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

ISSN 2409-4587

IUGNES
Vol. 25, No 2, 2017, pp 89-94

Accepted on (14-03-2017)

Kernel Estimation of the Conditional Density Function with Application of the Palestinian Merchandise Imports Data

Raid B. Salha[1,*]
Sahar Z. Alhourani[1]

[1]Department of Mathematics, Faculty of Science, Islamic University of Gaza, Gaza Strip, Palestine

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

Abstract

The relationship between a current observation and previous observations, where the conditional density function plays an important role, is the main subject of this paper. To study this relation, the unknown conditional density function must be estimated. In this paper, the kernel estimation of the mean and mode of the conditional density function will be studied, and the conditions under which these estimators are asymptotically normally distributed will be discussed.

The performance of the kernel estimator of the conditional mean and mode will be tested using simulated data. We will analyze the monthly data of the Palestinian merchandise imports using the kernel techniques to predict future observations.

Also, the Box and Jenknis methodology will be applied to the data and its results will be compared to that of the kernel techniques.

Keywords:
Conditional distribution,
Conditional mean,
Conditional mode,
Kernel estimation,
Time series.

1. Introduction:

Conditional probability density function (pdf) plays an important role in studying the relationship between a dependent variable Y and independent variable X. If the conditional pdf of Y given X, $f(y|x)$ is unknown, it must be estimated. Indeed, estimating the conditional pdf is actually much more informative, since it allows us not only to recalculate the expected value $E(Y|X)$, but also to provide the general shape of the conditional distribution.

Several nonparametric methods can be proposed for estimating the conditional pdf function based on data $(X_1, Y_1), \ldots, (X_n, Y_n)$. One of the most popular method to estimate $f(y|x)$ is the kernel method, where no assumptions on the distribution of (X, Y) are assumed. Conditional distribution estimation was introduced by Rosenblatt (1969). A bias correction was proposed by Hyndman et. al. (1996) and a modification of the kernel estimation of the conditional pdf has proposed by Cai (2002).

In this paper, we will introduce the kernel estimation of two aspects of the conditional pdf, the conditional mean and the conditional mode. Then their performance will be tested using simulated data. Also, we will analyze the monthly data of the Palestinian merchandise imports using them to predict future observations. Finally, the Box and Jenknis methodology will be applied to the data and its results will be compared to that of the kernel techniques.

The remaining of this paper is organized as follows. In Section 2, the kernel estimation of the conditional pdf is introduced. Then the kernel estimation of two aspects of the conditional pdf, the conditional mean and the conditional mode will be

presented and discussed in Section 3 and Section 4 respectively.
Section 5 contains a comparison between the two estimators using simulated data. In Section 6, the two estimators will be used to predict future observation of the Palestinian merchandise imports. The result of the kernel method will be compared to that of Box and Jenknis methodology. Finally, we closed this paper by some conclusion remarks.

2. Kernel estimation of the conditional probability density function:

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be $\boldsymbol{R} \times \boldsymbol{R}$ valued independent random variables with a joint pdf $f(x, y)$. Also assume that X admits a marginal density $g(x)$. The conditional pdf of Y given $X = x$ is denoted by $f(y|x)$ and it is defined by

$$f(y|x) = \frac{f(x,y)}{g(x)}, \quad g(x) > 0.$$

The conditional pdf plays an important role in studying the relationship between the two variable Y and X.
If $f(y|x)$ is unknown, it is must be estimated to study the relationship between the response variable Y and predictor variable X. One of the best method to estimate $f(y|x)$ is the kernel method. To estimate it, we consider the following kernel estimators of the joint density $f(x, y)$ and the marginal pdf $g(x)$,

$$\hat{f}(x,y) = \frac{1}{n}\sum_{i=1}^{n} K_{h_n}(x - X_i)\, K_{h_n}(y - Y_i),$$

and

$$\hat{g}(x) = \frac{1}{n}\sum_{i=1}^{n} K_{h_n}(x - X_i),$$

where $K_{h_n}(x) = \frac{1}{h_n} K\left(\frac{x}{h_n}\right)$.

