Location-Adaptive Density Estimation and Nearest-Neighbor Distance

P. Burman and D. Nolan*
University of California, Davis
and
University of California, Berkeley

Technical Report No. 220
October 1989

*Research partly supported by NSF Grant No. DMS-8701426

Department of Statistics
University of California
Berkeley, California

Location-Adaptive Density Estimation and Nearest-Neighbor Distance

P. Burman and D. Nolan*

University of California, Davis and University of California, Berkeley

Abstract: A location-adaptive hybrid of the fixed-bandwidth kernel density estimate and the nearest-neighbor density estimate is introduced in this paper. It is constructed via a simple adhoc truncation and smoothing of nearest-neighbor distance. Simulations show that the hybrid outperforms its parent estimators, according to quadratic loss. Empirical process techniques are employed to obtain rates of uniform convergence of the random location-adaptive bandwidth to a deterministic function, from which uniform consistency of the hybrid, rates of convergence of the ISE, and asymptotic optimality of the ISE for the cross validatory choice of the smoothing parameter are obtained.

keywords: density estimation, nearest-neighbor distance, location-adaptive bandwidth, integrated square error, cross-validation, empirical processes, Vapnik-Červonenkis class.

*Research partly supported by NSF Grant no. DMS-8701426

1. Introduction

The advantages of location adaptive density estimation, where the bandwidth depends on the local behavior of the density, are often outweighed by additional computational requirements and unmanageable asymptotic theory. The nearest-neighbor distance provides an attractive adaptive density estimator because of its natural determination of high and low density regions that does not rely on estimates of derivatives of the density. An unfortunate drawback, however, is the fast growth of the nearest neighbor distance in the tails of the density. The roughness of the estimate can also be a drawback.

In this paper, we propose two simple adhoc techniques to rid nearest neighbor estimation of these problems. The first technique truncates the nearest neighbor distance at an arbitrary level; the second smooths the nearest neighbor distance via an arbitrary smoothing procedure. These simple changes greatly improve the nearest-neighbor density estimate. We measure improvement according to integrated square error and improved visual appearance.

This first change also facilitates asymptotic theory for nearest-neighbor estimators. As reported by Devroye and Györfi (1985), the asymptotic properties of nearest-neighbor based estimators have 'eluded most researchers.' Here a fresh approach uses rates of uniform convergence for annuli to get good error bounds on the approximation of nearest-neighbor distance by a deterministic function. This approximation enables us to obtain asymptotic optimality of a cross-validatory choice of the number of nearest neighbors used to create the estimate, as well as rates of convergence.

The nearest-neighbor density estimate (Loftsgaarden and Quesenberry 1965, Mack and Rosenblatt 1979) is defined as follows. Let $X_1, \ldots, X_n$ be a sample from some distribution P with density p. Let $r_k(x)$ denote the distance from the point x to its k^{th} nearest-neighbor among the observations. That is, if $B_k(x)$ stands for the sphere centered at x with radius $r_k(x)$ then $B_k(x)$ is the smallest closed sphere about x to contain at least k observations. The k^{th} nearest-neighbor estimate of p is:

$$\hat{p}_k(x) = \frac{1}{nV(B_k(x))} \sum_{i=1}^{n} \omega(\frac{x - X_i}{r_k(x)}),$$

where V stands for volume and ω is a symmetric density function. The hybrid of the nearest-neighbor and fixed-bandwidth estimators that is proposed here substitutes $\min(\bar{r}_k(x), tk/n)$ for $r_k(x)$ in the definition of $\hat{p}_k$. The bar over r_k denotes a smoothed version of the nearest-neighbor distance.

Breiman, Meisel, and Purcell (1977) were the first to suggest a simple change to the nearest-neighbor estimate for the purpose of reducing the bias in low density regions. Their estimator uses k^{th}-nearest-neighbor spheres centered at the observations.

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{V(B_k(X_i))}\omega(\frac{x-X_i}{r_k(X_i)}).$$

Unfortunately, $r_k(X_i)$ can also be quite large for extreme values of the sample. Truncation and smoothing of r_k can improve this estimator as well.

