted to support the vulnerable by the government. Distributing cost probability can help policymakers to determine the amount of subsidy to those in need. In this section, the probability distribution function of the cost of one month of food items for urban households in Iran in 1398 AH (2019 AD) is estimated. The data of this section is related to 19827 urban households that have been randomly selected from all over the country by the Statistics Center of Iran and the information about food consumed during one month and the price of each unit in Rials using a detailed questionnaire of sixty-nine pages has been compiled. Data and questionnaires are retrieved from https://www.amar.org.ir/. For simplicity, prices are divided into ten million Rials (In that year on average, every 129183 Rials was equivalent to $1 in U.S. currency). Figure 2 shows the histogram of the data. As can be seen, the accumulation of a large part of the data is near the source and the positive skewness of the data is quite obvious.

Figure 3 demonstrates an estimate of the probability distribution function of the monthly household food expenses. In addition to the gamma estimator, the empirical distribution and the estimates using the Ordinary kernel method and the Boundary method are also presented. Figure 3 b zooms on the left boundary area to show the boundary problem in the Ordinary kernel estimation. The gamma estimator shows more agreement with the empirical distribution though very little. Using the probability distribution function estimated by the proposed method, the monthly cost of

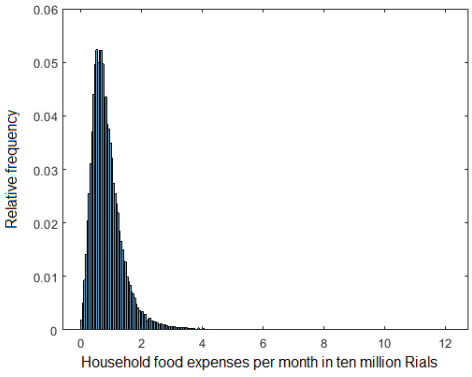

Figure 2. Histogram of the household food expenses per a month in ten million Rials in Iran (1398 AH, 2019 AD).

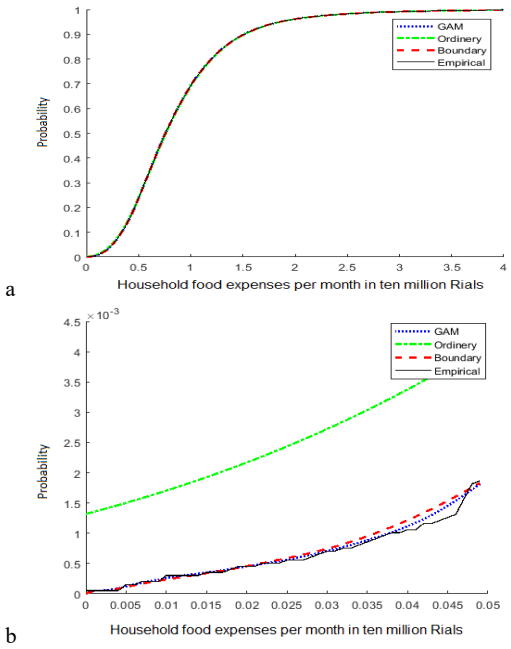

Figure 3. Estimating the cumulative probability distribution function of the household food expenses via four method. Plot b shows the left boundary in more details.

food for 95% of urban households was less than 18520000 Rials. During that year under review, the government subsidized 455,000 Rials for an individual in a low-income household per month. If each urban household has an average of 4 members, then the amount of 1,820,000 Rials monthly subsidy has just provided the cost of food 2.48 percent of urban households.

Acknowledgements

We are grateful to the Research Council of Shahid Chamran University of Ahvaz for financial support (99/3/02/18287).

