Iterated Transformation–Kernel Density Estimation

Lijian YANG and James S. MARRON

Transformation from a parametric family can improve the performance of kernel density estimation. In this article we give two data-driven estimators for the optimal transformation parameter. We demonstrate that multiple families of transformations can be used at the same time, and there can be benefits to iterating this process. The transformation scheme can be expected to first pick the right transformation family and then pick the optimal parameter. Insight as to the performance of the method comes from our analysis of a number of real datasets, two of which are included in this article. To illustrate the effectiveness and asymptotics of the transformation method, we also present results on one of the five target densities used in our simulation study. It is then proved that the Johnson family of transformations, when coupled with transformation-kernel density estimation, makes a wide variety of density shapes easier to estimate. The transformation method has overall better performance than the usual method and in many cases is much more effective.

KEY WORDS: Curvature; Global bandwidth; Johnson family; Normal mixture; Pilot estimator.

1. INTRODUCTION

Nonparametric density estimation extracts information about the underlying structure of a dataset when no appropriate parametric model is available. In estimating the density function f_X of an iid random sample $X_1, \ldots, X_n$, the kernel density estimator (KDE) is easy to understand intuitively; it puts "bumps" of the same shape at every observation and averages them,

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

where $K_h(u) = h^{-1}K(u/h)$. The kernel K is a symmetric probability density; $h > 0$ is the bandwidth.

Very often, when the true density has sharp features, such as high skewness or kurtosis, the KDE using a global bandwidth either obscures these important features or creates extra features. Here we use real data to illustrate; further analysis of this data is given in Section 3.

Figure 1 shows that the kernel estimates with smaller bandwidth (i.e., $\frac{1}{5}h_{\mathrm{SJ}}$) captures the peak on the left, yet is too wiggly; the one with larger bandwidth (i.e., $5h_{\mathrm{SJ}}$) is smooth, yet smooths away the peak; the one using the Sheather–Jones automatic bandwidth (h_{SJ}) is an attempt to balance the needs to capture the peak and to remain smooth. In the sense of overall performance, the Sheather–Jones bandwidth is a very popular global bandwidth, but it also does not produce a satisfactory density estimate for the data in Figure 1.

We have developed an algorithm that transforms a dataset first and then uses the Sheather–Jones bandwidth to estimate the density of the transformed dataset. That density estimate is then back-transformed to the original scale to estimate the original density. Figure 1 also shows the estimate obtained by this transformation method overlaid with the estimates not using transformation. The transformed estimate is better than the untransformed ones in terms of overall smoothness and capturing the left peak.

The transformed estimate works like a variable bandwidth estimate; it estimates the density at the left peak as if using a small global bandwidth (e.g., $\frac{1}{5}h_{\mathrm{SJ}}$), while using a large global bandwidth (e.g., $5h_{\mathrm{SJ}}$) toward the right tail. The advantage of using transformation is that it still allows the use of a global bandwidth, albeit on a transformed scale. The different amount of smoothing needed at different locations is absorbed in the transformation, making it possible to use a global bandwidth effectively. This is important because much is known about global bandwidth choice (see Jones, Marron, and Sheather 1992), but less for local choice. The transformation procedure can be iterated any number of times. We find that transforming the data twice yields an estimate much superior to the estimate without transformation, and in most cases, little improvement is achieved after two transformation steps. (For further references on the transformation method, see Devroye and Györfi 1985; Silverman 1986. Related recent works include Park, Chung, and Seog 1992; Ruppert and Cline 1994; Ruppert and Wand 1992; Wand, Marron, and Ruppert 1991; Yang 1995, 1999.)

In Section 2 we present the setting of transformation in density estimation, examine its functional analysis, and introduce a data-driven method for implementation. In particular, we discuss use of the Johnson family. We then show that the method is sensitive enough to select the best family of transformations for a given density. In Section 3 we give a real data example. In Section 4 we present key simulation results. In Section 5 we give a sufficient condition for an effective transformation and also prove that the Johnson family of transformations can improve the estimation of all density shapes that are $C^3(R)$ [we use the notation $C^k(R)$ for the space of functions with continuous kth derivative

Lijian Yang is Assistant Professor, Department of Statistics and Probability, Michigan State University, East Lansing, MI 48824. James S. Marron is Professor, Statistics Department, University of North Carolina, Chapel Hill, NC 27599. The results presented here are part of Lijian Yang's Ph.D. work under the direction of James Marron. The authors thank J. Fan, C. Ji, D. Ruppert, J. S. Simonoff, Y. Truong, an associate editor, and an anonymous referee for their many useful suggestions and J. Hannan for pointing out a technical error. Part of the revision was done while Lijian Yang was visiting Humboldt University in Berlin with the financial support of Sonderforschungsbereich 373 "Quantifikation und Simulation Ökonomischer Prozesse" Deutsche Forschungsgemeinschaft, which is gratefully acknowledged.

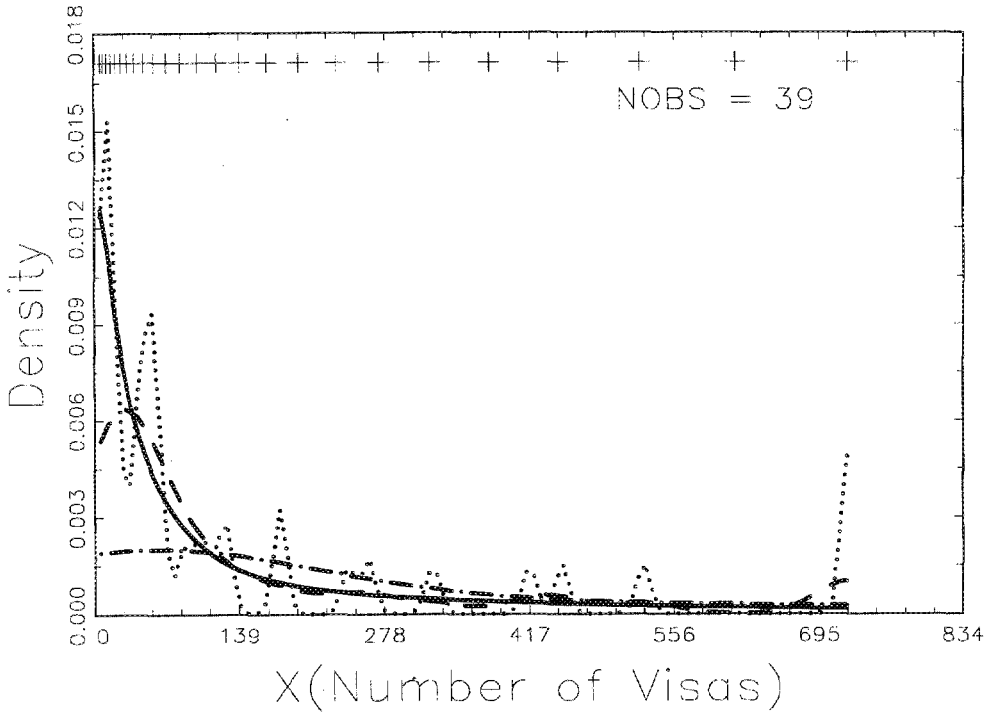

Figure 1. Kernel Density Estimates of the Adoption Visa Data. The following three bandwidths are used: the Sheather–Jones bandwidth h_{SJ} (– – –), $5h_{SJ}$ (– · · · –), and $\frac{1}{5}h_{SJ}$ (· · ·). ——, the estimate by transformation method; +, points that become equally spaced after transformation.

on R] and rapidly decreasing. We present our conclusions in Section 6, and technical proofs in the Appendix.

