COMPARISON OF THE BANDWIDTH SELECTION METHODS FOR KERNEL ESTIMATION OF PROBABILITY DENSITY FUNCTION

Öniz Toktamış and Serpil Gökçe Cula
Hacettepe University Faculty of Science,
Department of Statistics, Beytepe-Ankara

Serdar Kurt
Dokuz Eylul University Faculty of Art and Sciences,
Department of Statistics, Alsancak-İzmir

Abstract

In this study the bandwidth selection methods for the kernel estimation of the probability density function are discussed. Least-squares cross-validation method, biased cross-validation method and bootstrap method are reviewed, compared and their applications are presented.

Key Words: Kernel estimation, bandwidth, cross-validation, biased cross-validation, bootstrap.

1. Introduction

Let \(X_1, X_2, \ldots, X_n\) be a random sample from an unknown absolutely continuous distribution with probability density function \(f\). The kernel estimate derived from this sample is

\[
\hat{f}(x, h) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right)
\]  

(1)

Here, \(K\) is the kernel function such that it is usually a probability density function unimodal and symmetric around zero. \(h\) is the window width which is called
the bandwidth or also called smoothing parameter because it controls the degree of smoothing in data for the kernel estimation.

The kernel estimation at a given point is weighted mean which is calculated by overlapping the mean point of the kernel function with the given point and taking account the other observations with certain weights which are obtained according to the kernel function and bandwidth (Toktamiş, 1995).

Kernel estimation method is one of the non-parametric methods to estimate probability density function and it was suggested first in 1956 by Rosenblatt and theoretical properties were investigated by Parzen in 1962 (Rosenblatt, 1956; Parzen, 1962).

In application, \( K \) and \( h \) are selected by the users. For different choices of \( K \) and \( h \), the estimates of the density function differ. The way of choosing \( K \) and \( h \) has been the interest of many studies. The choice of the kernel function \( K \) was studied first by Epanechnikov in 1969. Epanechnikov showed that there exists an optimal kernel in some sense, but there are other kernels which give almost optimal results (Epanechnikov, 1969). It is quite satisfactory to choose a kernel for computational convenience or differentiability properties. For this reason, the choice of kernel function is not as important as the choice of bandwidth in application (Silverman, 1986).

The choice of bandwidth has a very important place in the kernel density estimation. Boneva and his colleagues showed that small changes in bandwidth could change estimates on large scale (Silverman, 1978). A lot of methods were suggested and investigated to select the bandwidth. But there has been no commonly acceptable method up to now. In this study, the most commonly used methods will be investigated, compared and some applications will be presented.

2. Bandwidth selection methods

To examine the performance of kernel estimator, several criterions related to the deviation of \( \hat{f} \) from the real probability density function \( f \) were considered. Commonly used one of these criterions was suggested by Rosenblatt and it is known as the mean integrated squared error (MISE). MISE is a preferable criterion because it is mathematically simple. It is defined as follows:

\[
MSE\left\{\hat{f}(x, h)\right\} = \int_{-\infty}^{\infty} E\left\{\hat{f}(x, h) - f(x)\right\}^2 \, dx
\]

\[= (nh)^{-1} \int_{-\infty}^{\infty} K(u)^2 du + \frac{1}{4} h^4 \left[ \int_{-\infty}^{\infty} u^2 K(u) du \right]^2 \int_{-\infty}^{\infty} f^4(x)^2 dx + o\left\{(nh)^{-1} + h^4\right\} \quad (2)\]
BANDWIDTH SELECTION FOR KERNEL ESTIMATION

And, the asymptotic mean integrated squared error (AMISE) is defined as follows (Wand and Jones, 1995):

\[
AMISE \left\{ \hat{f}(x, h) \right\} = (nh)^{-1} \int_{-\infty}^{\infty} K(u)^2 du + \frac{1}{4} h^4 \left\{ \int_{-\infty}^{\infty} u^2 K(u) du \right\}^2 \int_{-\infty}^{\infty} f''(x)^2 dx. \tag{3}
\]

The most appropriate way to select bandwidth is to find \( h \) value which minimizes MISE. Optimal \( h \) value obtained from (2) is given as, \( h_{opt} \),

\[
h_{opt} \approx \left[ \int_{-\infty}^{\infty} u^2 K(u) du \right]^{-2/5} \left\{ \int_{-\infty}^{\infty} K(u)^2 du \right\}^{1/5} \left\{ \int_{-\infty}^{\infty} f''(x)^2 dx \right\}^{-1/5} n^{-1/5} \tag{4}
\]

As it can be seen from this equation, the optimal \( h \) value depends on the second derivative of the unknown density function \( f \). For this reason different methods are suggested to select \( h \) value. Some of these methods will be given in the following sections.

