Journal of the Korean
Statistical Society
Vol. 20, No. 1, 1991

On Asymptotically Optimal Plug-in Bandwidth Selectors in Kernel Density Estimation+

Moon Sup Song*, Kyung Ha Seog and Sin sup Cho***

ABSTRACT

Two data-based bandwidth selectors which are optimal in the sense that they achieve $n^{-1/2}$ rate of convergence in kernel density estimation are proposed. The proposed bandwidth selectors are constructed by modifying Park and Marron's plug-in method. The first modification is taking Taylor expansion of the mean integrated squared error to two more terms than in the case of plug-in method. The second is estimating more accurately the functionals of the unknown density appeared in the minimizer of the expansion by using higher order kernels. The proposed bandwidth selectors were proved to be optimal in terms of convergence rate. According to small-sample Monte Carlo studies, the proposed bandwidth selectors showed better performance than all the other bandwidth selectors considered in the simulation.

1. Introduction

The area of nonparameteric density estimation has been remarkably progressed in recent years, both in theoretical and practical aspects. There are various methods available for univariate density estimation. See Silverman(1986) for the detailed description about density estimators. In this paper we are particularly interested in the kernel estimator which is known to be simple, intuitively appealing and best understood. To apply the kernel density estimator in practice, it is well known that the choice of smoothing parameter(bandwidth) is more important than the choice of kernel function. Various bandwidth selection methods have been studied. See Marron(1988) for a listing of bandwidth selelction methods.

+ This research was suported by SNU Daewood Research Fund 89-90.

* Department of Computer Science and Statistics, Seoul National University, Seoul, Korea.

** Department of Statistics, Inje University, Kimhae, Korea.

The plug-in method, which was proved to be a successful data-based bandwidth selector by Park and Marron(1990), has some weak aspects. For example, the existing data-based bandwidth selectors based on plug-in method do not achieve $n^{-1/2}$ rate of convergence. In this paper, we thus want to construct data-based bandwidth selectors which achieve the optimal rate of convergence by modifying the plug-in procedure.

The first modification, which is motivated by Hall and Marron(1989), is done by taking Taylor expansion of the mean integrated squared error to more terms than in the case of Park and Marron's plug-in method, and finding the approximate minimizer of the resulting approximate mean integrated squared error. The second is done by estimating the unknown functionals of the true density appeared in the approximate minimizer by using higher order kernels. The associated bandwidth selector at the stage of estimating the unknown functionals still depends on the underlying density. Two approaches are considered to resolve the dependence. One approach is to replace the underlying density by some given reference density. The other one is to take one more step to estimate the unknown functionals.

An independent study conducted by Hall, Sheather, Jones and Marron(1989) uses almost the same idea and yields the same bandwidth selector as the proposed one with the first approach above.

In Section 2, we introduce the appealing plug-in bandwidth selector. Our proposed bandwidth selectors are proposed in Section 3. Section 4 is devoted to the comparison of the proposed bandwidth selectors with various bandwidth selectors through small sample Monte Carlo studies. The proposed bandwidth selectors show very good behaviors and dominate other bandwidth selectors in all assumed underlying densities.

2. Plug-in Method

2-1. Mathematical Formulation and Notations

Let $X_1, X_2, \cdots X_n$ be a random sample from a probability density f. Then, using this sample, we are interested in estimating f based on kernel functions. The kernel density estimator of f, which was proposed by Rosenblatt(1956) and Parzen(1962), is given by

$$\hat{f}_h(x)=n^{-1}\sum_{i=1}^{n} K_h(x-X_i),$$

where $K_h(x)=K(x/h)/h$, K is called the kernel function and h is called the bandwidth or smoothing parameter. It is well known that the choice of the kernel function K is of essentially negligible concern compared to the choice of bandwidth h (see Chapter 3 of Silverman(1986)). For this reason, we often compare the density estimators through the performance of the corresponding bandwidth selectors. The limiting distributions of data-based bandwidth selectors are useful tools to compare the sample variability and characteristics.

The convergence rate of a bandwidth selector obtained from the limiting distribution is usually determined by the amount of smoothness of density f or by the order of the kernel function K. We thus introduce the definitions of the smoothness order of f and the order of K in the following.

Definition 2.1. (1) Let $\nu=l+\eta$, where l is an integer and $\eta \in (0, 1]$. The density function f is said to have *smoothness of order* ν when the condition that f has a Hölder continuous and square intergrable second derivatives is satisfied and there is a constant $M>0$ so that

$$|f^{(2+l)}(x)-f^{(2+l)}(y)| < M|x-y|^{\eta}, \text{ for all } x \text{ and } y. \tag{2.1}$$

(2) The kernel function K is said to have *order* r when

$$\int x^j K(x)\,dx = \begin{cases} 1, & j=0, \\ 0, & j=1,\cdots,r-1, \\ c, & j=r, (c\neq 0). \end{cases}$$

Widely considered performance measures of $\hat{f}_h$ are

$$ISE(h)=\int (\hat{f}_h - f)^2, \tag{2.2}$$

and

$$MISE(h)=E\int (\hat{f}_h - f)^2. \tag{2.3}$$

For another criterion, see Devroye and Gyorfi(1984). The fact that there is, for all reasonable sample sizes, a considerable difference between h_{ISE}, the minimizer of $ISE(h)$, and h_{MISE}, the minimizer of $MISE(h)$, has been established by Hall and Marron(1987b). In this paper, $MISE(h)$ is used as the performance measure of $\hat{f}_h$, because of its sample stability. Another reason for use of $MISE(h)$ is given in Hall and Marron(1989) who present an importance sense in which $ISE(h)$ is too difficult a goal.

2-2. Plug-in Method

The plug-in bandwidth selector which we review here was developed by Hall(1980) and Sheather (1983, 1986), and quantified asymptotically by Park and Marron(1990).

