Optimal Shapes for Kernel Density Estimation: An Historical Footnote

Michael W. Trosset

September, 1991

TR91-28

Optimal Shapes for Kernel Density Estimation: An Historical Footnote

Michael W. Trosset
Consultant, P.O. Box 40993, Tucson, 85717-0993, U.S.A.

October 4, 1991

Abstract

In the early years of kernel density estimation, Watson and Leadbetter (1963) attempted to optimize kernel shape for fixed sample sizes by minimizing the expected L^2 distance between the kernel density estimate and the true density. Perhaps out of technical necessity, they did not impose the constraint that the kernel be a probability density function. The present paper uses recent developments in the theory of infinite programming to successfully impose that constraint. Necessary and sufficient conditions for solution of the constrained problem are derived. These conditions are not trivial; however, they can be exploited to demonstrate that symmetric densities with sufficiently light tails have optimal kernels with compact support.

Keywords: Density estimation, kernel estimator, kernel shape.

1 Introduction

Kernel estimators of probability density functions were first proposed by Rosenblatt (1956),who analyzed the behavior of the so-called shifted histogram. For general kernel shapes, the detailed explication of these estimators is due to Parzen (1962). Subsequently, a number of authors have considered the problem of optimizing the shape of the kernel. Comprehensive surveys of this literature can be found in the books of Tapia and Thompson (1978), Wertz (1978), Prakasa Rao (1983), Silverman (1986), and Thompson and Tapia (1990).

One of the first investigations of the kernel shape problem was undertaken by Watson and Leadbetter (1963). Let $X_1, \ldots, X_n$ be independent and identically distributed random variables having density $f \in L^2(-\infty, +\infty)$. Given a sample $x_1, \ldots, x_n$, f is to be estimated by estimates of the form

$$\hat{f}_n(x) := \frac{1}{n} \sum_{i=1}^{n} K_n(x - x_i) \,, \tag{1.1}$$

where $K_n \in L^2(-\infty, +\infty)$. This is a very general formulation: note that there is no assumption, as is usual for kernel estimators, that $K_n(y) = K(y/h_n)/h_n$.

Watson and Leadbetter analyzed the unconstrained optimization problem

(UWL)

$$\underset{K_n \in L^2(-\infty,+\infty)}{\text{minimize}} \quad J_0(K_n) := E\|\hat{f}_n - f\|^2_{L^2(-\infty,+\infty)} \cdot$$

Because this problem admits any square-integrable function as a possible kernel, solutions to it may not be probability densities. As an example, for the Gamma density $f(x) = xe^{-x}$ Watson and Leadbetter give the optimal kernel as

$$K_n^*(x) = [n^{3/4}/2(n-1)^{1/2}] \exp(-xn^{1/4} \sin \alpha/2) \cos(|x|n^{1/4} \cos \alpha/2 - \alpha/2) \ .$$

This means, of course, that the corresponding probability density estimates may not be probability densities themselves.

The use of kernels that may assume negative values is sometimes advocated as a technique for reducing the bias of kernel density estimators. The relevant literature was reviewed by Silverman (1986), who concluded:

> However, if the interest is in discerning the general shape of the density, it may be better to tolerate some bias rather than risk the introduction of spurious fluctuations in the tails and the possible exaggeration of such things as modes in the data. In some applications it is essential that the density estimate is itself a probability density function and so the use of a non-negative

kernel is essential. Cautious users are best advised to stick to symmetric non-negative unimodal kernels. (p. 69)

How is it that Watson and Leadbetter did not insist that their kernels be probability density functions? Rosenblatt (1956) had proposed "a class of estimates of the density function" that is precisely those kernel estimates whose kernels are density functions, and had even noted that "all estimates of this form are themselves density functions." He had also argued in favor of kernels symmetric about zero. Parzen's (1962) results had assumed that the kernel be symmetric and integrate to unity, although not that it be nonnegative; nevertheless, the seven examples of kernels listed in his paper are all probability density functions. Watson and Leadbetter, however, referenced Parzen (1962) only in passing and Rosenblatt (1956) not at all. Instead, their work was built on the papers of Whittle (1958) and Parzen (1958).

