〈Research Paper〉

Journal of the Korean
Statistical Society
Vol. 18, No. 2, 1989

On the Plug-in Bandwidth Selectors in Kernel Density Estimation+

Byeong-Uk Park*

ABSTRACT

A stronger result than that of Park and Marron (1990) is proved here on the asymptotic distribution of the plug-in bandwidth selector. The new result is that the plug-in bandwidth selector may have the rate of convergence ($n^{-4/13}$) with less smoothness conditions on the unknown density functions than as described in Park and Marron's paper. Together with this, a class of various plug-in bandwidth selectors are considered and their asymptotic distributions are given. Finally, some ideas of possible improvements on those plug-in bandwidth selectors are provided.

1. Introduction

As a useful tool for exploring the distribution structure of unknown population, kernel density estimation is widely used by many data analysts. See Silverman (1986) for a variety of real data examples which illustrate the power of this method. But the practical implementation of this method requires to determine the amount of smoothing, As a consequence of the pressing need for determining the amount of smoothing using the data in kernel density estimation, various data-driven smoothing parameters or bandwidths have been studied so far. See Marron (1988) for a listing of proposed methods.

Recently, a comparison of several data-driven bandwidth selectors is provided by Park and Marron (1990). It has been seen that, when the underlying density is sufficiently

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

+ This research was partially supported by the Basic Science Research Institute program, Ministry of Education 1989.

smooth, the plug-in bandwidth selector (abbreviated as PI from now on) is superior to the least squares cross-validation and biased cross-validation in the asymptotic rate of convergence to the optimum. It also has been seen that the PI performs better than those two bandwidths in the simulation study. However, it is pointed out that there is still room for improvement

In this paper, we provide a stronger result than Theorem 3.3 of Park and Marron (1990) on the asymptotic distribution of the PI. Together with this, a class of various plug-in bandwidth selectors are considered and their asymptotic distributions are given in section 4. Finally, some ideas of possible improvements on those plug-in bandwidth selectors are also provided at the end of section 4.

2. Plug-in and Other Bandwidth Selectors

Kernel density estimators of the unknown probability density function $f(x)$ have the form,

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

where $K_h(x) = K(x/h)h$ and $X_1, X_2, \cdots, X_n$ is a random sample from f. The function K is assumed here to be a probability density function, called the kernel. It is well known, see section 3.3 of Silvermean (1986) for example, that performance of the estimator is far more sensitive to choice of the bandwidth h than to K. For the reasons discussed in Hall and Marron (1989), we take the optimal choice of h to be the minimizer of the Mean Integrated Squared Error,

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

Since h_{MISE}, the minimizer of $\mathrm{MISE}(h)$, is not available to us (it depends on the unknown density f), various data-driven bandwidth selectors have been invented as an attempt to estimate this optimal bandwidth. Those include the least squares cross-validated bandwidth, the biased cross-validated bandwidth and the PI.

The least squares cross-validated bandwidth has been introduced independently by Rudemo (1982) and Bowman (1984). The fact that this bandwidth is asymptotically correct has been demonstrated by Hall (1983), Stone (1984), etc. A drawback to least squares cross-validation is that the score function has a tendency towards having several local minima, which has been observed theoretically in Hall and Marron (1988). Another major

weakness of the bandwidth is that it is subject to a great deal of sample variability. This has been quantified asymptotically by Hall and Marron (1987a).

The biased cross-validated bandwidth, introduced by Scott and Terrell (1987) as an attempt to reduce sample variability of least squares cross-validation, is based on the asymptotic representation of MISE(h),

$$\text{AMISE}(h) = n^{-1}h^{-1}R(K) + h^4\sigma_K^4 R(f'')/4. \tag{2.1}$$

In the equation (2.1) and below, the functionals $R(\cdot)$ and σ_K^2 are defined by

$$R(K) = \int K^2(x)dx, \qquad \sigma_K^2 = \int x^2K(x)dx.$$

The biased cross-validated bandwidth is the minimizer of the estimator of AMISE(h), obtained by replacing the unknown $R(f'')$ by $\hat{R}_{f''}(h) \equiv R(\hat{f}_h'') + n^{-1}h^{-5}R(K'')$.