The asymptotic representation of $MISE(h)$ in (2.3) can be written as

$$AMISE(h)=(nh)^{-1}R(K)+h^4\mu_2^2 R(f'')/4, \tag{2.4}$$

where $R(K)$ and μ_j are defined as follows.

$$R(K)=\int K^2(x)\,dx \;; \qquad \mu_j=\int x^j K(x)\,dx.$$

The minimizer of $AMISE(h)$, h_{AMISE}, can be represented as

$$h_{AMISE}=\left\{\frac{R(K)}{R(f'')\,\mu_2^2}\right\}^{1/5} n^{-1/5}. \tag{2.5}$$

The idea is to replace the only unknown factor $R(f'')$ in h_{AMISE} by some suitable estimator. The estimator considered is

$$\hat{R}_a(f'')=R(\hat{f}'')-n^{-1}a^{-5}R(K''), \tag{2.6}$$

where a is a bandwidth differing from h. The minimizer of approximate mean squared error of $\hat{R}_a(f'')$, a_{AMSE}, have the relationship to h_{AMISE} such as

$$a_{AMSE}=C_1(K)\,C_2(f)\,h_{AMISE}^{10/13} \tag{2.7}$$

where

$$C_1(K)=\{18R(K^{(4)})\,\mu_2^4/4\mu_2^2 R(K)^2\}^{1/13},$$

$$C_2(f)=\{R(f)R(f'')^2/R(f^{(3)})^2\}^{1/13}.$$

Since the scale of f, denoted by λ, is considered to be more crucial than f itself on the determina-

tion of the value of $C_2(f)$, it seems to be enough to replace f in $C_2(f)$ by g_λ, where $g_\lambda(x)=g_1(x/\lambda)/\lambda$ and g_1 is any fixed probability density with unit scale factor. Since $C_2(g_\lambda)=\lambda^{3/13}C_2(g_1)$, the relation (2.7) motivates the equation

$$a_\lambda(h)=C_1(K)\,C_2(g_1)\,\lambda^{3/13}h^{10/13}. \tag{2.8}$$

Using the equations (2.5), (2.6) and (2.8), the plug-in bandwidth selector, $\hat{h}_{PI}$, is defined to be the solution of the equation

$$h=\left\{\frac{R(K)}{\mu_2^2\hat{R}_{a\hat{\lambda}(h)}(f'')}\right\}^{1/5}n^{-1/5},$$

where $\hat{\lambda}$ denotes a $\sqrt{n}$ - consistent estimator of λ (e.g. standard deviation, interquartile range).

According to the results of Park and Marron(1990) and Park(1989), under a strong assumption about the underlying smoothness ($\nu>1$), the rate of relative convergence of $\hat{h}_{PI}$ to h_{MISE} is $n^{-4/13}$. Thus, the sample variability of $\hat{h}_{PI}$ decreases much faster than that of $\hat{h}_{UCV}$ (the unbiased cross-validated bandwidth selector due to Rudemo(1982) or Bowman(1984)) or $\hat{h}_{BCV}$(the biased cross-validated bandwidth selector due to Scott and Terrel (1987)) whose convergence rate is $n^{-1/10}$. Note that the efficiency of $\hat{h}_{PI}$ depends on the underlying smoothness as the case of $\hat{h}_{BCV}$. If the underlying distribution is not smooth enough, $\hat{h}_{PI}$ has more sample variability than that of $\hat{h}_{UCV}$.

3. The Proposed Bandwidth Selectors

3-1. The Proposed Bandwidth Selectors

As discussed in Section 2.2, the plug-in method developed by Hall(1980) and Sheather(1983, 1986) represents a as a function of h and solve the resulting nonlinear equation. This approach, so called *solved equation* version of plug-in, makes the problem of bandwidth selection complicated to solve. Thus it seems natural to consider an alternative approach in which instead of representing a as a function of h, we plug $\hat{R}(f'')$ into $R(f'')$ directly. We may call this approach as a *direct plug-in method* and let $\hat{h}_{PID}$ be the resulting bandwidth selector. Note that this direct plug-in method gives the same asymptotic distribution as that of $\hat{h}_{PI}$. Accordingly, the direct plug-in method may be better in the sense of computation.

The direct plug-in bandwidth selector, $\hat{h}_{PID}$, is

$$\hat{h}_{PID}=\left\{\frac{R(K)}{\hat{R}(f'')\mu_2^2}\right\}^{1/5}n^{-1/5}, \tag{3.1}$$

where $\hat{R}(f'')$ is an estimator of $R(f'')$. Hall and Marron(1987a) discussed slightly more general problem of estimating $R(f^{(m)})$. The estimators considered are

$$\hat{R}_a^1(f^{(m)})=(-1)^m n^{-1}(n-1)^{-1}\,a^{-2m-1}\sum_{i\neq j}U^{(2m)}*U\{(X_i-X_j)/a\},$$

and

$$\hat{R}_a^2(f^{(m)})=(-1)^m n^{-1}(n-1)^{-1}\,a^{-2m-1}\sum_{i,j}U^{(2m)}\{(X_i-X_j)/a\},$$

where U is a kernel function of order r and a is a bandwidth. For simplicty of presentation, we let $\hat{R}_a(f^{(m)})$ denote either $\hat{R}_a^1(f^{(m)})$ or $\hat{R}_a^2(f^{(m)})$ and assume that

A1. U is symmetric and has $2m$ derivatives which vanish at $\pm\infty$.

According to Hall and Marron(1987a), we can construct a $\sqrt{n}$-consistent estimator of $R(f'')$ by using higher order kernel. Accordingly, since the other factors in (3.1) depend only on K and sample size, we can obtain a bandwidth selector $\hat{h}_{PID}$ whose relative rate of convergence to h_{AMISE} is $n^{-1/2}$. However, note that the $\sqrt{n}$-consistency is to h_{AMISE} not to h_{MISE}. The rate of relative convergence of h_{MISE} is given by Park(1989) as follows.

$$h_{AMISE}/h_{MISE}-1=\begin{cases} O(n^{-2\nu/5}), & \text{if } 0<\nu\leq 1 \\ O(n^{-2/5}), & \text{if } \nu>1 \end{cases}$$

Thus the resultant estimator $\hat{h}_{PID}$ can not have faster rate than $n^{-2/5}$ to h_{MISE}.

