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

Advances in Data-Driven Bandwidth Selection +

Byeong U. Park*

ABSTRACT

Considerable progress on the problem of data-driven bandwidth selection in kernel density estimation has been made recently. The goal of this paper is to provide an introduction to the methods currently available, with discussion at both a practical and a nontechnical theoretical level. The main setting considered here is global bandwidth kernel estimation, but some recent results on variable bandwidth kernel estimation are also included.

1. Introduction

The nonparameteric kernel density estimator uses a sample $X_1, \cdots, X_n$ from a density f, to estimte the curve $f(x)$ by

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i), \tag{1.1}$$

where $K_h(x) = K(x/h)/h$, K is called the kernel function, and h is called the *bandwidth* or smoothing parameter. Both K and h are to be selected by the user. When K is a probability density function, so is $\hat{f}_h$ and this is usually the case preferred in applications. It is understood that $\hat{f}_h$ puts probability mass, according to K_h, around the observed data points X_i's, so the bandwidth controls the degree of smoothing applied to the data by the kernel density estimate.

Like all other types of nonparametric curve estimators, the choice of the bandwidth is the central issue in the application of the kernel density estimator. This is demonstrated in Fig. 1. In Fig. 1a, the curve is the true underlying density function. Fig. 1a also includes a kernel density estimate at the bottom of the plot. This estimate is based on a very samll bandwidth. It is *not* given for comparison with the true curve, but as a descriptor of the 100 simulated data. It is scaled down to one sixth of its original height to prevent its graph from interfering with the graph of the true density. Fig. 1b, 1c and 1d show the same true density curve together with kernel density estimates, as shown as the thick curves, corresponding to different bandwidths, as shown by the curves representing K_h, which appear at the bottom of each plot. Note that in Fig. 1b, the bandwidth is quite narrow, with the result that there are not enough observations involved in the construction of $\hat{f}_h$ at each point x, and the resulting estimate is excessively subject to sample variability, i.e. is too wiggly. This is improved in Fig. 1c, where a larger bandwidth has been used. In Fig. 1d, the bandwidth is so large that observations from too far away are involved in the construction of $\hat{f}_h(x)$, with the effect of introducing some bias, or in other words the bimodal feature of the underlying density curve has been smoothed away.

+ This research was suported by Korean Science and Engineering Foundation 88-07-23-02

* Seoul National University

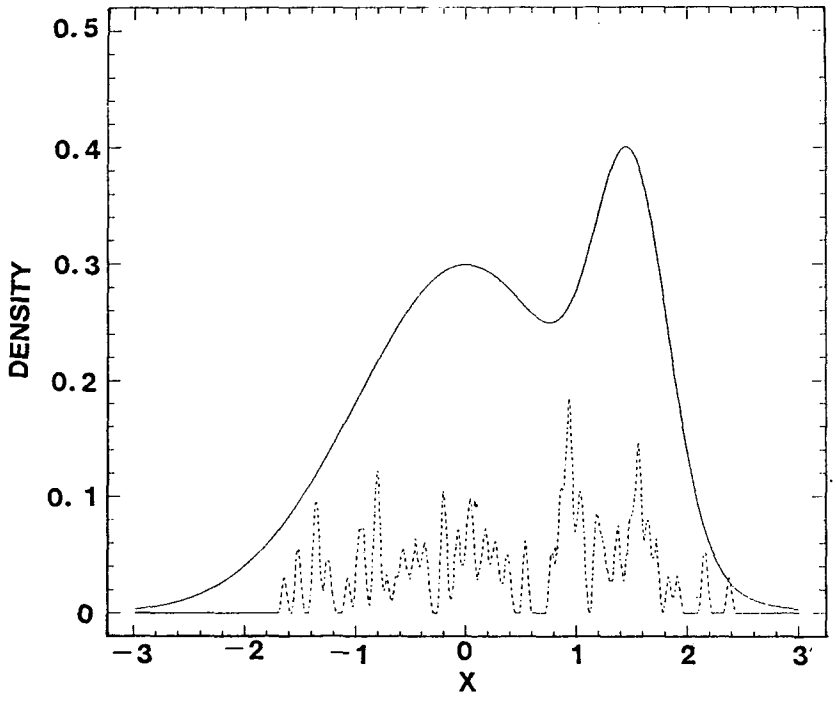
0.5
0.4
0.3
0.2
0.1
0
DENSITY
−3
−2
−1
0
1
2
3
X

Fig. 1a.

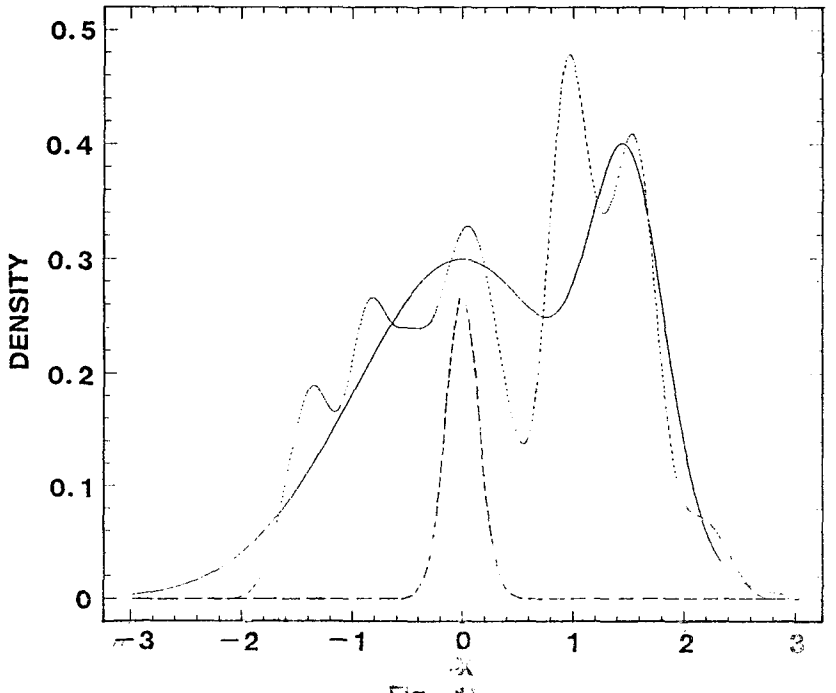
0.5
0.4
0.3
0.2
0.1
0
DENSITY
−3
−2
−1
0
1
2
3
X

Fig. 1b.

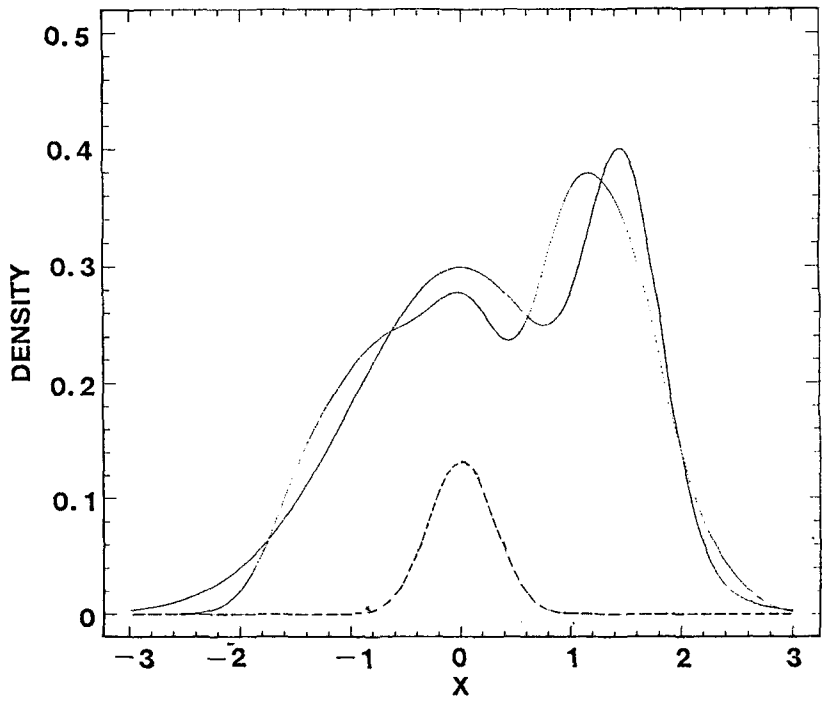
0.5
0.4
0.3
0.2
0.1
0
DENSITY
−3
−2
−1
0
1
2
3
X

Fig. 1c.

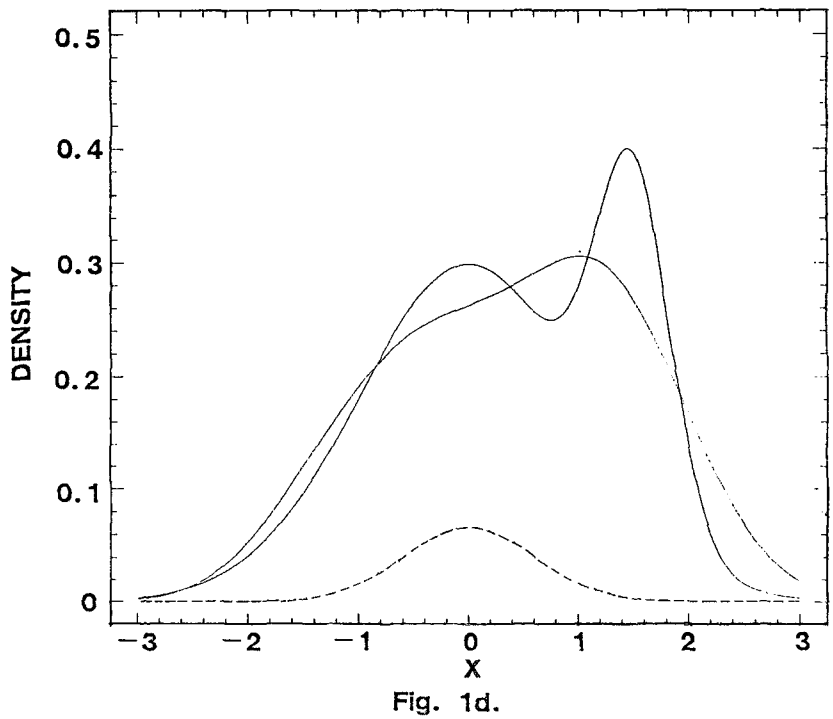
0.5
0.4
0.3
0.2
0.1
0
DENSITY
−3
−2
−1
0
1
2
3
X

Fig. 1d.

