Statistical Science
2004, Vol. 19, No. 4, 588–597
DOI 10.1214/088342304000000297

Density Estimation

Simon J. Sheather

Abstract. This paper provides a practical description of density estimation based on kernel methods. An important aim is to encourage practicing statisticians to apply these methods to data. As such, reference is made to implementations of these methods in R, S-PLUS and SAS.

Key words and phrases: Kernel density estimation, bandwidth selection, local likelihood density estimates, data sharpening.

1. INTRODUCTION

Density estimation has experienced a wide explosion of interest over the last 20 years. Silverman's (1986) book on this topic has been cited over 2000 times. Recent texts on smoothing which include detailed density estimation include Bowman and Azzalini (1997), Simonoff (1996) and Wand and Jones (1995). Density estimation has been applied in many fields, including archaeology (e.g., Baxter, Beardah and Westwood, 2000), banking (e.g., Tortosa-Ausina, 2002), climatology (e.g., Ferreyra et al., 2001), economics (e.g., DiNardo, Fortin and Lemieux, 1996), genetics (e.g., Segal and Wiemels, 2002), hydrology (e.g., Kim and Heo, 2002) and physiology (e.g., Paulsen and Heggelund, 1996).

This paper provides a practical description of density estimation based on kernel methods. An important aim is to encourage practicing statisticians to apply these methods to data. As such, reference is made to implementations of these methods in R, S-PLUS and SAS. Section 2 provides a description of the basic properties of kernel density estimators. It is well known that the performance of kernel density estimators depends crucially on the value of the smoothing parameter, commonly referred to as the bandwidth. We describe methods for selecting the value of the bandwidth in Section 3. In Section 4, we describe two recent important improvements to kernel methods, namely, local likelihood density estimates and data sharpening. We compare the performance of some of the methods that have been discussed using a new example involving data from the U.S. PGA tour in Section 5. Finally, in Section 6 we provide some overall recommendations.

Simon J. Sheather is Professor of Statistics, Australian Graduate School of Management, University of New South Wales and the University of Sydney, Sydney, NSW 2052, Australia (e-mail: simonsh@agsm.edu.au).

2. THE BASICS OF KERNEL DENSITY ESTIMATION

Let $X_1, X_2, \ldots, X_n$ denote a sample of size n from a random variable with density f.

The kernel density estimate of f at the point x is given by

$$\hat{f}_h(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - X_i}{h}\right), \tag{1}$$

where the kernel K satisfies $\int K(x)\,dx = 1$ and the smoothing parameter h is known as the bandwidth. In practice, the kernel K is generally chosen to be a unimodal probability density symmetric about zero. In this case, K satisfies the conditions

$$\int K(y)\,dy = 1,$$

$$\int yK(y)\,dy = 0,$$

$$\int y^2K(y)\,dy = \mu_2(K) > 0.$$

A popular choice for K is the Gaussian kernel, namely,

$$K(y) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{y^2}{2}\right). \tag{2}$$

Throughout this section we consider a small generated data set to illustrate the ideas presented. The data consist of a random sample of size $n = 10$ from a normal mixture distribution made up of observations from $N(\mu = -1, \sigma^2 = (1/3)^2)$ and $N(\mu = 1, \sigma^2 = (1/3)^2)$, each with probability 0.5. Figure 1 shows a kernel estimate of the density for these data using the

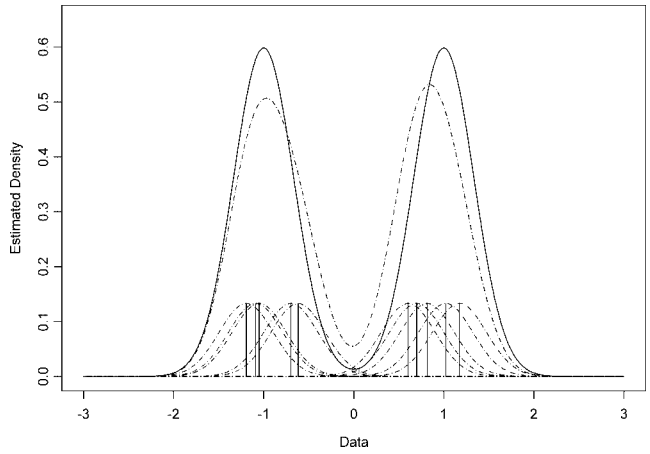

FIG. 1. *Kernel density estimate and contributions from each data point* (*dashed curve*) *along with the true underlying density* (*solid curve*).

Gaussian kernel with bandwidth $h = 0.3$ (the dashed curve) along with the true underlying density (the solid curve). The 10 data points are marked by vertical lines on the horizontal axis. Centered at each data point is each point's contribution to the overall density estimate, namely, $(1/(nh))K((x - X_i)/h)$ (i.e., $1/n$ times a normal density with mean X_i and standard deviation h). The density estimate (the dashed curve) is the sum of these scaled normal densities. Increasing the value of h widens each normal curve, smoothing out the two modes currently apparent in the estimate.

A Java applet that allows the user to watch the effects of changing the bandwidth and the shape of the kernel function on the resulting density estimate can be found at *http://www-users.york.ac.uk/~jb35/mygr2.htm*. It is well known that the value of the bandwidth is of critical importance, while the shape of the kernel function has little practical impact.

Assuming that the underlying density is sufficiently smooth and that the kernel has finite fourth moment, it can be shown using Taylor series that

$$\text{Bias}\{\hat{f}_h(x)\} = \frac{h^2}{2}\mu_2(K) f''(x) + o(h^2),$$

$$\text{Var}\{\hat{f}_h(x)\} = \frac{1}{nh}R(K) f(x) + o\left(\frac{1}{nh}\right),$$

where

$$R(K) = \int K^2(y)\, dy$$

(e.g., Wand and Jones, 1995, pages 20–21). In addition to the visual advantage of being a smooth curve, the kernel estimate has an advantage over the histogram in terms of bias. The bias of a histogram estimator with bin width h is of order h, whereas centering the kernel at each data point and using a symmetric kernel zeroes this term and as such produces a leading bias term for the kernel estimate of order h^2.