Bacause of such defect, Hall and Marron(1989) discussed an improved h_{AMISE}, which is based on

$$\begin{aligned} AMISE^*(h) &= (nh)^{-1}R(K)+h^4R(f'')\,\mu_2^2/4-h^6\,\mu_2\,\mu_4\,R(f^{(3)})/24-n^{-1}R(f) \\ &= AMISE(h)-h^6\,\mu_2\,\mu_4\,R(f^{(3)})/24-n^{-1}R(f) \end{aligned} \tag{3.2}$$

The minimizer of $AMISE^*(h)$, h_{AMISE}^*, can be represented bv

$$\begin{aligned} h_{AMISE}^* &= h_{AMISE}+h_{AMISE}^3\,\mu_4\,R(f^{(3)})/\{20\mu_2-R(f'')\}+O(n^{-1}) \\ &= \left\{\frac{R(K)}{\mu_2^2R(f'')}\right\}^{1/5}n^{-1/5}+\left\{\left[\frac{R(K)}{\mu_2^2R(f'')}\right]^{1/5}n^{-1/5}\right\}^3\frac{\mu_4R(f^{(3)})}{20\mu_2^2R(f'')}+O(n^{-1}). \end{aligned} \tag{3.3}$$

The great advantage of using $h_{AMISE}^*$to the direct plug-in method is

$$h_{AMISE}^*/h_{MISE}-1=O(^{-1/2}) \text{ for } \nu>5/4.$$

The asymptotic performance of the bandwidth selector based on the representation (3.3) highly depends on how good estimators of $R(f'')$ and $R(f^{(3)})$ we use. Therefore we now focus our attention on the estimation of $R(f^{(m)})$. As is shown in Hall and Marron(1987a), $\hat{R}_a(f'')$ is biased negatively, if the kernel U has order 2. Hence there will be tendency to positive bias in the estimator of h_{AMISE}^*if we use $\hat{R}_a(f'')$ in estimating h_{AMISE}^*, in view of (3.3). Moreover, $\hat{R}_a(f'')$, which is expected to be positive, often has negative value for small sample size. We thus propose to use $\tilde{R}_a(f^{(m)})$ as an estimator of in estimating $R(f^{(m)})$ defined by

$$\tilde{R}_a(f^{(m)})=n^{-1}a^{-2m-1}R(U^{(m)})+(-1)^m\,n^{-1}(n-1)^{-1}\sum_{i\neq j}U_a^{(m)}*U_a^{(m)}\,(X_i-X_j), \tag{3.4}$$

which is essentially equal to $R(\hat{f}^{(m)})$. Note that the leading bias term $n^{-1}a^{-2m-1}R(U^{(m)})$ in (3.4) actually dominate the mean squared error of $\tilde{R}_a(f^{(m)})$.

To introduce the minimum mean squared error of $\tilde{R}_h(f^{(m)})$ we use the following notations. Let

$$\begin{aligned} D_1 &= R(U^{(m)}) \\ D_2 &= 2R(f)R(U^{(m)}*U^{(m)}) \\ D_3 &= \int\{f^{(2m)}\}^2f-R(f^{(m)})^2 \\ D_4 &= 2(-1)^{r/2}R(f^{(m+r/2)})\int x^r\,U(x)\,dx/r\,! \end{aligned}$$

Theorem 3.1. If U satisfies the condition A1 and has order r, and if f has smoothness

of order $\nu > m+r/2-2$ and $\nu > 2m-2$, then the minimum mean squared error of $\tilde{R}_a(f^{(m)})$ is

$$E\{\tilde{R}(f_a^{(m)}) - R(f^{(m)})\}^2 = D_2 c^{\frac{-(4m+1)}{r+2m+1}} n^{\frac{-(2r+1)}{r+2m+1}}$$
$$+\left\{ D_1 c^{\frac{-(2m+1)}{r+2m+1}} + D_4 c^{\frac{r}{r+2m+1}} \right\}^2 n^{\frac{-2r}{r+2m+1}} + D_3 n^{-1}$$
$$+o\left(n^{\frac{-2r}{r+2m+1}} + n^{-1} \right), \tag{3.5}$$

which is achieved by taking $a=(cn^{-1})^{1/(r+2m+1)}$, where

$$c = \begin{cases} \dfrac{(2m+1)\ D_1}{rD_4} & \text{if } D_4 > 0 \\ \dfrac{-D_1}{D_4} & \text{if } D_4 < 0. \end{cases}$$

Remark 3.1. It can be easily seen from Theorem 3.1 that if $D_4 < 0$ and $a = -D_1/D_4$, then the leading term $O(n^{-2r/(r+2m+1)})$ in (3.5) is canceled out. Therefore $\tilde{R}_a(f^{(m)})$ is superior to $\hat{R}_a(f^{(m)})$ since the convergence rates of $\tilde{R}_a(f^{(m)})$ and $\hat{R}_a(f^{(m)})$ are $n^{-(2r+1)/(r+2m+1)}$ and $n^{-4r/(4r+2m+1)}$, respectively. We also know that if $D_4 < 0$, then $\tilde{R}_a(f^{(m)})$ is inferior to $\hat{R}_a$. However, this inferiority does not matter in density estimation($m=2$) when $r > 6$, since both $\tilde{R}_a(f^{(m)})$ and $\hat{R}_a(f^{(m)})$ have their convegence rate $n^{-1/2}$.

Remark 3.2. The minimizing a in Theorem 3.1 is unavailable as it stands since it depends on the unknown quantity D_4. However, as in Section 2.2, D_4 may be replaced by

$$D_4(\hat{\lambda}) = 2(-1)^{r/2} R(g_1^{(m+r/2)}) \int x^r U(x)\, dx / \{r!\, \hat{\lambda}^{r+2m+1}\},$$

where λ is a good estimator of the scale of f.

Remark 3.3. Sheather and Jones(1989) proposed to use what is called *diagonals-in estimators* as estimators of $R_a(f^{(m)})$. Indicating that the bias of the non-stochastic term(arising if $i=j$ in $\hat{R}_a(f^{(m)})$) usually has the opposite sign to the bias due to the smoothing, they reintroduce the non-stochastic term omitted in Hall and Marron(1987a). Accordingly, they choose a bandwidth selelctor to make these bias terms canceled out.