The great advantage of kernel density estimator, also possessed by all other types of nonparametric curve estimators, is that this estimator does not hide features in the true densities, i.e. does not impose structures on the data, by allowing the data to speak for themselves. For example, think of doing $N(\mu, \sigma^2)$ parametric estimation using the data in Fig. 1. In this case, one can not catch the bimodal feature of the true density since the parametric approach *does* impose unimodal structure on the data. For interesting collections of effective data analyses carried out by nonparametric density estimation in general and by the kernel method in particular, see Silverman(1986).

However there is an obstacle which one has to overcome for practical application of this powerful method, which is that the bandwidth must be chosen. Effective data analysis has often been done by a subjective, trial and error approach to the choice of the bandwidth, which consists of looking at several different plots representing different amounts of smoothness. While this approach certainly allows one to learn much about the data set, it can never be used to convince a skeptic since broad range of alternative choices are not considered. This leads one to search efficient and objective methods of using the data to determine the amount of smoothing.

This paper reviews considerable recent progress on the problem of data-based selection of the bandwidth in kernel density estimation. Attention will be focussed here on the methods proposed since 1988. However some of the methods proposed up until 1987 will be discussed here also to motivate the ideas and to demonstrate the effectiveness of the *new* methods. For detailed discussion on such *old* methods, see the survey by Marron(1988).

There is an important class of kernel density estimators other than the type defined in (1.1), which use different amounts of smoothing at different locations. Note that the estimator defined in (1.1) uses uniform(*or global*) bandwidth for all x values, and certainly this type of estimator does not perform well when the underlying density has features which require different amounts of smooting at different locations. The bandwidth selection problem for the *variable* bandwidth kernel estimators is much harder than for the global bandwidth kernel estimators since there are essentially infinitely many parameters to choose. Hence there have been relatively fewer attempts for searching data-driven methods to choose the variable bandwidth. Recently, several promising methods have been proposed, which are based on transformations of original data. These are discussed in Section 4.

Section 2 of this paper introduces two theoretically optimal global bandwidths, both based on squared error performance measures, that are often discussed in the literature and most data-driven bandwidth selectors aim for. Section 3 introduces and discusses various method for data-driven bandwidth selection in global bandwidth kernel density estimation. Section 5 discusses future research.

2. Theoretically optimal Bandwidths

The usual error criteria, which assess how well $\hat{f}_h$ in (1.1) estimates f, are the integrated squared error

$$ISE(h)=\int [\hat{f}_h(x)-f(x)]^2dx, \tag{2.1}$$

and its expected value(for fixed h), the mean integrated squared error

$$MISE(h)=E[ISE(h)]. \tag{2.2}$$

Note that the integration in both (2.1) and (2.2) reflects the global, rather than the local, nature of investigations. Related criteria are the integrated absolute error

$$IAE(h)=\int |\hat{f}_h(x)-f(x)|\,dx,$$

and its expected value

$$MIAE(h)=E[IAE(h)].$$

Although Devroye and Gyorfi(1984) point out a number of reasons for using these absolute error type criteria, their use has not gained wide acceptance, one reason being that squared error criteria are much easier to work with from a technical point of view. For this reason, all of the real theoretical breakthroughs in density estimation have come first from considering sqaured error criteria, in the hope that the idea may be extended to the absolute error case with much more work.

Most published theoretical work takes h_0, the bandwidth that minimizes MISE of $\hat{f}_h$, as the theoretically optimal bandwidth, rather than $\hat{h}_0$, the(random) bandwidth that minimizes ISE specific to the data set at hand. There is no consensus about which should be taken as the right bandwidth to aim for, though. Note that

$$E[ISE(\hat{h}_0)]\leq E[ISE(h_0)],$$

and this means that the ideal bandwidth, assuming that best estimation of f is truely the objective, is not h_0, but $\hat{h}_0$. However, $\hat{h}_0$, being a random quantity itself, is a much harder target to estimate than h_0. In fact, Hall and Marron(1991) show that the best possible relative error rate of convergence of any data-driven bandwidth selector to $\hat{h}_0$ is of order $n^{-1/10}$, much slower thatn $n^{-1/2}$, the rate to h_0. Furthermore, Jones and Kappenman(1991) argued that estimating h_0 well remains a particularly useful way to go about seeking data-driven bandwidth selectors which perform well in terms of $\hat{h}_0$, too. This outlook leads back to h_0 being suitable to aim for. See Jones(1991) for further discussion of this issue.

The usual way of accessing the theoretical performance of various data-driven bandwidth selectors($\hat{h}$'s) is to compare them with h_0. It is well established that direct comparison with h_0 is the key to understanding the performance of $\hat{f}_{\hat{h}}$ as an estimate of f relative to that of $\hat{f}_{h_0}$. In particular, by simple Taylor series expansions, the asymptotic properties of $\hat{h}/h_0$ can be directly translated into analogous asymptotic properties of $MISE(\hat{h})/MISE(h_0)$(see, for example, Park and Marron 1990).

For later use, let us introduce here the usual asymptotic version of (2.2) which we call AMISE (for asymptotic MISE) :

$$AMISE(h)=(nh)^{-1}R(K)+h^4\sigma_K^4R(f'')/4. \tag{2.3}$$

Here and below, $\sigma_K^2=\int x^2K(x)dx$ and $R(g)=\int g^2(x)dx$. Note that as $n\to\infty$ and $h\to 0$, with $nh\to\infty$, under some conditions on K and f,

$$MISE(h)=AMISE(h)+o(AMISE(h))$$

(see, for example, Silverman 1986, Section 3.3). The two terms in AMISE provides a very clean asymptotic summary of the smoothing problem. Recall from Fig. 1, that too small a bandwidth results in too much sample variability. Note that this is reflected by the first term(unusally called the variance term) in AMISE becoming too large. On the other hand, the fact that too large a bandwidth gives too much bias, is reflected by the second term(usually called the bias

term) which gets too large in this case.

Some of the data-driven bandwidth selectors are motivated from the asymptotic representation (2.3) of MISE. This is the result of a tendency to think of h_1, the bandwidth that minimizes AMISE, as being the same as h_0. It is seen by Marron and Wand(1991) that h_1 usually begins to provide a decent approximation to h_0 for sample sizes between 100 and 1000, but in some cases the sample size needs to be close to one million for good approximation.

3. Selection Methods

3-1. Preliminaries

Most of data-driven bandwidth selectors are divided into two groups according to their target functions, one of which is MISE and the other is the asymptotic representation of MISE. Note that none of these two target functions yields an immediately practicable method for choosing h since both of them have some dependence on the unknown f. We can, however, estimate the f-dependent quantities and then choose h on the basis of the corresponding estimated target functions. In particular, functionals of f of the form $\theta_m=R(f^{(m)})$, $m=0,1,2,\cdots$, where $f^{(m)}$ is the m-th derivative of f, prove to be of particular importance. We can estimate each θ_m by quantities involving $\hat{f}_g$ as investigated by Hall and Marron(1987b), where $\hat{f}_g$ is a kernel density estimator with bandwidth now represented by g and kernel L(allowed to be different from h and K because estimation of this integral is a different smoothing problem form estimation of f). Both alternatives derived as potentially good estimators by Hall and Marron are :

$$\begin{aligned}\hat{\theta}_m(g)&=(n-1)^{-1}nR(\hat{f}_g^{(m)})-(n-1)^{-1}g^{-2m-1}R(L^{(m)})\\&=\{(n-1)n\}^{-1}g^{-2m-1}(-1)^m\ \Sigma\Sigma_{i\neq j}(L*L)^{(2m)}\{g^{-1}(X_i-X_j)\}\end{aligned}\tag{3.1}$$

(where * denotes convolution) and

$$\begin{aligned}\hat{\hat{\theta}}_m(g)&=(-1)^m n^{-1}\sum_{i=1}^{n}\hat{f}_{g,i}^{(2m)}(X_i)\\&=\{(n-1)n\}^{-1}g^{-2m-1}(-1)^m\ \Sigma\Sigma_{i\neq j}L^{(2m)}\{g^{-1}(X_i-X_j)\}\end{aligned}\tag{3.2}$$

where $\hat{f}_{g,i}$ is a kernel density estimator using only $(n-1)$ of the sample values leaving out X_i.

Both of the estimators above have a "cross-validatory" element to them, in that a non-stochastic term arising from "i=j" terms is deleted, on the grounds that it causes unnecessary bias. Jones and Sheather(1991) investigates the reintroduction of the non-stochastic term and shows how it can be used to improve the cross-validatory estimators. Their "non-cross-validatory" estimators are ;

$$\tilde{\theta}_m(g)=R(\hat{f}_g^{(m)})\tag{3.3}$$

and

$$\tilde{\tilde{\theta}}_m(g)=(-1)^m n^{-1}\sum_{i=1}^{n}\hat{f}_g^{(2m)}(X_i).\tag{3.4}$$

The key to successful employment of such non-cross-validatory estimation procedures is the recognition that the non-stochastic term bias has the opposite sign to the bias due to smoothing. Further more, it is possible to utilize the freedom, not available if $g=h$, to choose the associated

bandwidth g to make these bias terms cancel and in this way to improve the MSE properties of the resulting estimators.

A kernel L is said to be of order r if

$$\int L(x)dx=1,\ \int x^j L(x)dx=0 \text{ for } j=1,\cdots r-1,\ \int x^r L(x)dx\neq 0.$$

Note that if L is a probability density function, it is of order 2. For the good asymptotic properties of the estimators defined in (3.1)-(3.4), higher order(greater than 2) kernels are often used.

3-2. MISE Based Methods

3-2-1. Least Squares Cross-Validation

The most widely studied bandwidth selector is least squares cross-validation, proposed by Rudemo(1982) and Bowman(1984). Noting that

$$MISE(h)=E[R(\hat{f}_h)]-2E[\int \hat{f}_h(x)f(x)dx]+R(f) \tag{3.5}$$

and the last term $R(f)$ does not depend on h, an unbiased estimate of $MISE(h)-R(f)$,

$$CV(h)=R(\hat{f}_h)-2\hat{\hat{\theta}}_0(h), \tag{3.6}$$

with $L=K$ and $g=h$ for $\hat{\hat{\theta}}_0$, is minimized to yield the least squares cross-validatory choice $\hat{h}_{CV}$, say. Here, $\hat{\hat{\theta}}_0(h)$ is thought of as an estimate $E\int \hat{f}_h(x)f(x)dx$ rather than of $R(f)$.

The main strength of this bandwidth is that it is asymptotically correct under very weak smoothness assumptions on the underlying density(see Hall 1983, and Stone 1984). However, it has been seen that CV suffers a great deal of sample variability in the sense that for different data sets from the same distributions, it will typically give much different answers. This has been quantified asymptotically by Hall and Marron(1987a) and Scott and Terrell(1987), who show that the relative rate of convergence of h_{CV} to either of h_0 or h_0 is of the excruciatingly slow order of $n^{-1/10}$. For other drawbacks discussion of this bandwidth, see Marron(1988).

