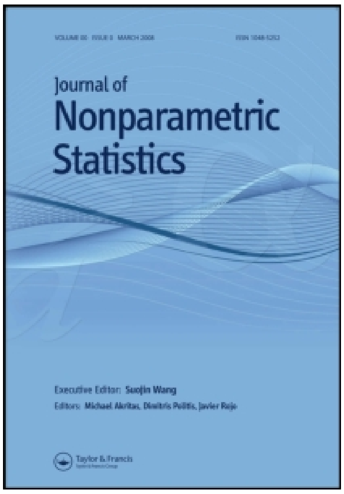

Journal of Nonparametric Statistics
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713645758

Root-n convergent transformation-kernel density estimation

Lijian Yang [a]
[a] Department of Statistics and Probability, Michigan State University, East Lansing, Michigan, USA

Online Publication Date: 01 January 2000

To cite this Article Yang, Lijian(2000)'Root-n convergent transformation-kernel density estimation',Journal of Nonparametric Statistics,12:4,447 — 474

To link to this Article: DOI: 10.1080/10485250008832818

URL: http://dx.doi.org/10.1080/10485250008832818

Nonparametric Statistics, Vol. 12, pp. 447–474
Reprints available directly from the publisher
Photocopying permitted by license only

ROOT-*n* CONVERGENT TRANSFORMATION-KERNEL DENSITY ESTIMATION

LIJIAN YANG*

Department of Statistics and Probability, Michigan State University, East Lansing, Michigan 48824, USA

(Received 3 September 1997; In final form 23 March 1999)

Transformation from a parametrized family can be combined with kernel density estimation for improved effectiveness. Pilot estimators had been proposed for the parameter that gives the optimal transformation, yet their rates of convergence had not been resolved. In this paper, the rates of convergence are given. An improved estimator is also proposed which achieves the desirable root-n rate of convergence.

Keywords and Phrases: Global bandwidth; improved estimator; optimal parameter; pilot bandwidth; vector parameter

1. INTRODUCTION

The kernel density estimator (KDE) has been a very useful tool in estimating probability density function [Scott (1992); Silverman (1986) and Wand and Jones (1995)]. Let f_X be such a function and $X_1, \ldots, X_n$ an i.i.d. random sample, the KDE of f_X is

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

whre $K_h(u) = h^{-1}K(u/h)$. The kernel K used here is a symmetric probability density, $h > 0$ is the bandwidth. When f_X has sharp features,

*e-mail: yang@stt.msu.edu

such as high skewness or kurtosis, the KDE very often is unable to capture these features. This is due to the use of a global bandwidth because different amount of smoothing is needed at different locations. Devroye and Györfi (1985), Silverman (1986) proposed to first transform the dataset and then use the KDE. Recent works in this direction include: Wand, Marron and Ruppert (1991), Park, Chung and Seog (1992), Ruppert and Wand (1992), Ruppert and Cline (1994). Algorithms had been developed by Wand, Marron and Ruppert (1991), Ruppert and Wand (1992), the dissertation of Yang (1995) and Yang and Marron (1997).

Wand, Marron and Ruppert (1991) proposed selection of a transformation from a parametric family $\{g_\lambda\}$, where $\lambda \in \Lambda$, a compact interval. Each g_λ is a strictly increasing transformation well-defined on $S(f_X)$, the support of f_X, and each transforms the density f_X to a new density

$$f_Y(y,\lambda) = f_X\{g_\lambda^{-1}(y)\}(g_\lambda^{-1})'(y). \tag{1.2}$$

From this family $\{f_Y(y,\lambda)\}_{\lambda\in\Lambda}$ of densities, we would select the one easiest to estimate with a global bandwidth. We need a functional $G(\cdot)$ of density functions that measures the difficulty of estimation with a global bandwidth. It should be scale-invariant since every density remains as easy to estimate under rescaling. Two such functionals were considered in Wand and Devroye (1993), each measures how well the kernel estimator converges to the true density in the L^2 and L^1 norms respectively. For convenience of obtaining asymptotic results and implementing the algorithm, we use the functional for L^2-theory, which is $G(f_X) = \sigma(f_X)(R(f_X''))^{1/5}$, where $\sigma(f_X)^2$ is the variance of the distribution whose density is f_X, and $R(\psi) = \int(\psi)^2$ for any function $\psi \in L^2(R)$. One can verify that $G(f_X)$ is invariant under rescaling. Small $G(f_X)$ entails a large scale free optimal bandwidth h_*, which results in small L^2 error for estimating f_X with $\widehat{f}_X$ [Jones, Marron and Sheather (1992)], *i.e.*, easier estimation. Geometrically speaking, $G(f_X)$ is a global measure of the curvature of the density, and less curvature makes estimation easier.

Our transformation-kernel density estimation (TKDE) method attempts to reduce $G(\cdot)$ as much as possible by transforming f_X. For simplicity of notation, the $\{g_\lambda\}_{\lambda\in\Lambda}$ are scaled so that $Y_i = g_\lambda(X_i)$ has the same variance as $X_i, i = 1,2,\ldots,n$, $\forall\lambda\in\Lambda$. The variables $Y_1,\ldots,Y_n$

are i.i.d. and each has the same density function f_Y given in (1.2). The following target function of $\lambda \in \Lambda$ is to be minimized

$$L(\lambda) = G(f_Y(\cdot, \lambda)) = \sigma(f_Y(\cdot, \lambda))(R(f''_Y(\cdot, \lambda)))^{1/5}.$$

But because $\sigma(f_Y(\cdot, \lambda)) \equiv \sigma(f_X), \forall \lambda \in \Lambda$, we use the following modified function instead

$$L(\lambda) = R(f''_Y(\cdot, \lambda)) = \int [f''_Y(y, \lambda)]^2 dy. \tag{1.3}$$

Suppose that $L(\lambda)$ is minimized at λ_0. If λ_0 is known, we can construct the kernel density estimate of $f_Y(\cdot, \lambda_0)$

$$\widehat{f}_Y(y, h_*, \lambda_0) = 1/n \sum_{j=1}^{n} \varphi_{h_*}(y - Y_j), \tag{1.4}$$

and the following estimate of f_X

$$\widehat{f}_X(x, h_*, \lambda_0) = n^{-1} \sum_{j=1}^{n} g'_{\lambda_0}(x) \varphi_{h_*}[g_{\lambda_0}(x) - g_{\lambda_0}(X_j)] \tag{1.5}$$

where we use the notation φ to denote the Gaussian kernel, *i.e.*, $\varphi(u) = (1/\sqrt{2\pi})e^{-u^2/2}, \varphi_h(u) = h^{-1}\varphi(u/h)$ and h_* is the optimal bandwidth. The theoretical value λ_0, however, is unknown because its definition involves the unknown density f_X. So we use $\widehat{\lambda}$, the minimizer of $\widehat{L}(\lambda) = \int [\widehat{f}''_Y(y, b, \lambda)^2] dy$ as an estimator of λ_0, in which

$$\widehat{f}''_Y(y, b, \lambda) = \frac{d^2}{dy^2} \widehat{f}_Y(y, b, \lambda) = 1/n \sum_{j=1}^{n} \varphi_b^{(2)}(y - Y_j)$$

and $\varphi_b^{(k)}(u) = b^{-(k+1)} \varphi^{(k)}(u/b)$. Here b is a pilot bandwidth used solely for $\widehat{\lambda}$, not to be confused with the bandwidth h_*.

Because Λ is compact and $\widehat{L}(\lambda)$ is continuous in $\lambda, \widehat{\lambda}$ clearly exists under assumptions A1–A4 given below. We do not need $\widehat{\lambda}$ to be unique. The assumptions are:

- A1: that λ_0 is the unique value that minimizes $L(\lambda)$;
- A2: that $L(\lambda)$ is locally convex at λ_0, *i.e.*, $\forall \varepsilon > 0, \exists D(\varepsilon) > 0$, such that for $\lambda \in \Lambda$, $L(\lambda) - L(\lambda_0) > D(\varepsilon)$ when $|\lambda - \lambda_0| > \varepsilon$;
- A3: that g_λ is jointly C^∞ for $x \in S(f_X)$ and $\lambda \in \Lambda$;

- A4: that $f_Y \in S(R)$, the space of rapidly decreasing functions, for every $\lambda \in \Lambda$. This means that f_Y is C^∞ and that for any $k = 1, 2, 3, \ldots$ and $l = 0, 1, 2, 3, \ldots$, $\lim_{|y| \to \infty} |y|^k f_Y^{(l)}(y) = 0$.

The uniqueness of λ_0 as minimizing point of the target function entails that the asymptotic theory is well-defined. When there are several minimizers, let λ_0 be the smallest of all the λ values that minimize $L(\lambda)$ and change λ_M to any number between λ_0 and the next smallest minimizer, then λ_0 is the unique minimizer of $L(\lambda)$ in the new shrunken interval $\Lambda = [0, \lambda_M]$. The local convexity condition A2 is met, for example, when $L''(\lambda_0) > 0$, which holds for most useful families of transformations. We could also weaken the assumptions A3 and A4, *e.g.*, all functions need to be only 4-th order smooth ($\in C^4(R)$). The term "rapidly decreasing" is from the theory of generalized functions, it means that the function, together with all of its derivatives, converge to zero at infinity with a rate faster than any rational function. Exponential and normal mixture densities are rapidly decreasing. It is possible to relax the condition that f_Y be rapidly decreasing: one only needs $\lim_{|y| \to \infty} |y|^k f_Y^{(l)}(y) = 0$ for $l = 0, 1, 2, 3, 4$. To highlight new ideas, however, we have not striven for generality.

