Statistica Sinica **6**(1996), 129-145

A COMPARATIVE REVIEW OF BANDWIDTH SELECTION FOR KERNEL DENSITY ESTIMATION

Shean-Tsong Chiu

Colorado State University

Abstract. In kernel density estimation, a crucial step is to select a proper smoothing parameter (bandwidth). The bandwidth considerably affects the appearance of the density estimate. The most studied procedure is cross-validation. It is well known that cross-validation is subject to large sample variation and often selects smaller bandwidth. Recently, some procedures have been proposed to remedy the difficulties. The implementation, the asymptotic properties and the empirical performance of several bandwidth selectors are investigated. Based on the sample characteristic function, it is shown that these bandwidth selectors have a similar form. The main difference is in the selection of a second bandwidth to estimate the mean integrated squared errors. Our simulation study indicates that the selection of the second bandwidth greatly affects the performance of the procedures.

Key words and phrases: Bandwidth selection, characteristic function, cross-validation, kernel density estimation.

1. Introduction

Given a random sample $X_1, ..., X_n$ from a distribution with the density function $f(x)$, one is often interested in estimating $f(x)$. Silverman (1986) discussed many important applications of density estimation. The most commonly used nonparametric method is the kernel estimate $\hat{f}_\beta(x) = (n\beta)^{-1} \sum_{j=1}^n w\{(x - X_j)/\beta\}$, (see Rosenblatt (1956)) where the kernel function $w(x)$ is assumed to be a symmetric probability density function and β is the bandwidth. The bandwidth controls the smoothness of the density estimate and greatly affects its appearance. Selecting a proper β is a crucial step in estimating $f(x)$. Although in practice, one may choose the bandwidth subjectively, there is a great demand for automatic (data-driven) bandwidth selection procedures. Some reasons for using automatic procedures were given in Silverman (1985). In Section 2, we give a brief background on automatic bandwidth selection.

The most studied automatic bandwidth selector is the least squares cross-validation (henceforth CV) proposed by Rudemo (1982) and Bowman (1984). It is well recognized that the bandwidth estimate has a very slow convergence rate, and is subject to large sample variation. In simulation studies, it is also

observed that the selector chooses smaller bandwidth much more frequently than predicted by asymptotic results. The difficulties of cross-validation severely limit its practicability.

Recently, several procedures have been proposed to remedy the difficulties of CV. In this paper, we compare the procedures of Scott & Terrell (1987), Park & Marron (1990), Chiu (1991b, 1992), Hall, Marron & Park (1992) (henceforth HMP), Sheather & Jones (1991), Hall, Sheather, Jones & Marron (1991) (henceforth HSJM) and Jones, Marron & Park (1991) (henceforth JMP). Except for the biased cross-validation (henceforth BCV) of Scott & Terrell (1987), these bandwidth estimates have a faster convergence rate. In particular, the procedures of Chiu (1991b, 1992), HSJM (1991), and JMP (1991) give $\sqrt{n}$ consistent estimates. Although the new procedures are asymptotically better than CV, it was found in simulation studies that some procedures perform much worse than CV.

The main purpose of this paper is to investigate the theoretic properties, the implementation, and most importantly, the actual performance of the procedures under various situations. We show similarities and point out differences between the procedures. We also explain why some procedures do not perform as well as indicated by the asymptotic results.

Based on Fourier transforms, we show that all the procedures have a similar form. Roughly speaking, the new procedures use another kernel to estimate the bias term in the mean integrated squared error (MISE). The high frequency components of the sample characteristic function are downweighted to reduce their effects. Some practical recommendations are also given in Section 5.

2. Background

A commonly used measure of the performance of $\hat{f}_\beta(x)$ is the mean integrated squared error $\mathrm{MISE}_n(\beta) = E\{\mathrm{ISE}_n(\beta)\}$, where $\mathrm{ISE}_n(\beta) = \int\{\hat{f}_\beta(x) - f(x)\}^2 dx$. Unless indicated otherwise, the integration is over the whole real line throughout the paper. Under some smoothness assumptions, $A_n(\theta) = n^{4/5}\mathrm{MISE}_n(n^{-1/5}\theta)$ converges to

$$A(\theta) = \theta^{-1}\int w^2(x)dx + 4^{-1}\theta^4\{\int x^2 w(x)dx\}^2 \int\{f''(x)\}^2 dx \tag{2.1}$$

which has a unique minimum at θ_0, where $\theta_0^5 = \int w^2(x)dx/[\{\int x^2 w(x)dx\}^2 \int\{f''(x)\}^2 dx]$.

In the following discussion, let β_{0n} denote the optimal bandwidth that minimizes $\mathrm{MISE}(\beta)$. We also let $\beta_0 = n^{-1/5}\theta_0$ be the asymptotic optimal bandwidth. In the discussion about theoretic properties, we should assume that the density function and the kernel satisfy the assumptions set in Chiu (1992).

Rudemo (1982) and Bowman (1984) proposed the least squares CV

$$\mathrm{CV}_n(\beta) = \int \hat{f}_\beta^2(x) - n^{-1}\sum_{j=1}^{n} \hat{f}_{\beta,j}(X_j),$$

where $\hat{f}_{\beta,j}(x)$ is the kernel density estimate without using the jth observation. The asymptotic properties of the bandwidth estimate were established in Scott & Terrell (1987) and Hall & Marron (1987). It was shown that the bandwidth estimate is consistent and is asymptotically normal. The estimate has a very slow relative convergence rate $n^{-1/10}$. In the simulation studies of Scott & Terrell (1987) and Chiu (1991b), it was found that CV often selects a very small bandwidth, and the resulting density estimate is very rough and shows too many false feathers.

As demonstrated in Rice (1984) and Chiu (1990, 1991a, 1991b, 1992), Fourier analysis is a powerful tool in the study of bandwidth selection. Following this approach, we use characteristic functions to compare bandwidth selection procedures. Let $\phi(\lambda) = \int \exp(i\lambda x) f(x)dx$ be the characteristic function of $f(x)$, and $\tilde{\phi}(\lambda) = (1/n)\sum \exp(i\lambda X_j)$ be the sample characteristic function. In the following discussing, we borrow the terminology "frequency" for λ in time series analysis. For smooth f, $|\phi(\lambda)|$ decays quickly, and $\mathrm{Var}\{\tilde{\phi}(\lambda)\} \approx 1/n$ at high frequencies. The information about f is concentrated at the low frequencies.