2. THE THEORY AND METHOD

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

$$f_Y(y, \lambda) = f_X\{g_\lambda^{-1}(y)\}(g_\lambda^{-1})'(y). \tag{2}$$

From the family $\{f_Y(y, \lambda)\}$ of densities, one for each value of the parameter λ, we would select the one easiest to estimate with a global bandwidth. Suppose that this is $f_Y(y, \lambda_0)$. We need a functional $G(\cdot)$ of density functions such that $f_Y(\cdot, \lambda_0)$ has the least $G(\cdot)$ values among the $f_Y(y, \lambda)$'s. This functional $G(\cdot)$ measures the difficulty of estimation with a global bandwidth. It should be scale invariant, because every density remains as easy to estimate under rescaling. Two such functionals were considered by Wand and Devroye (1993); each measures how well the kernel estimator converges to the true density in the L^2 and L^1 norms. We use the one for L^2 theory, for convenience of obtaining asymptotic results and implementing the algorithm. We now describe this functional in context.

The mean integrated squared error (MISE) of estimating f_X with the kernel estimator $\hat{f}_X$ in (1) is

$$\text{MISE}(h) = E \int \{\hat{f}_X(x) - f_X(x)\}^2 \, dx. \tag{3}$$

As $n \to \infty, h \to 0$, and $nh \to \infty$, if f_X has a uniformly continuous and L_2 second derivative, then we have

$$\text{MISE}(h) = \text{AMISE}(h) + o\left(\frac{1}{nh} + h^4\right),$$

where

$$\text{AMISE}(h) = \frac{R(K)}{nh} + \frac{1}{4}\sigma_K^4 h^4 R(f_X'') \tag{4}$$

is the asymptotic mean integrated squared error (AMISE), which is approximately the same as MISE, with $R(\psi) = \int(\psi)^2$ for any function ψ and $\sigma_K^2 = \int x^2 K(x)\,dx$ (Jones et al. 1992). The AMISE is minimized by the following asymptotically optimal bandwidth:

$$h_* = \left(\frac{R(K)}{\sigma_K^4 R(f_X'')}\right)^{1/5} n^{-1/5}. \tag{5}$$

Plugging this h_* back into (4), we have

$$\begin{aligned}\inf_{h>0} \text{AMISE}(h) &= C(K)n^{-4/5}R(f_X'')^{1/5} \\ &= C(K)n^{-4/5}\left[\int [f_X''(x)]^2\,dx\right]^{1/5}, \end{aligned} \tag{6}$$

where $C(K)$ is a constant depending on K only. It is clear from (6) that when $C(K)$ is fixed, smaller values of $R(f_X'')$ yields smaller L^2 errors of estimating f_X with $\hat{f}_X$. Therefore, it is "easier" to estimate f_X when $R(f_X'')$ is smaller. Another way of looking at this is from the viewpoint that the global asymptotic optimal bandwidth h_* in (5) is larger when $R(f_X'')$ is smaller; in general, a smaller optimal bandwidth is needed when the underlying density is wiggly. Finally, geometrically speaking, $R(f_X'')$ is a global measure

of the curvature of the density, and less curvature makes estimation easier.

Hence $R(f_X'')$ is a functional measuring how easy it is to estimate f_X. Because $R([(1/c)f_X(\frac{\cdot}{c})]'') = (1/c^5)R(f_X'')$, a scale-invariant version of $R(f_X'')$ is

$$G(f_X) = \sigma(f_X)(R(f_X''))^{1/5} = \sigma_{f_X}(R(f_X''))^{1/5}, \quad (7)$$

where $\sigma_{f_X}^2 = \sigma(f_X)^2$ is the variance of the distribution whose density is f_X. One can verify that $G(f_X)$ is invariant under rescaling. The smaller $G(f_X)$, the larger the scale-free optimal bandwidth h_*, and thus the easier the estimation.

Terrell (1990) proved the following general result:

> Among those $C^k(R)$ densities defined on the real line with specified variance σ^2, the member of the scale family of beta$(k+2, k+2)$ has the smallest value of $\int (f^{(k)})^2$; that is, the density that achieves the minimum is a rescaling of the function
>
> $$f(x) = \frac{(2k+3)!}{2^{2k+3}((k+1)!)^2}(1-x^2)^{k+1}1_{\{|x|\le 1\}}.$$

If $k = 2$, then Terrell's result says in particular that $G(\cdot)$ is minimized by beta(4, 4). The minimum is $(35/243)^{1/5} \approx .6787$. The normal density is close to attaining this lower bound value $(= (3/8\sqrt{\pi})^{1/5} \approx .733)$ and is similar to the beta(4,4) in shape.

Our transformation–kernel density estimation (TKDE) method attempts to reduce $G(\cdot)$ as much as possible by transforming f_X. The idea is to use $G(\cdot)$ as a kind of "oblique" functional to force the density toward the shape of the beta(4, 4) density. This raises the issue of whether $f_Y(\cdot, \lambda_0)$ is easier to estimate than $f_X(\cdot)$. Without loss of generality, we also assume that Λ is one-dimensional and that the left endpoint of Λ gives the identical transformation; that is, $g_{\min(\Lambda)}(x) \equiv x$. This means that $\lambda = \min(\Lambda)$ corresponds to the untransformed density $f_X(\cdot)$, which is thus included in the family of transformed densities $\{f_Y\}_{\lambda\in\Lambda}$. We assume here that Λ is compact, so that an optimal parameter exists. One can then further assume that $\Lambda = [0, \lambda_M]$, where $\lambda_M \in (0, 1)$, and that $g_0(x) \equiv x$. If this is not true, then we just do a linear reparametrization of $\Lambda, \lambda \to (\lambda - \min(\Lambda))/2 \times (\max(\Lambda) - \min(\Lambda))$ to make it true. Now if $\lambda_0 > 0$, then $G(f_Y(\cdot, \lambda_0)) < G(f_X(\cdot))$. That means transforming f_X into $f_Y(\cdot, \lambda_0)$ by $g_{\lambda_0}(\cdot)$ provides an easier estimation setting, which allows improved performance. If, on the other hand, $\lambda_0 = 0$, then either no transformation, or transformation by a different family is needed.

We now define the following target function of $\lambda \in \Lambda$:

$$L(\lambda) = G(f_Y(\cdot, \lambda)) = \sigma(f_Y(\cdot, \lambda))(R(f_Y''(\cdot, \lambda)))^{1/5} \quad (8)$$

and set $Y_i = g_\lambda(X_i), i = 1, 2, \ldots, n$. Then $Y_1, \ldots, Y_n$ are iid, and each has the same density function f_Y given in (2). Because $L(\lambda)$ is minimized at λ_0, an approach to estimation of λ_0 is based on estimation of $L(\lambda)$. We estimate λ_0 with $\hat{\lambda}$, the minimizer of

$$\hat{L}(\lambda) = \sigma_{\hat{Y}}(\lambda) \int [\hat{f}_Y''(y, b, \lambda)^2]\, dy, \quad (9)$$

where $\sigma_{\hat{Y}}(\lambda)$ is the standard deviation of $Y_1, \ldots, Y_n$,

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

and φ denotes the Gaussian kernel; that is, $\varphi(u) = (1/\sqrt{2\pi})e^{-u^2/2}$ with $\varphi_b^{(k)}(u) = b^{-(k+1)}\varphi^{(k)}(u/b)$. Here b is a pilot bandwidth used solely for $\hat{\lambda}$. Because Λ is compact and $\hat{L}(\lambda)$ is continuous in λ (under some mild assumptions; see Yang 1995, sec. 2.1); $\hat{\lambda}$ clearly exists. We do not need $\hat{\lambda}$ to be unique.