3-2-2. Complete Cross-Validation

Complete cross-validation(CCV), proposed by Jones and Kappenman(1991), estimates the entire MISE function, as opposed to CV's estimation of $MISE(h)-R(f)$. Taking $\hat{\hat{\theta}}_0(h)$ as an estimate of $R(f)$, one may estimate the entire MISE function by $R(\hat{f}_h)-\hat{\hat{\theta}}_0(h)$. But this has a bias since

$$E[\hat{\hat{\theta}}_0(h)]=R(f)-\sigma_K^2h^2R(f')/2+\delta_Kh^4R(f'')/24+o(h^4)$$

where $\delta_K=\int x^4K(x)dx$. Observing that

$$E[\hat{\hat{\theta}}_1(h)]=R(f')-\sigma_K^2h^2R(f'')/2+o(h^2)$$

and

$$E[\hat{\hat{\theta}}_2(h)]=R(f'')+o(1)$$

(see Hall and Marron 1987b), this bias can be reduced to $o(h^4)$ by using, as an estimate of $MISE(h)$,

$$CCV(h)=R(\hat{f}_h)-\hat{\hat{\theta}}_0(h)+\sigma_K^2h^2\hat{\hat{\theta}}_1(h)/2+(6\sigma_K^4-\delta_K)h^4\hat{\hat{\theta}}_2(h)/24. \tag{3.7}$$

The complete cross-validatory choice of h is the minimizer of (3.7). It has been shown by Jones and Kappenman(1991) that the relative rate of convergence of the CCV bandwidth selector to h_0(or $\hat{h}_0$) is the same as, but the constant multiplier of the rate of convergence is slightly less than(for $K=\phi$, the standard normal density), of h_{CV}.

3-2-3. Smoothed Cross-Validation

This was proposed and studied by Hall, Marron and Park(1991). The essential idea of smoothed cross-validation(SCV) is related to CV. Note that when there are no duplications among the data, which happens with probability one with truely continuous data, (3.6) can be written as

$$CV(h)=(nh)^{-1}R(K)+n^{-1}(n-1)^{-1}\ \Sigma\Sigma_{i\neq j}(K_h*K_h-2K_h+K_0)(X_i-X_j) \tag{3.8}$$

(modulo $n\simeq n-1$) where here and below K_0 denotes the Dirac delta function. This reveals that unacceptably large noise in $\hat{h}_{CV}$ is created by the second part in the right hand side of (3.8) since it is the only random part of $CV(h)$. The smoothed cross-validation criterion addresses this problem by modifying this term, by a type of "presmoothing" of the differences X_i-X_j which has better stability properties. In particular, $\hat{h}_{SCV}$ is defined to be the minimizer of

$$SCV_g(h)=(nh)^{-1}R(K)+\hat{B}_g(h), \tag{3.9}$$

where

$$\hat{B}_g(h)=n^{-1}(n-1)^{-1}\sum_{i=1}^{n}\sum_{j=1}^{n}\{(K_h*K_h-2K_h+K_0)*L_g*L_g\}(X_i-X_j), \tag{3.10}$$

for the possibly different kernel function L and bandwidth g.

Remark : The original version of (3.10) defined in Hall, Marron and Park(1991) does not have "i=j" terms. But Jones, Marron and Park(1991) have shown that there can be a substantial advantage to leaving these terms in.

The reason that L_g*L_g is used here(instead of using simply L_g) is that the second part in the right hand side of (3.8) can be viewed as an estimate of integrated squared bias, $B(h)=\int(K_h*f-f)^2$, and use of L_g*L_g yields an intuitive estimate of this quantity. In particular, note that

$$\hat{B}_g(h)=\int(K_h*\hat{f}_g-\hat{f}_g)^2$$

where $\hat{f}_g$ denotes the kernel estimator with kernel L and bandwidth g. Another compelling feature of SCV is that it is essentially the same as a smoothed bootstrap estimate of MISE. The idea of using a smoothed bootstrap estimate of MISE was first proposed by Faraway and Jhun(1990). However, Faraway and Jhun did not recognized that the smoothed bootstrap estimate could be calculated directly, and used simulation instead. Taylor(1989) proposed the same idea, and did point out the exact form of the estimate, but his derivatiion is only in the special case $L=K$ and $g=h$. It has been seen by Hall, Marron and Park(1991) that use of higher order kernel L and an elaborate choice of g entail the very fast $n^{-1/2}$ relative rate of convergence of $\hat{h}_{SCV}$ to h_0, which has been shown to be the best possible by Hall and Marron(1991).

3-2-4. Bandwidth Factorizied SCV

The idea of bandwidth factorization was introduced by Jones, Marron and Park(1991), where the main idea was illustrated using SCV criterion, but it was pointed out that essentially the same results are easily established for the methodology of Hall, Sheather, Jones and Marron (1991). Note that higher order kernels are required for h_{SCV} to get an $n^{-1/2}$ relative rate of convergence(L needs to be of order 6). But, as is discussed by Marron and Wand(1991), the use of higher order kernels is unappealing since, while they are excellent in the limit, quite large sample sizes(even the millions is not sufficient in some cases) seems to be required all too often before their beneficial effects begin to appear and become dominant. The main advantage of bandwidth factorization is that it allows the fastest possible rate of convergence with the use of only nonnegative kernels, i.e. kernels of order 2, at all stages of the selection process.

The main idea consists of allowing g to depend on h, which was not considered(except in one special case, $g=h$) by Hall, Marron and Park(1991). The dependence considered is the factorization

$$g=Cn^{p}h^{m} \tag{3.11}$$

for various constants C, p and m. Note that the case $m=0$ corresponds to the ordinary SCV discussed above. It has been shown by Jones, Marron and Park(1991) that when $m=-2$ and $p=-23/45$, there is an important type of cancellation in the bias of the resulting bandwidth, which is the key to $n^{-1/2}$ convergence even in the case of nonnegative kernel L.

3-2-5. Stabilized Bandwidth Selector

The stabilized bandwidth selector, as was proposed by Chiu(1991), is also related to CV and is based on the use of related Fourier transform method. Note that $CV(h)$ is approximately equal to

$$\pi^{-1}\int_0^{\infty} |\tilde{\phi}(\lambda)|^2[\tilde{K}(h\lambda)-2\tilde{K}(h\lambda)]d\lambda+2K(o)/(nh)$$

(see Silverman 1986, pp.62-63) where $\tilde{\phi}$ and $\tilde{K}$ are the characteristic functions of the sample distribution function F_n and K, respectively. This reveals that the large sample variability of $CV(h)$ is created by $|\tilde{\phi}(\lambda)|^2$, especially when $|\phi(\lambda)|$ is negligible. This observation suggests that the variation of h_{CV} can be reduced by modifying $|\tilde{\phi}(\lambda)|^2$ when $|\phi(\lambda)|$ becomes negligible. Chiu(1991) proposed to modify $|\tilde{\phi}(\lambda)|^2$ into $|\tilde{\phi}(\lambda)|^2 I(\lambda\leq\Lambda)+n^{-1}I(\lambda>\Lambda)$ where Λ is the first λ such that $|\tilde{\phi}(\lambda)|^2\leq c/n$ for some constant $c>1$. The reason of using the factor n^{-1} is that when $|\phi(\lambda)|$ is negligible, $|\phi(\lambda)|^2$ is approximately an exponential random variable with mean n^{-1}. So $n|\tilde{\phi}(\lambda)|^2$ is compared with a critical value c to decide when $|\phi(\lambda)|$ becomes negligible. For example $c=3\approx log_e(0.05)$ is approximately equal to the 95-th percentile of the exponential distribution with mean 1. The stabilized bandwidth selector is now defined to be the minimizer of

$$\begin{aligned} S(h)=&\pi^{-1}\int_0^{\Lambda} |\tilde{\phi}(\lambda)|^2\{\tilde{K}^2(h\lambda)-2\tilde{K}(h\lambda)\}d\lambda \\ &+(\pi n)^{-1}\int_{\Lambda}^{\infty}\{\tilde{K}^2(h\lambda)-2\tilde{K}(h\lambda)\}+d\lambda+2K(o)4(nh), \end{aligned}$$

which is equal to

$$(nh)^{-1}R(K)+\pi^{-1}\int_0^{\Lambda}\{|\tilde{\phi}(\lambda)|^2-n^{-1}\}\{\tilde{K}^2(h\lambda)-2\tilde{K}(h\lambda)\}d\lambda. \tag{3.12}$$

There is a close relationship between the stabilized criterion $S(h)$ and $SCV_g(h)$. Note that the smoothed cross-validation criterion can be written as

$$SCV_g(h)=(nh)^{-1}R(K)+\pi^{-1}\int_0^{\infty}\{|\widetilde{\phi}(\lambda)|^2-n^{-1}\}\{1-\widetilde{K}(h\lambda)\}^2\widetilde{L}^2(g\lambda)d\lambda, \qquad (3.13)$$

where $\widetilde{L}(\lambda)$ is the characteristic function of the kernel L. This observation reveals that the stabilized criterion $S(h)$ is equivalent to $SCV_g(h)$ when L is of infinite order, since the indicator function $I(-\Lambda\leq\lambda\leq\Lambda)$ can be viewed as the Fourier transform of an infinite order kernel with the bandwidth proportional to $1/\Lambda$. It has been proved by Chiu(1991) that the relative rate of convergence of the stabilized bandwidth selector to h_0 is again $n^{-1/2}$(as is expected from the close connection to SCV). The main strength of this bandwidth selector is that its constant multiplier of the rate of convergence is also best possible, which has been shown by Fan and Marron(1991). However, it should be noted that the best constant can be also achieved by SCV and the bandwidth factorized SCV with infinite order kernels L, although this fact is not mentioned in the corresponding papers.

3-3. AMISE Based Methods

3-3-1. Biased Cross-Validation

This was proposed and studied by Scott and Terrell(1987). The essential idea is to minimize the following estimate of $AMISE(h)$,

$$BCV(h)=(nh)^{-1}R(K)+h^4\sigma_K^4\hat{\theta}_2(h) \qquad (3.14)$$

where $\hat{\theta}_2(h)$ uses the same kernel K and bandwidth h as when estimating f itself. Scott and Terrell(1987) show that the biased cross-validated bandwidth selector has sample variability with the same rate of convergence as $\hat{h}_{CV}$, but with a typically much smaller constant multiplier.

