Youngnam Statistical Letters
June 1992, Vol. 3, No. 1, 17-24

Transformation in Kernel Density Estimation*

Kyung Ha Seog**

ABSTRACT

The problem of estimating symmetric probability density with high kurtosis is considered. Such densities are often estimated poorly by a global bandwidth kernel estimation since good estimation of the peak of the distribution leads to unsatisfactory estimation of the tails and vice versa. In this paper, we propose a transformation technique before using a global bandwidth kernel estimator. Performance of density estimator based on proposed transformation is investigated through simulation study. It is observed that our method offers a substantial improvement for the densities with high kurtosis. However, its performance is a little worse than that of ordinary kernel estimator in the situation where the kurtosis is not high.

Key word : transformed kernel density estimation

1. Introduction

Consider the problem of estimating symmetric probability density function f_X from real-valued random sample $X_1, \cdots, X_n$. The global bandwidth kernel estimator is

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

where the kernel K is symmetric density and $K_h(x) = K(x/h)/h$. Silverman (1986) provides a detailed account of the kernel estimator and its applications.

* This paper was supported by non-directed research fund 1990
** Department of Statistics, Inje University, Kimhae

The presence of high peak at the center will have an unsatisfactory effect on the performance of (1.1), since the amount of smoothing is uniform across the sample space. For this density, a good estimation of the sharp peak requires a relatively small bandwidth which is likely to induce artificial bumps in the tail parts. If a larger bandwidth is chosen to smooth out these bumps then the peak will almost certainly be oversmoothed. To overcome this problem we propose applying the transformation g_λ to the data $X_1, \cdots, X_n$ to obtain $Y_1, \cdots, Y_n$ with common density $f_Y(\cdot;\lambda)$. The parameter λ lies in some finite dimensional set Λ. The immediate goal of the transformation is to reduce the kurtosis of $f_Y(\cdot;\lambda)$ but the ultimate goal is a density that is easy to estimate using (1.1). The back transformation by change of variable from $\hat{f}_Y(\cdot;\lambda)$ to $\hat{f}_X(\cdot;\lambda)$ is our proposed estimator.

Our proposed estimator may be viewed as an alternative to the variable window width kernel estimator (Breiman, Meisel and Purcell (1977), Abramson (1982)) which permits the bandwidth to vary at each sample point X_i proportionally to $f_X(X_i)^{-\gamma}$ for some $0 < \gamma \leq 1$. This approach, however, requires the specification of a pilot estimator for f_X itself for effective implemention. The transformation algorithm we use in this paper is similar to that of Wand, Marron and Ruppert (1990) where the estimation of nonnegative skewed densities was the main concern.

In the next section we will cover the theory of transformed kernel density estimator. In Section 3, a suitable two-parametr family of transformation is proposed and in Section 4, the results of simulation study are given. For the densities with high kurtosis, the performance of the proposed transformed kernel estimator is much superior to that of ordinary kernel estimator, and is a little worse for the case where the kurtosis is not high.

2. Theory of Transformed Kernel Density Estimators

The transformation technique for density estimation is discussed extensively in Wand *et al.*(1990). In this section we briefly review the transformation ideas.

Let X be a random variable having density f_X and $\{\tilde{g}_\lambda : \lambda \in \Lambda\}$ be some parametric family of increasing transformations defined on the support of f_X. Put $\tilde{Y} = \tilde{g}_\lambda(X)$. So that scale is preserved we will take our transformation to be

$g_\lambda = (\sigma_X/\sigma_{\tilde{Y}})\tilde{g}_\lambda$ where σ_X and $\sigma_{\tilde{Y}}$ are standard deviations of X and $\tilde{Y}$ respectively. Our estimator of $f_Y(y;\lambda) = f_X(g_\lambda^{-1}(y))(g_\lambda^{-1})'(y)$ is the usual global bandwidth kernel estimator

$$\hat{f}_Y(y;\lambda) = n^{-1}\sum_{i=1}^{n} K_h(y - Y_i). \tag{2.1}$$

The transformed kernel density estimator is the back transform of (2.1), which is given by

$$\hat{f}_X(x;h,\lambda) = n^{-1}\sum_{i=1}^{n} g'_\lambda(x)K_h[g_\lambda(x) - g_\lambda(X_i)] \tag{2.2}$$

Note that, from the mean value theorem,

$$\hat{f}_X(x;h,\lambda) = n^{-1}g'_\lambda(x)\sum_{i=1}^{n} K_h[(x - X_i)g'_\lambda(\xi_i)]$$

where ξ_i lies between x and X_i. Comparing this formulation with (1.1) we see that the bandwidth at x is approximately $h/g'_\lambda(x)$.

