# On Bias Reduction in Kernel Density Estimation

Choongrak Kim[1], Byeong-Uk Park, Woochul Kim[2]

## Abstract

Kernel estimator is very popular in nonparametric density estimation. In this paper we propose an estimator which reduces the bias to the fourth power of the bandwidth, while the variance of the estimator increases only by at most moderate constant factor. The estimator is fully nonparametric in the sense of convex combination of three kernel estimators, and has good numerical properties.

## 1. Introduction

Nonparametric methods have received a lot of attention in density estimation, and the kernel density estimation is very popular among them. Best references in this area are Silverman(1986) and Wand and Jones(1995).

To reduce the bias in kernel density estimation, higher-order kernels are usually used. Recently Cheng, Choi, Fan and Hall(2000) suggested a bias reduction technuque by using a skewing method which is suggested and discussed by Rice(1984) and Choi and Hall(1998) in the regression context. In fact, Cheng, et.al(2000) used a locally parametric method discussed by Copas(1995), Hjort and Jones(1986) and Loader(1996), and they argued that the resulting estimator reduces the bias to the fourth power of the bandwidth, while variance of the estimator increases only by at most a moderate constant factor.

In this paper, we suggest an estimator which has same asymptotic properties, in the sense of bias and variance as the estimator suggested by Cheng, et.al(2000). As in the estimator by Cheng, et.al(2000), we also used a skewing method, hoewver, instead of the locally parametric estimation, we used a fully nonparametric method, i.e., a kernel estimator. One disadvantage of our estimator is that the nonnegativity is not guaranteed, but we

[1]Department of Statistics, Pusan National University, Pusan, 609-735
[2]Department of Statistics, Seoul National University, Seoul, 151-742

suggest two versions correcting for nonnegativity. In Section 2, the proposed estimator is introduced based on a motivational example, and its asymptotic properties are derived. Also two versions correcting for nonnegativity are discussed. Numerical properties of the proposed estimator will be illustrated in Section 3.

## 2. The Proposed Estimator

### 2.1 The Estimator

Let $X_1, \cdots, X_n$ be random sample from a distribution with an unknuwn density $f(\cdot)$,which we wish to estimate. The kernel estimator of $f$ at $x$ is

$$\hat{f}(x) = \frac{1}{nh}\sum_i K(\frac{x - X_i}{h}), \tag{1}$$

where $h$ is the bandwidth, and $K$ is the kernel function.

One typical feature of the nonparametric estimators including the kernel estimator underestimate at peaks and overestimate at troughs. This phenomenon is well illustrated in Figure 1 which shows the kernel density estimator $\hat{f}$ based on $n$=100 random rample from N(0,1) with the Gaussian kernel and the optimal bandwidth $h$=0.422.

Motivated by this phenomenon, we suggest as an estimator at $x$

$$\tilde{f}(x) = \frac{\lambda_1 \hat{f}_1(x) + \hat{f}(x) + \lambda_2 \hat{f}_2(x)}{\lambda_1 + 1 + \lambda_2}, \tag{2}$$

where $\lambda_1, \lambda_2 > 0$ are weights, $l_1 < 0, l_2 > 0$ are contants to be determined,

$$\hat{f}_j(x) = \hat{f}(x + l_j h) - l_j h \hat{f}'(x + l_j h), j = 1, 2, \tag{3}$$

and

$$\hat{f}'(x) = \frac{1}{nh^2}\sum K'(\frac{x - X_i}{h})$$

is the kernel estimator of $f'$, the first derivative of $f$, i.e., the suggested estimator $\tilde{f}(x)$ is a convex combination of $\hat{f}_1(x), \hat{f}(x)$ and $\hat{f}_2(x)$. The estimator $\hat{f}_j(x), j = 1, 2$, represent values of the tangent line evaluated at $x + l_j h$. See Figure 2 for clarity.