According to Parzen, Whittle's (1958) is the only paper on the subject of probability density estimation to appear between Rosenblatt (1956) and Parzen (1962). Whittle, who did not reference Rosenblatt (1956), analyzed density estimators of the form (1.1) from a somewhat controversial Bayesian perspective. In so doing he conceded: "This hypothesis introduces the idea of

an *a priori* distribution of the ordinates $f(x)$, an idea that often encounters a mixed reception." This was rather prescient of him, as Watson's and Leadbetter's agenda evidently was to replace Whittle's Bayesian methodology with the frequentist methodology developed by Parzen (1958) for the problem of estimating the spectral density function of a stationary time series. In fact, they explicitly stated:

> The estimators have the form considered by Whittle, but the criterion employed by Parzen [(1958)] will be used instead of the expectation (over the prior distributions) of the sampling mean square error. (p. 480)

Of course, Parzen (1958) also inspired Parzen (1962), which has since become the standard approach to kernel density estimation.

There is another possibility, and that is that Watson and Leadbetter simply posed the problem that they were able to solve. Their analysis uses Parseval's formula to transform problem (UWL) to the Fourier domain, where it can be solved in straightforward fashion. However, the constraints that the kernel be a probability density function are wholly unmanageable in the Fourier domain, so that imposing them requires that one retain the original

formulation of the problem.

The present paper imposes on problem (UWL) constraints that make the kernel a probability density function. The resulting infinite programming problem is then analyzed by application of a Lagrange multiplier theorem derived for such problems by Tapia and Trosset (1991). Because this approach differs from Watson's and Leadbetter's Fourier approach, we will first consider the unconstrained problem and note the connection between our characterization of optimal kernels and theirs.

It should be emphasized that the study of optimal kernel shape, with or without constraints, assumes knowledge of the unknown density f. Were such knowledge available, estimation would be unnecessary. Therefore, the reason to investigate optimal kernels is to acquire knowledge of general properties that will facilitate the selection of actual kernels in practice. In fact, it will be demonstrated that, for a large class of densities, the optimal kernel is compactly supported.

Finally, it should be noted that the sample size n is fixed in all of our results. Most results about the optimal choice of a kernel simplify the imposed optimality criterion with asymptotic approximations. Following Watson and Leadbetter, we do not do so here; however, retaining the dependence on n

predictably complicates the analyses.

2 The Unconstrained Problem

Because J_0 is a strictly convex functional, a necessary and sufficient condition for $K_n^* \in L^2(-\infty, +\infty)$ to be the unique solution of problem (UWL) is that

$$J_0'(K_n^*)(\eta) = 0 \qquad \forall\ \eta \in L^2(-\infty, +\infty)\ , \tag{2.1}$$

where $J_0'(K)(\eta)$ denotes the first Gâteaux variation of J_0 at K in the direction η. (An excellent reference for this and most of the optimization theory utilized in this paper is Appendix I of Tapia and Thompson (1978).) Define the operator Ψ by

$$\Psi K := f * \bar{f} - \frac{n-1}{n} K * f * \bar{f} - \frac{1}{n} K\ ,$$

where $\bar{f}(x) := f(-x)$. Some computation reveals that (2.1) is equivalent to

$$\Psi^*(x) := [\Psi K_n^*](x) = 0 \qquad \forall\ x \in (-\infty, +\infty)\ . \tag{2.2}$$

Equation (2.2) is an implicit characterization of the optimal kernel K_n^*. To represent K_n^* explicitly requires that K_n^* and $f * \bar{f}$ be deconvolved, an operation that is obviously most easily performed in the Fourier domain. It is easily verified that, if one Fourier transforms (2.2) and solves for the

characteristic function of K_n^* , then one obtains the explicit representation given by Watson and Leadbetter, viz.

$$\Phi_{K_n^*}(t) = \frac{|\Phi_f(t)|^2}{(1/n) + [(n-1)/n]|\Phi_f(t)|^2} ,$$

where Φ denotes the characteristic function of the density function that subscripts it. This observation makes clear the superiority of the Fourier approach, at least for analyzing the unconstrained problem.

