Sumer Journal for Pure Science (ISSN: 2790-7031)
2nd International Scientific Conference on Pure and Medical Sciences/University of Sumer 2024

# Comparison of Kernel Functions to Estimate Nonparametric Confidence Limits with Application

**Najlaa A. Al-Khairullah[1], Nada Hussein Tali[2,*], and Aseel Muslim Eesa[2]**

[1]*Department of Mathematics, Education College, University of Sumer, Thi Qar, Iraq*
[2] *Department of Statistic, College of Administration and Economics, University of Sumer, Thi Qar, Iraq*

* Email address of the corresponding author: nada.hussein@uos.edu.iq

**Abstract**

Nonparametric confidence bound estimation is a statistical technique used to estimate the probability density function, which works to smooth each point in the data of the variable to be studied. Nonparametric confidence intervals define an interval containing the core function based on the sample data, which is defined by an upper and lower bound. In this research, a comparison was made between the nonparametric kernel functions in the case of estimation with nonparametric confidence intervals using the plug-in approach method. It was noted that all the functions gave good results through the graph in the case of using real data, and the best functions were the Epanechnikov and the Tricube functions for estimating the kernel function with nonparametric confidence intervals, where the confidence intervals were narrow in the graph.

**Keywords**:The kernel Function, nonparametric confidence limits, bandwidth parameter, plug-in Approch method.

# مقارنة بين دوال اللب لتقدير حدود الثقة اللامعلمية مع التطبيق

**نجلاء علي هداب [1], ندى حسين تالي[2،*], اسيل مسلم عيسى[2]**

[1]قسم الرياضيات, كلية التربية ,جامعة سومر , ذي قار , العراق
[2]قسم الاحصاء , كلية الادارة والاقتصاد, جامعة سومر , ذي قار , العراق

**الخلاصة**

التقدير بحدود الثقة اللامعلمية هي تقنية احصائية تستخدم لتقدير دالة الكثافة الاحتمالية والتي تعمل على تنعيم كل نقطة من بيانات المتغير المراد دراسته. تقوم فترات الثقة اللامعلمية بتحديد فترة تحتوي على دالة اللب المعتمدة على بيانات العينة والتي تحدد بحدين اعلى وادنى . في هذا البحث تم المقارنة بين دوال اللب اللامعلمية في حالة التقدير بفترات الثقة اللامعلمية باستخدام طريقة Plug in Approch ولوحظ أن جميع الدوال قد اعطت نتائج جيده من خلال الرسم البياني في حالة استخدام بيانات حقيقية وكانت افضل الدوال هي دالة Epanechnikov و دالة Tricube لتقدير دالة اللب بفترات الثقة اللامعلمية حيث كانت فترات الثقة ضيقة في الرسم البياني.

**الكلمات المفتاحية** : دالة اللب ، حدود الثقة اللامعلمية , معلمة عرض الحزمة , طريقة Plug-in Approch .

## 1. Introduction

Many studies and research have examined the estimation of community parameter that are usually unknown and through samples are estimated using several statistical methods. It is known that estimation has three methods: parametric methods, nonparametric methods, and semi parametric methods. There are two ways to estimate parametric: point and confidence interval estimation. Confidence intervals can be defined as a range determined by a set of values based on sample data, which determine upper and lower limits. Confidence intervals are influenced by the sample size; the larger the sample size, the closer it will be to the confidence limits because it works to reduce the standard deviation, indicating the efficiency of the estimator [1]. The confidence limits have two types: parametric confidence limits and nonparametric confidence limits. The parametric confidence limits are defined as the recognition of the community's marker from the sample data by setting a period with a set of points. An estimate of the nonparametric confidence limits is more difficult because the estimate of the nonparametric function is biased, so there will be a problem in measuring the bias of the estimation of the function directly [2].

## 2. Research Objective

The aim of this paper is to use the most important and widespread core functions to estimate the nonparametric confidence limits through the use of the core method, which is one of the methods of preparation, and thus to compare these functions and indicate their best.