Adding the leading variance and squared bias terms produces the asymptotic mean squared error (AMSE)

$$\text{AMSE}\{\hat{f}_h(x)\} = \frac{1}{nh}R(K) f(x) + \frac{h^4}{4}\mu_2(K)^2[f''(x)]^2.$$

A widely used choice of an overall measure of the discrepancy between $\hat{f}$ and f is the mean integrated squared error (MISE), which is given by

$$\text{MISE}(\hat{f}_h) = E\left\{\int \left(\hat{f}_h(y) - f(y)\right)^2 dy\right\}$$
$$= \int \text{Bias}(\hat{f}_h(y))^2\, dy + \int \text{Var}(\hat{f}_h(y))\, dy.$$

Under an integrability assumption on f, integrating the expression for AMSE gives the expression for the asymptotic mean integrated squared error (AMISE), that is,

$$\text{AMISE}\{\hat{f}_h\} = \frac{1}{nh}R(K) + \frac{h^4}{4}\mu_2(K)^2 R(f''), \tag{3}$$

where

$$R(f'') = \int [f''(y)]^2\, dy.$$

The value of the bandwidth that minimizes the AMISE is given by

$$h_{\text{AMISE}} = \left[\frac{R(K)}{\mu_2(K)^2 R(f'')}\right]^{1/5} n^{-1/5}. \tag{4}$$

Assuming that f is sufficiently smooth, we can use integration by parts to show that

$$R(f'') = \int [f''(y)]^2\, dy = -\int f^{(4)}(y) f(y)\, dy.$$

Thus, the functional $R(f'')$ is a measure of the underlying roughness or curvature. In particular, the larger the value of $R(f'')$ is, the larger is the value of AMISE (i.e., the more difficult it is to estimate f) and the smaller is the value of h_{AMISE} (i.e., the smaller the bandwidth needed to capture the curvature in f).

3. BANDWIDTH SELECTION FOR KERNEL DENSITY ESTIMATES

In this section, we briefly review methods for choosing a global value of the bandwidth h. Where applicable, reference is made to implementations of these methods in R, S-PLUS and SAS.

In SAS, PROC KDE produces kernel density estimates based on the usual Gaussian kernel (i.e., the Gaussian density with mean 0 and standard deviation 1), whereas S-PLUS has a function density which produces kernel density estimates with a default kernel, the Gaussian density with mean 0 and standard deviation 1/4. Thus, the bandwidths described in what follows must be multiplied by 4 when used in S-PLUS. The program R also has a function density which produces kernel density estimates with a default kernel, the Gaussian density with mean 0 and standard deviation 1.

3.1 Rules of Thumb

The computationally simplest method for choosing a global bandwidth h is based on replacing $R(f'')$, the unknown part of h_{AMISE}, by its value for a parametric family expressed as a multiple of a scale parameter, which is then estimated from the data. The method seems to date back to Deheuvels (1977) and Scott (1979), who each proposed it for histograms. However, the method was popularized for kernel density estimates by Silverman (1986, Section 3.2), who used the normal distribution as the parametric family.

Let σ and IQR denote the standard deviation and interquartile range of X, respectively. Take the kernel K to be the usual Gaussian kernel. Assuming that the underlying distribution is normal, Silverman (1986, pages 45 and 47) showed that (3) reduces to

$$h_{\mathrm{AMISE}_{\mathrm{NORMAL}}} = 1.06\sigma n^{-1/5}$$

and

$$h_{\mathrm{AMISE}_{\mathrm{NORMAL}}} = 0.79\,\mathrm{IQR}\,n^{-1/5}.$$

Jones, Marron and Sheather (1996) studied the Monte Carlo performance of the normal reference bandwidth based on the standard deviation, that is, they considered

$$h_{\mathrm{SNR}} = 1.06Sn^{-1/5},$$

where S is the sample standard deviation. In SAS PROC KDE, this method is called the simple normal reference (METHOD = SNR). Jones, Marron and Sheather (1996) found that h_{SNR} had a mean that was usually unacceptably large and thus often produced oversmoothed density estimates.

Furthermore, Silverman (1986, page 48) recommended reducing the factor 1.06 in the previous equation to 0.9 in an attempt not to miss bimodality and using the smaller of two scale estimates. This rule is commonly used in practice and it is often referred to as Silverman's reference bandwidth or Silverman's rule of thumb. It is given by

$$h_{\mathrm{SROT}} = 0.9An^{-1/5},$$

where $A = \min\{\text{sample standard deviation}, (\text{sample interquartile range})/1.34\}$. In SAS PROC KDE, this method is called Silverman's rule of thumb (METHOD = SROT). In R, Silverman's bandwidth is invoked by bw = "bw.nrd0". In S-PLUS, Silverman's bandwidth with constant 1.06 rather than 0.9 is invoked by width = "nrd".

Terrell and Scott (1985) and Terrell (1990) developed a bandwidth selection method based on the maximal smoothing principle so as to produce oversmoothed density estimates. The method is based on choosing the "largest degree of smoothing compatible with the estimated scale of the density" (Terrell, 1990, page 470). Looking back at (3), this amounts to finding, for a given value of scale, the density f with the smallest value of $R(f'')$. Taking the variance σ^2 as the scale parameter, Terrell (1990, page 471) found that the beta(4, 4) family of distributions with variance σ^2 minimizes $R(f'')$. For the standard Gaussian kernel this leads to the oversmoothed bandwidth

$$h_{\mathrm{OS}} = 1.144Sn^{-1/5}.$$

In SAS PROC KDE, this method is called oversmoothed (METHOD = OS).