Therefore, $\tilde{f}(x)$ will be larger than $\hat{f}(x)$ where the point of interest $x$ is located at peak area. Similarly, $\tilde{f}(x)$ will be smaller than $\hat{f}(x)$ where $x$ is located at trough area. Therefore, we can expect that the bias of $\tilde{f}(x)$ is smaller than of $\hat{f}(x)$. In fact, by choosing $\lambda_1 = \lambda_2 = \lambda, l_1 = -l_2 = l(\lambda)$, say, and

$$l(\lambda) = \{(1+2\lambda)\mu_2/(2\lambda)\}^{1/2}, \tag{4}$$

where $\mu_l = \int u^l K(u)du$ , it can be shown that the bias of $\tilde{f}(x)$ is $O(h^4)$, while that of $\hat{f}(x)$ is $O(h^2)$. The following theorem, whose proof is in the Appendix, shows the bias of $\tilde{f}(x)$ in detail.

Theorem 2.1 (Bias of $\tilde{f}$). *Assume that $f$ has four bounded, continuous derivatives in a neighborhood of $x$; that the kernel $K$ is nonnegative, bounded, symmetric, with $\int K = 1$; and that $h \to 0$ and $nh \to \infty$. Take $\lambda_1 = \lambda_2 = \lambda > 0$ and $l_1 = -l_2 = l(\lambda)$. Then,*

$$E[\tilde{f}(x) - f(x)] = B(x)h^4 + o_p\{h^4 + (nh)^{-1/2}\},$$

*where*

$$B(x) = \frac{f^{(4)}(x)}{24}\{\mu_4 - \frac{3(1+5\lambda)}{2\lambda}{\mu_2}^2\}.$$

After tedious and lengthy algebra, although conceptually simple to derive, we obtain asymptotic variance of $\tilde{f}(x)$.

Theorem 2.2 (Variance of $\tilde{f}$). *Assume the same conditions imposed on $K$ and $h$ in Theoram 2.1, and $\lambda_1 = \lambda_2 = \lambda, l_1 = -l_2 = l(\lambda)$. Then*

$$Var[\tilde{f}(x)] = \frac{f(x)}{nh}V(\lambda) + o_p\{(nh)^{-1}\},$$

*where*

$$\begin{aligned} V(\lambda) &= (2\lambda+1)^{-2}[(2\lambda^2+1)\int K^2(t)dt + 4\lambda\int K(t)K(t+l)dt \\ &\quad +2\lambda^2\int K(t-l)K(f+l)dt \\ &\quad +\lambda(2\lambda+1)\mu_2\int\{K'(t)^2 - K(t-l)K(t+l)\}dt]. \end{aligned}$$

Remark 2.1 : The case $\lambda = \infty$. Choosing $\lambda = \infty$ in the definiton of $\tilde{f}$ we obtain $\tilde{f} = \frac{1}{2}(\hat{f}_1 + \hat{f}_2)$. We can easily show that the bias of $\tilde{f}$ with $\lambda = \infty$ is $O(h^3)$.

Remart 2.2 : The choice of $\lambda$. Choi and Hall(1998) suggested using $\lambda$ minimizing $V(\lambda)$, and their suggestion can be used in this situation, too. Another possibility is using $\lambda$ minimizing MISE, however, it is computationally difficult. The minimizer of $V(\lambda)$ varies from 0.1 to 1.0 depending on the kernel $K$. In general, the choice of $\lambda$ is not very sensitive to the estimator $\tilde{f}$.

## 2.2. Corrections for Nonnegativity

Since the estimator $\tilde{f}(x)$ in (2) is not guaranteed to be nonnegative, we consider some corrections for nonnegativity.

