Computational Statistics & Data Analysis 16 (1993) 311–323
North-Holland

How easy is a given density to estimate?

M.P. Wand

Australian Graduate School of Management, University of New South Wales, Kensington, N.S.W., Australia

Luc Devroye

School of Computer Science, McGill University, Montreal, Canada

Received January 1992
Revised June 1992

Abstract: In data analytic applications of density estimation one is usually interested in estimating the density over its support. However, common estimators such as the basic kernel estimator use a single smoothing parameter over the whole of the support. While this will be adequate for some densities there will be other densities that will be very difficult to estimate using this approach. The purpose of this article is to quantify how easy a particular density is to estimate using a global smoothing parameter. By considering the asymptotic expected L_1 error we obtain a scale invariant functional that is useful for measuring degree of estimation difficulty. Implications for the transformation kernel density estimators, which attempt to overcome the inadequacy of the basic kernel estimator, are also discussed.

Keywords: Kernel density estimator; L_1 loss; Mean integrated absolute error; Nonparametric curve estimation; Transformation density kernel estimator

1. Introduction

Suppose that $X_1, \ldots, X_n$ is a real-valued random sample with unknown probability density function f. In data analytic applications it is desirable to construct an estimate of f over its support. The simplest and best understood way of doing this is via the kernel estimator

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

Here $K_h(u) = K(u/h)/h$ and the kernel K is a function which integrates to one. However, throughout this article we will take K to be a symmetric density

Correspondence to: Dr. M.P. Wand, Australian Graduate School of Management, University of New South Wales, P.O. Box 1, Kensington, NSW 2033, Australia.

having finite second moment. This ensures that (1.1) is also a density. The smoothing parameter $h > 0$ is called the window width or bandwidth. Because h is fixed across the range of the estimation (1.1) will be called a global window width kernel estimator.

Kernel density estimators are a now becoming a popular tool for detecting and displaying distributional structure in populations and ongoing research, mainly concerned with the data-driven choice of the window width h, has made this methodology increasingly more practical (see e.g. Park and Marron, 1990). There are also many compelling reasons for using kernel estimators rather than the classical histogram density estimator. Several examples of practical kernel density estimation are given in Silverman (1986) and Izenman (1991). However, one problem which is very apparent in some of these examples is that certain density shapes can be difficult to estimate using (1.1). This notion is conveyed in Figure 1 where the underlying density is a strongly skewed density (density (n) as defined in Section 2). The underlying density corresponds to the dotted curve, while the three solid curves represent kernel density estimates based on a sample of size $n = 1000$ and window widths of 0.05, 0.15 and 0.45. The kernel function is the standard normal density. None of the estimates is close to the true density. The one with the smallest window width estimates the true density well near the peak, but perform terribly in the tails. The other estimates smooth the tail region much better, but severely underestimate the peak. It is clear that

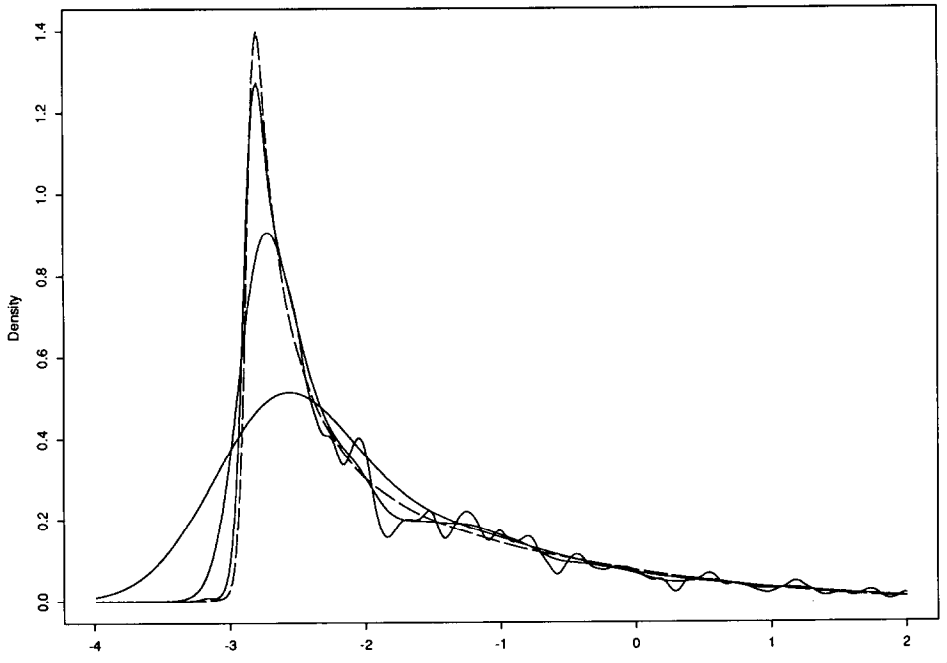

Fig. 1. Kernel density estimates of the Strongly Skewed Density based on a sample of size $n = 1000$ and with window widths $h = 0.05$, $h = 0.15$ and $h = 0.45$. The kernel is the standard normal density. The dashed line is the true density.

