IUG Journal of Natural Studies
Peer-reviewed Journal of Islamic University-Gaza

ISSN 2409-4587

IUGNES
Vol 25, No 3, 2017, pp 33-38

Received on (31-10-2016) Accepted on (26-03-2017)

On Improvement of Hazard Rate Function Estimation Using Adaptive Kernel Estimates

Raid B. Salha[1,*]
Hazem I. El Shekh Ahmed [2]
Iyad M. Alhoubi

[1]Department of Mathematics, Faculty of Science, Islamic University of Gaza, Palestine
[2]Department of Mathematics, Al Quds Open University Palestine
[3]Department of Mathematics, University College of Science and Technology, Palestine

*** Corresponding author**
e-mail address: rbsalha@iugaza.edu.ps

Abstract
In this paper, we use the adaptive kernel estimates method to improve nonparametically the estimator of the probability density function (pdf) using the Erlang kernel (Erlang estimator). In addition, the cumulative distribution (cdf) of the improved Erlang estimator and the related hazard rate function for independent and identically distributed (iid) data will be evaluated. The performance of improved Erlang estimator and the related hazard rate function are tested using a simulation study.

Keywords:
Adaptive kernel estimates, Erlang kernel, hazard rate function.

1. Introduction:

Suppose we have a set of observed data assumed to be a sample from unknown pdf, we can construct a density estimation of the pdf from this sample. There are several nonparametric methods to do this, such as histograms which is widely use. When constructing a histogram, two choices must be made, the bandwidth and the position of the bin edge, if we vary these two choices, the aspect of the histogram will vary too, then we will get different features of the histogram. To avoid these disadvantages, we may use naive method which depends on constructing a box with a fixed bandwidth at every point of the observed data.
However, naive method gives more stability aspect than histogram, it suffers from discontinuity and has jumps at points. To overcome these drawbacks, kernel method is available and regards as a commonly used estimator to estimate pdf from an observed data, a non-negative support pdf and a fixed bandwidth may be used to construct the estimator, this estimator is continuous and differentiable. But this estimator suffers from a slight drawback if the data was collected from a long-tail distribution.
If we estimated the pdf, then we can think of estimation of the hazard rate function which have been considered in the literature. Hazard rate function estimation by nonparametric methods has an advantage in flexibility because no formal assumptions are made about the mechanism that generates the sample order than the randomness.
Estimators of the hazard function based on kernel smoothing have been studied extensively. For instance, see [7], [9], [8], [4], [11] and [6].
In the preceding methods, the bandwidth plays an important role in estimation accuracy and

smoothing, if tends to zero or becomes large, all detail of the curve of the estimator spurious or otherwise obscured. So, and in order to deal with this difficulty, various adaptive methods have been proposed, for instance, the nearest neighbor and the adaptive kernel methods.

In [1] Alain Berlinet et al are interested in the data-based selection of a variable band-width within an appropriate parameterized class of functions. They presented an automatic selection procedure inspired by the combinatorial tools developed in Devroye and Lugosi (2001) see [2]. In addition they showed that the expected error of the corresponding selected estimate is up to a given constant multiple of the best possible error plus an additive term which tends to zero under mild assumptions.

In [5] Van Kerm describes the Stata module adaptive kernel density (akdensity). akdensity extends the official Stata command kdensity that estimates density functions by the kernel method. He cleared that the extensions are of two types. Firstly, akdensity allows the use of an adaptive kernel approach with varying, rather than fixed, bandwidths. Secondly, akdensity estimates point wise variability bands around the estimated density functions. In [3] Salgado-Ugarte et al explored the use of one implementation of a variable kernel estimator in conjunction with several rules and procedures for bandwidth selection applied to several real data sets.

In [7] Salha R. et al used kernel method in their study, a new kernel called Erlang kernel was proposed as follows ،

$$K_E(x,\frac{1}{h})(t)=\frac{1}{\Gamma(1+\frac{1}{h})}\left[\frac{1}{x}(1+\frac{1}{h})\right]^{\frac{h+1}{h}} t^{\frac{1}{h}} exp(-\frac{t}{x}(1+\frac{1}{h})),\ h>0,\ t\geq 0 \qquad (1)$$

and the estimator of the pdf which called Erlang estimator was in the form ،

$$\hat{f}(x)=\frac{1}{n}\sum_{i=1}^{n}K_E(x,\frac{1}{h})(X_i) \qquad (2)$$

In addition, the hazard rate function using Erlang estimator was evaluated and the performance was tested for the estimators. In this paper, we use the adaptive kernel estimates method to improve Erlang estimator and the related hazard rate function.

