Iraqi Journal of Statistical Science (17) 2010
The Second Scientific Conference of Mathematics-Statistics& Information
pp [81 -90]

Comparing nonparametric density estimators using different canonical kernel functions

Munaf Yousif H.* Dhafir H.Rashed **

Abstract

In this article we make a comparison between three different nonparametric density estimators .The first estimator is called fixed kernel which uses a fixed bandwidth.

The second estimator is called variable kernel, which uses variable bandwidth, then the third estimator, refers to a new version of nonparametric density estimator (this estimator is a hybrid of the above estimators.

This comparison depends on using different kernels, two of these functions are proposed in this article .To make this comparison suitable, we use standardization techniques, which are called canonical kernel.

Simulation is used to find kernel function based on the standardization techniques (canonical kernel).

Keywords: Fixed kernel, Variable kernel, Kernel functions, Plug-in, Canonical kernel, Canonical bandwidth.

الملخص

نقوم في هذه المقالة باجراء مقارنة بين ثلاثة مقدرات لا معلمية لدالة كثافة الاحتمال: المقـدر الاول يسمى (اللب المثبت) والذي يستخدم سعة قيد مثبتة. والمقدر الثاني يسمى اللب المتغيـر والذي يستخدم سعة قيد متغيرةز اما المقدر الثالث نسخة جديدة من المقدرات اللامعلمية ويكون مهجن من المقدرين السابقين.

تعتمد هذه المقارنة على استخدام دوال لبية مختلفة، اثنتان منها تقترح في هذه المقال. ولجعـل هذه المقارنة مناسبة استخدمت تقنية المعيارية وسميت المقدرات اللبية القانونية.

استخدمت طريقة المحاكاة لايجاد دالة لبية ترتكز على تقنيات معيارية (مقدرات لبية قانونية).

* **Assistant professor of Statistics, Department of Statistics, Baghdad University,(munaf_yousif@yahoo.com).**
** **Professor of Statistics, Head of the department of Statistics, Baghdad University.**
Received:1/7 /2009 ____________________Accepted: 6 /12 / 2009

1. Introduction

The main property that describes the behavior of the random variables X can be seen from the probability density function. From the density function can be known the number of observations that lie in some groups.

The probability can be calculated as an integral on that group

$$\Pr(a < x < b) = \int_a^b f(x)dx \qquad \textbf{… (1)}$$

This value x lies in the interval that mentioned above, this will lead to prediction that the observations will appear frequently; also, through using the graphical plots for probability density function. We will know that the observation may have long tails; also we can see that the observations have multimodality.

Estimating the unknown probability density function will provide the researcher are understanding for the behavior of the random variables, but in most studies the distribution is unknown and instead of that will depend on the given observation that regarded as identically and independent random variables which have unknown probability density.

The goal is to estimate the probability density function on the basis of these observations. The methods are used in this article is of a nonparametric kind that does not have any restrictions on the form of the probability density function. These methods are estimating the density completely based on observations.

Also in this article we want to compare different kernels, where two kernels are proposed here. This comparison relies on three nonparametric density estimators, one of these is a new version of estimators.

2. *Kernel density estimators*

The proposed kernel estimator is:

$$\hat{f}(x) = (1-m)\hat{f}_V(x) + m\hat{f}_F(x) \qquad \textbf{… (2)}$$

Where $\hat{f}_V(x)$ refers to variable kernel estimator that has the form

$$\hat{f}_V(x) = n^{-1}\sum_{i=1}^{n}\frac{1}{h(x_i)}K\left(\frac{x-X_i}{h(x_i)}\right) \qquad \textbf{… (3)}$$

This estimator relies on using variable bandwidth at each point through using big window at low density and using small window at high density .The idea of this estimator come back to Breiman, Meisel and Purcell (1977).