3-3-2. Improved Versions of BCV

Recall that $\hat{\theta}_2(h)$ in (3.14) does not include the diagonals "i=j" terms. The reason for this, as argued in Scott and Terrell(1987), is that inclusion of the diagonal terms introduces unnecessary bias, $n^{-1}h^{-5}K''*K''(0)$, which is not negligible since the optimal bandwidth is of order $n^{-1/5}$. However, this is true only when one sets $g=h$, i.e. uses the same bandwidth both for estimating θ_2 and f. Furthermore, it is observed that $\hat{\theta}_2(h)$ often has negative values(due to omission of the diagonal terms), which is unreasonable since $\theta_2>0$.

The reintroduction of the diagonal terms is investigated by Sheather and Jones(1991), where they show how the diagonal terms can be used to advantage to improve BCV(and other bandwidth selection procedures) in terms of theory, computation and simulation practice. In fact, the use of the non-cross-validatory estimators, $\widetilde{\theta}_m(g)$ or $\widetilde{\widetilde{\theta}}_m(g)$, is shown to improve the relative rate of convergence $n^{-1/10}$ of BCV to $n^{-5/14}$ with careful choices of g, even retaining the use of nonnegative kernels L. It is pointed out that the use of higher order kernels L affords a further improvement to $n^{-2/5}$, which is known to be the best possible based on the objective function (2.3). Their simulation results also reveal that, on the whole, the non-cross-validatory bandwidth selectors provide a worthwhile improvement over the cross-validatory counterparts.

As mentioned above, bandwidth selectos based on the asymptotic representation (2.3) of MISE have the rate of convergence $n^{-2/5}$ at their best. This is because such bandwidth selectors aim at h_1, not h_0, and the fastest relative rate of convergence of h_1 to h_0 is $n^{-2/5}$(see Lemma 5.2 of Park 1989, for example). In this respect, the asymptotic expansion of MISE to more terms than is the case of (2.3) may afford improvement. Noting that the first part of (2.3) is a very

good approximation of the integrated variance of $\hat{f}_h$(see Section 4 of Marron and Wand 1991, for example), one might attempt to expand one more term only in the integrated squared bias, which results in

$$AMISE(h) = (nh)^{-1}R(K) + h^4\sigma_K^4\theta_2/4 - h^6\sigma_K^2\delta_K\theta_3/24. \quad (3.15)$$

Hall, Sheather, Jones and Marron(1991) pursue this approach. They showed that any of the estimators of θ_2 and θ_3, defined in (3.1)-(3.4), ensure the fastest $n^{-1/2}$ relative rate of convergence to h_0 for the bandwidth which minimizes the corresponding estimated $AMISE(h)$. Furthermore, it can be shown that their bandwidth selector achieves the best constant also. For this, L needs to be of order 6 when a cross-validatory estimator of θ_2 is used, while a fourth order kernel is sufficient for the non-cross-validatory counterpart. As discussed above, here also, non-cross-validatory estimators are preferred for their practical performance.

The two major weeknesses of this approach are the fact that two bias terms are needed in the AMISE expression, and also the requirement of higher order kernels. Marron and Wand (1991) observed from MISE and AMISE comparsion that there is very little to be gained in practice through the use of the extra bias term. Since practical implementation requires the addition of noise through estimation of the extra quantity θ_3, it is highly unlikely that there is net gain from inclusion of this extra bias term. This fact has been born out in the simulation study of Hall, Sheather, Jones and Marron(1991).

These two objections can be overcome by means of bandwidth factorization, the idea of which was illustrated using SCV criterion in Section 3.2.4. In fact, the same choice $g = Cn^{-23/45}h^{-2}$, as of the bandwidth factorized SCV, if used for $\tilde{\theta}_2$(or $\tilde{\tilde{\theta}}_2$) being plugged into AMISE in (2.3), gives the same type of cancellation, and again results in an $n^{-1/2}$ rate of convergence. As in the case of SCV, this also needs only a nonnegative kerenel in the pilot estimator $\tilde{\theta}_2$(or $\tilde{\tilde{\theta}}_2$).

3-3-3. Solve-the-equation Methods

The so called "solve-the-equation" methods are closely related to BCV. Here, we still work with AMISE in (2.3) but minimization with respect to h, yielding the usual formula

$$h_1 = [R(K)/(\sigma_K^4\theta_2)]^{1/5}n^{-1/5}, \quad (3.16)$$

is done prior to estimation of θ_2. The most primitive approach in this direction is to plug $\hat{\theta}_2(h)$ or $\hat{\tilde{\theta}}_2(h)$, which uses the kernel K, into (3.16) and then, setting $h_1 = h$, solve the equation

$$h = [R(K)/(\sigma_K^4\hat{\theta}_2(h))]^{1/5}n^{-1/5}. \quad (3.17)$$

The original idea of this approach is due to Scott, Tapia and Thompson(1977) and its finite sample properties were investigated by Scott and Factor(1981). The relative rate of convergence of this bandwidth selector to h_0 is known to be of order $n^{-1/10}$(see Jones and Kappenman 1991, for example).

A more effective method was introduced by Park and Marron(1990), the essential idea of which is due to Sheather(1983, 1986) where the case of pointwise density estimation is considered. The idea is to use bandwidth g, different from h, in estimating θ_2, but in the form of a reasonable representation in terms of h. Using such a representation $g(h)$, say, Park and Marron's bandwidth selector is taken to be the root of the equation

$$h = [R(K)/\{\sigma_K^4\hat{\theta}_2(g(h))\}]^{1/5}n^{-1/5}. \quad (3.18)$$

This bandwidth selector is known to have a rather fast $n^{-4/13}$ relative rate of convergence, and

have good finite sample properties, compared to CV and BCV.

Effectiveness of the non-cross-validatory estimators $\breve{\theta}_2$ and $\tilde{\tilde{\theta}}_2$, discussed in the previous section, still applies to the present case. It was observed by Sheather and Jones(1991) that, when one use non-cross-validatory estimators with an appropriate representation for g, the relative rate of convergence can be improved again to $n^{-5/14}$. Finite sample property of this bandwidth selector was compared with Park and Marron's selector and Sheather and Jones's improved BCV version. The results indicate that it is better than both of them in all settings considered in the simulation study.

One may think that the bandwidth factorization ideas may be applicable to the above solve-the-equation methods for $n^{-1/2}$ rate of convergence even with nonnegative pilot kernel estimators. To say the truth, the same set of ideas do *not* work in the same way to the solve-the equation methods. The reason for this is that the cancellation effect, which is the key to $n^{-1/2}$ convergence, only applies to those methods which involve minimization.

3-3-4. Chiu's Adjusted Plug-in Bandwidth Selector

The adjusted plug-in method, proposed by Chiu(1991), is very similar to the method of Hall, Sheather, Jones and Marron(1991) in that it is based on two term bias expansion of $MISE(h)$. As the stabilized bandwidth selector, this is also based on Fourier transform methods. In particular, note that

$$\theta_2=(2\pi)^{-1}\int \lambda^4 \mid \phi(\lambda) \mid^2 d\lambda$$

and $E \mid \tilde{\phi}(\lambda) \mid^2=n^{-1}(n-1)^{-1} \mid \phi(\lambda) \mid^2+n^{-1}$. By the similar reason discussed in Section 3.2.5, a potentially good estimator of θ_2 is given by

$$\bar{\theta}_2=\pi^{-1}\int_0^{\Lambda} \lambda^4\{ \mid \tilde{\phi}(\lambda) \mid^2-1/n\}d\lambda$$

where Λ is defined in Section 3.2.5. This gives a bandwidth selector

$$\hat{h}_P=[R(K)/(\sigma_K^4\bar{\theta}_2)]^{1/5}n^{-1/5}.$$

However, as discussed in Section 3.3.2, this bandwidth aims at h_1 which itself converges to h_0 at only the rate $n^{-2/5}$. For this reason, h_P is adjusted according to inclusion of the second bias term in the MISE expansion, yielding the bandwidth selector

$$\hat{h}_{AP}=\hat{h}_P+A_1(\hat{h}_P)/A_2(\hat{h}_P)$$

where

$$A_1(h)=(24\pi)^{-1}n^{3/5}h^6\sigma_K^2\delta_K\int_0^{\Lambda} \lambda^6[\mid \tilde{\phi}(\lambda) \mid^2-1/n]d\lambda$$

and

$$A_2(h)=n^{4/5}h^4\bar{\theta}_2\sigma_K^4/4+n^{-1/5}h^{-1}R(K)-n^{1/5}A_1(h).$$

Chiu(1991) showed that $\hat{h}_{AP}$ and the stabilized bandwidth selector have the same limiting distribution.

3-4. Pilot Estimation

Most of the bandwidth selectors, which have relative rates of convergence faster than $n^{-1/10}$, require tuning for their effective use. For example, one needs to decide which constant C to

use to implement the bandwidth factorized SCV methodlogoy. This is particularly important here since the fast $n^{-1/2}$ rate of convergence relies upon the use of proper tuning constant C. Of course, there is a theoretically optimal choice of C which ensures the desired rate of convergence. However, it depends on the unknown f in temrs of θ_m's, so is unavailable. Another example in this direction is that effective representations $g(h)$, ensuring the desired rates of convergence of the solve-the-equation bandwidth selectors, are in the form of Ch^p with C depending on θ_m's.

A simple means of doing this tuning, in real data situations, is to use these best tuning values, but with the unknown functionals replaced by some reference distribution counterparts. An often considered choice of reference distribution is the normal distribution with mean zero and an estimated variance (for scale invariance reasons). This reference distribution approach is good enough for the asymptotic performance, in terms of rate of convergence, of all the bandwidth selectors, discussed above, except the non-cross-validatory bandwidth selectors(including the bandwidth factorized SCV), where the best constants C are chosen to cancel bias terms.

However, even if the normal reference distribution approach does not alter the asymptotic performance, one may have qualms when the data have a distribution that is quite far from the normal. Furthermore, it has been seen that quite often, the reference distribution does not provide adequate tuning of these bandwidth selection methods. This motivates more careful tuning, through estimation of the unknown θ_m's. This is particularly important for the non-cross-validatory methods. In fact, such methods need some sort of consistent estimation of the unknown constants to enjoy the full advantages of reintroducing the diagonal terms. However, this entails other difficulties, because these pilot estimates of θ_m's have bandwidth that need to be selected. As these bandwidth depend asymptotically also on θ_m's, a simple approach is to use the reference distribution at this stage. However, it has been seen that, in some situations, the reference distribution still has too strong an effect on the original tuning process. An obvious idea is to iterate the pilot estimation process by estimating θ_m's needed, and only using the reference distribution at some final step where its effect has finally become negligible. See Park and Marron (1991) for further discussion of this issue.