This paper is organized in five sections. In the second section, we state some information about that adaptive kernel method which will be used in our study. In the third section, we improve the Erlang kernel estimator, then its performance will be tested via a simulated data. In the fourth section, we evaluate the cdf of the improved estimator, then evaluate the improved hazard rate function. Also, the performance of the improved hazard rate function will be tested via the simulated data. In the fifth section, we introduce comments and conclusion.

2. The adaptive kernel method:

The main idea of this method is to construct a kernel estimator using a specific kernel and make the bandwidth variable from a point of the observed data to another. This procedure is based on the common sense notion that a natural way to deal with long tailed densities is to use a broader kernel in regions of low density, this able the observation in the tail to take appropriate mass over a wider range than those observations in the main part of the distribution.

In [10] a clear strategy is stated to apply this method consists of three steps as follows:

1- Find a pilot estimate $\tilde{f}(t)$ that satisfies $\tilde{f}(X_i)>0$ for all i where $\{X_i\}_{i=1}^{n}$ is a family of iid.

2- Define the local bandwidth factor λ_i by

$$\lambda_i=\left[\frac{\tilde{f}(X_i)}{g}\right]^{-\alpha},\quad 0\leq\alpha\leq 1$$

where $g=\left(\prod_{i=1}^{n}\tilde{f}(X_i)\right)^{\frac{1}{n}}$ the geometric mean of $\tilde{f}(X_i)$.

3- Define the adaptive kernel estimator by

$$\hat{f}(x)=\frac{1}{n}\sum_{i=1}^{n}K(x,\frac{1}{h_i})(X_i)$$

where $h_i=h*\lambda_i$, K is the kernel function and h is the bandwidth.

More detail of this method is available in [10].

3. The improvement of density estimation:

In this section, we will improve 1 -where 2 is considered- the Erlang kernel estimator denoted (EKE) in [7] using the strategy in section 2. In addition, the performance of the improved Erlang kernel estimator (IMEKE) using a simulation study will be tested, a sample is taken from the exponential distribution of size 400 for this aim. Finally, the mean squared error (MSE) will be computed.

Applying the strategy of improvement appears in the following steps.

1- The pilot estimate which we choose is the Gaussian kernel estimator, where the kernel is

$$K(t) = \frac{1}{\sqrt{2\pi}} exp(-0.5t^2)$$

2- In our study α was chosen to equal 0.5, so the local bandwidth factor λ_i which makes the bandwidth vary at every observation point is defined by

$$\lambda_i = \left[\frac{\hat{f}(X_i)}{g} \right]^{-0.5}$$

3- Now, the improved estimator IMKEK is defined to be as follows:

$$\hat{f}(x) = \frac{1}{n} \sum_{i=1}^{n} K_E(x, \frac{1}{h_i})(X_i)$$

where

$$h_i = h * \lambda_i$$

K_E is the Erlang kernel which is our choice in this paper and h is a bandwidth which will be evaluated by the relation

$$h_{opt} = 0.79 R n^{-\frac{1}{5}}, \qquad (3)$$

where R is the interquartile range. see [10] page 47.

A sample of 400 observation from the exponential distribution $f(x) = exp(-x)$ is generated and graphed using R program, the graph is given in Figure 1. This figure shows the curves of EKE and IMEKE together with the true density.