By Parseval's formula, $\mathrm{ISE}_n(\beta)$ can be written as

$$\mathrm{ISE}_n(\beta) = \int \{\hat{f}_\beta(x) - f(x)\}^2 dx = \frac{1}{2\pi}\int |\phi(\lambda) - \tilde{\phi}(\lambda)W(\beta\lambda)|^2 d\lambda. \tag{2.2}$$

Let $\tilde{\phi}_d(\lambda) = \tilde{\phi}(\lambda) - \phi(\lambda)$ denote the noise part of $\tilde{\phi}(\lambda)$. Note that $|\tilde{\phi}_d(\lambda)|^2$ is approximately an exponential random variable with mean $\{1 - |\phi(\lambda)|^2\}/n$. Expand (2.2) to obtain

$$\begin{aligned}\mathrm{ISE}_n(\beta) = {} & \frac{1}{2\pi}\int |\phi(\lambda)|^2\{1 - W(\beta\lambda)\}^2 d\lambda + \frac{1}{2\pi}\int |\tilde{\phi}_d(\lambda)|^2 W^2(\beta\lambda)d\lambda \\ & - \frac{2}{2\pi}\int \phi(\lambda)\tilde{\phi}_d(-\lambda)W(\beta\lambda)\{1 - W(\beta\lambda)\}d\lambda. \end{aligned}\tag{2.3}$$

Letting $w_2(x) = w * w(x)$, we have from (2.3)

$$\begin{aligned}\mathrm{MISE}_n(\beta) &= \frac{1}{2\pi}\int |\phi(\lambda)|^2\{1 - W(\beta\lambda)\}^2 d\lambda + \frac{1}{2\pi}\int W^2(\beta\lambda)\{1 - |\phi(\lambda)|^2\}/n\, d\lambda \\ &\approx \frac{1}{2\pi}\int |\phi(\lambda)|^2\{1 - W(\beta\lambda)\}^2 d\lambda + \frac{w_2(0)}{n\beta}. \end{aligned}\tag{2.4}$$

Since $W(\lambda) \approx 1 - \lambda^2 \int x^2 w(x)dx/2$ for λ near the origin,

$$\int |\phi(\lambda)|^2\{1 - W(\beta\lambda)\}^2 d\lambda \approx \frac{\beta^4}{8\pi}\{\int x^2 w(x)dx\}^2 \int \lambda^4 |\phi(\lambda)|^2 d\lambda. \tag{2.5}$$

From (2.5), we could obtain the asymptotic MISE given in (2.1).

Silverman (1986) shows that CV can be approximately expressed by

$$\mathrm{CV}(\beta) \approx \frac{1}{2\pi}\int\{|\tilde{\phi}(\lambda)|^2 - 1/n\}\{W^2(\beta\lambda) - 2W(\beta\lambda)\}d\lambda + \frac{w_2(0)}{n\beta}. \tag{2.6}$$

Comparing (2.4) and (2.6), we see that CV uses the first term in (2.6) to estimate the bias term in $\mathrm{MISE}_n(\beta)$. Applying a Taylor series expansion yields

$$\hat{\beta}_{\mathrm{CV}} - \beta_{0n} = \{\mathrm{CV}'(\hat{\beta}_{\mathrm{CV}}) - \mathrm{MISE}_n'(\beta_{\mathrm{CV}})\}/\mathrm{MISE}_n''(\tilde{\beta})$$

for some $\tilde{\beta}$ between $\hat{\beta}_{CV}$ and β_{0n}. Simple computation shows that $\hat{\beta}_{CV} - \beta_{0n}$ is proportional to

$$-\int[|\tilde{\phi}_d(\lambda)|^2 - E\{|\tilde{\phi}_d(\lambda)|^2\}]V(\beta_{0n}\lambda)/\beta_{0n}d\lambda - 2\int\phi(\lambda)\tilde{\phi}(-\lambda)V(\beta_{0n}\lambda)/\beta_{0n}d\lambda \tag{2.7}$$

plus some negligible terms, where $V(\lambda) = W(\lambda)W'(\lambda)\lambda$. The first term in (2.7) is the dominant one. Note that the amplitude of $V(\beta_{0n}\lambda)$ is significant only at $\lambda = O(n^{1/5})$. However, relative to the noise level in $\tilde{\phi}(\lambda)$, the characteristic function of a smooth density is negligible at $\lambda = O(n^{-1/5})$. From this, we see that the difficulty of CV is caused by including too much $\tilde{\phi}(\lambda)$ at high frequencies, which do not contain much information about f. Figure 1 shows $|\tilde{\phi}(\lambda)|^2$ of a data set ($n = 100$) simulated from the standard normal distribution. The heights of the horizontal lines are 3/100 and 1/100, respectively. The sidelobes around $\lambda = 4$ and $\lambda = 10$ are due to the sample variation. The characteristic function of the normal density has no sidelobes. The sidelobe around $\lambda = 10$ causes CV to select a very small bandwidth 0.181, while the optimal bandwidth is 0.445.

The new bandwidth selectors downweight the high frequency components to reduce their effects. According to the targets to be estimated, the bandwidth selectors can be classified into three groups. The first group includes the CV, the smoothed cross-validation (henceforth SCV) of HMP (1992), the stabilized selector of Chiu (1991b, 1992), and the procedure of JMP (1991). The bias term in $\mathrm{MISE}_n(\beta)$ is estimated by

$$\frac{1}{2\pi}\int\{|\tilde{\phi}(\lambda)|^2 - 1/n\}\{1 - W(\beta\lambda)\}^2U(\alpha\lambda)d\lambda, \tag{2.8}$$

where $U(\lambda)$ is a real and symmetric weighting function, and α is another smoothing parameter. To improve the convergence rate of the bandwidth estimate, the smoothing parameter α must converges to zero slower than β_{0n}. The procedure of JMP (1991) does not subtract $1/n$ from $|\tilde{\phi}(\lambda)|^2$, and sets α proportional to β^{-2} to make the leading bias term independent of β.