At this point, it is interesting to observe that the solution K_n^* to problem (UWL), while not necessarily nonnegative, must necessarily integrate to unity. This result can be obtained by integrating both sides of equation (2.2), which leads to

$$1 - \int_{-\infty}^{\infty} K_n^*(x)dx = 0 .$$

Alternatively, one can reach the same conclusion from Watson's and Leadbetter's formulation by computing

$$\int_{-\infty}^{\infty} K_n^*(x)dx = \Phi_{K_n^2}(0) = \frac{1}{(1/n) + (1-n)/n} = 1 .$$

Finally, before imposing the constraints that K_n be a probability density function, we demonstrate that none of these constraints are necessarily binding, i.e., that for some f, K_n^* is automatically a density. Suppose that

$K \in L^2(-\infty, +\infty)$ is a known probability density function symmetric about zero. Let $r := (n-1)/n$ and

$$g := (1-r)K * \{1 + rK + r^2 K * K + \cdots\} . \tag{2.3}$$

It is easily checked that g is a square-integrable probability density function, symmetric about zero and satisfying

$$g(x) - \frac{n-1}{n}[K * g](x) - \frac{1}{n}K(x) = 0 \qquad \forall\ x \in (-\infty, +\infty) . \tag{2.4}$$

If there exists a square-integrable probability density function f such that $f * \bar{f} = g$, then (2.4) is identical to (2.2) and we can conclude that the kernel K solves problem (UWL) for the density f.

The equation $f * \bar{f} = g$ can be solved whenever the square root of the characteristic function of g is itself a characteristic function. Unfortunately, the standard techniques for making such a determination are not well-suited to densities of form (2.3). Moreover, it is unclear how to characterize densities of this form, although it should be noted that they must be everywhere strictly positive.

As a tractable example of a density for which the constraints are not binding, consider the double exponential density $f(x) = \exp(-|x-a|/b)/2b$, which we denote by $DE(a,b)$. Let K_n be the double exponential density

$DE(0, 2bn^{-1/2})$. Then some computation establishes that $\Psi^*(x) \equiv 0$ for these choices of f and K_n, so that the unique optimal kernel for $DE(a, b)$ is $DE(0, 2bn^{-1/2})$.

3 The Constrained Problem

We now restrict attention to kernels that are probability density functions. For technical reasons that will become apparent, we also replace $L^2(-\infty, +\infty)$ with a better behaved space of possible kernels. Let $W = W^{1,2}(-\infty, +\infty)$ denote the classical Sobolev space consisting of all functions $u \in L^2(-\infty, +\infty)$ having generalized (in the sense of Schwartz distributions) derivative $D^1 u$ that is also square-integrable. It is well known that W is a Hilbert space if equipped with the inner product

$$\langle u, v \rangle_W := \int_{-\infty}^{\infty} u(x)v(x)dx + \int_{-\infty}^{\infty} D^1 u(x) D^1 v(x) dx \,, \tag{3.1}$$

and moreover that W is a proper functional Hilbert space, i.e. that point evaluation is a continuous operation. See Adams (1975) for an introduction to Sobolev spaces; the quoted results are in Tapia and Thompson (1978).

The constrained kernel shape problem (CWL) that we consider is

$$\begin{aligned} \underset{K_n \in W}{\text{minimize}} \quad & J_1(K_n) := E\|\hat{f}_n - f\|_W^2 \\ \text{subject to} \quad & K_n(x) \geq 0 \qquad \forall\ x \in (-\infty, +\infty) \\ & \int_{-\infty}^{\infty} K_n(x)dx = 1\ . \end{aligned}$$

The constraint set for this problem is obviously convex; since point evaluation is continuous, it is also closed. Furthermore, it is easily verified (by computing the second Gâteaux variation) that J_1 is a uniformly convex functional. It follows that problem (CWL) has exactly one solution, which we denote by K_n^*.

If one views problem (CWL) as an infinite programming problem, then it is natural to attempt to characterize K_n^* by appealing to the theory of Lagrange multipliers. It is this observation that motivated Tapia and Trosset (1991) to develop a multiplier theory for such problems. Their results, however, require certain hypotheses not satisfied in the present case. Therefore, our strategy will be to approximate problem (CWL) by problems for which we can obtain necessary and sufficient conditions for solution.