As in Wand *et al.*(1990) we aim to choose the parameter λ so that $f_Y(\cdot;\lambda)$ can be estimated with the smallest possible error. Under the assumptions that $f_Y(\cdot;\lambda)$ possesses two continuous derivatives and that $h = h(n) \to 0$ and $nh \to \infty$ as $n \to \infty$, Wand *et al.*(1990) showed that the lowest asymptotic mean integrated squared error of $\hat{f}_Y(\cdot : \lambda)$ can be achieved by taking λ to minimize

$$J_Y(\lambda) = [\int f''_Y(\cdot;\lambda)^2]^{\frac{1}{5}} \tag{2.3}$$

This result is the basis for our data-driven selection of λ as described in Section 3. Also, because (2.1) is simply a global bandwidth kernel estimator any one of the current bandwidth selection procedures (see e.g. Park and Marron(1990)) may be applied to select h.

3. The Proposed Transformation Method

In this section we will assume that the density f_X is symmetric and unimodal. In the theoretical development below, the center of symmetry is assumed to be 0 - in examples this is achieved by subtraction of the sample median from each observation.

Ruppert (1987) gives an account of several notions of comparative kurtosis. The most relevent to this work is that introduced by van Zwet(1964) who defined f_Y as having no more kurtosis than f_X if $Y = g(X)$ where g is convex on the negative half line and concave on the positive half line. Such a convex-concave transformation has the effect of taking probability mass from both the peak and the tails and moving it to shoulders (which reduces peakedness and lightens the tails).

It is important that our family of convex-concave transformation is sufficiently smooth so the $f_Y(\cdot;\lambda)$ inherits the smoothness properties of f_X. A class of transformations having the required properties is given by (with $\lambda = (\lambda_1, \lambda_2)$)

$$\tilde{g}_{\lambda_1,\lambda_2}(x) = [(|x| + \lambda_1)^{\lambda_2} - \lambda_1^{\lambda_2}]sgn(x)$$

for $0 < \lambda_2 \leq 1$ and $\lambda_1 \geq 0$ where $sgn(x) = -1, 0, 1$ when $x < 0, x = 0, x > 0$ respectively. For a fixed λ_1 in the vicinity of 0, the decreasing of λ_2 strenthens the reducing of the peakedness and the lightening of the tail and the increasing of λ_2 makes $\tilde{g}_{\lambda_1,\lambda_2}$ close to the identity transformation through the center region and shoulders and only the tail region is affected by the transformation. When $\lambda_2 = 1$, $\tilde{g}_{\lambda_1,\lambda_2}$ has no effect on distributional shape, since it is the identity transformation. For a fixed λ_2, the increasing of λ_1 makes $\tilde{g}_{\lambda_1,\lambda_2}$ close to the linear transformation which reduces $Y = X$.

4. Simulation Results

For complete automatic implementation of our transformed kernel estimator we require reliable choice of the scale estimators $\hat{\sigma}_X$ and $\hat{\sigma}_{\tilde{Y}}$, the transformation parameters λ_1,λ_2 and the bandwidth h. For the scale estimators we use the sample standard deviations.

Ideally we would like to choose (λ_1, λ_2) to minimize (2.3). However, since that quantity is unknown we instead choose (λ_1, λ_2) to minimize the diagonals-in kernel estimator of $J_Y(\lambda_1, \lambda_2)$ (Jones and Sheather(1991));

$$\hat{J}_Y(\lambda_1, \lambda_2) = [n^{-2}a^{-5}\sum_{i=1}^{n}\sum_{j=1}^{n}K^{(4)}(a^{-1}(Y_i - Y_j))]^{\frac{1}{5}}, \tag{3.1}$$

where K is the Gaussin kernel and the bandwidth a is chosen as

$$a = 1.241\hat{\sigma}_Y n^{-\frac{1}{7}}$$

where $\hat{\sigma}_Y$ is the sample standard deviation of $Y_1, \cdots, Y_n$.

In the examples below, $\hat{J}_Y(\lambda_1, \lambda_2)$ was minimized over the grid

$$\lambda_1 = \frac{i}{10}, i = 1, \cdots, 10$$

$$\lambda_2 = \frac{j}{20}, j = 1, \cdots, 20.$$

Once we have chosen (λ_1, λ_2) we propose the plug-in bandwidth (Park and Marron(1990)),

$$h_{PI} = \left[\frac{\int K(x)^2 dx}{(\int x^2 K(x)dx)^2 \hat{J}_Y(\lambda_1, \lambda_2)}\right]^{\frac{1}{5}} n^{-\frac{1}{5}}$$

to use in the kernel estimator (2.1).

For the simulation study, the 50,100 and 200 samples each of size 100, 100 and 50 respectively were generated by $IMSL$ at the following population densities;

(i) Standard Normal density

(ii) Standard Cauchy

(iii) $\frac{2}{3}N(0,1) + \frac{1}{3}N(0,(0.1)^2)$

We used the standard normal density as the population density to see how the transformed kernel density estimator performs for the case where the kurtosis is not high. The density (ii) and (iii) are used for testing the effectiveness of the proposed transformed kernel estimator in the high kurtotic density.