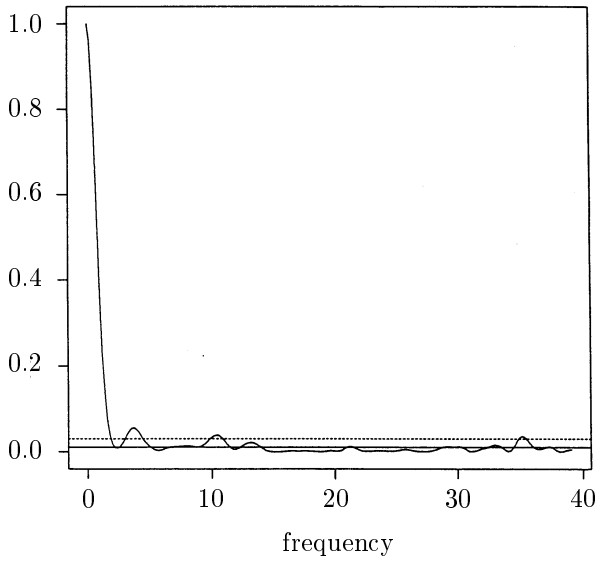

Figure 1. The plot of $|\tilde{\phi}(\lambda)|^2$ of a data set of size 100 simulated from the standard normal distribution.

The second group comprises the plug-in estimates which estimate the optimal bandwidth β_{0n} or β_0 by replacing the unknown quantities in AMISE(β) (cf. (2.1)) with estimates. The estimate of $\int\{f''(x)\}^2dx$ has the form

$$\frac{1}{2\pi}\int \lambda^4\{|\tilde{\phi}(\lambda)|^2 - 1/n\}U(\alpha\lambda)d\lambda. \tag{2.9}$$

By replacing $\int\{f''(x)\}^2dx$ in (2.1) with a $\sqrt{n}$-consistent estimate, we obtain a $\sqrt{n}$-consistent estimate of β_0. In order to obtain a $\sqrt{n}$-consistent estimate of the optimal bandwidth β_{0n}, we need to expand the approximation one more term, which depends on $\int\{f'''(x)\}^2dx$. (See HSJM (1991) for more details). The estimate of $\int\{f'''(x)\}^2dx$ is obtained in a similar way.

The third group includes the conventional plug-in methods: the BCV of Scott & Terrell (1987), and the procedures of Park & Marron (1990) and Sheather & Jones (1991). These procedures use $2\pi^{-1}\int \lambda^4\{|\tilde{\phi}(\lambda)|^2 - 1/n\}W^2\{\alpha(\beta)\lambda\}d\lambda$ to estimate $\int\{f''(x)\}^2dx$. Here $W^2\{\alpha(\beta)\lambda\}$ is used as the weighting function, and α is a function of β. The estimate replaces $\int\{f''(x)\}^2dx$ in AMISE(β) or AMISE$'(\beta)$. The procedure of Sheather & Jones (1991) does not subtract $1/n$ from $|\tilde{\phi}(\lambda)|^2$.

From the discussion above, it can be seen that all selectors have a similar form. A weighting function $U(\alpha\lambda)$ is used to reduce the variation caused by the

high frequency components of $\tilde{\phi}(\lambda)$. The main difference between the procedures is in the selection of $U(\lambda)$ and the smoothing parameter α.

As we shall see later, it is critical to select a proper α. There are two major approaches. One approach figures out the asymptotic optimal value of α. But since the optimal value depends on the unknown density, it is suggested to set α according to some "reference density". The main difficulty here is that the procedures would not perform well unless the true density is quite similar to the reference density. Also, in order to make the procedures scale equivariant, the approach needs a scale estimate of the density. We should point out that Sheather & Jones (1991) used two kernel estimates to estimate α, whose bandwidths are set according to the reference density. As shown in the simulation study in Section 5, this additional step greatly improves the performance of the procedures using this approach.

Noting that the variation in the CV is mainly caused by the high frequency components, Chiu (1991b, 1992) suggested another approach which cuts off the high frequency components in estimating $\int\{f''(x)\}^2dx$ and MISE. The practical issue here is the selection of the cut-off frequency.

3. Bandwidth Selectors

In this section, we express some bandwidth selectors using (2.8) and (2.9). From (2.6), suppose we treat $\int\{|\tilde{\phi}(\lambda)|^2 - 1/n\}d\lambda$ (which is undefined and independent of β) as a finite constant, and add it to (2.6), then

$$\mathrm{CV}_n(\beta) \approx \frac{1}{2\pi}\int\{|\tilde{\phi}(\lambda)|^2 - 1/n\}\{1 - W(\beta\lambda)\}^2 d\lambda + \frac{w_2(0)}{n\beta}.$$

Thus, we obtain the form (2.8) with $U(\lambda) \equiv 1$.

The BCV of Scott & Terrell (1987) replaces $\int\{f''(x)\}^2dx$ in AMISE(β) by the estimate $(2\pi)^{-1}\int\lambda^4\{|\tilde{\phi}(\lambda)|^2 - 1/n\}W^2(\beta\lambda)d\lambda$. The bandwidth estimate is the minimizer of

$$\mathrm{BCV}(\beta) = \frac{\beta^4}{8\pi}\{\int x^2 w(x)dx\}^2\int\lambda^4\{|\tilde{\phi}(\lambda)|^2 - 1/n\}W^2(\beta\lambda)d\lambda + \frac{w_2(0)}{n\beta},$$

such that $U(\lambda) = W^2(\lambda)$ and $\alpha = \beta$.

