A Bandwidth Selection for Kernel Density Estimation of Functions of Random Variables

A. R. Mugdadi
Department of Mathematics
Southern Illinois University
Carbondale, IL 62901, U.S.A.
amugdadi@math.siu.edu

and

Ibrahim A. Ahmad
Department of Statistics and Actuarial Sciences
University of Central Florida
Orlando, FL 32816, U.S.A.
iahmad@mail.ucf.edu

Abstract

In this investigation, the problem of estimating the probability density function of a function of \( m \) independent identically distributed random variables, \( g(X_1, X_2, ..., X_m) \), is considered. The choice of the bandwidth in the kernel density estimation is very important. Several approaches are known for the choice of bandwidth in the kernel smoothing methods for the case \( m = 1 \) and \( g \) is the identity. In this study we will derive the bandwidth using the least square cross validation and the contrast methods. We will compare between the two methods using Monte Carlo simulation and using an example from the real life.

Keywords: Density estimation, function of random variables, bandwidth, kernel contrast.

1 INTRODUCTION

The kernel estimation method is an important tool in nonparametric density and distribution functions fitting. Suppose that a data set \( X_1, X_2, ..., X_n \), denotes a random sample from an unknown probability density function (pdf) \( f(x) \), then the kernel density estimator of \( f(x) \) is defined by

\[
\hat{f}(x) = \frac{1}{nb_n} \sum_{i=1}^{n} k\left(\frac{x - X_i}{b_n}\right),
\]

(1.1)
where $k$ is a bounded nonnegative function satisfying $\int k(x)dx = 1$ and $b_n$ is a sequence of positive number usually called the bandwidth.

Consider the function $g(X_1, X_2, ..., X_m)$ that depends on $m \geq 1$ observations. The distribution function of $g(X_1, X_2, ..., X_m)$ is defined by

$$H(t) = P(g(X_1, X_2, ..., X_m) \leq t),$$

where $t \in R$. A function closely related to the distribution function is the density function, defined by $h(t) = H'(t)$, when it exists. The nonparametric kernel estimate of $h(t)$ is easily seen to be (cf. Frees (1994)):

$$\hat{h}(t) = \frac{1}{b_n} \sum_{(n,m)} w(t - g(X_{i_1}, ..., X_{i_m})/b),$$  \hspace{1cm} (1.2)

where $b = b_n$ is the bandwidth, $1 \leq i_1 < i_2 < ... < i_m \leq n$ is an ordered subset of $1, 2, ..., n$, $\sum_{(n,m)}$ denotes summation over all $\binom{n}{m}$ subsets and $w(.)$ is a kernel function. It is clear that if $m = 1$ and $g(x) = x$ then the estimator $\hat{h}(t)$ reduces to the estimator $\hat{f}(x)$.

The function of $m$ identical random variables $g(X_1, X_2, ..., X_m)$ have applications in real life such as the case where $g(X_1, X_2, ..., X_m) = \sum_{i=1}^{m} X_i$ which identifies in actuarial science the total claims of, for example, individual insurances. It also identifies, in life testing, the life of a parallel system of $m$ identical components. Other applications are indicated in Frees(1994), Ahmad and Fan (2001) and Ahmad and Mugdadi (2003 b). Also, the authors use the estimate (1.2) to derive a kernel based testing procedure for normality, cf. Ahmad and Mugdadi (2003 a).

Some asymptotic properties of the estimate (1.2) are known, for example, Frees(1994) discusses the consistency and the asymptotic normality of $\hat{h}(t)$. Precisely, he shows that under some conditions and if $b_n \rightarrow 0$ such that the bias $B_n(t) = Eh(t) - h(t) \rightarrow 0$ as $n \rightarrow \infty$, then $\hat{h}(t)$ is consistent estimate of $h(t)$. In addition Frees(1994) proves that under some conditions $n^{1/2}(\hat{h}(t) - h(t))$ is asymptotically normal with mean $B_n$ and variance $C$, a positive constant. The above work demonstrated that the asymptotic behavior of $\hat{h}(t)$ is different from that of $\hat{f}(x)$ and utilizes new methodologies. Ahmad and Fan (2001) obtain the optimal theoretical bandwidth $b$ in this general case. This work is expanded by Ahmad and Mugdadi (2003 b). Thus we are motivated to further the above work and discuss data-based choices of the bandwidth.

