STATISTICAL PROPERTIES OF THE NONPARAMETRIC KUMARASWAMY KERNEL DENSITY ESTIMATOR WITH APPLICATION

Samah M. Abo-El-Hadid
Department of Mathematics, Insurance and Applied Statistics,
Faculty of Commerce and Business Administration, Helwan University.
Cairo, Egypt.

ABSTRACT

The In this paper, new kernel density estimator so called Kumaraswamy kernel estimator is introduced. The asymptotic bias, variance, mean squared error (MSE), integrated mean squared error (IMSE), and the optimal smoothing parameter of the proposed estimator are obtained. A simulation study is used to compare the new estimator with other estimators. Finally, the proposed Kumaraswamy kernel estimator was applied to a real Egyptian Average monthly precipitation rate dataset.

Keywords: *Kumarawamy distribution, Kernel estimator, Precipitation rate.*

1. INTRODUCTION

Kumaraswamy distribution was originally developed by the Indian hydrologist Poondi Kumaraswamy in a paper published in 1980 to model hydrological phenomena (see Kumaraswamy 1980). The probability density function of the Kumaraswamy distribution with two shape parameters a, b is:

$$K(u) = a\, b\, u^{a-1}(1-u^a)^{b-1} \qquad 0 < u < 1, \qquad a, b > 0 \tag{1}$$

If the random variable u is distributed Kum (a, b), its moments around zero can be expressed as (Mitnik 2013):

$$E(u^r) = b\, Beta\left(\frac{r}{a}+1, b\right) \tag{2}$$

$$= \frac{b\,\Gamma\left(\frac{r}{a}+1\right)\Gamma(b)}{\Gamma\left(\frac{r}{a}+b+1\right)} \tag{3}$$

Thus, the expectation (first moment), the second moment, and variance of u are:

$$E(u) = \int_0^1 u\, K(u)\, du = \frac{b\,\Gamma\left(\frac{1}{a}+1\right)\Gamma(b)}{\Gamma\left(\frac{1}{a}+b+1\right)} \tag{4}$$

$$E(u^2) = \int_0^1 u^2\, K(u)\, du = \frac{b\,\Gamma\left(\frac{2}{a}+1\right)\Gamma(b)}{\Gamma\left(\frac{2}{a}+b+1\right)} \tag{5}$$

$$var(u) = \frac{b\,\Gamma\left(\frac{2}{a}+1\right)\Gamma(b)}{\Gamma\left(\frac{2}{a}+b+1\right)} - \left[\frac{b\,\Gamma\left(\frac{1}{a}+1\right)\Gamma(b)}{\Gamma\left(\frac{1}{a}+b+1\right)}\right]^2 \tag{6}$$

In this paper the Kumaraswamy distribution in equation (1), is used to introduce a new nonparametric kernel estimator. The kernel density estimator method was introduced by Rosenblatt (1956). He considered $\hat{f}(x)$ as an estimator of the unknown density $f(x)$:

$$\hat{f}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \qquad , -\infty \le x \le \infty \tag{7}$$

Where n is the sample size; $K(\cdot)$ and h are the kernel function and the bandwidth respectively, where the kernel function $K(\cdot)$ is assumed to be symmetric density function.

Parzen (1962) studied the statistical properties of kernel density estimator in equation (1), and proved that $\hat{f}(x)$ is biased and consistent estimator. He also obtained the optimal bandwidth h_{opt} which minimizes the integrated mean squared error of $\hat{f}(x)$.

Bagai and Rao (1995) studied the statistical properties of the kernel density estimator in the case of using asymmetric kernel function; also they found the optimal kernel function and the optimal bandwidth h_{opt} which minimizes the integrated mean squared errors (IMSE). Chen (1999) proposed using the density of Beta distribution as the kernel function when $x \in [0,1]$.