Then the kernel estimator of $f(y|x)$ is given by

$$\hat{f}(y|x) = \frac{\hat{f}(x,y)}{\hat{g}(x)} = \frac{\sum_{i=1}^{n} K_{h_n}(x - X_i)\, K_{h_n}(y - Y_i)}{\sum_{i=1}^{n} K_{h_n}(x - X_i)}, \qquad (1)$$

where $K(.)$ is a kernel function and h_n is a sequence of positive numbers converges to zero.
In the literature, the researchers usually study one of the following three aspects of the conditional pdf, the conditional mean function, the conditional mode function and the conditional quantiles function. For example, Salha and El Shekh Ahmed (2015) studied the Rewighted Nadaraya –Watson estimator of the conditional mean function at distinct points and established its asymptotic normality. Salha and Iqelan (2015) proposed an estimator of the conditional mode using the symmetrized nearest neighbor kernel estimator. They discussed the asymptotic properties of the estimator and compared its performance to that of the Nadaraya-Watson estimator. El Shekh Ahmed et al. (accepted, 2015) studied a new technique to improve the performance of the Nadaraya-Watson estimator of the conditional quintiles by replacing the constant bandwidth by varying bandwidth.

3. Kernel estimation of the conditional mean function:

In this section, we introduce the Nadaraya-Watson (NW) kernel estimation of the conditional mean function and its asymptotic properties.
Let $(X_i, Y_i), \ldots, (X_n, Y_n)$ be a random sample in which the relationship between Y and X can be written as,

$$Y_i = m(X_i) + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where $m(x)$ is the regression mean function

$$m(x) = E(Y|X = x),$$

and $\varepsilon_1, \ldots, \varepsilon_n$ are independent random variables for which,

$$E(\varepsilon_i) = 0, Var(\varepsilon_i) = \sigma^2 \text{ and } Cov(\varepsilon_i, \varepsilon_j) = 0, \qquad i \neq j.$$

Now, we will introduce the NW estimator. It is one of the popular nonparametric methods for estimating the conditional density function $f(y|x)$ and the conditional mean function.

Definition 2.1 (The Conditional Mean)

Let X and Y be a continuous random variables with joint pdf $f(x, y)$. The conditional mean function of Y given $X = x$, $E(Y|X = x)$, is defined as follows

$$m(x) = E(Y|X = x) = \int_{-\infty}^{\infty} y f(y|x)dy = \frac{\int_{-\infty}^{\infty} y f(x,y)dy}{\int_{-\infty}^{\infty} f(x,y)dy}.$$

The estimator $\hat{m}(x)$ of $m(x)$ is defined as follows

$$\hat{m}(x) = \int_{-\infty}^{\infty} y\hat{f}(y|x)dy = \frac{\int_{-\infty}^{\infty} y\hat{f}(x,y)dy}{\int_{-\infty}^{\infty} \hat{f}(x,y)dy}. \qquad (2)$$

The NW estimator for $m(.)$ is given by:

$$\widehat{m}_{NW}(x) = \frac{\sum_{i=1}^{n} K_h(x-X_i)Y_i}{\sum_{i=1}^{n} K_h(x-X_i)} = \frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)Y_i}{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)} \quad (3)$$

The asymptotic normality of the $\widehat{m}_{NW}(x)$ has been established under the following assumptions, Geenens (2014).

Assumptions:

We consider the following assumptions

A1. As $n \to \infty$, $h_n \to 0$, $n\,h_n \to \infty$.

A2. The kernel function *K(.)* is nonnegative probability density function with compact support satisfying:

$$\int_{-\infty}^{\infty} sK(s)\,ds = 0, \quad \int_{-\infty}^{\infty} K(s)ds = 1, \quad \int_{-\infty}^{\infty} s^2\,K(s)ds < \infty.$$

Theorem 1. Under assumptions **A.1 and A.2**, we have that

$$\sqrt{nh}\ [\widehat{m}\,(x) - m(x)] \xrightarrow{D} N\left(0, \sigma^2\,(x) f(x) \int_{-\infty}^{\infty} K^2(u)\,du\right),$$

where $\sigma^2\,(x) = Var(Y|X)$.

4. Kernel estimation of the conditional mode:

The problem of estimating the mode of the pdf is a matter of both theoretical and practical interest. Parzen (1962) first consider the problem of estimating the mode of a univariate pdf. Parzen (1962) and Nadaraya (1965) have shown that under regularity conditions the estimate of the population mode obtained by maximizing a kernel estimate of the pdf is strongly consistent and asymptotically normally distributed. Samanta (1973) has given multivariate versions of Parzen's results. Samanta and Thavaneswaran (1990) considered the problem of estimating the mode of a conditional pdf and they have shown under regularity conditions that the estimate of the population conditional mode is strongly consistent and asymptotically normally distributed. Salha and Ioanides (2007), generalized the results of Samanta and Thavaneswaran (1990), by considering the estimation of the conditional mode at a finite number of distinct points.