For $a < b$, let $W[a, b]$ denote the Sobolev space defined as was W, but on $[a, b]$ instead of on $(-\infty, +\infty)$. For $K_n \in W[a, b]$, let $\bar{K}_n$ be the extension

of K_n to $(-\infty, +\infty)$ defined by $\bar{K}_n(x) = 0$ if $x \notin [a, b]$. Note that $\bar{K}_n \in L^2(-\infty, +\infty)$, but that it may be discontinuous at $x = a, b$. Tapia and Trosset demonstrated the applicability of their multiplier theorem to the constrained optimization problem

$$\begin{aligned} \underset{K_n \in W[a,b]}{\text{minimize}} \quad & J_0(\bar{K}_n) \\ \text{subject to} \quad & K_n(x) \geq 0 \qquad \forall\ x \in [a, b] \\ & \int_a^b K_n(x)dx = 1 \ . \end{aligned}$$

This formulation retains Watson's and Leadbetter's L^2 criterion, but since J_0 is not uniformly convex with respect to the Sobolev norm, there is no guarantee that a solution exists.

We would like to replace J_0 with J_1 in Tapia's and Trosset's problem, but J_1 is not continuous on the set of $\bar{K}_n$. To remedy this difficulty, let

$$W[r] := \{w \in W : \text{supp } w \subseteq [-r, +r]\}$$

for $r > 0$. The set $W[r]$ is a closed subspace of W and hence inherits all of its properties. Later, we will let $r \to \infty$; for the present, we would like to

characterize the unique solution K_n^r of problem (CWL)[r]:

$$\begin{aligned} \underset{K_n \in W[r]}{\text{minimize}} \quad & J_1(K_n) \\ \text{subject to} \quad & K_n(x) \geq 0 \qquad \forall \ x \in [-r, +r] \\ & \int_{-\infty}^{\infty} K_n(x)dx = 1 \ . \end{aligned}$$

Unfortunately, because $K_n(\pm r) = 0$ is not free to vary, the set of constraint gradients will not satisfy one of Tapia's and Trosset's hypotheses.

In order that K_n be free to vary at each of the constraints, fix $0 < \epsilon < r$, let $W_\epsilon[r]$ denote the closed subspace of $W[r]$ in which w is linear on $(-r, -r + \epsilon)$ and also on $(r - \epsilon, r)$ and let $K_n^{r,\epsilon}$ denote the unique solution of problem (CWL[$r; \epsilon$]):

$$\begin{aligned} \underset{K_n \in W_\epsilon[r]}{\text{minimize}} \quad & J_1(K_n) \\ \text{subject to} \quad & K_n(x) \geq 0 \qquad \forall \ x \in [-r + \epsilon, r - \epsilon] \\ & \int_{-\infty}^{\infty} K_n(x)dx = 1 \ . \end{aligned}$$

Tapia's and Trosset's multiplier theorem does apply to this problem, and it is simply a matter of replacing $\nabla J_0(K_n)$ with $\nabla J_1(K_n)$ to modify their first order conditions from the case of J_0 to the case of J_1. We thus obtain the following result.

Lemma 3.1 *Fix $0 < \epsilon < r$. Given $K_n^{r,\epsilon} \in W_\epsilon[r]$, let $\Psi^{r,\epsilon}(x) := [\Psi K_n^{r,\epsilon}](x)$. For $K_n^{r,\epsilon}$ to be the unique solution of problem* (CWL$[r,\epsilon]$)*, it is both necessary and sufficient that there exist a function $u^{r,\epsilon}$ and a real number $\lambda^{r,\epsilon}$ such that*

(a) $\Psi^{r,\epsilon}(x) + D^1\Psi^{r,\epsilon}(x) = \lambda^{r,\epsilon} - u^{r,\epsilon}(x) \qquad \forall\ x \in (-\infty, +\infty)\,,$

(b) $K_n^{r,\epsilon}(x) \geq 0 \qquad \forall\ x \in [-r+\epsilon, r-\epsilon]\,,$

(c) $\int_{-\infty}^{\infty} K_n^{r,\epsilon}(x)dx = 1\,,$

(d) $K_n^{r,\epsilon}(x)u^{r,\epsilon}(x) = 0 \qquad \forall\ x \in [-r+\epsilon, r-\epsilon]\,,$

(e) $u^{r,\epsilon}(x) \geq 0 \qquad \forall\ x \in [-r+\epsilon, r-\epsilon]\,.$

Next, we approximate problem (CWL$[r]$) with problems (CWL$[r;\epsilon]$) by letting $\epsilon \to 0$.

