Bandwidth Selection for Nonparametric Distribution Estimation

Bruce E. Hansen*
University of Wisconsin†

www.ssc.wisc.edu/~bhansen

May 2004

Abstract

The mean-square efficiency of cumulative distribution function estimation can be improved through kernel smoothing. We propose a plug-in bandwidth rule, which minimizes an estimate of the asymptotic mean integrated squared error.

*Research supported by the National Science Foundation.
†Department of Economics, 1180 Observatory Drive, University of Wisconsin, Madison, WI 53706

1 Introduction

The classic nonparametric estimator of the distribution function is the empirical distribution function (EDF). This estimator is widely used in practice despite the known fact that smoothing can produce efficiency gains in finite samples. Proposals to use kernel smoothing for distribution estimation dates back to Nadaraya (1964), and demonstrations of the resulting asymptotic efficiency gains include Azzalini (1981), Reiss (1981), and Jin and Shao (1999).

Kernel smoothing requires the choice of a bandwidth parameter. This choice is critical, as under- or over-smoothing can substantially reduce precision. A cross-validation selection method was proposed by Sarda (1993), but this method has deficiencies as shown by Altman and Leger (1995). The latter proposed a plug-in estimate which minimizes an estimate of the mean weighted integrated squared error, using the density function as a weight function. This weighting choice can result in oversmoothing in the tails, whose sampling error has been downweighted.

Instead, our measure of precision is the unweighted mean integrated squared error (MISE). This results in a simple expression for the MISE, as shown by Jones (1990). The optimal bandwidth only depends on the roughness of the first derivative of the density, which is straightforward to estimate. We show how to construct reference and plug-in estimators of the optimal bandwidth. A simulation study illustrates the gains achieved by smoothing. Gauss code to implement the method is available on the author's webpage.

2 Distribution Estimation

For a random variable Y with random sample $\{Y_1, ..., Y_n\}$ let $F(y)$ denote its distribution function, $f(y)$ the density, and $f^{(m)}(y)$ its m'th derivative. Let

$$R_m = \int_{-\infty}^{\infty} \left(f^{(m)}(y)\right)^2 dy \tag{1}$$

denote the "roughness" of $f^{(m)}$. Furthermore, let $\phi(y)$ and $\Phi(y)$ denote the standard normal density and distribution, respectively.

Let $k(x)$ denote any symmetric $2'$nd order kernel normalized to have unit variance. (That is, $\int_{-\infty}^{\infty} x^2 k(x)dx = 1$.) For a bandwidth $h > 0$ let $k_h(u) = h^{-1}k(u/h)$. The kernel estimator of $f(y)$ is

$$\hat{f}_h(y) = \frac{1}{n}\sum_{i=1}^{n} k_h\left(y - Y_i\right). \tag{2}$$

The density estimator can be integrated to obtain the smoothed distribution function (SDF) estimator of $F(x)$

$$\hat{F}_h(y) = \int_{-\infty}^{y} \hat{f}_h(u)du = \frac{1}{n}\sum_{i=1}^{n} K_h(y - Y_i)$$

where

$$K(x) = \int_{-\infty}^{x} k(u)du$$

is the integrated kernel and $K_h(u) = K(u/h)$. For example, if $k(x) = \phi(x)$ then $K(x) = \Phi(x)$.

As shown by Jones (1990) if $hn^{1/2} \to \infty$ as $n \to \infty$ then

$$\begin{aligned} AMISE(h) &= \int_{-\infty}^{\infty} E\left(\hat{F}_h(y) - F(y)\right)^2 dy \\ &= \frac{V}{n} - \frac{h\psi}{n} + \frac{h^4 R_1}{4} + O(h^4) \end{aligned}$$

where $V = \int_{-\infty}^{\infty} F(y)(1 - F(y))dy$, R_1 is defined in (1) and

$$\psi = 2\int_{-\infty}^{\infty} xK(x)k(x)dx > 0$$

is a constant which depends only on the kernel. For example, if $k(x) = \phi(x)$, then $\psi = 1/\sqrt{\pi}$.