This paper addresses the estimation of λ_0 by $\hat{\lambda}$ and other proposed estimators. Section 2 contains the basic asymptotic results on $\hat{\lambda}$, which are crucial for the Wand-Marron-Ruppert (WMR) transformation approach. Although the WMR method (by using $\hat{\lambda}$ in place of λ_0) had been known for its good practical performance, the first theoretical result that it works is only given here. Also provided in Section 2 is an improved version of $\hat{\lambda}$ which achieves the surprisingly good $\sqrt{n}$ rate of convergence, much superior to the slow convergence rate of the WMR estimator $\hat{\lambda}$. This section concludes with a discussion of what these asymptotics mean in some simulation settings. Section 3 extends the theory to vector parameter case. Such extension is important as it makes possible the application of transformations from richer and more flexible families. Proofs of the main results are given in Section 4.

2. ASYMPTOTICS AND AN IMPROVED ESTIMATOR

In this section we study the estimation of λ_0 under the assumptions A1–A4 of Section 1.

THEOREM 1 *Let $B'_n = [n^{(-1+\delta)/5}, n^{-\delta}]$ for some $\delta \in (0, 1/6)$ and let the pilot bandwidth used in defining $\widehat{\lambda}$ be $b \in B'_n$, then $\widehat{\lambda}$ estimates λ_0 consistently, i.e.,*

$$\sup_{b \in B'_n} |\widehat{\lambda} - \lambda_0| \to 0 \text{ a.s. as } n \to \infty. \tag{2.1}$$

Moreover, if $\lambda_0 \in$ interior of Λ, then for n large enough, $\widehat{\lambda} \in$ interior of Λ also and therefore $\widehat{L}'(\widehat{\lambda}) = 0$.

Now assume that $\lambda_0 \in$ interior of Λ, then $\widehat{L}'(\widehat{\lambda}) = 0$ gives

$$0 = \widehat{L}'(\lambda_0) + (\widehat{\lambda} - \lambda_0)\widehat{L}''(\xi), \tag{2.2}$$

where ξ is a value between λ_0 and $\widehat{\lambda}$. We study the limiting distributions of $\widehat{L}'(\lambda_0)$ and $\widehat{L}''(\xi)$ respectively to decide that of $(\widehat{\lambda} - \lambda_0)$.

The following is the main result on the limiting distribution of $\widehat{\lambda}$:

THEOREM 2 *Suppose that $\lambda_0 \in$ interior of Λ, $b \in B'_n = [n^{(-1+\delta)/5}, n^{-\delta}]$ for some $\delta \in (0, 1/6)$. Under the assumptions $A1 - A4$, and as $n \to \infty$,*

$$\widehat{\lambda} = \lambda_0 + Bias(n, b) + Z,$$

in which the bias term is the following

$$Bias(n, b) = C(f_X)b^2 + O(b^4)$$

with coefficient

$$C(f_X) = \frac{\frac{d}{d\lambda}\Big|_{\lambda_0} \int (f_Y^{(3)}(y))^2 dy}{\frac{d^2}{d\lambda^2}\Big|_{\lambda_0} \int (f_Y^{(2)}(y))^2 dy} = \frac{\frac{d}{d\lambda}\Big|_{\lambda_0} \int (f_Y^{(3)}(y))^2 dy}{L''(\lambda_0)} \tag{2.3}$$

and where the variance term Z satisfies

$$\sqrt{n + n^2 b^9} Z \to N(0, \sigma^2(f_X))$$

for some functional $\sigma^2(f_X)$.

The estimator $\widehat{\lambda}$, therefore, has bias term of order $O(b^2)$ and a variance term of order $(1/\sqrt{n+n^2b^9})$, its rate of convergence much slower than the usually expected rate of $(1/\sqrt{n})$. To obtain an estimator with the $(1/\sqrt{n})$ rate, we use a pilot bandwidth b_V that makes the variance of $\widehat{\lambda}$ of order $O(1/n)$, which is done by requiring that $(1/nb_V^9) = O(1)$ as $n \to \infty$, or $b_V \in [n^{(-1+\delta)/9}, n^{-\delta}]$ for some $\delta \in (0, 1/10)$. This b_V gives a bias converging very slowly (at the rate of about $b^2 \propto n^{-(2/9)}$), but a variance that is of order $(1/n)$. This feature frees one from having to increase the convergence rates for both the bias and the variance terms at the same time, one only needs to correct the bias term in $\widehat{\lambda}$. Next we describe how the convergence rate of $\widehat{\lambda}$ can be improved in two steps. Throughout the rest of this section, the pilot bandwidth b_V is used and abbreviated simply as b.

Note that Eq. (2.2) can be written as

$$\widehat{\lambda} = \lambda_0 - \frac{\widehat{L}'(\lambda_0)}{\widehat{L}''(\xi)},$$

where ξ is in between $\widehat{\lambda}$ and λ_0, and ξ tends to λ_0 because of the consistency of $\widehat{\lambda}$. Thus asymptotically

$$\widehat{\lambda} \approx \lambda_0 - \frac{\widehat{L}'(\lambda_0)}{\widehat{L}''(\lambda_0)} \approx \lambda_0 - \frac{\widehat{L}'(\lambda_0)}{L''(\lambda_0)},$$

where $\widehat{L}'(\lambda_0)$ has a bias term of the form $[(d/d\lambda)|_{\lambda_0} \int (f_Y^{(3)}(y))^2 dy]$ $b^2 + O(b^4)$ which gives $\widehat{\lambda}$ a bias term of the form $((d/d\lambda)|_{\lambda_0}$ $\int (f_Y^{(3)}(y))^2 dy)(L''(\lambda_0))^{-1} b^2 + O(b^4)$ as in Theorem 2. This suggests modifying the pilot estimator $\widehat{L}(\lambda)$ to

$$\widehat{L}(\lambda) - \lambda \left[\frac{d}{d\lambda}\bigg|_{\lambda_0} \int (f_Y^{(3)}(y))^2 dy \right] b^2.$$

This modification, however, makes use of the unknown λ_0 and the unknown density $f_Y|_{\lambda_0}$, *etc.*, which need to be replaced by their pilot estimates. Hence we modify the pilot estimate of $L(\lambda)$ by introducing the following:

$$\widehat{L}_1(\lambda) = \widehat{L}(\lambda) - \lambda c_1(\widehat{f}_X, \widehat{\lambda}) b^2, \tag{2.4}$$

where the coefficient

$$c_1(\widehat{f}_X,\widehat{\lambda}) = -\frac{d}{d\lambda}\bigg|_{\lambda=\widehat{\lambda}} \int (\widehat{f}_Y^{(3)}(y,b))^2 dy \tag{2.5}$$

is evaluated at $\lambda = \widehat{\lambda}$. Let $\widehat{\lambda}_1$ be the minimizer of $\widehat{L}_1(\lambda)$, it has better convergence rate than $\widehat{\lambda}$. But because the modification of $\widehat{L}(\lambda)$ to $\widehat{L}_1(\lambda)$ was made using pilot estimates, the bias term of $\widehat{\lambda}_1$ is still not of order $(1/\sqrt{n})$, it actually is of order $b^4 \propto n^{-(4/9)}$. Hence we introduce yet another modification:

$$\widehat{L}_2(\lambda) = \widehat{L}(\lambda) - \lambda(c_1(\widehat{f}_X,\widehat{\lambda}_1)b^2 + 4c_2(\widehat{f}_X,\widehat{\lambda}_1)b^4),$$

where

$$c_1(\widehat{f}_X,\widehat{\lambda}_1) = -\frac{d}{d\lambda}\bigg|_{\lambda=\widehat{\lambda}_1} \int (\widehat{f}_Y^{(3)}(y,b))^2 dy,$$

$$c_2(\widehat{f}_X,\widehat{\lambda}_1) = \frac{1}{3}\frac{d}{d\lambda}\bigg|_{\lambda=\widehat{\lambda}_1} \int (\widehat{f}_Y^{(4)}(y,b))^2 dy,$$

respectively. Let $\widehat{\lambda}_2$ be the minimizer of this $\widehat{L}_2(\lambda)$, then we have:

THEOREM 3 *Suppose that* $\lambda_0 \in$ *interior of* Λ, $b = b_V$. *Under the assumptions* $A1 - A4$, *and as* $n \to \infty$

$$\sqrt{n}(\widehat{\lambda}_2 - \lambda_0) \to N(0, \sigma_0^2(f_X))$$

where $\sigma_0^2(f_X)$ *is an implicitly defined functional of* f_X. *Thus* $\widehat{\lambda}_2$ *has the desired* $(1/\sqrt{n})$ *rate of convergence.*

In summary, an improved version of $\widehat{\lambda}$, which is obtained by correcting the estimator $\widehat{\lambda}$ in two steps, has the fast parametric convergence rate. This superior asymptotic property of $\widehat{\lambda}_2$ should be viewed in the larger context of the estimation problem at hand. An important issue is how strongly the asymptotic lessons apply for reasonable sample sizes. Simulation showed that the lessons of the asymptotics did not take effect up to a sample size of 10000. For reasonable sample sizes there is not a significant improvement from $\widehat{\lambda}_2$ over some data-driven estimators of slower convergence rate. This phenomenon is expected. it is similar to the performance of a $\sqrt{n}$ bandwidth selector given in Kim, Park and Marron (1994). Another issue is that $\widehat{\lambda}_2$ is not automatic, because the pilot bandwidth b_V is only required to be of a certain order. This means a lot of uncertainty if one wants to develop an algorithm of TKDE using $\widehat{\lambda}_2$ as estimator of λ. For these considerations, $\widehat{\lambda}_2$ had not yet been used in simulation work and applications. At least at this point, $\widehat{\lambda}_2$ is of primarily theoretical interest. It is conceivable, however, that one come up with a suitable pilot bandwidth b_V. In that case, one can take the full advantage of $\widehat{\lambda}_2$'s superior rate.

