ESTIMATION OF MULTIVARIATE PROBABILITY DENSITY FUNCTION WITH KERNEL FUNCTIONS

S. Gökçe Cula and Ö. Toktamış
Hacettepe University, Department of Statistics

Abstract

In this study, multivariate kernel density estimation has been investigated. Also, the applicability of multivariate kernel density function, estimation of two variable probability density function whose geometric presentation is possible has been shown by using the earthquake data in Marmara region.

Key words: Multivariate density estimation, bandwidth choice, cross-validation, biased validation, bootstrap.

1. Introduction

A nonparametric modelling process in multivariate case is more complicated than one in univariate case, in order to determine the structure in data sets studies related to nonparametric probability density estimation are less. Therefore, in recent years it has needed that this issue has been taken account more frequently, in this study kernel density estimation method, which was first studied by Rosenblatt (1956) and Parzen (1962), which has extensive application field in univariate case and whose mathematical properties can be investigated very well is studied.

The univariate kernel density estimation has one bandwidth parameter. The specification of more bandwidth parameters than one is required for multivariate density estimation. Also, it has been faced with difficulties in geometric presentation in multivariate case. The multivariate density estimation is the generalization of univariate case.

The kernel estimation at a given point in one variable case defined as an weighted mean which is calculated by overlapping the mean point of the kernel function with the given point taking account the other observations with weights obtained according to the kernel function and the bandwidth is extended to multivariate case (Toktamış, 1995).
2. Multivariate kernel density estimation

Let \( X_1, X_2, \ldots, X_n \), denote a \( d \)-variate random sample having density \( f \), the component of \( b \) and the component of \( X_i \) vector be \( \mathbf{X}_i = (X_{i1}, X_{i2}, \ldots, X_{id})' \) and the component of \( x \) vector be \( \mathbf{x} = (x_1, x_2, \ldots, x_n)' \) and \( \mathbf{x} \in \mathbb{R}^d \). Also, \( f \) notation is shortland for \( \int_{\mathbb{R}^d} \cdots \int_{\mathbb{R}^d} dx \) is shortland for \( dx_1, dx_2, \ldots, dx_d \) and the \( d \times d \) identity matrix is denoted by \( I \). Under these notations, the \( d \) dimensional kernel density estimator is given as follows:

\[
\hat{f}(\mathbf{x}; \mathbf{H}) = \frac{1}{n |\mathbf{H}|^{1/2}} \sum_{i=1}^{n} K_d \left( \frac{\mathbf{x} - \mathbf{X}_i}{|\mathbf{H}|^{1/2}} \right) 
\]  

(2.1)

Where \( \mathbf{H} \) is a symmetric positive definite \( d \times d \) matrix called the bandwidth matrix and \( K \) is a \( d \)-variate kernel functions satisfying \( \int_{-\infty}^{\infty} K_d(\mathbf{x})\, d\mathbf{x} = 1 \). \( K_d \) is usually chosen to be a \( d \)-variate probability density function. Nevertheless, in order to construct a multivariate kernel function from a multivariate kernel function, two common techniques have been followed. One of them is to use (2.2) called product kernel function as a kernel function and the other one is to use (2.3) called spherical kernel function as a kernel function. These equations are given as follows:

\[
K_d(\mathbf{x}) = \prod_{i=1}^{d} K(x_i) 
\]  

(2.2)

\[
K_d(\mathbf{x}) = \frac{K \left\{ \left( \mathbf{x} \mathbf{x} \right)^{1/2} \right\}}{\int K \left\{ \left( \mathbf{x} \mathbf{x} \right)^{1/2} \right\} d\mathbf{x}} 
\]  

(2.3)

Another choice which is widely used is to use symmetric unimodal probability density function. The most widely used function for this purpose is standard \( d \)-variable normal density function which is given as follows:

\[
K_d(\mathbf{x}) = (2\pi)^{-d/2} \exp \left( -\frac{1}{2} \mathbf{x} \mathbf{x} \right) 
\]  

(2.4)