The AMISE is minimized by setting h equal to

$$h_0 = (\psi/R_1)^{1/3} n^{-1/3}. \tag{3}$$

The optimal AMISE is

$$AMISE(h_0) = \frac{V}{n} - \frac{3\psi^{4/3}}{n^{4/3} 4 R_1^{1/3}}$$

which is lower than that of the EDF. From this expression we learn the following. First, while the improvement over the SDF disappears as $n \to \infty$, it does so at the slow rate of $n^{-1/3}$, which suggests that the SDF may have meaningful finite sample gains over the EDF. Second, the improvement of the SDF is inversely proportional to R_1. Thus we expect the gains to be minimal when the density is steep. Third, the choice of kernel k only affects the AMISE through ψ (larger values reduce the AMISE). As shown by Jones (1990), this is maximized by the uniform kernel, but the difference relative to other standard kernels is small and is unlikely to have a small sample effect on efficiency.

3 Bandwidth Selection

The AMISE optimal bandwidth (3) depends on the unknown roughness R_1. A simple choice is a normal scale estimate. If $f = \phi_\sigma$ then

$$R_1 = \int_{-\infty}^{\infty} \left(\phi_\sigma^{(1)}(y)\right)^2 dx = \frac{1}{\sigma^3}\int_{-\infty}^{\infty} y^2 \phi(y)^2 dx = \frac{1}{\sigma^3 4\sqrt{\pi}}.$$

Thus a reference bandwidth is

$$\hat{h}_0 = \hat{\sigma}\left(4\sqrt{\pi}\psi\right)^{1/3} n^{-1/3} \tag{4}$$

where $\hat{\sigma}$ is the sample standard deviation. In particular, for the normal kernel ($k = \phi$) then $\hat{h}_1 = 1.59\hat{\sigma}n^{-1/3}$. The reference bandwidth, however, may work poorly for distributions which are far from the normal.

A plug-in bandwidth replaces R_1 in (3) by an estimate. Efficient estimation relies on knowledge of higher order derivatives.

By repeated integration by parts, R_m as defined in (1) equals the expectation

$$\begin{aligned} R_m &= (-1)^m \int_{-\infty}^{\infty} f^{(2m)}(y) f(y) dy \\ &= (-1)^m E f^{(2m)}(Y). \end{aligned} \tag{5}$$

An estimate of $f^{(2m)}(y)$ using a Gaussian kernel with bandwidth a is

$$\hat{f}_a^{(2m)}(y) = \frac{1}{n} \sum_{i=1}^{n} \phi_a^{(2m)} (y - Y_i). \tag{6}$$

Computationally, it is useful to observe that

$$\phi_a^{(2m)} (y) = a^{-(1+2m)} He_{2m} \left(\frac{y}{a} \right) \phi \left(\frac{y}{a} \right)$$

where $He_m(x)$ is the $m'th$ Hermite polynomial, the orthogonal polynomials with respect to $\phi(x)$. For example, $He_0(x) = 1$, $He_1(x) = x$, and $He_2(x) = x^2 - 1$. The Hermite polynomials satisfy the recursive relation $He_{m+1}(x) = xHe_m(x) - mHe_{m-1}(x)$.

Equations (5) and (6) motivate the Jones and Sheather (1991) estimator

$$\begin{aligned} \hat{R}_m(a) &= (-1)^m \frac{1}{n} \sum_{j=1}^{n} \hat{f}_a^{(2m)} (Y_j) \\ &= (-1)^m \frac{1}{n^2} \sum_{i,j=1}^{n} \phi_a^{(2m)} (Y_j - Y_i). \end{aligned} \tag{7}$$