Kernel smoothing provides a simple way of finding structures in data sets without knowing the probability density function. Theoretical and simulation analysis have shown that the choice of the kernel is not crucial for density estimation in the case of independent identically distributed (i.i.d.) random variables. The most important part in the kernel estimation method is to select the bandwidth. There are many situations where it is satisfactory to select the bandwidth by graphing many densities using different bandwidths, then select the most acceptable density. One of the strategies to do that by starting with small (or large) bandwidth then going up (or down) until you reach the suitable one.

Meanwhile, when we select the bandwidth, we need to consider the error in our selection. There are many methods to calculate the error such as the Mean Square Error
(MSE) and the Mean Integrated Square Error (MISE) where
\[
MSE(\hat{h}(t)) = E[\hat{h}(t) - h(t)]^2,
\]
and the
\[
MISE(\hat{h}(t)) = E \int [\hat{h}(t) - h(t)]^2 dt
\]

Also, there are many methods to select the bandwidth for the case of \( \hat{f}(x) \). Most of these methods are based on minimizing the MSE or the MISE. The following methods are used to select the bandwidth for \( \hat{f}(x) \):
1- Least Squares Cross-Validation (see Rudemo (1982) and Bowman (1984))
2- Biased Cross-Validation (see Scott and Terrell (1987))
3- Plug-in Bandwidth Selection (see Sheather and Jones (1991))
4- Smoothed Cross-Validation (see Hall, Marron and Park (1992))
5- Root-n Bandwidth Selection (see Hall, Sheather, Jones and Marron (1991))
6- The Contrast Method (see Ahmad and Ran (2003))

Jones, Marron and Sheather (1996) compared among the first generation methods (Rules of thumb, Least squares cross validation and biased cross validation) and the second generation methods (the plug-in approach and smoothed bootstrap) using real data examples, asymptotic analysis and simulation.

On the other hand, Ahmad and Fan (2001) derived the optimal bandwidth that minimizes the AMISE of \( \hat{h}(t) \) by:
\[
b_{opt} = \left( \frac{R(w)}{(\mu_2(w))^{2/3}} \right)^{1/5} \left( \frac{n}{m} \right) \frac{1}{5}
\]
where, \( R(f) = \int f^2(x) dx \) and \( \mu_2(w) = \int x^2 w(x) dx \), but \( b_{opt} \) depends as usual on the unknown density function \( h(t) \).

2 Least Square Cross Validation for \( \hat{h}(t) \).

The motivation of the least square cross validation (LSCV) comes from expanding the MISE of \( \hat{h}(t,b) \) to obtain
\[
MISE(\hat{h}(t,b)) = E \left( \int (\hat{h}(t,b))^2 dt \right) - 2E \left( \int (\hat{h}(t,b)h(t)) dt \right) + \int (h(t))^2 dt
\]
but, \( \int (h(t))^2 dt \) does not depend on \( b \). Therefore, minimization of \( MISE(\hat{h}(b)) \) is equivalent to minimization of
\[
MISE(\hat{h}(t,b)) - \int (h(t))^2 dt = E \left( \int (\hat{h}(t,b))^2 dt - 2 \int (\hat{h}(t,b)h(t)) dt \right).
\]
The LSCV is based on abstaining an unbiased estimate of (2.2). Let \( m_{(1)} \) be a fixed number equal to \( m \) and define the density estimate of \( g(X_1, X_2, ..., X_m) \) based on the \( \binom{n}{m} \) cases with \( X_1, X_2, ..., X_{m_{(1)}} \) deleted by:
\[
\hat{h}_{-m_{(1)}}(t,b) = \frac{1}{(\binom{n}{m} - 1)b} \sum_{(n,m),m\neq m_{(1)}} w(t - g(X_{i_1}, X_{i_2}, ..., X_{i_m})),
\]
where, \( w = \frac{(t - g(X_{i_1}, X_{i_2}, ..., X_{i_m}))}{b} \).
But,
\[ E \sum_{(n,m), m \neq m(1)} w\left(\frac{t - g(X_i, X_{i2}, \ldots, X_{in})}{b}\right) = \left(\binom{n}{m} - 1\right) E\left(\frac{t - g(X_1, X_2, \ldots, X_m)}{b}\right) \] (2.4)
and
\[ E(w\left(\frac{t - g(X_1, X_2, \ldots, X_m)}{b}\right)) = \int w\left(\frac{t - g(x_1, x_2, \ldots, x_m)}{b}\right) dH(g(x_1, x_2, \ldots, x_m)). \] (2.5)

