The Korean Communications
in Statistics Vol. 10 No. 2, 2003
pp. 607-617

On Practical Efficiency of Locally Parametric Nonparametric Density Estimation Based on Local Likelihood Function[1)]

Kee-Hoon Kang[2)] and Jung-Hoon Han[3)]

Abstract

This paper offers a practical comparison of efficiency between local likelihood approach and conventional kernel approach in density estimation. The local likelihood estimation procedure maximizes a kernel smoothed log-likelihood function with respect to a polynomial approximation of the log likelihood function. We use two types of data driven bandwidths for each method and compare the mean integrated squares for several densities. Numerical results reveal that local log-linear approach with simple plug-in bandwidth shows better performance comparing to the standard kernel approach in heavy tailed distribution. For normal mixture density cases, standard kernel estimator with the bandwidth in Sheather and Jones(1991) dominates the others in moderately large sample size.

Keywords : Kernel density estimation, bandwidth, local log likelihood, local linear estimator

1. Introduction

Let $\{X_1, \cdots, X_n\}$ denote independent and identically distributed random sample drawn from the population density f. In the parametric setting, the density is usually modelled by a finite dimensional vector of parameters which has the form $f(x, \theta)$. Then, an estimator of the density is $f(x, \hat{\theta})$, which is obtained by plugging in the usual maximum likelihood estimator $\hat{\theta}$. In contrast, nonparametric approaches do not assume the parametric family of densities and the standard nonparametric kernel estimator of f is

$$\hat{f}_h(x) = n^{-1} \sum_{i=1}^{n} K_h(x - X_i), \tag{1.1}$$

1) This work was supported by Korea Research Foundation Grant. (KRF-2001-003-D00024).
2) Assistant Professor, Department of Statistics, Hankuk University of Foreign Studies, Yongin 449-791, KOREA.
E-mail : khkang@hufs.ac.kr
3) Graduate Student, Department of Statistics, Hankuk University of Foreign Studies, Yongin 449-791, KOREA.

where $K_h(\cdot)=(1/h)K(\cdot/h)$ for a "kernel function" K, which is open taken to be a smooth, bounded and symmetric probability density and a "bandwidth" or "smoothing parameter" h.

The basic properties of such a kernel estimator are well known and these include

$$E\{\hat{f}_h(x)\}=f(x)+\frac{1}{2}h^2x_2f^{(2)}(x)+o(h^2),$$
$$Var\{\hat{f}_h(x)\}=(nh)^{-1}R(K)f(x)+o\{(nh)^{-1}\}, \tag{1.2}$$

where $x_2=\int z^2K(z)dz$ and $R(K)=\int K(z)^2dz$. See for example, Wand and Jones(1995, pp.20-21).

Local likelihood function is proposed by Tibshirani and Hastie(1987) as a method of approximating non-Gaussian regression model such as logistic regression and proportional hazards model by local polynomial. Extensions of local likelihood methods to the nonparmetric density estimation setting are made by Loader(1996) and Hjort and Jones(1996). The properties of local likelihood density estimation with large bandwidth is described in Eguchi and Copas(1998), which corresponds to the case when the underlying true density is parametric or near parametric. When the parametric estimator is unsuitable for the data, nonparametric aspect of the local likelihood estimation is needed and asymptotic properties of this approach with small bandwidth is presented in Park, Kim and Jones(2002).

Hall and Tao(2002) argue that standard kernel methods have comparable properties to the local likelihood approach in terms of asymptotic mean integrated squared error comparison. However, non of the aforementioned papers explore the practical comparison of performance of these estimators. Main purpose of this paper is to make such comparison and explore the behavior of local likelihood density estimator with data driven bandwidths. Our approaches and results have close relation with Hjort and Jones(1996) and Hall and Tao(2002).

The remaining sections of this paper are as follows. Section 2 gives a brief introduction of local likelihood density estimation and explores its properties. Two types of bandwidth selection methods for local log-linear density estimator and standard kernel density estimator are briefly described in section 3. Numerical comparison between these two estimators is summarized in section 4. Some concluding remarks are given in section 5.