The other nonparametric density estimator is $\hat{f}_F(x)$, which refers to fixed kernel estimator that has the form:

$$\hat{f}_F(x) = (nh)^{-1} \sum_{i=1}^{n} K(\frac{x - X_i}{h}) \qquad \textbf{... (4)}$$

This estimator is called fixed because its rely on using fixed bandwidth on the real line of the density function. This estimator is proposed from Rosenblatt (1956) and Parzen (1962).

We must note that the proposed estimator consists of two unknowns (K, h), that is kernel function and bandwidth.

Therefore we will display the estimator of h and the possible kernel functions, and then we compare the performance of the proposed nonparametric estimator with the other estimators depending on the different kernel functions and using the standardization or rescale techniques to find the best estimator and kernel function.

3. <u>Kernels</u>

Kernel functions have different names such as window function or weight function. These functions are real, symmetric, continuous and these integrals equal to one. That is:

1- $Sup|K(z)| < \infty -, \int_{-\infty}^{\infty} K(z)dz = 1.$

$$\int_{-\infty}^{\infty} |K(z)|dz < \infty -, Lim|z||K(z)| = 0 -, K(-z) = K(z) \ . \qquad \textbf{... (5)}$$

2- $\int_{-\infty}^{\infty} z^q K(z)dz = 0, q = 1,2,3,...r-1$

3- $\int_{-\infty}^{\infty} |z^q K(z)|dz < \infty, q = r$.

Be sure that these conditions are satisfied when r=2.The following table displays some kernels and their efficiencies:

Table (1)
Refers to some kernels with two proposed kernels and their efficiencies relative to Epanchnikov

Kernel	K(z)	C_K	d_k	$C_K^2 d_k$	$E(K_{opt}, K)$
Epanchnikov	$(3/4)(1-z^2)$	0.6	0.2	0.072	1
Quartic	$(15/16)(1-z^2)^2$	0.714	0.143	0.073	1.006
Triweight	$(35/32)(1-z^2)^3$	0.816	0.111	0.0739	1.0132
Gaussian	$(2\pi)^{-1/2}\exp(-z^2/2)$	$(2(\pi)^{1/2})^{-1}$	1	0.0796	1.0515
Proposed1	$(45/64)(1-z^4)^2$	0.611	0.195	0.0728	1.0055
Proposed2	$(72/89)/(1-z^4)^4$	0.692	0.151	0.0723	1.0021

The relative efficiency for the above kernels can be obtained from the following:

$$effi(K, K_{opt}) = \left[\frac{\left[\int K^2(z)dz\right]^2 \left[\int z^2 K(z)dz\right]}{\left[\int K_{opt}{}^2(z)dz\right]^2 \left[\int z^2 K_{opt}(z)dz\right]} \right]^{0.5}$$

$$\equiv \left[\frac{C_k^2 d_k}{C_{k_{opt}}^2 \; d_{k_{opt}}} \right]^{0.5} \qquad \textbf{… (6)}$$

4. Canonical Kernels

We must note that if one used a kernel of the form $K_\delta(.) = \frac{1}{\delta} K(./\delta)$, and rescale the bandwidth by the factor δ, one would obtain the estimate as with the original kernel estimator.

Kernel can therefore be seen as equivalence class of functions K with possible rescaling by δ.

A consequence of this scale dependence is that the bandwidth selection problem is not identifiable if the kernel is determined only up to scale.

Consider the situation in which two statisticians analyze the same data set but use different kernels for their estimators. Their bandwidths have been determined subjectively or automatically, but they have been computed for different kernels and therefore can't be compared directly.

In order to allow some comparison one needs a common scale for both bandwidths. To find such a scale, we must note that two kernel estimators with the same bandwidth should ascribe the same amount of smoothing to the data.