## 3. The kernel Function

The kernel function of nonparametric estimation has two types: the first is called Kernel Optimal, which operates on reducing the Mean Integrated Square Errors (MISE). That is where these functions are derived from the MISE for a kernel function [3]. The second type is the variance minimum kernel, which works to reduce the corresponding variation. The kernel function has several names, including weight function, window function, shape function, or core function [4]. The kernel function is defined as a real, similar, limited and continuous function, and its integrality is equal to the one. We can find the kernel functions with the lowest variation; assume that$\{x_1, x_2,..., x_n\}$ are independent and single-distributed variables and $n$ represents sample observations of community $x$ [5].

$$\hat{f}_K(u) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \quad (1)$$

$K$ indicates the kernel function, and it achieves the following characteristics:

$$\int \mathrm{K(u)du} = 1$$

$$\int \mathrm{uK(u)du} = 0$$

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

And h refers to the bandwidth parameter, which represents a function of the size of the sample and has a significant impact on the bias and variation by increasing the preparatory milestone, increasing the bias, reducing the variation, and vice versa, thus affecting the degree of preparatoryization of the estimate curve [1]. We can find the preface parameter mentioned by Silverman by using MISE, which is the most accurate measure [6]. It can be expressed as follows [7]:

$$MISE\left(\hat{\hat{f}}(x)\right) = \int E\left\{\hat{\hat{f}}(x) - f(x)\right\} 2dx \qquad (2)$$

The bandwidth parameter is obtained through the following formula [8]:

$$h = k_2^{2/5}\left\{\int k^2(u)du\right\}^{1/5}\left\{\int (f^2(x))^2\,dx\right\}^{-1/5} n^{-1/5} \qquad (3)$$

## 4. Selection of Kernel function

Most scientific studies suggest that the choice of kernel functions is less important than the choice of a bandwidth parameter for the performance of densities, where a few kernel functions are used. There is a set of kernel functions that belong to the Beta family, which are called kernels univariate non-normal, and one of the most famous is (triweight, biweight, Epanechnikov, and uniform) [8]. We assume that $K(x)$ represents a function of real value used to determine the local weights of the linear estimate, which refers to a function of real value used to determine the local weights of the estimate by fulfilling the requirement $\int K(x)\,dx = 1$, and *h* refers a bandwidth parameter, so the general formula is [9]:

$$K(x,\alpha) = \frac{1}{B(0.5,\alpha+1)}(1-u^2)^{\alpha} \quad I(|x| \le 1) \qquad (4)$$

Where $B(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ and $\Gamma(a) = (a-1)!$ , such that $\alpha = 0, 1, 2, 3.$

If $\alpha$ = 0, We will gain an uniform function, and if $\alpha$ = 1, 2, 3, the Beta function becomes are Epanechnikov, biweight, and triweight respectively. When $\alpha$ is big, the function of a beta will almost be close to a function Gaussian Kernel [2].

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \qquad (5)$$

The most important kernel functions [1, 10] used in this paper are shown in Table 1.

**Table 1-** Kernel functions are used

| Kernel Shape | K(u) | $\int u^2 K(u)du$ | $\int k^2(u)du$ |
|---|---|---|---|
| Uniform | $\frac{1}{2} I(\lvert u\rvert \le 1)$ | $\frac{1}{3}$ | $\frac{1}{2}$ |
| Epanechnikov | $\frac{3}{4}(1-u^2) I(\lvert u\rvert \le 1)$ | $\frac{1}{5}$ | $\frac{3}{5}$ |
| Biweight | $\frac{15}{16}(1-u^2)^2 I(\lvert u\rvert \le 1)$ | $\frac{1}{7}$ | $\frac{5}{7}$ |
| Triweight | $\frac{35}{32}(1-u^2)^3 I(\lvert u\rvert \le 1)$ | $\frac{1}{9}$ | $\frac{350}{429}$ |
| Triangular | $(1-\lvert u\rvert) I(\lvert u\rvert \le 1)$ | $\frac{1}{6}$ | $\frac{2}{3}$ |
| Tricube | $\frac{70}{81}(1-\lvert u\rvert^3)^3 I(\lvert u\rvert \le 1)$ | $\frac{35}{243}$ | $\frac{175}{247}$ |
| Gaussian | $\frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$ | 1 | $\frac{1}{\sqrt{2\pi}}$ |
| Logistic | $\frac{1}{e^u + e^{-u} + 2}$ | $\frac{\pi^2}{3}$ | $\frac{1}{6}$ |

