New approach for Bandwidth Selection in the Kernel Density Estimation Based on β-Divergence

Hamza Dhaker, Papa Ngom, Pierre Mendy, El Hadji Deme

New approach for Bandwidth Selection in the Kernel Density Estimation Based on β-Divergence.

Hamza Dhaker[a,b], Papa Ngom[a,*], Pierre Mendy[b], El Hadji Deme[c]

[a]LMA,Laboratoire de Mathematiques Appliquees
[b]LMDAN,Universit Cheikh Anta Diop, Dakar-Fann, BP 5005, Senegal
[c]LERSTAD, Universite Gaston Berger, UFR SAT Saint-Louis, BP 234, Senegal

Abstract

The choice of bandwidth is crucial to the kernel density estimation KDE. Various bandwidth selection methods for KDE least squares cross-validation $LSCV$ and Kullback-Leibler cross-validation are proposed. We propose a method to select the optimal bandwidth for the KDE. The idea behind this method is to generalize the $LSCV$ method, using the measure of β-divergence, and to see the importance of improving our method, we will compares these $\mathcal{D}_\beta(\widehat{f_h}, f)$ bandwidth selector with a normal reference(NR), the last squares cross-validation($LSCV$), the Sheather and Jones (SJ) method, and the generalized $LSCV(LSCV_g)$ bandwidth selector, on simulated data. The use of the various practical bandwidth selectors is illustrated on a real data example.

Keywords: nonparametric density estimation, β-divergence, integrated squared error, bandwidth
AMS Subject Classification : 62G07, 94A17 .

1. Introduction

The problem of choosing the bandwidth (window width or smoothing parameter) h is importantly in statistical estimation of the kernel density estimation. A vast amount of literatures has been devoted in choosing practical optimal bandwidth for techniques built on kernel estimation et some comparative studies have been made to these methods. Representative surveys of bandwidth selection techniques can be found in Bowman[3], Jones et al.[9], Loader[12], Peter Hall [5] Scott[18],and Wand and Jones[24].
Least-squares cross-validation ($LSCV$) is Among the earliest bandwidth selectors, this method was suggested by Rudemo [17] and Bowman [2], in the 80s ,it has been the method of reference, but in the early 90s, studies have shown that other methods performs better from the bias points of view and much better in reducing the variance. See Park and Turlach [15] for a detailed description.
Interesting comparative studies have been published. Bowman[2] compared two methods for selecting bandwidth, The first is the Kullback-Leibler Cross-Validation and the second is that of Integrated Squared Error Cross-Validation. Scott and Terrell [19] compared the two methods by theoretical calculation of the noise in the cross-validation function and corresponding cross-validated smoothing parameters, by Monte Carlo simulation, and by example.
Sheather and Jones [20] set up a plug-in type of three-step procedure. They choose to estimate $R(f'')$ (the term unknown in $AMISE$).
Jin Zhang [26]) have proposed a generalization the classical least squares cross-validation ($LSCV$) selector for its variability and under smoothing, He did a comparison of bandwidths for finite sample behavior.
For a more complete treatment, from a historical viewpoint, with complete references, and detailed discussion of variations that have been suggested, see Jones et al. [10] Quick access to implementation of most of the methods discussed here has been provided by park and Jones et al. [9] .

*Corresponding author
Email address: papa.ngom@ucad.edu.sn (Papa Ngom)

The main purpose of this paper is to investigate the optimal bandwidth h_β for minimizing criterion $\mathcal{D}_\beta(\widehat{f_n}, f)$ error. We will see that $\mathcal{D}_\beta(\widehat{f_n}, f)$ generalize the Integrated Square Error (ISE) and Kullback-Leiber divergence (KL).
After having introduced the $\mathcal{D}_\beta(\widehat{f_n}, f)$ selection method we study the finite sample performances of various bandwidth selectors via a simulation study. We compare five procedures: the normal reference (NR) method, the last squares cross-validation ($LSCV$), the Sheather and Jones (SJ) method, the generalized $LSCV(LSCV_g)$ and criterion $\mathcal{D}_\beta(\widehat{f_n}, f)$ error.
This paper is organized as follows. Section 2 describes the classical methods for bandwidth selections. Section 3 presents the new method proposed for bandwidth selector, which generalizes and provides improved for the least squares cross-validation ($LSCV$). In Section 4 we present some simulation results for estimation and comparison of the various methods. Section 5 applies the methods to real data. Finally, the conclusion and perspective is presented in Section 6.

2. Classical Methods for Bandwidth Selection

Given an n-sample $X_1, X_2, ..., X_n$ of independent random variables and same unknown density f. Consider the Parzen-Rosenblatt kernel estimator of the density f given by:

$$\widehat{f_h}(x) = \frac{1}{nh} \sum_{i=1}^{n} \mathcal{K}\left(\frac{x - X_i}{h}\right) \tag{1}$$

$h > 0$ where is the bandwidth and $\mathcal{K}$. a density function defined on $\mathbb{R}$ called kernel.
To estimate f, choose the kernel $\mathcal{K}$ and h parameter. If the choice of the kernel is not a problem, it is not the case for the choice of the width of the window h which essentially depends on the size n of the sample. There are two methods of families: the family of cross-validation methods and the family of plug-in methods.
The decision of an optimal choice for the bandwidth suppose the specification of an error criterion that can be optimized. The criterion is to minimize the Mean Integrated Square Error ($MISE$). In this case, [E. Parzen [16].] there is obtained

$$MISE(\widehat{f_h}(x)) = \int \mathbb{E}\left[\widehat{f_h}(x) - f(x)\right]^2 dx = \frac{h^4}{4}\mu_2(\mathcal{K})^2 \int \left(f^{''}(x)\right)^2 dx + \frac{R(\mathcal{K})}{nh} + O\left(h^5 + \frac{1}{n}\right) \tag{2}$$

where $\mu_2(\mathcal{K}) = \int x^2 \mathcal{K}(x)dx$ is the variance of kernel $\mathcal{K}$ and $R(g) = \int g^2(x)dx$ for any function g.
The Asymptotics Mean Integrated Square Error (AMISE) is then of the form:

$$AMISE(\widehat{f_h}) = \frac{h^4}{4}\mu_2(\mathcal{K})^2 R(f^{''}) + \frac{R(\mathcal{K})}{nh} \tag{3}$$

To find the closed form expression for h_{AMISE}, begin by differentiating (3) to obtain

$$\frac{\partial AMISE}{\partial h} = -(nh^2)^{-1} R(\mathcal{K}) + h^3 \mu_2(\mathcal{K})^2 R(f^{''})$$

Setting this equation equal to 0 and solving for h produces

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

It is found that the optimal width of h_{AMISE} window depends on the unknown density f through the parameter $R(f^{''})$, which has to be estimated before using h_{AMISE}.