Such a scale is given by canonical kernels in the class of kernels K_δ .It is based on the expansion of the $AMISE(K_\delta)$

$$AMISE(K_\delta) = n^{-1}\frac{\|K\|_2^2}{\delta} + 0.25\|f''\|_2^2\delta^4 M_2^2(K) \quad \text{… (7)}$$

Where:

$$\|K_\delta\|_2^2 = \frac{1}{\delta}\|K_\delta\|$$

$$\int_{-\infty}^{\infty} z^2 K_\delta(z)dz = \delta^2 \int_{-\infty}^{\infty} z^2 K(z)dz \quad \text{… (8)}$$

To Scale the bias and the variance equally we choose δ^* such that:

$$\frac{1}{\delta^*}\|K\|_2^2 = (\delta^*)^4 M_2^2(K) = T(K) \quad \text{… (9)}$$

We call the parameter δ^* the canonical bandwidth or canonical smoothing parameter for kernel K_{δ^*} and δ^* can be achieved by defining

$$\delta^* = \left(\frac{\|K\|_2^2}{M_2^2(K)}\right)^{0.2} \quad \text{… (10)}$$

Also we should be noting that:

$$T(K) = \left\|\delta^{*-1}K\right\|_2^2 = \left\|\left(\frac{M_2^2(K)}{\|K\|_2^2}\right)^{0.2} K\right\|_2^2$$

$$= \left[M_2^2(K)\|K\|_2^8\right]^{0.2} \quad \text{… (11)}$$

If we take $h' = \delta^* h$ then:

$$AMISE(K,h') = \left\{nh^{-1} + 0.25\|f''\|_2^2 h^4\right\} * T(K) \quad \text{… (12)}$$

And the canonical bandwidth is:

$$h_j = h_i \frac{\delta_j^*}{\delta_i^*} \quad \textbf{… (13)}$$

From the above equation we obtain the relationship between the bandwidth and kernel (K_i, h_i) with the coupled (K_j, h_j) .After we substitute the relation (13) in the form (12), we obtain

$$AMISE(K_j,h_j) = AMISE\left(K_j, h_i \frac{\delta_j^*}{\delta_i^*}\right) = \left\{\left(n \frac{h_i}{\delta_i^*}\right)^{-1} + 0.25\left(\frac{h_i}{\delta_i^*}\right)^4 \| f''\|_2^2\right\} * T(K_j)$$

$$= \left\{(nh_i)^{-1} \delta_i^* T(K_i) + 0.25 h_i^4 \| f''\|_2^2 \delta_i^{*-4} T(K_i)\right\} \frac{T(K_j)}{T(K_i)}$$

$$= AMISE(K_i,h_i) * \frac{T(K_j)}{T(K_i)} \quad \textbf{… (14)}$$

5. <u>Simulation</u>

To know the performance of each of the kernels density estimators (Fixed, Variable and the proposed) with the kernel functions we do some simulation experiments.

In these experiments we use three different sizes (40,125,160), each with different distributions (Normal, Bimodal and contaminated distributions).

We repeat each experiment 100 once for each sample size and distribution. Also we use a plug-in estimator for estimating the bandwidth .The common distance measure that it used is mean averaged squared error (MASE)

$$MASE(\hat{f}(x)) = n^{-1} \sum_{i=1}^{n} \left[\hat{f}(x_i) - f(x_i)\right]^2$$

The following tables illustrate the values of $MASE(\hat{f}(x))$ for different distributions and sample sizes.

Table (2)
Refers to MASE for different nonparametric estimators, kernels and distributions when using Normal distribution