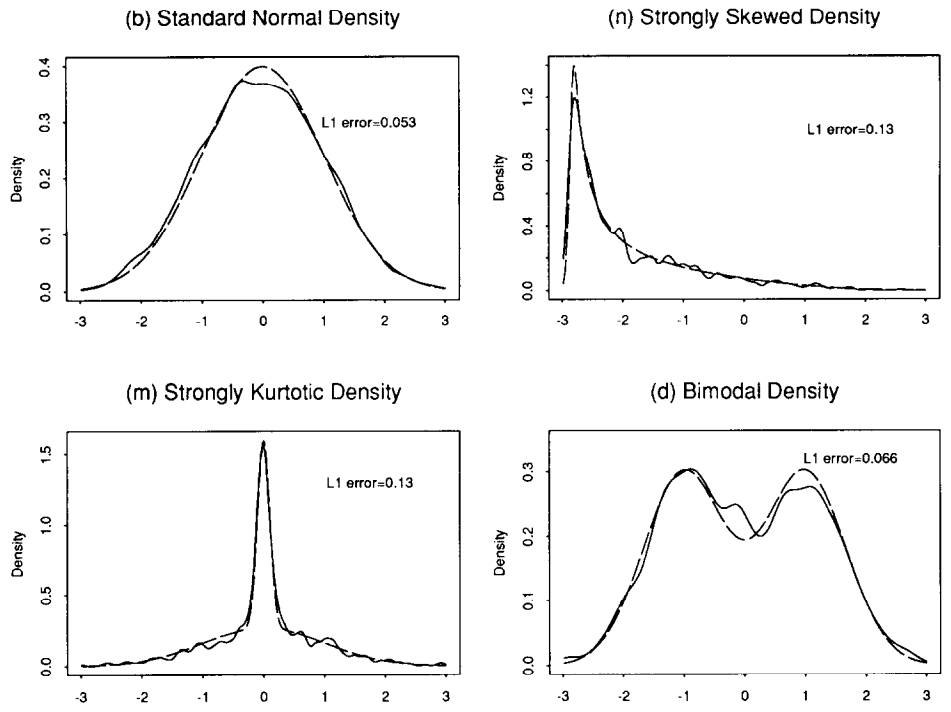

Fig. 2. Kernel density estimates using the L_1 optimal window width for $n = 1000$ based on the data set corresponding to the median of L_1 errors from a simulation having 500 replications. The solid line is the density estimate. The dashed line is the true density.

this density is very difficult to estimate using (1.1) since no choice of h will give a satisfactory estimate.

This point can be made more objectively through Figure 2. Each of these figures depict density estimates (solid line) based on samples of size $n = 1000$, but with h chosen to minimise the area, or L_1 distance, between (1.1) and the true density (dotted line). These densities are (b), (m), (n) and (d) as defined and discussed in Section 2. In an attempt to have each these figures represent a typical situation the sample used in each case was chosen to be the one that gave the median minimum area out of 500 replications. By looking at pictures like Figure 1 one gets the impression that smooth 'bell-shaped' densities, close to normality say, should be estimated fairly well by (1.1). On the other hand, we should expect (1.1) to perform less well if the target density is more complex in the sense that it has features such as high skewness, high kurtosis, multiple modes or discontinuous low-order derivatives.

This lack of flexibility of the kernel estimator is also shared by other delta sequence estimators such as the histogram and orthogonal series estimators since they are each based on local averaging with a global smoothing parameter. However, we will work with the kernel estimator because of its simplicity.

The purpose of this article is to develop a better understanding for when the global window width kernel estimate can be expected to perform well in

practice. We shall show that theory for the L_1 norm provides a particularly appealing way of measuring the complexity of a particular density in terms of how difficult it is to estimate using a global window width kernel estimator. The result of our L_1 analysis is a positive valued functional $Q(f)$ that depends only on the shape of f which can be thought of as measuring the degree of difficulty of estimating f via (1.1). We argue that measures of degree of difficulty based on the more popular and tractable L_2 loss are not as appealing since they depend on scale adjustment. These discussions are given in Section 2.