Therefore,
\[ E(\hat{h}_{m(1)}(t, b)) = \frac{1}{b} \int w\left(\frac{t - g(x_1, x_2, \ldots, x_m)}{b}\right) dH(g(x_1, x_2, \ldots, x_m)). \] (2.6)

Thus, \( \frac{1}{\binom{n}{m}} \sum_{(n,m(1))} \hat{h}_{m(1)}(g(x_1, x_2, \ldots, x_{m(1)}), b) \) is an unbiased estimator of \( f(h(t), h(t)) dt \).

From the above equations, we can conclude that
\[ \text{LSCV}(b) = \int (\hat{h}(t, b))^2 dt - \frac{2}{\binom{n}{m}} \sum_{(n,m(1))} \hat{h}_{m(1)}(g(X_i, X_{i2}, \ldots, X_{in(1)}), b) \] (2.7)
is an unbiased estimator of
\[ E(\int (\hat{h}(t, b))^2 dt - 2 \int (\hat{h}(t, b)h(t)) dt) \] (2.8)

Thus, it is reasonable to choose \( b \) that minimizes \( \text{LSCV}(b) \).

3 The Contrast Method for \( \hat{h}(t) \).

The contrast method for the kernel density estimator \( \hat{f}(x) \) proposed by Ahmad and Ran (2003). Hence we extend it to the case of the estimator \( \hat{h}(t) \). The first step in the contrast method, we define the kernel density estimations \( \hat{h}_j(t, b) \) based on \( q \) kernels, \( w_1, \ldots, w_q \).

Thus
\[ \hat{h}_j(t, b) = \hat{h}(t, b) = \frac{1}{b \binom{n}{m}} \sum_{(n,m)} w_j \left(\frac{t - g(X_{i1}, \ldots, X_{in})}{b}\right). \] (3.1)

After choosing the contrast coefficients \( p_1, \ldots, p_q \), where \( \sum_{j=1}^q p_j = 0 \), select the bandwidth that minimizing the \( \text{MISE}_{\text{cont}} \), where
\[ \text{MISE}_{\text{cont}}(\hat{h}(t, b)) = E\left[\int (\sum_{j=1}^q p_j \hat{h}_j(t, b) - \sum_{j=1}^q p_j h(t))^2 dx\right], \] (3.2)
but \( \sum_{j=1}^q p_j h(t) = h(t) \sum_{j=1}^q p_j = 0 \) Therefore
\[ \text{MISE}_{\text{cont}} = E\left[\int (\sum_{j=1}^k p_j \hat{h}_j(t, b))^2 dt\right]. \] (3.3)
Thus
\[ ISE_{\text{cont}} = ISE(b)_{\text{cont}} = \int \left( \sum_{j=1}^{k} p_j \hat{h}_j(t,b) \right)^2 dt \]  
(3.4)
is an unbiased estimator of \( MISE_{\text{cont}} \). Therefore a reasonable choice for estimating \( b \)
is to minimize \( ISE_{\text{cont}} \) which does not depend on the unknown density function \( h(t) \).
Ahmad and Mugdadi (2003) proved that the estimator based on the \( ISE_{\text{cont}} \) for \( \hat{h}(t) \) is consistent. Finally, define the density estimation using the kernel contrast approach by:
\[ \hat{h}(t) = \sum_{j}^{q} c_j \hat{h}_j(t,b), \]  
(3.5)
where \( \sum_{j}^{q} c_j = 1 \). We can have an equal weight for the kernels by choosing \( q \) as an even integer, \( p_j = -p_{2j} \), \( j = 1, ..., \frac{q}{2} \) and \( c_j = \frac{1}{q} \), for \( j = 1, ..., q \).