Distr.	Estimators (Kernel fn. / n)	Fixed Kernel 40	Fixed Kernel 125	Fixed Kernel 160	Variable Kernel 40	Variable Kernel 125	Variable Kernel 160	Suggested Kernel 40	Suggested Kernel 125	Suggested Kernel 160
N(-2,2)	Epanch.	0.0051	0.00174	0.00136	0.0039	0.00125	0.00111	0.00333	0.00112	0.00099
	Quartic	0.00523	0.00176	0.00138	0.0048	0.0015	0.00133	0.00388	0.0013	0.00113
	Triweight	0.00531	0.00178	0.0014	0.00554	0.00172	0.00151	0.00425	0.00143	0.00122
	Gaussian	0.00563	0.00186	0.00146	0.00377	0.00162	0.00126	0.00201	0.00077	0.00057
	Proposed1	0.005	0.00174	0.00136	0.004	0.00128	0.00114	0.00337	0.00114	0.00101
	Proposed2	0.00501	0.00174	0.00136	0.00458	0.00145	0.00129	0.00375	0.00127	0.0011
N(3,3)	Epanch.	0.003	0.00112	0.00093	0.00179	0.00071	0.00068	0.00164	0.00068	0.00065
	Quartic	0.00309	0.00114	0.00094	0.00228	0.00084	0.00079	0.00196	0.00079	0.00074
	Triweight	0.00316	0.00116	0.00095	0.00272	0.00096	0.00088	0.0022	0.00087	0.00081
	Gaussian	0.00339	0.00122	0.00098	0.00243	0.00098	0.00088	0.00108	0.00044	0.0004
	Proposed1	0.00293	0.00111	0.00093	0.00184	0.00074	0.00071	0.00166	0.00069	0.00066
	Proposed2	0.00293	0.00111	0.00093	0.00217	0.00083	0.00078	0.00189	0.00077	0.00073
N(0,2.5)	Epanch	0.00502	0.0013	0.00105	0.00339	0.00089	0.00079	0.00305	0.00082	0.00072
	Quartic	0.00516	0.00132	0.00106	0.0043	0.00103	0.0009	0.00368	0.00093	0.0008
	Triweight	0.00525	0.00134	0.00107	0.00506	0.00116	0.001	0.00413	0.00102	0.00087
	Gaussian	0.00558	0.00140	0.00112	0.0027	0.00146	0.00123	0.00143	0.0006	0.00052
	Proposed1	0.00488	0.00129	0.00105	0.00348	0.00092	0.00081	0.00308	0.00083	0.00073
	Proposed2	0.0049	0.00129	0.00104	0.00412	0.00102	0.00089	0.00351	0.00091	0.00078

Table (3)
Refers to MASE for different nonparametric estimators, kernels and distributions when using Bimodal and Contamination distributions

Distr.	Estimators	Fixed Kernel			Variable Kernel			Suggested Kernel		
	Kernel fn. / n	40	125	160	40	125	160	40	125	160
0.35N(-2,2)+0.65N(3,1)	Epanch.	0.00165	0.00071	0.00054	0.00142	0.00067	0.00047	0.00115	0.00055	0.00042
	Quartic	0.0017	0.00072	0.00055	0.00183	0.00082	0.00057	0.00138	0.00063	0.00048
	Triweight	0.00174	0.00073	0.00056	0.00222	0.00096	0.00067	0.00155	0.00068	0.00051
	Gaussian	0.00189	0.00078	0.00059	0.00144	0.00074	0.00061	0.00083	0.00041	0.0003
	Proposed1	0.0016	0.0007	0.00054	0.00142	0.00068	0.00048	0.00115	0.00056	0.00042
	Proposed2	0.0016	0.0007	0.00054	0.00168	0.00078	0.00055	0.00129	0.00061	0.00046
0.4 N(3,3)+0.6 N(-5,3/2)	Epanch.	0.00078	0.00034	0.00028	0.00058	0.00032	0.00025	0.00052	0.00027	0.00022
	Quartic	0.00081	0.00035	0.00029	0.00076	0.00039	0.0003	0.00065	0.00031	0.00025
	Triweight	0.00083	0.00035	0.00029	0.00093	0.00045	0.00035	0.00072	0.00033	0.00027
	Gaussian	0.0009	0.00037	0.00031	0.00083	0.00038	0.0003	0.00037	0.00019	0.00016
	Proposed1	0.00076	0.00034	0.00028	0.00059	0.00032	0.00026	0.00053	0.00027	0.00023
	Proposed2	0.00077	0.00034	0.00028	0.00071	0.00037	0.0003	0.00061	0.00029	0.00025
0.15N(0,2.5)+0.85 N(5,1.25)	Epanch.	0.00178	0.00081	0.00057	0.00142	0.00068	0.00046	0.00122	0.00062	0.00042
	Quartic	0.00184	0.00082	0.00057	0.0018	0.00081	0.00053	0.00143	0.00069	0.00047
	Triweight	0.00188	0.00083	0.00058	0.00213	0.00093	0.0006	0.00157	0.00074	0.00051
	Gaussian	0.00203	0.00087	0.0006	0.00164	0.00066	0.00049	0.00077	0.00036	0.00027
	Proposed1	0.00173	0.00081	0.00057	0.00145	0.0007	0.00047	0.00122	0.00063	0.00042
	Proposed2	0.00174	0.00081	0.00057	0.0017	0.00079	0.00052	0.00136	0.00068	0.00046