References

1. Nadaraya EA. Some new estimates for distribution functions. Theory Probab. Appl. 1964;9:497-500.
2. Watson G and Leadbetter M. Hazard analysis II. Sankhya Ser. A. 1964;26:101–116.
3. Singh RS, Gasser T and Prasad B. Nonparametric estimates of distribution functions. Commun Stat Simul Comput. 1983;12:2095-2108.
4. Falk M. Relative efficiency and deficiency of kernel type estimators of smooth distribution functions. Stat Neerl. 1983;37:73-83.
5. Reiss R D. Nonparametric estimation of smooth distribution functions. Scand Stat Theory Appl. 1981;8:116-119.
6. Rice J. Boundary modification for kernel regression. Commun Stat Theory Methods. 1984;13:893–900.
7. Gasser T and Muller H. Kernels estimation of regression functions. Lect. Notes Math. 1979;757:23-68.
8. Gasser T, Muller H and Mammitzsch V. Kernels for nonparametric curve estimation. J R Stat Soc Series B Stat Methodol. 1985;47:238–252.
9. Muller H. Smooth optimum kernel estimators near endpoints. Biometrika. 1991;78:521–530.
10. Geenens G. Probit transformation for kernel density estimation on the unit interval, J Am Stat Assoc. 2014;109:346–358.
11. Tenreiro C. Boundary kernels for distribution function estimation. Revstat Stat J. 2013;11:169-190.
12. Tenreiro C. A new class of boundary kernels for distribution function estimation. Commun Stat Theory Methods. 2018;47:5319-5332.
13. Chacón JE and Duong T. Multivariate Kernel Smoothing and Its Applications, Chapman & Hall, 2018.
14. Chen SX. Beta kernel estimators for density functions. Comput Stat Data Anal. 1999;31:131–145.
15. Chen SX. Probability density function estimation using gamma kernels. Ann Inst Stat Math. 2000;52:471-480.
16. Jin X and Kawczak J. Birnbaum–Saunders and lognormal kernel estimators for modelling durations in high frequency financial data. Ann Econ Finance. 2003;4:103-124.
17. Scaillet O. Density estimation using inverse and reciprocal inverse Gaussian kernels. J Nonparametr Stat. 2004;16:217-226.
18. Hirukawa M and Sakudo M. Nonnegative bias reduction methods for density estimation using asymmetric kernels. Comput Stat Data Anal. 2014;75:112–123.
19. Igarashi G. Bias reductions for beta kernel estimation. J Nonparametr Stat. 2016;28:1–30.
20. Bouezmarni T and Scaillet O. Consistency of asymmetric kernel density estimators and smoothed histograms with application to income data. Econ Theory. 2005;21:390-

412.
21. Mombeni HA, Mansouri B and Akhoond MR. Asymmetric kernels for boundary modification in distribution function estimation. Revstat Stat J Forthcoming papers. 2019.
22. Zhang S. A note on the performance of the gamma kernel estimators at the boundary. Stat Probab Lett. 2010;80:548-557.
23. Scott DW. Multivariate Density Estimation, Theory, Practice and Visualization, Second edition. John Wiley & Sons, 2015.
24. Liu R and Yang L. Kernel estimation of multivariate cumulative distribution function. J Nonparametr Stat. 2008;20:661-677.
25. Altman N and Leger C. Bandwidth selection for kernel distribution function estimation. J Stat Plan Inference. 1995;46:195-214.

Appendix

Using integral by parts, it is easy to show that $\frac{1}{\Gamma(k+1)}\int_{\theta}^{\infty} e^{-x}x^{k}dx = \sum_{j=0}^{k}\frac{e^{-\theta}\theta^{j}}{j!}$.

Let Z be a gamma random variable with parameters α and β i.e. $f(z) = \frac{\beta^{\alpha}}{\Gamma(\alpha)}z^{\alpha-1}e^{-\beta z}$. Now, we can deduce that $1 - F(z) = \frac{1}{\Gamma(\alpha)}\int_{z\beta}^{\infty} e^{-x}x^{\alpha-1}dx = \sum_{j=0}^{\alpha-1}\frac{e^{-z\beta}(z\beta)^{j}}{j!}$.

Define $g(z) = 2f(z)\big(1 - F(z)\big) = \frac{2e^{-2z\beta}}{\Gamma(\alpha)}\sum_{j=0}^{\alpha-1}\frac{z^{j+\alpha-1}\beta^{j+\alpha}}{j!}$.

Lemma 1. We have

$$\sum_{j=0}^{\alpha-1}\frac{\Gamma(j+\alpha)}{\Gamma(j+1)2^{j+\alpha}} = \frac{\Gamma(\alpha)}{2}. \qquad (A.1)$$

Proof. Since $\int_{0}^{\infty} 2f(z)\big(1 - F(z)\big)\,dz = 1$, the result follows.

Lemma 2. Let Z be a random variable with distribution $g(z)$. Then, we have

$$E(Z) = \frac{1}{\beta}(\alpha - \frac{\Gamma(2\alpha)}{\Gamma^{2}(\alpha)2^{2\alpha-1}})$$

and

$$E(Z^{2}) = \frac{1}{\beta^{2}}(\alpha + \alpha^{2} - \frac{(2\alpha+1)\Gamma(2\alpha)}{\Gamma^{2}(\alpha)2^{2\alpha-1}})$$

.