3. VECTOR PARAMETER CASE

It is worthwhile to extend all the results of the previous section to vector parameter. Such extensions would allow the use of transformation families with vector parameters that are able to effectively transform wider variety of density shapes [for example, the Johnson family and the g-and-h family in MacGillivray (1992), and the shifted power family in Wand Marron and Ruppert (1991)].

The assumptions need to be modified. The parameter becomes a vector $\lambda \in \Lambda$, a p-dimensional rectangle of the form $\Lambda_1 \times \Lambda_2 \times \cdots \times \Lambda_p$, where each Λ_j is a compact interval. Let λ_0 be the minimizer of $L(\lambda)$. The following are assumed

- A1′: that λ_0 is the unique value that minimizes $L(\lambda)$;
- A2′: that $L(\lambda)$ is locally convex at λ_0, *i.e.*, $\forall \varepsilon > 0, \exists D(\varepsilon) > 0$, such that for $\lambda \in \Lambda$, $L(\lambda) - L(\lambda_0) > D(\varepsilon)$ when $|\lambda - \lambda_0| > \varepsilon$. where $|\lambda - \lambda_0|$ is the p-dimensional Euclidean norm;
- A3′: that g_λ is jointly C^∞ for $x \in S(f_X)$ and $\lambda \in \Lambda$;
- A4′: that $f_Y \in \mathcal{S}(R)$, the space of rapidly decreasing functions, for every $\lambda \in \Lambda$. This means that f_Y is C^∞ and that for any $k = 1, 2, 3, \ldots$ and $l = 0, 1, 2, 3, \ldots$, $\lim_{|y| \to \infty} |y|^k f_Y^{(l)}(y) = 0$.

We make the usual convention that λ represents a p-dimensional column vector. We use $(\partial/\partial\lambda)$ to represent the gradient operator, therefore, if F is a function of λ, then

$$\frac{\partial}{\partial\lambda}F(\lambda) = \left(\frac{\partial}{\partial\lambda_1}F(\lambda)\ \frac{\partial}{\partial\lambda_2}F(\lambda)\ \cdots\ \frac{\partial}{\partial\lambda_p}F(\lambda)\right)'$$

Similarly, we use $(\partial^2/\partial\lambda^2)$ to represent the Hessian operator, *i.e.*,

$$\frac{\partial^2}{\partial\lambda^2}F(\lambda) = \begin{bmatrix} \frac{\partial^2}{\partial\lambda_1\partial\lambda_1}F(\lambda) & \frac{\partial^2}{\partial\lambda_1\partial\lambda_2}F(\lambda) & \cdots & \frac{\partial^2}{\partial\lambda_1\partial\lambda_p}F(\lambda) \\ \frac{\partial^2}{\partial\lambda_2\partial\lambda_1}F(\lambda) & \frac{\partial^2}{\partial\lambda_2\partial\lambda_2}F(\lambda) & \cdots & \frac{\partial^2}{\partial\lambda_2\partial\lambda_p}F(\lambda) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial\lambda_p\partial\lambda_1}F(\lambda) & \frac{\partial^2}{\partial\lambda_p\partial\lambda_2}F(\lambda) & \cdots & \frac{\partial^2}{\partial\lambda_p\partial\lambda_p}F(\lambda) \end{bmatrix}$$

Theorems 1 and 2 may be extended to

PROPOSITION 3.1 *Let* $B'_n = [n^{(-1+\delta)/5}, n^{-\delta}]$ *for some* $\delta \in (0, 1/6)$ *and let* $\widehat{\lambda}$ *be the minimizer of* $\widehat{L}(\lambda)$ *in* Λ *using* $b \in B'_n$, *then* $\widehat{\lambda}$ *estimates* λ_0 *consistently, i.e.,* $sup_{b \in B'_n}|\widehat{\lambda} - \lambda_0| \to 0$ *a.s.*

PROPOSITION 3.2 *Suppose that* $\lambda_0 \in$ *interior of* Λ, $b \in B'_n = [n^{(-1+\delta)/5}, n^{-\delta}]$ *for some* $\delta \in (0, 1/6)$. *Under the assumptions* $A1' - A4'$, *as* $n \to \infty$

$$\widehat{\lambda} = \lambda_0 + Bias(n, b) + \boldsymbol{Z}$$

in which the bias term is the following

$$Bias(n, b) = \boldsymbol{C}(\boldsymbol{f_X})b^2 + O(b^4)$$

with coefficient

$$\begin{aligned} \boldsymbol{C}(\boldsymbol{f_X}) &= \left[\frac{\partial^2}{\partial\lambda^2}\bigg|_{\lambda_0} \int (f_Y^{(2)}(y))^2 dy\right]^{-1} \left[\frac{\partial}{\partial\lambda}\bigg|_{\lambda_0} \int (f_Y^{(3)}(y))^2 dy\right] \\ &= \left[\frac{\partial^2}{\partial\lambda^2}\bigg|_{\lambda_0} L(\lambda)\right]^{-1} \left[\frac{\partial}{\partial\lambda}\bigg|_{\lambda_0} \int (f_Y^{(3)}(y))^2 dy\right] \end{aligned}$$

and where the variance term Z satisfies

$$\sqrt{n + n^2 b^9}Z \to N(0, \Sigma(f_X))$$

for some $p \times p$ *functional matrix* $\Sigma(f_X)$.

Now let $\widehat{\lambda}$ denote the estimator with $b = b_V$, the pilot estimate of $L(\lambda)$ is modified as

$$\widehat{L}_1(\lambda) = \widehat{L}(\lambda) - \lambda \cdot c_1(\widehat{f}_X, \widehat{\lambda})b^2,$$

where the coefficient

$$c_1(\widehat{f}_X, \widehat{\lambda}) = -\frac{\partial}{\partial\lambda}\bigg|_{\lambda=\widehat{\lambda}} \int (\widehat{f}_Y^{(3)}(y,b))^2 dy$$

is the gradient vector of $-\int(\widehat{f}_Y^{(3)}(y,b))^2 dy$ with respect to λ evaluated at $\lambda = \widehat{\lambda}$, and $\lambda \cdot c_1(\widehat{f}_X, \widehat{\lambda})$ is the inner product of the two p-dimensional vectors. Let $\widehat{\lambda}_1$ be the minimizer of $\widehat{L}_1(\lambda)$, then we introduce yet another modification:

$$\widehat{L}_2(\lambda) = \widehat{L}(\lambda) - \lambda \cdot (c_1(\widehat{f}_X, \widehat{\lambda}_1)b^2 + 4c_2(\widehat{f}_X, \widehat{\lambda}_1)b^4),$$

where

$$c_1(\widehat{f}_X, \widehat{\lambda}_1) = -\frac{\partial}{\partial\lambda}\bigg|_{\lambda=\widehat{\lambda}_1} \int (\widehat{f}_Y^{(3)}(y,b))^2 dy, c_2(\widehat{f}_X, \widehat{\lambda}_1)$$
$$= \frac{1}{3}\frac{\partial}{\partial\lambda}\bigg|_{\lambda=\widehat{\lambda}_1} \int (\widehat{f}_Y^{(4)}(y,b))^2 dy,$$

respectively. Let $\widehat{\lambda}_2$ be the minimizer of this $\widehat{L}_2(\lambda)$. Theorem 3 may be extended to

PROPOSITION 3.3 *Suppose that* $\lambda_0 \in$ *interior of* $\Lambda, b = b_V$. *Under the assumptions* $A1' - A4'$, *and as* $n \to \infty$:

$$\sqrt{n}(\widehat{\lambda}_2 - \lambda_0) \to N(0, \Sigma_0(f_X))$$

where $\Sigma_0(f_X)$ *is an implicitly defined matrix functional of* f_X.

4. PROOFS

This section contains the proof of Theorems 1, 2, and 3. To prove consistency of $\widehat{\lambda}$, our method is through the Borel-Cantelli Lemma for $|\widehat{\lambda} - \lambda_0|$, similar to Härdle and Marron (1990). We assume in this section that $b \in B'_n = [n^{(-1+\delta)/5}, n^{-\delta}]$ for some $\delta \in (0, 1/6)$, hence $nb^5 \to \infty$. The proof of the following lemma is found on page 136 of Yang (1995).