A very natural way to get around the problem of not knowing f'' is to use a standard family of distributions to assign a value of the term $R(f'')$ in expression (4). For example, assume that a density f belongs to the Gaussian family with mean μ and variance σ, then

$$\begin{aligned} R(f'') = \int \left(f''(x)\right)^2 dx &= \sigma^{-5} \int \left(\phi''(x)\right)^2 dx \\ &= \tfrac{3}{8}\pi^{-1} 2\sigma^{-5} \approx 0.212\sigma^{-5} \end{aligned} \tag{5}$$

where $\phi(x)$ is the standard normal density. If one uses a Gaussian kernel, then

$$\begin{aligned} h_{NR} &= (4\pi)^{-1/10} \left(\tfrac{3}{8}\pi^{-1/2}\right)^{-1/5} \sigma n^{-1/5} \\ &= \left(\tfrac{4}{3}\right)^{1/5} \sigma n^{-1/5} \end{aligned} \tag{6}$$

If we want to make this estimate more insensitive to outliers, we have to use a more robust estimate for the scale parameter of the distribution. Let $\widehat{R}$ be the sample interquartile, then one possible choice for h is

$$\begin{aligned} h_{NR} &= 1.06 min \left(\widehat{\sigma}, \tfrac{\widehat{R}}{(\Phi(3/4) - \Phi(1/4))}\right) n^{-1/5} \\ &= 1.06 min \left(\widehat{\sigma}, \tfrac{\widehat{R}}{1.349}\right) n^{-1/5} \end{aligned} \tag{7}$$

where Φ is the standard normal distribution function. To see more detail (e.g., Silverman, [21]; Härdle[8]; Scott, 1992).
The $LSCV$, sometimes called an unbiased cross-validation was proposed by Rudemo [17] and Bowman [2]. The criterion is to choose the bandwidth that minimizes an estimator of Integrated Square Error (ISE):

$$ISE = \int \widehat{f_h^2}(x)dx - 2\int \widehat{f_h}(x) f(x) dx + \int f^2(x) dx \tag{8}$$

The ideal choice of bandwidth is the one which minimizes:

$$L(h) = ISE - \int f^2(x)dx = \int \widehat{f_h^2}(x)dx - 2\int \widehat{f_h}(x) f(x) dx \tag{9}$$

The principle of the least squares cross-validation method is to find an estimate of $L(h)$ from the data and minimize it over h. Consider the estimator

$$LSCV(h) = \int \widehat{f_h^2}(x)dx - \frac{2}{n}\sum_{i=1}^{n} \widehat{f_{h(i)}}(X_i) \tag{10}$$

with

$$\int \widehat{f_h^2}(x)dx = \frac{1}{n^2 h}\sum_{i=1}^{n}\sum_{j=1}^{n} (k * k)\left(\frac{X_i - X_j}{h}\right)$$

and

$$\widehat{f_{h(i)}}(X_i) = \frac{1}{h(n-1)}\sum_{j \neq i}^{n} \mathcal{K}\left(\frac{X_i - X_j}{h}\right) \tag{11}$$

where $*$ represents the convolution.
Further discussion on this method can be found in Bowman [3], and Hall and Marron [7]. Under mild conditions, Hall [5] and Stone [22] proved that h_{LSCV} is asymptotically the best in the sense of minimizing $MISE(\widehat{f_h})$.

Sheather and Jones (1991) introduced a reliable bandwidth selector $\widehat{h}_{SJ}$, which is a plug-in estimator of h_{AMISE}, the idea of Sheather and Jones is to estimate the quantity $R(f^{''})$ by an estimator of $\mathbb{E}\left(f^{(4)}(X)\right)$, by remarking that $R(f^{''}) = \mathbb{E}\left(f^{(4)}(X)\right) = \int f^{(4)}(x)f(x)dx$.

Jin Zhang [26] introduced a generalization of classical least squares cross-validation($LSCV$), his method provides a significant improvement for ($LSCV$).
Used it as the case that $\mathcal{K}$ is the Gaussian kernel ϕ. According to Equation 10,

$$LSCV(h) = \frac{\phi_{\sqrt{2}h}(0)}{n} - \frac{2}{n(n-1)} \sum_{i<j} \left[2\phi_h(X_i - X_j) + \left(\frac{1}{n} - 1\right) \phi_{\sqrt{2}h}(X_i - X_j) \right] \tag{12}$$

Jin Zhang [26] is generalized $LSCV$ by:

$$LSCV_g(h) = \frac{\phi_{\sqrt{2}h}(0)}{n} + \frac{2}{n(n-1)} \sum_{i<j} \left[\frac{2}{g(g-2)} \phi_{\sqrt{g}h}(X_i - X_j) + \left(\frac{1}{n} - \frac{1}{g-2}\right) \phi_{\sqrt{2}h}(X_i - X_j) \right] \tag{13}$$

g with a positive number.
The generalized $LSCV$ bandwidth selector h_{LSCV_g} is defined as the minimize of $LSCV_g(h)$ over h

3. β-Divergence for Bandwidth Selection

The basic Beta-divergence was introduced by Basu et al. [1] and Minami and Eguchi [14].
The β-Divergence measure for bandwidth selection will be introduced in this section to improve the behavior of the choice for bandwidth.

$$\mathcal{D}_\beta(\widehat{f_h}, f(x)) = \frac{1}{\beta} \int_S \widehat{f}_h^\beta(x)dx - \frac{1}{\beta - 1} \int_S \widehat{f}_h^{\beta-1}(x) f(x)dx + \frac{1}{\beta(\beta - 1)} \int_S f^\beta(x)dx$$

in the case $\beta = 2$,

$$2\mathcal{D}_2(\widehat{f_h}, f(x)) = ISE(\widehat{f_h}) = \int_S (\widehat{f_h}(x) - f(x))^2 dx$$

So we can say that ISE is a special case of $\mathcal{D}_\beta$.
Note, the optimal bandwidth that minimizes $\mathcal{D}_\beta(\widehat{f_{n,h}}, f(x))$ is equivalent to the bandwidth that minimizes the expected value of the quantity:

$$\mathcal{D}_\beta(h) = \mathcal{D}_\beta(\widehat{f_h}, f(x)) - \frac{1}{\beta(\beta - 1)} \int_S f^\beta(x)dx = \frac{1}{\beta} \int_S \widehat{f}_h^\beta(x)dx - \frac{1}{\beta - 1} \int_S \widehat{f}_h^{\beta-1}(x) f(x)dx$$

The principle of the least squares cross-validation method is to find an estimate of $\mathcal{D}_\beta(h)$ from the data and minimize it over h. Consider the estimator,

$$\mathcal{D}_\beta CV(h) = \frac{1}{\beta} \int_S \widehat{f}_h^\beta(x)dx - \frac{1}{n(\beta - 1)} \sum_{i=1}^n \widehat{f}_{h(i)}^{\beta-1}(X_i) \tag{14}$$

with $\widehat{f}_{h(i)}^{\beta-1}(X_i)$ is defined in (11)
When we want to implement this technique on the computer, the computation of minimized $\mathcal{D}_\beta CV(h)$ for a of bandwidths h may be based on the following algorithm:

Algorithm 1 algorithm for minimize $\mathcal{D}_\beta CV(h)$

for $i = 1$ to n **do**

  $c_1 = \frac{1}{nh^{\beta-1}}$

  $c_2 = \frac{1}{\beta h n^{\beta-1}}$

  $c_3 = \frac{1}{(\beta-1)(n-1)^{\beta-1}}$

  $Sum = 0$

  $Sum1 = 0$

  for $j = 1$ to n **do**

    $Sum1 = Sum1 + \mathcal{K}\left(\frac{X_i - X_j}{h}\right)$

   $Sum = c_2 * (Sum1)^\beta - c_3 * (Sum1 - \mathcal{K}(0))^{\beta-1}$

  end for $\mathcal{D}_\beta CV(h) = c_1 * Sum$

end for

  $h_{\mathcal{D}_\beta CV} = arg \min_h \left(\mathcal{D}_\beta CV(h)\right)$

Theorem 1. *Let the following conditions on f be satisfied:*

(F1) f is compactly supported on I.