## 5. Nonparametric Confidence Limits

This paper uses a plug-in approach to estimate nonparametric confidence limits by estimating the nonparametric density function [5]. The process of estimating the nonparametric confidence limits is using a plug-in approach [11], twitches one of the simplest methods of estimating the nonparametric confidence limits, where the corresponding variation is replaced by the sign level of 1-∝ and the confidence limits of variable x are:

$$p(f(x) \in C_{1-\propto}(x)) = 1 - \alpha \tag{6}$$

Where $C_{1-\propto}(x)$ represents the zone of confidence for the density function is random $C_{1-\propto}(x)$, it is obtained from sample data, and we have taken a kernel estimate [2]:

$$\hat{f}_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h_n}\right) \tag{7}$$

The nonparametric confidence limits are calculated according to the following formula:

$$p\left(\hat{f}_n(x) - Z_{\frac{\alpha}{2}} * \sigma_n(x) < f(x) < \hat{f}_n(x) + Z_{\frac{\alpha}{2}} * \sigma_n(x)\right) = 1 - \alpha \tag{8}$$

Where: $\sigma_n^2(x) \simeq \frac{\hat{f}_n(x) \int \mathrm{K(u)}^2 \ \mathrm{du}}{nh_n}$ , and $Z_{\frac{\alpha}{2}}$ represents table value by level of statistical significance.

## 6. Bandwidth Parameter

Assessment of the best bandwidth parameter is selected by relying on the average Mean Integrated Square Error (MISE) and as follows [12]:

$$\mathrm{MISE}\left[\hat{f}(x,h)\right] = E\int\left[\hat{f}(x,h) - f(x)\right]^2 dx \tag{9}$$

$$\mathrm{MISE}\left[\hat{f}(x,h)\right] = var\left(\hat{f}(x,h)\right) + baise^2\left(\hat{f}(x,h)\right) \tag{10}$$

We take the case of one univariate variable and after the derivative of the variation and bias of the kernel function[2]:

$$\mathrm{MISE}\left[\hat{f}(x,h)\right] = n^{-1}h^{-1}R(K) + \frac{1}{4}h^4\mu_2(K)^2R((f)^2) \tag{11}$$

Where:

$$R(K) = \int (K(u))^2 du$$

$$\mu_2(K)^2 = \int u^2K(u)du$$

$$R((f)^2) = \int [f''(u)]^2 du$$

Then we find a square of $\mathrm{MISE}\left[\hat{f}(x,h)\right]$, and the derivative of *h* produces:

$$\hat{\mathrm{h}}_{\mathrm{MISE}} = \left[\frac{R(K)}{\mu_2(K)^2R((f)^2)}\right]^{1/5} \tag{12}$$

The final equation is to estimate the bandwidth parameter for a single variable, either in the case of multiple variables, the preface parameter is written as follows:

$$\hat{\mathrm{h}}_{\mathrm{MISE}} = \left[\frac{R(K)}{\mu_2(K)^2\widehat{\Psi}_4(g)}\right]^{1/5} \tag{13}$$

## 7. Practical aspect

Chronic kidney disease (kidney failure) indicates irreversible deterioration in kidney function, which has typically evolved over the years. At first, it only appears to be a biochemical anomaly but ultimately leads to a loss of kidney function. Clinical symptoms and signs of renal failure are collectively referred to as polyina and often affect people over 65 years of age. Blood samples were collected from 73 patients with chronic kidney failure treated with continuous dialysis. Blood samples were drawn for a group of patients before dialysis was

performed, which takes three to four hours in cooperation with the Ibn Sina Educational Hospital—the Industrial College Unit, which is between the ages of 20 and 80, with 38 males and 35 females. The study included six illustrative variables that are thought to have an impact on the response variable, which represents the number of dialysis cycles per month. Table 2 provides a description of the illustrative variables used in the study.