Once $\hat{\lambda}$ is obtained, one also gets $\hat{h}_*$, an estimate of h_*, the optimal bandwidth as in (5), except that the underlying density is $f_Y(\cdot, \lambda_0)$. This $\hat{h}_*$ can be the Sheather–Jones plug-in bandwidth h_{SJ}, the Park-Marron plug-in bandwidth h_{PM} or other versions, as discussed by Jones et al. (1992). Then one gets the following estimate of $f_X(x)$:

$$\hat{f}_X(x, \hat{h}_*, \hat{\lambda}) = n^{-1} \sum_{j=1}^{n} g_{\hat{\lambda}}'(x)\varphi_{\hat{h}_*}[g_{\hat{\lambda}}(x) - g_{\hat{\lambda}}(X_j)]. \quad (10)$$

Following Wand et al. (1991), the global performance of the estimator (10) is conveniently assessed by the AMISE of $\hat{f}_Y(\cdot, \hat{h}_*, \hat{\lambda})$, which is an estimate of

$$\begin{aligned} \mathrm{AMISE}_Y(h_*, \lambda_0) &= \inf_{h>0} \mathrm{AMISE}_Y(h, \lambda_0) \\ &= Cn^{-4/5}L(\lambda_0)^{1/5} \\ &= Cn^{-4/5} \min_{\lambda\in\Lambda} L(\lambda)^{1/5}, \end{aligned} \quad (11)$$

which is consistent with our view here, because both amount to minimizing the same function $L(\lambda)$.

This approach leads to a big computational advantage. The binning ideas of Fan and Marron (1994) and Scott and Terrell (1987) are very straightforward to implement for $\hat{f}_Y(\cdot, \hat{h}_*, \hat{\lambda})$. One can bin $Y_1, \ldots, Y_n$ and then estimate $\hat{f}_Y(\cdot, h_*, \lambda)$, which is simply a binned implementation of an ordinary kernel estimator. From $\hat{f}_Y(\cdot, \hat{h}_*, \hat{\lambda})$ to $\hat{f}_X(\cdot, \hat{h}_*, \hat{\lambda})$ takes only the simple step of (10). Computation of $\hat{f}_X(\cdot, \hat{h}_*, \hat{\lambda})$ is thus as interactive as the ordinary kernel estimation method, except for the steps needed to estimate λ_0 with $\hat{\lambda}$.

To complete the description of the implementation procedure, we now give two data-driven bandwidths b that can be used in the pilot estimator $\hat{L}(\lambda)$:

1. The diagonals-out bandwidth of Park-Marron, which minimizes the asymptotic mean squared error (AMSE) of estimating $\int f_Y''^2\, dy$ by $(1/n^2)\sum_{i\neq j} \varphi_{\sqrt{2}b}^{(4)}(Y_i - Y_j)$:

$$b_{\mathrm{PM}} = C_1(f_Y)D_1(\varphi)n^{-2/13}$$

2. The diagonals-in bandwidth of Sheather–Jones, which minimizes the AMSE of estimating $\int f_Y''^2\, dy$ by $(1/n^2)\sum_{i,j=1}^{n} \varphi_{\sqrt{2}b}^{(4)}(Y_i - Y_j)$:

$$b_{\mathrm{SJ}} = C_2(f_Y)D_2(\varphi)n^{-1/7}.$$

The b_{SJ} is used in computing the Sheather–Jones bandwidth, as b_{PM} is used in computing the Park-Marron bandwidth. An account of b_{SJ}, b_{PM}, and the functionals $C_1(\cdot), C_2(\cdot), D_1(\cdot)$, and $D_2(\cdot)$ has been given by Jones et al. (1992); the pilot bandwidth b_{PM} was used by Wand et al. (1991). Because the Sheather–Jones bandwidth is preferred over the Park-Marron bandwidth in overall performance, we use b_{SJ} instead of b_{PM} in our simulation studies and applications.

About the behavior of $\hat{\lambda}$ as an estimator of λ_0, we know that under certain regularity conditions on $f_X, g_\lambda(\cdot)$, and the pilot bandwidth $b, \hat{\lambda}$ converges consistently to λ_0. Also, $\hat{\lambda}$ converges to λ_0 in probability with asymptotic bias of order b^2 and asymptotic variance of order $1/(n + n^2 b^9)$. For our algorithm that uses $b_{SJ}, \hat{\lambda}$ has bias of order $n^{-2/7}$ and variance of order $n^{-5/7}$. Proofs of these results have been given by Yang (1995, chap. 6).

For the choice of transformation family, we explicitly define the Johnson families

$$g_{1,\lambda}(x) = \frac{1}{cJ}\ln(1 + cJx), \qquad 0 < \lambda \le \lambda_M, \tag{12}$$

$$g_{2,\lambda}(x) = \frac{1}{c}\ln(cx + \sqrt{1 + c^2x^2}), \qquad 0 < \lambda \le \lambda_M, \tag{13}$$

and

$$g_{3,\lambda}(x) = \frac{1}{2c}\ln[(1 + cx)/(1 - cx)], \quad 0 < \lambda \le \lambda_M, \tag{14}$$

where $J = \pm 1, c = \lambda^p/(1 - \lambda^p), g_{i,0}(x) \equiv x$, for $i = 1, 2, 3$, and $p \ge \frac{1}{2}$ is a tuning constant that is needed for smoothness of $g_{\gamma,\lambda}(x)$ as a function of λ, as can be seen in Theorem 3 and its proof. Another way of writing these is

$$g_{\gamma,\lambda}(x) = \begin{cases} \ln(1 + cJx)/cJ & 0 < \lambda \le \lambda_M, \quad \gamma = 1 \\ \ln(cx + \sqrt{1 + c^2x^2})/c & 0 < \lambda \le \lambda_M, \quad \gamma = 2 \\ \ln\left(\frac{1+cx}{1-cx}\right) \Big/ 2c & 0 < \lambda \le \lambda_M, \quad \gamma = 3 \\ x & \lambda = 0, \qquad \gamma = 1, 2, 3. \end{cases} \tag{15}$$

which is equivalent to pooling several families together by adding a discrete parameter (γ here). We use $c = \lambda^p/(1-\lambda^p)$ instead of λ, so that λ lies entirely in $[0, 1)$, yet c lies in the full interval $[0, \infty)$. This is important, because larger c means more drastic transformation. (For the Johnson family, see Johnson 1949.)

To provide some intuition as to why we use these three families, we include in Figure 2 two pictures, each overlaying six density curves that are the results of applying the inverse of Johnson families 2 and 3 to the standard normal density, using equally spaced λ values. The curve $\lambda = 0$ is the standard normal density.