The population conditional mode θ(x) is defined by

$$\theta(x) = \arg\max_{y\in \mathrm{R}} f(y|x), \qquad x \in \boldsymbol{R},$$

Definition 2. (The estimated conditional mode).

The estimated conditional mode is defined as the maximum of $\hat{f}$(y|x) over y$\in$ **R**,

$$\hat{\theta}(x) = \arg\max_{y\in \mathrm{R}} \hat{f}(y|x), \qquad x \in \boldsymbol{R},$$

where $\hat{f}(y|x)$ is given in Equation (1).

We call $\hat{\theta}(x)$ the sample conditional mode. $\hat{\theta}(x)$ is considered as an estimate $\theta(x)$.

Consider the following conditions from Samanta and Thavaneswaran (1990),

C1. $(X_1, Y_1), \ldots\ldots, (X_n, Y_n)$ is a sample of i.i.d. random variables with joint pdf $f(x,y)$, where the following hold:

i. The marginal probability density function of X, g(x) is uniformly continuous.

ii. $f^{(i,j)}(x,y) = \frac{\partial^{i+j} f(x,y)}{\partial x^i \partial y^j}$ exist and are bounded for $1 \le i+j \le 4$.

C2. The kernel K is a Borel function and satisfies the following:

i. $K(u)$ tends to zero as u tends to ±∞

ii. $K(u)$ and its first two derivatives are functions of bounded variation.

iii. $\lim_{|u|\to\infty} \left|u^2 K^{(i)}(u)\right| = 0, \quad (i = 0,1)$

iv. $\int_{-\infty}^{\infty} u^i K(u)\,du = 1, \ (if\ i = 0),$

v. $\int_{-\infty}^{\infty} u^i K(u)\,du = 0, \quad (if\ i = 1,2)$

vi. $\int_{-\infty}^{\infty} |u|^3\,K(u)du < \infty$

C3. h_n is a sequence of positive numbers tending to zero, and satisfies the following

$\lim_{n\to\infty} nh_n{}^8 = \infty, \qquad \lim_{n\to\infty} nh_n{}^{10} = 0$

Theorem 2. Suppose that $f(x,y) > 0$, then under the assumption **C1**, **C2** and **C3**, the following holds

$$(nh_n{}^4)^{\frac{1}{2}}\{\hat{\theta}(x) - \theta(x)\} \xrightarrow{D} N\left(0, \frac{f(x,\ \theta(x))}{\{f^{(0,2)}(x,\ \theta(x))\}^2} \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \{K(u)K^{(1)}(v)\}^2\ dudv\right),$$

where, $f^{(0,2)}(x,\ \theta(x))$ is the second derivative of $f(x,\ y)$ with respect to y and $K^{(1)}(v)$ denotes the first derivative of $K(v)$.

Proof. See Samanta and Thavaneswaran (1990).

5. Simulation study:

In this section, the performance of the conditional mode and mean estimators will be tested using a simulation study. Samples of sizes 50, 200 and 500 are simulated from the model

$$y = sin(2\pi x) + \exp(-16x^2) + e$$

$x \sim N(0,1)$, $e \sim N(0, 0.5)$.

The bandwidth h_n will be computed using the following formula from Silverman (1986)

$h_n = 1.06\, s\, n^{-\frac{1}{5}}$, where s is the sample standard division and n is the sample size.

Figure 1 shows the graph of the true curve of the simulated data from the model together with its estimation using the mode and mean estimators. The MSE for the two estimators are computed and listed in Table 1. The results of the comparison indicated data the conditional mean estimator performs better than the conditional mode.

Table 1 *The MSE for the conditional mode and mean estimators*

Sample size	Mode	Mean
50	0.0745390	0.0605896
200	0.0269099	0.0202534
500	0.0226929	0.0128929

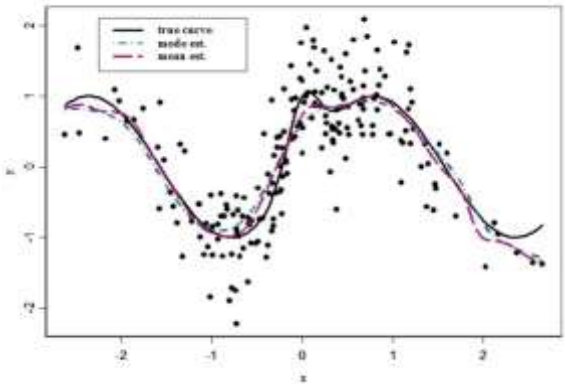

Figure 1 *Graph of the simulated data together with the true curve and its estimators*

6. Application:

We consider the total of values for Palestinian merchandise imports (unless the part of Jerusalem who Israel embrace by force after occupation the west bank in 1967) data. (Monthly totals in thousands of American dollars from January 2005 to December 2014). we have transformed the data by taking the logarithm. We used the first 109 observation to predict the last 11 observations. The plot of the time series is shown in Figure 2.