Comparing the oversmoothed bandwidth with the normal reference bandwidth h_{SNR}, we see that the oversmoothed bandwidth is 1.08 times larger. Thus, in practice there is often very little visual difference between density estimates produced using either the oversmoothed bandwidth or the normal reference bandwidth.

3.2 Cross-Validation Methods

A measure of the closeness of $\hat{f}$ and f for a given sample is the integrated squared error (ISE), which is given by

$$\begin{aligned}\mathrm{ISE}(\hat{f}_h) &= \int (\hat{f}_h(y) - f(y))^2\,dy \\ &= \int (\hat{f}_h(y))^2\,dy - 2\int \hat{f}_h(y)f(y)\,dy \\ &\quad + \int f^2(y)\,dy.\end{aligned}$$

Notice that the last term on the right-hand side of the previous expression does not involve h.

Bowman (1984) proposed choosing the bandwidth as the value of h that minimizes the estimate of the two other terms in the last expression, namely

$$\frac{1}{n}\sum_{i=1}^{n}\int(\hat{f}_{-i}(y))^2\,dy-\frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(X_i), \tag{5}$$

where $\hat{f}_{-i}(y)$ denotes the kernel estimator constructed from the data without the observation X_i. The method is commonly referred to as least squares cross-validation, since it is based on the so-called leave-one-out density estimator $\hat{f}_{-i}(y)$. Rudemo (1982) proposed the same technique from a slightly different viewpoint. Bowman and Azzalini (1997, page 37) provided an explicit expression for (5) for the Gaussian kernel.

Stone (1984, pages 1285–1286) provided the following straightforward demonstration that the second term in (5) is an unbiased estimate of the second term in ISE. Observe

$$\begin{aligned}E\Big[\int\hat{f}_h(y)f(y)\,dy\Big]&=\iint K\Big(\frac{y-x}{h}\Big)f(x)\,dx\,f(y)\,dy\\&=E\Big[K\Big(\frac{Y-X}{h}\Big)\Big].\end{aligned}$$

This leads to the unbiased estimate of $\int\hat{f}(y)\,f(y)\,dy$:

$$\frac{1}{n(n-1)h}\sum_{i\neq j}\sum K\Big(\frac{X_i-X_j}{h}\Big)=\frac{1}{n}\sum_{i=1}^{n}\hat{f}_{-i}(X_i).$$

Hall (1983, page 1157) showed that

$$\frac{1}{n}\sum_{i=1}^{n}\int(\hat{f}_{-i}(y))^2\,dy=\int(\hat{f}_h(y))^2\,dy+O_p\Big(\frac{1}{n^2h}\Big)$$

and hence changed the least squares cross-validation (LSCV) based criterion from (5) to

$$\mathrm{LSCV}(h)=\int(\hat{f}_h(y))^2\,dy-\frac{2}{n}\sum_{i=1}^{n}\hat{f}_{-i}(X_i),$$

since "it is slightly simpler to compute, without affecting the asymptotics." This version is the one used by most authors. We denote the value of h that minimizes $\mathrm{LSCV}(h)$ by h_{LSCV}. Least squares cross-validation is also referred to as unbiased cross-validation since

$$\begin{aligned}E[\mathrm{LSCV}(h)]&=E\Big[\int\big(\hat{f}_h(y)-f(y)\big)^2\,dy\Big]\\&\quad-\int f^2(y)\,dy\\&=\mathrm{MISE}-\int f^2(y)\,dy.\end{aligned}$$

In S-PLUS, h_{LSCV} is invoked by width = "bandwidth.ucv", while in R it is invoked by bw = "bw.ucv". The least squares cross-validation function $\mathrm{LSCV}(h)$ can have more than one local minimum (Hall and Marron, 1991). Thus, in practice, it is prudent to plot $\mathrm{LSCV}(h)$ and not just rely on the result of a minimization routine. Jones, Marron and Sheather (1996) recommended that the largest local minimizer of $\mathrm{LSCV}(h)$ be used as h_{LSCV}, since this value produces better empirical performance than the global minimizer. The Bowman and Azzalini (1997) library of S-PLUS functions contains the function $\mathrm{cv}(y,h)$ which produces values of $\mathrm{LSCV}(h)$ for the data set y over the vector of different bandwidth values in h.

For the least squares cross-validation based criterion, by using the representation

$$\begin{aligned}&\mathrm{LSCV}(h)\\&\quad=\frac{1}{nh}R(K)+\frac{2}{n^2h}\sum_{i<j}\sum\gamma\Big(\frac{X_i-X_j}{h}\Big),\end{aligned} \tag{6}$$

where $\gamma(c)=\int K(w)K(w+c)\,dw-2K(c)$, Scott and Terrell (1987) showed that

$$\begin{aligned}E[\mathrm{LSCV}(h)]&=\frac{1}{nh}R(K)+\frac{h^4}{4}\mu_2(K)^2R(f'')\\&\quad-R(f)+O(n^{-1})\\&=\mathrm{AMISE}\{\hat{f}_h\}-R(f)+O(n^{-1}).\end{aligned}$$

Thus, least cross-validation essentially provides estimates of $R(f'')$, the only unknown quantity in $\mathrm{AMISE}\{\hat{f}_h\}$.