2. Local likelihood Approach

The log-likelihood function of observations $X_1,\cdots,X_n$ with unknown density f is

$$L(f)=\sum_{i=1}^{n}\log(f(X_i))-n(\int f(x)dx-1). \tag{2.1}$$

By Loader(1996) and Hjort and Jones(1996), local likelihood function around each x can be defined as

$$L_n(f,x)=n^{-1}\sum_{i=1}^{n}K\left(\frac{X_i-x}{h}\right)\log f(X_i)-\int K\left(\frac{u-x}{h}\right)f(u)\,du. \tag{2.2}$$

This corresponds to the localized version of the log-likelihood, which gives more weight to data in the region of an estimating point x, and less weight to observations elsewhere in the sample space.

The local likelihood estimation procedure maximizes a kernel smoothed log-likelihood function (2.2) with respect to a polynomial approximation of the log likelihood function in a neighborhood of the fitting point x. That is, $\log f(u)\approx P(u-x)$(in one dimension) with $P(u-x)=\theta_0+\theta_1(u-x)+\cdots+\theta_p(u-x)^p$. By plugging this into (2.2), this approximation gives the following local likelihood function :

$$L_p(f,x)=n^{-1}\sum_{i=1}^{n}K\left(\frac{X_j-x}{h}\right)P(X_i-x)-\int K\left(\frac{u-x}{h}\right)\exp(P(u-x))du. \tag{2.3}$$

Then, local likelihood density estimate is defined by

$$\hat{f}(x)=\exp(\hat{\theta}_0), \tag{2.4}$$

where $(\hat{\theta}_0, \hat{\theta}_1, \cdots, \hat{\theta}_p)$ is the maximizer of (2.3). See Loader(1996) and Hjort and Jones(1996) for more precise definition.

In this paper, we shall focus our attention on the case of $p=1$, which results in local log-linear estimator denoted by $\hat{f}_L$. (Let us denote standard kernel density estimator (1.1) by $\hat{f}_K$.) That is, we consider the local model $a\exp(b(u-x))$ and the corresponding score function is $(1/a,\ u-x)'$, and the two equations to solve, in order to maximize the local likelihood, are

$$n^{-1}\sum_{i=1}^{n}K_h(X_i-x)\begin{pmatrix}1/a\\ X_i-x\end{pmatrix}=\int K_h(u-x)\begin{pmatrix}1/a\\ u-x\end{pmatrix}a\exp(b(u-x))du. \tag{2.5}$$

Note that the resulting local linear estimator $\hat{f}_L(x)=\hat{a}(x)$, where $\hat{a}(x)$ is the solution of equation (2.5).

As noted by Loader(1996) and Hjort and Jones(1996), $\hat{f}_L$ has the following bias and variance expression :

$$\begin{aligned}E\{\hat{f}_L(x)\}&=f(x)+\frac{1}{2}h^2x_2b(x)+o(h^2),\\ Var\{\hat{f}_L(x)\}&=(nh)^{-1}R(K)f(x)+o\{(nh)^{-1}\},\end{aligned} \tag{2.6}$$

where $b(x)=f^{(2)}(x)-f'(x)^2f(x)^{-1}$.

Comparing (2.6) with (1.2), variance is the same and the difference is in the bias. The local linear estimate is better when $|b(x)|<|f^{(2)}(x)|$. In the central part of the distribution, either of the bias can be the larger. The two bias terms are equal whenever $f'(x)=0$, suggesting the estimates would have similar performance near the modes and troughs in the density.