The kernel estimator which is given by (2.1) requires specification of the bandwidth matrix \( \mathbf{H} \), which has \( \frac{d(d+1)}{2} \) distinct entries. As dimension increases it is getting more difficult to control calculations in which \( \mathbf{H} \) matrix is used. In order to simplify \( \mathbf{H} \) matrix some restrictions are proposed. Therefore, three situation are considered. The simplest situation corresponds to the restriction \( \mathbf{H} \in S \) which means that \( \mathbf{H} = h^2 I \) (\( h > 0 \)) . This restriction, which is to use one constant \( h \) bandwidth, means that the amount of smoothing in each direction is the same. This is suitable if the scales of all variables are roughly the same. So this selection can be done only if each variable is standardized to be on a common scale (Simonoff, 1996). Another restriction is to take \( \mathbf{H} \in D \), \( \mathbf{H} = \text{diag}(h_1^2, h_2^2, \ldots, h_d^2) \). This restriction allows different
ESTIMATION OF DENSITY FUNCTIONS

amounts of smoothing in each coordinate direction. This approach is the practical
version of restriction $H \in S$. Let $F$ denote the class of symmetric, positive definite
$d \times d$ matrices, $D$ is the subclass of diagonal positive definite $d \times d$ matrices, $D \subseteq F$.
When the smoothing in different directions from the direction of coordinates are re-
quired, the full bandwidth matrix, $H \in F$, would be appropriate. In this case as the
number of different elements of $H$ matrix increases that is the number of parameters
to be estimated increases. This means that in case of $H \in F$ kernel density estimation
becomes more complicated. Jones and Wand has been done a detailed study related
to the three different choices given above, of bandwidth matrix $H$ which will be used
to estimate the bivariate density function (Wand and Jones, 1993).

Under $H \in D$, $H = \text{diag} (h_1, h_2^2, \ldots, h_d^2)$, multivariate kernel density estimation can
be written as follows (Wand and Jones, 1995):

$$
\hat{f}(x; h) = \frac{1}{n} \left( \prod_{j=1}^{d} h_j \right)^{-1} \sum_{i=1}^{n} K_d \left( \frac{x_1 - X_{i1}}{h_1}, \frac{x_2 - X_{i2}}{h_2}, \ldots, \frac{x_d - X_{id}}{h_d} \right) \tag{2.5}
$$

In 1996, the equation given above (2.5) was simplified by Cacoullos under $H \in S$
, $S = \{ h^2 I : (h > 0) \}$. This simplified form is defined in (2.6):

$$
\hat{f}(x; h) = \frac{1}{nh^d} \sum_{i=1}^{n} K_d \left( \frac{x - X_i}{h} \right) \tag{2.6}
$$

This selection of bandwidth means that the amount of smoothing is the same in
every direction. By using product kernel function as kernel function $K_d$, (2.6) can be
rewritten as (Sain et al., 1994):

$$
\hat{f}(x; h) = \frac{1}{nh^d} \sum_{i=1}^{n} \left\{ \prod_{j=1}^{d} K_d \left( \frac{x_j - X_{ij}}{h} \right) \right\} \tag{2.7}
$$

The use of only one bandwidth parameter $h$ in (2.6) shows that scaling the kernel
function which is placed at each observation is the same in different coordinate
directions. If spread of the data points on one coordinate axis is wider than the other
one, then it is necessary to use a different bandwidth for every variable. But, in
this situation, it is very difficult to obtain optimal bandwidth from mean integrated
squared error, MISE, related to kernel estimation. Because, to make MISE minimum
for every $h$ requires very complicated calculations. In the majority of multivariate
statistical processes, the data need to be standardized in order to make disappear the
difference among the ranges of variables (Wand and Jones, 1993). If the standardiza-
tion on the variables of multivariate kernel density estimation is carried out, then the
equation (2.6) which includes one smoothing parameter is used (Silverman, 1986).
3. Asymptotic MISE approximations

In the majority of studies related to the density estimation, the comment about estimation performance is made by measuring the closeness of the estimator to its target value. Rosenblatt stated that the use of MISE which is widely used for the kernel density estimation and easily followed criteria is preferred due to its matematically simpler (Rosenblatt, 1956).

An asymptotic approach for MISE of multivariate kernel density estimation can be obtained in a similar manner to the univariate kernel density estimation. While this approach is obtained, some assumptions as to density function \( f \), kernel function \( K_d \) and bandwidth matrix \( H \) are given as follows:

i) Each entry of \( Hf(\cdot) \) is piecewise continuous and square integrable;

ii) \( H = H_n \), is a sequence of bandwidth matrices such that \( n^{-1/2} \) and all entries of \( H \) approach zero as \( n \to \infty \);

iii) \( K_d \) is a bounded, compactly supported \( d \)-variate kernel satisfying

\[
\int_{-\infty}^{\infty} K_d(z) \, dz = 1, \quad \int_{-\infty}^{\infty} zK_d(z) \, dz = 0 \quad \text{and} \quad \int_{-\infty}^{\infty} z^2 K_d(z) \, dz = \mu_2(K_d) I
\]

(3.1)

where, \( \mu_2(K_d) = \int_{-\infty}^{\infty} z^2 K_d(z) \, dz \) is independent of \( i \).