The PI which we discuss here is due to Hall (1980) and Sheather (1983, 1986). In particular, h_{AMISE}, the minimizer of AMISE(h), can be written as

$$h_{\text{AMISE}} = \{R(K)/\sigma_K^4 R(f'')\}^{1/5} n^{-1/5}. \tag{2.2}$$

The idea is that one replaces $R(f'')$ by some suitable estimator of it, which may use a bandwidth different from h in the expression of h_{AMISE}. The estimator considered is

$$\hat{R}_{f''}(a) = R(\hat{f}_a'') + n^{-1}a^{-5}R(K''), \tag{2.3}$$

where a is used as the bandwidth to stress its difference from h. The relationship between h_{AMISE} and a_{AMSE}, the approximation of the minimizer of the Mean Squared Error for using $\hat{R}_{f''}(a)$ to estimate $R(f'')$ is seen to be

$$a_{\text{AMSE}} = C_1(K)C_2(f)h_{\text{AMISE}}^{10/13} \tag{2.4}$$

where,

$$C_1(K) = \{18R(K^{(4)} * K)/\sigma_{K*K}^4 R(K)^2\}^{1/13}$$
$$C_2(f) = \{R(f)R(f'')^2/R(f^{(3)})^2\}^{1/13},$$

and where $K*K$ denotes the convolution. See Hall and Marron (1987b) for this. Since the

scale of f (defined in terms of standard deviation, inter quartile range or other related measures) denoted by λ, is considered the most crucial characteristic of the dependence of $C_2(f)$ on f, $C_2(f)C_2(g_\lambda)$, where $g_\lambda(x)=g_1(x/\lambda)/\lambda$ and g_1 is any fixed probability density with unit scale factor. Motivated from (2.2), (2.4) and the above consideration, the PI, h_{PI} is defined to be the solution of the equation

$$h=\{R(K)/\sigma_K^4\hat{R}_{f''}(a_{\hat{\lambda}}(h))\}^{1/5}n^{-1/5}$$

where $\hat{\lambda}$ denotes a $\sqrt{n}$-consistent estimator of λ and

$$a_\lambda(h)=C_1(K)C_2(g_\lambda)h^{10/13}. \tag{2.5}$$

3. The Theorems

In this section we recall theorem 3.3 of Park and Marron (1990) and it is followed by the new improved theorem. Throughout this paper it is assumed that.

(A1) K is a symmetric probability density function with finite support.

(A2) K has four Hölder continuous derivatives.

The asymptotic results depend on the amount of the smoothness of the underlying density function f. A popular measure of this is $\nu=\xi+\eta$, where ξ is an integer and $\eta\in(0,1)$ such that for some $M>0$,

$$|f^{(2+\xi)}(x)-f^{(2+\xi)}(y)|\leq M|x-y|^{\eta} \tag{3.1}$$

for all x and y. We say f has smoothness of order ν when (3.1) is satisfied. The condition on f now is given by

(A3) The underlying density function f has smoothness of order $\nu>0$ with finite support. (The assumption of finite support can be loosened in such a way that f has exponential decreasing tails, which needs a little more complicated proof.)

Here is the old theorem.

Theorem 3.1(Park and Marron (1990)) Under conditions (A2), (A2) and (A3),

(a) when $0<\nu\leq 2$,

$$\hat{h}_{PI}/h_{MISE}-1=O_P(n^{-2\nu/13})$$

(b) when $\nu>2$,

$$n^{4/13}(\hat{h}_{PI}/h_{MISE}-1)\Rightarrow N(\mu_{PI},\sigma_{PI}^2)$$

where

$$\mu_{PI}=\frac{1}{5}\{C_1(K)C_2(g_\lambda)\}^2R(f'')R(K)^{4/13}\sigma_K^{10/13}R(f'')^{-17/13}$$

$$\sigma_{PI}^2=\frac{2}{25}\sigma_K^{72/13}R(\phi)R(f)R(K)^{-18/13}R(f'')^{-8/13}/\{C_1(K)C_2(g_\lambda)\}^9$$

$$\phi(x)=K''*K''(x)=\int K''(t)K''(x+t)dt.$$

The fact that the rather strong smoothness conditions on f to get the advertized rate of convergences are attributed to the method of the proof used in Park and Marron (1990) and are not intrinsic to the nature of the PI is demonstrated by the following new theorem.

Theorem 3.2 Under the same conditions as in Theorem 3.1,

(a) when $0<\nu\leq 1$,

$$\hat{h}_{PI}/h_{MISE}-1=O_P(n^{-4\nu/13})$$

(b) when $\nu>1$,

$$n^{4/13}(\hat{h}_{PI}/h_{MISE}-1)\Rightarrow N(\mu_{PI},\sigma_{PI}^2)$$

where μ_{PI} and σ_{PI}^2 are the same as in Theorem 3.1.
The proof of the theorerm is deferred to section 5.