4 Simulation Work
The \( LSCV \) and the contrast are data based bandwidth methods. We can use these two methods with any kernel function \( w_j(x) \) such that \( \int w_j(x) dx = 1 \) and \( w_j(x) \) is a bounded function.
In this section we will compare between the \( LSCV \) and the \( ISE_{\text{cont}} \) methods using Monte Carlo studies and a real data example as well.

4.1 Monte Carlo Studies
During our simulation study we will use the following three kernels
1- The normal kernel with mean 0 and variance \( c \).
2- The Epanechnikov kernel
3- The Biweight kernel

Also, we will simulate from the following populations:
I- The normal distribution with mean 0 and variance 1.
II- The exponential distribution with mean 1.
III- The Cauchy distribution with pdf
\[ f(x) = \frac{1}{\pi (1 + x^2)} \]
To evaluate the performance and to compare between the \( LSCV \) and the \( ISE_{\text{cont}} \) methods we simulate random samples of sizes 25 and 50 from the standard normal distribution and also from the exponential distribution with mean 1. In Figures 1 through 4 we compare between the exact densities and the estimated densities of \( q(X_1, X_2) = X_1 + X_2 \), where \( X_1 \) and \( X_2 \) are random variables having standard normal distribution (Figures 1 and 2) and having exponential distribution with mean 1 (Figures 3 and 4). From these Figures we conclude that both density estimates describe the data precisely, but the one based on the contrast method is smoother.
Tables 1 through 6 give us, in a more precise fashion, the differences between the $ISE_{cont}$ and the $LSCV$ methods.

We compare the variance of $\hat{b}$, when we select $\hat{b}$ minimizing the $ISE(b)_{cont}$, defined in (3.4) with the one by minimizing the $LSCV(b)$ defined in (2.7). Also, we compare the biases, where the biases are calculated by comparison with (theoretically) ”optimal” bandwidth choices of (1.3). We simulated $n$ values from each of the above pdf’s based on a set with 30 $\hat{b}$ values. To calculate $\hat{b}$ using the $ISE_{cont}$ and the $LSCV$ we used the following combinations:

1. $w_1$ and $w_2$ are two kernels distributed $N(0,1)$ and $N(0,4)$ respectively,
2. $w_1$ and $w_2$ are two kernels distributed Epanechnikov and Biweight, respectively,
3. $w$ is a kernel from $N(0,1)$, and
4. $w$ is a kernel from Epanechnikov distribution.

Also, in the contrast method we used $p_1 = -p_2$ and $c_1 = c_2 = \frac{1}{2}$.

Tables 1-6 provide us with more information about the differences between the contrast and the least square cross validation methods to derive the bandwidth. We can conclude from the Tables the following:

1- The variance of $\hat{b}$ based on the contrast method is less than the variance based on the $LSCV$ method when we simulate from normal and when we use two normal kernels(Table 1).

2- In most of the cases the variance of $\hat{b}$ based on the $ISE_{cont}$ method is less than the variance based on the $LSCV$ method when we simulate from normal distribution and
Table 3: Population II: using two normal kernels

<table>
<thead>
<tr>
<th>Population II</th>
<th>$ISE_{\text{cont}}$</th>
<th>LSCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>Var.</td>
<td>Bias</td>
</tr>
<tr>
<td>---</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>25</td>
<td>0.0728</td>
<td>-1.322</td>
</tr>
<tr>
<td>30</td>
<td>0.0729</td>
<td>-1.294</td>
</tr>
<tr>
<td>35</td>
<td>0.0192</td>
<td>-1.155</td>
</tr>
<tr>
<td>40</td>
<td>0.0248</td>
<td>-1.130</td>
</tr>
<tr>
<td>45</td>
<td>0.0519</td>
<td>-1.202</td>
</tr>
<tr>
<td>50</td>
<td>0.0219</td>
<td>-1.223</td>
</tr>
</tbody>
</table>