(F2) f is four times continuously differentiable on I.

(F3) $\lim_{x \longrightarrow + \inf I} f^{(i)}(x) = \lim_{x \longrightarrow - \sup I} f^{(i)}(x)\ ,\ 1 \leq j \leq 3.$

(F4) $\int_I f^{(2)}(x)^2 f(x)^{\beta-2} dx < \infty.$

As $n \longrightarrow \infty$, the window width $h_{\mathbb{E}D_\beta}$ that minimizes the mean β-divergence between a kernel estimator $\widehat{f_h}$ and density f satisfies

$$h_\beta = h_{\mathbb{E}\mathcal{D}_\beta} = \left\{ \frac{\int \mathcal{K}(t)^2 dt \int_I f(x)^{\beta-1} dx}{\left[\int t^2 \mathcal{K}(t) dt\right]^2 \int_I f(x)^{\beta-2} f^{(2)}(x)^2 dx} \right\}^{1/5} n^{-1/5} \tag{15}$$

in the particular case

- $\beta = 2$ *this case the Mean, integrated square error*

$$h_2 = h_{MISE(\widehat{f_h})} = \left\{ \frac{\int \mathcal{K}(t)^2 dt}{\left[\int t^2 \mathcal{K}(t) dt\right]^2 \int_I f^{(2)}(x)^2 dx} \right\}^{1/5} n^{-1/5}.$$

- $\beta = 1$ *this case the Kullback-Libler,*

$$h_1 = h_{\mathbb{E}(KL)} = \left\{ \frac{\int \mathcal{K}(t)^2 dt \int_I dx}{\left[\int t^2 \mathcal{K}(t) dt\right]^2 \int_I f(x)^{-1} f^{(2)}(x)^2 dx} \right\}^{1/5} n^{-1/5}$$

Theorem1 is derived from the following proposition by assuming $(F4)$ and by balancing the first two terms in (16).

Proposition 1. *Under* $(F1) - (F3)$ *we have*

$$\mathbb{E}\mathcal{D}_\beta(\widehat{f_h}, f) = \frac{h^4}{8} \left\{ \int_I t^2 \mathcal{K}(t) dt \right\}^2 \int f(x)^{\beta-2} f^{(2)}(x)^2 dx + \frac{1}{2nh} \int_I (\mathcal{K}(t))^2\, dt \int f(x)^{\beta-1} dx + O(n^{-1} + h^6) \tag{16}$$

Choosing β: the β value that minimizes equation (16)

$$\frac{\partial \mathbb{E}\mathcal{D}_\beta(\widehat{f_h}, f)}{\partial\beta} = (\beta-2)\frac{h^4}{8}\left\{\int_I t^2\mathcal{K}(t)dt\right\}^2\int f(x)^{\beta-3}f^{(2)}(x)^2dx + \frac{\beta-1}{2nh}\int_I(\mathcal{K}(t))^2\,dt\int f(x)^{\beta-2}dx + O(n^{-1}+h^6)$$

we pose $\quad \Delta_1 = \frac{h^4}{8}\left\{\int_I t^2\mathcal{K}(t)dt\right\}^2\int f(x)^{\beta-3}f^{(2)}(x)^2dx \quad$ and $\quad \Delta_2 = \frac{1}{2nh}\int_I(\mathcal{K}(t))^2\,dt\int f(x)^{\beta-2}dx$

$$\frac{\partial \mathbb{E}\mathcal{D}_\beta(\widehat{f_h}, f)}{\partial\beta} = (\beta-2)\Delta_1 + (\beta-1)\Delta_2 = 0$$

$$\beta = 1 + \frac{\Delta_1}{\Delta_1+\Delta_2}$$

we know that $\Delta_1 \geq 0 \quad$ and $\quad \Delta_2 \geq 0 \quad$ This implies that:

$$1 < \beta < 2$$

4. Proof

Proof. **Proposition 1**

With a random variable $\xi = O_p(1)$ whose expectation is 0 and variance 1, we can write $\widehat{f_h}(x)$ as (see [11])

$$\widehat{f_h}(x) = f(x)\left[1 + \frac{h^2}{2}\frac{f^{(2)}(x)}{f(x)}\int_I t^2\mathcal{K}(t)dt + \frac{h^4}{24}\frac{f^{(4)}(x)}{f(x)}\int_I t^4\mathcal{K}(t)dt + O(h^6) + \left\{\frac{\int_I\mathcal{K}(t)^2dt}{nhf(x)}\right\}^{1/2}\xi + O_p(n^{-1/2})\right], \quad (17)$$

Where the $O(h^6)$ terms depend upon x. Using $(1+z)^\beta = 1 + \beta z + \frac{\beta(\beta-1)}{2}z^2 + O(z^3)$

$$\begin{aligned}\widehat{f_h^\beta}(x) = \quad & f^\beta(x)[1 + \beta\left(\tfrac{1}{2}h^2\tfrac{f^{(2)}(x)}{f(x)}\int_I t^2\mathcal{K}(t)dt + \tfrac{1}{24}h^4\tfrac{f^{(4)}(x)}{f(x)}\int_I t^4\mathcal{K}(t)dt + \left\{\tfrac{\int_I\mathcal{K}(t)^2dt}{nhf(x)}\right\}^{1/2}\xi\right)\\ & +\tfrac{\beta(\beta-1)}{2}\left[\tfrac{1}{4}h^4\tfrac{f^{(2)}(x)^2}{f(x)^2}\left(\int_I t^2\mathcal{K}(t)dt\right)^2 + \tfrac{\int_I\mathcal{K}(t)^2dt}{nhf(x)}\xi^2\right] + O(h^6) + O_p(n^{-1/2})]\end{aligned} \quad (18)$$

