A Comparative Study for Bandwidth Selection in Kernel Density Estimation

Omar m. Eidous, M. Abd Alrahem Shafeq Marie, M. H. Baker Al-Haj Ebrahem

Department of Statistics, Faculty of Science,
Yarmouk University, Irbid, Jordan
omarm@yu.edu;jomohammadmarie@yahoo.com;malhaj@yu.edu.jo

ABSTRACT

Nonparametric kernel density estimation method makes no assumptions on the functional form of the curves of interest and hence allows flexible modeling of the data. Many authors pointed out that the crucial problem in kernel density estimation method is how to determine the bandwidth (smoothing) parameter.
In this paper, we introduce the most important bandwidth selection methods. In particular, least squares cross-validation, biased cross-validation, direct plug-in, solve-the-equation rules and contrast methods are considered. These methods are described and their expressions are presented. Our main practical contribution is a comparative simulation study that aims to isolate the most promising methods. The performance of each method is evaluated on the basis of the mean integrated squared error and for small-to-moderate sample size. The simulation results showed that the contrast method is the most promising methods based on the simulated families that are considered.

Key Words: Probability Density Function, Bandwidth, Least Squares Cross-Validation, Biased Cross-Validation, Contrast Method, Direct Plug-In, Solve-The-Equation Rules.

1. Introduction

Kernel method is widely used in nonparametric density estimation. It produces a kernel estimator for the unknown probability density function (*p.d.f*) $f(x)$. Many authors pointed out that the choice of the bandwidth (smoothing) parameter, *h* is crucial for the effective performance of the kernel estimator (e.g. Scott (1992)). A method that uses the data $X_1, X_2, ..., X_n$ to produce a value for the bandwidth *h* is called a bandwidth selector or data-driven selector.
Various data-driven methods for choosing the bandwidth *h* have been proposed and studied over the years. Most of these methods are based on minimizing the *MISE* or *AMISE*. Unfortunately, none of these bandwidths is available in practice since all of them depend on the unknown probability density function. See for example, Bowman (1984), Stone (1984), Hall and Marron (1985), Scott and Terrell (1987), and Sheather and Jones (1991). Marron (1988) gave a list of some methods with discussion. A survey of smoothing methods for density estimation may be found in Titterington (1985).
Sheather (1992) applied several bandwidth selectors to the Old Faithful data. Janssen, et al. (1995) developed and improved scale measures for use in bandwidth selection. Ahmad and Fan (2001) obtained the optimal theoretical bandwidth *h* in the general case. Ahmad and Mugdadi (2003) discussed data-based choices of the bandwidth and analyze the kernel density estimation.

2. Choosing the bandwidth

In this paper, a few important and popular data-driven methods for choosing h are discussed. In particular, least squares cross-validation (Rudemo (1982) and Bowman (1984)), biased cross-validation (Scott and Terrell (1987)), direct plug-in (Sheather and Jones (1991)), solve-the-equation rules (Sheather and Jones (1991)) and contrast method (Ahmad and Ran (2004)).

Let $X_1, X_2, ..., X_n$ be a random sample of size n from a continuous univariate distribution with an unknown *p.d.f* $f(x)$, then the kernel density estimator of $f(x)$, $x \in R$ is defined by (Silverman (1986)) as

$$\hat{f}(x;h) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - X_i).$$

Where $K_h(u) = h^{-1}K(uh^{-1})$, K is called kernel function assumed to be symmetric (Silverman (1986)) and h is called the bandwidth (or the smoothing parameter) that controls the degree of smoothing applied to the data. Both K and h are under the control of the user, and therefore, we need to determine them for practical purposes to consider the results about kernel estimator.

The bandwidth can be chosen to minimize the asymptotic mean integrated square error *AMISE* (Silverman (1986)). In this case, h can be obtained by minimizing .

$$MISE = \int_{-\infty}^{\infty} \left[\left(Bias\ \hat{f}(x) \right)^2 + Var\ (\hat{f}(x)) \right] dx$$

The practical implementation of the kernel density estimator requires the specification of the bandwidth h. A widely used criterion is to choose h that minimizes the *AMISE*. The bandwidth h controls the smoothness of the fitted density curve. We note here that, a larger h gives a smoother estimate with smaller variance and larger bias, while a smaller h produces a rougher estimate with larger variance and smaller bias.

3. Simulation study

A simulation study is preformed to compare among the several methods for selecting the bandwidth of a kernel density estimator. The methods which compared to estimate the bandwidth h and consequently *f(x)*. It is important to know the effects of the different methods for the estimator of *f(x)* for different values of the sample size, *n*. In this study, four different normal mixture densities are simulated, these densities are (Marron and Wand (1992)):

a. Gaussian: $f_1(x) = \phi(x)$.
b. Kurtotic unimodal: $f_2(x) = \frac{2}{3}\phi(x) + \frac{1}{3}\phi_{1/10}(x)$.
c. Bimodal: $f_3(x) = \frac{1}{2}\phi_{2/3}(x+1) + \frac{1}{2}\phi_{2/3}(x-1)$.
d. Strongly skewed:

$$f_4(x) = \sum_{l=1}^{8} \frac{1}{8}\phi_{(2/3)^{l-1}}\{x - 3[(2/3)^{l-1} - 1]\}.$$

where $\phi_A(u) = A^{-1}\phi(u/A)$ and ϕ denotes the *p.d.f* of a standard normal variable (i.e. $\phi_A(u) = \frac{1}{\sqrt{2\pi}A}e^{\frac{-1}{2A^2}u^2}$).