Hall and Tao(2002) made a global comparison by introducing the mean integrated squared error (MISE) expression of each estimator. From (1.2) and (2.6), both estimators have the following MISE expression :

$$\text{MISE}(h) = E\int(\hat{f}(x) - f(x))^2 dx = h^4 \int b(x)^2 dx + (nh)^{-1} R(K), \tag{2.7}$$

where $b(x) = b_K(x) \equiv x_2 f^{(2)}(x)/2$ for $\hat{f} = \hat{f}_K$ and $b(x) = b_L(x) \equiv x_2 \{f^{(2)}(x) - f'(x)^2 f^{-1}(x)\}/2$ for $\hat{f} = \hat{f}_L$. Simple calculation results in

$$\int b_L(x)^2 dx = \int b_K(x)^2 dx + \frac{1}{12} x_2^2 \int \frac{f'(x)^4}{f(x)^2} dx, \tag{2.8}$$

which implies that $\int b_L(x)^2 dx > \int b_K(x)^2 dx$, and so the kernel estimator is superior in terms of global performance. The variance term in the formula (2.7) is the same for both of kernel and local likelihood estimator.

When we use the Gaussian kernel function, the solution of equation (2.5) has a simple expression, which relates to standard kernel estimator. This can be obtained by letting $\psi(u) = \int \exp(uz) K(z) dz$, which is the moment generating function of K. Then, the two equations in (2.5) becomes

$$\begin{aligned} n^{-1} \sum_{i=1}^{n} K_h(X_i - x) &= a\psi(bh), \\ n^{-1} \sum_{i=1}^{n} K_h(X_i - x)(X_i - x) &= ah\psi'(bh). \end{aligned} \tag{2.9}$$

In fact, $\psi(u) = \exp(u^2/2)$ for Gaussian kernel. In this case, solving two equations in (2.9) results in

$$\hat{f}_L(x) = \hat{a}(x) = \hat{f}_K(x) \exp[-\frac{1}{2} h^2 \{\hat{f}_K'(x)/\hat{f}_K(x)\}^2]. \tag{2.10}$$

See section 5.2 of Hjort and Jones(1996) for more details. We shall compare the empirical performance of this local log-linear estimator $\hat{f}_L$ with standard kernel estimator $\hat{f}_K$ in section 4.

3. Bandwidth Selection

Both of the standard kernel estimation and local likelihood estimation procedure described in the previous section highly depend on the selection of bandwidth. There is an extensive literature on the bandwidth choice in the standard kernel density estimation. A brief survey of those selection methods is presented in Jones et al.(1996). However, there are few results on the bandwidth selection for the local likelihood density estimation. This section describes some of bandwidth selectors which are used as data driven bandwidths in this paper.

First, we begin with cross-validatory bandwidth selector for standard kernel density estimation. Cross-validation is a popular, utilitarian method for choosing bandwidth in curve estimation. Not least among its attractive features are the very wide range of contexts where it may be applied. Comparing to the other selection criteria, it does not require any auxiliary stage of estimation. It also has demerits that the resulting bandwidth is more highly variable than that selected by a plug-in rule; see for example Park and Marron(1990) and Park and Turlach(1992). However in many circumstances, for example where oversmoothing can obscure important features of a curve, it performs well; see for example Loader(1999).

As is well known, the least square cross-validation function for standard kernel density estimator is defined by

$$CV(h)=\int \hat{f}_h(x)^2\,dx-2n^{-1}\sum_{i=1}^{n}\hat{f}_{h,-i}(X_i), \tag{3.1}$$

where $\hat{f}_{h,-i}(x)=(n-1)^{-1}\sum_{j\neq i}^{n}K_h(x-X_j)$ is the density estimate based on the sample with X_i deleted. Then, the cross-validation bandwidth is obtained by minimizing (3.1) with respect to h and denoted by $\hat{h}_{CV}$.

The second one for standard kernel density estimate is proposed by Sheather and Jones(1991), which belongs to the second generation classified by Jones et al.(1996). This is known to be the best one for the standard kernel estimator in terms of overall performance. The idea is motivated by the formula for minimizing asymptotic MISE and to take $\hat{h}_{SJ}$ to be the solution of the following equation :

$$h=\left[\frac{R(K)}{x_2^2\hat{\psi}_4(\gamma(h))n}\right]^{1/5}, \tag{3.2}$$

where $\hat{\psi}_4(\gamma(h))$ is the estimate of $\psi_4=\int f^{(4)}(x)f(x)\,dx=\int f^{(2)}(x)^2\,dx$ with the bandwidth γ, which is a function of h. Detailed procedure and example are described in Wand and Jones(1995).(See, pp. 74-75).