4. Transformation Methods

So far, what we have discussed is the bandwidth selection problem for the global bandwidth kernel density estimator. A drawback of this estimator is that it performs quite poorly when the true density has features that require far different amount of smoothing at different locations. Silverman(1986) discusses nearest neighbor kernel estimators and adaptive kernel estimators which use variable amount of smoothing. Recently, several promising and intuitively appealing methods have been proposed in this direction, which are based on trnasformations of original data. The essential idea is to first transform the data, X_i from f_x, to $Y_i=\alpha(X_i)$ where α is a smooth and monotone function. The function α is chosen so that, for the density of Y_i,

$$f_Y(y\,;\,\alpha)=f_x(\alpha^{-1}(y))(d\alpha^{-1}(y)/dy),$$

a global bandwidth is more appropriate. The kernel density estimator

$$\hat{f}_Y(y\,;\,\alpha,h)=n^{-1}\sum_{i=1}^{n}K_h(y-Y_i)$$

is then back-transformed by change-of-variables to an estimator of f_x:

$$\hat{f}_x(x\,;\,\alpha,h)=\hat{f}_Y(\alpha(x)\,;\,\alpha,h)\alpha'(x).$$

The transformation ideas in density estimation are first introduced by Wand, Marron and Ruppert(1991). They worked with positive, skewed data, and considered a two-parameter shifted power transformation family, from which α is chosen. An effective method for selecting α, which is in some sense optimal, was proposed together with suggestion of a data-driven bandwidth selector for $\hat{f}_Y$. An empirical assessment of this methodology is done by Park, Chung and Seog (1991). From their simulation study, it is observed that the method works quite well in general, but not so good(even worse than the untransform estimator) for the populations with relatively high density near their support boundaries. This difficulty was resolved by making a fixed preliminary shift transformation.

Ruppert and Wand(1991) dealt with the problem of estimating an approximately symmetric probability density with high kurtosis. Note that this kind of densities also are estimated poorly by a global bandwidth kernel estimator since good estimation of the peak of the distribution leads to unsatisfactory estimation of the tails and vice versa. A two-parameter convex-concave transformation family was considered here since a convex-concave transformation(convex on the left half-line and concave on the right half-line) has the effects of taking probability mass proposed. Their simulation results confirm the superiority of the method over ordinary untransform kernel estimation.

A nonparametric way of choosing a transformation α was proposed by Ruppert and Cline(1991). The idea is first to transform the original data using a smooth estimate of their cdf. The transformed data is further transformed by the inverse cdf of some target distribution, such as the uniform or normal distribution. This allows one to think of the final transformed data as being a random sample from the target distribution under consideration. Good theoretical properties and practical significance of this method were observed.

5. Future Research

It should be clear from the above that the field of bandwidth selection in density estimation has been progressing very rapidly in recent 3 or 4 years. While many methods have been proposed, there is still room for improvement. Two important open problems in global bandwidth kernel estimation are : 1. Is there any bandwidth selector which has the fastest $n^{-1/2}$ rate of convergence together with the best constant multiplier, but using only nonnegative kernels at all stages of the selection process ? Recall that the improved BCV proposed by Hall, Sheather, Jones and Marron(1991), and the two Chiu's selectors attain the best rate and constant. But their methos rely upon the use of higher order kernels at functional estimation stages. On the other hand, the bandwidth factorized SCV uses nonnegative kernels at all stages and achieves the best rate, but does not have the best constant. 2. How may times does one need to iterate the pilot estimation process by estimating the integrated squared density derivatives ? Neglecting the computational aspects, the crucial issue is a need to balance the reduction in bias, from using higher stages, with the increased variance, which higher stages entail.

The field of transformation based variable bandwidth kernel density estimation is in its infancy. None of the methods proposed so far has emerged as clearly superior, and there is much comparison to be done, and many other possibilities to be investigated. Also, the high technologies developed in density estimation setting may be carried over to the related fields, such as regression function, spectral density and quantile function estimation. Extension to the multivariate case is another possibility.

References

1. Bowman, A. (1984). An Alternative Method of Cross-validation for the Smoothing of Density Estimates, *Biometrika*, 71, 353-360.
2. Chiu, S-T. (1991). Bandwidth Selection for Kernel Density Estimation, to Appear in *Annals of Statistics*.
3. Devroye, L. and Gyorfi, L. (1984). *Nonparametric Density Estimation : The L_1 View*, Wiely, New York.
4. Fan, J. and Marron, J.S. (1991). Best Possible Constant for Bandwidth Selection, Unpublished Manuscript.
5. Faraway, J. and Jhun, M. (1990). Bootstrap Choice of Bandwidth for Density Estimation, *Journal of American Statistical Association*, 85, 1119-1122.
6. Hall, P. (1983). Large Sample Optimality of Least Square Cross-validation in Density Estimation, *Annals of Statistics*, 11, 1156-1174.
7. Hall, P. and Marron, J.S. (1987a). Extent to Which Least-squares Cross-validation Minimizes Integrated Squared Error in Nonparametric Density Estimation, *Probability Theory and Related Fields*, 74, 567-581.
8. Hall, P. and Marron, J.S. (1987b). Estimation of Integrated Squared Density Derivatives, *Statistics and Probability Letters*, 6, 109-115.
9. Hall, P. and Marron, J.S. (1991). Lower Bounds for Bandwidth Selection in Density Estimation, to Appear in *Probability Theory and Related Fields*.
10. Hall, P., Marron, J.S. and Park, B.U. (1991). Smoothed Cross-validation, to Appear in *Probability Theory and Related Fields*.
11. Hall, P., Sheather, S.J., Jones, M.C. and Marron, J.S. (1991). On Optimal Data Based Bandwidth Selection in Kernel Density Estimation, to Appear in *Biometrika*.
12. Jones, M.C. (1991). The Roles of ISE and MISE in Density Estimation, to Appear in *Statistics and Probability Letters*.
13. Jones, M.C. and Kappenman, R.F. (1991). On a Class of Kernel Density Estimate Bandwidth Selectors, to Appear in *Scandinavian Journal of Statistics*.
14. Jones, M.C. and Sheather, S.J. (1991). Using Non-stochastic Terms to Advantage in Kernel-based Estimation of Integrated Squared Density Derivatives, to Appear in *Statistics and Probability Letters*.
15. Jones, M.C. and Sheather, S.J. (1991). Using Non-stochastic Terms to Advantage in Kernel-based Estimation of Integrated Squared Density Derivatives, to Appear in *Statistics and Probability Letters*.
16. Marron, J.S. (1988). Automatic Smoothing Parameter Selection : A Survey, *Empirical Economics*, 13, 187-108.
17. Marron, J.S. and Wand, M.P. (1991). Exact Mean Integrated Squared Error, to Appear in *Annals of Statistics*.
18. Park, B.U. (1989). On the Plug-in Bandwidth Selectors in Kernel Density Estimation, *Journal of the Korean Statistical Society*, 18, 107-117.
19. Park, B.U. and Marron, J.S. (1990). Comparison of Data-driven Bandwidth Selectors, *Journal of American Statistical Association*, 85, 66-72.
20. Park, B.U., Chung, S. and Seog, K.H. (1991). An Empirical Investigation of Shifted Power Transformation Method in Density Estimation, Unpublished Manuscript.
21. Park, B.U. and Marron, J.S. (1991). On the Use of Pilot Estimators in Bandwidth Selection, Unpublished Manuscript.
22. Rudemo, M. (1982). Empirical Choice of Histograms and Kernel Density Estimators, *Scandinavian Journal of Statistics*, 9, 65-78.

23. Ruppert, D. and Cline, D.B.H. (1991). Transformation-kernel Density Estimation : Bias Reduction by Empirical Transformations, Unpublished Manuscript.
24. Ruppert, D. and Wand, M.P. (1991). Correcting for Kurtosis in Density Estimation, to Appear in *Australian Journal of Statistics*.
25. Scott, D.W., Tapia, R.A. and Thompson, J.R. (1977). Kernel Density Estimation Revisited, *Nonlinear Anal. Theory Meth. Applic.* 1, 339-372.
26 Scott, D.W. and Factor, L.E. (1981). Monte Carlo Study of Three Data-based Nonparameteric Probability Density Estimators, *Journal of American Statistical Association*, 76, 9-15.
27. Scott, D.W. and Terrell, G.R. (1987). Biased and Unbiased Cross-validation in Density Estimation, *Journal of American Statistical Association*, 82, 1131-1146.
28. Sheather, S.J. (1983). A Data-based Algorithm for Choosing the Window Width when Estimating the Density at a Point, *Computational Statistics and Data Analysis*, 1, 229-238.
29. 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.
30. Sheather, S.J. and Jones, M.C. (1991). A Reliable Data-based Bandwidth Selection Method for Kernel Density Estimation, to Appear in *Journal of Royal Statistical Society, Series B*.
31. Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York.
32. Stone, C.J. (1984). An Asymptotically Optimal Window Selection Rule for Kernel Density Estimates, *Annals of Statistics*, 12, 1285-1297.
33. Taylor, C.C. (1989). Bootstrap Choice of the Smooting Paramter in Kernel Density Estimation, *Biometrika*, 76, 705-712.
34. Wand, M.P., Marron, J.S. and Ruppert, D. (1991). Transformations in Density Estimation, with Discussion, to Appear in *Journal of American Statistical Association*.

Discussion

Ja-Yong Koo*

Park deserves commendation for the excellent expository review in the present paper. I offer two comments at first.

1. In a mathematical treatment of density estimation, it is convenient to use integrated squared error $\int(f-f)^2$ as a measure of inaccuracy. But this measure does not reliably reflect qualitative fidelity. See Fig. 1 of Kooperberg and Stone(1990). Basically most bandwidth selectors use ISE or MISE as their motives. Such selectors may not reliably reflect qualitative fidelity and thus we need to check their performance by pictures.

2. Most data analysts are content with IBM PC and do not want to go CRAY. Is there any study on computational problem of selectors? On the other hand, one basic assumption for selection methods is that f is twice differentiable. Practically, however, we don't have such

* Hallym University

information beforehand. Thus a selection method may not work well if it is derived under a specific smoothness assumption. One nice feature of the least squares cross-validation is that it does not depend on any smoothness assumption. A simulation study would be very useful on the performance of selectors for a variety of class of densities, say, bimodal, skewed, unsmooth densities.

Recently I have been working on a different approach to fitting more or less the same class of densities f but using polynomial cubic splines to estimate $\log(f)$ by maximum likelihood estimation. In order to avoid artificial end effects of polynomial fits, the splines are constrained to be linear to the left of the first knot and to the right of the last knot. To avoid multiple representation of f, one linear constraint is imposed. Thus, if there are N knots, there are $N+4$ degrees of freedom for the unconstrained spline and $N-1$ degrees of freedom for the estimate. This approach is referred to as the logspline density estimation to distinguish it from the smoothing spline approach favored by Wahba and others in which smoothing is achieved by a roughness penalty instead of by confining attention to spline models with a modest number of degrees of freedom. Preliminary study on the asymptotic optimal rules for selecting N based on the data is currently on the way.