First, note that $\hat{f}_1(x)$ in (3) can be regarded as the approximated form of

$$\hat{f}_1^*(x) = \hat{f}(x + l_1 h) exp\{-l_1 h \hat{f}'(x + l_1 h)/\hat{f}(x + l_1 h)\} \tag{5}$$

by the Taylor expansion of $\hat{f}_1(x)$ to the linear term. Note that $\hat{f}_1^*(x)$ is very similar to the estimator $\hat{f}_\infty(x)$ suggested by Cheng, et.al (2000). In fact,

$$\hat{f}_\infty(x) = \frac{1}{2}(\hat{f}_+(x) + \hat{f}_-(x)), \tag{6}$$

where

$$\hat{f}_\pm(x) = \hat{f}(x \pm lh) exp[\frac{1}{2}l^2 - \frac{1}{2}\{l \pm h\hat{f}'(x \pm lh)/\hat{f}(x \pm lh)\}^2].$$

Therefore,

$$\tilde{f}^*(x) = \frac{\lambda_1 \hat{f}_1^*(x) + \hat{f}(x) + \lambda_2 \hat{f}^*(x)}{\lambda_1 + 1 + \lambda_2}$$

is always nonnegative, and we can find $l_1, l_2, \lambda_1, \lambda_2$ such that the bias of $\tilde{f}(x)$ is $O(h^4)$. However, it turned out that the corresponding $l = -l_1 = l_2$ depends on the unknown density $f$ which is also undesirable.

Recently, Glad, Hjort and Ushakov(1999) suggested correction of density estimators which may not be nonnegative or/and do not integrate to one. They showed that the corrected estimators has smaller mean integrated squared error than the original estimator.

Remark 2.3 : In our limited experience, we didn't get negative values of $\tilde{f}(x)$ so far except a very extreme case such as the separated bimodal.

## 3. Numerical Results

For the Gaussian density, we compare $\tilde{f}$ with a standard second-order kernel estimator $\hat{f}$ in (1), two-parameter locally parametrie estimator by Cheng, et.al (2000) $\hat{f}_\infty$ in (6), and a fourth-order kernel estimator $\hat{f}_{(4)}$. N(0,1) is used as a kernel for $\tilde{f}, \hat{f}, \hat{f}_\infty$, and $\frac{1}{2}(3-x^2)\phi(x)$ is used as a kernel for $\hat{f}_{(4)}$, where $\phi(x)$ is N(0,1) pdf. Also, we used sample size $n$=100.

Figure 3(a) shows the mean integrated squared error(MISE) performances of four estimators.

As anticipated, $\hat{f}$ performs worst, and $\tilde{f}, \hat{f}_\infty, \hat{f}_{(4)}$ perform similarly. Among them, $\tilde{f}$ is slightly better than $\hat{f}_\infty$ and $\hat{f}_{(4)}$. Figure 3(b) shows the four estimators with the true density. The peak is captured better by $\tilde{f}$ than by others.

## Appendix : Proof of Theoran 2.1

Let $\nu_l = \int t^l K'(t)dt$ , then $\nu_2 = \nu_4 = 0$. Now, by a Taylor expansion, it is easy to show that

$$E[\hat{f}(x)] = f(x) + \frac{1}{2}h^2 f''(x)\mu_2 + \frac{1}{24}h^4 f^{(4)}(x)\mu_4 + O(h^5),$$

$$\begin{aligned} E[\hat{f}(x+lh)] &= f(x) + hlf'(x) + \frac{1}{2}h^2(l^2+\mu_2)f''(x) + \frac{1}{6}h^3(l^3+3l\mu_2)f'''(x) \\ &\quad + \frac{1}{24}h^4(l^4+6l^2\mu_2+\mu_4)f^{(4)}(x) + O(h^5), \end{aligned}$$

and

$$\begin{aligned} E[\hat{f}'(x+lh)] &= -\nu_1 f'(x) - \nu_1 hlf''(x) - \frac{1}{6}h^2(3l^2\nu_1+\nu_3)f'''(x) \\ &\quad - \frac{1}{24}h^3(4l^3\nu_1+4l\nu_3)f^{(4)}(x) + O(h^4). \end{aligned}$$