Table 4: Population II: using Epanechnikov and Biweight kernels

<table>
<thead>
<tr>
<th>Population II</th>
<th>$ISE_{\text{cont}}$</th>
<th>LSCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>Var.</td>
<td>Bias</td>
</tr>
<tr>
<td>---</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>25</td>
<td>0.0787</td>
<td>0.3149</td>
</tr>
<tr>
<td>30</td>
<td>0.0433</td>
<td>0.3916</td>
</tr>
<tr>
<td>35</td>
<td>0.0573</td>
<td>0.3649</td>
</tr>
<tr>
<td>40</td>
<td>0.0203</td>
<td>0.4796</td>
</tr>
<tr>
<td>45</td>
<td>0.0165</td>
<td>0.4669</td>
</tr>
<tr>
<td>50</td>
<td>0.0282</td>
<td>0.4029</td>
</tr>
</tbody>
</table>

Table 5: Using two normal kernels

<table>
<thead>
<tr>
<th>Population III</th>
<th>$ISE_{\text{cont}}$</th>
<th>LSCV</th>
</tr>
</thead>
<tbody>
<tr>
<td>n</td>
<td>Var.</td>
<td>Bias</td>
</tr>
<tr>
<td>---</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>25</td>
<td>0.0120</td>
<td>1.2465</td>
</tr>
<tr>
<td>30</td>
<td>0.0119</td>
<td>1.2105</td>
</tr>
<tr>
<td>35</td>
<td>0.0152</td>
<td>1.1919</td>
</tr>
<tr>
<td>40</td>
<td>0.0070</td>
<td>1.1665</td>
</tr>
<tr>
<td>45</td>
<td>0.0115</td>
<td>1.1912</td>
</tr>
<tr>
<td>50</td>
<td>0.0092</td>
<td>1.1645</td>
</tr>
</tbody>
</table>
we use Epanechnikov and Biweight kernels (Table 2). Also, when we simulate from exponential with two normal kernels (Table 3).

3- The variance of \( \hat{b} \) based on the LSCV method is less than the variance based on the ISE\(_{\text{cont}} \) when we simulate from Cauchy distribution (Table 5 and 6) and when we simulate from exponential and we use Epanechnikov and Biweight kernels for the contrast methods.

4- The absolute value of the bias of \( \hat{b} \) based on the ISE\(_{\text{cont}} \) method is less than the one based on based on the LSCV method when we simulate from normal and when we simulate from exponential (Tables 1, 2, 3 and 4) and when we simulate from Cauchy when we use Epanechnikov and Biweight kernels using the ISE\(_{\text{cont}} \) method (Table 6).

5- The absolute value of the bias of \( \hat{b} \) based on the LSCV method is less than the one using the contrast with two normal kernels.

6- The mean square error using the ISE\(_{\text{cont}} \) is less than the one using the LSCV when we simulate from normal and exponential distributions (Tables 1, 2, 3 and 4), however, the mean square error using the LSCV is less than the one using the ISE\(_{\text{cont}} \) when we simulate from Cauchy (Tables 5 and 6). Also, it is clear that in all the cases the mean square error is decreasing as the sample size is increasing.

Based on these simulation the two methods are provide us different bandwidths as well appropriate choices of the bandwidth for \( \hat{h}(t) \).

### 4.2 Real Data Example

An important measure of the performance of any bandwidth selection method is how well it performs in practice. There are many applications for \( \hat{h}(t) \), cf. Frees (1994) and Ahmad and Fan (2001). During our study we will choose the bandwidth using the contrast method and the least square cross validation method, and we want to check whether our conclusion in the real life example about the smoothness of the density function is consistent with the conclusion in the simulations examples.

Let \( X_1, X_2, \ldots, X_n \) be a random sample of insurance claims, a particular line of business. Table 2.7 represents the claims in thousands of dollars received by one of the Auto Insurance companies in the United States in September 2001.