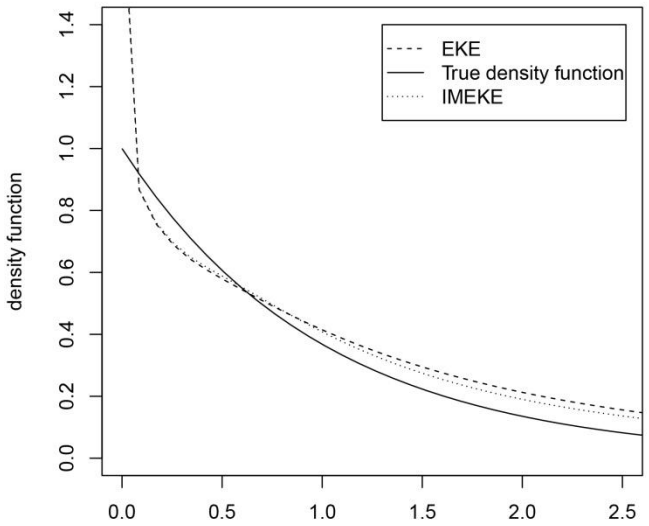

Figure 1: The EKE and IMEKE estimators of the exponential density underlying the generated sample.

The figure shows that the curve of IMEKE is closer to the true density than the curve of EKE, which indicates for the success of the improvement process.

Moreover, we calculated MSE for different numbers of points to test the performance of old and new estimators, and show the effect of the number of points in MSE. The results are shown in Table1.

Table 1: *MSE for different numbers of points*

Number of points	EKE	IMEKE	Ratio of decreasing
100	0.009147	0.008785	3.96%
150	0.007237	0.006946	4.02%
200	0.005909	0.005522	6.54%
300	0.004659	0.004174	10.4%

The table shows decreasing in MSE in all choices of numbers of points which mean that IMEKE more accuracy than EKE. Moreover, we calculated in the fourth row of the table the ratio of decreasing in MSE using the equation:

(((MSE of EKE) - (MSE of IMEKE))/(MSE of EKE))*100%

which gives the ratio of decreasing in MSE of EKE as a result of improving process. Table 1 shows that the bigger number of observation points, the bigger ratio in reduction MSE.
These ratios in addition with the results in the middle two rows in Table 1 ensure the effect of the sample size used in estimating process.

4. The improvement of the hazard rate function:

In this section, we will evaluate the cdf of IMEKE then evaluate the improved hazard rate function which will be denoted through this paper by (IMHEKE). The hazard rate function related to EKE will be denoted through this paper by (HEKE). In addition, the performance of IMHEKE using the same sample in Section 3 will also be tested for values close to zero which we concern in, and MSE will be computed for different numbers of observation points closed to zero.

Definition 1: The cdf of IMEKE is defined by:

$$\hat{F}(x)=\int_0^x \hat{f}(u)du=\frac{1}{n}\sum_{i=1}^{n}\int_0^x K_E(u,\frac{1}{h_i})(X_i)du$$

Proposition 1: The cdf of IMEKE is given by:

$$\hat{F}(x)=\frac{1}{n}\sum_{i=1}^{n}[(1-h_i)(1-F(\alpha_i)]$$

where, $F(\alpha_i)$ is the value of the Gamma cdf at $\alpha_i=\frac{t}{x}(1+\frac{1}{h_i})$.

Proof:

$$\int_0^x K_E(u,\frac{1}{h_i})(t)du=\int_0^x \frac{1}{\Gamma(1+\frac{1}{h_i})}\left[\frac{1}{u}(1+\frac{1}{h_i})\right]^{\frac{h_i+1}{h_i}} t^{\frac{1}{h_i}} exp(-\frac{t}{u}(1+\frac{1}{h_i}))du$$

$$=\frac{\left[(1+\frac{1}{h_i})\right]^{\frac{h_i+1}{h_i}} t^{\frac{1}{h_i}}}{\Gamma(1+\frac{1}{h_i})}\int_0^x u^{-\frac{h_i+1}{h_i}} exp(-\frac{t}{u}(1+\frac{1}{h_i}))du,\quad let\ w=\frac{t}{u}(1+\frac{1}{h_i})$$

$$=\frac{\left[(1+\frac{1}{h_i})\right]^{\frac{h_i+1}{h_i}} t^{\frac{1}{h_i}}}{\Gamma(1+\frac{1}{h_i})}\int_{\infty}^{\frac{t}{x}(1+\frac{1}{h_i})}\left[\frac{t}{w}(1+\frac{1}{h_i})\right]^{-\frac{h_i+1}{h_i}} exp(-w).\frac{-h_i}{t(1+h_i)}.\frac{t^2}{w^2}(1+\frac{1}{h_i})^2dw$$