$$\begin{aligned}\widehat{f_h^{\beta-1}}(x) = \quad & f^{\beta-1}(x)[1 + (\beta-1)\left(\tfrac{1}{2}h^2\tfrac{f^{(2)}(x)}{f(x)}\int_I t^2\mathcal{K}(t)dt + \tfrac{1}{24}h^4\tfrac{f^{(4)}(x)}{f(x)}\int_I t^4\mathcal{K}(t)dt + \left\{\tfrac{\int_I\mathcal{K}(t)^2dt}{nhf(x)}\right\}^{1/2}\xi\right)\\ & +\tfrac{(\beta-1)(\beta-2)}{2}\left[\tfrac{1}{4}h^4\tfrac{f^{(2)}(x)^2}{f(x)^2}\left(\int_I t^2\mathcal{K}(t)dt\right)^2 + \tfrac{\int_I\mathcal{K}(t)^2dt}{nhf(x)}\xi^2\right] + O(h^6) + O_p(n^{-1/2})]\end{aligned} \quad (19)$$

$$\begin{aligned}\mathcal{D}_\beta(\widehat{f_h}, f) \quad &= \int f(x)^\beta\left[\tfrac{1}{\beta} - \tfrac{1}{\beta-1} + \left(\tfrac{\beta-1}{2} - \tfrac{\beta-2}{2}\right)\left[\tfrac{1}{4}h^4\tfrac{f^{(2)}(x)^2}{f(x)^2}\left(\int_I t^2\mathcal{K}(t)dt\right)^2 + \tfrac{\int_I\mathcal{K}(t)^2dt}{nhf(x)}\xi^2\right] + O(h^2) + O_p(n^{-1/2})\right]dx\\ &\quad + \tfrac{1}{\beta(\beta-1)}\int f(x)^b dx \quad (20)\\ &= \tfrac{1}{2}\int f(x)^\beta\left[\tfrac{1}{4}h^4\tfrac{f^{(2)}(x)^2}{f(x)^2}\left(\int_I t^2\mathcal{K}(t)dt\right)^2 + \tfrac{\int_I\mathcal{K}(t)^2dt}{nhf(x)}\xi^2 + O(h^2) + O_p(n^{-1/2})\right]dx \quad (21)\end{aligned}$$

$$\begin{aligned}\mathbb{E}\mathcal{D}_\beta(\widehat{f_h}, f) \quad &= \tfrac{1}{2}\mathbb{E}\int f(x)^\beta\left[\tfrac{1}{4}h^4\tfrac{f^{(2)}(x)^2}{f(x)^2}\left(\int_I t^2\mathcal{K}(t)dt\right)^2 + \tfrac{\int_I\mathcal{K}(t)^2dt}{nhf(x)}\xi^2 + O(h^2) + O_p(n^{-1/2})\right]dx\\ &= \tfrac{h^4}{8}\left\{\int_I t^2\mathcal{K}(t)dt\right\}^2\int f(x)^{\beta-2}f^{(2)}(x)^2dx + \tfrac{1}{2nh}\int_I(\mathcal{K}(t))^2\,dt\int f(x)^{\beta-1}\mathbb{E}(\xi^2)dx + O(n^{-1}+h^6)\end{aligned}$$

we know that $\mathbb{E}(\xi^2) = 1$

$$\mathbb{E}\mathcal{D}_\beta(\widehat{f_h}, f) = \frac{h^4}{8}\left\{\int_I t^2\mathcal{K}(t)dt\right\}^2\int f(x)^{\beta-2}f^{(2)}(x)^2dx + \frac{1}{2nh}\int_I(\mathcal{K}(t))^2\,dt\int f(x)^{\beta-1}dx + O(n^{-1}+h^6) \quad (22)$$

as required.

5. Simulation

We approximate the true density f by a normal mixture.

$$f(x) = \sum_{j=1}^{J} \omega_j \phi_{\sigma_j}(x - \mu_j) \tag{23}$$

where J is a positive integer. $\omega_1, ..., \omega_J$ is a set of positive numbers that sum to one, and for each j, $-\infty < \mu_j < \infty$ and $\sigma_j > 0$. The family of normal mixture densities used by (Marron and Wand [13]) is extremely rich, and, in fact, any density can be approximated arbitrarily well by a member of this family.
for f given by Equation 23, the $MISE$ in Equation (2) of the kernel density estimator in Equation (1) have explicit forms. In fact,

$$MISE(\widehat{f_h}(x)) = (2\pi^{1/2}nh)^{-1} + w^T\{(1-n^{-1})\Omega_2 - 2\Omega_1 + \Omega_0\}w \tag{24}$$

(Marron and Wand [13]) where $w = (\omega_1, ..., \omega_J)$ and Ω_a is the $J \times J$ matrix having $(j \times j)$ entry equal to

$$\phi_{\sqrt{ah^2+\sigma_j^2+\sigma_{j'}^2}}(\mu_j - \mu_{j'})$$

.
We consider the normal mixture in the case of $J = 2$ and $\omega_1 = \omega_2 = 0.5$, similar similation were performed by Jin Zhang [26]. Therefore the true density is:

$$f(x) = 0.5\phi(x) + 0.5\phi_\sigma(x - \mu) \tag{25}$$

Based on $50, 200, 700$ draws from f in the case where $\mu = 0, 1, 5$ and $\sigma = 1, 0.5, 0.1$. **Table 1** give the exhibits the simulated relative efficiency $RE\left(\widehat{h}\right) = MISE\left(\widehat{f}_{h_{MISE}}\right)/MISE\left(\widehat{f_h}\right)$ of the kernel estimator, using bandwidths $\widehat{h}_{NR}$, $\widehat{h}_{LSCV}$, $\widehat{h}_{SJ}$, $\widehat{h}_{LSCV_4}$ and $\widehat{h}_{\mathcal{D}_\beta CV}$(with $\beta = 1.1, 1.5, 4$ and 4), it is lower wherever than 1, because the optimal bandwidth h_{MISE} minimize $MISE$. Each bandwidth, mean $E\left(\widehat{h}\right)$ and mean relation error $E|\widehat{h}/h_{MISE} - 1|$ are obtained, these values are given by respectively **Tables 2** and **3**.

1. It can be seen that when the density f is not very far from normal, such as that cases of $(\mu, \sigma) = (0, 1), (0, 0.5), (1, 1)$ and $(1, 0.5)$, bandwidth which are obtained by using NR criterion performs well. In other cases it usually has the smallest $RE(\widehat{h})$ and largest $E(\widehat{h})$,tending to over smooth its kernel density estimate the most.

2. From these tables it can be seen that bandwidth $\widehat{h}_{LSCV}$ which are obtained by using $LSCV$ performs poorly if the true density is close to normal usually having the smallest $RE(\widehat{h})$ and $E(\widehat{h})$, then the kernel density estimator is undersmoothed. But in the contrary case, $\widehat{h}_{LSCV}$ can perform very well, and in many situations, in **Table 2** it is seen that $E(\widehat{h}_{LSCV})$ is close to the optimal $\widehat{h}_{MISE}$, but the corresponding $E(\widehat{h}_{LSCV}/\widehat{h}_{MISE} - 1)$ is large, which means that the bias of $\widehat{h}_{LSCV}$ is small but its variation is large.