Now, let us consider the bandwidth choice for the local likelihood approach. Recall that the asymptotic mean integrated squared error of the local likelihood approach has the following form :

$$\text{MISE}(\hat{f}_L;h)=\frac{1}{4}h^4x_2^2\int\{f^{(2)}(x)-f'(x)^2f^{-1}(x)\}^2dx+(nh)^{-1}R(K). \tag{3.3}$$

The bandwidth which minimizes the above MISE is expressed by

$$h=\{R(K)x_2^2\}^{1/5}R(f^{(2)}-f'^2f^{-1})^{-1/5}n^{-1/5}. \tag{3.4}$$

By plugging an estimate of $R(f^{(2)}-f'^2f^{-1})$ into equation (3.4), we obtain the bandwidth estimate and denoted by $\hat{h}_{LP}$.

Cross-validation function for the local likelihood estimate is similar to equation (3.1), which is defined by

$$CV_L(h) = \int \hat{f}_L(x;h)^2 dx - 2n^{-1} \sum_{i=1}^{n} \hat{f}_{L,-i}(X_i;h), \tag{3.5}$$

where $\hat{f}_{L,-i}(X_i;h)$ is the local log-linear density estimate with i-th observation is deleted. Then, the cross-validation bandwidth for local likelihood estimate is the minimizer of equation (3.5), which is denote by $\hat{h}_{LC}$. Comparing to $\hat{h}_{LP}$, it has the advantage that does not involve other estimation stage related to function of f.

4. Numerical Comparison

This section is devoted to comparing the small sample performance of the ordinary kernel density estimators ($\hat{f}_K$) with that of the local log-linear estimators ($\hat{f}_L$) on simulated data sets. Two versions of bandwidth selector, which is described in section 3, are considered for each estimator. The standard Gaussian kernel was used throughout.

We shall summarize results obtained from simulated data drawn from four target distributions proposed by Marron and Wand(1992) and two other distributions. See Table 1 for the distributions, and Figure 1 for graphs of the corresponding densities. Distributions #1 and #5 were chosen because they represent opposite extremes in terms of tail weight, distribution #2 represents bimodal, distribution #3 is separated bimodal, density #4 is trimodal, and density #6 is highly skewed with a long, relatively flat portion. The results given here are for cases of sample sizes $n=25$, 50, 100 and 200, and derived from 500 simulations in each setting.

We employ a grid search over a fine grid of h, for estimating cross-validation bandwidth in both settings of local likelihood and standard kernel estimators. We use MISE as a measure of performance of estimators, which is defined by $\mathrm{MISE}(h) = E\int(\hat{f}_h - f)^2$, since it is preferred because of its simplicity and mathematical conveniences. Instead of arranging quite a lot of tables which contain MISEs and their standard errors, we shall display some figures which is visually interpreted. Tables on this simulation result may be available from the first author.

Let us denote local log-linear density estimator with plug-in bandwidth(from equation (3.4)) and cross-validation bandwidth(from (3.5)) by $\hat{f}_{LP}$ and $\hat{f}_{LC}$, respectively. And standard kernel estimator with Sheather and Jones(1991) bandwidth(from (3.2)) and cross-validation bandwidth(from (3.1)) are denoted by $\hat{f}_{SJ}$ and $\hat{f}_{CV}$, respectively.

Figure 2 depicts the relative efficiencies of $\hat{f}_{LP}$ to $\hat{f}_{LC}$, $\hat{f}_{SJ}$ and $\hat{f}_{CV}$ in each sample size. That is, vertical values correspond to the ratios of MISE of $\hat{f}_{LP}$ relative to the other three estimators. Except distribution #3, $\hat{f}_{LP}$ shows the best behavior when the sample size is relatively small. As the sample size grows $\hat{f}_{SJ}$ has better performance in all target densities

except $t(2)$ distribution. $\hat{f}_{LC}$ is not comparable to the others except for density #5 and $\hat{f}_{CV}$ is the best for density #3. This means both of local likelihood estimator and kernel estimator are highly dependent on the choice of the bandwidths.