2.1. Least squares cross-validation

As a data method, the least-squares cross-validation (LSCV) was suggested by Rudemo (1982) and Bowman (1984), independently each other. Let \( \hat{f} \) be a kernel estimator of a probability density function \( f \). Then MISE can be written as follows:

\[
MSE \left\{ \hat{f}(x, h) \right\} = E \int \hat{f}(x, h)^2 dx - 2E \int \hat{f}(x, h)f(x)dx + \int f(x)^2 dx. \tag{5}
\]

Here the aim is to find \( h \) value which minimizes MISE. Selection of \( h \) value which minimizes MISE is equivalent to the \( h \) value which minimizes the following expression

\[
MSE \left\{ \hat{f}(x, h) \right\} - \int f(x)^2 dx = E \int \hat{f}(x, h)^2 dx - 2E \int \hat{f}(x, h)f(x)dx. \tag{6}
\]

The right-hand side of the (6) depends on \( f \), it is not known. But it can be shown that an unbiased estimator for the right-hand side is,

\[
LSCV(h) = \int \hat{f}(x, h)^2 dx - 2n^{-1} \sum_{i=1}^{n} \hat{f}_{-i}(X_i, h). \tag{7}
\]

Here

\[
\hat{f}_{-i}(X_i, h) = \frac{1}{(n-1)h} \sum_{j \neq i} K \left( \frac{X_i - X_j}{h} \right)
\]

and it is a kernel estimate obtained by using all observations with \( X_i \) deleted. This is the reason for the term “cross-validation” which refers to the use of part of a sample to obtain information about another part. (7) is called least-squares cross-validation function and because of (7) gives an unbiased estimator of (6) it is also called unbiased cross-validation function (Cula, 1998).
2.3. The bootstrap

The bootstrap (B) is based on the following base: Main simple was assumed to be a population. Several samples were drawn from main sample with replacement. For each of the drawn sample, related estimators are calculated. If the bootstrap is used to estimate MISE, then bias component can not be calculated. For this reason, the procedure to choice of the bandwidth \( h \) is different than the ordinary bootstrap and it is called as smoothed bootstrap. In this method the bandwidth \( h \) for sample \( \{X_i\} \) by using one of the preceding methods used to find the kernel estimate \( \hat{f} \). The bootstrap sample \( \{X_i^*\} \) is chosen from density \( \hat{f} \) using the following algorithm (Faraway and Jhun, 1990).

Step 1: An integer \( j \) is chosen from \( \{1, 2, \ldots, n\} \) with equal probability.

Step 2: A random variable \( \Phi \) is derived from the probability density function \( K \).

Step 3: Set \( X_j^* \rightarrow X_j^* + h \Phi \)

Then many samples are obtained by repeating the preceding procedure and taking expected value MISE is calculated. Taylor said that if the standard probability density function is taken as the kernel function, the for bootstrap estimate of MISE it is not necessary to take a sample over again. If the standard normal probability density function is used for the kernel function, then bootstrap estimate of MISE is,

\[
B(h) = \frac{1}{2n^2 h \sqrt{2\pi}} \left[ \sum_{i,j} \exp \left\{ -\frac{(x_j - x_i)^2}{8h^2} \right\} - \frac{4}{\sqrt{3}} \sum_{i,j} \exp \left\{ -\frac{(x_j - x_i)^2}{6h^2} \right\} 
+ \sqrt{2} \sum_{i,j} \exp \left\{ -\frac{(x_j - x_i)^2}{4h^2} \right\} + n \sqrt{2} \right] \tag{10}
\]

(Taylor, 1989). It can be seen that if observed values are replaced in (10), then only a bootstrap function depending on \( h \) is obtained. The bandwidth which minimizes this function is found and this bandwidth is shown as \( h_B \).

2.4. Comparison of bandwidth selection methods

In the theoretical point of view, various estimators are compared according to the rate of convergence of some non-random error criterion, such as MISE, to zero. The concept of rate of convergence is an asymptotic concept. For this reason, the concept of rate of convergence is used for large sample sizes. For small samples, comparisons
The basic principle of least-squares cross-validation is to find the kernel estimates from the data for various $h$ values and to select $h$ value which minimizes (7). Bandwidth obtained by this strategy will be shown as $\hat{h}_{LSCV}$.