3. The plug-in bandwidth $\widehat{h}_{SJ}$ seen to be the best existing bandwidth selectors (as commented by Venables and Ripley [23]). $\widehat{h}_{SJ}$ among the best bandwidth selectors in most situations. However it behaves very poorly for small σ (the density curve is sharp), oversmoothing its kernel density curve by overestimating $\widehat{h}_{MISE}$.

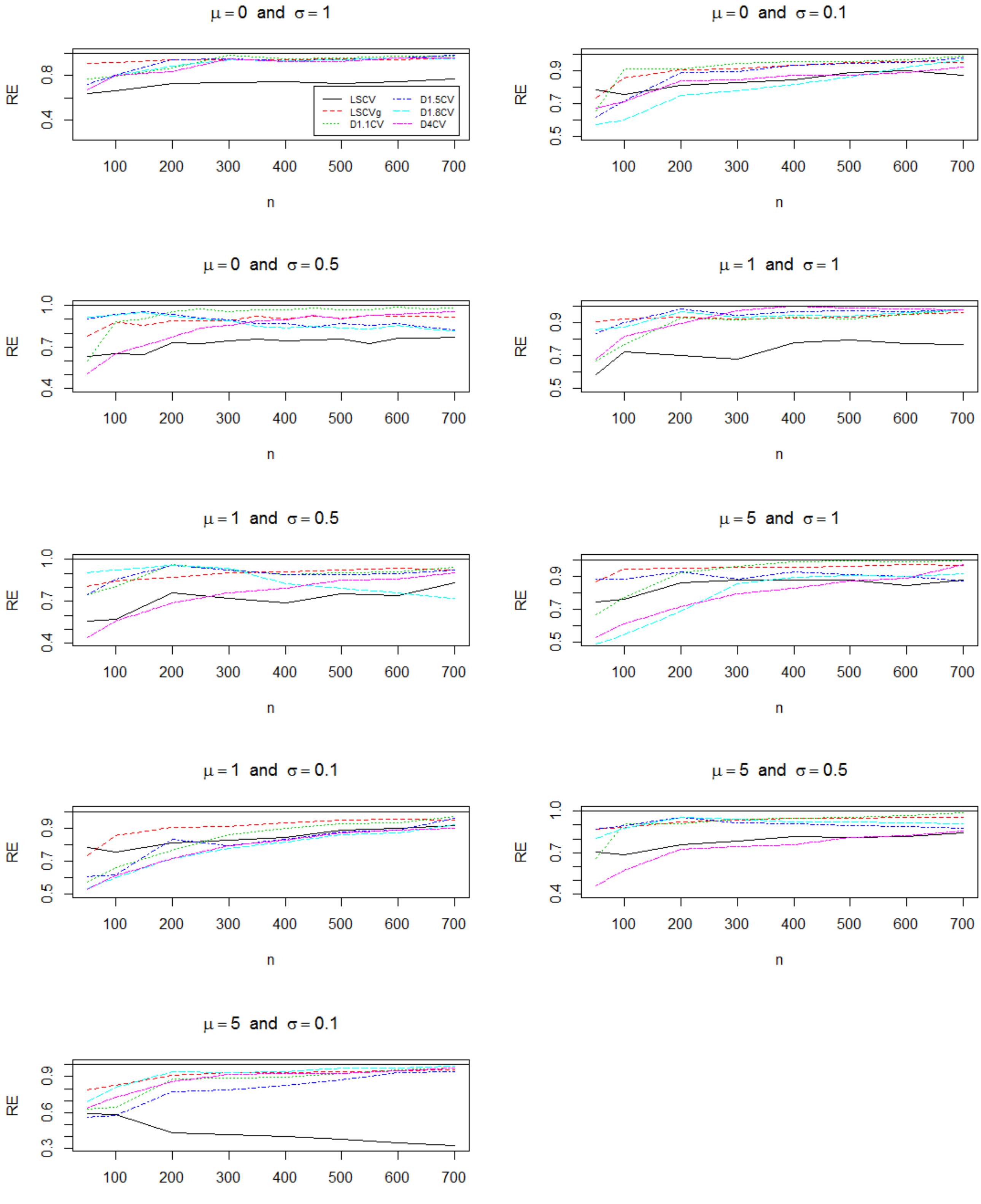
4. As commented by J. Zhang [26], The generalized $LSCV$ bandwidth $\widehat{h}_{LSCV_4}$ seems to be bandwidth selectors that is always among the best for having large $RE(\widehat{h})$ and small $E|\widehat{h}/h_{MISE} - 1|$.

5. The proposed bandwidth selector method $(\widehat{h}_{\mathcal{D}_\beta CV})$ completely dominate the selection methods Bandwidth as n increases, otherwise $\widehat{h}_{\mathcal{D}_\beta CV}$ (with $\beta = 1.1, 1.5, 1.8$ and 4) are with $\widehat{h}_{LSCV_4}$ bandwidths selectors that are always among the best. for having large $RE(\widehat{h}_{\mathcal{D}_\beta CV})$ and small $E|\widehat{h}_{\mathcal{D}_\beta CV}/h_{MISE} - 1|$ for most cases. They significantly improve the classical $\widehat{h}_{LSCV}$. Indeed Figue1 shows that the increase in n causes a value of $RE(\widehat{h}_{D_\beta CV})$ close to 1

and greater than $RE(\widehat{h}_{LSCV})$

Figure 2 show that for the $\mathcal{D}_\beta CV$, this bias does not have a serious effect on the efficiency of the method, since the $RE(\widehat{h}_{\mathcal{D}_\beta CV})$ is relatively close to 1. We can conclude from Figure 2 our practical selection procedures $\mathcal{D}_\beta CV$ have a performance close to the one for the $LSCV_4$ bandwidth selector (since the ratios are relatively close to 1).

Figure 1: $RE(\widehat{h})$ using bandwidths $\widehat{h}_{LSCV}$, $\widehat{h}_{LSCV_4}$ and $\widehat{h}_{D_\beta CV}$ with $\beta = 1.1, 1.5, 1.8, 4$