**Table 2-** Description of the variables for patients with kidney failure

| Variable | Description of variable | Unit |
|---|---|---|
| $x_1$ | Age | Years |
| $x_2$ | Period of illness | Days |
| $x_3$ | Urea concentration | mmol/l |
| $x_4$ | Total protein concentration | g/100ml |
| $x_5$ | Albumin concentration | g/100ml |
| $x_6$ | Globulins concentration | g/100ml |

The theoretical part was applied by using the MATLAB package, where nine kernel functions were applied at a mental level of 0.05. We estimated the kernel function, given that the prep parameter was selected using a plug-in method, and gave the following results, as shown in Table 3.

**Table 3-** Shows the sum of the pulp function vector and the sum of the upper limit and the lower limit

| Kernel Shape | $\hat{f}_n(x)$ | upper bound | lower bound |
|---|---|---|---|
| Uniform | 0.0198 | 0.0228 | 0.0169 |
| Epanechnikov | 0.0198 | 0.0228 | 0.0169 |
| Quadratic | 0.0198 | 0.0229 | 0.0168 |
| Triweight | 0.0198 | 0.0229 | 0.0168 |
| Triangular | 0.0198 | 0.0226 | 0.0171 |
| Tricube | 0.0198 | 0.0227 | 0.0170 |
| Gaussian | 0.0198 | 0.230 | 0.0167 |
| Logistic | 0.0198 | 0.0220 | 0.0177 |

Table 3 shows the kernel function used to estimate the kernel density function. We note that the shapes of the nonparametric kernel function curves as well as the curve of periods of higher and lower confidence using real data show that all kernel functions have yielded good results and are close to the kernel function axis as shown in Figure 1. The confidence periods for the upper and lower limits were close to the probability density axis. The reason is that the size of the sample is large, so the size of the sample affects the axle capacity. If the sample size of the axle is large, the axle capacity is close, and if the sample size of the axle is small, the axle capacity increases.