The least-squares cross-validation function can have more than one local minimum. Researchers show that in this case it is appropriate to take the largest local minimizer of LSCV. Because the largest local minimizer is the nearest bandwidth to the optimal bandwidth obtained from MISE (Wand and Jones, 1995).

### 2.2. Biased cross-validation

AMISE, which is a simple formula with respect to MISE, also depend on $\int_{-\infty}^{\infty} f'(x)^2 dx$ like MISE. Scott, Tapia and Thompson have taken for $\int_{-\infty}^{\infty} f'(x)^2 dx$ the integral of $\int_{-\infty}^{\infty} f'(x)^2 dx$ by using kernel estimator $\hat{f}$ in order to obtain the bandwidth. Scott and Terrell said that this estimator is deficient asymptotically and it is appropriate to use

$$\int_{-\infty}^{\infty} \hat{f}'(x)^2 dx = \int_{-\infty}^{\infty} f'(x)^2 dx - \frac{1}{nh^5} \int_{-\infty}^{\infty} K'(x)^2 dx$$

(8)

for the estimate of $\int_{-\infty}^{\infty} f'(x)^2 dx$ (Scott and Terrell, 1987). (8) is the adjusted value of $\int_{-\infty}^{\infty} \hat{f}'(x)^2 dx$. Then biased cross-validation (BCV) function is obtained by substituting (8) into the asymptotic expression and is given as follows:

$$BCV(h) = \frac{1}{nh} \int_{-\infty}^{\infty} K(u)^2 du + \frac{1}{4} h^4 \left[ \int_{-\infty}^{\infty} u^2 K(u) du \right]^2 \int_{-\infty}^{\infty} \hat{f}'(x)^2 dx$$

(9)

In this study, the bandwidth value which minimizes the function given by (9) will be shown as $\hat{h}_{BCV}$.  

110
are made with respect to simulation studies. The commonly used method to choose a bandwidth was LSCV criterion between the years 1982-1988. General performance of this method is not satisfactory very well. The rate of convergence of the estimator of LSCV is $O(n^{-1/10})$ (Wand and Jones, 1995). To obtain optimal bandwidth, very large sample size is needed (Chiu, 1992).

An unpleasant aspect of LSCV and BCV, which has been noticed in simulation studies and in applications to real data sets, is that the LSCV function and BCV function often have more than one local minima. BCV selection method has also the same rate of convergence of order $n^{-1/10}$ like LSCV method (Hall and Marron, 1991). The studies show that the bandwidth which is obtained by BCV is larger than the bandwidth which is obtained by LSCV. One advantage of BCV method is to have small variance with respect to LSCV. Scott and Terrell showed that an attractive property of $h_{BCV}$ with respect to $h_{LSCV}$ is that $h_{BCV}$ has minimum asymptotic variance. In this case, $h_{BCV}$ is more stable than $h_{LSCV}$ (Chao et al., 1994). From the simulation studies, it was seen that these selectors are to be directed to select small bandwidth with respect to asymptotic theorems. Density estimate which is obtained by using small bandwidths show spurious structure (Chiu, 1991). Researchers show that these selectors have an unsatisfactory performance with theoretical and practical points of view. Performance of B method is better than the performance of LSCV and BCV for some distributions because of the standard deviation of bootstrap’s bandwidth selectors is small (Faraway and Jhun, 1990).

The bandwidth value which is obtained by B is larger than the bandwidth values which are obtained by LSCV and BCV. The bootstrap bandwidth has smaller variance but computational cost of bootstrap is higher than LSCV and BCV criterion.

3. Application

The optimal bandwidth, $h_{opt}$, which minimizes (4) is

$$h_{opt} \equiv \left[ \int_{-\infty}^{\infty} u^2 K(u) du \right]^{-2/5} \left\{ \int_{-\infty}^{\infty} K(u)^2 du \right\}^{1/5} \left( \int f^4(x)^{2} dx \right)^{-1/5} n^{-1/5}$$

If it is assumed that probability density function is known as a normal distribution with sample mean 100 and variance 4 and kernel function is assumed as a standard
normal distribution, then optimal bandwidths for sample sizes \( n = 50, 100, 250 \) and 500 can be obtained as follows by using the formula above:

\[
\begin{align*}
h_{opt} &= 0.969240 \quad \text{for } n = 50 \\
h_{opt} &= 0.843773 \quad \text{for } n = 100 \\
h_{opt} &= 0.702486 \quad \text{for } n = 250 \\
h_{opt} &= 0.611549 \quad \text{for } n = 500
\end{align*}
\]