Abramson (1982) employs a preliminary local quadratic fit to the density in constructing an adaptive bandwidth. The purpose of the estimator is to reduce bias. Abramson also truncates the adaptive bandwidth if it becomes too large, but the truncation is employed more as a matter of convenience. Hall and Marron (1989) show that mean integrated square error (MISE) of Abramson's estimator achieves a very fast rate of convergence when restricted to a compact interval. The rate is typical of estimates of densities with 4 rather than 2 derivatives, where the estimate employs a kernel which is not a probability density function.

Abramson (1984) also proposes a location-adaptive estimator based on nearest-neighbor distance. It uses a two pass method. The data is split in two parts; the first part provides a preliminary estimate of the density; the second part estimates the nearest-neighbor distance. These two functions are then combined to produce a location-adaptive bandwidth.

Apart from uniform consistency (Devroye and Wagner (1977), Devroye and Penrod (1982), Nolan and Marron (1989)), few theoretical results have been proved for the nearest-neighbor estimator or for other estimators based on nearest-neighbor distance. In particular, Mack and Bhattacharya (1987) find the limit distribution of the nearest-neighbor density estimate at a point. They treat the estimator $\hat{p}_k(x)$ as a stochastic process indexed by the bandwidth. Muller and Stadtmuller (1987) explore the adaptive bandwidth for nonparametric regression in the fixed design case. They obtain rates of convergence for the MSE and MISE as well as pointwise convergence for the regression function. Mack and Muller (1987) estimate the mean function in nonparametric regression, and obtain pointwise results similar to those of Muller and Stadtmuller.

The asymptotic results presented here include asymptotic optimality of the cross-validated nearest-neighbor density estimate as well as rates of convergence for ISE and uniform consistency. Empirical process methodology is used to obtain these results.

Empirical process techniques were first employed by Devroye and Wagner (1977) to obtain the uniform consistency of the nearest-neighbor estimate. Their argument is based on the treatment of $\{B_k(x)\}$ as a Vapnik-Červonenkis class of sets, which leads to the uniform convergence of k/n, the empirical probability content of $B_k(x)$, to its expected probability content $\int_{B_k(x)} p(y)dy$. From there, it can be shown that $r_k(x)$ is uniformly close to the deterministic function $k/(2np(x))$ for x in high density regions. The error in the approximation is $o(k/n)$. We continue in this same vein, employing rates of uniform convergence for another Vapnik-Červonenkis collection of sets, the annuli formed by the symmetric difference $B_k(x)\Delta S_k(x)$ where $S_k(x)$ is the ball centered at x such that $\int_{S_k(x)} p(y)dy = k/n$. This rate of convergence yields a much smaller, more manageable error when $r_k(x)$ is approximated by the radius of $S_k(x)$, rather than $k/(2np(x))$. If P_n represents the empirical distribution based on the sample, the error can be expressed as:

$$\int_{S_k(x)} d(P_n - P)(y) + O(\log n(\frac{k^{3/2}}{n^2} + \frac{k^{1/4}}{n})).$$

We also employ rates of convergence for the empirical process of U-statistic structure (Nolan and Pollard, 1987) to obtain the above mentioned asymptotics.

The rest of this paper is organized as follows. First a heuristic justification for truncating $r_k(x)$ is given in Section 2. Then simulations compare the MISE of our hybrid estimator with the optimal kernel and nearest-neighbor estimators (Section 3). Theoretical results are presented in Section 4 with proofs in Section 5 and the Appendix.

2. Truncation

For the heuristics only, we assume the density p has two derivatives. Also assume the dimension is one, and the kernel function ω is a symmetric density function with mean 0 and finite variance v. Write $h(x)$ for the truncated bandwidth $\min(r_k(x), tk/n)$ and write $\hat{p}_h$ for the density estimate constructed from h. The mean integrated square error can then be approximated as follows:

$$\begin{aligned} MISE(k) &= \mathbf{E}\int(\hat{p}_h(x)-p(x))^2dx \\ &\approx (\int\omega^2/n)\int p(x)h(x)^{-1}dx+v^2/4\int h(x)^4p''(x)^2\,dx. \end{aligned}$$

The random bandwidth $\min(r_k(x), tk/n)$ is approximately: $k/n\ \min(\frac{1}{2}p(x)^{-1}, t)$, because

$$\int_{B_k(x)} p(x)dx \approx p(x)2r_k(x)$$

and

$$\int_{B_k(x)} p(x)dx \approx \frac{1}{n}\sum_{i=1}^{n}\{X_i\varepsilon B_k(x)\} = \frac{k}{n}.$$