As λ increases, the curve becomes more and more kurtotic under Johnson family 2; that is, having large probability mass near the mean and in the tails. This suggests that when a density f_X has high kurtosis, an optimal transformation from the Johnson family 2 could transform it into a near-normal shape. Similarly, Johnson family 3 deals with density shapes having small kurtosis; that is, those having less probability mass in the center and in the tails. Due to lack of space, Johnson family 1 is not included in Figure 2; we simply point out that it will transform skewed density into a near-normal shape, with $J = -1$ and 1 handling skewness of both orientations. In Section 5 we show that the Johnson families are indeed able to improve the estimation of most density functions.

Because there are three Johnson families, we describe the TKDE method when several families are used at the same time. Suppose that we have m families of transformations $\{g_j\}, j = 1, 2, 3, \ldots, m$, each with its own range of parameters $\Lambda_j = [0, \lambda_{j,M}], g_j(x, 0) \equiv x$, for $j = 1, 2, 3, \ldots, m$. Suppose also that $\lambda_j = \text{argmin}_{\lambda \in \Lambda_j} L_j(\lambda)$, where $L_j(\lambda)$ is defined as in (8) and $L_j = L_j(\lambda_j), j = 1, 2, 3, \ldots, m$. Let $\hat{\lambda}_j = \text{argmin}_{\lambda \in \Lambda_j} \hat{L}_j(\lambda), j = 1, 2, 3, \ldots, m$, where $\hat{L}_j(\lambda)$ is defined as in (9). Let $\hat{l}_j = \hat{L}_j(\hat{\lambda}_j)$ be the pilot estimator of $L, j = 1, 2, \ldots, m$. To decide which family is more suitable for transforming density f_X, we use the following rules.

Family Selection Rules.

- Theoretical rule: Use family j if $L_j = \min\{L_k, 1 \le k \le m\}$.
- Data-driven rule: With a given sample, use family j if $\hat{l}_j = \min\{\hat{l}_k, 1 \le k \le m\}$.

So, theoretically, one chooses family j over family k if $L_j < L_k$. Here, the convention is that when ties exist among the L_k's, one can order the tied families arbitrarily. The data–driven rule has the same convention about ties.

Theorem 1.

$$P[(\hat{l}_j - \hat{l}_k)(L_j - L_k) < 0 \quad \text{for some} \quad 1 \le j, k \le m, j \ne k] \to 0, \quad \text{as} \quad n \to \infty.$$

Therefore, the probability of not selecting the family with the lowest L value is asymptotically zero.

This theorem guarantees that the data-driven rule is asymptotically the same as the theoretical rule. The proof is based on the fact that $\hat{l}_j \to L_j$ in probability, for $j = 1, 2, \ldots, m$, as $n \to \infty$. (For a proof, see Yang 1995a, sec. 7.2.)

The method that we have just finished describing is used throughout the rest of the article.

3. APPLICATION

In this section we analyze the adoption visa data first introduced in Section 1. (See Yang 1995, chap. 1, for more real data examples.) The data comprise the numbers of international adoption visas granted to U.S. residents by the Immigration and Naturalization Service in 1991, by the country of origin of the adoptee. Only countries with nonzero values are included (see Chatterjee, Handcock, and Simonoff 1995, p. 287).

Recall that $L(\lambda) = G(f_Y(\cdot, \lambda))$ as defined in (8). Our analysis finds that $G(\cdot)$ is reduced from about 1.81 to about

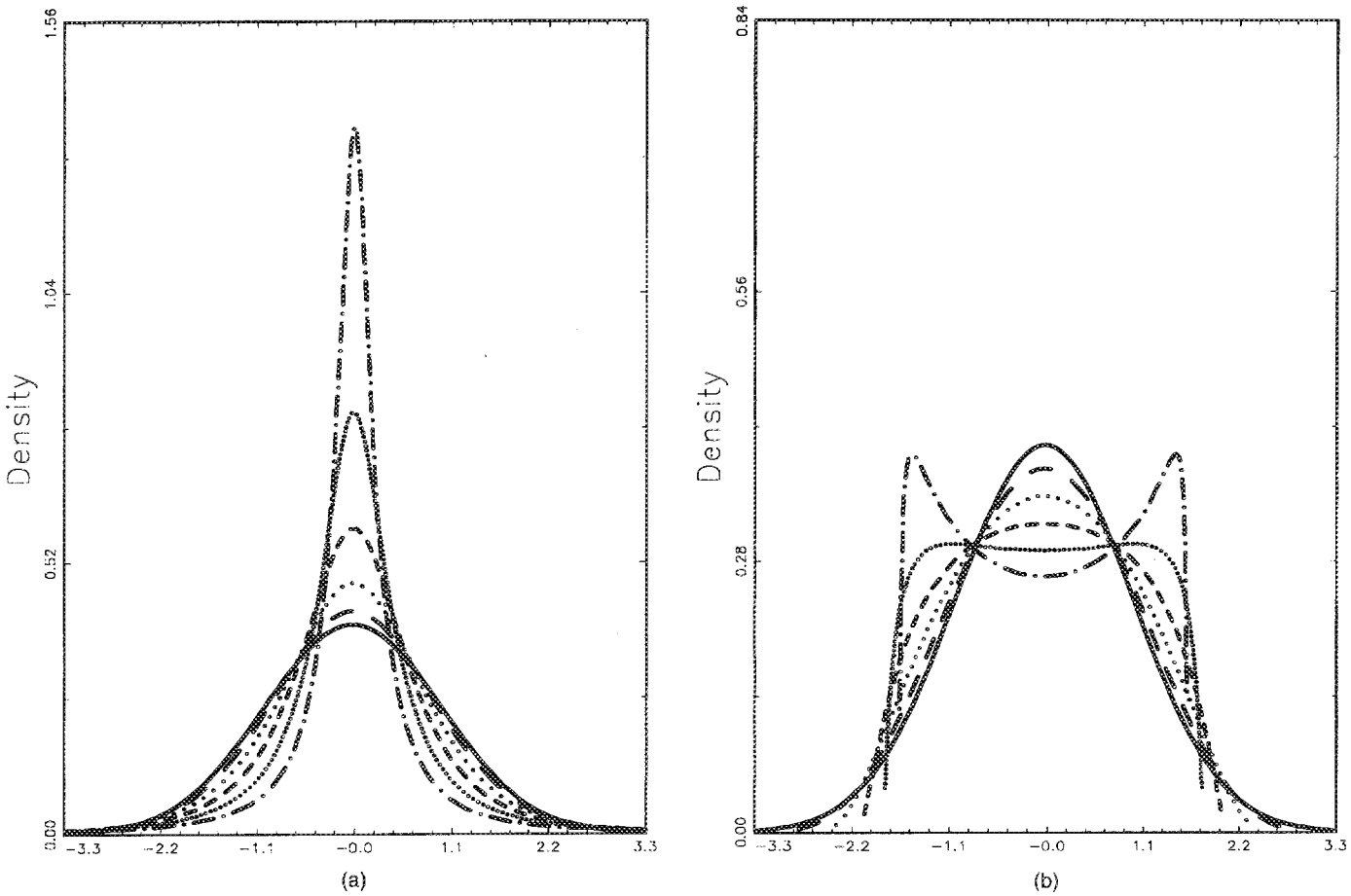

Figure 2. The Effects of Applying the Inverse of Johnson Family 2(a) and Family 3(b) to Standard Normal Density. η values as follow: (a) ——, .000; —, .268; · · ·, .379; – – –, .465; · · ·, .537; – · – · –, .600; (b) ——, .000; —, .215; · · ·, .304; – – –, .372; · · ·, .429; – · – · –, .480.