Lemma 3.2 *Fix $r > 0$. Let $K_n^r \in W[r]$ denote the unique solution of problem* (CWL$[r]$) *and let $\Psi^r(x) := [\Psi K_n^r](x)$. Then*

(i) $\lim_{\epsilon \to 0} \|K_n^{r,\epsilon} - K_n^r\|_w = 0.$

(ii) *K_n^r is the unique element of $W[r]$ for which there exists a func-*

tion u^r and a real number λ^r such that

$$\begin{aligned}
&\text{(a)} \quad \Psi^r(x) + D^1\Psi^r(x) = \lambda^r - u^r(x) && \forall\ x \in (-\infty, +\infty)\,, \\
&\text{(b)} \quad K_n^r(x) \geq 0 && \forall\ x \in [-r, +r]\,, \\
&\text{(c)} \quad \textstyle\int_{-\infty}^{\infty} K_n^r(x)dx = 1\,, && \\
&\text{(d)} \quad K_n^r(x)u^r(x) = 0 && \forall\ x \in [-r, +r]\,, \\
&\text{(e)} \quad u^r(x) \geq 0 && \forall\ x \in [-r, +r]\,.
\end{aligned}$$

Proof. Let $W^+[r]$ denote the feasible set for problem (CWL[r]) and let $W_\epsilon^+[r]$ denote the feasible sets for problems (CWL[r, ϵ]) respectively. We first demonstrate that $\underset{\epsilon>0}{\cup} W_\epsilon^+[r]$ is dense in $W^+[r]$.

Given $w \in W^+[r]$, let

$$W_\epsilon(x) := \left\{ \begin{array}{ll} (x+r)\ c\ w(-r+\epsilon)/\epsilon & x \in [-r, -r+\epsilon] \\ c\ w(x) & x \in [-r+\epsilon, r-\epsilon] \\ (x-r)\ c\ w(r-\epsilon)/(-\epsilon) & x \in [r-e, r] \end{array} \right\}\ .$$

It is easily verified that $w_\epsilon \in W_\epsilon^+[r]$ and that $\lim_{\epsilon\to 0} \|w_\epsilon - w\|_W = 0$.

Now, as $\underset{\epsilon>0}{\cup} W_\epsilon^+[r]$ is dense in $W^+[r]$, choose $w_\epsilon \in W_\epsilon^+[r] \to K_n^r$ as $\epsilon \to 0$. By continuity, $\lim_{\epsilon\to 0} J_1(w_\epsilon) = J_1(K_n^r)$. However, $J_1(w_\epsilon) \geq J_1(K_n^{r,\epsilon}) \geq J_1(K_n^r)$, so we must also have $\lim_{\epsilon\to 0} J_1(K_n^{r,\epsilon}) = J_1(K_n^r)$. As K_n^r is the unique minimizer of J_1 on $W^+[r]$, it follows that $\lim_{\epsilon\to 0} \|K_n^{r,\epsilon} - K_n^r\| = 0$, which is (i).

To prove (ii), we begin by fixing $\delta > 0$ and letting $I(\delta) := \{x \in [-r, r] : K_n^r(x) \geq \delta\}$. As convergence in Sobolev norm entails convergence in sup norm, there exists $\bar{\epsilon} > 0$ such that $K_n^{r,\epsilon}(x) \geq \delta/2 \quad \forall\ \epsilon \leq \bar{\epsilon}$. In light of conditions (a) and (d) of Lemma 3.1, we can therefore write

$$\Psi^{r,\epsilon}(x) + D^1\Psi^{r,\epsilon}(x) = \lambda^{r,\epsilon} \qquad \forall\ x \in I(\delta)\,, \quad \forall\ \epsilon \leq \bar{\epsilon}\,. \tag{3.2}$$

Because $K_n^{r,\epsilon} \to K_n^r$ as $\epsilon \to 0$, the left hand side of (3.2) must converge to $\Psi^r(x) + D^1\Psi^r(x)$. But the right hand side of (3.2) can only converge if there exists λ^r such that $\lambda^{r,\epsilon} \to \lambda^r$ as $\epsilon \to 0$. Moreover, since $\delta > 0$ is arbitrary, we obtain

$$\Psi^r(x) + D^1\Psi^r(x) = \lambda^r \qquad \forall\ x : K_n^r(x) > 0\,.$$

Now let

$$u^r(x) := \lambda^r - \Psi^r(x) + D^1\Psi^r(x)\,,$$

so that conditions (ii)(a) and (ii)(d) hold. By construction, $u^{r,\epsilon} \to u^r$, so that condition (ii)(e) must hold as well. Conditions (ii)(b) and (ii)(c) merely restate the feasibility of K_n^r. Thus, K_n^r must satisfy (ii)(a)–(ii)(e), i.e. the conditions are necessary. Finally, Tapia and Trosset (1991) give a general argument that the first order necessity conditions are also sufficient in the case of a convex program. This proves (ii). □

If it is desired *a priori* to use a compactly supported kernel, then Lemma 3.2 may be of some independent interest. We will simply employ it to approximate problem (CWL) with problems (CWL[r]) by letting $r \to \infty$.