The performance of local likelihood estimator $\hat{f}_{LP}$ and $\hat{f}_{LC}$ is the worst for the distribution #3 among our simulations. This density has a trough which has very low density and relatively sharp two modes. When the density is very close to zero, there is a difficulty in modelling by a local polynomial since the zero is a singularity of the log density. The greatest benefit of using local likelihood estimator is enjoyable for the distribution #5, which is Student's $t(2)$ case. This density has relatively heavy tails and the pointwise bias term in (2.6) is smaller than that for standard kernel estimator for wide range. This property can be conveyed to the other densities which have heavy tails. For all distributions, we have examined the integrated variance (IVAR) and the integrated bias square (IBIAS) separately. We found that IVAR of the local likelihood estimator is almost the same or smaller than that for standard kernel estimator. But, the IBIAS is relatively large except for the density #3.

To have a more careful comparison between the distribution #1 and #5, we plot the MISE, integrated variance (IVAR) and integrated bias square (IBIAS) as a function of h in the logarithmic scale, for sample size 100. Figure 3 shows the result. From this figure, the better performance of $\hat{f}_{LP}$ in distribution #5 came from smaller bias than the standard kernel estimator although it's not so big. Note that the scale of vertical axes in variance and bias is different.

We note that there is room for improvement of performance of local log-linear density estimator. This may be possible by computing $\hat{f}_K'(x)$, which is in the definition of $\hat{f}_L$ in equation (2.10), separately from $\hat{f}_K(x)$ with a somewhat larger bandwidth and a different kernel. This is because local slope estimations typically require larger bandwidth than for local level estimation. We do not pursue this here.

5. Concluding Remarks

We have examined the performance of the local log-likelihood density estimator with data driven bandwidths via numerical simulation. Overall, the results in section 4 coincide with the theoretical arguments in Hall and Tao(2002). Note that $\hat{h}_{SJ}$ is the best bandwidth for standard kernel estimation in terms of theory and practice. If one suggest a good data driven bandwidth for the local likelihood approach, the performance can be more improved. This problem has not been dealt yet and should be studied in the future. One other possibility of enhancing the performance of local likelihood estimation is allowing the bandwidth h to vary as a function of x or X_i.

References

[1] Eguchi, S. and Copas, J. (1998). A class of local likelihood methods and near-parametric asymptotics, *Journal of the Royal Statistical Society*, B, 60, 551-563.

[2] Gasser, T. and Müller, H.G. (1979). Kernel estimation of regression functions. In *Smoothing Techniques for Curve Estimation* (eds. T. Gasser and M. Rosenblatt), Springer-Verlag, Heidelberg, pp. 23-68.

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

[4] Hall, P. and Tao, T. (2002). Relative efficiencies of kernel and local likelihood density estimators, *Journal of the Royal Statistical Society*, B, 64, 537-547.

[5] Kim, W.C, Park, B.U. and Kim, Y.G. (2001). On Copas' local likelihood density estimator. *Journal of the Korean Statistical Society*, 30, 77-87.

[6] Loader, C. (1996). Local likelihood density estimation, *The Annals of Statistics*, 24, 1602-1618.

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

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

[9] Park, B.U., Kim, W.C. and Jones, M.C. (2002). On local likelihood density estimation, *The Annals of Statistics*, 30, 1480-1495.

[10] Park, B.U. and Marron, J.S. (1990). Comparison of data-driven bandwidth selectors, *Journal of the American Statistica Association*, 85, 66-72.

[11] Park, B.U. and Turlach, B.A. (1992). Practical performance of several data driven bandwidth selectors, *Computational Statistics*, 7, 251-270.

[12] 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*, B, 53, 683-690.

[13] Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation, *Journal of the American Statistical Association*, 82, 559-567.