From the tables above we see the following:

- ❖ For table (2) and for all the parameters are proposed we find that the proposed nonparametric estimator is the best for the probability density function, then variable kernel and lastly fixed kernel.
- ❖ The Gaussian kernel is the best kernel when we are using the proposed density estimator, then Epanchnikov and the two proposed kernels respectively.
- ❖ As a general case we see that the best kernel is Epanchnikov when we are using variable kernel, then the two proposed kernels except at one case when using the distribution N(-2,2) with sample size n=40.
- ❖ For the case N (0, 2.5) and n=40, the results show that the Gaussian kernel is the best, then Epanchnikov and the two proposed kernels respectively.
- ❖ The best kernels when we are using fixed kernel estimator are the two proposed kernels, then Epanchnikov.
- ❖ Refer to the results we see that the two kernels *Quartic and Triweight) are less efficient relative to the other kernels when we are using the proposed kernel estimator.

From the table (3) we obtain the following:

- ❖ The best estimator for the probability density function is the proposed kernel estimator (for all the kernels).
- ❖ The Gaussian kernel is the best kernel when we are using the proposed density estimator, then Epanchnikov and the first proposed kernels equally.
- ❖ When we use the variable kernel, the results show that the best kernels are the first proposed kernel with Epanchnikov, then the second proposed kernels.
- ❖ When we use the fixed kernel, the results show that the best kernels are the proposed kernels, then Epanchnikov.

6. <u>Conclusions</u>

From the results that we are mentioned we see that the best estimator for the probability density function is the proposed kernel estimator for all kernels.

Also we conclude that we must rely on the proposed kernels and Gaussian kernel especially when the proposed kernel estimator is applied.

For the fixed and variable kernel estimators we conclude that the proposed kernels and Epanchnikov should be used.

References

1. Munaf Yousif H. (2005)”Comparing nonparametric estimators for probability density functions” Ph.D. dissertation, Department of Statistics, Baghdad University.

2. Breiman, L., Meisel, W. and Purcell, E. (1977)” Variable kernel estimators of multivariate densities” Technometrics, Vol.19, No.2, PP135-144.

3. Gasser, T., Muller, H.G. & Mammitzsch, R. (1985)” Kernels for Nonparametric curve Estimation” J.Royal S.S., series B, Vol.47, No.2, PP238-252.

4. Hardle, W. (1990)” Applied Nonparametric regression” Cambridge, Cambridge University press.

5. Hardle, W. (1991)” Smoothing techniques with implementation in S“Springer- verlage, New York.

6. Parzen, E. (1962)” On Estimation of probability density function and mode “Ann. Stat., Vol.15, No.1, PP 1065-1076.

7. Rao, P.B.L.S. (1983)” Nonparametric functional estimation” Academic press.

8. Rosenblatt, M. (1956)”Remarks on some nonpaparametric estimates of a density function “Ann. Math. Stat., Vol.27, PP832-837.

9. Silverman, B.W. (1986)”Density Estimation for statistics and data analysis” Chapman and Hall, London.

10. Wand, M.P. & Jones, M.C. (1995)” Kernel Smoothing” Chapman and Hall, London.