We estimate

$$MISE_{TR} = E[ISE_{TR}] \tag{3.2}$$

by averaging out the integrated squared errors,

$$ISE_{TR} = \int [\hat{f}_X(x; \lambda_1, \lambda_2) - f_X(x)]^2 dx. \tag{3.3}$$

These values are compared with Monte Carlo estimator of

$$MISE_{UN} = E[ISE_{UN}] \tag{3.4}$$

where

$$ISE_{UN} = \int [\hat{f}_X(x) - f_X(x)]^2 dx, \tag{3.5}$$

and $\hat{f}_X(x)$ is given in (1.1) and its bandwidth h_X is chosen as

$$h_X = (0.77639/\hat{J}_X)n^{-\frac{1}{5}}$$

where

$$\hat{J}_X = [n^{-2}a^{-5}\sum_{i=1}^{n}\sum_{j=1}^{n}\phi^{(4)}(a^{-1}(X_i - X_j))]^{\frac{1}{5}},$$

where a is chosen as

$$a = 1.241\hat{\sigma}_X n^{-\frac{1}{7}}.$$

and $\hat{\sigma}_X$ is the sample standard deviation of the original sample $X_1, \cdots, X_n$. The integrals in (3.3) and (3.5) are approximated by the composite rectangle rule,

Note that $MISE_{TR}$ and $MISE_{UN}$ in (3.2) and (3.4) are the mean integrated squared errors of the transformed and the untransformed kernel density estimators respectively. Let $\widehat{MISE}_{TR}$ and $\widehat{MISE}_{UN}$ be the Monte Carlo estimators of $MISE_{TR}$ and $MISE_{UN}$ respectively. To give some understanding for how much accuracy there is in these values, we constructed 95% confidence intervals for the difference between the $MISE's$, $(MISE_{UN} - MISE_{TR})$ using the differences between the two $ISE's$. The confidence intervals are given in tables. For further comparison of the two estimators we computed $RATIO = ISE_{TR}/ISE_{UN}$. Tables also give the 95% confidence intervals for the expected value of $RATIO$ and the proportions of case where $RATIO$ is less than or equal to 1.

From the tables we can see the transformed estimators significantly dominate the untransformed estimators except the standard normal case. Table 1 reveals that our proposed transformed kernel estimators perform very well in Cauchy density. As sample sizes growing up, the better performances are shown. It is observed that the transformed estimator had a larger ISE in 7 and 4 out of 100 samples when

the sample size is 50 and 100 respectively and had no larger ISE out of 50 samples when the sample size is 200.

We also see from Table 2 that the performance of the transformed kernel estimator is superior to the untransformed kernel estimator in normal mixture case. Among samples, it had no larger ISE than the untransformed kernel estimator in any sample size.

As expected, the transformed estimators show a little worse performance than the untransformed estomator in the standard normal case. The transformed estimator had a larger ISE in 37, 43 out of 100 samples when the sample size is 50 and 100 respectively. When the sample size is 200 it had a larger ISE in 19 out of 50 samples.

Table 1. In the Cauchy case

sample size	# of samples	95% conf. int. for ($MISE_{UN} - MISE_{TR}$)	95% conf. int. for $E(RATIO)$	# of $RATIO \leq 1$
50	100	(0.035,0.051)	(0.271,0.444)	93
100	100	(0.049,0.065)	(0.155,0.309)	96
200	50	(0.057,0.082)	(0.075,0.123)	50

Table 2. In the Normal Mixture case

sample size	# of samples	95% conf. int. for ($MISE_{UN} - MISE_{TR}$)	95% conf. int. for $E(RATIO)$	# of $RATIO \leq 1$
50	100	(0.104,0.115)	(0.257,0.311)	100
100	100	(0.090,0.098)	(0.225,0.263)	100
200	50	(0.071,0.079)	(0.217,0.247)	50

Table 3. In the Standard Normal case

sample size	# of samples	95% conf. int. for $(MISE_{UN} - MISE_{TR}) \times 10^2$	95% conf. int. for $E(RATIO)$	# of $RATIO \leq 1$
50	100	(-0.386,-0.109)	(1.214,1.601)	63
100	100	(-0.147,-0.038)	(1.160,1.381)	57
200	50	(-0.064,-0.003)	(1.046,1.496)	31

References

1. Abramson, I. S. (1982). On bandwidth variation in kernel estimates - a squared root law. *Annals of Statistics*, 9, 168-176.
2. Breiman, L. Meisel, W. and Purcell, E. (1977). Variable kernel estimates of probability density estimates. *Technometrics*, 19, 135-144.
3. Jones, M. C. and Sheather, S. (1991). Using nonstochatic terms to advantage in kernel-density estimation of integrated squared density derivatives. *Statistics and Probability Letters*, to apper.
4. Park, B. U. and Marron, J. S. (1990). Comparison of data driven bandwidth selectors. *Journal of the American Statistical Association*, 85, 66-72.
5. Ruppert, D. (1987). What is kurtosis?: an influence function approach. *The American Statistician*, 41, 1-5.
6. Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis.* New York: Chapman and Hall.
7. van Zwet, W. R. (1964). *Convex Transformation of Random Variables.* Amsterdam: Mathematisch Centrum.
8. Wand, M. P., Marron, J. S. and Ruppert, D. (1991). Transformation in density estimation, with discussion. *Journal of the American Statistical Association*, to appear.