Therefore,

$$E[\hat{f}_1(x)] = f(x) + hlf'(x)(1+\nu_1) + \frac{1}{2}h^2(l_2+\mu_2+2\nu_1 h^2)f''(x)$$

$$+\frac{1}{6}h^3 l(l^2+3\mu_2+3l^2\nu_1+\nu_3)f'''(x)$$
$$+\frac{1}{24}h^4(l^4+6l^2\mu_2+\mu_4+4l^4\nu_1+4l^2\nu_3)f^{(4)}(x)+O(h^5).(A.1)$$

By substituting $l=0,l_1,l_2$ in (A.1), and combining them to produce a formula for bias, we see that the terms in $h^2$ and $h^3$ disappear if and only if

$(i)$ $\lambda_1 l_1+\lambda_2 l_2=0$,
$(ii)$ $\lambda_1\{{l_1}^2(1+2\nu_1)+\mu_2\}+\mu_2+\lambda_2\{{l_2}^2(1+2\nu_1)+\mu_2\}=0$,
$(iii)$ $\lambda_1 l_1\{{l_1}^2(1+3\nu_1)+3\mu_2+\nu_3\}+\lambda_2 l_2\{{l_2}^2(1+3\nu_1)+3\mu_2+\nu_3\}=0$.

If we assume that $\lambda_1,\lambda_2>0$ and $\lambda_1,\lambda_2\neq 0$, (i) and (iii) imply that $\lambda_1=\lambda_2=\lambda$ and $\lambda_1=-\lambda_2=\lambda$, say. Now (ii) gives $l=l(\lambda)$ given in (4).

## References


Cheng, M-Y., Choi, E., Fan, J. and Hall, P.(2000) Skewing-methods for two parameter locally-parametric density estimation, *Bernoulli*, 6,169-182

Choi,E. and Hall,P (1998) On bias reduction in local linear smoothing, *Biometrika*, 85, 333-345.

Copas, J.B. (1995) Local likelihood based on kernel smmothing, *Journal of the Royal Statistical Society*, Ser.B. 57, 221-235.

Glad, I.K.,Hjort, N.L. and Ushakov, N.G. (1999) Correction of density estimators which are not densities, Manuscript

Hjort, N.L. and Jones, M.C. (1996) Locally parametric nonparametric density estimation, *The Annals of Statistics*, 24, 1619-1647.

Loader, C.R.(1996) Local likelihood density estimation, *The Annals of Statis tics*, 24,1602-1618.

Marron, J.S and Wand, M.(1992) Exact mean integrated squared error, *The Annals of Statistics*,20, 712-736.

Rice, J.A (1984) Boundary modification for nonparametric regression, *Communicationa in Statistics -Theory and Methods*, 13, 893-900.

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

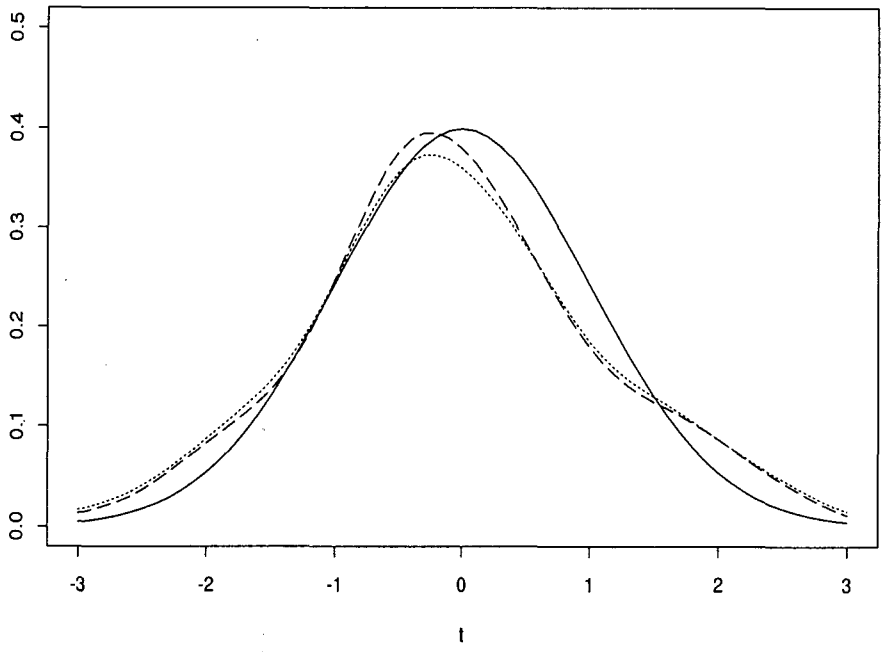
Wand, M.P. and Jones, M.C.(1995) *Kernel Smoothing*, Chapman and Hall, London.