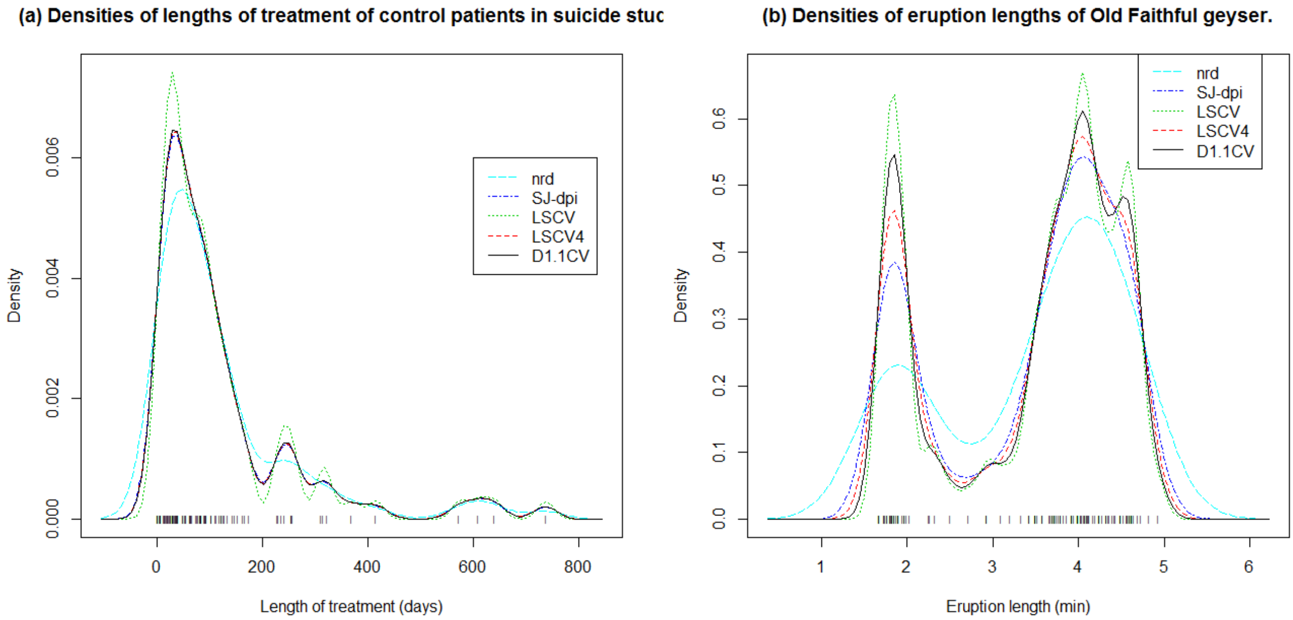


Figure 2: Kernel density estimates with different bandwidths. NR, SJ, $LSCV$, $LSCV_4$, $\mathcal{D}_{1.1}CV$

6. Examples

In this Section, we will provided two examples to evaluate performance of our method compared to several classical bandwidth selection methods for Gaussian kernel density. The two data sets in the examples have been analyzed by many authors (Silverman [21] and J. Zhang [26]) to illustrate various kinds of methods in density estimation.

The first example comprises the lengths of 86 spells of psychiatric treatment undergone by control patients used as controls in a study of suicide risks reported by Copas and Fryer [4]; Silverman [21].
Figure 2a plot the data points and the kernel density estimates for the suicide study data, when we using commonly used bandwidths $\widehat{h}_{NR} = 35.78$, $\widehat{h}_{SJ} = 23.16$, $\widehat{h}_{LSCV} = 15.69$, $\widehat{h}_{LSCV_4} = 22.57$ and $h_{\mathcal{D}_{1.1}CV} = 24.25$.
It seems that the density for the length of treatment is a unimodal curve heavily skewed to the right. it seems clear than the bandwidth h_{NR} oversmooth her kernel density curve and underestimate the peak near 20.
The bandwidths were obtained by using the methods of SJ, $LSCV_4$ and $\mathcal{D}_{1.1}CV$, are better because $\widehat{h}_{SJ}$, $\widehat{h}_{LSCV_4}$ and $\widehat{h}_{\mathcal{D}_{1.1}CV}$ well balance the two situations and seem to capture the true shape of the data. In this example, $\widehat{h}_{LSCV_4} = 27.57$ and $\widehat{h}_{\mathcal{D}_{1.1}CV} = 24.25$ are closer to $\widehat{h}_{SJ} = 23.16$.

The second comprises the lengths in minutes of 107 eruption lengths in minutes for the Old Faithful geyser in Yellowstone National Park, USA (source: Weisberg [25]; Silverman [21]).
Figure 2b plot the data points and the kernel density estimates for Old Faithful geyser data, we using bandwidths $\widehat{h}_{NR} = 0.4331$, $\widehat{h}_{SJ} = 0.2250$, $\widehat{h}_{LSCV} = 0.1030$, $\widehat{h}_{LSCV_4} = 0.1740$ and $\widehat{h}_{\mathcal{D}_{1.1}CV} = 0.1480$.
An important point to note that the density curve for eruption length is similar to bimodal normal density (normal mixture). As commented by Zing, the bandwidthS $\widehat{h}_{NR}$ is heavily over smoothes its kernel density curve, underestimating the two peaks of the curve but overestimating the valley between them. On the other hand, $\widehat{h}_{LSCV}$ seems to undersmooth the curve too much, overestimating the two peaks but underestimating for the valley.
However $\widehat{h}_{SJ}$, $\widehat{h}_{LSCV_4}$ and $\widehat{h}_{\mathcal{D}_{1.1}CV}$, especially the later, are proper bandwidths for their density estimates to be able to capture the feature of the true density curve.

7. Conclusions

The kernel estimator is the most used in density estimation. The main issue is bandwidth selection, which is a hot topic and is still frustrating statisticians. A various bandwidth selection strategies have been proposed such as normal reference $\widehat{h}_{NR}$, unbiased cross-validation $\widehat{h}_{LSCV}$, cross validation, Sheather-Jones method $\widehat{h}_{SJ}$ and and most recently Zing proposed a bandwidth $\widehat{h}_{LSCV_g}$.

The normal reference bandwidth $\widehat{h}_{NR}$ method is limited the practical use, since they are restricted to situations where a pre-specified family of densities is correctly selected. The popularity of the $LSCV$ method is due to the intuitive motivation and the fact that $\widehat{h}_{LSCV}$ is asymptotically optimal under low conditions. Bandwidth $\widehat{h}_{SJ}$ seems to be the best existing bandwidth selector. Unfortunately, it also tends to oversmooth the density estimate when the true density curve is sharp.

We have attempted to evaluate choice the optimal bandwidth $\widehat{h}_{LSCV}$, using β-divergence. Compared to traditional bandwidth selection methods designed for kernel density estimation, our proposed $\mathcal{D}_\beta CV$ bandwidth selection method is always one of the best for having large $RE\left(\widehat{h}\right)$ and small $E|\widehat{h}/h_{MISE}-1|$. Simulation studies showed that our proposed novel optimal bandwidth method designed for kernel density estimation significantly improves the classical $\widehat{h}_{LSCV}$ for its variability and undersmoothing, adapts to different situations, and outperforms other bandwidths.

References

[1] Basu, A.; Harris, I.R.; Hjort, N.; Jones, M. (1998), Robust and efficient estimation by minimising a density power divergence. Biometrika , 85, 549–559.

[2] Bowman, A. W. (1984), An alternative method of cross-validation for the smoothing of density estimates, Biometrika 71, 353–360.

[3] Bowman, A. W. (1985), A comparative study of some kernel-based nonparametric density estimators, Journal of Statistical Com- putation and Simulation, **21**, 3-4, 313–327.

[4] Copas, J. P. and M. J. Fryer (1980), Density estimation and suicide risks in psychiatric treatment, Journal of the Royal Statistical Society, **Series B** 143, 167–176.