If a particular density is difficult to estimate then one way to overcome this problem is to transform the data so that the density of the transformed data is easy to estimate. The kernel estimator can then be applied to the transformed data and the density estimator of the original data can be obtained by back transformation. This procedure was first discussed in Devroye and Györfi (1985, Chapter 9), Silverman (1986, p. 28), but for more recent discussion see Wand, Marron and Ruppert (1991) and Ruppert and Wand (1991) where the transformation is chosen from an appropriate parametric family. In Section 3 we show how the $Q(f)$ functional can be used to assess the flexibility of parametric families of transformations by measuring the extent to which they can lower the degree of difficulty of the density that is estimated by (1.1).

2. Measuring ease of estimation

As seen in Figure 1 it is clear that certain densities are easier to estimate than others. We now address the question as to whether there is a natural formal way of ordering probability densities according to their estimation complexity.

The traditional way of studying the global performance of kernel density estimators is with the respect to the mean integrated squared error (MISE) given by

$$\mathrm{MISE}(h) = E\int \{\hat{f}(x; h) - f(x)\}^2 \, \mathrm{d}x.$$

Under the assumptions that f has a continuous, square integrable second derivative we have as $n \to \infty$,

$$\inf_{h>0} \mathrm{MISE}(h) \sim (5/4)G(f)A(K)n^{-4/5},$$

where $A(K) = \{(\int K(x)^2 \, \mathrm{d}x)^2 \int x^2 K(x) \, \mathrm{d}x\}^{1/5}$ and

$$G(f) = \left\{\int f''(x)^2 \, \mathrm{d}x\right\}^{1/5}.$$

It follows that densities with smaller values of the functional $G(f)$ will be easier to estimate in terms of asymptotic MISE. However, $G(f)$ is not scale-invariant so scale must be fixed in some way before densities can be compared through

this measure. Terell (1990) considers several scale measurements and derives minimum values of $G(f)$ when the scale is fixed. In the case of standard deviation, this is equivalent to replacing $G(f)$ by $\sigma(f)G(f)$ where $\sigma(f)$ is the standard deviation of f. Terrell's results show that the lowest possible value of $\sigma(f)G(f)$ is $\frac{1}{3}35^{1/5}$ and that this is achieved for the Beta(4, 4) density. However, the fact that certain scale measurements are more appropriate than others for a particular density makes it desirable to obtain a functional for measuring ease of estimation that is not dependent on the choice of scale measurement. This goal can be achieved by working with the L_1 norm as we will now show.

The expected L_1 loss or mean integrated absolute error (MIAE) of $\hat{f}(\cdot; h)$ is

$$\text{MIAE}(h) = E\int |\hat{f}(x; h) - f(x)|\,dx.$$

The L_1 error $\int |\hat{f}(x; h) - f(x)|\,dx$ has the simple interpretation of being the area between the two curves. In addition it is invariant under monotone transformations of the coordinate axes which makes it an appealing quantity for measuring the performance of transformation kernel density estimators (see Section 3). See Devroye and Györfi (1985) for a detailed exposition of L_1 loss for the density estimation problem.

Consider the class of densities possessing two continuous, integrable derivatives and a finite moment of order $1+\epsilon$ for some $\epsilon > 0$. Then Theorem 5.1 of Devroye and Györfi (1985, p. 78) and Theorem 2.1 of Hall and Wand (1988) imply that

$$\inf_{h>0} \text{MIAE}(h) \sim (1/2)^{1/5} Q(f) A(K) n^{-2/5}, \tag{2.1}$$

where

$$Q(f) = \inf_{u>0} u^{-1} \int f^{1/2}(x) \psi\left(\frac{u^5 f''(x)}{f(x)^{1/2}}\right) dx. \tag{2.2}$$

The function ψ is given by $\psi(t) = E|N - t|$ where N is a standard normal random variable.