For a given set of data, denote the bandwidth that minimizes $\mathrm{ISE}(\hat{f}_h)$ by $\hat{h}_{\mathrm{ISE}}$. A number of authors (e.g., Gu, 1998) argued that the ideal bandwidth is the random quantity $\hat{h}_{\mathrm{ISE}}$, since it minimizes the ISE for the given sample. However, $\hat{h}_{\mathrm{ISE}}$ is an inherently difficult quantity to estimate. In particular, Hall and Marron (1987a) showed that the smallest possible relative error for any data based bandwidth $\hat{h}$ is

$$\frac{\hat{h}}{\hat{h}_{\mathrm{ISE}}}-1=O_p(n^{-1/10}).$$

Hall and Marron (1987b) and Scott and Terrell (1987) showed that the least squares cross-validation bandwidth h_{LSCV} achieves this best possible convergence rate. In particular, they showed that

$$n^{1/10}\Big(\frac{h_{\mathrm{LSCV}}}{\hat{h}_{\mathrm{ISE}}}-1\Big)$$

has an asymptotic $N(0, \sigma^2_{\mathrm{LSCV}})$ distribution. The slow $n^{-1/10}$ rate of convergence means that h_{LSCV} is highly variable in practice, a fact that has been demonstrated in many simulation studies (see Simonoff, 1996, page 76 for references). In addition to high variability, least squares cross-validation "often undersmooths in practice, in that it leads to spurious bumpiness in the underlying density" (Simonoff, 1996, page 76). On the other hand, a major advantage of least squares cross-validation over other methods is that it is widely applicable.

Figure 2 shows Gaussian kernel density estimates based on two different bandwidths for a sample of 500 data points from the standard normal distribution. The 500 data points are marked as vertical bars above the horizontal axis in Figure 2. The dramatically undersmoothed density estimate (depicted by the dashed line in Figure 2) is based on the bandwidth obtained from least squares cross-validation, in this case $h_{\mathrm{LSCV}} = 0.059$, while the density estimate depicted by the solid curve in Figure 2 is based on the Sheather–Jones plug-in bandwidth, h_{SJ} (which is discussed below). In this case, $h_{\mathrm{SJ}} = 0.277$. Since both the data and the kernel in this example are Gaussian, it is possible to perform exact MISE calculations (see Wand and Jones, 1995, pages 24–25 for details). The bandwidth which minimizes MISE in this case is $h_{\mathrm{MISE}} = 0.315$.

Following the advice given above, Figure 3 contains a plot of the least squares cross-validation function $\mathrm{LSCV}(h)$ against h. It is clear from this figure that the value $h_{\mathrm{LSCV}} = 0.059$ is the unambiguous minimizer of $\mathrm{LSCV}(h)$. Thus, least squares cross-validation has performed poorly by depicting many modes in a situation in which the underlying density is easy to estimate

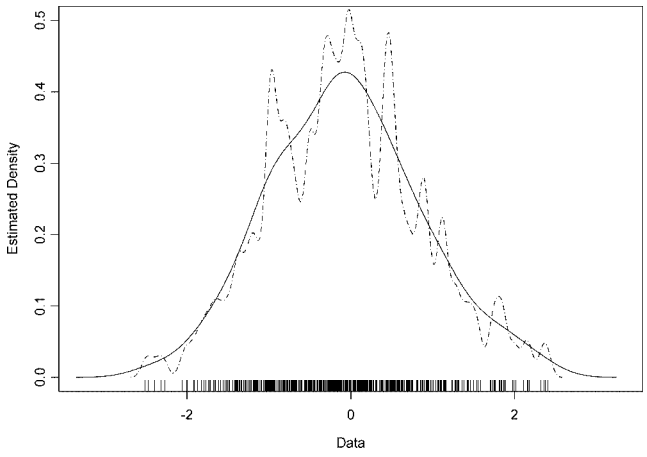

FIG. 2. *Kernel density estimates based on LSCV (dashed curve) and the Sheather–Jones plug-in (solid curve) for 500 data points from a standard normal distribution.*

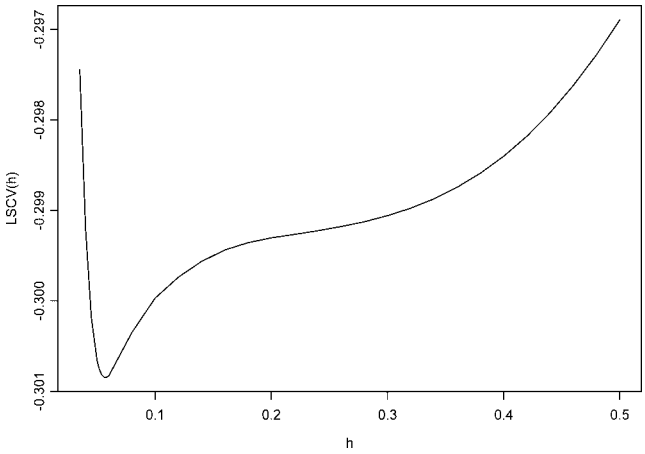

FIG. 3. *Plot of the least squares cross-validation function* $\mathrm{LSCV}(h)$ *against* h.

(see Wand and Jones, 1995, pages 36–39 for further material on measuring how difficult a density is to estimate).

Scott and Terrell (1987) proposed a method called biased cross-validation (BCV), which is based on choosing the bandwidth that minimizes an estimate of AMISE rather than an estimate of ISE. The BCV objective function is just the estimate of AMISE obtained by replacing $R(f'')$ in (3) by

$$R(\hat{f}''_h) - \frac{1}{nh^5} R(K''), \tag{7}$$

where $\hat{f}''_h$ is the second derivative of the kernel density estimate (1) and the subscript h denotes the fact that the bandwidth used for this estimate is the same one used to estimate the density f itself. The reason for subtracting the second term in (6) is that this term is the positive constant bias term that corresponds to the diagonal terms in $R(\hat{f}''_h)$.

The BCV objective function is thus given by

$$\begin{aligned}\mathrm{BCV}(h) &= \frac{1}{nh} R(K) + \frac{h^4}{4} \mu_2(K)^2 \\ &\qquad \cdot \left[R(\hat{f}''_h) - \frac{1}{nh^5} R(K'') \right] \\ &= \frac{1}{nh} R(K) + \frac{\mu_2(K)^2}{2n^2 h} \sum_{i<j} \sum \phi\left(\frac{X_i - X_j}{h} \right),\end{aligned}$$

where $\phi(c) = \int K''(w) K''(w+c)\, dw$ (Scott and Terrell, 1987). Notice the similarity of this last equation and the version of least squares cross-validation given in (6).