Jones and Sheather (1991) Result 2 show that the bandwidth a which minimizes the asymptotic mean-square error of $\hat{R}_m(a)$ is

$$\begin{aligned} a_m(R_{m+1}) &= \left(\frac{(-1)^m 2\phi^{(2m)}(0)}{R_{m+1} n} \right)^{1/(2m+3)} \\ &= \left(\frac{(2m)!}{m! 2^{m-1/2} \sqrt{\pi} R_{m+1} n} \right)^{1/(2m+3)} \\ &= \left(\frac{2^{m+1/2} \Gamma\left(m + \frac{1}{2}\right)}{\pi R_{m+1} n} \right)^{1/(2m+3)} \end{aligned} \tag{8}$$

where the second equality uses equation (7.2) of Marron and Wand (1992), and the third is an algebraic

simplification. We write $a_m(R_{m+1})$ as a explicit function of R_{m+1} to emphasize the dependence on the unknown roughness.

(7) and (8) can be combined to write the estimate of R_m as a function of the input R_{m+1}.

$$\hat{R}_m(R_{m+1}) = \hat{R}_m(a_m(R_{m+1})).$$

This relationship suggests the sequential plug-in rule. Fix $J \geq 1$ and take R_{J+1} as given. Then estimate R_J by $\hat{R}_J = \hat{R}_J(R_{J+1})$, R_{J-1} by $\hat{R}_{J-1} = \hat{R}_{J-1}\left(\hat{R}_J\right)$, etc, until we obtain an estimate $\hat{R}_1 = \hat{R}_1\left(\hat{R}_2\right)$ for R_1. We can write this as an explicit function of the input R_{J+1}:

$$\hat{R}_1(R_{J+1}) = \hat{R}_1\left(\hat{R}_2\left(\hat{R}_3\left(\cdots \hat{R}_J(R_{J+1})\right)\right)\right).$$

As this sequential plug-in estimator depends on the input R_{J+1}, we suggest a Gaussian reference estimate. When $f = \phi_\sigma$ then

$$R_{J+1} = R^0_{J+1} = \frac{(2J+2)!}{\sigma^{2J+3}(J+1)!2^{2J+3}\sqrt{\pi}} = \frac{\Gamma\left(J+\frac{3}{2}\right)}{\sigma^{2J+3}2\pi}$$

(see equations (7.1) and (7.2) of Marron and Wand (1992)). The scale σ can be replaced by the sample standard deviation. This yields the J'th-step estimate of R_1

$$\hat{R}_{1,J} = \hat{R}_1\left(R^0_{J+1}\right)$$

and the J'th-step plug-in bandwidth

$$\hat{h}_J = \left(\psi/\hat{R}_{1,J}\right)^{1/3} n^{-1/3} \tag{9}$$

4 Simulation

The finite sample performance of the SDF estimator is explored in a simple simulation experiment. The observations are generated by the first nine mixture-normal test densities of Marron and Wand (1992), normalized to have unit variance. These densities cover a broad range of shapes and characteristics.

Samples of size $n = 10$, 30, 60 and 120 are investigated. The Gaussian kernel and 10,000 simulation draws are used for each sample size and model.

We first investigate the accuracy of the AMISE approximation

$$AMISE(h) \simeq \frac{V}{n} - \frac{h}{n\sqrt{\pi}} + \frac{h^4 R_1}{4}. \tag{10}$$

The constant V is calculated by numerical integration using 51 gridpoints over the region [-2.5, 2.5]. The roughness R_1 is calculated analytically using Theorem 4.1 of Marron and Wand (1992). The exact MISE is calculated by numerical integration (using the same gridpoints) over the simulation draws.

The exact MISE and AMISE are displayed in Figure 1 for $n = 30$. The graphs for other sample sizes are qualitatively similar. The MISE and AMISE have been normalized by the MISE for $h = 0$ (the EDF), and are graphed as functions of the bandwidth h. The solid lines are the exact MISE (computed by simulation), the dashed lines the AMISE (10). The asymptotic approximation (10) should equal the exact MISE for $h = 0$. The slight differences between the curves at $h = 0$ observed in Figure 1 are therefore due to simulation and integration error.