Theorem 3.1 *Let $K_n^* \in W$ denote the unique solution of problem* (CWL) *and let $\Psi^*(x) := [\Psi K_n^*](x)$. Then*

(i) $\lim_{r \to 0} \|K_n^r - K_n^*\|_W = 0.$

(ii) *K_n^* is the unique element of W for which there exists a function u^* and a real number λ^* such that*

(a) $\Psi^*(x) + D\Psi^*(x) = \lambda^* - u^*(x) \qquad \forall\ x \in (-\infty, +\infty)\,,$

(b) $K_n^*(x) \geq 0 \qquad \forall\ x \in (-\infty, +\infty)\,,$

(c) $\int_{-\infty}^{\infty} K_n^*(x)dx = 1\,,$

(d) $K_n^*(x)u^*(x) = 0 \qquad \forall\ x \in (-\infty, +\infty)\,,$

(e) $u^*(x) \geq 0 \qquad \forall\ x \in (-\infty, +\infty)\,.$

Proof. Let W^+ denote the feasible set for problem (CWL) and let $W^+[r]$ denote the feasible sets for problems (CWL[r]) respectively. After verifying that $\bigcup_{r>0} W^+[r]$ is dense in W^+, the proof of Theorem 3.1 is exactly analogous to the proof of Lemma 3.2. □

The remainder of this paper will explore consequences of Theorem 3.1. We note, however, that the above results can easily be generalized. For example,

suppose that it is desired that the criterion for measuring the performance of the estimator $\hat{f}_n$ place less (or more) emphasis on the smoothness of the estimate. Then we can replace (3.1) with a weighted inner product

$$\langle u, v \rangle = \int_{-\infty}^{\infty} u(x)v(x)dx + \gamma \int_{-\infty}^{\infty} D^1 u(x) D^1 v(x) dx \, ,$$

where $\gamma > 0$, and our results remain unchanged except that condition (ii)(a) becomes

$$\Psi^*(x) + \gamma D^1 \Psi^*(x) = \lambda^* - u^*(x) \qquad \forall \; x \in (-\infty, +\infty) \, .$$

See Kufner (1985) for an introduction to weighted Sobolev spaces.

4 Properties of Optimal Kernels

In this section we assume that the probability density f is symmetric about some number $\theta \in (-\infty, +\infty)$. It then follows from the form of J_1 and the uniqueness of the global solution to problem (CWL) that K_n^* is symmetric about zero. In turn, this fact helps to establish a simplification of condition (ii)(a) in Theorem 3.1.

Lemma 4.1 *If f is symmetric about $\theta \in (-\infty, +\infty)$, then*

$$D^1 \Psi^*(x) = 0 \qquad \forall \; x : K_n^*(x) > 0 \, .$$

Proof. For any f, $f * \bar{f}$ is symmetric about zero. If f is symmetric about θ, then, by the above remark, K_n^* is also symmetric about zero. Then $\Psi^*(x) = \Psi^*(-x)$, and it follows that

$$D^1\Psi^*(x) = -D^1\Psi^*(-x) \tag{4.1}$$

and

$$D^2\Psi^*(x) = D^2\Psi^*(-x)\ . \tag{4.2}$$

Next, conditions (ii)(a) and (ii)(d) in Theorem 3.1 combine to give

$$\Psi^*(x) + D^1\Psi^*(x) \equiv \lambda^* \qquad \forall\ x : K_n^*(x) > 0\ . \tag{4.3}$$

Differentiating (4.3) gives

$$D^1\Psi^*(x) = -D^2\Psi^*(x) \qquad \forall\ x : K_n^*(x) > 0\ . \tag{4.4}$$

We now apply (4.1), (4.4), (4.2), and (4.4) again to obtain

$$D^1\Psi^*(x) = -D^1\Psi^*(-x) = D^2\Psi^*(-x) = D^2\Psi^*(x) = -D^1\Psi^*(x)$$
$$\forall\ x : K_n^*(x) > 0\ .$$

□

Lemma 4.1 allows us to establish the main result of this section.