We denote the bandwidth that minimizes $\mathrm{BCV}(h)$ by h_{BCV}. In S-PLUS, h_{BCV} is invoked by width =

"bandwidth.bcv", while in R it is invoked by bw = "bw.bcv". Scott (1992, page 167) pointed out that $\lim_{h\to\infty} \mathrm{BCV}(h) = 0$ and hence he recommended that h_{BCV} be taken as the largest local minimizer less than or equal to the oversmoothed bandwidth h_{OS}. On the other hand, Jones, Marron and Sheather (1996) recommend that h_{BCV} be taken as the smallest local minimizer, since they claim it gives better empirical performance.

Scott and Terrell (1987) showed that

$$n^{1/10}\left(\frac{h_{\mathrm{BCV}}}{h_{\mathrm{AMISE}}} - 1\right)$$

has an asymptotic $N(0, \sigma^2_{\mathrm{BCV}})$ distribution. A related result holds for least squares cross-validation, namely, that

$$n^{1/10}\left(\frac{h_{\mathrm{LSCV}}}{h_{\mathrm{AMISE}}} - 1\right)$$

has an asymptotic $N(0, \sigma^2_{\mathrm{LSCV}})$ distribution (Hall and Marron, 1987a; Scott and Terrell, 1987). According to Wand and Jones (1995, page 80), the ratio of the two asymptotic variances for the Gaussian kernel is

$$\frac{\sigma^2_{\mathrm{LSCV}}}{\sigma^2_{\mathrm{BCV}}} \simeq 15.7,$$

thus indicating that bandwidths obtained from least squares cross-validation are expected to be much more variable than those obtained from biased cross-validation.

3.3 Plug-in Methods

The slow rate of convergence of LSCV and BCV encouraged much research on faster converging methods. A popular approach, commonly called plug-in methods, is to replace the unknown quantity $R(f'')$ in the expression for h_{AMISE} given by (3) with an estimate. The method is commonly thought to date back to Woodroofe (1970), who proposed it for estimating the density at a given point. Estimating $R(f'')$ by $R(\hat{f}''_g)$ requires the user to choose the bandwidth g for this so-called pilot estimate. There are many ways this can be done. We next describe the "solve-the-equation" plug-in approach developed by Sheather and Jones (1991), since this method is widely recommended (e.g., Simonoff, 1996, page 77; Bowman and Azzalini, 1997, page 34; Venables and Ripley, 2002, page 129).

Different versions of the plug-in approach depend on the exact form of the estimate of $R(f'')$. The Sheather and Jones (1991) approach is based on writing g, the pilot bandwidth for the estimate $R(\hat{f}'')$, as a function of h, namely,

$$g(h) = C(K)\left[\frac{R(f'')}{R(f''')}\right]^{1/7} h^{5/7},$$

and estimating the resulting unknown functionals of f using kernel density estimates with bandwidths based on normal rules of thumb. In this situation, the only unknown in the following equation is h:

$$h = \left[\frac{R(K)}{\mu_2(K)^2 R(\hat{f}''_{g(h)})}\right]^{1/5} n^{-1/5}.$$

The Sheather–Jones plug-in bandwidth h_{SJ} is the solution to this equation. In S-PLUS, h_{SJ} is invoked by width = "bandwidth.sj", while in R it is invoked by bw = "bw.SJ". In SAS PROC KDE, this method is called Sheather–Jones plug-in (METHOD = SJPI).

Under smoothness assumptions on the underlying density,

$$n^{5/14}\left(\frac{h_{\mathrm{SJ}}}{h_{\mathrm{AMISE}}} - 1\right)$$

has an asymptotic $N(0, \sigma^2_{\mathrm{SJ}})$ distribution. Thus, the Sheather–Jones plug-in bandwidth has a relative convergence rate of order $n^{-5/14}$, which is much higher than that of BCV. Most of the improvement is because BCV effectively uses the same bandwidth to estimate $R(f'')$ as it does to estimate f, while the Sheather–Jones plug-in approach uses different bandwidths. However, it is important to note that the Sheather–Jones plug-in approach assumes more smoothness of the underlying density than either LSCV or BCV.

Jones, Marron and Sheather (1996) found that for easy to estimate densities [i.e., those for which $R(f'')$ is relatively small], the distribution of h_{SJ} tends to be centered near h_{AMISE} and has much lower variability than the distribution of h_{LSCV}. For hard to estimate densities [i.e., those for which $|f''(x)|$ varies widely], they found that the distribution of h_{SJ} tends to be centered at values larger than h_{AMISE} (and thus oversmooths) and again has much lower variability than the distribution of h_{LSCV}.

A number of authors recommended that density estimates be drawn with more than one value of the bandwidth. Scott (1992, page 161) advised looking at "a sequence of (density) estimates based on the sequence of smoothing parameters

$$h = h_{\mathrm{OS}}/1.05^k \quad \text{for } k = 0, 1, 2, \ldots,$$

starting with the sample oversmoothed bandwidth h_{OS} and stopping when the estimate displays some instability and very local noise near the peaks." Marron and Chung (2001) also recommend looking at a family of density estimates for the given data set based on different values of the smoothing parameter. Marron and Chung (2001, page 198) advised that this family be based around a "center point" which is "an effective choice of the global smoothing parameter." They recommended the Sheather–Jones plug-in bandwidth for this purpose. Silverman (1981) showed that an important advantage of using a Gaussian kernel is, in this case, that the number of modes in the density estimate decreases monotonically as the bandwidth h increases. This means that the number of features in the estimated density is a decreasing function of the amount of smoothing.

4. POTENTIAL IMPROVEMENTS TO KERNEL DENSITY ESTIMATES

In this section, we consider recent improvements to kernel density estimates. We focus on two such improvements, namely local likelihood density estimates and data sharpening for density estimation.