Instead of using the same bandwidth β, Park & Marron (1990) use a different bandwidth to estimate $\int\{f''(x)\}^2dx$ in the AMISE$'(\beta)$. The bandwidth estimate is the root of

$$\frac{\beta^3}{2\pi}\{\int x^2 w(x)dx\}^2\int\lambda^4\{|\tilde{\phi}(\lambda)|^2 - 1/n\}W^2\{\alpha(\beta)\lambda\}d\lambda - \frac{w_2(0)}{n\beta^2} = 0.$$

The bandwidth α is set as $\alpha(\beta) = C_{\rm PM}\hat{\sigma}^{3/13}\beta^{10/13}$, where $C_{\rm PM}$ is a constant and $\hat{\sigma}$ is the sample standard deviation. The normal density is used as the reference density to set $C_{\rm PM}$. There is a type error in the formula of $C_{\rm PM}$, but we will use the correct constant in the simulation study of Section 5. The approach of using a reference density is also suggested in Sheather & Jones (1991), HMP (1992), HSJM (1991), and JMP (1991).

Sheather and Jones (1991) proposed a modification to Park & Marron (1990). Their estimate is the root of

$$\frac{\beta^3}{2\pi}\{\int x^2 w(x)dx\}^2 \int \lambda^4 |\tilde{\phi}(\lambda)|^2 W\{\alpha(\beta)\lambda\}d\lambda - \frac{w_2(0)}{n\beta^2} = 0.$$

Note that $1/n$ is not subtracted from $|\tilde{\phi}(\lambda)|^2$. By setting α properly, the effect of the leading bias terms would be cancelled. The bandwidth α is set as $\alpha(\beta) = \beta^{-5/7}\hat{C}_{\rm SJ}$. There is a major difference between this and other procedures that use a reference density. At the first stage, the constant $C_{\rm SJ}$ is estimated instead of being set according to the reference density. The reference density is used in the second stage to obtain the bandwidths for estimating $C_{\rm SJ}$. The sample inter-quartile is used as the scale estimate.

The next group of selectors estimates MISE(β). HMP (1992) propose the SCV,

$$\mathrm{SCV}(\beta) = \frac{1}{2\pi}\int \{|\tilde{\phi}(\lambda)|^2 - 1/n\}\{1 - W(\beta\lambda)\}^2 U(\alpha\lambda)d\lambda + \frac{w_2(0)}{n\beta}.$$

They discuss some theoretic properties of the general procedure. For the implementation, they consider the case $U(\lambda) = W^2(\lambda)$ with $w(x)$ the normal kernel, and α set as $\alpha = \hat{\sigma}C_{\rm HMP}n^{-2/13}$, where $\hat{\sigma}$ is the sample standard deviation. The constant $C_{\rm HMP}$ is obtained by using the normal density as the reference density.

JMP (1991) suggested the bandwidth estimate that minimizes

$$\frac{1}{2\pi}\int |\tilde{\phi}(\lambda)|^2\{1 - W(\beta\lambda)\}^2 W^2\{\alpha(\beta)\lambda)\}d\lambda + \frac{w_2(0)}{n\beta}.$$

Note that the procedure does not subtract $1/n$ from $|\tilde{\phi}(\lambda)|^2$. The effects of the leading bias terms would cancel when α is set as $\alpha(\beta) = \beta^{-2}n^{-23/45}C_{\rm JMP}$, where $C_{\rm JMP}$ is a constant obtained by using $N(0, \hat{\sigma}^2)$ as the reference density, and $\hat{\sigma}$ is some scale estimate.

Noting that the difficulties of CV are caused by including too much $\tilde{\phi}(\lambda)$ at high frequencies, Chiu (1991b) suggests that $\tilde{\phi}(\lambda)$ be ignored at high frequencies,

and proposes the stabilized criterion,

$$S(\beta) = \frac{1}{2\pi}\int_{-\Lambda}^{\Lambda}\{|\tilde{\phi}(\lambda)|^2 - 1/n\}\{1 - W(\beta\lambda)\}^2 d\lambda + \frac{w_2(0)}{n\beta}.$$

For the stabilized procedure, $U(\lambda) = 1_{[-1,1]}(\lambda)$, and $\alpha = 1/\Lambda$. Here $1_{[-1,1]}$ is the indicator function on $[-1, 1]$. Chiu (1991b) proposes to selecting the cut-off frequency Λ as the first frequency such that $|\tilde{\phi}(\lambda)|^2 > c/n$ for some constant c. Since, at high frequencies, $|\tilde{\phi}(\lambda)|^2$ is approximately exponentially distributed with mean $1/n$, it is suggested setting c between 2 and 3. The procedure works well when $|\phi(\lambda)|^2$ decays monotonically. However, when $|\phi(\lambda)|$ has significant sidelobes, the procedure may ignore the sidelobes, and seriously overestimate the bandwidth.

To overcome this difficulty, Chiu (1992) proposes selecting Λ as the minimizer of

$$\mathrm{CV}_n^{\infty}(\Lambda) = -(2\pi)^{-1}\int_{-\Lambda}^{\Lambda}|\tilde{\phi}(\lambda)|^2 d\lambda + \frac{4\Lambda}{2\pi n}. \tag{3.1}$$

Similar to the difficulties in $\mathrm{CV}_n(\beta)$, $\mathrm{CV}_n^{\infty}(\Lambda)$ also selects large Λ occasionally. In order to reduce this chance, Chiu (1992) also suggests a modification. The basic idea of the modified procedure is to select the cut-off frequency as a smaller local minimizer of $\mathrm{CV}_n^{\infty}(\Lambda)$ unless we are sure that the higher frequency components contain significant information about f. The modified estimate is the global minimizer of

$$\mathrm{CV}_n^{\mathrm{m}}(\Lambda) = \mathrm{CV}_n^{\infty}(\Lambda) + 1.65\{2\max(0, \Lambda - \hat{\Lambda}_1)\textstyle\int f^2(x)dx/\pi\}^{1/2}/n. \tag{3.2}$$

The constant 1.65 is used because it is the 95th percentile of the standard normal distribution. Figure 2 compares $\mathrm{CV}_n^{\infty}(\beta)$ and $\mathrm{CV}_n^{m}(\beta)$ for the data set used in Figure 1. The stabilized procedures are scale equivariant and do not need a scale estimate. Note that the rate of Λ is not fixed, and is adaptive to the smoothness of the density f.