.743. Most of the reduction happens in the first step of the transformation. Both steps were by Johnson family 1, because the data are very skewed. The "transformed-twice" density in the Z space has little curvature, because $G(f_Z)$ has nearly achieved the minimum of .68. Thus there is little to be gained in further transforming Z.

Our transformed estimates of the density clear identifies a peak on the left and a long tail to the right. This suggests that except all but a small number of the 39 countries have very few immigration visas granted to their citizens. A closer look at the data identifies Romania, South Korea, Peru, Columbia, India, Philippines, and Guatemala as having a large number of visas, with all other countries having visa numbers below 270. This might suggest either a bias in U.S. immigration law or a greater availability for adoption from these seven countries. The high number for India is not so impressive considering its huge population, whereas the high number for Romania clearly relates to the economic hardship following the political changes of 1990. It seems that geographical affinity for the South American countries is a significant factor as well.

4. SIMULATION STUDY

To understand the effectiveness and the asymptotics of the transformation method, we present here the results of a simulation study on one normal mixture density; the skewed M-shape density shown in Figure 3. This is of the form

$$f_X(x) = \sum_{j=1}^{k} w_j \varphi_{\sigma_j}(x - \mu_j) = \sum_{j=1}^{k} w_j \frac{1}{\sqrt{2\pi}\sigma_j} e^{-[(x-\mu_j)^2/2\sigma_j^2]}, \tag{16}$$

where

$$(w_1, w_2, \ldots, w_k) = \frac{1}{16}(1, 1, \ldots, 1, 8)_{1\times 9},$$

$$(\sigma_1, \sigma_2, \ldots, \sigma_k) = \left(1, \frac{2}{3}, \ldots, \left(\frac{2}{3}\right)^7, 1\right)_{1\times 9}$$

and

$$(\mu_1, \mu_2, \ldots, \mu_k) = \left(3\times 1 - 1, 3\times\frac{2}{3} - 1, \ldots, 3\times\left(\frac{2}{3}\right)^7 - 1, 0\right)_{1\times 9}.$$

The simulation results on four other normal mixtures—the M-shape, the strongly skewed, the standard normal, and the kurtotic—are not included here. The lessons of those parallel that of the skewed M-shape (Yang 1995a, chap. 4). The normal mixtures constitute a very broad class. The

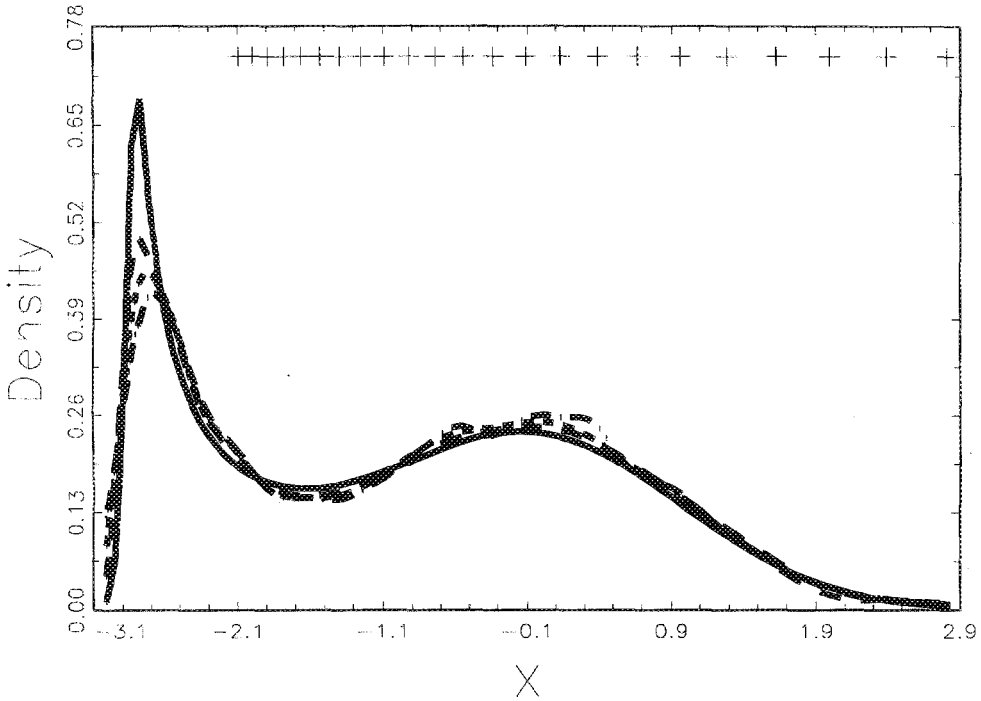

Figure 3. Kernel Density Estimates Based on a Pseudodataset Generated From the Skewed M-Shape Density. Optimally selected transformations from the Johnson families are applied to the data twice (– – –) and once (- - -) to compare to the estimate without transformation (– · – · –) and the true underlying curve (——). The + 's represent points that become equally spaced after the first transformation. SSE2 = .082; SSE1 = .152; SSEU = .234; NOBS = 1,000.

skewed M-shape and the M-shape are new additions to the 15 normal mixtures of Marron and Wand (1992), which included the other three mentioned earlier. Figure 3 shows the skewed M-shape density ("true") together with estimates by three methods. The "untransformed" refers to the global Sheather–Jones bandwidth estimate, the "transformed" refers to the estimate by applying transformation from X to Y with Johnson family 1, and the "transformed twice" refers to the estimate by applying transformation in two consecutive steps, from X to Y and then from Y to Z by Johnson family 3.

In Figure 3, the values "SSEU," "SSE1," and "SSE2" are the sum of the squared errors of the "untransformed," the "transformed," and the "transformed twice" estimates, with respect to the "true" density. The number "NOBS" is the number of observations. Here we have applied the binned implementation with 401 bins (Fan and Marron 1994). The +'s represent 26 points on the X scale that become equally spaced in the Y scale after transformation. We see that SSEU > SSE1 > SSE2, which shows that the "transformed" estimator is better than the "untransformed" and the "transformed twice" is the best of all three.

The theoretical $G(\cdot)$ values of transformed densities under all three Johnson families are plotted in Figure 4 (with the parameters for all three families adjusted to the same scale). This figure shows why Johnson family 1 is the family used to transform the skewed M-shape density, because it reduces the $G(\cdot)$ value the most: from about 5.6 to 1.4. This is because the skewed M-shape density is more skewed than kurtotic compared to the beta(4, 4).

For the pseudodataset used in Figure 3, transformation by Johnson family 1 first corrected skewness and transformation by Johnson family 3 then corrected the kurtosis. Our algorithm was able to detect that skewness is the more dominant feature than kurtosis and must be dealt with first. During the two transformation steps, the $G(\cdot)$ value was reduced from 1.566 to 1.311 and then to 1.146.

To better understand the variability of our estimates across datasets, we generated 500 pseudodatasets from the skewed M-shape density, with sample sizes 100, 1,000, and 10,000, and applied transformation method to these 1,500 datasets. The conclusion is that both across samples and in terms of the average, there is a huge reduction of $G(\cdot)$ values and the log-integrated squared errors (log (ISE))'s from "untransformed" to "transformed" and a smaller, yet nonnegligible reduction from "transformed" to "transformed twice."