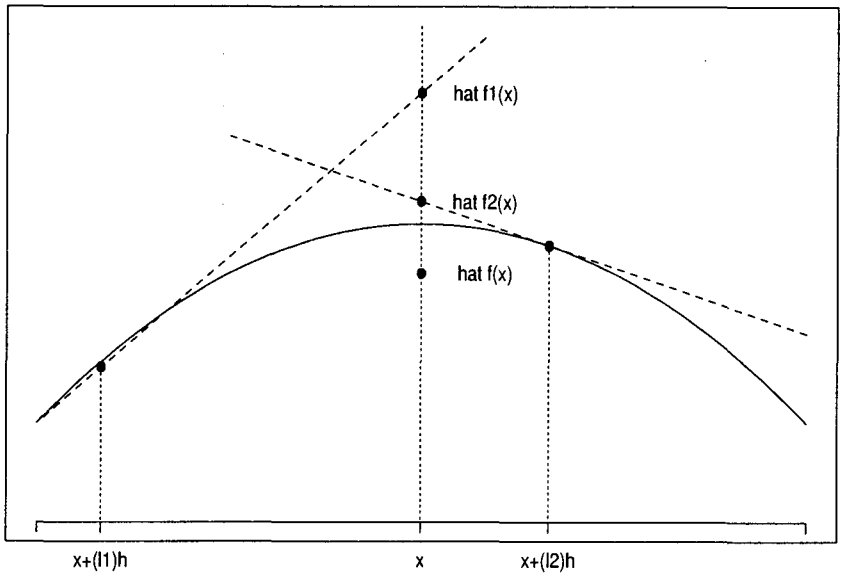
Figure Legends

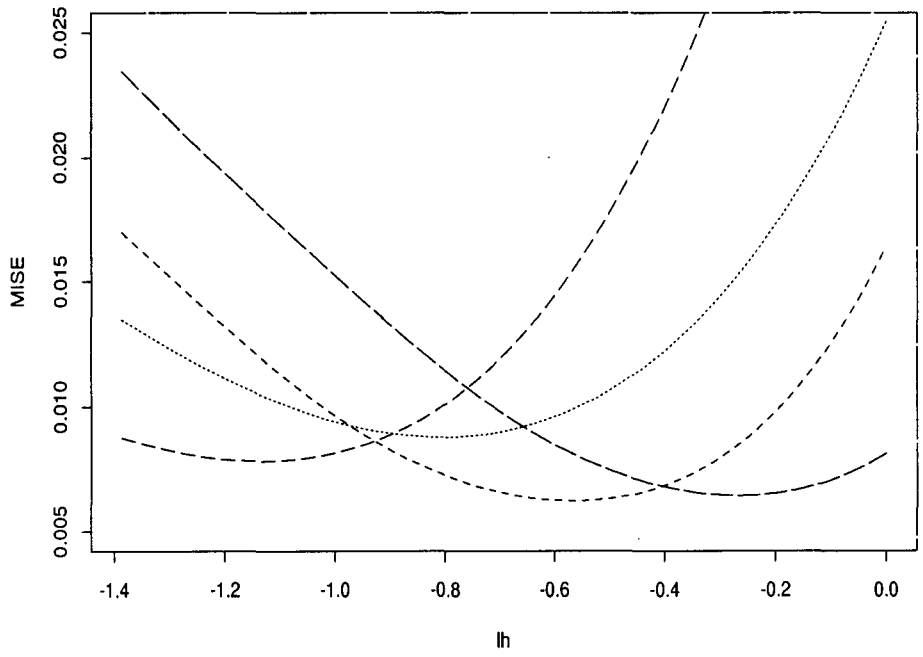
Figure 1 : N(0,1) (——), kernel estimator $\hat{f}(x)$ (· · · · · ·), and the proposed estimotor $\tilde{f}(x)$ (- - - - -).