Chen (2000) suggested using the density of Gamma distribution as the kernel function for density estimation when $x \in [0, \infty$. Scaillet (2004) used inverse Gaussian and reciprocal inverse Gaussian probability density functions as kernels for densities defined on [0,1) support.
Bouezmarni et al. (2011) suggested using the gamma kernels for the density and the hazard rate functions for right censored data; they also studied IMSE, the asymptotic normality and the law of iterated logarithm of this estimator. Markovich (2015) estimated the derivative of a probability density function defined on $[0,\infty)$ using the asymmetric gamma kernel functions; and obtained the optimal bandwidth for the case of dependent data. Markovich (2016) introduced new kernel estimator as a combination of the asymmetric gamma and Weibull kernels, also the theoretical asymptotic properties of the proposed density estimator and the optimal bandwidth selection for the estimate as a MISE are derived.
The paper is organised as follows: In Section 2, we introduce the new Kumaraswamy kernel density estimator. In Section 3 the smoothing parameter selection problem is discussed. Section 4 provides the results of a simulation study in which the behaviour of the Kumaraswamy kernel estimator is compared with the other density estimators. Real data application is introduced in section 5. Finally, in section 6, a brief conclusion is provided.

2. THE KUMARASWAMY KERNEL ESTIMATOR

We propose the use of Kumaraswamy distribution for density estimation in this paper. The Kumaraswamy kernel density estimator is as follows:

$$\hat{f}(x) = \frac{a\,b}{nh}\sum_{i=1}^{n}\left(\frac{x-x_i}{h}\right)^{a-1}\left[1-\left(\frac{x-x_i}{h}\right)^{a}\right]^{b-1} \tag{8}$$

It can be shown that the expectation of the kernel density estimator is:

$$E\left[\hat{f}(x)\right] = E\left[\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{x-x_i}{h}\right)\right] = \frac{1}{h}\int K\left(\frac{x-x_i}{h}\right)f(x)\,dx \tag{9}$$

Let $u = \frac{x-x_i}{h}$

$$\therefore E\left[\hat{f}(x)\right] = \frac{1}{h}\int K(u)\,f(x-uh)h\,du \tag{10}$$

Using Taylor expansion for $f(x-uh)$ yields that:

$$Bias\left[\hat{f}(x)\right] \simeq -hf'(x)\int uK(u)\,du + \frac{h^2}{2}f''(x)\int u^2K(u)\,du \tag{11}$$

Using equations (4) and (5), the asymptotic bias of the Kumaraswamy kernel density estimator is:

$$Bias\left[\hat{f}(x)\right] \simeq \frac{-h\,b\,\Gamma\left(\frac{1}{a}+1\right)\Gamma(b)}{\Gamma\left(\frac{1}{a}+b+1\right)}f'(x) + \frac{h^2\,b\,\Gamma\left(\frac{2}{a}+1\right)\Gamma(b)}{2\,\Gamma\left(\frac{2}{a}+b+1\right)}f''(x) \tag{12}$$

Also, it can be shown that the variance of the kernel density estimator is:

$$Var\left[\hat{f}(x)\right] = \frac{1}{nh^2}\left\{E\left[K\left(\frac{x-x_i}{h}\right)\right]^2 - \left[EK\left(\frac{x-x_i}{h}\right)\right]^2\right\} \tag{13}$$

$$= \frac{1}{nh^2}\left\{\int\left[K\left(\frac{x-x_i}{h}\right)\right]^2 f(x_i)dx_i - \left[\int K\left(\frac{x-x_i}{h}\right)f(x_i)dx_i\right]^2\right\}$$

let $u = \frac{x-x_i}{h}$, then

$$Var\left[\hat{f}(x)\right] = \frac{1}{nh^2}\{h\int K^2(u)f(x-uh)\,du - [h\int K(u)\,f(x-uh)\,du]^2\}$$

Again by Taylor expansion:

$$f(x-uh) = f(x) - uhf'(x) + \frac{u^2h^2}{2!}f''(x) + \cdots$$

Then

$$Var\left[\hat{f}(x)\right] \simeq \frac{1}{nh}f(x)\int K^2(u)du \tag{14}$$

Using the Kumaraswamy kernel function:

$$\int_0^1 K^2(u)\,du = \int_0^1 a^2\,b^2\,u^{2(a-1)}(1-u^a)^{2(b-1)}du \tag{15}$$

$$= \int_0^1 a^2\,b^2\,u^{a\left(\frac{2(a-1)}{a}\right)}(1-u^a)^{2(b-1)}du$$

Let $u^a = z$, then $u = z^{\frac{1}{a}}$, $du = \frac{1}{a}z^{\frac{1}{a}-1}\,dz$, then