The functional $Q(f)$ is both scale and location invariant and returns a positive number depending only on the shape of f. To see this, suppose that X has density f and $Y = (X-b)/a$ is a linear transformation of X having density g, where a and b are real constants. Then, noting that $g(x) = |a| f(ax+b)$ we have

$$Q(g) = \inf_{u>0} u^{-1} \int |a|^{1/2} f(ax+b)^{1/2} \psi\left(\frac{u^5 |a|^3 f''(ax+b)}{|a|^{1/2} f(ax+b)^{1/2}}\right) dx$$

$$= \inf_{|a|^{1/2}u>0} (|a|^{1/2}u)^{-1} \int f(y)^{1/2} \psi\left(\frac{(|a|^{1/2}u)^5 f''(y)}{f(y)^{1/2}}\right) dy = Q(f).$$

Because of this invariance and (2.1), $Q(f)$ provides a very appealing measure of how well f can be estimated. Of course, it is desirable to measure the complexity of *all* probability densities, not just those satisfying the above regularity conditions, so we should extend the definition of $Q(f)$ to accommodate all f. This can be done in the same way as was done by Devroye and Györfi (1985) for their functional $B(f)$ by setting

$$Q(f) = \inf_{u>0} \limsup_{a \to 0} u^{-1} \int f^{1/2}(x) \psi \left(\frac{u^5 (f * \varphi_a)''(x)}{f(x)^{1/2}} \right) \mathrm{d}x, \tag{2.3}$$

where $*$ denotes ordinary convolution and φ is a compactly supported infinitely differentiable density and $\varphi_a(x) = \varphi(x/a)/a$. The function φ is sometimes called a *mollifier*. This definition can be shown to be equivalent to (2.2) for densities satisfying the above regularity conditions and to also be independent of the choice of the mollifier. Theorem 1 of Devroye and Wand (1993) shows that the following extension of (2.1),

$$\lim_{n \to \infty} n^{2/5} \inf_{h>0} \mathrm{MIAE}(h) = (1/2)^{1/5} Q(f) A(K), \tag{2.4}$$

holds for all densities that can be written as a finite mixture of unimodal densities, or that have a finite moment of order $1 + \epsilon$ for some $\epsilon > 0$. The limit in (2.4) also caters for cases where $Q(f) = \infty$ which indicates that the best possible rate of convergence of (1.1) for such densities is slower than $n^{-2/5}$. Such a situation arises if either $\lim \sup_{a \to 0} \int |(f * \varphi_a'')(x)| \, \mathrm{d}x = \infty$ or $\int f^{1/2}(x) \, \mathrm{d}x = \infty$. For example, the former of these conditions is satisfied by the uniform density or any other density with a simple discontinuity, and the latter by the Cauchy density and other densities with a heavier tail.

The functional $Q(f)$ is a theoretically pleasing alternative to L_2-based measures of complexity. Of course, its form is not as simple as $G(f)$, yet for many important families of densities it is not difficult to calculate $Q(f)$ numerically.

We computed values of $Q(f)$ for a collection of common densities and some of the normal mixture densities used as examples in Marron and Wand (1992). The values are shown with plots of each density in Figure 3. Most of the densities are either defined in Marron and Wand, or have well known definitions. The extreme value density has formula $f(x) = \mathrm{e}^x \, \mathrm{e}^{-\mathrm{e}^x}$ while the lognormal density is that of an exponentiated standard normal random variable. For most of the densities in Figure 3 $Q(f)$ was computed using trapezoidal integration with successive doubling of the integration mesh until convergence was obtained. This also involved the truncation of the tails of the density which was not a problem for the light-tailed densities. The exceptions were the heavy-tailed Student's t_3 and lognormal densities for which $Q(f)$ was computed by simulation with 5×10^7 random pairs. The values of $Q(f)$ for the Laplace and isosceles triangular density were computed using analytic arguments as in Devroye and Györfi (1985, Chapter 5).

Out of all of the common density families it appears that the symmetric bell-shaped Beta densities give the lowest value of $Q(f)$ and within this family the parameters (5.3, 5.3) are approximately those for which $Q(f)$ is lowest with