Break up the range of integration according to the region $\{x : p(x) \geq (2t)^{-1}\}$ and its complement. Call these two regions H and L, respectively. Then

$$\begin{aligned} MISE(k) &\approx \frac{1}{kt}\int\omega^2\Big[1+\int_H(2tp(x)-1)p(x)dx\Big]+\frac{1}{4}v^2(\frac{tk}{n})^4\int p''(x)^2dx \\ &\Big[1+(\int p''(y)^2dy)^{-1}\int_H\{p''(x)^2(2tp(x))^{-4}-p''(x)^2\}dx\Big] \\ &= \frac{1}{kt}\int\omega^2[1+C_1]+\frac{1}{4}v^2(\frac{tk}{n})^4\int p''(x)^2dx\,[1-C_2] \end{aligned}$$

The constants C_1 and C_2 depend on the density and the truncation level. Take the k that minimizes the approximation above and plug it back into the MISE to find

$$MISE(k) \approx n^{-4/5}(\int\omega^2)^{4/5}v^{2/5}(\int p''(x)^2dx)^{1/5}\frac{5}{4}\,[1+C_1]^{4/5}\,[1-C_2]^{1/5}$$

If the truncation level is such that $[1+C_1]^4\,[1-C_2] \leq 1$ then this estimator can improve upon the fixed bandwidth kernel estimator, in addition to greatly improving the nearest-neighbor estimator.

3. Simulations

In general, the authors found that when the truncation level is chosen in an adhoc manner, both the ISE and the visual appearance are improved. The adaptive estimator is at least as good as the optimal kernel estimator, and often times better.

Figure 1 shows the effect of smoothing and truncation on the nearest neighbor distance. The data are lengths in days of psychiatric treatment of 86 patients in a suicide study (reported by Copas and Fryer (1980), as appeared in Silverman (1986)). Figure 2a displays the resulting density estimate, Figure 2b shows a fixed bandwidth estimate, and 2c a nearest-neighbor estimate, each uses the Epanechnikov kernel. The kernel estimate over smooths the high density area, combining the two peaks; whereas, the nearest-neighbor estimate is a very rough curve with a large right tail. The smoothed truncated nearest-neighbor estimate appears a good balance between the two.

From the previous section, the best truncation level depends on both the excess probability mass in the high density region: $\int_H (2tp - 1)$ and a function of the second derivative: $\int_H p''^2[(2tp)^{-4} - 1]$. To estimate these quantities would defeat the purpose of the hybrid estimator. Instead, as a rule of thumb, truncation levels near the standard deviation have in practice produced good estimates. In the multimodal case, the radius of the largest mode, as measured by the distance between quantiles or the radius of the smallest k^{th} nearest neighbor ball, works well.

Simulations show that truncation is very effective over a wide range of values. Figures 3a-h compare the ISE of the optimal fixed bandwidth estimate against the hybrid, over a range of truncation levels. The comparison with the nearest-neighbor estimator is even more striking, for in our examples its ISE is much larger than that of the kernel estimator. Extremely small truncation levels result in far-from-optimal fixed bandwidth estimates; alternatively, extremely large truncation levels produce nearest-neighbor estimates. It is evident from the figures that these extreme cases are easily avoided. In the case of the double exponential distribution and the mixture of double exponentials, the improvement on ISE is very satisfactory. The authors note, however, that the ISE, a popular measure of a density estimator's performance, is not necessarily a good measure. Even though truncation brings the nearest-neighbor estimator within the range of the kernel estimator in terms of ISE, it does not satisfactorily emphasize the good job it is doing in estimating the high density regions, as evident from figures 2a,b,c.

All simulations are based on 100 repeats. For each repeat, a sample of 100 observations is drawn and the ISE is minimized for the hybrid estimator at a fixed truncation level and for the kernel estimator. Figures 3a-h show the median, quartiles, and extremes of the 100 differences: ISE(best hybrid estimator) - ISE(best kernel estimator). The distributions included in the simulation are the N(0,1), the mixture .5 N(-1,1) + .5 N(2,1), the DE(0,1), and the mixture .3 DE(-1.5,.5) + .7 DE(1.5,.5), where DE(0,1) indicates the double exponential distribution with mean 0 and variance 1.