To see how often the right transformation family is selected, we note that for the 1,500 datasets, in step 1 Johnson family 1 was selected with relative frequencies of $107/500 = .214, 440/500 = .88$, and $500/500 = 1$ as sample size increases from 100 to 1,000 to 10,000. Likewise, in step 2 the relative frequencies of selecting Johnson family 3 are $347/500 = .694, 440/500 = .88$, and $500/500 = 1$. This shows that the probability of selecting the right family tends to 1 rather rapidly.

The $\hat{\lambda}$ values are also found to converge rapidly as the sample size goes from 100 to 1,000 to 10,000.

5. THE EFFECTIVENESS OF JOHNSON FAMILIES

In this section we give results that for any density f satisfying certain rather mild conditions, $G(f)$ can be reduced by applying at least one of the three Johnson families.

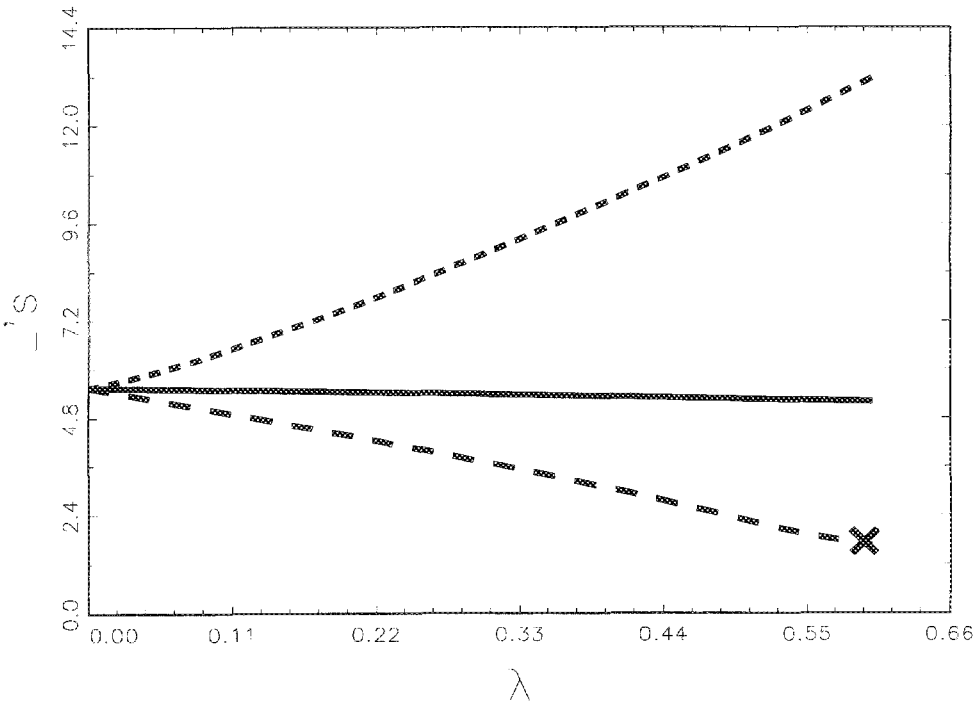

Figure 4. Theoretical G( ) Values of the Skewed M-Shape Density When Transformations From the Johnson Families (– – –; Family 1; - - -, Family 2; ——, Family 3) are Applied Once.

Recall from Section 2 that

$$G(f_X) = \sigma(f_X)(R(f_X''))^{1/5} = \sigma_{f_X}(R(f_X''))^{1/5}$$

is the functional to be minimized through transformation. The univariate function to be minimized is $L(\lambda) = \int f_Y''(y)^2\,dy = G(f_Y)^5$, where $Y = g_\lambda(X)$. Denote the minimizer by λ_0. Now the parameters are in $\Pi = [0, \lambda_M]$, where $\lambda_M \in (0,1)$, $g_0(x) \equiv x$, and $L(0) = G(f_X)^5$. A sufficient condition for $\{g_\lambda\}$ to reduce $G(f)$ is that $L'(0) < 0$, which implies that there is a small enough ε such that $L(\varepsilon) < L(0)$, and thus $L(\lambda_0) \le L(\varepsilon) < L(0)$. To simplify notation, we use in this section

$$\begin{aligned} L(\lambda) &= \sigma_Y^{10} R(f_Y'')^2 = G(f_Y)^{10} \\ &= \left[\int g_\lambda(x)^2 f_X(x)\,dx - \left(\int g_\lambda(x) f_X(x)\,dx\right)^2\right]^5 \\ &\quad \times \left(\int f_Y''(y)^2\,dy\right)^2, \end{aligned}$$

which is equivalent to the previous definition of L. We assume the following:

A1. g_λ is jointly C^3 for $x \in S(f_X)$ and $\lambda \in \Lambda$.

A2. $f_Y \in S^3(R)$, the space of third-order rapidly decreasing functions, for every $\lambda \in \Lambda$. This means that f_Y is C^3 and that for any $k = 1, 2, 3, \ldots$ and $l = 0, 1, 2, 3$, $\lim_{|y|\to\infty} |y|^k f_Y^{(l)}(y) = 0$.

A3. $\mu_X = \int x f_X(x)\,dx = 0$.

Assumption A1 is satisfied by the Johnson families. Assumption A2 is satisfied by densities such as normal mixture and exponential. Assumption A3 is just for convenience; if it is not satisfied, a translation of f_X can achieve it.

Theorem 2. Under Assumptions A1, A2, and A3,

$$\begin{aligned} L'(0) = {} & 2\sigma_X^8 R(f_X'')\left\{5R(f'')\int x\frac{d}{d\lambda}g_\lambda(x)\Big|_{\lambda=0} f_X(x)\,dx \right. \\ & + 2\sigma_X^2 \int f_X''(x)\left[f_X^{(3)}(x)x + 3f_X^{(2)}(x)\frac{d}{d\lambda}g_\lambda^{-1}(x)\right. \\ & + 3f_X'(x)\frac{d}{d\lambda}(g_\lambda^{-1})''(x) \\ & \left.\left. + f_X(x)\frac{d}{d\lambda}(g_\lambda^{-1})'''(x)\right]\Big|_{\lambda=0} dx\right\}; \end{aligned}$$

that is,

$$\begin{aligned} L'(0) = {} & 2\sigma_X^8 R(f_X'')\left\{5\int x\frac{d}{d\lambda}g(x)\Big|_{\lambda=0} f_X(x)\,dx \right. \\ & \times \int f''(x)^2\,dx + 2\int x^2 f_X(x)\,dx \\ & \times \int f_X''(x)\left[f_X^{(3)}(x)x\right. \\ & + 3f_X^{(2)}(x)\frac{d}{d\lambda}g_\lambda^{-1}(x) + 3_X'(x) \\ & \left.\left. \times \frac{d}{d\lambda}(g_\lambda^{-1})''(x) + f_X(x)\frac{d}{d\lambda}(g_\lambda^{-1})'''(x)\right]\Big|_{\lambda=0} dx\right\}. \end{aligned}$$