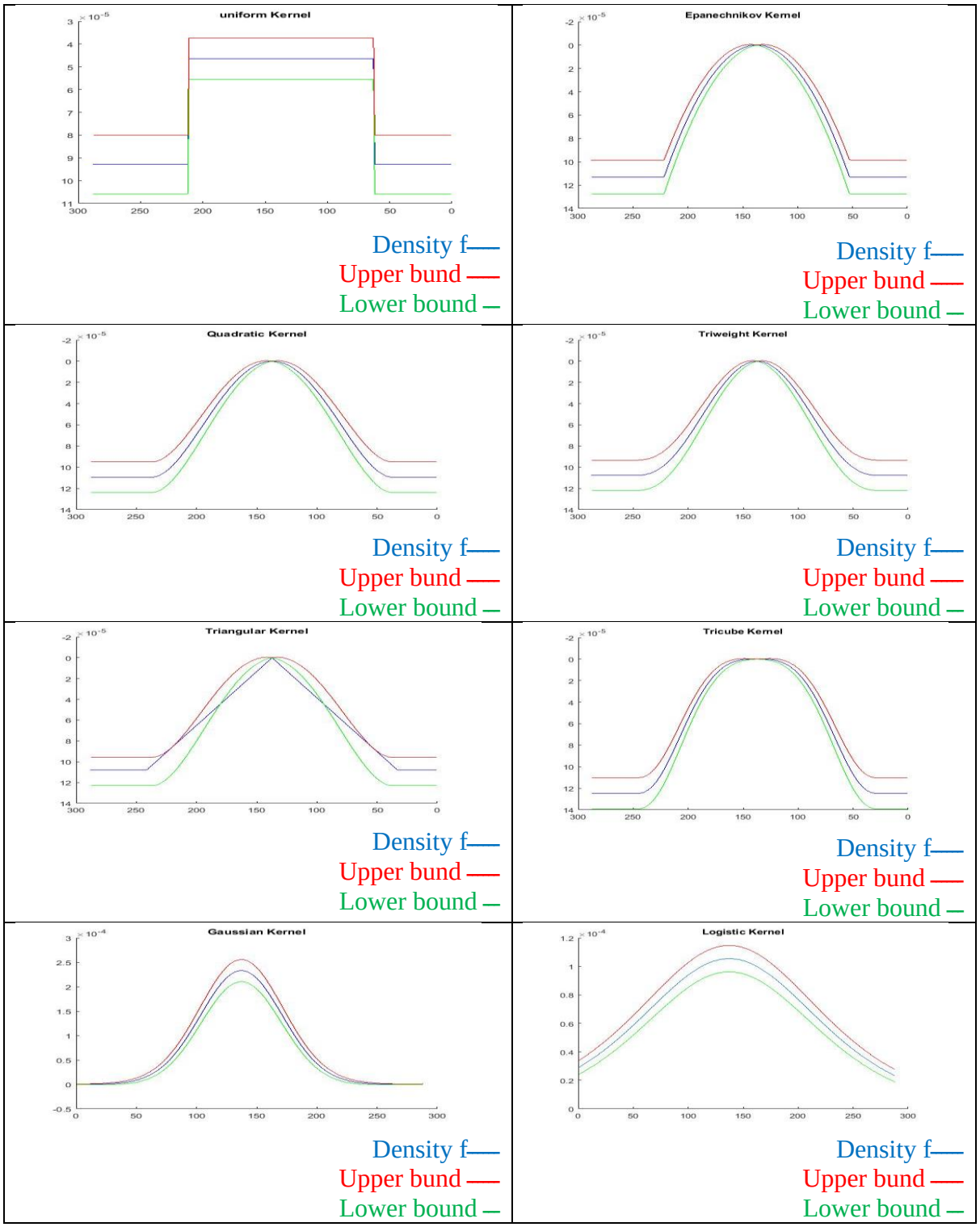


**Figure -1** Nonparametric kernel functions curves

## 8. Conclusions

From the practical aspect, the results can be summarized as follows:

I. All the kernel functions have yielded satisfactory results in estimating the nonparametric confidence limits.

II. The functions were compared with the knowledge of the best function in the figure by narrowing the confidence intervals. It was observed that when real data were used, the best function was the Epanechnikov function and then the Tricube function. This underscores the importance and preference of using nonparametric discretion.

## References:

[1] H. Munaf, "Comparison of kernel's nonparametric estimation of regression," *School of Administration and Economics, Baghdad University*, 2004.

[2] E. Parzen, "On estimation of a probability density function and mode," *The annals of mathematical statistics*, vol. 33, no. 3, pp. 1065-1076, 1962.

[3] N. S. S. Meti, "Assessment of the nonparametric regression role using some of the methods of preparation," *Iraqi Journal of Statistical Sciences (number of proceedings of the Fourth Scientific Conference of the Faculty of Computer Sciences and Mathematics)*, vol. 20, pp. 371-390, 2011.

[4] B. A. e. M. F.W . Mitras, "Using the kernel estimator and the k-rate clustering method to recognize the hand gesture," *Al-Rafidain Journal of Computer Science and Mathematics*, vol. 10, no. 1, 2013.

[5] D. Aydin, "A comparison of the nonparametric regression models using smoothing spline and kernel regression," *World Academy of Science, Engineering and Technology*, vol. 36, pp. 253-257, 2007.

[6] B. W. Silverman, *Density estimation for statistics and data analysis*. Routledge, 2018.

[7] J. Fan and R. Li, "Variable selection via nonconcave penalized likelihood and its oracle properties," *Journal of the American statistical Association*, vol. 96, no. 456, pp. 1348-1360, 2001.

[8] J. Fan and I. Gijbels, "Variable bandwidth and local linear regression smoothers," *The Annals of Statistics*, pp. 2008-2036, 1992.

[9] C. V. Fiorio, "Confidence intervals for kernel density estimation," *The Stata Journal*, vol. 4, no. 2, pp. 168-179, 2004.

[10] M. A. Delgado, "Applied Nonparametric RegressionW. Härdle Cambridge University Press, 1990," *Econometric Theory*, vol. 8, no. 3, pp. 413-419, 1992.

[11] J. D. Hart and T. E. Wehrly, "Kernel regression estimation using repeated measurements data," *Journal of the American Statistical Association*, vol. 81, no. 396, pp. 1080-1088, 1986.

[12] C. R. Loader, "Bandwidth selection: classical or plug-in?," *The Annals of Statistics*, vol. 27, no. 2, pp. 415-438, 1999.