For example, if the probability density function is an exponential distribution with parameter \( \lambda = 1 \), then for \( n = 50, 100, 250 \) and 500 optimal bandwidths will be obtained as follows:

\[
\begin{align*}
h_{opt} &= 0.407868 \quad \text{for } n = 50 \\
h_{opt} &= 0.355070 \quad \text{for } n = 100 \\
h_{opt} &= 0.295615 \quad \text{for } n = 250 \\
h_{opt} &= 0.257348 \quad \text{for } n = 500
\end{align*}
\]

In fact, the kernel estimation was used when the sample was taken from an unknown distribution. In this study, samples with several sizes were drawn from known distribution and optimal bandwidths were obtained using (4). The bandwidths which were obtained by using the methods of LSCV, BCV, B were both compared with each other and investigated the closeness of them to optimal bandwidth.

First 350 samples were drawn from the normal distribution with mean 100 and variance 4 for each sizes \( n = 50, 100, 250 \) and 500. By taking standard normal distribution as kernel function, for each 350 samples with \( n = 50 \) bandwidths were obtained by using the methods of cross-validation, biased cross-validation and bootstrap and their distributions were found. These procedures were repeated for \( n = 100, 250 \) and 500. The distributions of bandwidth are given in the following Figure 3.1.
Ö. TOKTAMIŞ, S. G. CULA AND S. KURT

\[ n = 50 \]

![Graph with lines labeled LSCV, BCV, and B.]

\[ n = 100 \]
Figure 3.1. The distributions of bandwidths for each 350 samples from the normal distribution with sample sizes $n = 50, 100, 250,$ and $500$ by using the methods of LSCV, BCV, and B.

For each bandwidth distribution, mean and variance are obtained. These values are given in following Table 3.1.
Table 3.1. Means and variances of the bandwidth distributions which are found for samples from normal distribution with various methods.

<table>
<thead>
<tr>
<th></th>
<th>LSCV</th>
<th>BCV</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>n=50</td>
<td>$\hat{h} = 1.821407$</td>
<td>$\hat{h} = 1.943505$</td>
<td>$\hat{h} = 3.529612$</td>
</tr>
<tr>
<td></td>
<td>$Var(\hat{h}) = 0.649172$</td>
<td>$Var(\hat{h}) = 0.721859$</td>
<td>$Var(\hat{h}) = 0.127460$</td>
</tr>
<tr>
<td>n=100</td>
<td>$\hat{h} = 1.328485$</td>
<td>$\hat{h} = 1.765337$</td>
<td>$\hat{h} = 2.981763$</td>
</tr>
<tr>
<td></td>
<td>$Var(\hat{h}) = 0.282590$</td>
<td>$Var(\hat{h}) = 0.495502$</td>
<td>$Var(\hat{h}) = 0.383692$</td>
</tr>
<tr>
<td>n=250</td>
<td>$\hat{h} = 0.822881$</td>
<td>$\hat{h} = 1.387032$</td>
<td>$\hat{h} = 2.430861$</td>
</tr>
<tr>
<td></td>
<td>$Var(\hat{h}) = 0.071172$</td>
<td>$Var(\hat{h}) = 0.376796$</td>
<td>$Var(\hat{h}) = 0.06767$</td>
</tr>
<tr>
<td>n=500</td>
<td>$\hat{h} = 0.623642$</td>
<td>$\hat{h} = 0.953251$</td>
<td>$\hat{h} = 1.837037$</td>
</tr>
<tr>
<td></td>
<td>$Var(\hat{h}) = 0.02995$</td>
<td>$Var(\hat{h}) = 0.106302$</td>
<td>$Var(\hat{h}) = 0.04729$</td>
</tr>
</tbody>
</table>

From these figures, it can be seen that when sample size increases, then the bandwidths which are obtained from each of 350 samples have more smooth distribution. It is seen that bandwidths which are obtained by using BCV criterion are greater than the bandwidths which are obtained by using LSCV criterion and the bandwidths which are obtained by using B criterion are greater than the bandwidth which are obtained by using BCV criterion. From these figures and table it can be seen that the variance of the bandwidths of B is small. This shows that the estimation is more stable. When the sample size is increasing, then the variance of the bandwidth is decreasing for all methods. But when $n = 50$, in samples out of 350 samples bandwidth can not be obtained with bootstrap method. For the other samples, bandwidth can not be obtained with bootstrap method. For this reason when $n = 50$, the variance of the distribution of bandwidths is smaller than others in the table and this is a fallacy.
BANDWIDTH SELECTION FOR KERNEL ESTIMATION

As seen from the table the method which gives the faraway bandwidth from the optimal bandwidth is the bootstrap method. For example, for $n = 100$, $h_{opt} = 0.843773$, $h_{LSCV} = 1.328485$, $h_{BCV} = 1.765337$, and $h_B = 2.981763$. When the sample size is small, then the variance of LSCV selectors is greater than the variance of BCV estimators. But when the sample size is increasing LSCV selectors is closer to the optimal bandwidth and have small variance with respect to BCV selectors.