Theorem 3. Under Assumptions A1, A2, and A3, define

$$A(f_X) = -\frac{5}{3}\sigma_X^8 R(f_X'')\left\{\int x^4 f_X R(f_X'') - 3\sigma_X^2\left[\int x^2 (f_X'')^2 - 2\int (f_X')^2\right]\right\}$$

$$= -\frac{5}{3}\sigma_X^8 R(f_X'')\left\{\int x^4 f_X(x)\,dx \int (f_X''(x))^2\,dx - 3\int x^2 f_X(x)\,dx\left[\int x^2 (f_X''(x))^2\,dx - 2\int (f_X'(x))^2\,dx\right]\right\}$$

and

$$B(f_X, J) = -J\frac{5}{2}\sigma_X^8 R(f_X'')\left\{\int x^3 f_X R(f_X'') - 2\sigma_X^2\int x (f_X'')^2\right\}$$

$$= -J\frac{5}{2}\sigma_X^8 R(f_X'')\left\{\int x^3 f_X(x)\,dx\int (f_X''(x))^2\,dx - 2\int x^2 f_X(x)\,dx\int x (f_X''(x))^2\,dx\right\}.$$

a. Let $p = 1$, if $J = 1$ and $S(f_X)$ is bounded from below, or if $J = -1$ and $S(f_X)$ is bounded from above; then for family $\{g_1\}$,

$$L'(0) = B(f_X, J) = -B(f_X, -J).$$

b. Let $p = \frac{1}{2}$; then, for family $\{g_2\}$,

$$L'(0) = A(f_X).$$

c. Let $p = \frac{1}{2}$, if $S(f_X)$ is bounded; then, for family $\{g_3\}$,

$$L'(0) = -2A(f_X).$$

Thus if $S(f_X)$ is bounded and $A(f_X)^2 + B(f_X, 1)^2 > 0$, then at least one family out of $\{g_1\}$ (with both options of J available), $\{g_2\}$ and $\{g_3\}$ can reduce the value of $\sigma_X^{10} R(f_X'')^2$.

Note that $A(f_X)^2 + B(f_X, 1)^2 = 0$ happens very rarely, because it requires both $A(f_X)$ and $B(f_X, 1)$ to be 0. Theorem 3 guarantees that every compactly supported and third-order smooth density f_X can be made easier to estimate by one of the Johnson families. As for infinitely supported densities, their truncation can be used. We have the following generalized version of Theorem 3.

Theorem 4. Under Assumptions A1, A2, and A3, and $A(f_X)^2 + B(f_X, 1)^2 > 0$, there exist constants depending on $f_X, R_0 > 0, \varepsilon > 0$, and a certain Johnson family $\{g_{\gamma,\lambda}\}$ such that for every $R > R_0$ if we denote by $(f_X)_R$ the truncation of f_X to the interval $[-R, R]$ and by $(f_Y)_R$ the optimally transformed density from $(f_X)_R$ by $\{g_{\gamma,\lambda}\}$, then the following is true:

$$|G(f_X) - G((f_X)_R)| < \frac{\varepsilon}{2}$$

and

$$G((f_Y)_R) < G((f_X)_R) - \varepsilon.$$

Therefore,

$$G((f_Y)_R) < G(f_X) - \frac{\varepsilon}{2}.$$

This result provides a theoretical basis for applying all three Johnson families to infinitely supported densities.

As an example, consider densities beta $(k, k), k \geq 5$. It is easily shown that their $B(\cdot)$'s are 0. A direct calculation shows that their $A(\cdot)$'s are all < 0; therefore, Johnson family 3 is able to improve their kernel estimation. This is consistent with the fact that family 3 reduces the kurtosis, as shown in Figure 3. As a density with high k is transformed to a shape close to the optimal $k = 4$, its kurtosis $[(2k+1)/(2k+3)]$ decreases as k decreases. A special case $(k = \infty)$ (i.e., the standard normal) has been a target of our simulation work.

6. CONCLUSIONS

In view of the findings on the simulated and real examples in this paper and Chapters 4 and 5 of Yang (1995), we conclude the following:

1. One of the Johnson families is suitable for transforming any given density, as indicated by Theorem 3; as the sample size increases, our algorithm is able to select the right Johnson family in most cases, as indicated by Theorem 1.
2. The estimates of the optimal parameter λ_0 converge with satisfactory speed.
3. The $G(\cdot)$ values and the ISEs are reduced by each step of transformation, more in the first step than the second, and in most cases little can be gained to transform more.
4. The transformation is more effective on densities with sharp features; the effects are marginal otherwise, yet there is some improvement.

Based on these conclusions, we recommend the iterated transformation method for kernel density estimation in most situations.

We also want to point out that although in this article we have used the Johnson families exclusively, any number of other families can be used simultaneously as well. In particular, it may be possible to combine the power transformation family used by Wand et al. (1991), the kurtosis reducing family used by Ruppert and Wand (1992), and a third family that increases kurtosis. In such a setting, Theorem 1 guarantees that the best family will be automatically selected for any density. Further investigation will be needed, however, to establish the analog of Theorem 3 to show that such a union of families can effectively transform a wide range of density functions.

APPENDIX: PROOFS

We give here the proofs of Theorems 2 and 3. Proof of Theorem 4 is straightforward from Theorem 3 and is not included here.

Proof of Theorem 2

This is rather straightforward. Note that for $\lambda > 0$ and $Y = g_\lambda(X)$,

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

and thus

$$\begin{aligned}\frac{d}{d\lambda} f_Y''(y) &= f_X'''(g_\lambda^{-1}(y))\frac{d}{d\lambda} g_\lambda^{-1}(y)((g_\lambda^{-1})'(y))^3 \\ &\quad + 3f_X''(g_\lambda^{-1}(y))((g_\lambda^{-1})'(y))^2 \frac{d}{d\lambda}(g_\lambda^{-1})'(y) \\ &\quad + 3f_X''(g_\lambda^{-1}(y))(g_\lambda^{-1})''(y)(g_\lambda^{-1})'(y)\frac{d}{d\lambda}(g_\lambda^{-1})(y) \\ &\quad + 3f_X'(g_\lambda^{-1}(y))(g_\lambda^{-1})''(y)\frac{d}{d\lambda}(g_\lambda^{-1})'(y) \\ &\quad + 3f_X'(g_\lambda^{-1}(y))(g_\lambda^{-1})'(y)\frac{d}{d\lambda}(g_\lambda^{-1})''(y) \\ &\quad + f_X'(g_\lambda^{-1}(y))(g_\lambda^{-1})'''(y)\frac{d}{d\lambda}(g_\lambda^{-1})(y) \\ &\quad + f_X(g_\lambda^{-1}(y))\frac{d}{d\lambda}(g_\lambda^{-1})'''(y).\end{aligned}$$

When $\lambda = 0$, we have $g_\lambda^{-1}(y) \equiv y, (g_\lambda^{-1})'(y) \equiv 1$, and $(g_\lambda^{-1})^{(k)}(y) \equiv 0$ for $k > 1$; therefore,

$$\begin{aligned}f_Y''(y)|_{\lambda=0} &= f_X''(y)(1)^3 + 3f_X'(y) \times 1 \times 0 + f_X(y) \times 0 \\ &= f_X''(y), \qquad \text{(A.1)}\end{aligned}$$

and similarly,

$$\begin{aligned}\frac{d}{d\lambda} f_Y''(y)\bigg|_{\lambda=0} &= f_X'''(y)\frac{d}{d\lambda}\, g_\lambda^{-1}(y) + 3f_X''(y)\frac{d}{d\lambda}\,(g_\lambda^{-1})'(y) \\ &\quad + 3f_X'(y)\frac{d}{d\lambda}\,(g_\lambda^{-1})''(y) + f_X(y)\frac{d}{d\lambda}(g_\lambda^{-1})'''(y). \qquad \text{(A.2)}\end{aligned}$$