$$\int_0^1 K^2(u)\,du = \int_0^1 a^2\,b^2\,z^{\frac{2(a-1)}{a}}(1-z)^{2(b-1)}\left(\frac{1}{a}z^{\frac{1}{a}-1}\right)dz$$

$$= ab^2\int_0^1 z^{\frac{2(a-1)+1}{a}-1}(1-z)^{2(b-1)}\,dz$$

$$= ab^2\,Beta\left(\frac{2(a-1)+1}{a}, 2b-1\right)$$

$$\therefore \int_0^1 K^2(u)\,du = \frac{ab^2\,\Gamma\left(2-\frac{1}{a}\right)\Gamma(2b-1)}{\Gamma\left(1+2b-\frac{1}{a}\right)} \quad (16)$$

Substitute equation (16) into equation (14), then the asymptotic variance of the Kumaraswamy kernel density estimator is:

$$Var\left[\hat{f}(x)\right] = \frac{ab^2 f(x)\,\Gamma\left(2-\frac{1}{a}\right)\Gamma(2b-1)}{nh\,\Gamma\left(1+2b-\frac{1}{a}\right)} \quad (17)$$

Combining (11) and (17), the mean squared errors for $\hat{f}(x)$ is:

$$MSE\left[\hat{f}(x)\right] = Var\left[\hat{f}(x)\right] + Bias^2\left[\hat{f}(x)\right]$$

$$= \frac{ab^2 f(x)\,\Gamma\left(2-\frac{1}{a}\right)\Gamma(2b-1)}{nh\,\Gamma\left(1+2b-\frac{1}{a}\right)} + h^2b^2\left(f'(x)\right)^2\left[\frac{\Gamma\left(\frac{1}{a}+1\right)\Gamma(b)}{\Gamma\left(\frac{1}{a}+b+1\right)}\right]^2 + o(h^2) \quad (18)$$

Where $o(h^2)$ higher than second order terms of h. Also, the asymptotic IMSE for $\hat{f}(x)$ is:

$$IMSE\left[\hat{f}(x)\right] \simeq \frac{ab^2\,\Gamma\left(2-\frac{1}{a}\right)\Gamma(2b-1)}{nh\,\Gamma\left(1+2b-\frac{1}{a}\right)} + h^2b^2\left[\frac{\Gamma\left(\frac{1}{a}+1\right)\Gamma(b)}{\Gamma\left(\frac{1}{a}+b+1\right)}\right]^2\int\left(f'(x)\right)^2\,dx \quad (19)$$

3. THE OPTIMAL SMOOTHING PARAMETER

The optimal smoothing parameter (h) which minimize the IMSE for $\hat{f}(x)$ is obtained as follows:

$$\frac{\partial\,IMSE[f(x)]}{\partial h} = \frac{-ab^2\,\Gamma\left(2-\frac{1}{a}\right)\Gamma(2b-1)}{nh^2\,\Gamma\left(1+2b-\frac{1}{a}\right)} + 2hb^2\left[\frac{\Gamma\left(\frac{1}{a}+1\right)\Gamma(b)}{\Gamma\left(\frac{1}{a}+b+1\right)}\right]^2\int\left(f'(x)\right)^2\,dx = 0 \quad (20)$$

$$\text{then } 2h^3b^2\left[\frac{\Gamma\left(\frac{1}{a}+1\right)\Gamma(b)}{\Gamma\left(\frac{1}{a}+b+1\right)}\right]^2\int\left(f'(x)\right)^2\,dx = \frac{ab^2\,\Gamma\left(2-\frac{1}{a}\right)\Gamma(2b-1)}{nh^2\,\Gamma\left(1+2b-\frac{1}{a}\right)}$$

and hence:

$$h_{opt} = \left[\frac{a\,\Gamma\left(2-\frac{1}{a}\right)\Gamma(2b-1)}{2n\Gamma\left(1+2b-\frac{1}{a}\right)\left[\frac{\Gamma\left(\frac{1}{a}+1\right)\Gamma(b)}{\Gamma\left(\frac{1}{a}+b+1\right)}\right]^2\int\left(f'(x)\right)^2\,dx}\right]^{1/3} \quad (21)$$