Theorem 4.1 *Suppose that the probability density function f is symmetric about $\theta \in (-\infty, +\infty)$ and satisfies $f'(x)/f(x) \to \infty$ as $x \to -\infty$. Then the*

optimal (in the sense of problem (CWL)*) kernel* K_n^* *for the kernel density estimator* $\hat{f}_n$ *has compact support.*

Proof. For simplicity of notation, we write K for K_n^*. Suppose that K is not compactly supported. Then there exists a sequence $x_j \to -\infty$ as $j \to \infty$ for which $K(x_j) > 0$ for all $j = 1, 2, \ldots$. By Lemma 4.1, $\Psi^*(x_j) \equiv \lambda^*$ for all j, and since $\Psi^*(x) \to 0$ as $x \to -\infty$, it follows that $\lambda^* = 0$. Hence, it must be the case that

$$\frac{1}{n}K(x_j) = g(x_j) - \frac{n-1}{n}[K * g](x_j) \qquad j = 1, 2, 3, \ldots, \tag{4.5}$$

where $g = f * \bar{f}$.

We now consider the function $(n-1)K * g(x)/ng(x)$. Making use of L'Hopital's Rule, it is easily computed that

$$\begin{aligned}
\lim_{x\to\infty} \frac{n-1}{n} \frac{K * g(x)}{g(x)} &= \frac{n-1}{n} \lim_{x\to-\infty} \int_{-\infty}^{x} K * g(y)dy \Big/ \int_{-\infty}^{x} g(y)dy \\
&= \frac{n-1}{n} \lim_{x\to-\infty} \int_{-\infty}^{\infty} [G(x-z)/G(x)]K(z)dz \\
&= \frac{n-1}{n} \lim_{x\to-\infty} \int_{0}^{\infty} [(G(x+z) + G(x-z))/G(x)]K(z)dz \, .
\end{aligned} \tag{4.6}$$

We claim that the last expression in (4.6) is infinite.

Choose any interval $[a, b] \subseteq [0, +\infty)$ on which K is strictly positive, and let $\gamma := \inf\{K(x) : a < x < b\}$. Then

$$\int_0^\infty [(G(x+z) + G(x-z))/G(x)]K(z)dz \leq \gamma(b-a)G(x+b)/G(x)\ , \quad (4.7)$$

and, for $x < -b$,

$$\begin{aligned} G(x+b)/G(x) &= \exp\left[\log G(x+b) - \log G(x)\right] \\ &= \exp\left[\frac{x}{x+b}\log G(x+b) - \log G(x)\right]\exp\left[\frac{b}{x+b}\log G(x+b)\right] \\ &\geq \exp\left[\log G(x+b) - \log G(x)\right]\exp\left[\frac{b}{x+b}\log G(x+b)\right] \\ &\geq \exp\left[\frac{b}{x+b}\log G(x+b)\right]\ . \end{aligned} \quad (4.8)$$

Evidently the last expression in (4.8) tends to infinity as $x \to -\infty$ if and only if $\log(G(x))/x \to +\infty$ as $x \to -\infty$. By L'Hopital's Rule, this is equivalent to $g'(x)/g(x) \to -\infty$ as $x \to -\infty$. But $g(x) = f * \bar{f}(x) = f(\theta + x/2)$, so this last condition is true by hypothesis. Thus, the right hand side of (4.7) tends to infinity as $x \to -\infty$, and therefore (4.6) is infinite, as claimed.

It follows that there exists j_0 such that

$$g(x_j) - \frac{n-1}{n}[K * g](x_j) < 0 \qquad \forall\ j \geq j_0\ .$$

However, it follows from (4.5) and the choice of the sequence $\{x_j\}$ that

$$g(x_j) - \frac{n-1}{n}[K * g](x_j) > 0 \qquad j = 1, 2, 3, \ldots$$

This is a contradiction, and we conclude that $K = K_n^*$ has compact support.