$$=\frac{\left[(1+\frac{1}{h_i})\right]^{\frac{h_i+1}{h_i}} t^{\frac{1}{h_i}}}{\Gamma(1+\frac{1}{h_i})}.\frac{-h_i t^2(1+\frac{1}{h_i})^2 t^{-\frac{h_i+1}{h_i}}(1+\frac{1}{h_i})^{-\frac{h_i+1}{h_i}}}{t(1+h_i)}\int_{-\infty}^{\alpha_i} w^{1+\frac{1}{h_i}} w^{-2} exp(-w)dw,$$

$$\alpha_i=\frac{t}{x}(1+\frac{1}{h_i})$$

$$=\frac{1+h_i}{\Gamma(\frac{1}{h_i})}\int_{\alpha_i}^{\infty} w^{\frac{1}{h_i}-1} exp(-w)dw$$

$$=\frac{1+h_i}{\Gamma(\frac{1}{h_i})}\left[\int_0^{\infty} w^{\frac{1}{h_i}-1} exp(-w)dw-\Gamma(\frac{1}{h_i})\int_0^{\alpha_i}\frac{w^{\frac{1}{h_i}-1}}{\Gamma(\frac{1}{h_i})}exp(-w)dw\right]$$

$$=\frac{1+h_i}{\Gamma(\frac{1}{h_i})}\left[\Gamma(\frac{1}{h_i})-\Gamma(\frac{1}{h_i})F(\alpha_i)\right],$$

$$=(1+h_i)(1-F(\alpha_i))$$

where $F(\alpha_i)$ is the value of the Gamma cdf at α_i

Now, let the kernel estimator of the survivor function $S(\cdot)$ be $\hat{S}(x)=1-\hat{F}(x)$, then we define IMHEKE as follows:

$$\hat{r}(x)=\frac{\hat{f}(x)}{\hat{S}(x)}$$

Figure 2 shows the performance of HEKE and IHMEKE together with the true hazard rate function.

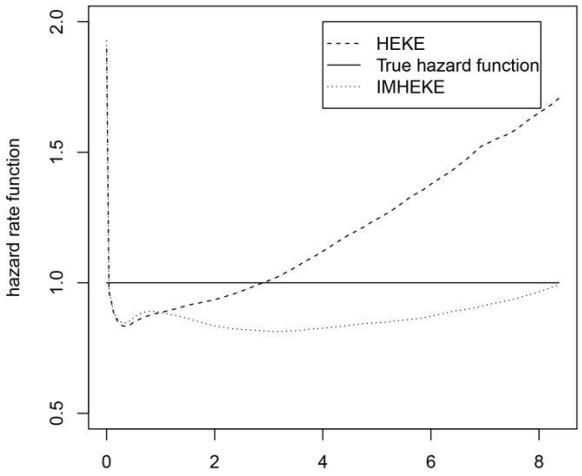

Figure 2: *The HEKE and IMHEKE estimators of the exponential density underlying the generated sample.*

It is clear that the curve of IMHEKE near zero is closer to the true hazard rate function than the curve of HEKE, which indicates that IMHEKE gives more accurate results than HEKE.

Table 2: *MSE for two proportion of points near zero*

Percentage of points	HEKE	IMHEKE	Ratio of decreasing
15%	0.033189	0.032114	3.24%
10%	0.043186	0.041263	4.45%

In addition, MSE for the boundary points near zero is evaluated using 15% and 10% of the observation points that closed to zero. Table 2 shows the measures of MSE at the two proportions. It is obvious that when we become close to zero we get less MSE for IMHEKE. In the fourth row of the table we calculated the ratio of decreasing in MSE using the equation:

(((MSE of HEKE) - (MSE of IMHEKE))/(MSE of HEKE))*100%

which gives the ratio of decreasing in MSE of HEKE as a result of improving process. The ratios ensure the improvement of the performance of the estimator HEKE.

5. Comments and Conclusion:

In this paper, we have improved EKE using the adaptive kernel estimates method. The new estimator IMEKE shows a satisfactory improvement of performance.