Now let us replace the unknown term $\int\left(f'(x)\right)^2$ in (21) by the Kumaraswamy density as a reference distribution. Let:

$$f(x) = a\,b\,x^{a-1}(1-x^a)^{b-1} \qquad ,\ 0 < x < 1 \quad (22)$$

then

$$f'(x) = -a^2\,b\,(b-1)\,x^{2a-2}(1-x^a)^{b-2} + a\,b(a-1)\,x^{a-2}(1-x^a)^{b-1} \quad (23)$$

and hence

$$\left(f'(x)\right)^2 = a^4\,b^2\,(b-1)^2\,x^{4a-4}(1-x^a)^{2b-4} + a^2\,b^2(a-1)^2\,x^{2a-4}(1-x^a)^{2b-2} - 2a^3\,(a-1)$$

$$(b-1)\,x^{3a-4}\,(1-x^a)^{2b-3}$$

$$\therefore \int_0^1\left(f'(x)\right)^2dx = a^3b^2\,(b-1)^2Beta\left(\frac{4a-3}{a}, 2b-3\right) + ab^2\,(a-1)^2Beta\left(\frac{2a-3}{a}, 2b-1\right)$$

$$-2a^2\,b^2(a-1)(b-1)Beta\left(\frac{3a-3}{a},2b-2\right)$$

$$\therefore \int_0^1 \left(f'(x)\right)^2 dx = \frac{a\,b^2(a-1)(b-1)(2ab-b-3)\,\Gamma\left(2-\frac{3}{a}\right)\Gamma(2b-3)}{\Gamma\left(2b-\frac{3}{a}+1\right)} \quad (24)$$

Substituting (24) into (21), we get

$$h_{opt} = \left[\frac{a\Gamma\left(2-\frac{1}{a}\right)\Gamma(2b-1)}{2n\Gamma\left(1+2b-\frac{1}{a}\right)\left[\frac{\Gamma\left(\frac{1}{a}+1\right)\Gamma(b)}{\Gamma\left(\frac{1}{a}+b+1\right)}\right]^2\left[\frac{a\,b^2(a-1)(b-1)(2ab-b-3)\,\Gamma\left(2-\frac{3}{a}\right)\Gamma(2b-3)}{\Gamma\left(2b-\frac{3}{a}+1\right)}\right]}\right]^{1/3} \quad (25)$$

4. SIMULATION

In this section, the influence of Kumaraswamy kernel estimator is examined using a simulation study. Kumaraswamy distribution is asymmetric distribution widely used to model wind velocity, because wind velocity is analyzed into its orthogonal 2-dimensional vector components. The Kumaraswamy kernel was given in equation (2).

In this section, the Kumaraswamy kernel is compared with the most widely used kernel functions: 1) The Gaussian kernel which is symmetric about zero:

$$K(u) = \frac{1}{\sigma\sqrt{2\pi}}\; e^{\left(-\frac{1}{2}u^2\right)} \qquad ,0 < u < \infty$$

2) The Gamma kernel which is asymmetric kernel:

$$K(u) = \frac{\beta^r u^{r-1} e^{-\beta u}}{\Gamma r} \qquad r,\beta > 0 \qquad ,0 < u < \infty$$

To evaluate the suggested Kumaraswamy estimator, we select the exponential random variable with ($\theta = 0.1$, θ =0.5, and θ =1). We generate the exponential i.i.d samples with sample sizes $n \in \{50, 100, 1000\}$, then, the actual and the estimated densities are plotted; and the following errors' measures are computed:

Mean squared error $(\mathrm{MSE}) = \frac{\sum_{i=1}^{n}\left(f(x_i)-f(x_i)\right)^2}{n}$ (26)

Mean absolute error $(\mathrm{MAE}) = \frac{\sum_{i=1}^{n}|f(x_i)-\hat{f}(x_i)|}{n}$ (27)

Mean absolute percentage error $(\mathrm{MAPE}) = \sum_{i=1}^{n}\left|\frac{f(x_i)-f(x_i)}{n\cdot f(x_i)}\right|$ (28)

These above measures are used to compare the fits obtained by different kernels. For all three measures, smaller values indicate a better fitting model. MSE is commonly-used measure of accuracy of fitted values but it is highly affected by outliers than MAE. MAPE expresses accuracy as a percentage of the error. The values of the above goodness of fit measures are given in tables (1,2,3) below:

Table 1. Goodness of fit measure's of the difference between the actual density and the estimated densities with $\theta = 0.1$

Sample Size	Measure	Estimated density		
		Kumaraswamy	Gamma kernel	Gaussian kernel
n = 50	MSE	0.000297	0.001984	0.002138
	MAE	0.010783	0.032501	0.033996
	MAPE	0.462362	0.577281	0.639893
n=100	MSE	0.000290	0.001958	0.002007
	MAE	0.008873	0.030345	0.032121
	MAPE	0.452804	0.563831	0.608663
n=1000	MSE	0.000158	0.001741	0.001801
	MAE	0.006583	0.028840	0.030649
	MAPE	0.153580	0.543387	0.671348

Table 2. Goodness of fit measure's of the difference between the actual density and the estimated densities with $\theta = .5$

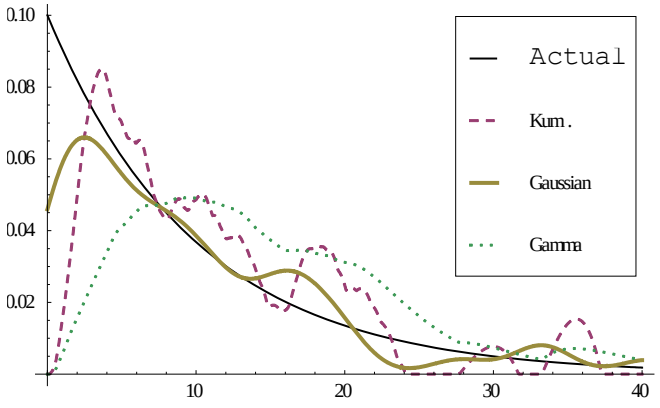
Sample Size	Measure	Estimated density		
		Kumaraswamy	Gamma kernel	Gaussian kernel
n = 50	MSE	0.03433	0.05086	0.07084
	MAE	0.14731	0.18319	0.21969
	MAPE	0.50237	0.94199	1.06986
n=100	MSE	0.03027	0.03744	0.05636
	MAE	0.13534	0.13842	0.18854
	MAPE	0.43670	0.47747	0.95795
n=1000	MSE	0.02907	0.03298	0.06604
	MAE	0.12114	0.12660	0.20475
	MAPE	0.39872	0.42491	0.99330

Table 3. Goodness of fit measure's of the difference between the actual density and the estimated densities with $\theta = 1$

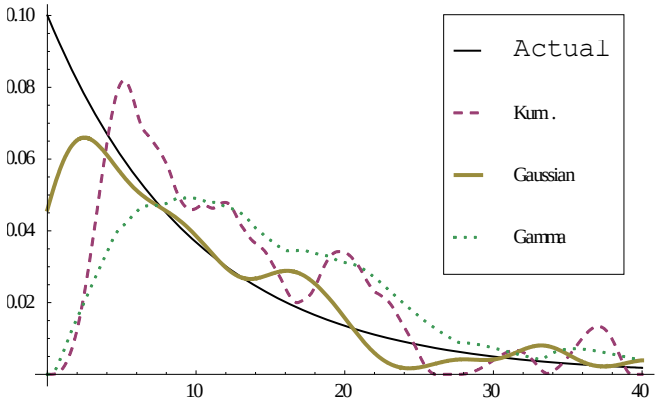
Sample Size	Measure	Estimated density		
		Kumaraswamy	Gamma kernel	Gaussian kernel
n = 50	MSE	0.203374	0.191002	0.308636
	MAE	0.318196	0.331753	0.458756
	MAPE	0.707607	0.820106	1.811820
n=100	MSE	0.172030	0.192857	0.296948
	MAE	0.314446	0.351150	0.430159
	MAPE	0.627280	0.755616	1.718690
n=1000	MSE	0.133558	0.191454	0.277108
	MAE	0.234049	0.318650	0.417522
	MAPE	0.386175	0.703782	0.979273

The above tables shows that under the generated exponential i.i.d samples, the estimated densities get closer to the original density function as the parameter θ decreases; and also the estimated densities get closer to the actual density as the sample size increases, and the suggested Kumaraswamy kernel always outperforms the others kernels, while the Gaussian kernel is the worst one .