From Figure 1 we observe the following. First, the MISE and AMISE have similar shapes across models. In many cases the two curves are quite close, especially for small values of h. For large bandwidths, however, the can be quite divergent. A common observation is that for large h, the exact MISE curve lies significantly below the AMISE curve. Second, the bandwidth which minimizes the exact MISE and the AMISE are typically quite close. A notable exception is the Strongly Skewed distribution, for which the AMISE is a poor approximation to the exact MISE, and the bandwidth which minimizes the former is much smaller than that which minimizes the latter. Third, we observe that the MISE curves vary considerably across distributions.

These observations suggest the following lessons. The diversity of the MISE curves across distributions indicates that reference (rule-of-thumb) bandwidths may perform quite poorly if the true distribution is not close to the reference distribution. Thus data-based bandwidth selection rules are necessary for implementation. The similarity of the exact MISE and AMISE curves indicates that bandwidth selection can be based on estimates of the AMISE, such as the plug-in rules of section 3.

We next examine the performance of the SDF estimator with empirically-selected bandwidths. Again, our performance measure is the MISE, expressed as a ratio to the MISE for the EDF ($h = 0$). We calculate the SDF using the reference bandwidth (4) and the plug-in bandwidths (9) for $j = 1$ through 8. Table 1 reports the results.

The first column reports the relative MISE for the reference bandwidth. As expected, we see that the results depend strongly on the true distribution. For six of the nine distributions, the MISE of the SDF is lower than that of the EDF, but for three other distributions (Kurtotic, Outlier, and Separated Bimodal) the MISE of the SDF can be substantially higher.

The remaining columns report the relative MISE for the plug-in bandwidths $\hat{h}_J$. For the distributions where the reference bandwidth had a low MISE, the plug-in bandwidths yield similar MISE results. The notable differences arise for the three other distributions. In these cases, the use of a plug-in bandwidth substantially reduces the MISE. If a plug-in bandwidth of order four or higher is used, the MISE of the SDF dominates that of the EDF for all distributions and sample sizes investigated. Increasing the order of the plug-in bandwidth above four has mixed benefits.

In summary, we conclude that the fourth-order plug-in bandwidth $\hat{h}_4$ yields a distribution function estimator which has uniformly excellent MISE, and is recommended for empirical practice. The bandwidth $\hat{h}_4$ and estimator $\hat{F}_h(y)$ are quite straightforward to calculate. Gauss code is available on the author's webpage.

References

[1] Altman, Naomi and Christian Leger (1995): "Bandwidth selection for kernel distribution function estimation," *Journal of Statistical Planning and Inference*, 46, 195-214.

[2] Azzalini, A. (1981): "A note on the estimation of a distribution function and quantiles by a kernel method," *Biometrika*, 68, 326-328.

[3] Jin, Zhezhen and Yongzhao Shao (1999): "On kernel estimation of a multivariate distribution function," *Statistics and Probability Letters*, 41, 163-168.

[4] Jones, M.C. (1990): "The performance of kernel density functions in kernel distribution function estimation," *Statistics and Probability Letters*, 9, 129-132.

[5] Jones, M.C. and S. J. Sheather (1991): "Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives," *Statistics and Probability Letters*, 11, 511-514.

[6] Marron, J.S. and M.P. Wand (1992): "Exact mean integrated squared error," *Annals of Statistics*, 20, 712-736.

[7] Reiss, R.D. (1981): "Nonparametric estimation of smooth distribution functions," *Scandinavian Journal of Statistics*, 8, 116-119.

[8] Sarda, Pascal (1993): "Smoothing parameter selection for smooth distribution functions," *Journal of Statistical Planning and Inference*, 35, 65-75.

[9] Shao, Yongzhao and Xiaojing Xiang (1997): "Some extensions of the asymptotics of a kernel estimator of a distribution function," *Statistics and Probability Letters*, 334, 301-308.