These densities represent Symmetric, Kurtotic unimodal, Bimodal and strongly skewed distributions respectively. Figure 1.1 displays the shapes of these densities. In fact, they are a small subset of fifteen normal mixtures used by Marron and Wand (1992).

The general normal mixtures density is given by (see Marron and Wand, 1992):

$$f(x)= \sum_{l=1}^{k} w_l \phi_{\sigma_l}(x-\mu_l)$$

Where $-\infty < \mu_l < \infty, \sigma_l > 0$ and w_l is a vector with positive entries summing to unity (weight), for *l=1, 2 ,…, k*. We say that f has a normal *k*-mixture density with parameters $\{(w_l, \mu_l, \sigma_l^2) : l = 1, 2, \ldots, k\}$.

Fryer (1976) and Deheuvels (1977) first showed that the *MISE* could be calculated exactly when both the underlying density and the kernel function are Gaussian. The integrated squared error (*ISE*) of the estimator if the true underlying density is known to be *f*(*x*) as in equation (1.37) is given by (Marron and Wand (1992))

$$ISE(\hat{f}) = \int_{-\infty}^{\infty}\left[\hat{f}(x;h) - f(x)\right]^2 dx$$

$$= \frac{1}{n^2}\sum_{i_1=1}^{n}\sum_{i_2=1}^{n}\phi_{h\sqrt{2}}(X_{i_1} - X_{i_2}) - \frac{2}{n}\sum_{i=1}^{n}\sum_{l=1}^{k} w_l \phi_{(h^2+\sigma_l^2)^{1/2}}(X_i - \mu_l) + U(h,0)$$

where $U(h,q) = \sum_{l_1=1}^{k}\sum_{l_2=1}^{k} w_{l_1} w_{l_2} \phi_{(qh^2+\sigma_{l_1}^2+\sigma_{l_2}^2)^{1/2}}(\mu_{l_1} - \mu_{l_2})$ and the kernel function *K* is the standard normal. Thus, it will be more appropriate to analyze the expected value of the *ISE*, called the *MISE*.

For each normal mixture density and each sample size *n* = 50, 100, 200, 500 that simulated from *f*(*x*), 1000 samples were artificially repeated from each *f*(*x*). Then, for each sample, the bandwidth *h* based on *LSCV*, *BCV*, *DPI*, *CONT* and *STE* methods are obtained. Subsequently, for each sample the *ISE* values are obtained by using according to the simulated density *f*(*x*). Subsequently, the *MISE* values are empirically determined as the mean of the *ISE* values obtained in each sample. Table 1.1 displays the simulation results and the *MISE*s against the sample sizes for the different underling normal mixture densities. Moreover, the relative efficiencies of the contrast (*CONT*) method against *LSCV*, *BCV*, *DPI* and *STE* methods are given in Table 1.2. The rule of relative efficiency is given by

$$RE(\hat{h}) = \frac{MISE(\hat{h}_*)}{MISE(\hat{h}_{CONT})}$$

where $\hat{h}_*$ is the bandwidth which computed from the other methods (see Table 1.2).

4. Discussion

We have shown in the previous section that the *CONT* method is a useful technique for choosing the bandwidth of the kernel estimator. The *CONT* method produces sensible estimates for *f*(*x*) in almost all cases considered (see Table 1.2). While our conclusions based on four different densities, there are many other candidate shapes for the densities in which, we may assume that the data are came from (Marron and Wand, 1992). Therefore, we do not claim that the *CONT* method performs better than the other methods for any set of data. However, based on our simulation study, we can rank the different methods in descending order according to their performances as the following: *CONT*, then *DPI*(2-stage) and *STE*(2-stage), followed by *LSCV* and then *BCV*.