We thus conclude from (A.1) and (A.2) that

$$\begin{aligned}&\frac{d}{d\lambda}\bigg|_{\lambda=0} R(f_Y'') \\ &= \int 2f_X''(x)\left(f_X'''(x)\frac{d}{d\lambda}g_\lambda^{-1}(x) + 3f_X''(x)\frac{d}{d\lambda}(g_\lambda^{-1})'(x)\right. \\ &\quad \left. + 3f_X'(x)\frac{d}{d\lambda}(g_\lambda^{-1})''(x) + f_X(x)\frac{d}{d\lambda}(g_\lambda^{-1})'''(x)\right)\bigg|_{\lambda=0} dx. \qquad \text{(A.3)}\end{aligned}$$

Now also note that

$$\begin{aligned}\frac{d}{d\lambda}\bigg|_{\lambda=0}(\sigma_Y^2) &= \frac{d}{d\lambda}\bigg|_{\lambda=0}\left(\int y^2 f_Y(y)\,dy - \left(\int y f_X(y)\,dy\right)^2\right) \\ &= \int 2x\frac{d}{d\lambda}g_\lambda(x)\bigg|_{\lambda=0} f_X(x)\,dx, \qquad \text{(A.4)}\end{aligned}$$

because by Assumption A3, $\mu_X = 0$. Now $L(\lambda) = \sigma_Y^{10} R(f_Y'')^2$, so

$$\begin{aligned}&\frac{d}{d\lambda}\bigg|_{\lambda=0} L(\lambda) \\ &= 5\sigma_Y^8\frac{d}{d\lambda}\bigg|_{\lambda=0}(\sigma_Y^2)R(f_Y'')^2 + 2\sigma_Y^{10}\frac{d}{d\lambda}\bigg|_{\lambda=0} R(f_Y'')R(f_Y'').\end{aligned}$$

Plugging in (A.3) and (A.4) directly completes the proof.

To prove Theorem 3, we need some preliminary results. As a convention, set $\lambda_p = p/\lambda(1-\lambda^p)$ in what follows. The next lemma has been proved by Yang (1995, lem. 7.1.2, p. 160).

Lemma A.1. For Johnson family 1, $g_{1,\lambda}(x) = (1/cJ)\ln(1 + cJx)$,

$$\frac{d}{d\lambda}\, g_{1,\lambda}(x) = \begin{cases}\lambda_p(xg_{1,\lambda}'(x) - g_{1,\lambda}(x)), \\ -\frac{J}{2}x^2, \\ 0,\end{cases}$$

$$\frac{d}{d\lambda}\, g_{1,\lambda}^{-1}(x) = \begin{cases}\lambda_p(x(g_{1,\lambda}^{-1})'(x) - g_{1,\lambda}^{-1}(x)) & \lambda > 0 \\ \frac{J}{2}x^2 & \lambda = 0, \quad p = 1 \\ 0 & \lambda = 0, \quad p > 1,\end{cases}$$

$$\frac{d}{d\lambda}\,(g_{1,\lambda}^{-1})'(x) = \begin{cases}\lambda_p x(g_{1,\lambda}^{-1})''(x), \\ Jx, \\ 0,\end{cases}$$

$$\frac{d}{d\lambda}\,(g_{1,\lambda}^{-1})''(x) = \begin{cases}\lambda_p(x(g_{1,\lambda}^{-1})'''(x) + (g_{1,\lambda}^{-1})''(x)) & \lambda > 0 \\ J & \lambda = 0, \quad p = 1 \\ 0 & \lambda = 0, \quad p > 1,\end{cases}$$

and

$$\frac{d}{d\lambda}\,(g_{1,\lambda}^{-1})'''(x) = \begin{cases}\lambda_p(x(g_{1,\lambda}^{-1})''''(x) + 2(g_{1,\lambda}^{-1})'''(x)) & \lambda > 0 \\ 0 & \lambda = 0, \quad p = 1 \\ 0 & \lambda = 0, \quad p > 1.\end{cases}$$

Proof of Theorem 3

We prove only part a of the theorem; the other parts are similar. Substituting for the $(d/d\lambda)$ terms in the formula of Theorem 2 their respective expressions in Lemma A.1, we have

$$\begin{aligned}L'(0) &= J2\sigma_X^8 R(f_X'')\left[-\frac{5}{2}\int x^3 f_X(x)\,dx \int f''(x)^2\,dx\right. \\ &\quad + 2\int x^2 f_X(x)\,dx \int \left(f_X''(x)f_X'''(x)\frac{1}{2}\,x^2\right. \\ &\quad \left.\left. + 3f_X''(x)^2 x + 3f_X'(x)f_X''(x)\right)dx\right].\end{aligned}$$

Note that $\int 3f_X'(x)f_X''(x)\,dx = \frac{3}{2}f_X'(x)^2|_{-\infty}^{+\infty} = 0$ by Assumption A2, which makes $\lim_{x\to\pm\infty} f_X'(x)^2 = 0$. By the same token, integrating by parts gives

$$\begin{aligned}&\int (f_X''(x)f_X'''(x)\frac{1}{2}x^2\,dx \\ &= \frac{1}{2}f_X''(x)^2\frac{1}{2}\,x^2|_{-\infty}^{+\infty} + \int \frac{1}{2}\,(-f_X''(x)^2)x\,dx \\ &= \int \frac{1}{2}(-f_X''(x)^2)x\,dx.\end{aligned}$$

Now a little reduction gives the result,

$$\begin{aligned}L'(0) &= -J\frac{5}{2}\sigma_X^8 R(f_X'')\left\{\int x^3 f_X R(f_X'') - 2\sigma_X^2\int x(f_X'')^2\right\} \\ &= B(f_X, J),\end{aligned}$$

which proves part a of the theorem.

[Received December 1995. Revised November 1998.]

REFERENCES

Chatterjee, S., Handcock, M. S., and Simonoff, J. S. (1995), *A Casebook for a First Course in Statistics*, New York: Wiley.

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

Fan, J., and Marron, J. S. (1994), "Fast Implementations of Nonparametric Curve Estimators," *Journal of Computational and Graphical Statistics*, 3, 35–56.

Johnson, N. L. (1949), "Systems of Frequency Curves Generated by Methods of Translation," *Biometrika*, 36, 149–176.

Jones, M. C., Marron, J. S., and Sheather, S. J. (1992), "Progress in Data-Based Bandwidth Selection for Kernel Density Estimation," Mimeo Series 2088, University of North Carolina, Institute of Statistics.

Marron, J. S., and Wand, M. P. (1992), "Exact Mean Integrated Squared Error," *The Annals of Statistics*, 20, 712–736.

National Hockey League (1992), *The National Hockey League Official Guide and Record Book 1992–93*, Chicago, IL: Triumph Books.

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*, 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*, New York: Wiley.

Scott, D. W., and Terrell, G. R. (1987), "Biased and Unbiased Cross-Validation in Density Estimation," *Journal of the American Statistical Association*, 82, 1131–1146.

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

Terrell, G. R. (1990), "The Maximal Smoothing Principle in Density Estimation," *Journal of the American Statistical Association*, 85, 470–477.

Wand, M. P., and Devroye, L. (1993), "How Easy is a Given Density to Estimate?" *Computational Statistics and Data Analysis*, 16, 311–323.

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," unpublished dissertation, University of North Carolina, Dept. of Statistics.

——— (in press), "Root-n Convergent Transformation–Kernel Density Estimation," *Journal of Nonparametric Statistics*.