It is claimed that the relative rate of convergence of selection rule based on AIC for N_0, which is an analogue of h_0, is of order $n^{-1/10}$. This phenomenon also happens to the kernel density estimation. Another claim is that this interesting rate $n^{-1/10}$ is best possible for a variety of density estimation techniques, not just for kernel density estimation.

In numerical implementation, one important problem is the dependence of knot placement. One remedy is this—put down many knots. Do the logspline density estimation. Now delete least contributory knots from the fit. Continue the deletion under the best fit is found. I have implemented this idea using B-splines. This procedure has the effect of using a locally adaptive window size—the feature that made supersmoother in regression problem so attractive. One interesting featurue is that this method is not designed to handle a specific feature of unknown density. The first two transformation methods in page 14-15 have a specific goal. One procedure designed to handle positive skewed data may not work well for symmetric density with high kurtosis. I wonder if Ruppert and Cline(1991) have done such work.

References

1. Kooperberg, C. and Stone, C.J.(1990). A Study on Logspline Density Estimation. Technical Report No. 238, Department of Statistics, Univ. of Calif., Berkeley.
2. Stone, C.J.(1990). Large-sample Inference for Log-spline Models. *Ann. Statist.* 18, 717-741.
3. Stone, C.J. and Koo, C.-Y.(1986). Logspline Density Estimation. *Contemporary Mathematics*, 59, 1-15.

Myoungshic Jhun*

Park has provided a successful updated introduction to the data-driven bandwidth selection in kernel density estimation. Certainly, we thank for his work. Actually, bandwidth selection

* Korea University

is a very crucial and popular problem in the most of smoothing techniques. Since his paper is a kind of survey for recents theoretical results mainly on gloval choice of bandwidth for L^2 −norm, I would like to discuss about some other aspects.

L^1−norm

In estimating density function $f(x)$ by kernel estimator $f_h(x)$, as a loss, we may consider IAE(integrated absolute error)

$$\int \mid f_h(x)-f(x) \mid dx$$

By using IAE, we may have some advantages including very natural and geometric interpretation and invariance property under rich class of transformations. However, mathematically it is hard to track. Maybe, that's why not so many works on this direction. But, asymptotically we can derive MIAE(Mean IAE) as following. Let, $K_2=\int x^2K(x)dx$ and $\|K\|^2=\int K^2(y)dy$. Then, for $h=cn^{-1/5}$,

$$Z_n(x,c)=n^{2/5}[f_h(x)-f(x)] \quad \rightarrow \quad \text{normal } (c^2K_2f''(x)/2,\ f(x)\|K\|^2).$$

Therefore, we can have

$$E\int n^{2/5} \mid f_h(x)-f(x) \mid dx \quad \rightarrow \quad \|K\|c^{-1/5}\int \psi\left[\frac{(cK_2/2\|K\|) \mid f''(x) \mid}{\sqrt{f(x)}}\right]\sqrt{f(x)}\ dx \quad (1)$$

where $\psi(y)=E \mid Z+y \mid$ with $Z\sim N(0,1)$. Devroye and Gyorfi(1984) obtained the upper limit of the LHS of (1). But, we way find $c=c(f)$ minimizing the limit, which is quite different from minimizing Devroye and Gyorfi's upper limit. Fortunately, RHS of (1) is a convex function of c so we may use bisection algorithm to find the minimizer say $c^*(f)$. Now, for given data $X_1, X_2, \cdots, X_n$, we can have a initial estimator $f_o(x)$ and use the bandwidth $h^*=c^*(f_o)n^{-1/5}$. In fact, $c(f)$ is continuous for f in a suitable metric space consists of density functions and we can have asymptotic results for the choice $h^*=c^*(f_o)n^{-1/5}$. For details see Woodroofe and Jhun (1988).

Variable Bandwidth

The idea of varing bandwidth to be larger in regions of data sparsity and smaller in regions where the data is plentiful has been proposed in the hope of producing less regged estimates of density. Our variable kernel density estimate will take the form

$$f_{h,\alpha}(x)=\frac{1}{n}\Sigma\frac{1}{h_i}K\left(\frac{x-X_i}{h_i}\right)$$

where $h_i=h\tilde{f}(X_i)^{-\alpha}$ with $\tilde{f}(X_i)$ is some initial estimate of density at X_i, for example by a fixed kernel density estimate. h is the smoothing parameter and represents the overall amount of smoothing and α is the sensitivity parameter representing the degree of adaption to sparsity. Proper selection of h and α is key to getting good variable kernel density estimates. See Silverman (1985) for details.

In order to investigate the behavior of the estimator over simultaneously varying choice of α and the bandwidth, a simulation study was carried out. We computed the estimate over a

20×20 grid of α×h, 100 replications were made for sample size 50 and 100. Test distributions are (1) Standard Normal ; N(0,1), (2) Bimodal Normal ; 1/2N(−1,1/4)+1/2N(1,1/4), (3) Contaminated Normal ; 1/2N(0,4)+1/2N(0,1/4), (4) Standard Lognormal, (5) Cauchy, (6) Beta(2.2). We display our results in two ways. Firstly, the resultsare displayed in Table 1 and 2. For 'Fixed choice' w give the fixed values of α and h which minimise the MISE over all replicatioins. For the 'ISE choice', the ISE is minimised over α and bandwidth for each sample and statistics for these choices are given. In the column marked 'corr' we give the correlation between the choice of α and h.

Table 1. Variable Kernel Density Estimates n=50

Distribution	Fixed choice			ISE choice						
	mise	alpha	bw	mise	se	alpha	sd	bw	sd	corr
normal	0.73	0.20	0.90	0.46	0.05	0.30	0.25	0.90	0.30	-0.85
bimodal	1.91	0.30	0.50	1.71	0.09	0.23	0.47	0.58	0.27	-0.95
cont. normal	0.88	0.70	0.55	0.62	0.05	0.69	0.24	0.53	0.13	-0.83
lognormal	2.93	0.10	0.41	2.42	0.12	0.33	0.46	0.39	0.18	-0.79
cauchy	3.06	0.70	0.35	2.04	0.18	0.75	0.25	0.33	0.06	0.20
beta	0.83	-0.60	2.23	0.72	0.06	-0.44	0.55	2.11	1.02	-0.93

Table 2. Variable Kernel Density Estimates n=100

Distribution	Fixed choice			ISE choice						
	mise	alpha	bw	mise	se	alpha	sd	bw	sd	corr
normal	0.47	0.3	0.82	0.32	0.03	0.30	0.26	0.79	0.28	-0.91
bimodal	1.04	0.30	0.45	0.89	0.05	0.36	0.36	0.44	0.17	0.91
cont. normal	0.59	0.60	0.55	0.41	0.03	0.64	0.21	0.51	0.12	-0.80
lognormal	1.95	0.20	0.33	1.68	0.08	0.30	0.41	0.32	0.12	-0.65
cauchy	1.52	0.60	0.33	1.03	0.08	0.64	0.21	0.31	0.05	0.38
beta	0.53	-0.50	1.65	0.44	0.03	-0.40	0.54	1.76	0.95	-0.92

Secondly, we give contour plots for sample size 50 of the IMSE over α and the bandwidth the heights of the contours are displayed on the plots. See Fig. 1.

The optimal choice of α varies greatly. It is large for the contaminated normal and cauchy, our long tailed distributions. It is quite negative for the beta. This most interesting since negative choices for α would not seem reasonable at first glance. However, with a little thought, one can see how this is the reasonable for the beta(2,2). On emight expect the same effect for a uniform distribution. Note also that the sample standard deviations on the choice of α are relatively large. These observations lead us to propose a database method of simultaneous selection of α and bandwidth. A marked negative correlation is shown between choice of α and bandwidth except in the case of the Cauchy. Of course we know that α and bandwidth are linked in some non-simple way but it is curious that the Cauchy should behave in so different a manner. The

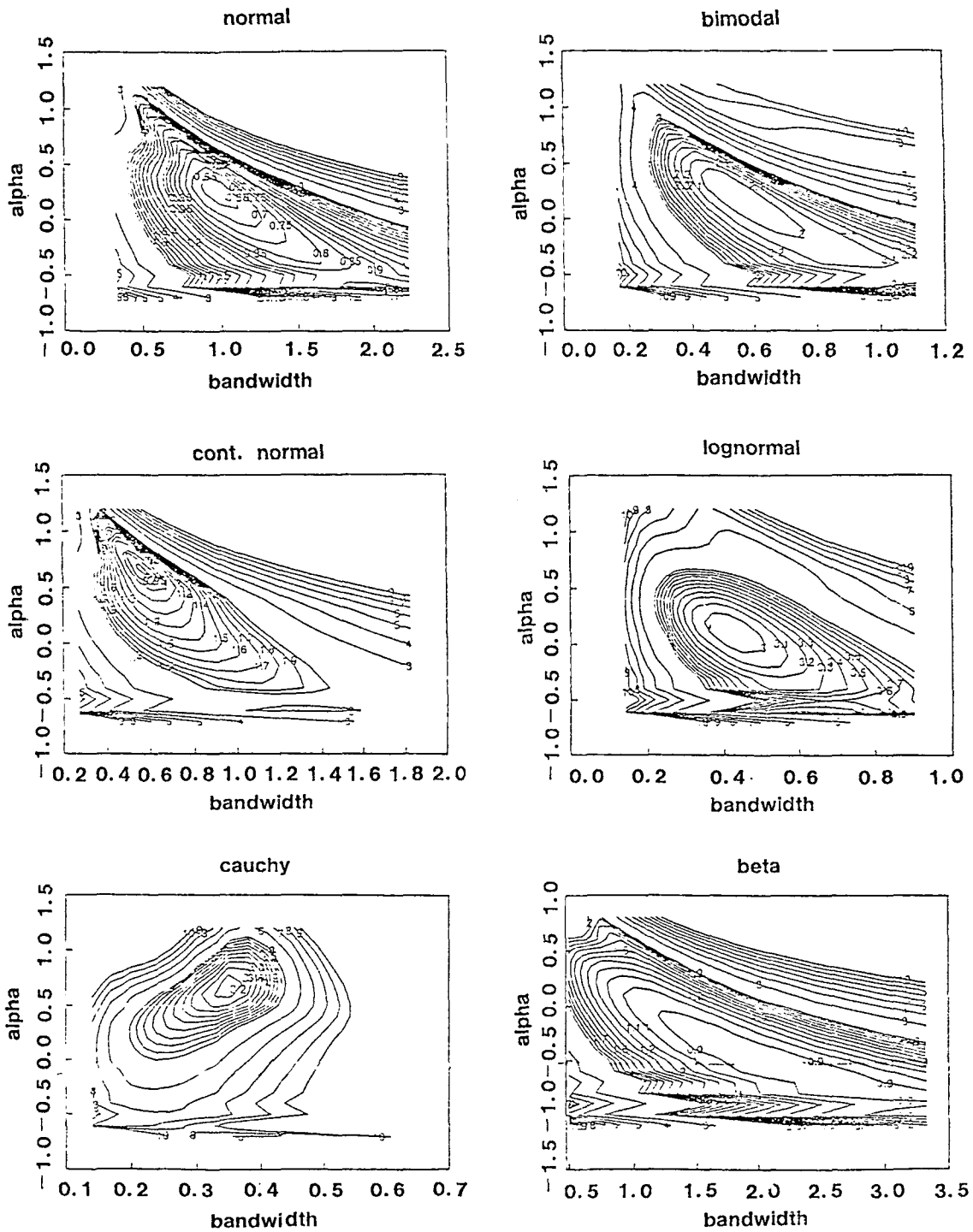

Fig. 1. Contour plots of MISE.

contour plots give some intereting shapes. Convexity is not quite there which may lead to some problems in some methods of data-based selection of the smoothing parameters. Note the peculiar effects as α becomes negative. The MISE's of the ISE choices show that, if we can only choose α and the bandwidth well, we may be able to obtain superior estimates. For details see Faraway and Jhun(1988).