[5] Hall, P. (1983). Large sampie optimality of least squares cross-validation in density estirnation, Ann. Statist., **11**, 1156–1174.

[6] Hall, P. (1985). Asymptotic theory of minimum integrated square error for multivariate density estimation. In Multivariate Analysis - **VI**, ed. P.R. Krishnaiah, Elsevier Science, Amsterdam, 289–309.

[7] Hall, P. and J. S. Marron (1991), Local minima in cross-validation functions, Journal of the Royal Statistical Society, **Series B 53**, 245–252.

[8] Hrdle, W. (1991), Smoothing techniques with implementation in S, Springer-Verlag, New York.

[9] Jones,M. C., Marron, J. S., and Sheather, S. J. (1996), A brief survey of bandwidth selection for density estimation, Journal of the American Statistical Association, **91**, 433, 401–407,.

[10] Jones, M. C. and R. F. Kappenman (1991), On a class of kernel density estimate bandwidth selectors, Scandinavian Journal of Statistics 19, 337?349.

[11] Kanazawa, Y. (1993), Hellinger distance and Kullback-Leibler loss for the kernel density estimator. Statistics and Probability Letters **18** 315–321

[12] Loader, C. R. (1999), ?Bandwidth selection: classical or plug-in?? The Annals of Statistics, **27**, 2, 415–438, .

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

[14] Minami, M.; Eguchi, S. Robust blind source separation by Beta-divergence. Neural Comput. 2002, **13**, 1859–1886.

[15] Park, B, U., Turlach, B, S. (1992) Practical performance of several data driven bandwidth selectors, Computational Statistics,**7**, 251–270

[16] Parzen, E. (1962), On estimation of a probability density function and mode, Annals of Mathematical Statistics **33**, 1065–1076.

[17] Rudemo, M. (1982), Empirical choice of histograms and kernel density estimators, Scandinavia Journal of Statistics **9**, 65–78.

[18] Scott, W. D (1992), Multivariate density estimation theory, practice, and visualization, Wiley, New York.

[19] Scott, D. W. and Terrell, G. R. (1987), Biased and unbiased cross-validation in density estimation, Journal of the American Statistical Association **82**, 1131–1146.

[20] Sheather, S. J. and Jones, M. C. (1991), A reliable data-based bandwidth selection method for kernel density estimation, Journal of the Royal Statistical Society series B 53, 683–690.

[21] Silverman, B. W. (1986), Density estimation for statistics and data analysis, Chapman and Hall, London.

[22] Stone, C. J. (1984), An asymptotically optimal window selection rule for kernel density esti- mates, Annals of Statistics, **12**, 1285–1297.

[23] Venables, W. N. and B. D. Ripley (2002), Modern applied statistics with S 4th Edition, Springer, New York.

[24] Wand, M. P. and M. C. Jones (1995), Kernel Smoothing, Chapman and Hall, London, UK.

[25] Weisberg, S. (1980), Applied linear regression, Wiley, New York.

[26] Zhang, J. (2015), Generalized least squares cross-validation in kernel density estimation,Statistica Neerlandica, 69, **3**, 315–328.

Table 1: $RE\left(\widehat{h}\right)$ for normal mixture $f(x) = 0.5\phi(x) + 0.5\phi_{\sigma}(x-\mu)$

n	$\widehat{h_{NR}}$	$\widehat{h_{LSCV}}$	$\widehat{h_{SJ}}$	$\widehat{h_{LSCV_4}}$	$\widehat{h_{\mathcal{D}_{1.1}CV}}$	$\widehat{h_{\mathcal{D}_{1.5}CV}}$	$\widehat{h_{\mathcal{D}_{1.8}CV}}$	$\widehat{h_{\mathcal{D}_4CV}}$
			$\mu = 0$	$\sigma = 1$				
50	0.877	0.633	0.798	0.909	0.768	0.716	0.694	0.770
200	0.928	0.730	0.867	0.944	0.978	0.944	0.924	0.833
700	0.974	0.767	0.942	0.956	0.978	0.981	0.973	0.959
			$\mu = 0$	$\sigma = 0.5$				
50	0899	0.633	0.850	0.776	0.600	0.900	0.917	0.502
200	0.948	0.644	0.895	0.908	0.884	0.987	0.949	0.709
700	0.952	0.767	0.939	0.914	0.980	0.821	0.816	0.967
			$\mu = 0$	$\sigma = 0.1$				
50	0.562	0.781	0.779	0.748	0.700	0.618	0.573	0.670
200	0.550	0.810	0.902	0.891	0.930	0.890	0.747	0.834
700	0.535	0.872	0.966	0.957	0.993	0.977	0.963	0.920
			$\mu = 1$	$\sigma = 1$				
50	0.871	0.585	0.791	0.904	0.665	0.830	0.855	0.680
200	0.955	0.699	0.905	0.949	0.928	0.989	0.992	0.892
700	0.977	0.763	0.943	0.960	0.981	0.973	0.977	0.975
			$\mu = 1$	$\sigma = 0.5$				
50	0.855	0.562	0.856	0.806	0.749	0.751	0.906	0.439
200	0.867	0.760	0.929	0.874	0.965	0.975	0.989	0.690
700	0.823	0.831	0.955	0.922	0.942	0.923	0.720	0.902
			$\mu = 1$	$\sigma = .1$				
50	0.2370	0.781	0.439	0.755	0.544	0.606	0.531	0.528
200	0.1010	0.810	0.419	0.904	0.756	0.834	0.717	0.716
700	0.0503	0.912	0.557	0.949	0.968	0.980	0.921	0.899
			$\mu = 5$	$\sigma = 1$				
50	0.411	0.741	0.875	0.864	0.618	0.883	0.490	0.528
200	0.285	0.861	0.947	0.948	0.852	0.981	0.688	0.716
700	0.215	0.876	0.972	0.967	0.997	0.870	0.914	0.979
			$\mu = 5$	$\sigma = 0.5$				
50	0.2390	0.707	0.696	0.871	0.657	0.871	0.806	0.460
200	0.1330	0.760	0.747	0.923	0.907	0.986	0.983	0.726
700	0.0804	0.842	0.846	0.957	0.990	0.800	0.912	0.858
			$\mu = 5$	$\sigma = 0.1$				
50	0.1360	0.588	0.1760	0.791	0.630	0.563	0.693	0.640
200	0.0523	0.458	0.0946	0.905	0.878	0.774	0.940	0.858
700	0.0205	0.341	0.0665	0.969	0.968	0.955	0.989	0.980