[Received March 2003, Accepted August 2003]

Table 1. Definitions of six distributions

Distribution	Definition
#1 Gaussian	$N(0,1)$
#2 Bimodal	$\frac{1}{2}N(-1,(\frac{2}{3})^2)+\frac{1}{2}N(1,(\frac{2}{3})^2)$
#3 Separated bimodal	$\frac{1}{2}N(-\frac{3}{2},(\frac{1}{2})^2)+\frac{1}{2}N(\frac{3}{2},(\frac{1}{2})^2)$
#4 Trimodal	$\frac{9}{20}N(-\frac{6}{5},(\frac{3}{5})^2)+\frac{9}{20}N(\frac{6}{5},(\frac{3}{5})^2)+\frac{1}{10}N(0,(\frac{1}{4})^2)$
#5 Heavy tailed	$t(2)$
#6 Highly skewed	$\frac{7}{20}N(-1,(\frac{3}{5})^2)+\frac{1}{2}N(1,(\frac{5}{2})^2)+\frac{3}{20}N(5,(\frac{3}{2})^2)$

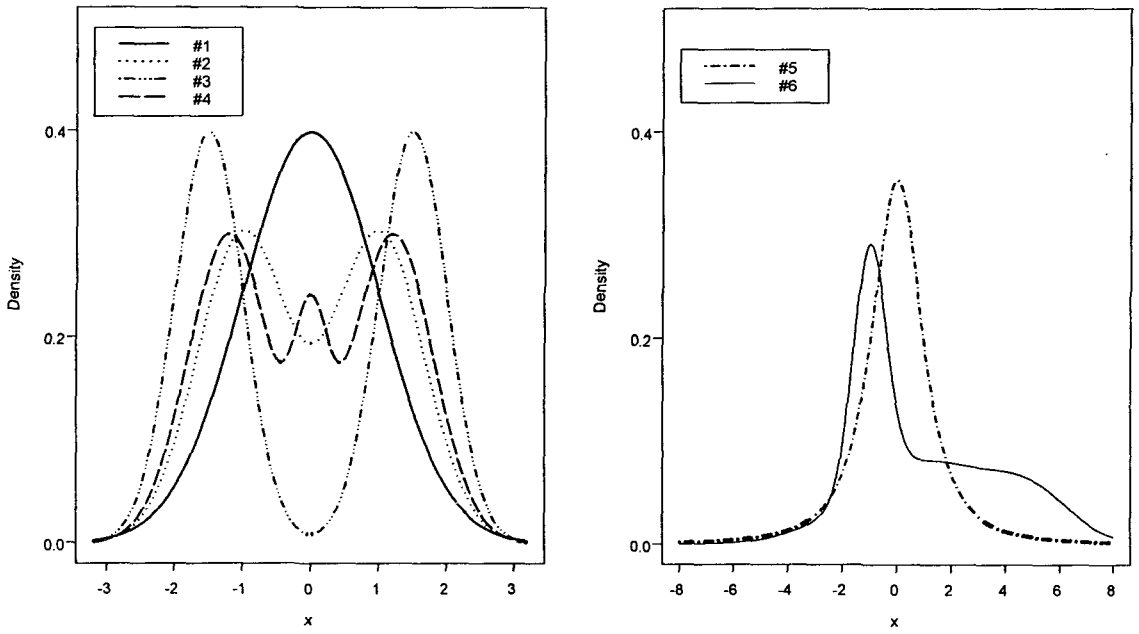

Figure 1. *Graphs of true distributions.* Graphs of six Normal mixture densities and Density #1 is standard Normal, density #2 is bimodal, density #3 is separated bimodal, density #4 is trimodal, density #5 is Student's t density with two degrees of freedom, and density #6 is highly skewed with a long, relatively flat portion.