□

Theorem 4.1 is reminiscent of a result due to Epanechnikov (1969), who minimized an objective function called the asymptotic relative global error. Restrict attention to kernels of the form $K_n(y) = K(y/h_n)/h_n$ and impose the additional constraint that

$$\int_{-\infty}^{\infty} y^2 K(y) dy = 1 \, .$$

Then the optimal K, which does not depend on f, is supported on the compact interval $[-5^{\frac{1}{2}}, +5^{\frac{1}{2}}]$. It should be emphasized that, unlike Theorem 4.1, this is an asymptotic result.

Theorem 4.1 addresses a substantive issue in the choice of kernel shape. It is generally considered desirable that kernels be reasonably smooth. For this reason, the Normal kernels $N(x) = (2\pi)^{-\frac{1}{2}} \exp[-(x/h)^2/2]/h$ are quite popular. However, the unbounded support of these kernels complicates computation of the corresponding estimates. Thus, Scott (1976) has argued that the quartic kernels $Q(x) = 15[1 - (x/h)^2]^2/16h$ are nearly as smooth as

Normal kernels, but enjoy a computational advantage due to their compact supports.

What are the statistical ramifications of compactly supported kernels? It is obvious that, if the density f itself has compact support, then so must the optimal kernel K_n^*. Theorem 4.1 establishes that this is also the case for all densities with sufficiently light tails.

5 Discussion

The purpose of this paper was to re-examine the approach of Watson and Leadbetter to optimizing kernel shape. Of interest was what could be stated about kernels that are required to be probability density functions. Because imposing these constraints effectively forces one to abandon the Fourier domain, it is difficult to state much. Theorem 3.1 characterizes the optimal kernels, and these conditions may be of use to other researchers. Theorem 4.1 is a substantive result about the desirability of kernels with compact support. Although it must be conceded that the hypotheses of this theorem are rarely satisfied in the practice of nonparametric density estimation (if a density is symmetric, then it likely belongs to some common parametric family), the result reinforces other arguments for using such kernels.

Acknowledgements

James R. Thompson stimulated my interest in the kernel shape problem, which in turn motivated my research with Richard A. Tapia on a Lagrange multiplier theory for infinite programming problems, on which the present results are based. David W. Scott provided encouragement and many helpful suggestions. I was supported in part by NSF Coop. Agr. No. CCR-8809615 as a visiting member of the Center for Research on Parallel Computation, Rice University, Houston, Texas 77251-1892, July 1991.

References

[1] Adams, R.A. (1975). *Sobolev Spaces.* Academic Press, New York.

[2] Epanechnikov, V.A. (1969). Nonparametric estimates of a multivariate probability density. *Theory of Probability and its Applications*, 14:153-158.

[3] Kufner, A. (1985). *Weighted Sobolev Spaces.* Wiley, New York.

[4] Parzen, E. (1958). On asymptotically efficient consistent estimates of the spectral density of a stationary time series. *Journal of the Royal*

Statistical Society, Series B, 20:303-322.

[5] Parzen, E. (1962). On estimation of a probability density function and mode. *Annals of Mathematical Statistics*, 33:1065-1076.

[6] Prakasa Rao, B.L.S. (1983). *Nonparametric Functional Estimation.* Academic Press, New York.

[7] Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. *Annals of Mathematical Statistics*, 27:832-837.

[8] Scott, D.W. (1976). Nonparametric probability density estimation by optimization theoretic techniques. Doctoral dissertation at Rice University, Houston, Texas.

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

[10] Tapia, R.A., and J.R. Thompson (1978). *Nonparametric Probability Density Estimation.* Johns Hopkins University Press, Baltimore.

[11] Tapia, R.A., and M.W. Trosset (1991). Extending the Farkas lemma approach to necessity conditions to infinite programming. Technical Report

91-25, Department of Mathematical Sciences, Rice University, Houston, Texas. Submitted for publication.

[12] Thompson, J.R., and R.A. Tapia. *Nonparametric Function Estimation, Modelling, and Simulation.* SIAM, Philadelphia.

[13] Watson, G.S., and M.R. Leadbetter (1963). On the estimation of the probability density, I. *Annals of Mathematical Statistics*, 34:480-491.

[14] Wertz, W. (1978). *Statistical Density Estimation: A Survey.* Vandenhoeck and Ruprecht, Göttingen.

[15] Whittle, P. (1958). On the smoothing of probability density functions. *Journal of the Royal Statistical Society, Series B*, 20:334-343.