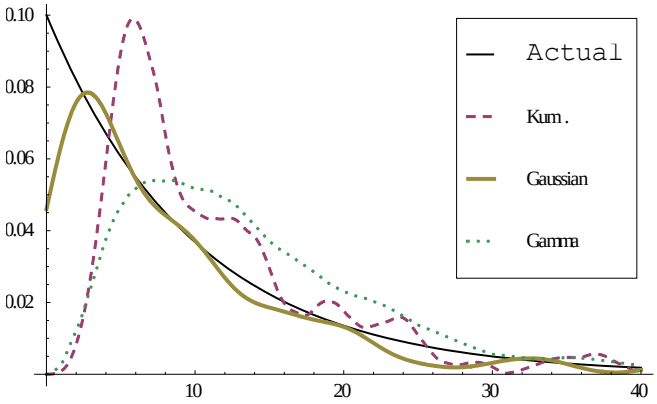
Figures (1), (2) and (3), present the actual density with the estimated densities using: Kumaraswamy kernel; Gaussian Kernel and Gamma kernel at the different values of parameter θ and different sample size.



(a)

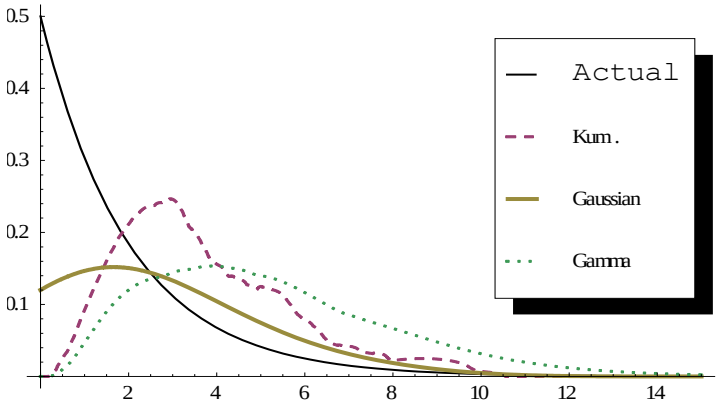


(b)

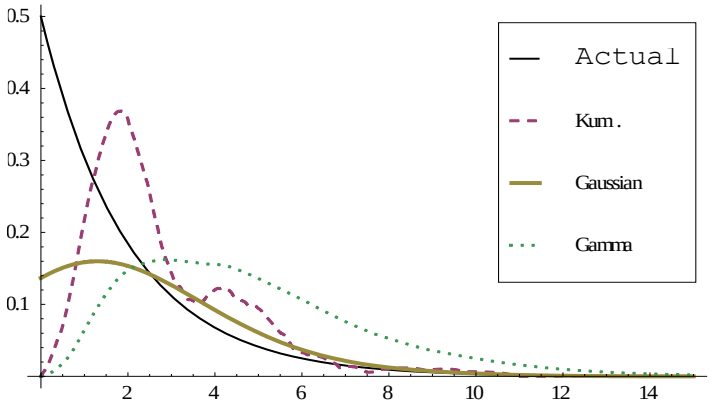


(c)

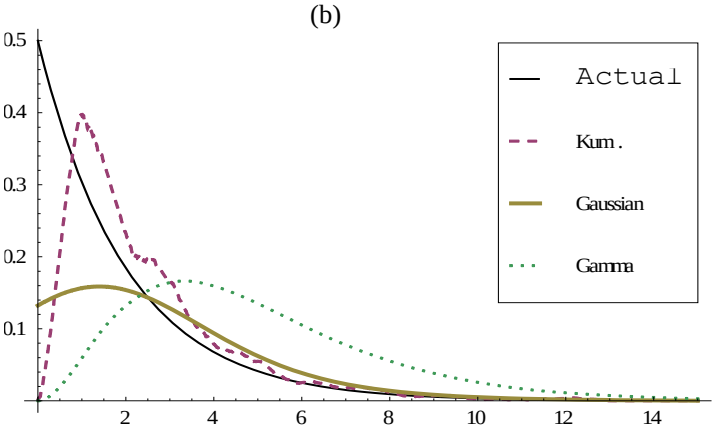
Figure 1. The actual Exponential density (θ =0.1) with its kernel estimation using Kumaraswamy; Gaussian l; and Gamma kernels with sample sizes (a) n=50, (b) n=100, (c) n=1000