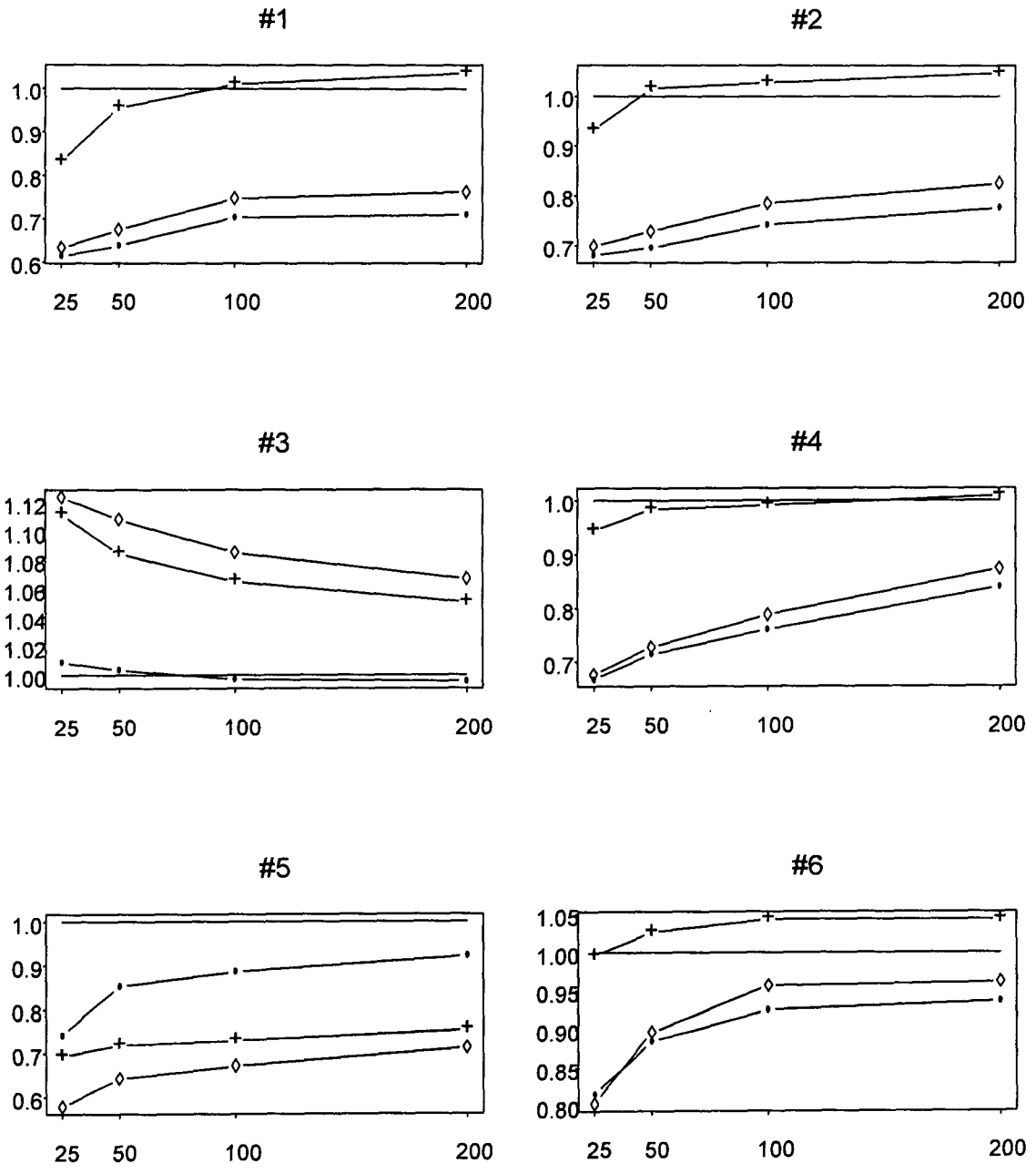

Figure 2. *Efficiency plots.* Relative efficiency versus sample size, where vertical values corresponds to the ratios of MISE of the $\hat{f}_{LP}$, relative to $\hat{f}_{LC}(\cdot)$, $\hat{f}_{CV}(\diamond)$ and $\hat{f}_{SJ}(+)$.