Table 1:
Distribution Function Estimation
Normalized MISE of Smooth Distribution Function

Sample Size	Density	Bandwidth Method								
		$\hat{h}_0$	$\hat{h}_1$	$\hat{h}_2$	$\hat{h}_3$	$\hat{h}_4$	$\hat{h}_5$	$\hat{h}_6$	$\hat{h}_7$	$\hat{h}_8$
$n = 10$	Gaussian	0.69	0.70	0.71	0.72	0.72	0.72	0.73	0.73	0.74
	Skewed	0.71	0.71	0.71	0.72	0.72	0.73	0.73	0.74	0.74
	Strongly Skewed	0.92	0.90	0.89	0.88	0.88	0.88	0.88	0.88	0.88
	Kurtotic	0.97	0.96	0.94	0.93	0.92	0.91	0.90	0.89	0.89
	Outlier	2.21	1.56	1.22	1.06	0.98	0.94	0.92	0.91	0.90
	Bimodal	0.67	0.68	0.69	0.69	0.70	0.71	0.71	0.72	0.72
	Separated Bimodal	0.93	0.92	0.88	0.85	0.83	0.83	0.83	0.83	0.83
	Asymmetric Bimodal	0.67	0.68	0.69	0.70	0.70	0.71	0.72	0.72	0.73
$n = 30$	Gaussian	0.76	0.77	0.78	0.78	0.78	0.78	0.78	0.78	0.78
	Skewed	0.78	0.78	0.78	0.78	0.78	0.79	0.79	0.79	0.79
	Strongly Skewed	0.93	0.90	0.90	0.90	0.90	0.90	0.91	0.91	0.91
	Kurtotic	1.22	1.12	1.07	1.04	1.01	0.99	0.98	0.97	0.96
	Outlier	3.02	1.36	1.01	0.93	0.91	0.90	0.89	0.89	0.89
	Bimodal	0.75	0.76	0.76	0.76	0.77	0.77	0.77	0.77	0.78
	Separated Bimodal	1.20	1.01	0.91	0.89	0.88	0.88	0.88	0.88	0.88
	Asymmetric Bimodal	0.76	0.77	0.77	0.77	0.77	0.78	0.78	0.78	0.78
$n = 60$	Gaussian	0.79	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.82
	Skewed	0.82	0.81	0.82	0.82	0.82	0.82	0.82	0.82	0.82
	Strongly Skewed	0.93	0.90	0.91	0.91	0.92	0.92	0.93	0.93	0.93
	Kurtotic	1.41	1.19	1.10	1.04	1.01	0.99	0.97	0.96	0.96
	Outlier	3.27	1.18	0.95	0.91	0.90	0.90	0.90	0.89	0.89
	Bimodal	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.81	0.82
	Separated Bimodal	1.36	1.00	0.92	0.91	0.90	0.90	0.90	0.90	0.90
	Asymmetric Bimodal	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82	0.82
$n = 120$	Gaussian	0.83	0.84	0.84	0.84	0.84	0.84	0.84	0.84	0.84
	Skewed	0.84	0.84	0.84	0.84	0.84	0.85	0.85	0.85	0.85
	Strongly Skewed	0.92	0.91	0.92	0.93	0.94	0.94	0.94	0.95	0.95
	Kurtotic	1.59	1.21	1.07	1.01	0.98	0.97	0.96	0.96	0.96
	Outlier	3.38	1.07	0.93	0.91	0.91	0.91	0.91	0.91	0.91
	Bimodal	0.85	0.84	0.84	0.84	0.84	0.84	0.84	0.85	0.85
	Separated Bimodal	1.43	0.98	0.93	0.92	0.92	0.92	0.92	0.92	0.92
	Asymmetric Bimodal	0.86	0.85	0.85	0.85	0.85	0.85	0.85	0.85	0.85

Figure 1
Exact and Asymptotic MISE
As a Function of Bandwidth
n=30