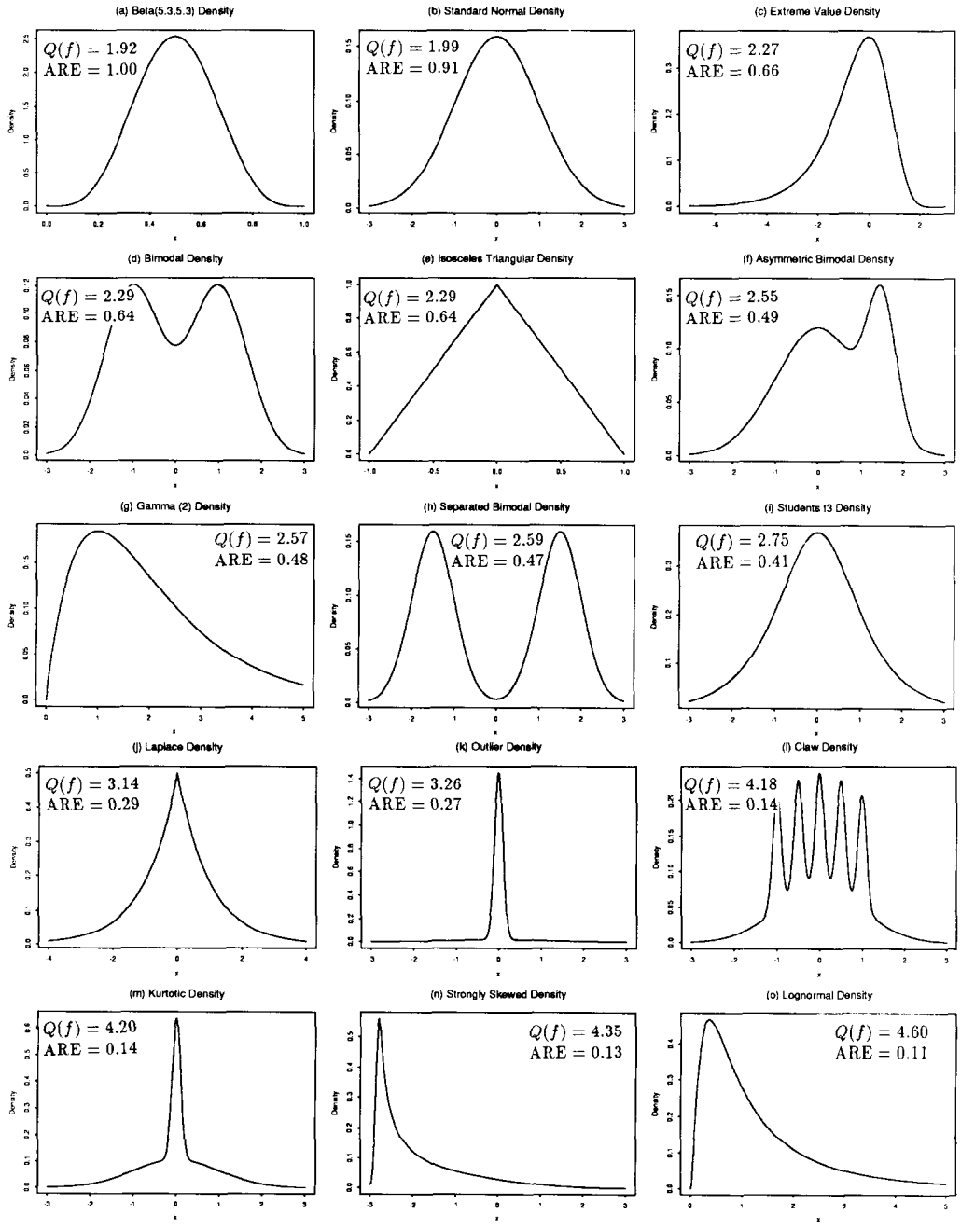

Fig. 3. Plots of 15 densities along with corresponding values of $Q(f)$ and $\mathrm{ARE}(g, f)$ where g is the Beta(5.3, 5.3) density.

a value of about 1.92. The Gaussian density is not far behind with $Q(f) \approx 1.99$. It is interesting to see from Figure 3 how $Q(f)$ is affected by features such as skewness, kurtosis, "corners" and discontinuities.

An appropriate measure of relative ease of estimation is the asymptotic relative efficiency of g compared to f given by

$$\mathrm{ARE}(g, f) = \{Q(g)/Q(f)\}^{5/2}$$

which, in view of (2.1) has a simple equivalent sample size interpretation. For example, if $\mathrm{ARE}(g, f) = 0.25$ then only a quarter as many observations are required to achieve a certain minimum asymptotic MIAE when estimating g than are needed for estimating f with the same precision. In Figure 3 we also include values of ARE of the Beta(5.3, 5.3) density compared to each other density.

It is particularly interesting to note how high $Q(f)$ is for the Kurtotic Density (m), the Strongly Skewed Density (n) and the Lognormal Density (o) and is even higher than for the Claw Density (l). There is definitely a high price to pay for estimating heavily skewed or kurtotic densities, with these examples showing a six to ten fold increase in asymptotically equivalent sample size compared to the Gaussian density.