Based on Theorem 3.1 and Remark 3.2, we propose $\hat{h}_{prop1}$ which has $n^{-1/2}$ rate of convergence toward h_{AMISE}. In particular,

$$\hat{h}_{prop1} = \hat{h}_{AMISE} + \hat{h}_{AMISE}^3\ \tilde{R}_{a_2}(f^{(3)})\ \mu_4/20\mu_2\tilde{R}_{a_1}(f''), \tag{3.6}$$

where

$$\hat{h}_{AMISE} = \left\{ \frac{R(K)}{\mu_2^2 \tilde{R}_{a_1}(f'')} \right\}^{1/5} n^{-1/5},$$

and

$$\tilde{R}_{a_1}(f'') = n^{-1} a_1^{-5} R(U_1'') + 2n^{-1}(n-1)^{-1} a_1^{-5} \sum_{i<j} U_1'' * U_1''\{(X_i - X_j)/a_1\},$$

with U_1 a kernel of order 6, a_1 a bandwidth selelctor defined by

$$a_1=a_1(\hat{\lambda})=\hat{\lambda}\left\{\frac{6!\,5R(U_1')\,n^{-1}}{12\int x^6U_1(x)\,dx\,R(g_1^{(5)})}\right\}^{1/11},$$

$$\hat{\lambda}_1\left\{\frac{-6!\,R(U_1')\,n^{-1}}{12\int x^6U_1(x)\,dx\,R(g_1^{(5)})}\right\}^{1/11},$$

and where

$$\tilde{R}_{a_2}(f^{(3)})=n^{-1}a_2^{-7}R(U_2^{(3)})-2n^{-1}(n-1)^{-1}a_2^{-7}\sum_{i<j}U_2^{(3)}*U_2^{(3)}\{(X_i-X_j)/a_2\},$$

with U_2 a kernel of order 2, a_2 a bandwidth selector defined by

$$a_2=a_2(\hat{\lambda})=\hat{\lambda}\left\{\frac{R(U_2^{(3)})\,n^{-1}}{\int x^2U_2(x)\,dx\,R(g_1^{(4)})}\right\}^{1/9},$$

and where $\hat{\lambda}$ is any $\sqrt{n}$-consistent estimator of the sale of f.

The following theorem is essentially due to the fact that $MSE\{\tilde{R}_{a_1}(f'')\}=O(n^{-1})$ and $MSE\{\tilde{R}_{a_2}(f^{(3)})\}=O(n^{-5/9})$.

Theorem 3.2. If f has smoothness of order $\nu>2.25$, and if U_1 and U_2 statisfy the condition A1 with $m=2$ and $m=3$, respectively, then

$$n^{1/2}(\hat{h}_{prop1}/h_{MISE}-1)\to N\{0,\ 4(R(f'')^{-1}\int(f^{(4)})^2f-1)/25\}.$$

Remark 3.4. An independent study conducted by Hall, Sheather, Jones and Marron(1989) uses almost the same idea and yields the same bandwidth selector as the proposed bandwidth selector $\hat{h}_{prop1}$. As estimators of $R(f^{(m)})$, they use the result of Sheather and Jones(1989).

Recall that a_1 and a_2 which appear in constructing $\hat{h}_{prop1}$ depend on the functionals of the derivatives of the underlying density f. Also note that

$$bias\{\tilde{R}_{a_1}(f'')\}=n^{-1}a_1^{-5}R(U_1')-2a_1^6R(f^{(5)})\int x^6U_1(x)\,dx/6!+o(a_1^6), \tag{3.9}$$

and

$$bias\{\tilde{R}_{a_2}(f^{(3)})\}=n^{-1}a_2^{-7}R(U_1^{(3)})-a_2^2R(f^{(4)})\int x^2U_2(x)\,dx+o(a_2^2), \tag{3.10}$$

where $a_1=O(n^{-1/11})$ and $a_2=O(n^{-1/9})$. According to (3.9) and (3.10), even $N(0,\hat{\lambda}^2)$ pilot for f in $R(f^{(4)})$ and $R(f^{(5)})$ are sufficient to give the proposed bandwidth selector $\sqrt{n}$-consistent in relative sense. So far, we have replaced the unknown density f by the reference density $g_1(\cdot/\hat{\lambda})/\hat{\lambda}$.

However, differences between the reference density and the true density f may yield bandwidth selectors which have poor performance in small sample sizes. We now propose a new bandwidth selector which is expected to have better performance in small smaple sizes as well as is optimal in the sense of convergence rate. New proposed bandwidth selector $\hat{h}_{prop2}$ can be obtained by the same method as that of $\hat{h}_{prop1}$ except that a_1 in (3.7) and a_2 in (3.8) are replaced by a_1^* and a_2^*, respectively, which use estimators of $\tilde{R}_{a_3}(f^{(5)})$ and $\tilde{R}_{a_4}(f^{(4)})$ instead of $R(f^{(5)})$ and $R(f^{(4)})$. In particular,

$$a_1^* = R_{a_3}(f^{(5)})^{-1/11}\left\{\frac{-6!\,R(U_1')}{2\int x^6 U_1(x)\,dx}\right\}^{1/11} n^{-1/11},$$

and

$$a_2^* = R_{a_4}(f^{(4)})^{-1/9}\left\{\frac{R(U_2^{(3)})}{\int x^2 U_2(x)\,dx}\right\}^{1/9} n^{-1/9},$$

where

$$\tilde{R}_{a_3}(f^{(5)}) = n^{-1}a_3^{-11}R(U_3^{(5)}) - 2n^{-1}(n-1)^{-1}a_3^{-11}\sum_{i<j} U_3^{(5)} * U_3^{(5)}\{(X_i-X_j)/a_3\},$$

and

$$\tilde{R}_{a_4}(f^{(4)}) = n^{-1}a_4^{-9}R(U_4^{(4)}) + 2n^{-1}(n-1)^{-1}a_4^{-9}\sum_{i<j} U_4^{(4)} * U_4^{(4)}\{(X_i-X_j)/a_4\},$$

with kernel functions U_3 and U_4 of order 2, a_3 and a_4 being bandwidth selectors defined by

$$a_3 = \hat{\lambda}R(g_1^{(6)})^{-1/13}\left\{\frac{R(U_3^{(5)})}{\int x^2 U_3(x)\,dx}\right\}^{1/13} n^{-1/13},$$

and

$$a_4 = \hat{\lambda}R(g_1^{(5)})^{-1/11}\left\{\frac{R(U_4^{(4)})}{\int x^2 U_4(x)\,dx}\right\}^{1/11} n^{-1/11},$$

respectively. The asymptotic distibution of $\hat{h}_{prop2}$ is the same as that of $\hat{h}_{prop1}$ given in Theorem 3.2.