(a)

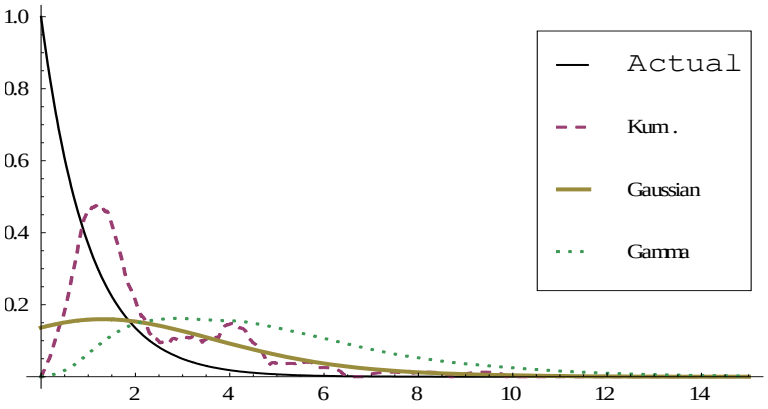


(b)

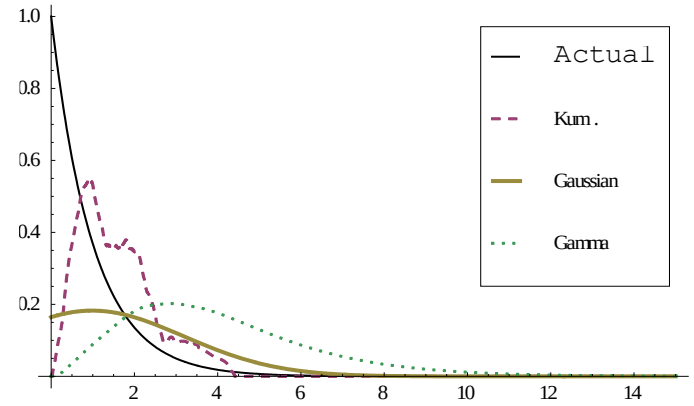


(c)

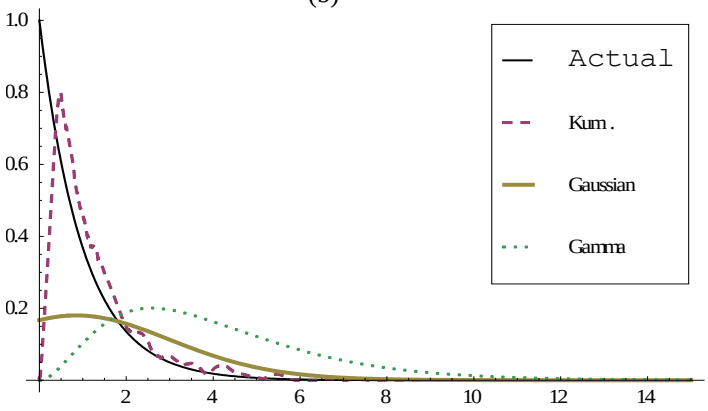
Figure 2. The actual Exponential density (θ =.5) with its kernel estimation using Kumaraswamy; Gaussian l; and Gamma kernels with sample sizes (a) n=50, (b) n=100, (c) n=1000



(a)



(b)



(c)

Figure 3. The actual Exponential density (θ =1) with its kernel estimation using Kumaraswamy; Gaussian l; and Gamma kernels with sample sizes (a) n=50, (b) n=100, (c) n=1000

The above figures show that under the generated exponential i.i.d samples, the three kernel functions gets closer to the original density function as the sample size increases for different values of parameter θ. Also, among the three kernel functions, the suggested Kumaraswamy kernel always outperforms the others for different values of θ and n while the Gaussian kernel is the worst one.

5. APPLICATION

Average monthly precipitation rate is determined by using rain gauges to measure the depth of rain that falls over a specific area. Rainfall measurements are taken every three hours at locations around the world. These observations are made at the same time at every location.

In this section, we apply the proposed Kumaraswamy kernel estimator to precipitation rate dataset, given in Climate Change Knowledge Portal for Development Practitioners and Policy Makers, the World Bank group (http://sdwebx.worldbank.org/climateportal/index.cfm?ThisCCode=EGY&page=country_historical_climate), which were collected (mm).