4.1 Local Likelihood Density Estimates

One of the shortcomings of kernel density estimates is increased bias at and near the boundary. Wand and Jones (1995, pages 46–47) provided a detailed discussion of this phenomenon. One way to overcome this bias is to use a so-called boundary kernel, which is a modification of the kernel at and near the boundary. An alternative and more general approach is to use a local likelihood density estimate, which we discuss next.

The local log-polynomial likelihood density estimate of $f(x)$ is produced by fitting, via maximum likelihood, a density of the form

$$\begin{aligned}\psi(u) &= \psi(u-x|\theta_0,\theta_1,\ldots,\theta_p)\\ &= \exp\left[\sum_{j=0}^{p}\theta_j(u-x)^j\right]\end{aligned}$$

in the neighborhood of x. As such, $\theta=(\theta_0,\theta_1,\ldots,\theta_p)$ is chosen to maximize

$$\begin{aligned}&\frac{1}{nh}\sum_{i=1}^{n}K\left(\frac{X_i-x}{h}\right)\log\left(\psi(X_i-x|\theta)\right)\\ &\quad-\frac{1}{h}\int K\left(\frac{u-x}{h}\right)\psi(u-x|\theta)\,du.\end{aligned}\tag{8}$$

If the bandwidth h is large, then the second term in (8) is close to zero and the first term in (8) is close to proportional to the log-likelihood, assuming that the density of x is ψ.

Let $\hat{\theta}=\hat{\theta}(x)=(\hat{\theta}_0,\hat{\theta}_1,\ldots,\hat{\theta}_p)$ denote this maximum. Then the local log-polynomial likelihood density estimate of $f(x)$ of degree p is given by

$$\hat{f}_{\mathrm{LLPE}}(x)=\exp(\hat{\theta}_0)$$

(Hjort and Jones, 1996; Loader, 1996). The Loader (1999) library of S-PLUS functions, LOCFIT, contains functions that calculate local likelihood density estimates.

Taking $p=1$ in (8) produces a density estimator with asymptotic bias of the same order as a kernel density estimator and thus it too suffers from the boundary bias problem described above. Taking $p=2$ in (8) produces a density estimator with asymptotic bias identical to that of a boundary kernel, which corrects for boundary bias. In other words, any local likelihood density estimator based on $p=2$ automatically corrects for boundary bias without having to explicitly define a boundary kernel.

Another advantage of local likelihood density estimators is that choosing a high value of p in (8) produces density estimators with optimal rates of convergence without the spurious bumps and wiggles, and without the problem of taking negative values that are characteristic of higher-order kernel estimators.

On the other hand, Hall and Tao (2002) argued that kernel density estimators can have distinct advantages over local likelihood density estimators when edge effects are not present. In the log-linear case (i.e., $p=1$), Hall and Tao (2002) showed that "the asymptotic integrated squared bias (ISB) of a local log-linear estimator is strictly greater than its counterpart in the kernel case, whereas the asymptotic integrated squared variances are identical. Moreover, the ISB for local log-linear estimators can be up to four times greater, for densities that have two square integrable derivatives. Furthermore, this excess of bias occurs in cases where the bias is already large, and that fact tends particularly to exacerbate global performance difficulties experienced by local log-linear likelihood."

4.2 Data Sharpening

A new general method for reducing bias in density estimation recently was proposed by Hall and Minnotte (2002). The method is known as data sharpening, since it involves moving the data away from regions where

they are sparse toward regions where the density is higher.

The method of Hall and Minnotte (2002) is based on the following result: If H_r is a smooth distribution function with the property $\int K(u)H_r'(x - hu)\,du = f(x) + O(h^r)$ and $\gamma_r = H_r^{-1}(\bar{F})$, where $\bar{F}$ is the distribution function that corresponds to the density $\bar{f} = E(\hat{f}_h)$, then under smoothness conditions on f,

$$\frac{1}{h}E\left[K\left(\frac{x-\gamma_r(X_i)}{h}\right)\right] = f(x) + O(h^r).$$

Comparing this last result with (1), we see that replacing the data X_i by $\gamma_r(X_i)$, its so-called sharpened form, produces an estimator of $f(x)$ which is always positive and for which the bias is $O(h^r)$ $(r = 4, 6, 8, \ldots)$ rather than the $O(h^2)$ bias of $\hat{f}_h(x)$.

In practice, γ_r has to be estimated. Using a Taylor series expansion on H_r and plug-in estimators, Hall and Minnotte (2002) produced the estimators

$$\hat{\gamma}_4 = I + h^2\frac{\mu_2(K)}{2}\frac{\hat{f}'}{\hat{f}},$$

$$\hat{\gamma}_6 = \hat{\gamma}_4 + h^4\left\{\left(\frac{\mu_4(K)}{24} - \frac{\mu_2(K)^2}{2}\right)\frac{\hat{f}'''}{\hat{f}} + \frac{\mu_2(K)^2}{2}\frac{\hat{f}''\hat{f}'}{\hat{f}^2} - \frac{\mu_2(K)^2}{8}\frac{(\hat{f}')^3}{\hat{f}^3}\right\},$$

where I denotes the identity function. Thus, the data sharpened density estimator of order r is given by

$$\hat{f}_{r,h}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-\hat{\gamma}_r(X_i)}{h}\right).$$

Using a different approach, Samiuddin and El-Sayyad (1990) obtained the expression for $\hat{f}_{4,h}$. Note that the same bandwidth h is used for the original estimate $\hat{f}$, all necessary derivatives and the final sharpened estimate. This ensures that bias terms cancel.

Finally, Hall and Minnotte (2002) discovered by Monte Carlo simulation that the optimal bandwidth for a sharpened density estimator is larger than the optimal bandwidth for a second-order kernel density estimator.

5. REAL DATA EXAMPLE

In this section we compare the performance of a number of the bandwidth selection methods described in Section 3 on a new example that involves data from the PGA golf tour. A number of other examples can be found in Sheather (1992).