References

1. Devroye, L. and Gyorfi, L.(1984). Nonparametric Density Estimation : The L_1 View, Wiley, New York.
2. Faraway, J. and Jhun, M.(1988). Variable Kernel Estimates of Density, Unpublished Prepirnt.
3. Silverman, B.(1985). Density Estimation for Statistics and Data Analysis, Chapman Hall, London.
4. Woodroofe, M. and Jhun, M.(1988). Adaptive Bandwidth Selectiion in Density Estimation, Unpublished Preprint.

Yunkee Ahn* and Tae-Yoon Kim**

Byung U. Park's paper is excellent survey and thought provoking paper about the data-driven bandwidth selection problems. Recently, noticeable progresses have been made to obtain better data-driven bandwidth selector. In this discussion, we will mainly focus on fundamental issues about kernel density estimator, its data-dependent bandwidth selector, cross validation method and the case of dependent variables.

Kernel Density Estimators

In kernel density estimator, the choice of smoothing parameter h is very critical issue. It is well known that the choice of bandwidth determines and controls the smoothness of the data. In this sense, the choice of bandwidth could be considered to select the density model to be imposed on the data. Indeed, small h imposes a complex model while large h imposes a smoothing model on the data. The condition always referred for kernel density estimation to be effective is that $h \to 0$ as $n \to \infty$, which implies that very complicated model is desirable with addition of data. Recall that there would be effectively no bias if no model (complicated model) is imposed. However, $nh \to \infty$ implicitly suggests that we still favor somewhat smoothing model.

We also would like to mention one of the interesting features of kernel estimation. That is, kernel estimations presents the model in terms of a control parameter h and thus requring no effort to estimate parameters in the selected model. This gives extensive leeway for us to select the appropriate model from infinitely many candiates even with one data point even though we

* Department of Applied Statistics, Yonsei University
** Department of Statistics, Keimyung University

need more data than number of parameters in most statistical methods.

It is well known that the choice of kernel is minor matter in selecting optimal bandwidth. But our true object of density estimation is to find the estimator which is close to true density not just finding bandwidth. In this sense, Samiuddin and El-Sayyao(1990) have shown that we can improve estimation by using Epanzchnikov kernel for any fixed bandwidth in admissibility sense. Thus we suggest the density estimation should be considered as two-stage procedure by using proper kernel first and then optimal bandwidth.

Cross-Validation Technique

Until recent times, many seemingly promising methods have been presented to estimate smoothing parameter h. Our main interest here goes to cross validation which have attracted many statistical analysts for its practical convenient use. However, recent investigations show that least square cross validation is subject to sampling variability in the sense that for different data sets from the same distribution will typically gives much different answers. Thus kernel estimaor with cross validation bandwidth selector couldn't entertain the expected fast convergence rate to true density f, $n^{-4/5}$ if n is not big enough.

In Park's paper, there is not enough discussion about the cause of sampling variability which cross validation suffers from. More intuitive illustration of the cause would be helpful to understand the discussed methods. One may mistakenly think that a long tail of true density may be main reason of smapling variability. However, from his paper it seems that sampling variability is a inherent drawback of cross validation.

Basic spirit under discussed methods is to smoothe cross validation to reduce the sampling variability whether AMISE based methods or MISE based methods. Very often than not, these smoothing tuning procedure of cross validation require pilot estimations and it may cause difficulty to wide use of proposed methods. Therefore, it would be desirable if this tuning procedure could be done without much effort, e.g. any convenient estimate can be used. As another method of smoothing cross validation, Marron(1987) proposed partitioned cross validation(PCV) for kernel density estimation to eliminate the sample noise of cross validation. The idea of PCV is to split the observation into g subgroups by taking every g-th observation and to calculate the ordinary cross validation score $CV_{o,k}(h)$ of the k-th subgroup of observations, $k=1, 2, \cdots g$ and minimize the average of these score functions

$$CV^*(h)=g^{-1}\sum_{k=1}^{g} CV_{o,k}(h).$$

The minimizer of $CV^*(h)$ is denoted by h_{PCV}. Since h_{PCV} is appropriate for sampled size only n/g, the partitioned cross-validated bandwidth is defined to be the rescaled $g^{-1/5}\ h_{PCV}$.

One of active statistical areas where correct density estimation is necessary(but not sufficient) is discriminant analysis. Kullback-Leibler loss is employed as a popular measure of distance in discriminant analysis. Log likelihood cross validation has been considered to be a good estimate minimizing Kullback Leibler loss, $L(f,g)=\int f(x)\ \log\{f(x)/g(x)\}\ dx$. Here, loglikelihood cross validation is defined to be

$$n^{-1}\sum_{i=1}^{m} \log\ f_{-i}(X_i)$$

However, it has been pointed out that the loglikelihood cross validation are profoundly influnced

by tail properties of the kernel and of the unknown density. Hall(1987) shows that a kernel with a thick tail smoothe the behaviour of loglikelihood cross validation and thus minimizing the loglikelihood across validation over h achieves the minimum of Kullback-Leibler loss when the underlying density has a thick tail. Our question is that some other techniques may be developed to smoothe the behaviour of loglikelihood cross validation.

Kernel Estimation with Dependent Variables

As another field of possible future studies, we like to add one area which is attracting a lot of research interests. It is kernel density estimation for dependent variables. As obvious application is to the analysis of time series data. Unlike density estimation for independent r.v.'s, it doesn't have many classical density estimation method avilable. So it is open area nonparametric density estimation is expected to work well. Strong consistency for various delta sequence estimators with dependent data including Kernel estimator have been established(e.g., Collomb and Hardle(1986), Robinson(1983) and Troung and Stone(1988))but still many research is underway to improve those results.

Bandwidth selection problem has been addressed by Hart and Vieu(1990) for density estimation and T. Kim(1990) for regression estimation. Difficulty in dependent variable cases arises from the fact that we should handle the dependence. Since dependence of data used to influence variance, we are likely to draw unstable estimates. Also the dependence cause troubles for the bandwidth selectors designed for independent observations. For instance, if the observations are positively correlated, then cross validation will produce under-smoothed estimates. On the other hand, if the observations are negatively correlated, then cross validation will produce oversmoothed estimates. See Hart and Wehrly(1987), for a detailed discussion of dependence effect on bandwidth selection.

To handle the dependence, several methods have been suggested but each has its own drawbacks. One adjustment is modified cross validation(MCV) and is simply the "leave$-(2j+1)-$out" version of cross$-$validation. See Hart and Vieu and Hardle and Vieu(1988) for earlier results on the application of this method. But the problem is how many data points(choice of j) should be deleted, which adds another difficulty to our estimation problem. Another possibility is to use PCV discussed above. Problems with PCV is that it is so effective at removing the dependence i.e., it handles dependent data as independent one and thus resulting much biased bandwidth selectors. See Chu and Marron(1988). ■

References

1. Chu, C.K. and Marron, J.S.(1989). Comparison of Two Bandwith Selectors with Dependent Errors. Mimeo Series #2007 Dept. of Stat. University of North Carolina at Chapel Hill.
2. Collomb, G. and Hardle, W.(1986). Strong Uniform Convergence Rate in Robust Nonparanetric Time Series Analysis. Stochastic Processes and their Applications. 23, 77-89.
3. Hardle, W. and Vieu, P.(1988). *Nonparametric Prediction by the Kernel Method. Preprint.*
4. Hart, J. and Vieu, P.(1990). Data-driven Bandwidth Choice for Density Estimatiion Based on Dependent Data. Annals of Statistics, 18, 873-890.
5. Hart, J. and Wehrly, T.(1986). Kernel Regression Estimation Using Repeated Measurements Data. Journal of the American Statistical Association, 81, 1080-1088.

6. Hall, P.(1987). On Kullback Liebler Loss and Density Estimation. Annals of Statistics, 15, 1491-1519.
7. Kim, T.(1990). Optimal Bandwidth Selection Rule for Kernel Regression Estimator with Dependent Variables. Ph. D. Thesis, Department of Statistics, University of Illinois at Urbana-Champaign.
8. Marron, J.S.(1987). Partitioned Cross Validation. Econometric Reviews, 6, 271-284.
9. Robinsion, P.M.(1983). Nonparametric Estimates for Time Series, Journal of Time Series Analysis, 4, 185-201.
10. Samiuddin, M. and El-Sayyad, G.M.(1990). On Nonparametric Kernel Density Estimates, Biometrika, 77, 865-874.

Byung H. Kim* and Kyung J. Cha**

It is a pleasure to read this introductory paper which introduces several methods for global bandwidth kernel estimator. Also, it is a pleasant news that there is a considerable progress on the problem of an adaptive(i. e., depending only on the data and then directly computable in practice) bandwidth selection in kernel density stimation.

Since the paper is to provide an introduction for global bandwidth kernel estimation, we would like to make a few comments and to introduce another method which is under investigation by several statisticians.

As the author pointed out, the choice of the bandwidth is the central issue in the application of the kernel density estimation. Even though the choice of the optimal bandwidth that minimizes MISE or AMISE is the main goal of the kernel estimation, selecting a kernel function has been discussed by a few authors. Results in (Gasser and Müller(1979)) reveal that the solution is a quadratic kernel(i.e., Epanechnikov kernel). Also, Benedetti((1977) pointed out the optimality of the qudratic kernel for nonparametric regression.