3-2. Proofs

Proof of Teorem 3.1. From Hall and Marron(1987a, 1989),

$$MSE\{\tilde{R}_a(f^{(m)})\} = \{bias(\tilde{R}_a(f^{(m)}))\}^2 + var\{\tilde{R}_a(f^{(m)})\}$$
$$= \{n^{-1}a^{-2m-1}D_1 + a^rD_4\}^2 + n^{-2}a^{-4m-1}D_2 + 4n^{-1}D_3 + o(n^{-1}+a^{2r}+n^{-2}a^{-4m-2}).$$
$$MSE'\{\tilde{R}_a(f^{(m)})\} = (-4m-2)n^{-2}a^{-4m-3}D_1^2 + 2(r-2m-1)n^{-1}a^{r-2m-2}D_1D_4 + 2ra^{2r-1}D_4^2$$
$$= \{2rD_4a^{r+2m+1} + (-4m-2)n^{-1}D_1\}\{D_4a^{r+2m+1} + n^{-1}D_1\} = 0$$

Immediately, the solutions are given by

$$a = \left\{\frac{(2m+1)D_1n^{-1}}{rD_4}\right\}^{1/(r+2m+1)}$$

or $$\left\{\frac{-D_1n^{-1}}{D_4}\right\}^{1/(r+2m+1)}$$

In order to prove Theorem 3.2 we state and prove a lemma which uses Theorem 1 in Hall (1984).

Lemma 3.1. If U is a kernel function satisfying the condition A1, with $a \to 0$ but $na \to \infty$, then $\Sigma_{i<j}\, aU^{(m)} * U^{(m)}\{(X_i-X_j)/a\}$ is asymptotically normally distributed.

Proof of Lemma 3.1. First note that $aU^{(m)} * U^{(m)}\{(X_i-X_j)/a\}$ can be decomposed into

$$\int\left[U^{(m)}\left(\frac{t-X_i}{a}\right)-\mu(t)\right]\left[U^{(m)}\left(\frac{t-X_j}{a}\right)-\mu(t)\right]dt$$
$$+\int\mu(t)\left[U^{(m)}\left(\frac{t-X_i}{a}\right)+U^{(m)}\left(\frac{t-X_j}{a}\right)\right]dt-\int\mu(t)^2\,dt, \qquad (3.11)$$

where

$$\mu(t)=E\left\{U^{(m)}\left(\frac{t-X}{a}\right)\right\}.$$

We denote $H_n(X_i, X_j)$ to be the first term in (3.11), then

$$E\{H_n(X_1, X_2)\}=0,$$

and

$$E\{H_n(X_1, X_2)\mid X_1\}=0.$$

So, H_n is degenerate martingale. Furthermore, we can obtain $EH_n^2=O(a^3)$ and $EH_n^4=O(a^5)$. Finally, if we let

$$G_n(x, y)=E\{H_n(X_1, x)H_n(X_1, y)\}$$
$$=\int H_n(u, x)H_n(u, y)f(u)\,dy,$$

then we have $EG_n^2=O(a^7)$.

Therefore, asymptotic normality follows from Theorem 1 of Hall(1984), since

$$E\{G_n^2(X_1, X_2)\}+n^{-1}E\{H_n^4(X_1, X_2)\}/\mid E\{H_n^2(X_1, X_2)\}\mid^2$$
$$=O(a^7)+n^{-1}O(a^5)/O(a^3)\overset{n\to\infty}{\to}0.$$

Proof of Theorem 3.2. Let $\hat{h}_{\widehat{AMISE}}{}^*$ be the minimizer of

$$\widehat{AMISE}^*(h)=(nh)^{-1}R(K)+h^4\tilde{R}_a(f'')\,\mu_2^2/4-h^6\,\mu_2\mu_4\tilde{R}_a(f^{(3)})/24,$$

then Taylor expansion of $\hat{h}_{\widehat{AMISE}}{}^*$ yields

$$\hat{h}_{\widehat{AMISE}}{}^*=\hat{h}_{AMISE}+\hat{h}_{AMISE}\,\hat{h}^2_{\widehat{AMISE}}{}^*\;\mu^4\tilde{R}_a(f^{(3)})/20\mu_2\tilde{R}_a(f'')+O(n^{-1}), \qquad (3.12)$$

where

$$\hat{h}_{AMISE}=\left\{\frac{R(K)}{\tilde{R}_a(f'')\,\mu_2^2}\right\}^{1/5}n^{-1/5}.$$

We may put

$$\tilde{h}_{\widehat{AMISE}}{}^*=\hat{h}_{AMISE}+h_{AMISE}\tilde{h}_{\widehat{AMISE}}{}^*\;\mu^4\tilde{R}_a(f^{(3)})/20\mu_2\tilde{R}_a(f'').$$

From the above equation, we can easily get

$$\hat{h}_{prop1}-\tilde{h}_{\widehat{AMISE}}{}^*=O(n^{-1}),$$
$$\hat{h}_{\widehat{AMISE}}{}^*-\tilde{h}_{\widehat{AMISE}}{}^*=O(n^{-1}).$$

Furthermore, we can obtain

$$\widehat{AMISE}^{*\prime}(h_{AMISE}{}^*)=\widehat{AMISE}^{*\prime}(h_{AMISE}{}^*)-AMISE^{*\prime}(h_{AMISE}{}^*)$$

$$=h_{AMISE}^{3*}\mu_2^2\,[\tilde{R}_a(f'')-R(f'')]-h_{AMISE}^{5*}\mu_2\mu_4[\tilde{R}_a(f^{(3)})-R(f^{(3)})]$$
$$=h_{AMISE}^{3*}\mu_2^2\,[\tilde{R}_a(f'')-R(f'')]+o(n^{-11/10}),$$

thus

$$\widehat{AMISE}^{*\prime}(\hat{h}_{\widehat{AMISE}}{}^*)=AMISE^{*\prime}(h_{AMISE}{}^*)+\widehat{AMISE}^{*\prime\prime}(h^*)(h_{AMISE}{}^*-\hat{h}_{\widehat{AMISE}}{}^*)$$
$$=h_{AMISE}^{3*}\mu_2^2\,[\tilde{R}_a(f'')-R(f'')]+O(h_{AMISE}^{2*})\;(h_{AMISE}{}^*-\hat{h}_{\widehat{AMISE}}{}^*)$$
$$+o\,(n^{-11/10}),$$

where h^* lies between $h_{AMISE}{}^*$ and $\hat{h}_{\widehat{AMISE}}{}^*$. From (3.3) and (3.12), the distribution of $\hat{h}_{prop1}/h_{MISE}-1$ is determined by that of $\tilde{R}_a(f'')$. Then the result is immediate from Hall and Marron (1987a), and Theorem 3.1 and Lemma 3.1 in this paper.

4. Simulation Results

We have seen the asymptotic behaviors of the proposed bandwidth selelctors. But to get some ideas of how good the proposed bandwidth selectors are in practical situation, it is quite useful to investigate their small-sample properties. For this purpose, simulation studies are conducted. In this section, the simulation results are presented and a brief discussion on the results is also given.