Finally, we review the plug-in estimates. Chiu (1991b) used $\int_{-\Lambda}^{\Lambda}\lambda^4\{|\tilde{\phi}(\lambda)|^2 - 1/n\}d\lambda$ to estimate $\int\{f''(x)\}^2dx$, where Λ is selected in the same way as the stabilized procedure above. HSJM (1991) proposed the estimate $\int\lambda^4\{|\tilde{\phi}(\lambda)|^2 - 1/n\}U(\alpha\lambda)d\lambda$, where $U(\lambda)$ is the Fourier transform of a $14th$ order polynomial, and $\alpha = n^{-1/11}(\hat{\sigma}/1.349)C_{\mathrm{HSJM}}$, with $\hat{\sigma}$ the sample inter-quartile range and C_{HSJM} being set by using the standard normal density as the reference. The above estimates of θ_0 are $\sqrt{n}$ consistent. To obtain $\sqrt{n}$ consistent estimates of $n^{1/5}\beta_{0n}$, we need an estimate of $\int\{f'''(x)\}dx$, which can be obtained in a similar fashion. Table 1 provides a summary of the implementation of the procedures.

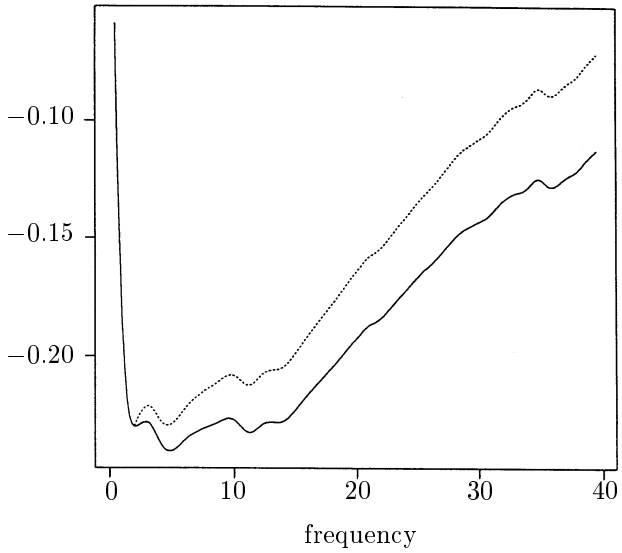

Figure 2. Comparison of $\mathrm{CV}_n^{\infty}(\beta)$ $\mathrm{CV}_n^{m}(\beta)$ for the data set used in Figure 1.

Table 1. A comparison of the implementation of the procedures. In each column, "√" means that the procedure needs to set the item, and "-" means that the procedure does not need the item.

			Requires		
Procedure	$u(x)$	α	Reference density	Scale estimate	Preset constant
CV	-	-	-	-	-
BCV	$w * w(x)$	β	-	-	-
PM	$w * w(x)$	$c\beta^{10/13}$	√	√	-
SJ	$w(x)$	$\hat{c}\beta^{-5/7}$	√	√	-
HSJM	14th order poly.	$cn^{-1/11}$	√	√	-
HMP, SCV	$w * w(x)$	$cn^{-2/13}$	√	√	-
JMP	$w * w(x)$	$c\beta^{-2}n^{-23/45}$	√	√	-
CS	Inf. order kernel	Estimated	-	-	√
CSI	Inf. order kernel	Estimated	-	-	-
CSM	Inf. order kernel	Estimated	-	-	√

4. Asymptotic Properties

In this section, we compare the asymptotic properties of the bandwidth selectors discussed in the previous section. To reduce some technical details, we use c as a generic constant, the meaning of which depends on the context in which

it is used. By a Taylor expansion, $\hat{R}'(\hat{\beta}) - \mathrm{MISE}'(\beta) = -(\hat{\beta} - \beta_{0n})\mathrm{MISE}''(\tilde{\beta})$ for some $\tilde{\beta}$ lies between $\hat{\beta}$ and β_{0n}. Similar expressions can be obtained for the procedures that estimate $\mathrm{AMISE}(\beta)$ or $\mathrm{AMISE}'(\beta)$. The asymptotic behavior of $\hat{\beta}$ is dominated by $\hat{R}'(\beta_{0n}) - \mathrm{MISE}'(\beta_{0n})$. For each bandwidth selector, we indicate, below, the dominant terms in $\hat{R}'(\beta) - \mathrm{MISE}'(\beta)$ when $\beta = O(n^{-1/5})$.

Except for the procedures of Scott & Terrell (1987), Sheathers & Jones (1991) and JMP (1991), the dominant terms are

$$A = c\beta^3 \int \lambda^4 [|\tilde{\phi}_d(\lambda)|^2 - E\{|\tilde{\phi}_d(\lambda)|^2\}] U(\alpha\lambda) d\lambda = O(n^{-1}\beta^3\alpha^{-9/2}), \tag{4.1}$$

$$B = c\beta^3 \int \lambda^4 |\phi(\lambda)|^2 \{1 - U(\alpha\lambda)\} d\lambda = O(\beta^3\alpha^{2k}),$$

and

$$C = c\beta^3 \{\textstyle\int x^2 w(x) dx\}^2 \displaystyle\int \lambda^4 \phi(\lambda) \tilde{\phi}_d(-\lambda) U(\alpha\lambda) d\lambda = O(n^{-1/2}\beta^3) = O(n^{-11/10}).$$

Here $2k$ is the order of the kernel $w(x)$. In the following discussion, we use the same notation to denote similar terms.

The CV has no bias, and the dominant variation term is A, which is of order $n^{-1}\beta^{-3} = O(n^{-2/5})$. The term C defines the lower bound of the variance of nonparametric bandwidth estimates. (see Fan & Marron (1992)). For sufficiently smooth $f(x)$, C becomes dominant when $\alpha = o(n^{-1/9})$ and $k \geq 3$. Since C is the dominant term for the procedures of Chiu (1991b, 1992), and HJMP (1991), these procedures are asymptotically optimal.