LEMMA 4.1 *For all* $\varepsilon > 0, \sum_{n=1}^{\infty} P[\sup_{\lambda\in\Lambda} \sup_{b\in B'_n} |\widehat{L}(\lambda) - L(\lambda)| > \varepsilon] < \infty$

Proof of Theorem 1 For all $\varepsilon > 0$, we see from local convexity of L at λ_0:

$$\sum_{n=1}^{\infty} P\left[\sup_{b\in B'_n} |\widehat{\lambda} - \lambda_0| > \varepsilon\right] \le \sum_{n=1}^{\infty} P\left[\sup_{b\in B'_n} (L(\widehat{\lambda}) - L(\lambda_0)) > D(\varepsilon)\right].$$

Now notice that $\widehat{L}(\lambda_0) - \widehat{L}(\widehat{\lambda}) \ge 0$ because $\widehat{L}$ is minimized at $\widehat{\lambda}$, therefore:

$$\begin{aligned}
&\sum_{n=1}^{\infty} P\left[\sup_{b\in B'_n} (L(\widehat{\lambda}) - L(\lambda_0)) > D(\varepsilon)\right] \\
&\quad \le \sum_{n=1}^{\infty} P\left[\sup_{b\in B'_n} (L(\hat{\lambda}) - \widehat{L}(\widehat{\lambda}) + \widehat{L}(\lambda_0) - L(\lambda_0)) > D(\varepsilon)\right]. \\
&\quad \le \sum_{n=1}^{\infty} P\left[\sup_{b\in B'_n} |L(\widehat{\lambda}) - \widehat{L}(\widehat{\lambda})| > D(\varepsilon)/2\right] \\
&\qquad + \sum_{n=1}^{\infty} P\left[\sup_{b\in B'_n} |\widehat{L}(\lambda_0) - L(\lambda_0)| > D(\varepsilon)/2\right] \\
&\quad \le 2\sum_{n=1}^{\infty} P\left[\sup_{\lambda\in\Lambda}\sup_{b\in B'_n} |\widehat{L}(\lambda) - L(\lambda)| > D(\varepsilon)/2\right] < \infty,
\end{aligned}$$

because of Lemma 4.1. Now Borel-Cantelli lemma shows the consistency. Q.E.D.

A series of auxiliary lemmas are developed on the limiting distributions of statistics of the form $(d^l/d\lambda^l)|_\lambda \int (\widehat{f}_Y^{(k)}(y))^2 dy$, where $k = 2, 3, 4$ and $l = 1, 2$. The case $l = 0$ was studied in Hall and Marron (1987), Jones and Sheather (1991). Our method here is similar but more terms will be involved by taking derivatives with respect to λ. As an illustration of how these statistics are analyzed, we begin with $k = 2$ and $l = 1$, the statistics is

$$\widehat{L}'(\lambda) = 2\int \widehat{f}''_Y(y, b, \lambda)\widehat{f}''_{Y,\lambda}(y, b, \lambda)dy$$

where

$$\begin{aligned}
\widehat{f}''_Y(y, b, \lambda) = 1/n\sum_{j=1}^{n} \varphi_b^{(2)}(y - Y_j),\ \widehat{f}''_{Y,\lambda}(y, b, \lambda) &= \frac{d}{d\lambda}\widehat{f}''_Y(y, b, \lambda) \\
&= -1/n\sum_{j=1}^{n} \varphi_b^{(3)}(y - Y_j)\frac{dg_\lambda(X_j)}{d\lambda}\bigg|_\lambda.
\end{aligned}$$

Define

$$A_n(X_i) = \varphi_b^{(2)}(y - Y_j), \quad B_n(X_j) = -\varphi_b^{(3)}(y - Y_j)\frac{dg_\lambda(X_j)}{d\lambda}\bigg|_\lambda,$$
$$a_n(b, y) = EA_n(X_1), \quad b_n(b, y) = EB_n(X_1).$$

Then rewrite $\widehat{L}'(\lambda)$ as

$$\widehat{L}'(\lambda) = C_b + 2\sum_{j=1}^{n} Z_{nj} + 2\sum_{1\le i<j\le n} H_n(X_i, X_j) + R_n, \tag{4.1}$$

where

$$C_b = \int 2a_n(b, y)b_n(b, y)dy, \tag{4.2}$$

$$Z_{nj} = \int \frac{[a_n(b,y)B_n(X_j) + b_n(b,y)A_n(X_j) - 2a_n(b,y)b_n(b,y)]}{n}dy, \tag{4.3}$$

$$H_n(X_i, X_j) = \int \frac{1}{n^2}[(A_n(X_i) - a_n(b, y))(B_n(X_j) - b_n(b, y)) + (A_n(X_j) - a_n(b, y))(B_n(X_i) - b_n(b, y))]dy, \tag{4.4}$$

$$R_n = \int 2/n^2 \sum_{j=1}^{n}(A_n(X_j) - a_n(b, y))(B_n(X_j) - b_n(b, y))dy. \tag{4.5}$$

The terms C_b, $2\sum_{j=1}^{n} Z_{nj}$, $2\sum_{1\le i<j\le n} H_n(X_i, X_j)$, and R_n are now analyzed one by one.

LEMMA 4.2 *We have the following asymptotic expressions*

$$a_n^{(k)}(b, y) = f_Y^{(k+2)}(y) + \frac{1}{2}f_Y^{(k+4)}(y)h^2 + b^4\left(\frac{1}{4!}f_Y^{(k+6)}(y) + A_k(y, b)b^2\right), \tag{4.6}$$

$$b_n^{(k)}(b, y) = f_{Y,\lambda}^{(k+2)}(y) + \frac{1}{2}f_{Y,\lambda}^{(k+4)}(y)b^2 + b^4\left(\frac{1}{4!}f_{Y,\lambda}^{(k+6)}(y) + B_k(y, b)b^2\right) \tag{4.7}$$

where the derivatives $f_Y^{(k)}(y) = (\partial^k/\partial y^k)f(y,b,\lambda)$, $f_{Y,\lambda}^{(k)}(y) = \big((d/d\lambda)(\partial^k/\partial y^k)f(y,b,\lambda)\big), k = 0,1,2,3,\ldots$ *etc., are all evaluated at* $\lambda = \lambda$, *and the remainder coefficients have the following uniform square integrability property:*

$$\sup_n \sup_{b \in B'_n} \int \{[A_k(y,b)]^2 + [B_k(y,b)]^2\}dy < \infty \tag{4.8}$$

for every fixed $k = 0,1,2,3,\ldots$.

In particular, we have

$$\begin{aligned} C_b &= b^2 \int (f_Y''(y)f_{Y,\lambda}^{(4)}(y) + f_{Y,\lambda}''(y)f_Y^{(4)}(y))dy \\ &\quad + b^4 \int \left[\frac{1}{2}f_Y^{(4)}(y)f_{Y,\lambda}^{(4)}(y) + \frac{1}{12}f_Y''(y)f_{Y,\lambda}^{(6)}(y) + \frac{1}{12}f_{Y,\lambda}''(y)f_Y^{(6)}(y)\right]dy \\ &\quad + 2\int f_Y''(y)f_{Y,\lambda}''(y)dy + O(b^6) \\ &= 2\int f_Y''(y)f_{Y,\lambda}''(y)dy + c_1(f_X)b^2 + c_2(f_X)b^4 + O(b^6), \end{aligned} \tag{4.9}$$

where

$$\begin{aligned} c_1(f_X) &= c_1(f_X,\lambda) = -\frac{d}{d\lambda}\bigg|_\lambda \int (f_Y^{(3)}(y))^2 dy, \\ c_2(f_X) &= c_2(f_X,\lambda) = \frac{1}{3}\frac{d}{d\lambda}\bigg|_\lambda \int (f_Y^{(4)}(y))^2 dy, \end{aligned} \tag{4.10}$$

hence C_b*'s bias is of order* b^2.

Proof By definition

$$a_n(b,y) = \int \varphi_b^{(2)}(y-z)f_Y(z)dz, a_n^{(k)}(b,y) = \int \varphi_b^{(k+2)}(y-z)f_Y(z)dz.$$

Repeating integration by parts $(k+2)$ times and using the substitution $z = y - bu$, one can immediately get

$$\begin{aligned} a_n^{(k)}(b,y) &= f_Y^{(k+2)}(y) + \frac{1}{2}f_Y^{(k+4)}(y)b^2 \\ &\quad + b^4\left(\frac{1}{4!}f_Y^{(k+6)}(y) + A_k(y,h)b^2\right), \end{aligned} \tag{4.11}$$

which is (4.6). Now notice that $b_n(y) = (d/d\lambda)|_\lambda a_n(y)$, hence

$$\begin{aligned} b_n^{(k)}(b,y) &= \frac{d}{d\lambda}\bigg|_\lambda \int \varphi(u) f_Y^{(k+2)}(y-bu)du \\ &= \int \varphi(u) \frac{d}{d\lambda}\bigg|_\lambda f_Y^{(k+2)}(y-bu)du \\ &= \int \varphi(u) f_{Y,\lambda}^{(k+2)}(y-bu)du. \end{aligned}$$

Then this leads to (4.7) just the same way (4.11) leads to (4.6).