Another interesting feature of the results in Figure 3 is that the Separated Bimodal Density (h) is more difficult to estimate than the related Bimodal Density (d). The separation of the modes means that the estimation of (h) should be roughly equivalent to estimating two Gaussian peaks, although with just one data set. Thus, from a sample size point of view we would expect it to be about twice as difficult to estimate (h) than (b), and density (d) to be somewhere intermediate between the two. The effect of several modes on the functional Q can be seen more clearly by taking f to be a twice continuously differentiable density having support on [0,1] (such as one of the smooth Beta densities, for example). Consider the m-modal density based on f and given by

$$f_m(x) = m^{-1} \sum_{i=1}^{m} f(x - 2i)$$

that is made up of m juxtaposed rescaled versions of f. It is a trivial exercise to show that

$$Q(f_m) = m^{2/5} Q(f)$$

which implies that

$$\mathrm{ARE}(f, f_m) = 1/m.$$

Thus, unsurprisingly, m times as much data are needed to estimate m similar modes using the kernel estimator than just one such mode. Devroye and Györfi (1985, p. 111) derive an analogous result for their functional $B^*(f)$.

One could also compute the L_2-based measure $\sigma(f)G(f)$ for the densities in Figure 3. However, because of the inappropriateness of standard deviation for scale adjustment of many of these densities the results would tend to be

misleading. This problem is most easily seen by considering the separated bimodal density (h). It is easy to see that the value of $\sigma(f)$ can be made arbitrarily large by moving the modes of the mixture sufficiently far apart. This means that $\sigma(f)G(f)$ could be made to equal any large number by increasing the separation of the modes, even though the degree of estimation difficulty of the density remains the same. Similar comments apply to other common scale measures such as the interquartile range. On the other hand, $Q(f)$ would hardly change from the value of 2.59, no matter how much the modes are separated. The inappropriateness of standard deviation also applies to other densities exhibiting a departure from normality, such as heavy skewness and kurtosis.

An intriguing question is: What is the lowest possible value of $Q(f)$ and for which density is the minimum attained? An answer to this would tell us which density is easiest to estimate in terms of asymptotic L_1 error and at the same time provide a sharp universal lower bound for this error in the sense that for all $f \in \mathscr{D}$, the class of all probability densities,

$$\liminf_{n\to\infty} n^{2/5} \inf_{h>0} \mathrm{MIAE}(h) \geq (1/2)^{1/5} A(K) \inf_{f\in\mathscr{D}} Q(f)$$

and this bound is obtainable for some $f \in \mathscr{D}$. Furthermore, we have from Theorem 5.2 of Devroye and Györfi (1985, p. 79) that for all non-negative kernels K, $A(K)$ has minimum value $(9/125)^{1/5}$ when K is the Bartlett–Epanechnikov kernel $K(x) = (3/4)(1-x^2)$, $|x| \leq 1$. This fact leads to the lower bound

$$\liminf_{n\to\infty} n^{2/5} \inf_{h>0} \mathrm{MIAE}(h) \geq (9/250)^{1/5} \inf_{f\in\mathscr{D}} Q(f) \tag{2.5}$$

for all non-negative kernels. Results of Devroye and Györfi (1985, pp. 78–79) show that for all densities f

$$C_L B^*(f) \leq Q(f) \leq C_U B^*(f) \tag{2.6}$$

where $C_L \approx 1.18$ and $C_U \approx 1.61$ are universal constants satisfying $C_U/C_L \approx 1.34$ and

$$B^*(f) = \left[\frac{1}{2}\left\{\int f(x)^{1/2}\,\mathrm{d}x\right\}^4 \sup_{a>0} \int |(f * \varphi_a)''(x)|\,\mathrm{d}x\right]^{1/5}.$$