4-1. Design of the simulation

In the simulation study, we include the following typical density functions :

1. standard normal density, N(0, 1).
2. mean mixture, 0.5N(1.5, 1)+0.5N(-1.5, 1).
3. variance mixture, 0.9N(0, 1)+0.1N(0,25).
4. asymmetric distribution, 0.75N(0.5, 1)+0.25N(-0.5, 1).

The reason we used these densities is that they are variants of normal densities so that we can compute the exact $MISE(h)$ and $\tilde{R}_a(f^{(m)})$ easily due to Marron, Park and Wand(1989).

The sample sizes considered here are $n=25$, 50 and 100, and the number of repetitions used is $nrep=300$. We investigate small samle properties of the bandwidth selectors $\hat{h}_{PI}$, $\hat{h}_{UCV}$, $\hat{h}_{BCV}$, and the proposed bandwidth selectors $\hat{h}_{prop1}$ and $\hat{h}_{prop2}$. As kernel functions we took K, U_2, U_3 and U_4 be standard normal density ϕ. To construct a kernel function of order 6, we may apply the method of Silverman(1986, Section 3.6.2) in which kernel function of order 4 is derived. In our simulation study, we used the kernel function U_1 of order 6 defined by

$$U_1(x)=\phi(x)-\phi''(x)/2+\phi^{(4)}(x)/8$$
$$=(15/8-5x^2/4+x^4/8)\phi(x).$$

We restrict the range of $\hat{h}$ to $(3\hat{h}_{prop1},\ \hat{h}_{prop1}/3)$. The values of $\hat{h}_{UCV}$ and $\hat{h}_{BCV}$ are chosen to be the largest local minimizers of their object functions if there are more than one local minimum. If there are no local mimima in the range under consideration, then either the right endpoint or the left of the range is chosen. In this simulation study, $BCV(h)$ sometimes fails to have a local minimum when sample size is 25. In particular, in the normal case 27/300 and in the case of variance mixture 81/300 of $BCV(h)$ had no local minimum. In order to find $\hat{h}_{PI}$ we used an iteration method using $\hat{h}_{prop1}$ as an initial value.

In order to take the Monte Carlo variability properly into account, we computed the pivo-

Table 1. 95% Confidence Intervals for the Mean of the Distribution of (4.1) for Various Bandwidth Selectors.

STANDARD NORMAL		MEAN MIXTURE		VARIANCE MIXTURE		ASYMMERTIC DISTRIBUTON	
n=25		n=25		n=25		n=25	
AMISE	(0.01720, 0.02376)	PROP1	(0.03257, 0.04497)	AMISE	(0.02181, 0.03012)	AMISE	(0.01715, 0.02368)
PROP1	(0.09165, 0.12657)	PROP2	(0.04484, 0.06192)	PROP2	(0.11778, 0.16266)	PROP1	(0.08520, 0.11767)
PROP2	(0.12702, 0.17543)	PI	(0.07061, 0.09752)	PROP1	(0.17297, 0.20592)	PROP2	(0.12035, 0.16620)
PI	(0.21035, 0.29050)	AMISE	(0.11007, 0.15201)	UCV	(0.28625, 0.39532)	PI	(0.22134, 0.30568)
UCV	(0.33663, 0.46490)	UCV	(0.25198, 0.34799)	PI	(0.39678, 0.54797)	UCV	(0.29596, 0.35235)
BCV	(0.47490, 0.65586)	BCV	(0.39147, 0.54063)	BCV	(1.38190, 1.90847)	BCV	(0.41836, 0.57777)
n=50		n=50		n=50		n=50	
AMISE	(0.01037, 0.01432)	PROP2	(0.04280, 0.05911)	AMISE	(0.01323, 0.01827)	AMISE	(0.01038, 0.01434)
PROP1	(0.04993, 0.06895)	AMISE	(0.05788, 0.07993)	PROP2	(0.05834, 0.08057)	PROP1	(0.04034, 0.05572)
PROP2	(0.07082, 0.09781)	PROP1	(0.06495, 0.08969)	PROP1	(0.11659, 0.16102)	PROP2	(0.05917, 0.08172)
PI	(0.12362, 0.17072)	PI	(0.13885, 0.19176)	BCV	(0.17865, 0.24673)	PI	(0.12996, 0.17948)
BCV	(0.13831, 0.19101)	UCV	(0.21027, 0.29039)	UCV	(0.21242, 0.29335)	BCV	(0.14795, 0.20432)
UCV	(0.24427, 0.33735)	BCV	(0.55662, 0.76872)	PI	(0.23530, 0.32496)	UCV	(0.25247, 0.34867)
N=100		N=100		N=100		N=100	
AMISE	(0.00616, 0.00851)	AMISE	(0.03147, 0.04347)	AMISE	(0.00806, 0.01114)	AMISE	(0.00648, 0.00893)
PROP1	(0.02225, 0.03073)	PROP2	(0.04971, 0.06865)	PROP2	(0.03205, 0.04426)	PROP1	(0.02287, 0.03158)
PROP2	(0.03153, 0.04354)	PROP1	(0.09238, 0.12758)	BCV	(0.07553, 0.10431)	PROP2	(0.03309, 0.04570)
BCV	(0.07004, 0.09672)	UCV	(0.17560, 0.24251)	PROP1	(0.08136, 0.11237)	BCV	(0.06948, 0.09595)
PI	(0.07223, 0.09976)	PI	(0.24032, 0.33189)	PI	(0.13821, 0.19088)	PI	(0.07430, 0.10261)
UCV	(0.18526, 0.25585)	BCV	(0.87710, 1.21131)	UCV	(0.15008, 0.20727)	UCV	(0.20299, 0.28033)