4. Class of Plug-in Bandwidth Selectors

In the previous sections and Park and Marron (1990), we used

$$a_\lambda(h)=C_1(K)C_2(g_\lambda)h^{10/13}$$

for the PI. However, we may use instead

$$a_\lambda(h,p)=C_1(K,p)C_2(g_\lambda,p)h^pn^{-q} \qquad (4.1)$$

where $q = p/5 - 2/13$ and

$$C_1(K,p) = \{18R(K^{(4)}, K)\sigma_K^{52p/5} / \sigma_{K*K}R(K)^{13p/5}\}^{1/13}$$
$$C_2(f,p) = \{R(f)R(f')^{13p/5} / R(f'')^2\}^{1/13}.$$

The fact that we may use $a_\lambda(h,p)$, instead of $a_\lambda(h)$, is motivated from the equality

$$a_{AMSE} = C_1(K,p)C_2(f,p)h_{AMSE}{}^p n^{-q} \tag{4.2}$$

where a_{AMSE} is defined in (2.4). Using $a_\lambda(h,p)$ defines a class of various PI's when p ranges over the real line. A particular choice of $p = 10/13$ corresponds to the PI considered in sections 2 and 3.

From (4.1) and the same considerations discussed in section 2, a PI, $\hat{h}_{PI}(p)$ is defined to be the solution of the equation

$$h = \{R(K) / \sigma_K{}^4 \hat{R}_{f^\nu}(\hat{a_\lambda}(h,p))\}^{1/5} n^{-1/5}.$$

How close these PI's are to their theoretical optimum is asymptotically quantified by the following theorem

Theorem 4.1 Under the same conditions as in theorem 3.2,

(a) when $0 < \nu < 1$,

$$\hat{h}_{PI}(p) / h_{MISE} - 1 = O_p(n^{-4\nu/13})$$

(b) when $\nu > 1$,

$$n^{4/13}(\hat{h}_{PI}(p) / h_{MISE} - 1) \Rightarrow N(\mu_{PI}(p), \sigma_{PI}{}^2(p))$$

where

$$\mu_{PI}(p) = \frac{1}{5}\{C_1(K,p)C_2(g_\lambda,p)\}^2 R(f''')R(K)^{2p/5}\sigma_K{}^{2-(8p/5)}R(f')^{1-(2p/5)}$$

$$\sigma_{PI}{}^2(p) = \frac{2}{25}\sigma_K{}^{36p/5}R(\phi)R(f)R(K)^{-9p/5}R(f')^{-2+(9p/5)} / \{C_1(K,p)C_2(g_\lambda,p)\}^9.$$

The proof of the theorem is very similar to that of Theorem 3.2. Hence it is deleted.

Remark 4.1 Plugging p=10/13 in $\mu_{PI}(p)$ and $\sigma_{PI}^2(p)$ reduces them to those in Theorem 3.2 as is expected.

The asymptotic MSE can be used to assess the performance of $\hat{h}_{PI}(p)$. To compare more effectively the asymptotic MSE's of various PI's with p and g_1 varying, first note that $\sigma_{K*K}^2=2\sigma_K^2$ and $R(K^{(4)}*K)=R(\phi)$. Then the asymptotic MSE of $\hat{h}_{PI}(p)$ can be expressed in the following simple form:

$$\mathrm{MSE}(p)=\sigma_{PI}^2(p)+\mu_{PI}^2(p)=C(f,K)\{Q_P^4+\frac{4}{9}Q_P^{-9}\}$$

where

$$C(f,K)=\frac{1}{25}\sigma_K^4R(f)^{4/13}R(f')^{-2}R(f''')^{18/13}\{18R(\phi)/\sigma_{K*K}^4\}^{4/13}$$

$$Q_p=C_2(g_\lambda,p)/C_2(f,p)=C_2(g_1,p)/C_2(f_1,p)$$

and

$$f(\cdot)=f_1(\cdot/\lambda)/\lambda.$$

Hence the MSE depends on the choices of p and g_1 only through Q_p. The MSE is minimized when $Q_p=1(f_1\equiv g_1)$ and getting larger as Q_p gets away from 1.

An important issue here is which p makes the MSE(p) insensitive to the choice of the reference density g_1. To see this, first observe that

$$Q_p^{13}=\{R(g_1)/R(f_1)\}\{R(f_1''')/R(g_1''')\}^2\{R(g_1'')/R(f_1'')\}^{13p/5}.$$

For insensitivity of the MSE(p) to reference density, it seems that we need $p>0$. For, Q_p^{13} depends on the roughness of g_1 through $T_1\equiv\{R(f_1''')/R(g_1''')\}^2$ and $T_2\equiv\{R(g_1'')/R(f_1'')\}^{13p/5}$, and increase (decrease) in T_1 is somehow balanced by decrease (increase) in T_2, which is impossible for the choice of $p\leq 0$. However, which positive p is more relevant can not be determined at present stage. More analysis on this may be done in furture study.