Proof. $E(Z) = \sum_{j=0}^{\alpha-1}\int_{0}^{\infty}\frac{2\beta^{j+\alpha}}{\Gamma(\alpha)\Gamma(j+1)}z^{j+\alpha}e^{-2z\beta}dz$

$$= \frac{2}{\Gamma(\alpha)}\sum_{j=0}^{\alpha-1}\frac{\Gamma(j+\alpha+1)\beta^{j+\alpha}}{\Gamma(j+1)(2\beta)^{j+\alpha+1}}\int_{0}^{\infty}\frac{(2\beta)^{j+\alpha+1}}{\Gamma(j+\alpha+1)}z^{j+\alpha}e^{-2z\beta}dz$$

$$= \frac{1}{\beta\Gamma(\alpha)}\sum_{j=0}^{\alpha-1}\frac{\Gamma(j+\alpha+1)}{\Gamma(j+1)2^{j+\alpha}} = \frac{1}{\beta\Gamma(\alpha)}\sum_{j=0}^{\alpha-1}\frac{(j+\alpha)\Gamma(j+\alpha)}{\Gamma(j+1)2^{j+\alpha}}.$$

Using Lemma 1, we get

$$E(Z) = \frac{1}{\beta\Gamma(\alpha)}\left\{\alpha\frac{\Gamma(\alpha)}{2} + \sum_{j=0}^{\alpha-1}\frac{j\Gamma(j+\alpha)}{\Gamma(j+1)2^{j+\alpha}}\right\}. \qquad (A.2)$$

Define $E_1 = \sum_{j=1}^{\alpha-1}\frac{j\Gamma(j+\alpha)}{\Gamma(j+1)2^{j+\alpha}}$. It is easy to show that $E_1 = \alpha\frac{\Gamma(\alpha)}{2} - \frac{\Gamma(2\alpha)}{2^{2\alpha-1}\Gamma(\alpha)}$. By substituting E_1in (A2), the proof is complete. Now we turn to $E(Z^{2})$.

$$E(Z^{2}) = \sum_{j=0}^{\alpha-1}\int_{0}^{\infty}\frac{2\beta^{j+\alpha}}{\Gamma(\alpha)\Gamma(j+1)}z^{j+\alpha+1}e^{-2z\beta}dz$$

$$= \frac{1}{2\beta^{2}\Gamma(\alpha)}\sum_{j=0}^{\alpha-1}\frac{(j+\alpha+1)(j+\alpha)\Gamma(j+\alpha)}{\Gamma(j+1)2^{j+\alpha}}$$

$$= \frac{1}{2\beta^{2}\Gamma(\alpha)}\left\{(\alpha^{2}+\alpha)\frac{\Gamma(\alpha)}{2} + \sum_{j=0}^{\alpha-1}\frac{j^{2}\Gamma(j+\alpha)}{\Gamma(j+1)2^{j+\alpha}} + (2\alpha+1)\sum_{j=0}^{\alpha-1}\frac{j\Gamma(j+\alpha)}{\Gamma(j+1)2^{j+\alpha}}\right\}$$

$$= \frac{1}{2\beta^{2}\Gamma(\alpha)}\left\{(\alpha^{2}+\alpha)\frac{\Gamma(\alpha)}{2} + E_2 + (2\alpha+1)E_1\right\}, \qquad (A.3)$$

and

$$E_2 = \sum_{j=0}^{\alpha-1}\frac{j^{2}\Gamma(j+\alpha)}{\Gamma(j+1)2^{j+\alpha}} = \sum_{j=1}^{\alpha-1}\frac{j\Gamma(j+\alpha)}{\Gamma(j)2^{j+\alpha}}$$

$$= \frac{1}{2}\sum_{j=0}^{\alpha-2}\frac{(j+1)(j+\alpha)\Gamma(j+\alpha)}{\Gamma(j+1)2^{j+\alpha}}$$

$$= \frac{1}{2}\sum_{j=0}^{\alpha-1}\frac{(j+1)(j+\alpha)\Gamma(j+\alpha)}{\Gamma(j+1)2^{j+\alpha}} - \frac{1}{2}\frac{\alpha\Gamma(2\alpha)}{2^{2\alpha-1}\Gamma(\alpha)}$$

$$= \frac{1}{2}\left\{E_2 + (\alpha+1)E_1 + \alpha\frac{\Gamma(\alpha)}{2}\right\} - \frac{1}{2}\frac{\alpha\Gamma(2\alpha)}{2^{2\alpha-1}\Gamma(\alpha)},$$

$$\Rightarrow E_2 = (\alpha+1)E_1 + \alpha\frac{\Gamma(\alpha)}{2} - \frac{\alpha\Gamma(2\alpha)}{2^{2\alpha-1}\Gamma(\alpha)}.$$

By substituting $E_1$and E_2 in (A.3) and simplifying it, the proof is complete.