The inequality (4.8) follows by examining the regularity assumptions A3–A4 in Section 1. Now using $C_b = \int 2a_n(b,y)b_n(b,y)dy$ and plugging in what had been got for $a_n(b,y)$ and $b_n(b,y)$

$$\begin{aligned} C_b = \int 2 &\left[f_Y^{(2)}(y) + \frac{1}{2} f_Y^{(4)}(y)b^2 + \frac{1}{4!} f_Y^{(6)}(y)b^4 + A_0(y,b)b^6 \right] \\ &\left[f_{Y,\lambda}^{(2)}(y) + \frac{1}{2} f_{Y,\lambda}^{(4)}(y)b^2 + \frac{1}{4!} f_{Y,\lambda}^{(6)}(y)b^4 + B_0(y,b)b^6 \right] dy, \end{aligned}$$

the uniform square integrability property in (4.8) and the above expression of C_b immediately yields (4.9). Q.E.D.

LEMMA 4.3 *For the linear terms Z_{nj}'s*

$$n\sum_{j=1}^{n} EZ_{nj}^2 = n^2 EZ_{n1}^2 \to \sigma_1^2, \quad \sum_{j=1}^{n} E|\sqrt{n}Z_{nj}|^3 \to 0 \text{ as } n \to \infty, \tag{4.12}$$

where σ_1^2 is the variance of $(d/d\lambda) f_Y^{(4)}(Y)$

$$\begin{aligned} \sigma_1^2 &= E\left[\frac{d}{d\lambda} f_Y^{(4)}(Y)\right]^2 - \left(2\int f_Y''(y) f_{Y,\lambda}''(y)dy\right)^2 \\ &= \sigma^2\left(\frac{d}{d\lambda} f_Y^{(4)}(Y)\right). \end{aligned} \tag{4.13}$$

Hence, $\sqrt{n}(2\sum_{j=1}^{n} Z_{nj}) \xrightarrow{D} N(0, 4\sigma_1^2)$, as $n \to \infty$.

Proof By the definition of Z_{nj}

$$
\begin{aligned}
n^2 E Z_{n1}^2 &= E\left\{ \int [a_n(b,y)B_n(X) + b_n(b,y)A_n(X) - 2a_n(b,y)b_n(b,y)]dy \right\}^2 \\
&= E\left\{ \int [a_n(b,y)B_n(X) + b_n(b,y)A_n(X)]dy \right\}^2 - C_b^2 \\
&= E\left\{ \int \left[- a_n(b,y)\varphi_b^{(3)}(y-Y)\frac{dg_\lambda(X)}{d\lambda}\bigg|_\lambda \right.\right. \\
&\quad \left.\left. + b_n(b,y)\varphi_b^{(2)}(y-Y)\right] dy \right\}^2 - C_b^2.
\end{aligned}
$$

Integrating by parts

$$
\begin{aligned}
n^2 E Z_{n1}^2 = E\left\{ \int \left[a_n^{(3)}(b,y)\varphi_b(y-Y)\frac{dg_\lambda(X)}{d\lambda}\bigg|_\lambda \right.\right. \\
\left.\left. + b_n^{(2)}(b,y)\varphi_b(y-Y)\right] dy \right\}^2 - C_b^2.
\end{aligned}
$$

Using substitution $y = Y + bu$

$$
\begin{aligned}
= E\left\{ \int \left[a_n^{(3)}(b, Y+bu)\varphi(u)\frac{dg_\lambda(X)}{d\lambda}\bigg|_\lambda \right.\right. \\
\left.\left. + b_n^{(2)}(b, Y+bu)\varphi(u)\right] du \right\}^2 - C_b^2.
\end{aligned}
$$

Now apply (4.6), (4.7) and (4.8)

$$
\begin{aligned}
n^2 E Z_{n1}^2 &= E\left\{ \int \varphi(u)\left[f_{Y,\lambda}^{(4)}(Y+bu) + f_Y^{(5)}(Y+bu)\frac{dg_\lambda(X)}{d\lambda}\bigg|_\lambda \right] du \right\}^2 \\
&\quad + o(b^2) - C_b^2.
\end{aligned}
$$

Then apply Taylor expansions

$$
\begin{aligned}
n^2 E Z_{n1}^2 &= E\left\{ \int \varphi(u)\left[f_{Y,\lambda}^{(4)}(Y) + f_Y^{(5)}(Y)\frac{dg_\lambda(X)}{d\lambda}\bigg|_\lambda \right] du \right\}^2 \\
&\quad + o(b^2) - C_b^2 = \sigma_1^2 + o(b^2) \to \sigma_1^2,
\end{aligned}
$$

as $n \to \infty$ because of (4.9).

Similarly

$$\sum_{j=1}^{n} E|\sqrt{n}Z_{nj}|^3 = n^{5/2}E|Z_{n1}|^3 \leq n^{5/2}(E|Z_{n1}|^4)^{3/4}$$
$$= \frac{1}{\sqrt{n}}(E|nZ_{n1}|^4)^{3/4}.$$

By (4.6), (4.7)

$$\frac{1}{\sqrt{n}}(E|nZ_{n1}|^4)^{3/4} = \frac{1}{\sqrt{n}}\left\{E\left[\int \varphi(u)\left[f_{Y,\lambda}^{(4)}(Y+bu) + f_Y^{(5)}(Y+bu)\frac{dg_\lambda(X)}{d\lambda}\bigg|_\lambda\right]du\right]^4\right\}^{3/4}.$$

Apply Taylor expansion and use (4.8)

$$\frac{1}{\sqrt{n}}(E|nZ_{n1}|^4)^{3/4} \propto \frac{1}{\sqrt{n}}.$$

We then only need to note that the Z_{nj}'s are i.i.d. for each n and that $EZ_{nj} = 0$, therefore (4.12) and array type central limit theorem [Härdle and Marron (1990)] give the asymptotic result about $\sqrt{n}(2\sum_{j=1}^{n} Z_{nj})$. Q.E.D.

PROPOSITION 4.1 *For the quadratic term* $U_n = \sum_{1\leq i<j\leq n} H_n(X_i, X_j)$

$$var(\sqrt{n}U_n) = E(nU_n^2) = \frac{1}{2}n^2(n-1)E\{H_n(X_1,X_2)\}^2 \propto \frac{1}{nb^9}\sigma^2(f_X, g_\lambda),$$

where $\sigma^2(f_X, g_\lambda)$ *is implicitly defined and depends only on* f_X *and* g_λ.

Proof According to Hall (1984), U_n is a degenerate U-statistic. Therefore we have

$$\text{var}(\sqrt{n}U_n) = E(nU_n^2) = \frac{1}{2}n^2(n-1)E\{H_n(X_1,X_2)\}^2. \quad (4.14)$$

Now rewrite the definition of $H_n(X_1, X_2)$ as

$$\begin{aligned}\int \frac{1}{n^2}[(A_n(X_1) - a_n(b,y))(B_n(X_2) - b_n(b,y)) \\ + (A_n(X_2) - a_n(b,y))(B_n(X_1) - b_n(b,y))]dy \\ = \int \frac{1}{n^2}[A_n(X_1)B_n(X_2) + A_n(X_2)B_n(X_1)]dy \\ - \frac{1}{n}(Z_{n1} + Z_{n2}) - \frac{1}{n^2}C_b. \end{aligned} \tag{4.15}$$

Since we have had information about the linear and constant terms, now we calculate the cross terms. By definition

$$\begin{aligned}\int A_n(X_1)B_n(X_2)dy &= \int \varphi_b^{(2)}(y - Y_1)\left[- \varphi_b^{(3)}(y - Y_2)\frac{dg_\lambda(X_2)}{d\lambda}\bigg|_\lambda\right]dy \\ &= \frac{dg_\lambda(X_2)}{d\lambda}\bigg|_\lambda \int \varphi_b^{(2)}(y - Y_1)\varphi_b^{(3)}(Y_2 - y)dy.\end{aligned}$$

Which is, by definition, the following convolution

$$\int [A_n(X_1)B_n(X_2)]dy = \frac{dg_\lambda(X_2)}{d\lambda}\bigg|_\lambda [\varphi_b^{(2)} * \varphi_b^{(3)}](Y_2 - Y_1).$$

Then by the convolution formula for normal densities found in Aldershof, Marron, Park and Wand (1990)

$$\int [A_n(X_1)B_n(X_2)]dy = \frac{dg_\lambda(X_2)}{d\lambda}\bigg|_\lambda \varphi_{\sqrt{2}b}^{(5)}(Y_2 - Y_1).$$

Similar result holds if we switch the roles of X_1 and X_2 in the above calculation. Therefore

$$\begin{aligned}H_n(X_1, X_2) = \frac{1}{n^2}\left[\frac{dg_\lambda(X_2)}{d\lambda}\bigg|_\lambda - \frac{dg_\lambda(X_1)}{d\lambda}\bigg|_\lambda\right]\varphi_{\sqrt{2}b}^{(5)}(Y_2 - Y_1) \\ - \frac{1}{n}(Z_{n1} + Z_{n2}) - \frac{1}{n^2}C_b.\end{aligned} \tag{4.16}$$

Now $(1/2)n^2(n-1)E\{H_n(X_1, X_2)\}^2$ is of the same order as $(1/2n)$ $E\{n^2H_n(X_1, X_2)\}^2$, or $(1/2n)E\{n^2H_n(X_1, X_2)\}^2$ and

$$\frac{1}{2n}E\{n^2H_n(X_1,X_2)\}^2$$
$$=\frac{1}{2n}E\left\{\left[\left.\frac{dg_\lambda(X_2)}{d\lambda}\right|_\lambda-\left.\frac{dg_\lambda(X_1)}{d\lambda}\right|_\lambda\right]\varphi^{(5)}_{\sqrt{2}b}(Y_2-Y_1)-n(Z_{n1}+Z_{n2})-C_b\right\}^2.$$