Under the above assumptions, the asymptotic mean integrated square error of a multivariate kernel density estimator, AMISE, can be obtained as follows:

\[
AMISE\left\{\hat{f}(x; H)\right\} = \frac{1}{n} |H|^{-1/2} \int_{-\infty}^{\infty} K_d(z)^2 \, dz + \frac{1}{4} \mu_2(K_d)^2 \int_{-\infty}^{\infty} tr^2\{HH_f(x)\} \, dx
\]

(Wand and Jones, 1995). Generally, an explicit expression for the AMISE optimal bandwidth matrix of the multivariate kernel density estimator is not available and the numerical value for this quantity can only be obtained by simulation. The most important problem in (3.2), is how the multivariate integrals in (3.2) are evaluated. Nevertheless, it can be obtained for AMISE simpler formula under \( H \in D \) and \( H \in S \).

For example, in the case where \( H = h^2I \), AMISE of the multivariate kernel density estimator can be obtained as follows (Wand and Jones, 1995):

\[
AMISE\left\{\hat{f}(x; H)\right\} = \frac{1}{nh^d} \int_{-\infty}^{\infty} K_d(z)^2 \, dz + \frac{1}{4} h^4 \mu_2(K_d)^2 \int_{-\infty}^{\infty} \left\{\nabla^2 f(x)\right\}^2 \, dx
\]

(3.3)

where \( \nabla^2 f(x) = \sum_{i=1}^{d} \left( \frac{\partial^2}{\partial x_i^2} \right) f(x) \).
ESTIMATION OF DENSITY FUNCTIONS

When a specific error criteria, for example AMISE, is fixed at a predetermined value, the required sample size increases rapidly with the number of dimension. The study which is related to this issue was carried out by Scott and Wand (Scott and Wand, 1991). For example, when density function \( f \) and kernel function \( K_d \) are taken normally distributed with mean zero and variance \( I_d \), the sample sizes necessary to achieve the given AMISE=0.393 have been obtained and it has been given in Table 1 (Simonoff, 1996).

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Required Sample Size, ( n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>480</td>
</tr>
<tr>
<td>7</td>
<td>5382</td>
</tr>
<tr>
<td>10</td>
<td>299149</td>
</tr>
</tbody>
</table>

Table 1. Sample size required for each dimension to achieve the given AMISE=0.393

4. Bandwidth selection

Kernel density estimators are affected from the bandwidth \( h \) very much. When the bandwidth \( h \) is chosen very small, the variance of estimator decreases while the bias of estimator decreases. When bandwidth \( h \) is chosen very big, bias increases while variance decreases. Therefore, we have to choose such a bandwidth that the optimal bandwidth is obtained. Some error criteria are used to obtained optimal bandwidth \( h \). The optimal bandwidth which makes AMISE minimum for multivariate kernel density estimation given in (3.3) is:

\[
h_{AMISE} = \left( \frac{d \int_{-\infty}^{\infty} K_d(z)^2 \, dz}{n \mu_d (K_d)^2 \int_{-\infty}^{\infty} \{\nabla^2 f(x)\}^2 \, dx} \right)^{1/(d+4)}
\]  

(4.1)

As seen from (4.1) The optimal bandwidth which makes AMISE minimum depends on second derivatives of an unknown density \( f \). As this (4.1) involves an unknown function, the optimal bandwidth can not be obtained by using this equation. That's why other methods a few of which are mentioned below are suggested. Some of these methods are given as follows.

4.1. Choice of bandwidth for a standard distribution

The aim of this method is to find an expression instead of \( \int_{-\infty}^{\infty} \{\nabla^2 f(x)\}^2 \, dx \) in (4.1). For this purpose the density function which is known is taken and the optimal bandwidth \( h \) is obtained. For example if standard \( d \)-variate normal density function is used instead of unknown density function \( f \), then the integral which was taken place in (4.1) becomes:
\[ \int_{-\infty}^{\infty} \left( \nabla^2 f(x) \right)^2 \, dx = (2\sqrt{\pi})^{-d} \left( \frac{1}{2} + \frac{1}{4d^2} \right) \]

(Silverman, 1986). By substituting \( \int_{-\infty}^{\infty} \left( \nabla^2 f(x) \right)^2 \, dx \) into (4.1) the optimal bandwidth is obtained.