However, Silverman(1986) pointed out that there is very little to choose between the various kernels on the basis of mean integrated square error. Recently, Eubank(1989) discussed the problem of selecting a kernel for working on nonparametric regression. He concluded that the actual choice of a kernel is not very important on the basis of mean square error.

In Section 4 the author mentioned the boundary problem. The boundary problem is such that x is less than the bandwidth h, thus the scaled support of the kernel is no longer completely in the interior. Let us look at $E[f_h(x)]$ and $Var[f_ha(x)]$. By simple Taylor expansions,

$$E[f_h(x)]=\int^q k(z)\{f(x)-zhf'(x)+[(zh)^2/2]f''(x)\}dz+o(h^2)$$

$$Var[f_h(x)]=\frac{1}{nh}\ f(x)\int^q k^2(z)dz+o(\frac{1}{nh})$$

where $q=x/h$ and assuming that there is a boundary effect. Since the symmetry of a kernel is lost by the boundary effect, $\int^q k(z)dz\neq 1$ and $\int^q zk(z)dz\neq 0$. Therefore, we could not get the same rate of convergence as the methods introduced in the paper. It would be suggested that

* Hanyang University
** Sejong University

the modification at the boundary is necessary to have decent approximation of the true density function.

Without considering boundary effect, lets look at asymptotic mean integrated square error because the bandwidth that minimizes AMISE provides a decent approximation to h_0 as the auther pointed in Section 3.3. Then,

$$\begin{aligned}\mathrm{AMISE}(h) &= \mathrm{Var}[f_h(x)] + \mathrm{Bias}^2[f_h(x)] \\ &= \frac{1}{nh}\int k^2(t)dt + \frac{1}{4}h^4\{\int x^2k(x)dx\}^2\{\int f_h''(x)^2dx\}. \qquad (1)\end{aligned}$$

Therefore, by using the lemma 4a of Parzen(1962), the optimal bandwidth which minimizes the AMISE is

$$h_{opt} = \left[\frac{\int k^2(x)dx}{n\{\int x^2k(x)dx\}^2\{\int f''(x)^2dx\}}\right]^{1/5}.$$

Also, by plugging h_{opt} into (1), it can be easily shown that

$$\mathrm{Var}[f_{hopt}(x)] = 4 \cdot \mathrm{Bias}^2[f_{hopt}(x)]\ . \qquad (2)$$

From equation(2), it can be seen that the optimal bandwidth balances the asymptotic $\mathrm{Var}[f_h(x)]$ and $4 \cdot \mathrm{Bias}^2[f_h(x)]$.

For some positive number A and B, it can be rewritten

$$\mathrm{Var}[f_h(x)] \sim \frac{A}{nh}$$
$$\mathrm{Bias}^2[f_h(x)] \sim Bh^2$$

for large n. Then the problem that we need to overcome is how to estimate A and B. There are several ways to estimate A and B, such as robust M-estimation and the least square method. Let us suppose that the least square method is employed. Then by substituting estimates into and solving equation(2), we find the adaptive bandwidth as

$$h = \left\{\frac{A}{4nB}\right\}^{1/5}.$$

Eubank and schucany(1990) developed the local version of adaptive bandwidth selection technique by using this method which can be directly extended to global bandwidth kernel estimation. In order to estimate A and B by the least square method, Eubank and Schucany used the grid of bandwidths. Cha and Schucany(1990) show that permissible values of the rate of convergence for grid of bandwidths are from 1/7 to 1/5 by simulation study. Thus, when the rate of convergence for grid of bandwidths is 4/25, the ratio of the bandwidth estimate to the optimal bandwidth converges to one at the rate $n^{-1/10}$ as Härdle, Hall, and Marron(1988) showed for cross-validation global bandwidth estimates.

With all of the above, it seems to be reasonable to find the bandwidth that minimizes criteria such as MISE, AMISE or mean square error. It also would be possible to extend cross-validation to local bandwidth kernel estimator. Hall and Schucany(1989) develop a local cross-validation algorithm. They propose a local version of square error cross-validation, suitable for estimating a probability density at a given point and show it is asympotically optimal in the sense of minimizing

mean square error.

We would like to end this by thanking to the author for opportunity to read this paper. As statisticians who are interested in this area, we also appreciate his time to spend for introducing new areas.

References

1. Benedetii, J.K. (1977). On the Nonparametric Estimation of Regression Functions, Journal of the Royal Statistical Society, B39, 248-253.
2. Cha, K.J. and Schucany, W.R. (1990). Adaptive Bandwidth Selection for a Boundary Kernel in Nonparametric Regression, Doctoral Dissertation, Southern Methodist University, Dallas.
3. Eubank, R.L. (1988). Spline Smoothing and Nonparametric Regression, Marcel Dekker, New York.
4. Eubank, R.L. and Schucany, W.R. (1990). Adaptive Bandwidth Choice for Kernel Regression, Prepinrt, to Appear, Journal of the American Statistical Association.
5. Gasser, Th. and Müller, H.G. (1979). Kernel Estimatioin of Regression Functions, In Smoothing Techniques for Curve Estimation, 23-68, Heidelberg : Spring-Verlag.
6. Hall, P. and Schucany, W.R. (1989). A Local Cross-Validation Algorithm, Statistics and Probability Letters, 8, 109-117.
7. Härdle, W., Hall, P., and Marron, S. (1988). How Far Are Automatically Chosen Regression Smoothers From their Optimum, Journal of the American Statistical Association, 83, 86-101.
8. Parzen, E. (1962). On Estimation of a Probability Density and Mode, The Annals of Mathematical Statistics, Vol. 33, 1065-1076.
9. Silverman, B.W. (1986). Density Estimation for Statistics and Data Analysis, Chapman and Hall, New York.

Rejoinder

First of all, I would like to thank Professor Yunkee Ahn for giving me an opportunity to write a survey paper on recent developments in bandwidth selection. I also thank all the discussants for their efforts to have some healthy interactions.

Not all the issues raised by the discussants are responded since response to some of the comments goes beyond the scope of the present paper. I will respond to the four comments separately in aribitrary order.

1. Koo

Koo points out that ISE and MISE do not reflect qualitative fidelity of density estimates. He suggests that we need to check the performance of density estimates by pictures. In principle, this is a good idea, but can not be accomplished in practice since we do not know the true density. ISE and MISE also depend on the unknown true density, but the advantage of using

these quantitative measures is that this type of dependency can be resolved by estimating the related functionals.

Koo asks me whether there is any study on computational problems of bandwidth selectors. There has been some study on this. One of the most frequently used and efficient algorithms for calculating related quantities of kernel density estimates is the binned implementation introduced by Scott(1985). See also Silverman(1982), and Jones and Lotwick(1984) for algorithms based on the Fast Fourier transform.

Koo says that one nice feature of the least squares cross-validation is that its derivation does not depend on any smoothness assumptions on the true density. This is right, but this feature is shared by other bandwidth selectors too, such as SCV. The real advantage of the least squares cross-validation is that it is more robust(of course, at some cost in efficiency) than any other bandwidth selectors, as pointed out in Park and Marron(1990), and reconfirmed in a recent extensive simulation study conducted by Steve Marron.

Koo says something about logspline density estimation. It has been known that this type of estimator, as well as kernel estimator, enjoys the asymptotically optimal rate of convergence to the true density. One great advantage of logspline density estimator is that it is computationally cheaper than kernel density estimator. However, compared to the latter, the former is less intuitively appealing and relatively few theoretically justified data-based smoothing parameter selection methods are available.

Finally, I would like to mention that the Ruppert and Cline(1991)'s transformation method has a parallel advantage with the stepwise knot deletion procedure in that is is not designed to handle a specific feature of the unknown density.

2. Jhun

Jhun makes out a case for L^1-norm as a loss of density estimate. I agree that it has a natural geometric interpretation and is invariant under some class of transformation. He proposes to minimize the limit of MIAE to get a bandwidth, instead of minimizing the upper bound as was proposed by Devroye and Gyorfi(1984). There are two real problems here. One is that minimization of the limit is computationally too expensive. It actually involves double numerical integration. The other is how to choose a pilot estimate f_0. If we want to use another kernel density estimate, we again need to choose a proper bandwidth at this stage. But I do not think there is any objective method to choose a bandwidth for this pilot estimate since the quantity which we want to estimate is too complicated, and it is an implicit functional of f.

3. Ahn and Kim

Ahn and Kim refer to Samiuddin and El-Sayyad(1990) arguing the importance of choice of kernel function. In fact, choice of kernel function and that of bandwidth are coupled problems and need to be considered simultaneously. Marron and Nolan(1989) consider canonical kernels to uncouple these two problems, which enables us to choose a kernel function irrespective of bandwidth. However, looking at AMISE, the performance of kernel density estimator is crucially dependent on the choice of bandwidth, and the choice of kernel does not really matter, only changing slightly the constant factors in AMISE(see Epanechnikov 1969).

Ahn and Kim advocate use of the partitioned cross-validation proposed by Marron(1987). A drawback to this approach is that we must decide on the number of groups to use, and this problem seems to closely parallel that of smoothing parameter selection. Also, it is known that this bandwidth selector even with the theoretically optimal number of groups has a slow $n^{-1/4}$ rate of convergence to h_0. One possible improvement suggested by N.I. Fisher is discussed in Marron(1987), but the idea has not been analyzed yet.

4. Kim and Cha

Kim and Cha comment on the boundary effect which kernel density estimates often have when the true density is supported by a finite interval. But the problem I raise in Section 4 is not quite related to this. What I mean is that the method to select a particular transformation may break down when the population has relatively high density near its boundary. The reason for this is that this kind of populations often lead us to choose a transformation which is stringent at the bounary, and this results in a density estimate with a big spike near the boundary. Kim and Cha mention that the usual asymptotic results are no longer valid when the boundary problem is present. I do not think so since the boundary problem is of finite sample matter, and is gone as the sample size increases.

References

1. Epanechnikov, V.A.(1969). Nonparametric Estimation of a Multivariate Probality Density, Theory of Proabability and Its Application, **14**, 153-158.
2. Jones, M.C. and Lotwick, H.W.(1984). A Remark on Algorithm AS 176 : Kernel Density Estimation Using the Fast Fourier Transform, Applied Statistics, **33**, 120-122.
3. Marron, J.S. and Nolan, D.(1989). Canonical Kernels for Density Estimation, Statistics and Probability Letter, **7**, 95-99.
4. Scott, D.(1985). Average Shifted Histograms : Effective Nonparametric Density Estimators in Several Dimensions, Annals of Statistics, **13**, 1024-1040.
5. Silverman, B.W.(1982). Algorithm AS 176 : Kernel Density Estimation Using the Fast Fourier Transform, Applied Statistics, **31**, 93-99.