For the procedures of Park & Marron (1990) and HMP (1992) (with $U(\lambda) = W^2(\lambda)$ and $k = 1$), the dominant terms are B and A with the rate $n^{-59/65}$.

The dominant term in the BCV is

$$A = c \int \lambda^4 [|\tilde{\phi}_d(\lambda)|^2 - E\{|\tilde{\phi}_d(\lambda)|^2\}] \{\beta^3 W^2(\beta\lambda) + \beta^4 W(\beta\lambda) W'(\beta\lambda)\lambda)/2\} d\lambda. \tag{4.2}$$

The order of (4.2) is $n^{-2/5}$, which is of the same order as the term for the CV.

The procedure of Sheather & Jones (1991) does not subtract $1/n$ from $|\tilde{\phi}(\lambda)|^2$, and thus has an additional bias term $D = n^{-1}\beta^3 c \int \lambda^4 W(\alpha\lambda) d\lambda$, which is of the order $n^{-1}\beta^3\alpha^{-5}$ with $\alpha = C_{\mathrm{SJ}}\beta^{-5/7}$. They attempt to set C_{SJ} to cancel out the bias term B and D. The term A (4.1) becomes the dominant term when the bias terms could be cancelled.

The procedure of JMP (1991) uses some interesting tricks. Noting that the leading bias term in $\hat{R}(\beta) - \mathrm{MISE}(\beta)$ is of order $\beta^4\alpha^2$, they set $\alpha = c_n\beta^{-2}$ to make the leading bias term be independent of β, and thus does not affect, asymptotically, the bandwidth estimate. Now the leading bias term becomes

$E = c\beta^3\alpha^4 \int \{f^{(4)}(x)\}^2 dx$, which is of order $n^{-47/45}$ with $\alpha = C_{\mathrm{JMP}} n^{-23/45}\beta^{-2}$. They claim that by selecting C_{JMP} to offset the bias terms D and E, the bias becomes negligible, and the dominant terms are

$$A = c\beta^3 \int \lambda^4 [|\tilde{\phi}_d(\lambda)|^2 - E\{|\tilde{\phi}_d(\lambda)|^2\}]\{2W^2(\alpha\lambda) + W(\alpha\lambda)W'(\alpha\lambda)\alpha\lambda\} d\lambda$$

and

$$C = c\beta^3 \int \lambda^4 \phi(\lambda)\tilde{\phi}_d(-\lambda)\{2W^2(\alpha\lambda) + W(\alpha\lambda)W'(\alpha\lambda)\alpha\lambda\} d\lambda.$$

The order of the terms above are $n^{-11/10}$. We should point out that, in general, the constant C_{JMP} is different from the constant obtained from the reference density, and the bias terms D and E would not cancel.

In Table 2, we provide a summary to compare the convergence rates and the dominant terms of the procedures.

Table 2. Theoretic comparison of the procedures; the procedures are arranged in the order of the relative convergence rate of $\hat{\beta}$. The cross-validation does not use α. In the "Bias" column, "-" means that the bias term is negligible.

Procedure	Target	Rate of α	Convergence Rate of $\hat{\beta}$	Dominated	
				Bias	Variance
CS,CSI,CSM	MISE	Adaptive	$n^{-1/2}$	-	C
CPI	$\int\{f''\}^2$	Adaptive	$n^{-1/2}$	-	C
HSJM [1]	$\int\{f''\}^2$	$n^{-1/11}$	$n^{-1/2}$	-	C
JMP [2]	MISE	$n^{-1/9}$	$n^{-1/2}$	$B-D$	$A+C$
SJ [3]	AMISE	$n^{-1/7}$	$n^{-5/14}$	$B-D$	A
PM	AMISE	$n^{-2/13}$	$n^{-4/13}$	B	A
HMP, SCV [4]	MISE	$n^{-2/13}$	$n^{-4/13}$	B	A
BCV	AMISE	$n^{-1/5}$	$n^{-1/10}$	-	A
CV	MISE	-	$n^{-1/10}$	-	A

[1] Another bandwidth of order $n^{-1/9}$ is used in estimating $\int\{f'''\}^2$.
[2] The convergence rate holds when $C_{\mathrm{JMP}}(f)$ is the same as the constant based on the reference density. Otherwise, B and D does not cancel, and the convergence rate is $n^{-4/9}$.
[3] The convergence rate holds when a consistent estimate of $C_{\mathrm{SJ}}(f)$ is available, otherwise $B-D$ is not negligible and the convergence rate is $n^{-2/7}$.
[4] The authors also discussed using a higher order kernel and an other setting of α, but provided no suggestion for the implementation.

5. Simulation and Remarks

As mentioned earlier, the new bandwidth selection procedures have better asymptotic properties. But the empirical evidence is quite different. It is observed in simulation studies that most of these procedures perform much worse than the CV. We will provide some explanation for the inconsistency.

In this section, we summarize the results from an extensive simulation study. We consider 11 densities to cover various situations. The first four densities are (1) the standard normal density, (2) the Cauchy density, (3) the normalized χ^2_4: $(\chi^2_4 - 4)/\sqrt{8}$ and (4) the log-normal density. We also consider seven mixtures of normal distributions.

(1) $0.75N(0,1) + 0.25N(2,1/9)$
(2) $0.5N(0,1) + 0.5N(8,1)$
(3) $0.9N(0,1) + 0.1N(0,100)$
(4) $0.25N(-2,1/16) + 0.5N(0,1) + 0.25N(2,1/16)$
(5) $0.25N(-4,1/16) + 0.5N(0,1) + 0.25N(4,1/16)$
(6) $0.25N(-2,1/16) + 0.5N(0,1) + 0.25N(1,0.01)$
(7) $0.25N(-2,1/16) + 0.5N(0,1) + 0.25N(1,1/16)$

For the first four densities, the amplitude of $\phi(\lambda)$ decays monotonically. The Cauchy density has heavy tails. The χ^2_4 and the log-normal densities are not very smooth. The log-normal density also has a very sharp peak. For the mixtures, the characteristic functions have sidelobes.