Note that the Z_{nj}'s are i.i.d. with $EZ^2_{nj}=O(1/n^2), EZ_{nj}=0$ and also note that $C_b=2\int f''_Y(y)f''_{Y,\lambda}(y)dy+O(b^2)$, thus

$$\begin{aligned}\frac{1}{2n}E\{n^2H_n(X_1,X_2)\}^2=&\frac{1}{2n}E\left\{\left[\left.\frac{dg_\lambda(X_2)}{d\lambda}\right|_\lambda-\left.\frac{dg_\lambda(X_1)}{d\lambda}\right|_\lambda\right]\varphi^{(5)}_{\sqrt{2}b}(Y_2-Y_1)\right\}^2\\&-2E\left\{\left[\left.\frac{dg_\lambda(X_2)}{d\lambda}\right|_\lambda-\left.\frac{dg_\lambda(X_1)}{d\lambda}\right|_\lambda\right]\varphi^{(5)}_{\sqrt{2}b}(Y_2-Y_1)Z_{n1}\right\}\\&-\frac{C_b}{n}E\left\{\left[\left.\frac{dg_\lambda(X_2)}{d\lambda}\right|_\lambda-\left.\frac{dg_\lambda(X_1)}{d\lambda}\right|_\lambda\right]\varphi^{(5)}_{\sqrt{2}b}(Y_2-Y_1)\right\}\\&+O\left(\frac{1}{n}+b^2\right).\end{aligned}\tag{4.17}$$

The first term in (4.17) can be calculated as:

$$\begin{aligned}&\frac{1}{2n}E\left\{\left[\left.\frac{dg_\lambda(X_2)}{d\lambda}\right|_\lambda-\left.\frac{dg_\lambda(X_1)}{d\lambda}\right|_\lambda\right]\varphi^{(5)}_{\sqrt{2}b}(Y_2-Y_1)\right\}^2\\&=\frac{1}{2n}\int\left\{\left[\left.\frac{dg_\lambda(g_\lambda^{-1}(y_2))}{d\lambda}\right|_\lambda-\left.\frac{dg_\lambda(g_\lambda^{-1}(y_1))}{d\lambda}\right|_\lambda\right]\right.\\&\qquad\left.\varphi^{(5)}_{\sqrt{2}b}(y_2-y_1)\right\}^2f_Y(y_1)f_Y(y_2)dy_1dy_2.\end{aligned}$$

Applying the substitution: $y_2-y_1=\sqrt{2}bu, y_2+y_1=v$

$$\begin{aligned}&\frac{1}{2n}E\left\{\left[\left.\frac{dg_\lambda(X_2)}{d\lambda}\right|_\lambda-\left.\frac{dg_\lambda(X_1)}{d\lambda}\right|_\lambda\right]\varphi^{(5)}_{\sqrt{2}b}(Y_2-Y_1)\right\}^2\\&=\frac{1}{2n}\int\left\{\left[\left.\frac{dg_\lambda\left(g_\lambda^{-1}\left(\frac{v+\sqrt{2}bu}{2}\right)\right)}{d\lambda}\right|_\lambda\right.\right.\end{aligned}$$

$$- \frac{dg_\lambda\left(g_\lambda^{-1}\left(\frac{v-\sqrt{2}bu}{2}\right)\right)}{d\lambda}\Bigg|_\lambda\Bigg] \varphi^{(5)}(u) \frac{1}{(\sqrt{2}b)^6}\Bigg\}^2$$

$$f_Y\left(\frac{v+\sqrt{2}bu}{2}\right) f_Y\left(\frac{v-\sqrt{2}bu}{2}\right) \frac{b}{\sqrt{2}} dudv.$$

Applying again Taylor expansions

$$\frac{1}{2n} E\left\{\left[\frac{dg_\lambda(X_2)}{d\lambda}\bigg|_{\lambda_o} - \frac{dg_\lambda(X_1)}{d\lambda}\bigg|_{\lambda_o}\right] \varphi^{(5)}_{\sqrt{2}b}(Y_2 - Y_1)\right\}^2 \propto \frac{1}{nb^9}.$$

The second term in (4.17) is

$$- 2E\left\{\left[\frac{dg_\lambda(X_2)}{d\lambda}\bigg|_\lambda - \frac{dg_\lambda(X_1)}{d\lambda}\bigg|_\lambda\right] \varphi^{(5)}_{\sqrt{2}b}(Y_2 - Y_1) Z_{n1}\right\}$$

$$= -2 \int \Bigg\{ \Bigg[\frac{dg_\lambda\left(g_\lambda^{-1}\left(\frac{v+\sqrt{2}bu}{2}\right)\right)}{d\lambda}\Bigg|_\lambda$$

$$- \frac{dg_\lambda\left(g_\lambda^{-1}\left(\frac{v-\sqrt{2}bu}{2}\right)\right)}{d\lambda}\Bigg|_\lambda\Bigg] \varphi^{(5)}(u) \frac{1}{(\sqrt{2}b)^6}\Bigg\}$$

$$f_Y\left(\frac{v+\sqrt{2}bu}{2}\right) f_Y\left(\frac{v-\sqrt{2}bu}{2}\right) \frac{1}{n} \chi\left(\frac{v-\sqrt{2}bu}{2}\right) \frac{b}{\sqrt{2}} dudv \propto \frac{1}{nb^4},$$

where we introduced the function $\chi(\cdot)$ as such

$$Z_{n1} = \frac{1}{n} \chi(Y_1).$$

The third term in (4.17) is similarly treated:

$$- \frac{C_b}{n} E\left\{\left[\frac{dg_\lambda(X_2)}{d\lambda}\bigg|_\lambda - \frac{dg_\lambda(X_1)}{d\lambda}\bigg|_\lambda\right] \varphi^{(5)}_{\sqrt{2}b}(Y_2 - Y_1)\right\}$$

$$= - \frac{C_b}{n} \int \Bigg[\frac{dg_\lambda\left(g_\lambda^{-1}\left(\frac{v+\sqrt{2}bu}{2}\right)\right)}{d\lambda}\Bigg|_\lambda - \frac{dg_\lambda\left(g_\lambda^{-1}\left(\frac{v-\sqrt{2}bu}{2}\right)\right)}{d\lambda}\Bigg|_\lambda\Bigg]$$

$$\varphi^{(5)}(u) \frac{1}{(\sqrt{2}b)^6} \frac{b}{\sqrt{2}} dudv \propto \frac{b \cdot b}{nb^6} \propto \frac{1}{nb^4}.$$

Thus we have shown that the second and third terms in (4.17) go to zero at rate $1/nb^4$ (because $nb^5 \to \infty$), while the first term is of order $1/nb^9$. Q.E.D.

LEMMA 4.4 *For the remainder term*

$$\begin{aligned} ER_n^2 &= \frac{2}{n^4} E\left\{ \sum_{j=1}^{n} \int (A_n(X_j) - a_n(b,y))(B_n(X_j) - b_n(b,y))dy \right\}^2 \\ &\leq \frac{2}{n^2} E\left\{ \int (A_n(X_1) - a_n(b,y))(B_n(X_1) - b_n(b,y))dy \right\}^2 \\ &= O\left(\frac{1}{n^2}\right) \end{aligned} \tag{4.18}$$

Hence $\sqrt{n}R_n \xrightarrow{L^2} 0$, *and of course also* $\sqrt{n}R_n \xrightarrow{P} 0$.

Proof Because X_j's are i.i.d. and using Schwarz inequality, the inequality part of (4.18) is obvious. The last part of (4.18) is worked out as

$$\begin{aligned} &\frac{2}{n^2} E\left\{ \int (A_n(X_1) - a_n(b,y))(B_n(X_1) - b_n(b,y))dy \right\}^2 \\ &= \frac{2}{n^2} E\left\{ \int A_n(X_1)B_n(X_1)dy - nZ_{n1} - \frac{1}{2}C_b \right\}^2 \\ &= \frac{2}{n^2} E\left\{ \frac{dg_\lambda(X_1)}{d\lambda}\bigg|_\lambda \varphi^{(5)}_{\sqrt{2}b}(Y_1 - Y_1) - nZ_{n1} - \frac{1}{2}C_b \right\}^2 \\ &= \frac{2}{n^2} E\left\{ \frac{dg_\lambda(X_1)}{d\lambda}\bigg|_\lambda \varphi^{(5)}_{\sqrt{2}b}(0) - nZ_{n1} - \frac{1}{2}C_b \right\}^2 \\ &= \frac{2}{n^2} E\left\{ nZ_{n1} + \frac{1}{2}C_b \right\}^2 = O\left(\frac{1}{n^2}\right) \to 0. \end{aligned}$$

Q.E.D.

PROPOSITION 4.2 *We have*

$$\sqrt{n + n^2b^9}(\widehat{L}'(\lambda) - C_b) \xrightarrow{D} N(0, 4\sigma_2^2)$$

uniformly for $\lambda \in \Lambda$, $b \in B_n = [n^{(-1+\delta)/5}, n^{-\delta}]$, *where*

$$C_b = L'(\lambda) + c_1(f_X)b^2 + c_2(f_X)b^4 + O(b^6)$$

is defined in (4.9) *and where* $\sigma_2^2 = \sigma_2^2(f_X, \lambda)$ *is implicitly defined and depends only on* f_X *and* λ.