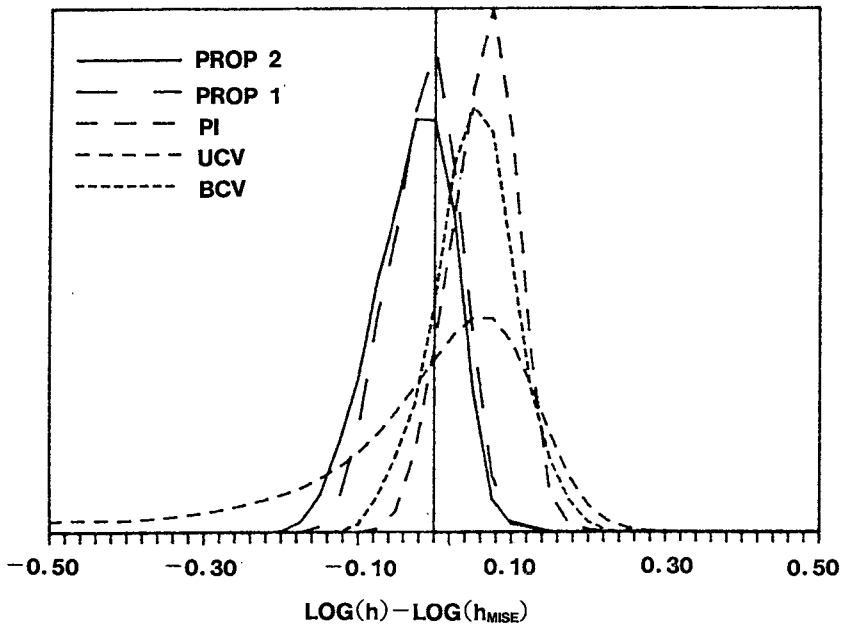

Fig. 1.a. Overlay of Kernel Density Estimators of Various Bandwidth Selectors in Standard Normal Case with n=100.

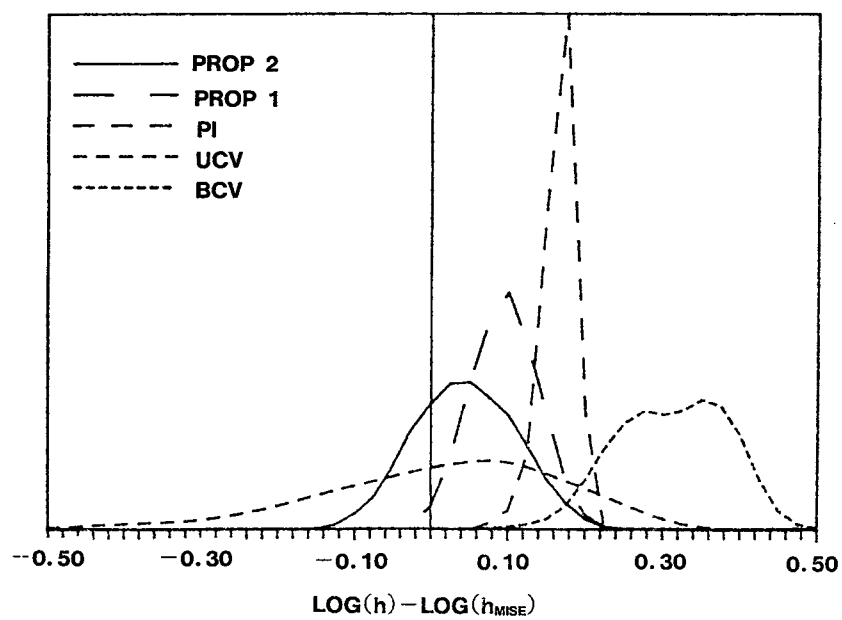

Fig. 1.b. Overlay of Kernel Density Estimators of the Distributions of Various Bandwidth Selectors in Mean Mixture Case with n=100.

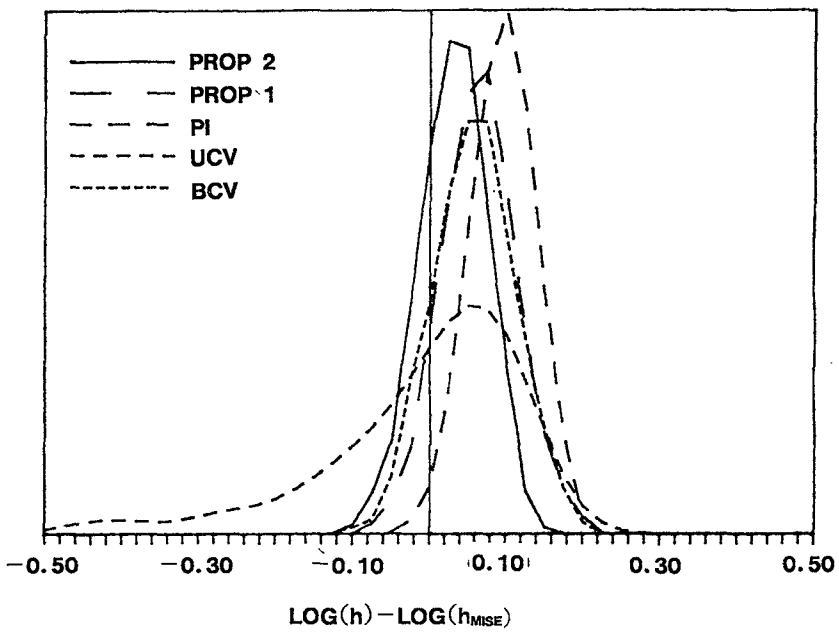

Fig. 1.c. Overlay of Kernel Density Estimators of the Distributions of Various Bandwidth Selectors in Variance Mixture Case with n=100.

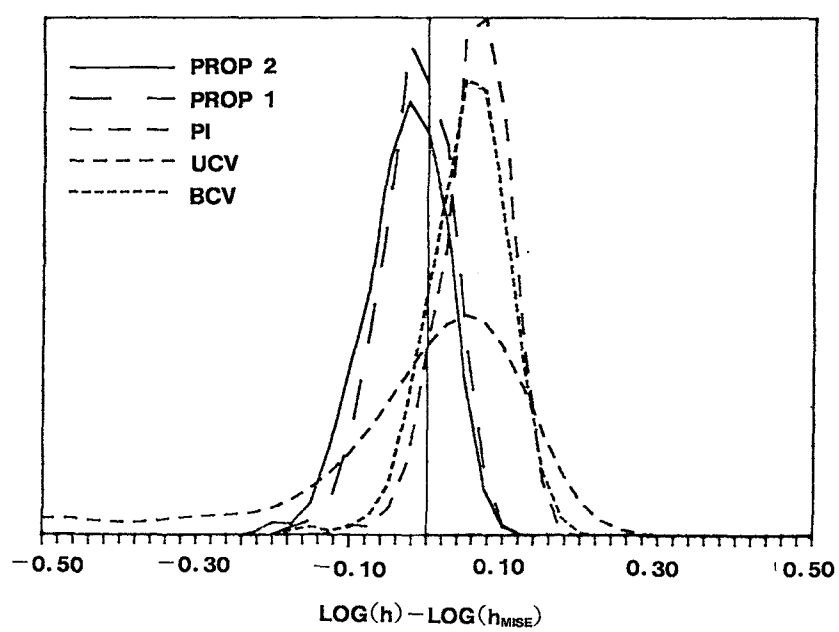

Fig. 1.d. Overlay of Kernel Density Estimation of the Distributions of Various Bandwidth Selectors in Asymmetric Case n=100.