Table 2: $E\left(\widehat{h}\right)$ for normal mixture $f(x) = 0.5\phi(x) + 0.5\phi_\sigma(x-\mu)$

n	$\widehat{h_{NR}}$	$\widehat{h_{LSCV}}$	$\widehat{h_{SJ}}$	$\widehat{h_{LSCV_4}}$	$\widehat{h_{\mathcal{D}_{1.1}CV}}$	$\widehat{h_{\mathcal{D}_{1.5}CV}}$	$\widehat{h_{\mathcal{D}_{1.8}CV}}$	$\widehat{h_{\mathcal{D}_4CV}}$	h_{MISE}
				$\mu = 0$	$\sigma = 1$				
50	0.455	0.480	0.452	0.516	0.323	0.330	0.379	0.360	0.520
200	0.354	0.373	0.349	0.385	0.321	0.328	0.376	0.348	0.383
700	0.283	0.288	0.282	0.301	0.308	0.309	0.307	0.308	0.293
				$\mu = 0$	$\sigma = 0.5$				
50	0.318	0.362	0.306	0.382	0.223	0.286	0.286	0.174	0.343
200	0.250	0.262	0.231	0.266	0.193	0.280	0.285	0.168	0.248
700	0.197	0.191	0.184	0.200	0.186	0.244	0.245	0.166	0.188
				$\mu = 0$	$\sigma = 0.1$				
50	0.1320	0.0897	0.0915	0.1040	0.510	0.0409	0.0356	0.0403	0.0752
200	0.0887	0.0557	0.0572	0.0597	0.485	0.0377	0.0331	0.0400	0.0530
700	0.0668	0.0406	0.0406	0.0422	0.421	0.0370	0.0329	0.0403	0.0398
				$\mu = 1$	$\sigma = 1$				
50	0.510	0.569	0.511	0.573	0.429	0.426	0.435	0.360	0.373
200	0.403	0.418	0.403	0.440	0.395	0.423	0.429	0.359	0.265
700	0.317	0.329	0.315	0.337	0.354	0.345	0.346	0.341	0.199
				$\mu = 1$	$\sigma = 0.5$				
50	0.409	0.395	0.366	0.446	0.326	0.342	0.283	0.155	0.373
200	0.328	0.278	0.270	0.296	0.280	0.282	0.283	0.152	0.265
700	0.256	0.197	0.203	0.212	0.233	0.239	0.281	0.151	0.199
				$\mu = 1$	$\sigma = .1$				
50	0.351	0.0877	0.1780	0.1030	0.0422	0.0451	0.0523	0.0413	0.0752
200	0.280	0.0553	0.1050	0.0614	0.0380	0.0380	0.0401	0.0357	0.0530
700	0.222	0.0408	0.0657	0.0426	0.0343	0.0314	0.0311	0.0309	0.0398
				$\mu = 5$	$\sigma = 1$				
50	1.300	0.636	0.753	0.742	0.420	0.475	0.315	0.308	0.608
200	0.986	0.456	0.495	0.472	0.330	0.470	0.285	0.276	0.441
700	0.770	0.342	0.361	0.261	0.275	0.336			
				$\mu = 5$	$\sigma = 0.5$				
50	1.260	0.407	0.590	0.448	0.310	0.295	0.251	0.217	0.369
200	0.963	0.274	0.378	0.296	0.210	0.286	0.250	0.171	0.262
700	0.750	0.201	0.255	0.210	0.209	0.270	0.244	0.146	0.197
				$\mu = 5$	$\sigma = 0.1$				
50	1.260	0.0871	0.523	0.1040	0.045	0.0415	0.0476	0.0497	0.0752
200	0.956	0.0560	0.292	0.0603	0.040	0.0385	0.0427	0.0395	0.0530
700	0.745	0.0406	0.172	0.0420	0.039	0.0339	0.0426	0.0387	0.0398

Table 3: $E|\widehat{h}/h_{MISE}-1|$ for normal mixture $f(x)=0.5\phi(x)+0.5\phi_{\sigma}(x-\mu)$

n	$\widehat{h_{NR}}$	$\widehat{h_{LSCV}}$	$\widehat{h_{SJ}}$	$\widehat{h_{LSCV_4}}$	$\widehat{h_{\mathcal{D}_{1.1}CV}}$	$\widehat{h_{\mathcal{D}_{1.5}CV}}$	$\widehat{h_{\mathcal{D}_{1.8}CV}}$	$\widehat{h_{\mathcal{D}_4CV}}$
			$\mu=0$	$\sigma=1$				
50	0.1350	0.1240	0.1640	0.0874	0.3790	0.3650	0.2710	0.328
200	0.0785	0.0829	0.1050	0.0578	0.1620	0.1390	0.0177	0.145
700	0.0396	0.0717	0.0572	0.0436	0.0509	0.0531	0.0486	0.052
			$\mu=0$	$\sigma=0.5$				
50	0.1370	0.1510	0.1670	0.1510	0.4560	0.179	0.166	0.342
200	0.0655	0.1360	0.0882	0.1010	0.2490	0.135	0.150	0.264
700	0.0559	0.0729	0.0537	0.0818	0.0104	0.299	0.300	0.188
			$\mu=0$	$\sigma=0.1$				
50	0.772	0.2530	0.3000	0.3990	0.4410	0.4980	0.562	0.463
200	0.674	0.1210	0.1250	0.1520	0.2070	0.2870	0.378	0.244
700	0.679	0.0772	0.0506	0.0726	0.585	0.0503	0.171	0.015
			$\mu=1$	$\sigma=1$				
50	0.1430	0.1040	0.1630	0.0748	0.398	0.2760	0.2610	0.387
200	0.0774	0.0800	0.0931	0.0501	0.184	0.0249	0.0124	0.171
700	0.0483	0.0626	0.0600	0.0361	0.037	0.0426	0.0416	0.0466
			$\mu=1$	$\sigma=0.5$				
50	0.172	0.1930	0.1400	0.2260	0.4560	0.3580	0.2420	0.584
200	0.236	0.1530	0.0899	0.1460	0.121	0.0965	0.0668	0.415
700	0.285	0.0989	0.0506	0.0794	0.169	0.2010	0.4100	0.223
			$\mu=1$	$\sigma=.1$				
50	3.67	0.2620	1.380	0.3980	0.544	0.507	0.586	0.589
200	4.29	0.1250	0.986	0.1720	0.353	0.300	0.413	0.416
700	4.58	0.0838	0.652	0.0878	0.137	0.068	0.218	0.221
			$\mu=5$	$\sigma=1$				
50	1.14	0.1450	0.2390	0.2430	0.4580	0.2840	0.570	0.546
200	1.23	0.0815	0.1210	0.0899	0.2530	0.0147	0.409	0.376
700	1.29	0.0686	0.0745	0.0540	0.0203	0.2970	0.224	0.179
			$\mu=5$	$\sigma=0.5$				
50	2.40	0.1860	0.600	0.2510	0.4340	0.2660	0.3380	0.6033
200	2.68	0.1180	0.444	0.1440	0.2020	0.0315	0.0673	0.440
700	2.80	0.0743	0.296	0.0804	0.0597	0.3800	0.2390	0.237
			$\mu=5$	$\sigma=0.1$				
50	15.7	0.884	5.95	0.3980	0.4810	0.549	0.433	0.484
200	17.1	1.021	4.51	0.1570	0.2630	0.359	0.194	0.267
700	17.7	1.104	3.34	0.0656	0.0178	0.146	0.074	0.0573