Proof This follows immediately from all the previous Lemmas and also the fact that

$$\widehat{L}'(\lambda) = C_b + 2\sum_{j=1}^{n} Z_{nj} + 2\sum_{1\le i<j\le n} H_n(X_i, X_j) + R_n,$$

where the terms $\sum_{1\le i<j\le n} H_n(X_i, X_j)$ and $2\sum_{j=1}^{n} Z_{nj}$, both being asymptotically normal of their respective orders of variance, are uncorrelated, and their joint distribution is asymptotically bivariate normal, a fact which can be easily shown through martingale central limit theorem [Hall (1984)]. Q.E.D.

The proof of the next lemma is on page 153 of Yang (1995).

LEMMA 4.5 *As* $n \to \infty$, $\widehat{L}''(\xi) = L''(\lambda_0) + O(b^2)$.

Proof of Theorem 2 This follows immediately by combining Proposition 4.2 and Lemma 4.5. Because we can write $\widehat{\lambda}$ as

$$\begin{aligned}\widehat{\lambda} = \lambda_0 - \frac{\widehat{L}'(\lambda_0)}{\widehat{L}''(\xi)} &= \lambda_0 - \frac{L'(\lambda_0) + c_1(f_X, \lambda_0)b^2 + c_2(f_X, \lambda_0)b^4 + Z}{L''(\lambda_0) + O(b^2)} \\ &= \lambda_0 - \frac{c_1(f_X, \lambda_0)b^2 + c_2(f_X, \lambda_0)b^4 + Z}{L''(\lambda_0) + O(b^2)},\end{aligned}$$

where

$$\sqrt{n + n^2 b^9}(Z) \xrightarrow{D} N(0, 4\sigma_2^2(f_X, \lambda_0))$$

and $L'(\lambda_0) = 0$ because λ_0 is the minimizer of $L(\lambda)$ and $\lambda_0 \in$ interior of Λ. Thus

$$\widehat{\lambda} = \lambda_0 - \frac{c_1(f_X, \lambda_0)b^2 + c_2(f_X, \lambda_0)b^4 + Z}{L''(\lambda_0)} + O\left(b^4 + \frac{b^2}{\sqrt{n + n^2 b^9}}\right)$$

and so

$$\sqrt{n+n^2b^9}\left(\widehat{\lambda}-\lambda_0+\frac{c_1(f_X,\lambda_0)b^2}{L''(\lambda_0)}+O(b^4)\right)\xrightarrow{D}N\left(0,\frac{4\sigma_2^2(f_X,\lambda_0)}{L''(\lambda_0)^2}\right).$$

This completes the proof of Theorem 2. Q.E.D.

To prove Theorem 3, we need to study several other $(d^l/d\lambda^l)|_\lambda \int(\widehat{f}_Y^{(k)}(y))^2dy$ statistics. First, we give the result for $k=2$ and $l=2$. By definition, it is $\widehat{L}''(\lambda)=(d^2/d\lambda^2)\widehat{L}(\lambda)|_\lambda=2\int[\widehat{f}''_Y\widehat{f}''_{Y,\lambda\lambda}+(\widehat{f}''_{Y,\lambda})^2]|_\lambda dy$. We note also that $L''(\lambda)=2\int[f''_Yf''_{Y,\lambda\lambda}+(f''_{Y,\lambda})^2]|_\lambda dy$.

PROPOSITION 4.3 *We have*

$$\sqrt{n+n^2b^9}(\widehat{L}''(\lambda)-(L''(\lambda)+c'_1(f_X)b^2+c'_2(f_X)b^4+O(b^6)))\xrightarrow{D}N(0,4\sigma_3^2)$$

uniformly for $\lambda\in\Lambda$, $b\in B_n=[n^{(-1+\delta)/5},n^{-\delta}]$, *where*

$$c'_1(f_X)=\frac{d}{d\lambda}c_1(f_X,\lambda),c'_2(f_X)=\frac{d}{d\lambda}c_2(f_X,\lambda)$$

and where $\sigma_3^2=\sigma_3^2(f_X,\lambda)$ *is implicitly defined and depends only on* f_X *and* λ.

Proof This is essentially the same as Proposition 4.2, the only difference being the order which is caused by an increase of the order of derivatives. Specifically, in this case, the

$$\begin{aligned}\operatorname{var}(\sqrt{n}U_n)&\propto\frac{1}{2n}E\left\{\left[\frac{dg_\lambda(X_2)}{d\lambda}\bigg|_\lambda-\frac{dg_\lambda(X_1)}{d\lambda}\bigg|_\lambda\right]^2\varphi^{(6)}_{\sqrt{2}b}(Y_2-Y_1)\right.\\&\qquad\left.-\left[\frac{d^2g_\lambda(X_2)}{d\lambda^2}\bigg|_\lambda-\frac{d^2g_\lambda(X_1)}{d\lambda^2}\bigg|_\lambda\right]\varphi^{(5)}_{\sqrt{2}b}(Y_2-Y_1)\right\}^2\\&\propto\frac{1}{nb^9},\end{aligned}$$

where the 6-th derivative term is being controlled by a term that gives out b^2, rendering it no less controlled than a 5-th order derivative with a term that gives out an b. Similarly

$$ER_n^2 \propto \frac{1}{n^2}$$

because

$$\left[\frac{dg_\lambda(X_1)}{d\lambda}\bigg|_\lambda - \frac{dg_\lambda(X_1)}{d\lambda}\bigg|_\lambda\right]^2 \varphi^{(6)}_{\sqrt{2}b}(Y_1 - Y_1) = 0$$

and

$$-\left[\frac{d^2g_\lambda(X_1)}{d^2\lambda}\bigg|_\lambda - \frac{d^2g_\lambda(X_1)}{d^2\lambda}\bigg|_\lambda\right]\varphi^{(5)}_{\sqrt{2}b}(Y_1 - Y_1) = 0.$$

Q.E.D.

From this point on, we have to assume that $b \in [n^{(-1+\delta)/9}, n^{-\delta}]$ for some $\delta \in (0, 1/10)$, in other words, $b = b_V$, the special bandwidth used for deriving the improved estimator $\widehat{\lambda}_2$. The next two propositions deal with the case of $k = 3$ or 4, $l = 1$ and $k = 3$ or 4, $l = 2$.

PROPOSITION 4.4 *The asymptotic variance of*

$$c_1(\widehat{f}_X, \lambda) = -\frac{d}{d\lambda}\bigg|_\lambda \int (\widehat{f}^{(3)}_Y(y))^2 dy \qquad (4.19)$$

is given by

$$E|c_1(\widehat{f}_X, \lambda) - c_1(f_X, \lambda) + 3c_2(f_X, \lambda)b^2 - c_3(f_X, \lambda)b^4|^2 = O\left(\frac{1}{n^2b^{13}}\right)$$

uniformly for $\lambda \in \Lambda$, $b \in [n^{(-1+\delta)/9}, n^{-\delta}]$; *similarly the following is true*

$$E|c_2(\widehat{f}_X, \lambda) - c_2(f_X, \lambda) - c_3(f_X, \lambda)b^2 - c_4(f_X, \lambda)b^4|^2 = O\left(\frac{1}{n^2b^{17}}\right)$$

uniformly for $\lambda \in \Lambda$, $b \in [n^{(-1+\delta)/9}, n^{-\delta}]$, *where* $c_3(f_X, \lambda)$ *is similarly defined as*

$$c_3(f_X, \lambda) = -const. \times \frac{d}{d\lambda}\bigg|_\lambda \int (f^{(5)}_Y(y))^2 dy$$

and so is with $c_4(f_X, \lambda)$.

Proof Similar to Proposition 4.1 before, note the variance that comes from the cross terms in those bilinear sums are, for example, in the case of $c_1(\widehat{f}x, \lambda)$ of the form:

$$\frac{1}{n^2}E\left\{\left[\frac{dg_\lambda(X_2)}{d\lambda}\bigg|_\lambda - \frac{dg_\lambda(X_1)}{d_\lambda}\bigg|_\lambda\right]\varphi^{(7)}_{\sqrt{2}b}(Y_2 - Y_1)\right\}^2,$$

therefore, a change of variables done in the same way as in the proof of Proposition 4.1 shows that this is of the order $(1/n^2b^{13})$. Consider also the case of $c_1'(\widehat{f}_X, \lambda)$. Here the term is of the form:

$$\frac{1}{n^2}E\left\{\left[\frac{dg_\lambda(X_2)}{d\lambda}\bigg|_\lambda + \frac{dg_\lambda(X_1)}{d\lambda}\bigg|_\lambda\right]^2 \varphi^{(8)}_{\sqrt{2}b}(Y_2 - Y_1) - \left[\frac{d^2g_\lambda(X_2)}{d\lambda^2}\bigg|_\lambda - \frac{d^2g_\lambda(X_1)}{d\lambda^2}\bigg|_\lambda\right]\varphi^{(7)}_{\sqrt{2}b}(Y_2 - Y_1)\right\}^2,$$

which by the same technique as used above is easily seen to be of the order $(1/n^2b^{13})$ as well, *etc.* Q.E.D.