© 2008 University of Salento - SIBA http://siba2.unile.it/ese

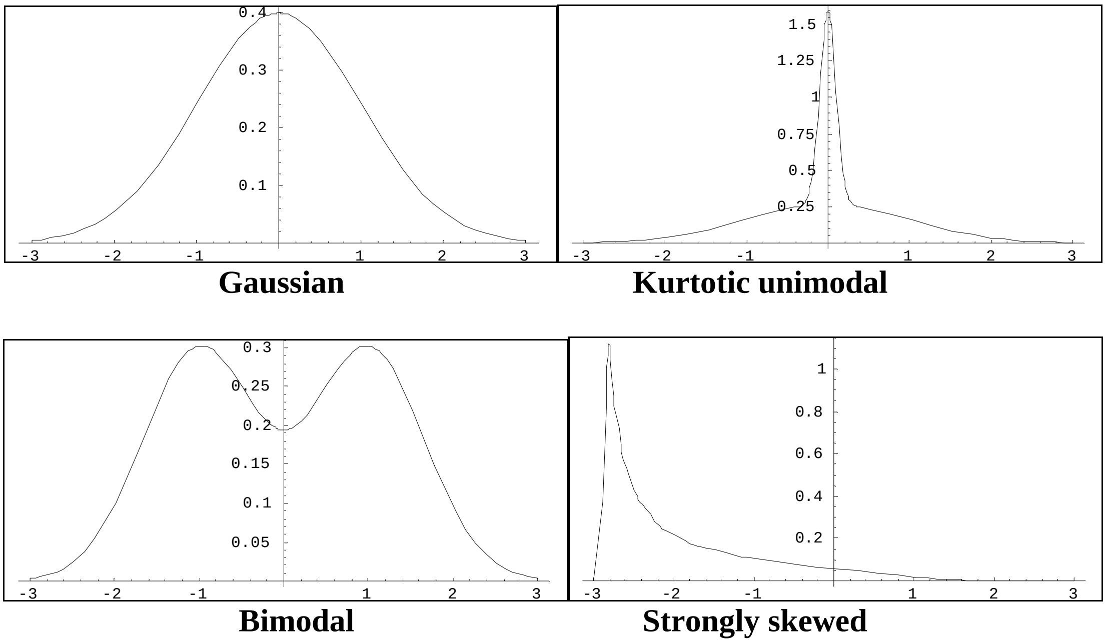

Figure 1.1 Some normal mixture densities.

METHOD	$MISE(\hat{f})$ / SAMPLE SIZE	$f_1(x)$	$f_2(x)$	$f_3(x)$	$f_4(x)$
DPI(2-stage)	50	0.12846	0.23448	0.15910	0.76199
CONT		0.12481	0.22572	0.10643	0.58228
LSCV		0.19144	0.28647	0.25236	0.94230
BCV		0.40578	0.43591	0.39941	0.76748
STE(2-stage)		0.13070	0.23831	0.17873	0.78334
DPI(2-stage)	100	0.12730	0.21665	0.15063	0.75133
CONT		0.12373	0.22057	0.09582	0.56732
LSCV		0.16841	0.26467	0.23532	0.88068
BCV		0.31693	0.38360	0.30008	0.71768
STE(2-stage)		0.12530	0.23352	0.12739	0.76518
DPI(2-stage)	200	0.12215	0.20491	0.14057	0.74043
CONT		0.11160	0.21296	0.09271	0.55792
LSCV		0.15314	0.25512	0.21145	0.83185
BCV		0.25271	0.30184	0.23965	0.60122
STE(2-stage)		0.11947	0.19695	0.12676	0.73857
DPI(2-stage)	500	0.11903	0.19237	0.13929	0.73197
CONT		0.11208	0.20337	0.09019	0.55088
LSCV		0.13948	0.24905	0.19998	0.78332
BCV		0.20559	0.28995	0.16698	0.56810
STE(2-stage)		0.10785	0.18467	0.12599	0.71857

Table 1.1 The $MISE(\hat{f})$ for different methods to choose thevalue of bandwidth.

RELATIVE EFFICIENCY	SAMPLE SIZE	$f_1(x)$	$f_2(x)$	$f_3(x)$	$f_4(x)$
$RE(h)=\frac{MISE(\hat{h}_{DPI(2-stage)})}{MISE(\hat{h}_{CONT})}$	50	1.02924	1.03880	1.49487	1.30863
	100	1.02885	0.98222	1.57201	1.32435
	200	1.09453	0.96219	1.51623	1.32712
	500	1.06200	0.94591	1.54440	1.32872
$RE(h)=\frac{MISE(\hat{h}_{LSCV})}{MISE(\hat{h}_{CONT})}$	50	1.53385	1.26913	2.37113	1.61829
	100	1.36110	1.19993	2.45585	1.55235
	200	1.37222	1.19797	2.28076	1.49098
	500	1.24446	1.22461	2.21731	1.42194
$RE(h)=\frac{MISE(\hat{h}_{BCV})}{MISE(\hat{h}_{CONT})}$	50	3.25118	1.93119	3.75279	1.31806
	100	2.56146	1.73913	3.13170	1.26503
	200	2.26442	1.41735	2.58494	1.07761
	500	1.83431	1.42572	1.85142	1.03125
$RE(h)=\frac{MISE(\hat{h}_{STE(2-stage)})}{MISE(\hat{h}_{CONT})}$	50	1.04719	1.05577	1.67932	1.34529
	100	1.01268	1.05871	1.32947	1.34876
	200	1.07052	0.92482	1.36727	1.32379
	500	0.96225	0.90804	1.39694	1.30440

Table 1.2 The relative efficiency (*RE*) for different sample sizes and different normal mixture densities.