In their Theorem 5.3 they also show that $B^*(f)$ is minimal when f is the Isosceles Triangular Density. On face value, this fact combined with (2.6) suggests that $Q(f)$ may be minimised when f is the isosceles triangular, however, as shown in Figure 3, this is not the case and in fact for this density $Q(f)$ and its upper bound $C_U B^*(f)$ can be shown to coincide. A consequence of the minimum value of $B^*(f)$ and (2.6) is that $\inf_{f\in\mathscr{D}} Q(f) \geq 1.708$ for all f, yet inspection of Figure 3 suggests that this bound is far from sharp. It is possible to treat the minimisation of $Q(f)$ as a variational calculus problem. However, the Euler–Lagrange equation is highly non-linear and does not appear to have an analytic solution. The numerical solution is also an open

problem. At this stage all we can do is speculate that $\inf_{f \in \mathcal{D}} Q(f)$ is approximately 1.9 and that the easiest density to estimate is bellshaped and close to the Beta(5, 5) density. If this speculation is proven to be true then from (2.5) we would get that all non-negative kernel estimators can have an MIAE rate of convergence no faster than about $0.98n^{-2/5}$ which is slightly higher than the existing theoretical lower bound of $0.87n^{-2/5}$ given in Theorem 5.2 of Devroye and Györfi (1985, p. 79).

3. Parametric transformation kernel estimators

A simple proposal for increasing the flexibility of the kernel density estimator is the transformation kernel estimator introduced by Devroye and Györfi (1985, Chapter 9) and Silverman (1986). Wand, Marron and Ruppert (1991) and Ruppert and Wand (1991) recently investigated choosing the transformation from an appropriate parametric family. Formally, let $\{T_\lambda: \lambda \in \Lambda\}$ be a parametric family of real-valued, increasing transformations on the support of f that includes the identity transformation. Smoothness assumptions such as the differentiability of T_λ and its inverse are usually required. For a particular choice of λ we obtain the transformed sample $Y_1, \ldots, Y_n$ where $Y_i = T_\lambda(X_i)$. The density of the transformed sample is

$$g(x; \lambda) = f\{T_\lambda^{-1}(x)\}(T_\lambda^{-1})'(x)$$

which may be estimated by the kernel estimator (1.1) applied to $Y_1, \ldots, Y_n$:

$$\hat{g}(x; \lambda, h) = n^{-1} \sum_{i=1}^{n} K_h(x - Y_i).$$

Then our estimate of f is the inverse transform

$$\hat{f}(x; \lambda, h) = n^{-1} \sum_{i-1}^{n} K_h\{T_\lambda(x) - T_\lambda(X_i)\}T_\lambda'(x).$$

The transformation invariance property of the L_1 metric works to our advantage here since

$$\int |\hat{f}(x; \lambda, h) - f(x)| \, dx = \int |\hat{g}(x; \lambda, h) - g(x)| \, dx \tag{3.1}$$

so that for finite $Q(g(\cdot; \lambda))$,

$$\inf_{h>0} E\int |\hat{f}(x; \lambda, h) - f(x)| \, dx \sim (1/2)^{1/5} Q(g(\cdot; \lambda)) A(K) n^{-2/5}. \tag{3.2}$$

Therefore, the transformation kernel estimator inherits the L_1 performance of the kernel estimate of $g(\cdot; \lambda)$. Clearly λ should be chosen to minimise $Q(g(\cdot \lambda))$ if L_1 minimisation is the goal.

Let $\mathcal{G}$ be the collection of all densities $g(x; \lambda)$, and define

$$\mathrm{ARE}(\mathcal{G}, f) = \inf_{\lambda} \mathrm{ARE}(g(\cdot; \lambda), f) = \left[\frac{\inf_{\lambda \in \Lambda} Q(g(\cdot; \lambda))}{Q(f)} \right]^{5/2}.$$

Then the minimum asymptotic MIAE is the same whether we use n observations and estimate f directly or whether we use $\mathrm{ARE}(\mathcal{G}, f)n$ observations and estimate f using the transformation kernel estimator with the optimal transformation from $\{T_\lambda: \lambda \in \Lambda\}$.

We shall illustrate this with two examples. Firstly, suppose f is the lognormal density for which $Q(f) = 4.58$ and that the family of transformations is the Box–Cox family

$$T_\lambda(x) = \begin{cases} (x^\lambda - 1)/\lambda, & \lambda \neq 0 \\ \ln(x), & \lambda = 0. \end{cases}$$

Then the value of λ that minimises $Q(g(\cdot; \lambda))$ is $\lambda = 0$ for which the density $g(\cdot; 0)$ is the Gaussian density having $Q(g(\cdot; 0)) = 1.99$. Therefore

$$\mathrm{ARE}(\mathcal{G}, f) = 0.12$$

which indicates that considerable gains can be made by the transformation kernel estimator in this case. A second example comes from the kurtosis-reducing family of transformations proposed by Ruppert and Wand (1992) given by (with $\lambda = (\alpha, \sigma)$)

$$T_\lambda(x) = \alpha x + (1 - \alpha)(2\pi)^{1/2}\sigma\{\Phi(x/\sigma) - \tfrac{1}{2}\}. \tag{3.3}$$