95% confidence intervals of

$$MISE(\hat{h})/MISE(h_{MISE})-1,$$

which is given by

$$(\hat{P}/(1+T),\ P/(1-T)),$$

where $\hat{P}$ is an estimator of

$$P=E\{MISE(\hat{h})/MISE(h_{MISE})\}-1, \tag{4.1}$$

and

$$T=1.96(2/nrep)^{1/2}.$$

See Marron(1989) for derivation and discussion of these intervals.

Table 1 contains the results of the confidence intervals, which are arranged from the best to the worst for each sample sizes, for the comparision number P in (4.1). This allows comparision of various bandwidth selectors, and at the same time gives an idea of the sample variability involved.

Figure 1. a~Figure 1. d contain a more visual method of illustrating the comparison of bandwidth selectors. The figures show an overlay of the kernel density estimators of the densities of $\log_{10}(\hat{h})-\log_{10}(h_{MISE})$ for $n=100$ observations from (*a*) the standard normal, (*b*) the mean mixture, (*c*) variance mixture, and (*d*) asymmetric distribution. The bandwidth used for the kernel estimators was the oversmoother which is due to Terrel and Scott(1985).

4-2. Simulation Results

From the table and figures, we can see that our proposed bandwidth selectors $\hat{h}_{prop1}$ and $\hat{h}_{prop2}$ have good performances for all underlying densities and for all sample sizes. As expected, $\hat{h}_{prop1}$ has better performances than $\hat{h}_{prop2}$ in the standard normal case and also in the asymmetric case. However, the performance of $\hat{h}_{prop2}$ is better than that of $\hat{h}_{prop1}$ when the assumed densities have somewhat different structure from standard normal, such as variance mixture and mean mixture. Especially, in the variance mixture case, $\hat{h}_{prop2}$ behaves much better than $\hat{h}_{prop1}$. We can also see from the table that the biased cross-validated bandwidth selector $\hat{h}_{BCV}$ exhibits very poor behaviors for $n=25$ in all cases. But for $n=100$ the performances of $\hat{h}_{BCV}$ is usually better than that of $\hat{h}_{UCV}$ is usually better than that of $\hat{h}_{UCV}$ and sometimes better than that of $\hat{h}_{PI}$. The performance of $\hat{h}_{UCV}$ is poor for all sample sizes in the normal and asymmetric cases. The plug-in bandwidth selector $\hat{h}_{PI}$ has relatively good performance in all cases. Especially in the mean mixture case, it exihibits better performance than that of h_{AMISE} for $n=25$ although not significantly so.

From the figures, we can also see that the distributions of $\hat{h}_{UCV}$ have long tails and are skewed to the right for all the cases. However, the other bandwidth selectors have almost symmetric distributions. In particular, the distributions of $\hat{h}_{pl}$ have short tails, which is difficult to see in the talbe.

References

1. Bowman, A.(1984), An alternative method of cross-validation for the smoothing of density estimates, *Biometrika*, 71, 353-360.
2. Devroye, L. and Györfi, L.(1984), *Nonparametric Density Estimation : The L_1-View*, Wiley, New York.
3. Hall, P.(1980), Objective methods for the estimation of window size in the nonparametric estimation of a density, *unpublished manuscript*.
4. Hall, P.(1984), Central limit theorem for intgrated squared error of multivariate density estimators, *Journal of Multivariate Analysis*, 14, 1-16.
5. Hall, P. and Marron, J.S.(1987a), Estimation of integrated squared density derivatives, *Statistics and Probability Letters*, 6, 109-115.
6. Hall, P. and Marron, J.S.(1987b), Extent to which least squares cross-validation minimizes integrated squared error in nonparametric density estimation, *Probability Theory and Related Fields*, 74, 567-581.
7. Hall, P. and Marron, J.S.(1989), Lower bounds for bandwidth selection in density estimation, *unpublished manuscript*.
8. Hall, P., Sheather, S.J., Jones, M.C. and Marron, J.S.(1989), On optimal data-based bandwidth selection in kernel density estimation, *unpublished manuscript*.
9. Marron, J.S.(1988), Automatic smoothing parameter selection : a survey, *Empirical Economics*, to appear.
10. Marron, J.S.(1989), Comments on a data based bandwidth selector, *Computational Statistics and Data Analysis*, 8, 155-170.
11. Marron, J.S., Park, B.U. and Wand, M.P.(1989), Facts about the normal density, *Technical Report*, Cornell University.
12. Park, B.U.(1989), On the plug-in bandwidth selectors in kernel density estimation, *Journal of the Korean Statistical Society*, 18, 107-117.
13. Park, B.U. and Marron, J.S.(1990), Comparision of data-driven bandwidth selectors, *Journal of the American Statistical Association*, 85, 66-72.
14. Parzen, E.(1962), On estimation of a probability density function and mode, *Annals of Mathematical Statistics*, 33, 1065-1076.
15. Rosenblatt, M.(1956), Remarks on some non-parametric estimates of a density function, *Annals of Mathematical Statistics*, 27, 832-837.
16. Rudemo, M.(1982), Empirical choice of histogram and kernel density estimators, *Scandinavian Journal of Statistics*, 9, 65-78.
17. Scott, D.W. and Terrell, G.R.(1987), Biased and unbiased cross-validation in density estimation, *Journal of the American Statistical Association*, 82, 1131-1146.
18. Sheather, S.J.(1983), A data-based algorithm for choosing the window width estimating the density at a point, *Computational Statistics and Data Analysis*, 1, 229-238.
19. Sheather, S.J.(1986), An improved data-based algorithm for choosing the window width when estimating the density at a point, *Computational Statistics and Data Analysis*, 4, 61-65.
20. Sheather, S.J. and Jones, M.C.(1989), Reliable data-based bandwidth selection for kernel density estimation, with emphasis on a successful non-cross-validatory approach, *unpublished manuscript*.
21. Silverman, B.W.(1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York.
22. Terrell, G.R. and Scott, D.W.(1985), Oversmoothed nonparametric density estimates, *Journal of the American Statistical Association*, 80, 209-214.