Remark 4.2 (personal communication with J.S.Marron and C.Jones)
As we discussed early in this section, $a_\lambda(h,p)$ in (4.1) is motivated from the equation (4.2). But keeping in mind that our final goal is to estimate f effectively, not to estimate $R(f'')$, we may consider to use

$$a_\lambda(h,p,q)= C_1(K,p,q)C_2(f,p,q)h^p n^{-q} \tag{4.3}$$

where q is no longer $p/5-2/13$ and instead set to be arbitrary. And then a pair (p,q) can be chosen to minimize the asymptotic MSE of the PI's defined through using $a_\lambda(h,p,q)$ in (4.3). With this optimal choice of (p,q), the rate of convergence is expected to be improved, i.e., faster than $n^{-4/13}$.

5. Proofs

We first introduce a lemma which has been developed by Hall and Marron (1989) and has a major role in expanding the bias of the PI.

Lemma 5.1 If the function H vanishes outside a compact set and satisfies

$$|H(x+y)-H(x)| \le c|y|^q \quad \text{for } -\infty<x,y<\infty$$

where $0\le q\le 1$, then

$$\int_{-\infty}^{\infty} H(x)\{H(x+\varepsilon)-H(x)\}dx= O(|\varepsilon|^{2q}) \quad \text{as } \varepsilon\to 0.$$

The proof of lemma 5.1 is given in detail in Hall and Marron (1989).

The following lemma is an improvement of Lemma 6.1 of Park and Marron (1990). The improvement is attributed to Lemma 5.1.

Lemma 5.2 Under condition (A1), (A2) and (A3),

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

Proof Observe that

$$\text{MISE}'(h) = -n^{-1}h^{-2}R(K) - 2(1-n^{-1}) \int\int\int uK(u)K(u+w)f(x)f'(x+hw)\,dudwdx + 2\int\int uK(u)f(x)f'(x-hu)\,dudx.$$

By a Taylor expansion with an exact remainder, we can deduce that

$$\begin{aligned}&\int f(x)f'(x+hw)\,dx = \int f(x)\{f'(x+hw)-f'(x)\}\,dx \\ &= \sum_{i=1}^{1+\xi} \frac{(-1)^i}{(2i-1)!} h^{2i-1} R(f^{(i)}) w^{2i-1} \\ &+ \frac{(-1)^{2+\xi}}{(2\xi+2)!} h^{2\xi+3} \left\{ \int\int\int_0^1 (1-t)^{2\xi+2} f^{(2+\xi)}(x) f^{(2+\xi)}(x+thw)\,dtdx \right\} w^{2\xi+3}.\end{aligned} \tag{5.1}$$

Hence

$$\begin{aligned}&\int\int\int uK(u)K(u+w)f(x)f'(x+hw)\,dudwdx \\ &= \sum_{i=1}^{\xi+2} \frac{(-1)^i}{(2i-1)!} h^{2i-1} R(f^{(i)}) \int\int u\, w^{2i-1} K(u)K(u+w)\,dudw \\ &+ \frac{(-1)^{\xi+2}}{(2\xi+2)!} h^{2\xi+3} \int\int\int\int_0^1 (1-t)^{2\xi+2} f(2+\xi)(x) \\ &\{f^{(2+\xi)}(x+thw) - f^{(2+\xi)}(x)\} w^{2\xi+3} u\, K(u)K(u+w)\,dtdudwdx.\end{aligned} \tag{5.2}$$

By the condition (A2) and Lemma 4.1, the second term of the right hand side of (5.2) is of order $h^{2\nu+3}$. Similarly we have

$$\begin{aligned}&\int\int u\, K(u)f(x)f'(x-hu)\,dudx \\ &= \sum_{i=1}^{\xi+2} \frac{(-1)^{i+1}}{(2i-1)!} h^{2i-1} R(f^{(i)}) \int u^{2i} K(u)\,du + O(h^{2\nu+3}).\end{aligned} \tag{5.3}$$

Plugging (5.2) and (5.3) into (5.1) entails

$$\begin{aligned}\text{MISE}'(h) = &-n^{-1}h^{-2}R(K) + h^3\sigma_K^4 R(f'') + 2n^{-1}h\sigma_K^2 R(f') \\ &+ O(n^{-1}h^3 + h^{2m+3})\end{aligned}$$

where $m = min(1,\nu)$.