If the probability density function shows a symmetric and smooth distribution as in the Figure 3.1, it can be said that LSCV's bandwidth in large sample is closer to the optimal bandwidth.

Secondly, 250 samples were drawn from a non-symmetric distribution with parameter $\lambda = 1$ for each sample sizes $n = 50, 100, 250$, and 500. By taking standard normal distribution as kernel function, for each 250 samples by using the methods of cross-validation, biased cross validation, and bootstrap, the bandwidths were obtained and their distributions were found. These distributions are given in the following Table 3.2.

$n = 50$
$n = 100$

$n = 250$
Figure 3.2. The distributions of bandwidths for each 250 random samples from the exponential distribution with sample sizes $n = 50, 100, 250, \text{ and } 500$ by using the methods of CV, BCV, and B.

For each bandwidth distribution the mean and variances are obtained. These values are given in Table 3.2.

Table 3.2. : Means and variances of the bandwidth distributions which are found for samples from exponential distribution

<table>
<thead>
<tr>
<th>$n$</th>
<th>LSCV</th>
<th>BCV</th>
<th>B</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>$\hat{h} = 0.18028$</td>
<td>$\hat{h} = 0.313036$</td>
<td>$\hat{h} = 0.705556$</td>
</tr>
<tr>
<td></td>
<td>$\text{Var}(\hat{h}) = 0.006111$</td>
<td>$\text{Var}(\hat{h}) = 0.029497$</td>
<td>$\text{Var}(\hat{h}) = 0.047312$</td>
</tr>
<tr>
<td>100</td>
<td>$\hat{h} = 0.14644$</td>
<td>$\hat{h} = 0.210843$</td>
<td>$\hat{h} = 0.539259$</td>
</tr>
<tr>
<td></td>
<td>$\text{Var}(\hat{h}) = 0.002624$</td>
<td>$\text{Var}(\hat{h}) = 0.008596$</td>
<td>$\text{Var}(\hat{h}) = 0.036186$</td>
</tr>
<tr>
<td>250</td>
<td>$\hat{h} = 0.10672$</td>
<td>$\hat{h} = 0.130600$</td>
<td>$\hat{h} = 0.953080$</td>
</tr>
<tr>
<td></td>
<td>$\text{Var}(\hat{h}) = 0.000632$</td>
<td>$\text{Var}(\hat{h}) = 0.001119$</td>
<td>$\text{Var}(\hat{h}) = 0.022534$</td>
</tr>
<tr>
<td>500</td>
<td>$\hat{h} = 0.0820$</td>
<td>$\hat{h} = 0.10060$</td>
<td>$\hat{h} = 1.03492$</td>
</tr>
<tr>
<td></td>
<td>$\text{Var}(\hat{h}) = 0.000255$</td>
<td>$\text{Var}(\hat{h}) = 0.000712$</td>
<td>$\text{Var}(\hat{h}) = 0.014567$</td>
</tr>
</tbody>
</table>
In this case it can be seen that biased cross-validation’s bandwidths are larger than cross-validation’s bandwidths, and bootstrap’s bandwidths are larger than biased cross-validation’s bandwidths. From the table and figures, we can see that the variance of cross-validation’s bandwidth are smaller than the variances of the other method’s bandwidth distributions. But the method which gives the nearest bandwidth to the optimal bandwidth is biased cross-validation method. For example, for $n = 100$, $\hat{h}_{opt} = 0.35507$, $\hat{h}_{LSCV} = 0.14644$, $\hat{h}_{BCV} = 0.210843$, and $\hat{h}_B = 0.539259$.

If the probability density function shows a non-symmetric distribution, for large samples BCV method’s bandwidth gets closer to the optimal bandwidth.

References


ÖZET

Bu çalışmada olasılık yoğunluk fonksiyonlarının çekirdek tahmin edicileri için bant genişliği seçimine yönelik yöntemler ele alınmıştır. En küçük kareler, çapraz doğrulama ve bootstrap yöntemleri tanıtılmış ve karşılaştırılarak uygulamalar yapılmıştır.