From the stand point of the insurer, of interest is the distribution of the sum of claims \( X_1, X_2, \ldots, X_m \), interpret \( m \) to be the expected number of claims in a specified
Table 7: The Insurance Data

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>0.981</td>
<td>2.040</td>
<td>8.734</td>
<td>1.340</td>
<td>6.578</td>
</tr>
<tr>
<td>0.503</td>
<td>6.089</td>
<td>2.345</td>
<td>1.110</td>
<td>1.005</td>
</tr>
<tr>
<td>1.134</td>
<td>2.670</td>
<td>0.503</td>
<td>6.089</td>
<td>2.345</td>
</tr>
<tr>
<td>1.110</td>
<td>2.457</td>
<td>1.774</td>
<td>1.005</td>
<td>2.457</td>
</tr>
<tr>
<td>1.134</td>
<td>2.670</td>
<td>0.503</td>
<td>6.089</td>
<td>2.345</td>
</tr>
</tbody>
</table>

financial period, for example a day or month. In this example, we discuss the case
$g(x_1, ..., x_m) = x_1 + ... + x_m$, when $m = 2$. Using the standard normal kernel in the LSCV
the bandwidth $b = 0.29$. In the contrast method the choice of the bandwidth doesn’t
depend on the constants $c_j, j = 1, ..., q$, thus in Figures 6 and 7 the bandwidth is the
same which is $b = 0.42$, while in Figure 8 the two bandwidths are $b = 0.42$ and $b = 0.24$.
In the contrast method, we used the normal kernel with mean 0 and with variances 1, 4, 9,
and 16 when the contrast coefficients are $p_1$, $p_2$, $p_3$ and $p_4$, respectively.

From Figures 5 through 8, we conclude the following:
1- The LSCV do not provide a smooth density estimation, which is consistent with the
case for $m = 1$, see Jones, Marron and Sheather (1996) (Figure 5).
2- When the contrast coefficients are 1/2 and 1/2 we have almost an identical density
estimations with the case of 1/3 and 2/3, but they are close to case 1/10 and 9/10, but
not identical (Figures 6 and 7).
4- The density estimations based on two or four kernels when the kernels have the same
weight is almost identical (Figure 8).
5- The bandwidth when we use four contrast coefficients is smaller than the one using the
LSCV, but because we used the four kernels in density estimation, the contrast method
provides us with a smoother bandwidth.

From Figures 1 through 9, we conclude that the choice of the bandwidth based on the
real data is consistent with the one based on simulations which is the $ISE_{cont}$ method
gives us a smoother density estimation.

References


*Journal of Nonparametric Statistics*, **15**, 273-288

functions of random variables. *Journal of Nonparametric Statistics*, to appear


---

Figure 1: Simulation from normal distribution with $n = 25$, columns: exact distribution, dots: using LSCV, solid: using $ISE_{cont}$
Figure 2: Simulation from normal distributing with $n = 50$, columns: exact distribution, dots: using LSCV, solid: using $ISE_{cont}$

Figure 3: Simulation from exponential distribution with $n = 25$, columns: exact distribution, dots: using LSCV, solid: using $ISE_{cont}$
Figure 4: Simulation from exponential distribution with $n = 25$, columns: exact distribution, dots: using LSCV, solid: using $\text{ISE}_{cont}$

Figure 5: Density estimation for the insurance data using LSCV and
Figure 6: Density estimation for the insurance data using $ISE_{cont}$

***: $p_1 = 1, p_2 = -1, c_1 = c_2 = 1/2$

—: $p_1 = 1, p_2 = -1, c_1 = 1/3, c_2 = 2/3$

Figure 7: Density estimation for the insurance data using $ISE_{cont}$

***: $p_1 = 1, p_2 = -1, c_1 = c_2 = 1/2$

—: $p_1 = 1, p_2 = -1, c_1 = 1/10, c_2 = 9/10$
Figure 8: Density estimation for the insurance data using $ISE_{cont}$

***: $p_1 = 1, p_2 = -1, c_1 = c_2 = 1/2$

−: $p_1 = p_3 = 1, p_2 = p_4 = -1, c_1 = c_2 = c_3 = c_4 = 1/4$