Three sample sizes 100, 400, and 1600 were considered. For each case, 200 samples were simulated by using FORTRAN on a Sun-Spark computer. We applied the procedures on each sample to obtain the bandwidth estimates. For each procedure, we followed the simulation setting or the suggestion given by the authors. In particular, we used the normal reference density for the procedures that require a reference density. For all cases, the kernel w is the standard Gaussian density and the bandwidth is the standard deviation.

For ease of discussion, here, we list all procedured considered. In the sequel, w and β are the kernel and the bandwidth for estimating $f(x)$ and u and α are the kernel and the bandwidth for estimating MISE(β), AMISE(β) or $\int\{f''(x)\}^2dx$. Unless indicated otherwise, the kernels w and u are the standard normal density.

CV The least square cross-validation of Bowman (1984) and Rudemo (1982).
CS The stabilized bandwidth selector Chiu (1991b), based on $\hat{\Lambda}$ with $c = 3$.
CSI The stabilized bandwidth selector of Chiu (1992), base on $\hat{\Lambda}_\infty$, see (3.1).
CSM The modified stabilized bandwidth selector of Chiu (1992), based on $\hat{\Lambda}_m$.
CPI The $\sqrt{n}$ consistent plug-in bandwidth estimate (Chiu (1991b)) of β_0, with Λ selected as in SM.

CPA The $\sqrt{n}$ consistent adjusted plug-in bandwidth estimate (Chiu (1991b)) of β_{0n}, with Λ selected as in SM.
BCV The biased cross-validation of Scott & Terrell (1987).
SCV The smoothed cross-validation of HMP (1992).
JMP The procedure of JMP (1991), with $\alpha = Cn^{-23/45}\beta^{-2}$.
PM The procedure of Park & Marron (1990).
SJ The procedure of Sheather & Jones (1991).
HSJM The plug-in estimate of HSJM (1991), with u as given in the paper.

For the procedures SCV, JMP, PM, SJ and HSJM that require a scale estimate, we tried both the sample standard deviation and the sample interquartile range. For CV, CS, CSI, CSM, BCV, SCV and JMP, the minima were searched inside the interval (0.001, 2). For PM and SJ, the roots were searched inside the interval (0.001, 2). These functions were evaluated based on the sample characteristic functions, which were obtained by applying the fast Fourier transform to properly discretized data. The discretization intervals are very small. (See Chiu (1991b, 1992) for more details about the implementation.)

We now briefly describe the simulation results. A more detailed report is available from the author. For the normal density, all new procedures perform much better than the cross-validation. For the Cauchy density, as one might expect, using the sample standard deviation as the scale estimate causes serious problem for SCV, JMP, PM, SJ and HSJM. The procedures HMP, JMP and HSJM have a sizable bias. For the χ^2_4, The biases of HMP, JMP and HSJM become much larger, and they do not decrease as the sample size increases. Except for the mixture $0.9N(0,1)+0.1N(0,100)$, the procedures HMP, JMP and HSJM have a huge bias for all the remaining cases. For the log-normal density, SCV has a large bias for smaller sample sizes. For the first four densities, CS works exceptionally well. These densities have a monotonically decaying $|\phi(\lambda)|$.

The first mixture $0.75N(0,1) + 0.25N(2,1/9)$ has been considered in Scott & Terrell (1987). CSI based on $\hat{\Lambda}_\infty$ has the best performance for larger sample sizes. The procedure SJ.sd (SJ using the sample standard deviation as the scale estimate) has the best performance for smaller sample sizes. Note that $|\phi(\lambda)|^2$ has weak sidelobes, and so CS ignores the sidelobes and has a large bias. The procedure CSM sometimes over-estimates the bandwidth. For this density, the local maximum of the first sidelobe of $|\phi(\lambda)|^2$ is about 0.036. The standard deviation of $|\tilde{\phi}(\lambda)|^2$ is about 0.01, and so it is hard to detect the sidelobe. However, as the sample size increases, the procedure CSM has no trouble to detect the sidelobes. For the mixtures with higher sidelobes, the procedure CSM always works very well.

The results for the third mixture is similar to the case of the Cauchy density.

SJ works well for the first three mixtures but is biased for the last four mixtures. Except for the first and the sixth mixtures, the performance of PM is similar to that of SJ. For the sixth mixture and sample size $n = 1600$, PM has no root in $(0.001, 2)$ in about half of the samples. Both SJ and BCV are seriously biased for the sixth mixture. For the fifth mixture, the procedure BCV almost always does not have a minimum in $(0.001, 2)$.

Table 3. A comparison of the performance of the bandwidth selectors in the simulation study. The procedures are arranged according to the ability of handling various densities. In each column, the number is the number of cases (out of 11 cases) in which the ratio of the sample mean of $\hat{\beta}$ to β_{0n} is bigger than the heading of the column, but less than the heading of the next column, "-" indicates no such case. The numbers in the columns with the heading "10" include the cases that the minimum or the root is outside the interval $[0, 2]$, or when the estimate of $\int \{f^{(k)}(x)\}^2 dx$ is not valid. For the procedures Hall, Jones, Marron, Park and Sheather, two scale estimates, standard deviation and inter-quartile-range, are used, "*" indicated the scale estimate used or suggested in their papers.

	$n = 100$					$n = 400$					$n = 1600$				
Procedure	1.5	2	3	5	10	1.5	2	3	5	10	1.5	2	3	5	10
CSM	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CSI	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CPI	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
CV	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-
SJ, IQR*	2	2	-	-	-	2	-	-	-	-	1	-	-	-	-
SJ, SD	3	2	-	-	-	2	1	-	-	-	2	1	-	-	-
PM, SD*	5	2	-	-	-	3	2	-	-	-	2	1	-	-	-
PM, IQR	3	3	1	-	-	2	1	-	-	-	-	-	1	-	1
CS	-	2	1	3	-	-	3	1	2	-	-	4	2	-	-
BCV	2	1	-	3	2	1	1	-	1	3	-	-	-	-	2
SCV; IQR	1	2	2	3	-	-	3	2	2	-	1	2	2	2	-
SCV; SD*	2	2	4	2	-	4	1	3	2	-	3	2	3	-	1
JMP; SD*	2	2	2	2	1	2	2	2	2	1	2	2	1	3	1
JMP; IQR	-	2	1	3	1	1	2	1	4	-	1	2	1	4	-
HSJM; IQR*	1	1	1	2	2	-	2	2	2	1	1	2	2	1	1
HSJM; SD	2	1	3	1	2	3	-	3	1	2	2	1	2	2	1

In Tables 3, we provide a brief summary regarding the bias of the bandwidth selectors. For each sample size, the table lists the number of times, for the 11 densities, that the ratios of the sample average of the bandwidth estimate to the

optimal bandwidth is between a certain range.