Figure 4, shows the estimated distribution of Average monthly precipitation rate data using both the parametric Kumaraswamy distribution and the nonparametric Kumaraswamy distribution.

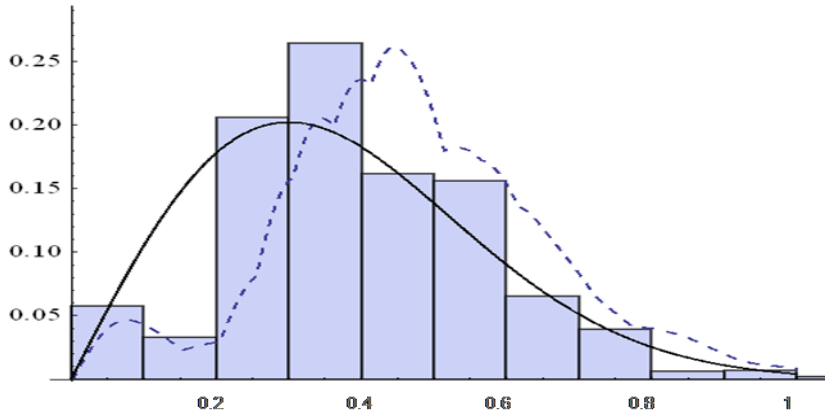


Figure 4. The histogram; parametric and nonparametric Kumaraswamy estimator of precipitation rate data

The above figure indicates the flexibility of Kumaraswamy kernel in modelling the precipitation rate distributions. It shows that the nonparametric Kumaraswamy estimator agrees well with the Average monthly precipitation rate data.

6. CONCLUSION

A new nonparametric kernel estimator so called the Kumaraswamy kernel estimator is introduced; some statistical properties of the new estimator were introduced. Finally, a simulation study and real data application were introduced. The suggested Kumaraswamy kernel always outperforms the others for different values of θ and n while the Gaussian kernel is the worst one.

7. REFERENCES

[1] Bagai, I. and P. Rao. (1995): "Kernel Type Density Estimates for Positive Valued Random Variables", *The Indian Journal of Statistics*, **57**, 56- 67.
[2] Bouezmarni, T., A. El Ghouch and M. Mesfioui (2011): "Gamma Kernel Estimators for Density and Hazard Rate of Right Censored Data", *Journal of Probability and Statistics*, 1-16.
[3] Chen, S. X. (1999): "Beta Kernel Estimators for Density Functions", *Computational Statistics and Data Analysis*, **31**, 131-145.
[4] Chen, S. X. (2000): "Probability Density Function Estimating Using Gamma Kernels", *Annals Institute of Mathematical Statistics*, **52**, 471-480.
[5] Climate Change Knowledge Portal For Development Practitioners and Policy Makers, the world bank group, http://sdwebx.worldbank.org/climateportal/index.cfm?ThisCCode=EGY&page=country_historical_climate.
[6] Kumaraswamy, P. (1980). "A generalized probability density function for double-bounded random processes", *Journal of Hydrology*. **46** (1–2): 79–88.
[7] Mitnik, P. A. (2013): "New Properties of the Kumaraswamy Distribution", *Communications in Statistics - Theory and Methods*, **42**, 741–755.
[8] Markovich, L. A. (2015): "Gamma Kernel Estimation of Multivariate Density and its Derivative on the Nonnegative Semi-axis by Dependent Data", Cornell University Library, arXiv preprint arXiv:1410.2507.
[9] Markovich, L. A. (2016): "Gamma-weibull kernel estimation of the heavy tailed densities", Cornell University Library, arXiv preprint arXiv:1604.06522v1.
[10] Parzen, E. (1962): "On Estimation of a Probability Density Function and Mode", *The Annals of Mathematical Statistics*, **33**, 1065-1076.
[11] Rosenblatt, M. (1956): "Remarks on Some Nonparametric Estimates of Density Function", *the Annals of Mathematical Statistics*, **27**, 832-837.
[12] Scaillet, O. (2004): "Density Estimation Using Inverse and Reciprocal Inverse Gaussian Kernels", *Journal of Nonparametric Statistics*, **16**, 217-226.