4.2. Data-driven methods for bandwidth selections

The least squared cross-validation method, LSCV, for the choice of bandwidth matrix \( H \) is exactly a data driven method. Rudemo and Bowman both separately from each other suggested this method for the choice of bandwidth of kernel estimator in a univariate density function (Rudemo, 1982; Bowman, 1984). This method is generalized to obtain the bandwidth matrix \( H \) in multivariate case. As a result of generalization, least squared cross-validation function (LSCV(H)) is found as follows:

\[ \text{LSCV}(H) = \int \hat{f}(x; H)^2 \, dx - 2n^{-1} \sum_{i=1}^{n} \hat{f}_{-i}(X_i; H) \]

(4.2)

where \( \hat{f}_{-i}(X_i; H) \) is the kernel estimator based on the sample with \( X_i \) deleted. Here, the main purpose is to obtain the optimal bandwidth matrix \( H \) which minimizes the expression given by (4.2) and to use this in the multivariate kernel density estimation. Under the restriction \( H \in D \) if the standard normal density function is used instead of kernel function \( K_d \), then (4.2) becomes:

\[ \text{LSCV}(h_1, h_2, \ldots, h_d) = \frac{1}{(2\sqrt{\pi})^d n h_1 h_2 \ldots h_d} + \frac{1}{(2\sqrt{\pi})^d n^2 h_1 h_2 \ldots h_d} \sum_{i=1}^{n} | \sum_{j \neq i} \exp \left\{ \frac{1}{4} \sum_{k=1}^{d} \left( \frac{x_{ik} - x_{jk}}{h_k} \right)^2 \right\} | - (2 \times 2^d/2) \exp \left\{ \frac{1}{2} \sum_{k=1}^{d} \left( \frac{x_{ik} - x_{jk}}{h_k} \right)^2 \right\} \]

(4.3)

As seen from (4.3) the function \( \text{LSCV}(h_1, h_2, \ldots, h_d) \) is a data driven one. The optimal bandwidth which makes (4.3) minimum are obtained and used in kernel density estimation. The least square cross-validation function can have more than one minimum. The studies have shown that the use of the bandwidth which has the largest local minimum is appropriate.

For multivariate case, Sain and his colleagues generalized biased cross-validation method, BCV, which is developed by Scott ve Terrell (1987) to obtained the bandwidth \( h \) in a univariate kernel density estimation. Sain and his colleagues used
ESTIMATION OF DENSITY FUNCTIONS

the estimation $\frac{1}{n} \sum_{i=1}^{n} f_{\hat{w}}(x_i)$ which has smaller bias instead of $\int f^2(x) dx$ (Sain et al., 1994). They used standard normal density function instead of kernel function $K_d$ under the restriction $H \in D$ and they found biased cross-validation function, $(BCV(h_1, h_2, ..., h_d))$, for multivariate product kernel estimation as follows:

$$BCV(h_1, h_2, ..., h_d) = \frac{1}{(2\sqrt{\pi})^d nh_1h_2...h_d} +$$
$$+ \frac{1}{4n(n-1)h_1h_2...h_d} \sum_{i=1}^{n} \sum_{j \neq i} \left\{ \sum_{k=1}^{d} \left( \frac{x_{ik} - x_{jk}}{h_k} \right)^2 \right\}^2 -$$
$$-(2d+4) \left\{ \sum_{k=1}^{d} \left( \frac{x_{ik} - x_{jk}}{h_k} \right)^2 \right\} + (d^2 + 2d) \times$$
$$\prod_{k=1}^{d} \Phi \left( \frac{x_{ik} - x_{jk}}{h_k} \right)$$

where $\Phi$ is standard normal density function. To obtain the optimal bandwidth, they found the bandwidth which minimizes (4.4) (Sain et al., 1994).

They also developed bootstrap method which is used by Taylor (1989) to find the bandwidth $h$ in a univariate case for multivariate density estimation. MISE for multivariate case is:

$$MISE(h) = \int_{R^d} E^* \left\{ \hat{f}^*(x) - \hat{f}(x) \right\}^2 dx$$

where $\hat{f}(x)$ is the multivariate kernel estimator, $\hat{f}^*(x)$ is a multivariate kernel estimator calculated with data sample from $\hat{f}(x)$, and the expectation, $E^*$, is taken with respect to the density $\hat{f}(x)$. Under restriction $H \in D$ by using the standard normal kernel function for $K_d$, bootstrap function, $(B(h_1, h_2, ..., h_d))$, for multivariate product kernel estimation as follows:

$$B(h_1, h_2, ..., h_d) = \frac{1}{(2\sqrt{\pi})^d nh_1h_2...h_d} +$$
$$+ \frac{1}{(2\sqrt{\pi})^d n^2h_1h_2...h_d} \sum_{i=1}^{n} \sum_{j \neq i} \sum_{k=1}^{n-1} \exp \left\{ \frac{1}{8} \sum_{k=1}^{d} \left( \frac{x_{ik} - x_{jk}}{h_k} \right)^2 \right\} +$$
$$+ \exp \left\{ \frac{1}{4} \sum_{k=1}^{d} \left( \frac{x_{ik} - x_{jk}}{h_k} \right)^2 \right\} - \frac{2}{3} \sum_{k=1}^{d} \exp \left\{ \frac{1}{6} \sum_{k=1}^{d} \left( \frac{x_{ik} - x_{jk}}{h_k} \right)^2 \right\}$$