Also, we used the Box et al. (1994) methodology to analyze the data using ARIMA models. Figure 3 shows the first difference of the data. The suitable model to predict future observations is ARIMA(2,1,1) and its equation is given by

$$\hat{Y}_t = 0.008 - 1.337\hat{Y}_{t-1} - 0.506\hat{Y}_{t-2} + 0.960\,\varepsilon_{t-1}$$

Figure 4 shows the time series of the data together with the forecast values of the last 11 observations using the ARIMA(2,1,1) model. Table 2 contains the last 11 true observations of the data together with their estimation using the three different methods. The MSE for the three estimators is listed in Table 3. In Figure 5, the tail of the time series is plotted together with the forecasting values for the last 11 observations from the three estimators.

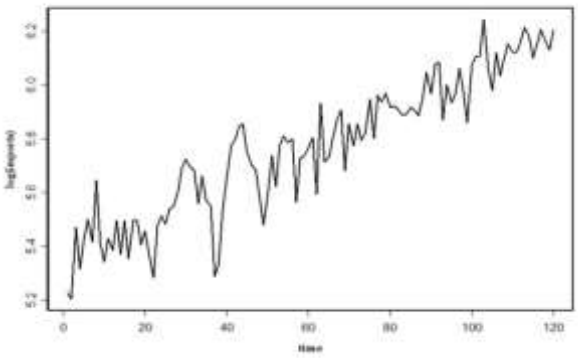

Figure 2 *The time series of the transformed Palestinian merchandise imports*

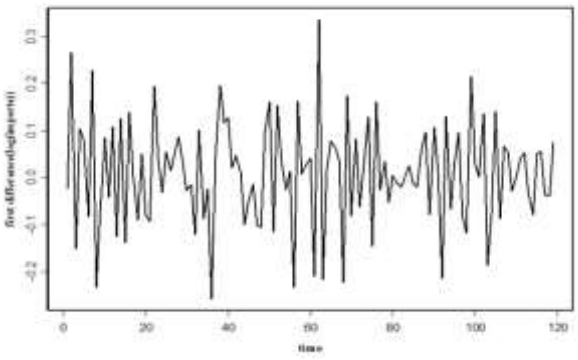

Figure 3 *The first difference of the transformed Palestinian merchandise imports*

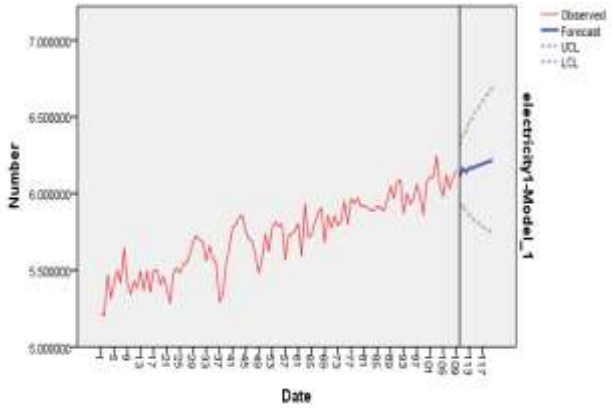

Figure 4 *The transformed Palestinian merchandise imports and the forecast values*

Table 2 *The forecasting values for the Palestinian merchandise imports*

Obs. #	True obs.	Mean est.	Mode est.	ARIMA
110	6.12	6.10	6.04	6.11
111	6.12	5.96	6.04	6.16
112	6.16	6.15	6.06	6.14
113	6.2	6.20	6.07	6.17
114	6.18	6.18	6.07	6.16
115	6.10	6.10	6.03	6.18
116	6.15	6.20	6.06	6.19
117	6.21	6.13	6.07	6.19
118	6.17	6.17	6.06	6.20
119	6.13	6.06	6.05	6.21
120	6.21	6.17	6.07	6.22

Table 3 *The MSE for the three estimators of the Palestinian merchandise imports*

Estimator	Mean	Mode.	ARIMA(2,1,1)
MSE	0.0038012	0.0112768	0 .0017790

7. Conclusions:

In this paper, the kernel estimation of two aspects of the conditional pdf, the conditional mean and mode, has been discussed and compared the conditional mean. A simulation study indicated that the conditional mean performs better than the conditional mode. Also, we used the two estimators to analyze the Palestinian merchandise imports and to forecast a future observations. Then we compared their performance to that of the ARIMA models. The comparison indicated that, the ARIMA(2,1,1) model was the best one. Also, the comparison indicated that the performance of the conditional mean estimator is close to the performance of the ARIMA(2,1,1) model.