Finally, we make a few remarks about the procedures and provide some practical suggestions. Although CV provides an unbiased estimate, it has a large variation and often selects a very small bandwidth. Therefore, we do not recommend using the CV alone in practice.

Based on the results, the most highly recommended selector is the modified stabilized procedure of Chiu (1992). The procedure works well for most cases, but may have some difficulty in detecting weak sidelobes of $|\phi(\lambda)|$. In such cases, it is usually difficult to tell whether the sidelobes of $\tilde{\phi}(\lambda)$ are real or are due to the sample variation. We prefer to be more conservative. That is, unless we are sure that the sidelobes are real, we would ignore them. The stabilized procedure based on $\hat{\Lambda}_\infty$ always outperforms the CV. The main concern here is that the procedure still selects a smaller bandwidth occasionally, although much less frequently than the CV. We recommend this procedure to the users who are more aggressive and would not want to miss possible feature in the density. In practice, we could use both procedures. When they disagree, there are some marginal sidelobes, and one may prefer to select either procedure.

For densities that are not far different from the normal density, the procedure SJ performs quite well. It could be used if one is sure that the density is not much different from the reference density. In the simulation study, we found that SJ rarely gives very small bandwidths. We also found that SJ does not severely overestimate the optimal bandwidth. This could be a desired property for the users who want to avoid selecting a small bandwidth.

As the simulation results indicate, the idea of using a reference density to set the bandwidth α does not work well when the true density is different from the reference density. The true optimal constant could be very different from the optimal one for the reference density. In this case, we could not expect the procedures HMP, HSJM and JMP to work well. The procedure SJ thus makes a significant improvement over other similar procedures by estimating the bandwidth α in the first stage. Being a modification of the procedure PM, and also using a reference density in setting α, SJ still has a better performance. The improvement shows the importance of using a proper α, which is interesting and merits further study.

Selecting a proper scale estimate is also a critical issue for the procedures that use a reference density. The sample standard deviation should not be used when the density has heavy tails. In other cases, there is no clear indication about the choice between the two scale estimates.

The BCV of Scott & Terrell (1987) does not work well for smaller sample sizes. It also fails for some mixtures even when the sample size is quite large.

Finally, we would like to point out that the sample characteristic function is

a very helpful tool for selecting a bandwidth. It is highly recommended that one should plot $|\tilde{\phi}(\lambda)|^2$ when estimating a density.

Acknowledgement

The helpful comments by the referee are greatly appreciated. This work was partially supported by U.S.A. National Science Foundation Grant DMS-9205684. Part of this work was done while the author visited the Department of Applied Mathematics, National Sun Yat-sen University, supported by the National Science Council, Taiwan.

References

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. *Biometrika* **71**, 353-360.

Brillinger, D. R. (1981). Time Series: Data Analysis and Theory. Expanded edition, Holt, Rinehart, and Winston, New York.

Chiu, S. T. (1990). On the asymptotic distributions of bandwidth estimates. Ann. Statist. **18**, 1696-1711.

Chiu, S. T. (1991a). Some stabilized bandwidth selectors for nonparametric regression. Ann. Statist. **19**, 1528-1546.

Chiu, S. T. (1991b). Bandwidth selection for kernel density estimation. Ann. Statist. **19**, 1883-1905.

Chiu, S. T. (1992). An automatic bandwidth selector for kernel density estimation. Biometrika **79**, 771-782.

Fan, Jianqing and Marron, James S. (1992). Best possible constant for bandwidth selection. Ann. Statist. **20**, 2008-2036.

Hall, P. (1983). Large sample optimality of least squares cross-validation in density estimation. *Ann▷Statist▷***11**, 1156-1174.

Hall, P. and Marron J. S. (1987). Extent to which least-squares cross-validation minimises integrated square error in nonparametric density estimation. *Probab.TheoryRelatedFields***74**, 567-581.

Hall, P., Marron, J. S. and Park, B. U. (1992). Smoothed cross-validation. *Probab▷TheoryRelated Fields***92**, 1-20.

Hall, P., Sheather, S. J., Jones, M. C. and Marron, J. S. (1991). On optimal data-based bandwidth selection in kernel density estimation. *Biometrika***78**, 263-269.

Jones, M. C., Marron, J. S. and Park, B. U. (1991). A simple root n bandwidth selector. Ann. Statist. **19**, 1919-1932.

Park, B. U. and Marron, J. S. (1990). Comparison of data-driven bandwidth selectors. *J▷Amer▷ Statist▷Assoc▷***85**, 66-72.

Rice, J. (1984). Bandwidth choice for nonparametric regression. *Ann▷Statist▷***12**, 1215-1230.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann▷ Math▷Statist▷***27**, 832-837.

Rosenblatt, M. (1971). Curve estimates. *Ann▷Math▷Statist▷***42**, 1815-1842.

Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators. *Scand▷J▷ Statist▷***9**, 65-78.

Scott, D. W. and Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation. *J▷Amer▷Statist▷Assoc▷***82**, 1131-1146.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. J. Roy. Statist. Soc. Ser.B **53**, 683-690.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *J▷Roy▷Statist▷Soc▷Ser▷B***47**, 1-21.

Silverman, B. W. (1986). *DensityEstimationforStatisticsandDataAnalysis.* Chapman and Hall, London.

Department of Statistics, Colorado State University, Fort Collins, CO 80523-1877, U.S.A.

(Received July 1993; accepted June 1995)