This family was shown by these authors to allow very good estimation of approximately symmetric kurtotic densites such as the Kurtotic Density (m) that has $Q(f) = 4.20$. Searching over a grid of (α,σ) values with each parameter

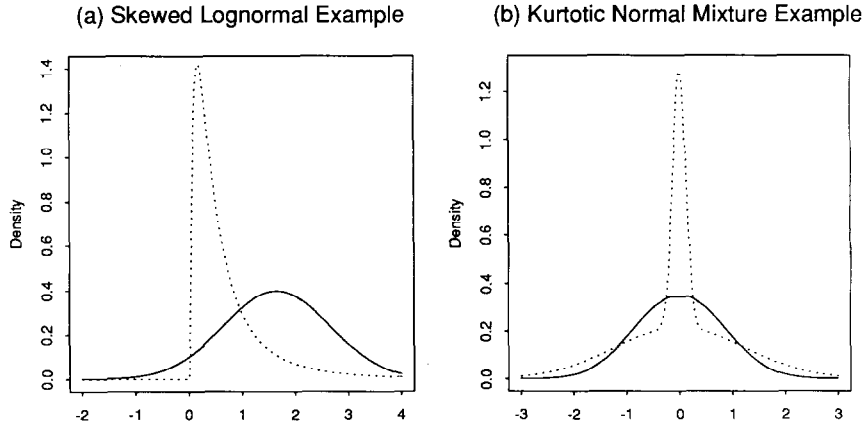

Fig. 4. Plots of densities corresponding to (a) the lognormal example and (b) the kurtotic normal mixture example. Broken lines show the original density and solid lines show the optimally transformed density. All densities have been scaled to have unit variance.

ranging over {0.01, 0.02, 0.03,...} it was found that the value of λ minimising $Q(g(\cdot;\lambda))$ was $\lambda = (0.17, 0.10)$ and the corresponding value of $Q(g(\cdot; 0.17, 1.10))$ is 2.05. This relatively low value of Q indicates that it is possible to get high quality estimation of (m) by using family (3.3). From this we obtain the approximation

$$\mathrm{ARE}(\mathscr{S}, f) = 0.17$$

which, once again, represents a considerable improvement through the use of transformation strategies. We also plotted the densities f and corresponding optimal $g(\cdot;\lambda)$ in Figures 4a and 4b. All densities have been scaled to have unit variance. Further appreciation of the possible improvements due to using the parametric transformation kernel estimator can be obtained by inspection of Figure 3 of Wand et al. (1992) and Figure 3b of Ruppert and Wand (1992) which respresent average case estimates of each of the densities in the above example using the transformation kernel estimator and data-driven selection of the parameters when $n = 200$.

Acknowledgements

This research was supported by NSERC Grant A3456 and ONR Grant N00014-90-J-1176. we are also grateful to Professor Steve Marron for allowing the use of several of his $\mathrm{GAUSS}_{\mathrm{TM}}$ routines in the simulation and to he and Dr. Chris Jones for their detailed comments.

References

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

Devroye, L. and M.P. Wand, On the effect of density shape on the performance of its kernel estimate, *Statistics* (1993) to appear.

Hall, P. and J.S. Marron, Estimation of integrated squared density derivatives, *Statist. Probab. Lett.*, **6** (1987) 109–115.

Hall, P. and M.P. Wand, Minimizing L_1 distance in nonparametric density estimation, *J. Mult. Analysis.*, **26** (1988) 59–88.

Izenman, A.J., Recent developments in nonparametric density estimation, *J. Amer. Statist. Assoc.*, **86** (1991) 205–224.

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

Marron, J.S. and M.P. Wand, Exact mean integrated squared error, *Ann. Statist.*, **20** (1992) 712–736.

Park, B.U. and J.S. Marron, Comparison of data-given bandwidth selectors, *J. Amer. Statist. Assoc.*, **85** (1990) 66–72.

Ruppert, D. and M.P. Wand, Correcting for kurtosis in density estimation, *Austral. J. Statist.* (1992) 19–29.

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

Terrell, G.R., The maximal smoothing principle in density estimation, *J. Amer. Statist. Assoc.*, **85** (1990) 470–477.

Wand, M.P., J.S. Marron and D. Ruppert, Transformations in density estimation, *J. Amer. Statist. Assoc.*, **86** (1991) 343–352.