The proof of the next proposition is similar.

PROPOSITION 4.5 *The derivatives of* $c_1(\widehat{f}_X, \lambda)$, *i.e.*,

$$c_1'(\widehat{f}_X, \lambda) = -\frac{d^2}{d\lambda^2}\bigg|_\lambda \int (\widehat{f}^{(3)}_Y(y))^2 dy$$

has the following asymptotic variance

$$E|c_1'(\widehat{f}_X, \lambda) - c_1'(f_X, \lambda) + 3c_2'(f_X, \lambda)b^2 - c_3'(f_X, \lambda)b^4|^2 = O\left(\frac{1}{n^2b^{13}}\right)$$

uniformly for $\lambda \in \wedge$, $b \in [n^{(-1+\delta)/9}, n^{-\delta}]$; *the derivatives of* $c_2(\widehat{f}x, \lambda)$ *has the following asymptotic variance*

$$E|c_2'(\widehat{f}_X, \lambda) - c_2'(f_X, \lambda) - c_3'(f_X, \lambda)b^2 - c_4'(f_X, \lambda)b^4|^2 = O\left(\frac{1}{n^2b^{17}}\right)$$

uniformly for $\lambda \in \Lambda$, $b \in [n^{(-1+\delta)/9}, n^{-\delta}]$.

To prove Theorem 1, we need one last lemma,

LEMMA 4.6 *The asymptotics for* $c_1(\widehat{f}_X, \widehat{\lambda})$ *is*

$$b^2(c_1(\widehat{f}_X,\widehat{\lambda}) - c_1(f_X,\lambda_0) - (\widehat{\lambda}-\lambda_0)c_1'(\widehat{f}_X,\zeta) + 3c_2(f_X,\lambda_0)b^2) = o_{L^2}\left(\frac{1}{\sqrt{n}}\right),$$

where ζ *is some point between* $\widehat{\lambda}$ *and* λ_0. *Furthermore,*

$$\begin{aligned}\left|(c_1(\widehat{f}_X, \widehat{\lambda}) - c_1(f_X, \lambda_0))b^2 - 3c_2(f_X, \lambda_0)b^4 + o_{L^2}\left(\frac{1}{\sqrt{n}}\right)\right| \\ \leq \textit{const.}|\widehat{\lambda} - \lambda_0|b^2 \\ \leq (\textit{const.})b^4 + o_{L^2}\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

Consequently, $\widehat{\lambda}_1$ *is an improved version of* $\widehat{\lambda}$

$$\sqrt{n}(\widehat{\lambda}_1 - \lambda_0 - Bias_1(n,b)) \to N(0, \sigma_0^2),$$

where

$$Bias_1(n,b) = O(b^4).$$

Proof

$$\begin{aligned}b^2(c_1(\widehat{f}_X, \widehat{\lambda}) - c_1(f_X, \lambda_0)) = b^2(c_1(\widehat{f}_X, \widehat{\lambda}) \\ - c_1(\widehat{f}_X, \lambda_0) + c_1(\widehat{f}_X, \lambda_0) - c_1(f_X, \lambda_0)).\end{aligned}$$

Apply mean value theorem to the first term, we have

$$\begin{aligned}b^2(c_1(\widehat{f}_X, \widehat{\lambda}) - c_1(f_X, \lambda_0)) = c_1'(\widehat{f}_X, \zeta)(\widehat{\lambda} - \lambda_0)b^2 \\ + (c_1(\widehat{f}_X, \lambda_0) - c_1(f_X, \lambda_0))b^2,\end{aligned}$$

apply the asymptotics of Proposition 4.4, we see that the first term is

$$c_1'(f_X, \zeta)(\widehat{\lambda} - \lambda_0)b^2$$

plus a term of order

$$b^{2(2+2)}\frac{1}{n^2b^{13}} = \frac{1}{n^2b^5} = o\left(\frac{1}{n}\right),$$

because $(\widehat{\lambda} - \lambda_0)$ is of order b^2; similarly, the variance of the second term is of order

$$b^{2\times 2}\frac{1}{n^2 b^{13}} = \frac{1}{n^2 b^9} = o\left(\frac{1}{n}\right)$$

while the bias in the second term is of the form

$$(-3c_2(f_X, \lambda_0)b^2 + c_3(f_X, \lambda_0)b^4)b^2 = -3c_2(f_X, \lambda_0)b^4 + o\left(\frac{1}{\sqrt{n}}\right).$$

These established the asymptotics of $c_1(\widehat{f}_X, \widehat{\lambda})$. Note that:

$$\widehat{L}_1'(\lambda) = \widehat{L}'(\lambda) - c_1(\widehat{f}_X, \widehat{\lambda})b^2$$

and also that:

$$\widehat{L}_1''(\lambda) = \widehat{L}''(\lambda),$$

we have already proved that:

$$\text{Bias}_1(n, b) = O(b^4)$$

and therefore $\widehat{\lambda}_1$ is an improvement over $\widehat{\lambda}$. Q.E.D.

Proof of Theorem 3 The proof is essentially the same as in Lemma 4.6. Note only two things in particular:

1. The coefficient of $c_2(\widehat{f}_X, \widehat{\lambda}_1)$ is 4 instead of 3, because as we see in Lemma 4.6, the subtraction of $c_1(\widehat{f}_X, \widehat{\lambda}_1)b^2$ from $\widehat{L}'(\lambda_0)$ produces an extra term $c_2(f_X, \lambda_0)b^4$;
2. To verify that the variances caused by the introduction of $c_2(\widehat{f}_X, \widehat{\lambda}_1)b^4$ are of order $o(1/n)$, note that $(\widehat{\lambda}_1 - \lambda_0)$ is of order $O(b^4)$ and therefore the variance in

$$c_2'(\widehat{f}_X, \zeta)(\widehat{\lambda}_1 - \lambda_0)b^4$$

is of the order

$$(b^8)^2\frac{1}{n^2 b^{17}} = \frac{1}{n^2 b} = o\left(\frac{1}{n}\right)$$

similarly, the variance in the term

$$(c_2(\widehat{f_X}, \lambda_0) - c_2(f_X, \lambda_0))b^4$$

is of the order

$$(b^4)^2 \frac{1}{n^2 b^{17}} = \frac{1}{n^2 b^9} = o\left(\frac{1}{n}\right)$$

thus we have proved Theorem 3. Q.E.D.

Acknowledgment

The results presented here are part of the Ph. D. work of the author, who thanks his advisor, J. S. Marron, for stimulating questions that motivated him to search hard for an improved estimator. Part of the paper was finished while the author was visiting Humboldt University in Berlin with the financial support of Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse" Deutsche Forschungsgemeinschaft, which is gratefully acknowledged. The suggestions of the associate editor are also appreciated.

References

Aldershof, B., Marron, J. S., Park, B. U. and Wand, M. P. (1990) Facts about the Normal Density, unpublished manuscript.

Devroye, L. and Györfi, L. (1985) *Nonparametric Density Estimation: The L_1 View*, John Wiley, New York.

Hall, P. (1984) Central Limit Theorem for Integrated Square Error of Multivariate Nonparametric Density Estimators, *Journal of Multivariate Analysis*, **14**, 1–16.

Hall, P. and Marron, J. S. (1987) Estimation of Integrated Squared Density Derivatives, *Statistics and Probability Letters*, **6**, 109–115.

Härdle, W. and Marron, J. S. (1990) Semiparametric Comparison of Regression Curves, *The Annals of Statistics*, **18**, 63–89.

Jones, M. C., Marron, J. S. and Park, B. U. (1991) A simple root n bandwidth selector, *The Annals of Statistics*, **19**, 1919–1932.

Jones, M. C. and Sheather, S. J. (1991) Using non-stochastic terms to advantage in kernel-based estimation of integrated squared density derivatives, *Statistics and Probability Letters*, **11**, 511–514.

Kim, W. C., Park, B. U. and Marron, J. S. (1994) Asymptotically best bandwidth selectors in kernel density estimation, *Statistics and Probability Letters*, **19**, 119–127.

MacGillivray, H. L. (1992) Shape Properties of the g-and-h and Johnson Families, *Commun. Statist.-Theory Meth.*, **21**, 1233–1250.

Park, B. U., Chung, S. S. and Seog, K. H. (1992) An Empirical Investigation of the Shifted Power Transformation Method in Density Estimation, *Computational Statistics and Data Analysis, North Holland,* **14**, 183–191.

Ruppert, D. and Cline, D. B. H. (1994) Bias Reduction in Kernel Density Estimation, *The Annals of Statistics*, **22**, 185–210.

Ruppert, D. and Wand, M. P. (1992) Correcting for Kurtosis in Density Estimation, *Australian Journal of Statistics*, **34**, 19–29.

Scott, D. W. (1992) *Multivariate Density Estimation*, John Wiley & Sons, Inc.

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

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

Wand, M. P., Marron, J. S. and Ruppert, D. (1991) Transformations in Density Estimation, *Journal of the American Statistical Association*, **86**, 343–361.

Yang, L. (1995) *Transformation-Density Estimation, Ph. D. Dissertation*, University of North Carolina, Institute of Statistics Mimeo Series #2337.

Yang, L. and Marron, J. S. (1999) Iterated Transformation-Kernel Density Estimation, *Journal of the American Statistical Association*, **94**, 580–589.