Hence

$$D(h) = \text{AMISE}'(h) - \text{MISE}'(h) = 2n^{-1}h\sigma_K^2 R(f') + O(n^{-1}h^3 + h^{2m+3}).$$
$$= O(n^{-1}h + h^{2m+3}).$$

Now it is not so difficult to show that

$$\text{MISE}''(h) \asymp h^2.$$

Therefore

$$h_{\text{AMSE}} - h_{\text{MISE}} \asymp \frac{D(h_{\text{AMISE}})}{\text{MISE}''(h_{\text{AMISE}})}$$
$$\asymp n^{-(1+2m)/5}.$$

This concludes the proof of lemma. ■

Proof of Theorem 3.2 This proof is essentially the same as the proof of Theorem 3.3 of Park and Marron (1990). One big difference is the way of expanding the bias term of $\hat{R}_{f'}(a)$. Here we make use of Lemma 5.1. In particular, using a similar argument as in the derivation of (5.2), we obtain that for $C_{ij}(a) = (X_i - X_j)/a$,

$$\begin{aligned} E\phi(C_{ij}(a)) &= a^5 \int\int\int K(w)K(w+u)f''(x)f''(x-au)dwdudx \\ &= a^5 R(f'') + a^5 \int\int\int K(w)K(w+u)f''(x)\{f''(x-au) - f''(x)\}dwdudx \\ &= a^5 R(f'') \sum_{i=1}^{\xi} \frac{(-1)^i}{(2i)!} a^{5+2i} R(f^{(i+2)}) \left\{ \int\int\int u^{2i} K(u)K(u+w)dudw \right\} \qquad (5.4) \\ &\quad + \frac{(-1)^{\xi}}{(2\xi-1)!} a^{5+2\xi} \int\int\int\int_0^1 (1-t)^{2\xi-1} f^{(2+\xi)}(x)\{f^{(2+\xi)}(x-tau) \\ &\quad - f^{(2+\xi)}(x)\}u^{2\xi}K(u)K(u+w)dtdudwdx. \end{aligned}$$

Now by Lemma 5.1, the last term of (5.4) is of order $a^{5+2\nu}$. Therefore, we can see that

$$E\phi(C_{ij}(a)) = \begin{cases} a^5 R(f'') - a^7 \sigma_K^2 R(f''') + o(a^7) & \text{if } \nu > 1 \\ a^5 R(f'') + O(a^{5+2\nu}) & \text{if } 0 < \nu \leq 1. \end{cases}$$

Now we proceed exactly as in the proof of Theorem 3.3 of Park and Marron (1990), but utilizing Lemma 5.2 instead of Lemma 6.1 in their paper. ■

References

(1) Bowman, A. (1984). An alternative method of cross-validation for the smoothing of density estimates, *Biometrika,* 65, 521-528.

(2) Hall, P. (1980). Objective methods for the estimation of window size in the nonparametric estimation of a density, unpublished manuscript.

(3) Hall, P. (1983). Large sample optimality of least square cross-validation in density estimtion, *Annals of Statistics*, 11, 1156-1174.

(4) Hall, P. and Marron, J.S. (1987a). Extent to which least-squares cross-validation minimizes integrated square error in nonparametric density estimation, *Probability Theory and Related Fields,* 74, 567-581.

(5) Hall, P. and Marron, J.S.(1987b). Estimation of Integrated squared density derivatives, *Statistics and Probability Letters,* 109-115.

(6) Hall, P. and Marron, J.S. (1988). Local minima in cross-validation functions, unpublished manuscript.

(7) Hall, P. and Marron, J.S. (1989). Lower bounds for bandwidth selection in density estimation, unpublished manuscript.

(8) Marron, J.S. (1988). Automatic smoothing parameter selection: a survey, to appear in *Empirical Economics.*

(9) Park, B.U. and Marron, J.S. (1990). Comparison of data-driven bandwidth selectors, To appear in *Journal of the American Statistical Association.*

(10) Rudemo, M. (1982). Empirical choice of histograms and kernel density estimators, *Scandinavian Journal of Statistics,* 9, 75-78.

(11) Scott, D.W. and Terrell, G.R. (1987). Biased and unbiased cross-validation in density estimation, *Journal of the American Statistical Association,* 76, 9-15.

(12) 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.

(13) Sheather, S.J. (1986). An improved data-based algorithm for choosing the window width when estimating the density at point, *Computationl Statistics and Data Analysis* 4, 61-65.

(14) Silverman, B.W. (1986). *Density Estimation for Statistics and Data Analysis,* Chapman and Hall, New York.

(15) Stone, C. (1984). An asymptotically optimal window selection rule for kernel density estimates, *Annals of Statistics*, 12, 1285-1297.