This appear clearly in sketching the curves of estimators and the true density, and the evaluation of MSE. Elaborated further, adaptive kernel method produced good improvement in the performance of EKE, graphically it makes the curve of IMEKE closer to true density than EKE. Further, MSE when we use IMEKE is less than it if EKE was used. In addition, the simulation study shows and ensures the effect of the number of the observation points used in estimating process, we notice that the larger number of points, the smaller MSE.

Moreover, the improvement of HEKE was so obvious when the curves of HEKE, IMHEKE with true hazard rate function have been drawn. The new hazard rate function IMHEKE shows acceptable improvement of performance for the values close to zero which we are concern in.

The effect of the improvement process also appeared when MSE was evaluated for two proportions of the observation points near zero, it was clear that the small proportion close to zero gave the lowest MSE.

From the above study we conclude that the adaptive kernel estimates method gives a good improvement in the performance of the estimators.

References:

[1] Alain Berlinet, Gerard Biau and Laurent Rouvi'ere, Optimal L_1 Bandwidth Selection for Variable Kernel Density Estimates. Institut de Mathematiques et de Modelisation de Montpellier, UMR CNRS 5149, Equipe de Probabilites et Statistique, Universite

Montpellier II, Cc 051, Place Eug'ene Bataillon, 34095 Montpellier Cedex 5, France.

[2] Devorye and Lugosi, Combinatorial Methods in Density Estimation, springer-verlag, New York, Berlin, Heidelberg, (2001).

[3] Isaas H. Salgado-Ugarte and Marco A. Perez-Hern, Exploring the use of variable bandwidth kernel density estimators, The Stata Journal (2003) 3, Number 2, pp. 133147.

[4] O. Scaillet (2004). Density estimation using inverse and reciprocal inverse Gaussian kernels. Nonparametric Statistics, 16, 217-226.

[5] Philippe Van Kerm, Adaptive kernel density estimation, Philippe Van Kerm CEPS/INSTEAD, G.-D Luxembourg. The Stata Journal (2003) 3, Number 2, pp. 148156

The Stata Journal (2003) 3, Number 2, pp. 148156

[6] Rice, J. and Rosenblatt, M. (1976). Estimation of the log survivor function and hazard function. Sankhya Series A 38, 60-78.

[7] Salha R., El Shekh Ahmed H. and Alhoubi I., Hazard Rate Function Estimation Using Erlang Kernel. International Publishers of Science, Technolory and Medicine. HIKARI LTD, P.O. Box 85, Ruse 7000, Bulgaria, Vol. 3, 2014.

[8] Salha, R. (2013). Estimating the density and hazard rate functions using the reciprocal inverse Gaussian kernel. 15th Applied Stochastic Models and Data Analysis International Conference, Mataro (Barcelona) Spain, June 25-28, 2013.

[9] Salha, R. (2012). Hazard rate function estimation using inverse Gaussian kernel. The Islamic university of Gaza journal of Natural and Engineering studies, 20(1), 73-84.

[10] Silverman, B. W. (1986). Density estimation for statisticas and data analysis. Chapman and Hall, London.

[11] Watson, G. and Leadbetter, M. (1964). Hazard analysis I, Biometrika, 51, 175-184.

تحسين أداء مقدر دالة المخاطرة بطريقة تكييف النواة

اقترح الباحث في بحث سابق توزيع احتمالي كنواة لتقدير دالة كثافة احتمالية مجهولة وهو توزيع إيرلنج وتم تركيب مقدر جديد هو مقدر إيرلنج وتم حساب دالة المخاطرة للمقدر. تم اختبار أداء المقدر لدالة الكثافة المجهولة ولدالة المخاطرة وأثبت الاختبار أن المقدر ودالة المخاطرة يعملان بشكل جيد حيث كان الخطأ في التقدير صغيراً جداً. في هذا البحث تم استخدام طريقة تكييف النواة لتحسين أداء المقدر الجديد ودالة المخاطرة المعرفة منه وأنتجت العملية تحسيناً ملحوظاً على أداء المقدر ودالة المخاطرة وذلك عبر تطبيق محاكاة لعينة حجمها 400 تم توليدها من توزيع أسي باستخدام برنامج R.

كلمات مفتاحية:
تكييف مقدر النواة،
نواة إيرلنج،
دالة المخاطرة،