In this example we look at data on putts per round, which is the average number of putts per round played for the top 175 players on the 1980 and 2001 PGA tours. The data were taken from *http://www.golfweb.com/stats/leaders/r/1980/119* and *http://www.golfweb.com/stats/leaders/r/2001/119*. Interest centers on comparing the results for the two years to determine if there has been any improvement.

Figure 4 shows Gaussian kernel density estimates based on two different bandwidths for the data from 1980 and 2001 combined. The resulting 350 data points are marked as vertical bars above the horizontal axis in Figure 4. The density estimate depicted by the dashed line in Figure 4 is based on the bandwidth obtained from least squares cross-validation. In this case, $h_{\mathrm{LSCV}} = 0.054$, producing an estimate with at least four modes, while the density estimate depicted by the solid curve in Figure 4 is based on the Sheather–Jones plug-in bandwidth h_{SJ}. In this case, $h_{\mathrm{SJ}} = 0.154$, producing an estimate with just two modes, which we see below correspond to the fact that the data come from two separate years.

Figure 5 shows Gaussian kernel density estimates based on two different bandwidths for the separate data sets from 1980 and 2001. The density estimates depicted by the dashed line in Figure 5 are based on the bandwidths obtained from least squares cross-validation. In this case, $h_{\mathrm{LSCV}} = 0.061$ (for 2001) and 0.187 (for 1980), while the density estimates depicted by the solid curve in Figure 5 are based on the Sheather–Jones plug-in bandwidth h_{SJ}. In this case, $h_{\mathrm{SJ}} = 0.121$ (for 2001) and 0.158 (for 1980). While the two density estimates are very similar for 1980, the same is not true for 2001. The density estimate for

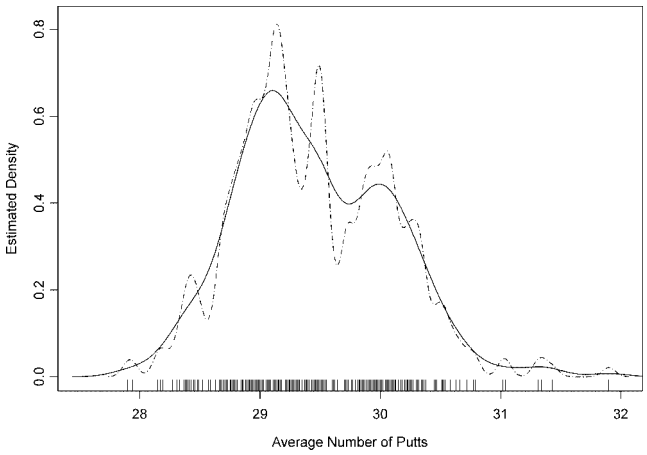

FIG. 4. *Kernel density estimates based on LSCV (dashed curve) and the Sheather–Jones plug-in (solid curve) for the data from 1980 and 2001 combined.*

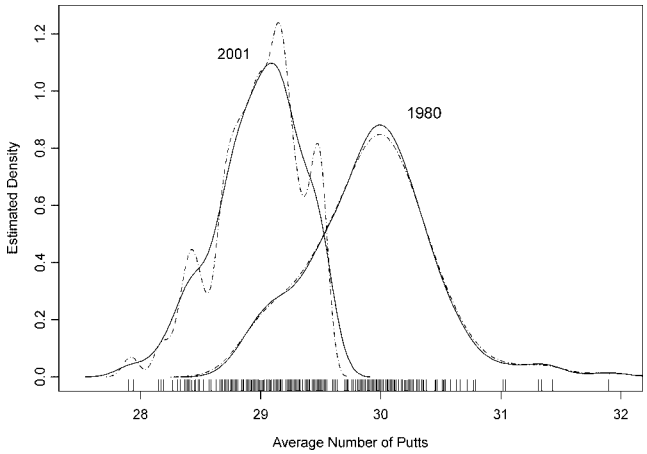

FIG. 5. *Kernel density estimates based on LSCV (dashed curve) and the Sheather–Jones plug-in (solid curve) produced separately for the data from 1980 and 2001.*

2001 based on LSCV produces at least three modes. When one considers that the data are in the form of averages taken over at least 43 rounds, the density estimates based on the Sheather–Jones plug-in bandwidth seem to be more reasonable.

Figure 6 shows a Gaussian kernel density estimate and a sixth-order sharpened estimate for the PGA data from 1980 and 2001 combined. The density estimate depicted by the solid curve in Figure 6 is based on the Sheather–Jones plug-in bandwidth h_{SJ}. The density estimate depicted by the dashed line in Figure 6 is based on the sharpened estimate with bandwidth equal to h_{SJ}. In this case, the sharpened estimate displays the two modes more distinctly.

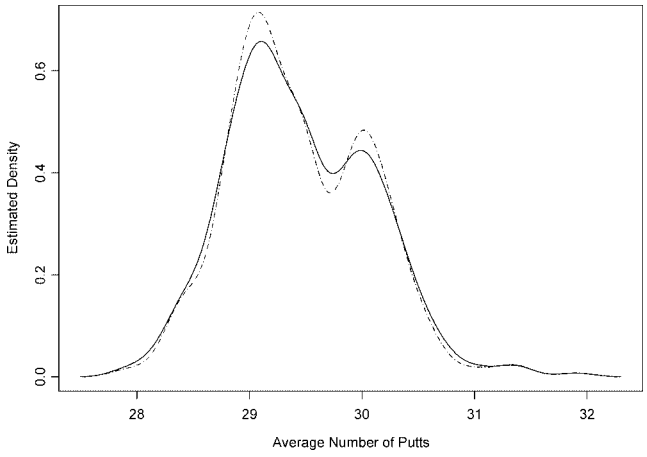

FIG. 6. *Kernel density estimate based on the Sheather–Jones plug-in (solid curve) and a sixth-order sharpened estimate for the data from 1980 and 2001 combined.*