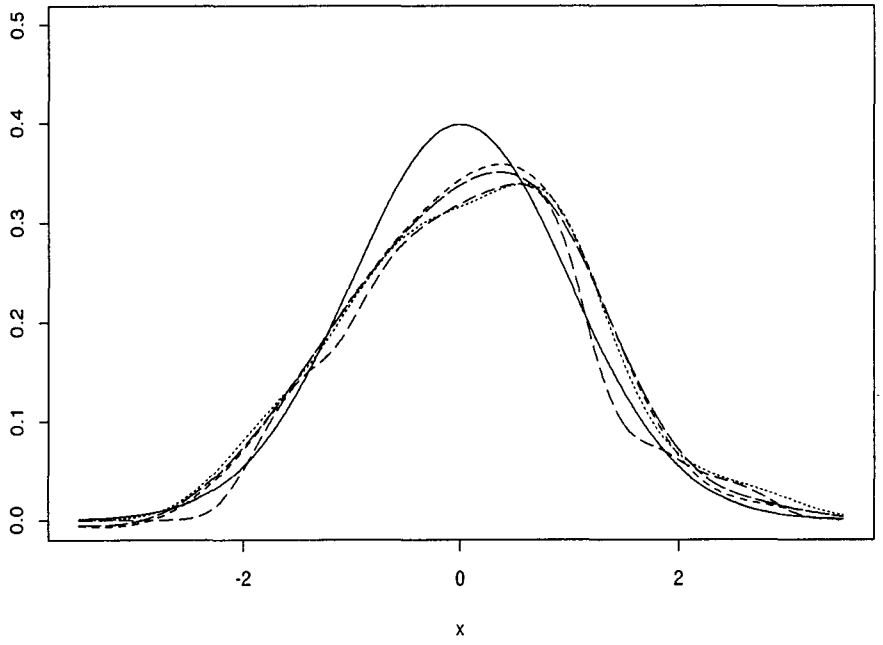
Figure 2 : Convex combination of three kernel estimators.

Figure 3 : (a) MISE of four estimators : kernel estimator $\hat{f}(x)$ (· · · · · ·), the proposed estimator $\tilde{f}(x)$ (- - - - -), the locally-parametric estimaotr $\hat{f}_{\infty}(x)$ (– – –), and the fourth-order kernel estimator $\hat{f}_{(4)}(x)$ (— —). (b) Plot of four estimators for the true density N(0,1).


0.0
0.1
0.2
0.3
0.4
0.5
-3
-2
-1
0
1
2
3
t


hat f1(x)
hat f2(x)
hat f(x)
x+(l1)h
x
x+(l2)h


MISE
0.005
0.010
0.015
0.020
0.025
-1.4
-1.2
-1.0
-0.8
-0.6
-0.4
-0.2
0.0
lh


0.0
0.1
0.2
0.3
0.4
0.5
-2
0
2
x