(Sain et al., 1994). The bandwidths which minimize this function are taken and these values are used to obtain multivariate kernel density estimation.
5. Application

In this study firstly the computer programs have been written to obtain the values of kernel estimates of multivariate density function since no standard computer programs available. Secondly some computer programs have been written for some methods (LSCV, BCV, B) which have developed in order to obtain the bandwidths. The programs have been coded by Delphi 3 for methods (Cula, 1998).

In this application, the data are consist of 255 earthquakes with magnitude at least 4 on Richter scale occurring between 1900-1999 years in Marmara region. This data was used to obtain kernel estimation of bivariate probability density function. Here, bivariate standard normal density function is used as a kernel function. As two variables were measured by using the same scale, the raw data were used without making any standardisation. In this simulation firstly, the bandwidth \( h \) was increased by 0.01 between 0 and 2 and then the values of the functions BCV\((h_1, h_2, ..., h_d)\), LSCV\((h_1, h_2, ..., h_d)\) and B\((h_1, h_2, ..., h_d)\) were found and their distribution was obtained.

The bandwidth \( h \) which makes the function BCV\((h_1, h_2, ..., h_d)\) minimum was obtained as 1.36 and the value of the function BCV\((h_1, h_2, ..., h_d)\) which corresponds to this value was obtained as 0.02234. The graph of the function BCV\((h_1, h_2, ..., h_d)\) which corresponds to the bandwidths were given in Figure 1.

The bandwidth \( h \) which makes the function BCV\((h_1, h_2, ..., h_d)\) minimum is substituted into the bivariate kernel density estimator, obtaining the value of estimation. Figure 2 gives surface plot and contour plot of the kernel estimate for the earthquake data.

Figure 1: The graph of the function BCV\((h_1, h_2, ..., h_d)\) for the bandwidth's value between 1.30-1.42 for the earthquake data set

Optimal bandwidth \( h \) value couldn't be obtained from cross-validation and bootstrap methods. The graph related to this is given in Figure 3.

As seen from Figure 3, as \( h \) increases the value of the function LSCV\((h_1, h_2, ..., h_d)\) also increases and as \( h \) increases the value of the function B\((h_1, h_2, ..., h_d)\) decreases. However, minimum value couldn't be obtained for both of the functions, in other words optimal bandwidth couldn't be found.
ESTIMATION OF DENSITY FUNCTIONS

Figure 2: The graphs of the bivariate kernel density estimation values related to the earthquake data for the Marmara region when $h=1.36$ a) Surface Plot b) Contour Plot

Figure 3: The graph of a) the function $LSCV(h_1,h_2,\ldots,h_d)$, b) the function $B(h_1,h_2,\ldots,h_d)$ for the bandwidth's value between 0.1-1.4 for the earthquake data set
6. Conclusion

In the application of earthquake data the bandwidth value which was obtained by using the method BCV has been found as 1.36. It couldn’t be obtained the bandwidth value for both of the methods LSCV and B. By putting the bandwidth which is obtained from method LSCV into the bivariate kernel density function, the estimation values have been calculated and the graphs related to these values have been drawn (Figure 2). According to the data of 255 earthquakes with magnitude at least 4 on the Richter scale occurring between 1900-1999 years in Marmara region, the density related to the observation between longitude 40.30-41.20 and between latitude 27.75-30.40 is found the highest. It can be said that Istanbul, İzmit, Yalova cities which fall into these coordinates have higher probability of occurring earthquake with magnitude at least 4 on the Richter scale in Marmara region than the other places in Marmara region.

References


**ÖZET**

Bu çalışmada, çok değişkenli olasılık yoğunluk fonksiyonunun çekirdekkestirim yöntemi incelemiştir. Ayrıca çok değişkenli olasılık yoğunluk fonksiyonunun uygulanabilirliği, geometrik gösterimlere olanak sağlayan iki değişkenli olasılık yoğunluk fonksiyonunun kestirimini Marmara Bölgesi için elde edilen deprem verileri kullanılarak yapılan uygulama ile gösterilmiştir.