We note from the two comparisons that, the estimator based on the conditional mean is better than that based on the conditional mode. This can be attributed to the nature of the mean which gives values more closer to the average of the data than that gives the mode.

We suggest to use modified kernel estimators of the conditional mean such that were proposed and discussed by Salha and El Shekh Ahmed (2015) and, Salha and Iqelan (2015) to analyze and forecast the Palestinian merchandise imports. The modified estimators will give better results and their performance will be close to that of the ARIMA models.

References:

Box, G. E. P., Jenkins, G. M., and Reinsel, G. C. (1994). *Time series analysis, forecasting and control*. (3rd ed.). Englewood: Prentice Hall.

Cai, Z. (2002). Regression Quantile for time series. *Econometric Theory*, *18*(01), 169-192.

El Shekh Ahmed, H., Salha, R. and EL-Sayed H. (accepted, 2015). Adaptive Weighted Nadaraya-Watson Estimation of the Conditional Quantiles by Varying Bandwidth, *Communications in Statistics - Simulation and Computation*

Geenens, G. (2014). Explicit formula for asymptotic higher moments of the Nadaraya-Watson Estimator. *The Indian Journal of Statistics. Vol. 76 -A, (1) , 77-100.*

Hyndman, R. J., Bashtannyk, D. M., & Grunwald, G. K. (1996). Estimating and visualizing conditional densities. *Journal of Computational and Graphical Statistics*, *5*(4), 315-336.

Nadaraya, E. A. (1965). On non-parametric estimates of density functions and regression curves. *Theory of Probability & Its Applications*, *10*(1), 186-190.

Parzen, E. (1962). On estimation of a probability density function and mode. *The annals of mathematical statistics*, *33*(3), 1065-1076.

Rosenblatt, M. (1969). Conditional probability density and regression estimates. In P. R. Krishnaiah (Ed.), *Multivariate Analysis II* (pp. 25-31). New York: Academic Press.

Salha, R., and El Shekh Ahmed, H. (2015). Reweighted Nadaraya-Watson Estimator of the Regression Mean. International *Journal of Statistics and Probability, 4(1), 138-147.*

Salha, R., and Ioanides, D. (2007). The asymptotic distribution of the estimated conditional mode at a finite number of distinct points under dependence conditions. *The Islamic University Journal*, *15*(2),199-214.

Salha, R., and Iqelan, B. (2015). Estimating the conditional mode using the symmetrized nearest neighbor kernel estimator. *The Islamic University Journal*, *23*(2), 12-20.

Samanta, M. (1973). Nonparametric estimation of the mode of a multivariate density. *South African Statistical Journal*, *7* (2), 109-117.

Samanta, M., and Thavaneswaran, A. (1990). Nonparametric estimation of the conditional mode. *Communications in Statistics-Theory and Methods*, *19*(12), 4515-4524.

تقدير النواة لدالة الكثافة الشرطية مع تطبيق على بيانات واردات البضائع الفلسطينية

الموضوع الرئيسي لهذا البحث هو دراسة العلاقة بين المشاهدة الحالية والمشاهدات السابقة، حيث تلعب دالة الكثافة الشرطية دورا هاما، ولدراسة هذه العلاقة، يجب تقدير دالة الكثافة الشرطية المجهولة.
في هذه البحث، سيتم دراسة تقدير النواة لوسط ومنوال دالة الكثافة الشرطية، وسيتم مناقشة الشروط التي بموجبها يتقارب توزيع هذين المقدرين للتوزيع الطبيعي.
سيتم اختبار أداء مقدري النواة للوسط الشرطي والمنوال الشرطي باستخدام المحاكاة. سنقوم بتحليل البيانات الشهرية لواردات البضائع الفلسطينية باستخدام تقدير النواة للتنبؤ بالمشاهدات المستقبلية. أيضا، سيتم تطبيق منهجية بوكس وجنكنيز على هذه البيانات وسيتم مقارنة نتائجها مع تلك التي حصلنا عليه من طريقة تقدير النواة.

كلمات مفتاحية:
التوزيع الشرطي،
الوسط الشرطي،
المنوال الشرطي،
تقدير النواة،
المتسلسلة الزمنية.