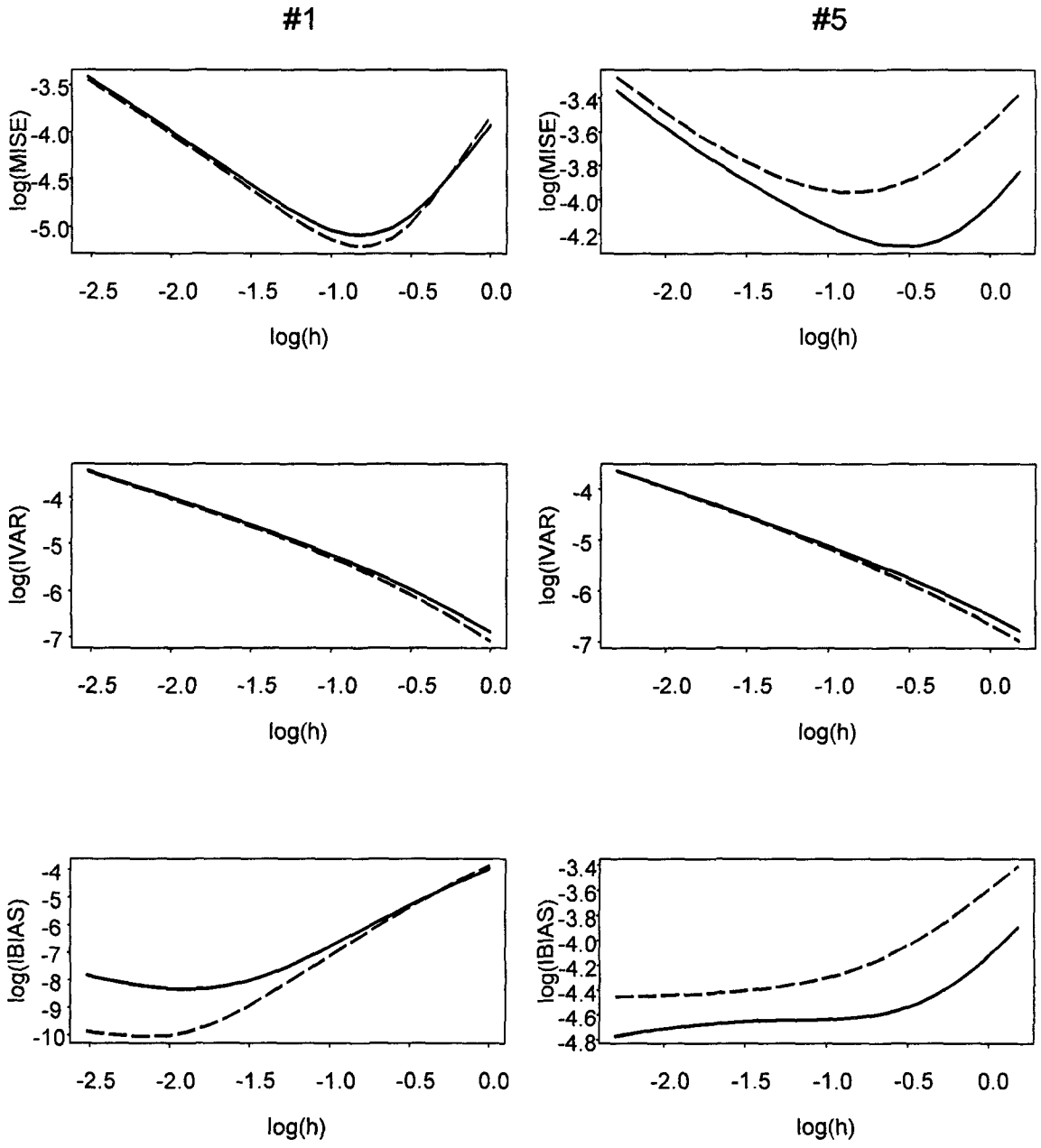

Figure 3. *Comparison between N(0,1) and t(2) density.* Mean integrated squared errors(MISE), integrated variances(IVAR) and integrated bias squares(IBIAS) plots versus $\log(h)$. Vertical axes are in logarithmic scale. The left panel corresponds to $N(0,1)$ and the right panel to $t(2)$ density. Solid line corresponds to the local likelihood density estimator and the dashed line corresponds to the standard kernel estimator.