6. RECOMMENDED APPROACH

We conclude with the following recommended approach to density estimation: Always produce a family of density estimates based on a number of values of the bandwidth. Following Marron and Chung (2001), we recommend that this set of estimates be based around a "center point" bandwidth. Natural choices of this center point bandwidth include the Sheather–Jones plug-in bandwidth h_{SJ} and least squares cross-validation h_{LSCV}. The Sheather–Jones plug-in bandwidth is widely recommended due to its overall good performance. However, for hard to estimate densities [i.e., those for which $|f''(x)|$ varies widely] it tends to oversmooth. In these situations, least squares cross-validation often provides some protection against this, due to its tendency to undersmooth. Recall that it is important to plot the least squares cross-validation function LSCV(h) and not just rely on the result of a minimization routine. Finally, density estimates with more than one mode can be generally improved by using a higher-order sharpened estimate.

ACKNOWLEDGMENT

Michael Minnotte kindly provided his S-PLUS functions for calculating sharpened density estimates.

REFERENCES

BAXTER, M. J., BEARDAH, C. C. and WESTWOOD, S. (2000). Sample size and related issues in the analysis of lead isotope data. *J. Archaeological Science* **27** 973–980.

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

BOWMAN, A. and AZZALINI, A. (1997). *Applied Smoothing Techniques for Data Analysis*: *The Kernel Approach with S-Plus Illustrations*. Oxford Univ. Press.

DEHEUVELS, P. (1977). Estimation nonparamétrique de la densité par histogrammes généralisés. *Rev. Statist. Appl.* **25** 5–42.

DINARDO, J., FORTIN, N. M. and LEMIEUX, T. (1996). Labor market institutions and the distribution of wages, 1973–1992: A semiparametric approach. *Econometrica* **64** 1001–1044.

FERREYRA, R. A., PODESTA, G. P., MESSINA, C. D., LETSON, D., DARDANELLI, J., GUEVARA, E. and MEIRA, S. (2001). A linked-modeling framework to estimate maize production risk associated with ENSO-related climate variability in Argentina. *Agricultural and Forest Meteorology* **107** 177–192.

GU, C. (1998). Model indexing and smoothing parameter selection in nonparametric function estimation (with discussion). *Statist. Sinica* **8** 607–646.

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

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

HALL, P. and MARRON, J. S. (1987b). Estimation of integrated squared density derivatives. *Statist. Probab. Lett.* **6** 109–115.

HALL, P. and MARRON, J. S. (1991). Local minima in cross-validation functions. *J. Roy. Statist. Soc. Ser. B* **53** 245–252.

HALL, P. and MINNOTTE, M. C. (2002). Higher order data sharpening for density estimation. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 141–157.

HALL, P. and TAO, T. (2002). Relative efficiencies of kernel and local likelihood density estimators. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 537–547.

HJORT, N. L. and JONES, M. C. (1996). Locally parametric nonparametric density estimation. *Ann. Statist.* **24** 1619–1647.

JONES, M. C., MARRON, J. S. and SHEATHER, S. J. (1996). A brief survey of bandwidth selection for density estimation. *J. Amer. Statist. Assoc.* **91** 401–407.

KIM, K.-D. and HEO, J.-H. (2002). Comparative study of flood quantiles estimation by nonparametric models. *J. Hydrology* **260** 176–193.

LOADER, C. R. (1996). Local likelihood density estimation. *Ann. Statist.* **24** 1602–1618.

LOADER, C. R. (1999). *Local Regression and Likelihood*. Springer, New York.

MARRON, J. S. and CHUNG, S. S. (2001). Presentation of smoothers: The family approach. *Comput. Statist.* **16** 195–207.

PAULSEN, O. and HEGGELUND, P. (1996). Quantal properties of spontaneous EPSCs in neurones of the Guinea-pig dorsal lateral geniculate nucleus. *J. Physiology* **496** 759–772.

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

SAMIUDDIN, M. and EL-SAYYAD, G. M. (1990). On nonparametric kernel density estimates. *Biometrika* **77** 865–874.

SCOTT, D. W. (1979). On optimal and data-based histograms. *Biometrika* **66** 605–610.

SCOTT, D. W. (1992). *Multivariate Density Estimation: Theory, Practice and Visualization*. Wiley, New York.

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

SEGAL, M. R. and WIEMELS, J. L. (2002). Clustering of translocation breakpoints. *J. Amer. Statist. Assoc.* **97** 66–76.

SHEATHER, S. J. (1992). The performance of six popular bandwidth selection methods on some real data sets (with discussion). *Comput. Statist.* **7** 225–281.

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

SILVERMAN, B. W. (1981). Using kernel density estimates to investigate multimodality. *J. Roy. Statist. Soc. Ser. B* **43** 97–99.

SILVERMAN, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

SIMONOFF, J. S. (1996). *Smoothing Methods in Statistics*. Springer, New York.

STONE, C. J. (1984). An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.* **12** 1285–1297.

TERRELL, G. R. (1990). The maximal smoothing principle in density estimation. *J. Amer. Statist. Assoc.* **85** 470–477.

TERRELL, G. R. and SCOTT, D. W. (1985). Oversmoothed nonparametric density estimates. *J. Amer. Statist. Assoc.* **80** 209–214.

TORTOSA-AUSINA, E. (2002). Financial costs, operating costs, and specialization of Spanish banking firms as distribution dynamics. *Applied Economics* **34** 2165–2176.

VENABLES, W. N. and RIPLEY, B. D. (2002). *Modern Applied Statistics with S*, 4th ed. Springer, New York.

WAND, M. P. and JONES, M. C. (1995). *Kernel Smoothing*. Chapman and Hall, London.

WOODROOFE, M. (1970). On choosing a delta-sequence. *Ann. Math. Statist.* **41** 1665–1671.