4. Asymptotics

In this section, we present asymptotic theory for the location adaptive density estimator. In particular, uniform consistency, rates of convergence of integrated square error, and the cross-validatory choice of k are treated. These results depend on the uniform approximation of $r_k(x)$ by a deterministic function where the density is bounded below by a positive constant. Truncation allows a uniform approximation to hold over the entire range of x, and this approximation continues to hold when r_k is replaced by its smoothed version.

All results presented are for univariate densities; the analogous multidimensional results hold as well. Define the following location-adaptive truncated bandwidths:

$$h_k(x) = \min(r_k(x), \frac{k}{n}t),$$

Also define the smoothed counterpart to h_k:

$$\bar{h}_k(x) = \min(\bar{r}_k(x), \frac{k}{n}t),$$

where $\bar{r}_k$, a smoothed version of r_k, may be smoothed according to a nonparametric kernel regression technique. That is, for kernel function ν, scale parameter σ, and grid $y_1, \ldots, y_N$:

$$\bar{r}_k(x) = \frac{\sum_{j=1}^{N} r_k(y_j)\nu(\frac{x-y_j}{\sigma})}{\sum_{i=1}^{N} \nu(\frac{x-y_i}{\sigma})}.$$

The density estimate constructed from $h_k(x)$ is defined as follows:

$$\hat{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n} h_k^{-1}(x)\omega(\frac{x-X_i}{h_k(x)}),$$

where ω is a symmetric density function. The estimate $\hat{p}_{\bar{h}}$ is defined similarly. Also, define the truncated version of Breiman, Meisel and Purcell's estimator, called BMP from now on, as follows:

$$\tilde{p}_h(x) = \frac{1}{n}\sum_{i=1}^{n} h_k^{-1}(X_i)\omega(\frac{x - X_i}{h_k(X_i)}),$$

and the 'expected value' of $\hat{p}_h$:

$$\bar{p}_h(x) = \int h_k^{-1}(x)\omega(\frac{x - y}{h_k(x)})p(y)dy,$$

The estimates $\bar{p}_{\bar{h}}$, $\bar{\tilde{p}}_h$ are defined accordingly.

The results presented below are for one dimension, but their multi- dimensional analogs are similarly obtained. For the results that follow, assume ω is a density function that is symmetric, twice differentiable, bounded by 1, decreasing in $|x|$ and with support on $[-1, 1]$. Also assume ω and its two derivatives are of bounded variation. We place the constraint on k that $n^{\delta} \leq k \leq n^{1-\delta}$, for some arbitrary, small, positive δ. As for notation, P_n stands for the empirical distribution based on a sample of size n from the distribution P; the indicator function for a set is denoted by the set itself; and linear functional notation is used. So, the probability content, with respect to the distribution P, of the ball centered at x with radius r is expressed as

$$PB(x, r) = \int_{B(x,r)} dP(y).$$

When there may be confusion over the argument of integration, write P^y to indicate that y is the argument of integration; for example, $P^y\{y \in B(x, r)\} = PB(x, r)$.

Two approximations to h_k are needed to obtain the theoretical results for $\hat{p}_h$ and $\tilde{p}_h$. The first approximation originates from the Taylor series expansion:

$$PB(x, r_k(x)) \approx 2r_k(x)p(x).$$

This approximation provides uniform consistency and rate results for the integrated square error. However a second, finer approximation is required of the more difficult asymptotic optimality. There, we make use of the function $s_k(\cdot)$ that is defined such that

$$PB(x, s_k(x)) = \frac{k}{n}.$$

That is, $s_k(x)$ is the radius of the ball centered at x with P-measure k/n. Here, r_k is approximated by s_k as follows:

$$\begin{aligned} 2p(x)(s_k(x) - r_k(x)) &\approx PB(x, s_k(x)) - PB(x, r_k(x)) \\ &= (P_n - P)B(x, r_k(x)) \\ &\approx (P - P_n)B(x, s_k(x)). \end{aligned}$$

The key to the approximation is that the error $(P - P_n)B(x, s_k(x))$ is a linear functional of a nonrandom indicator function $B(x, s_k(x))$. The details of this approximation appear in the proof of the following lemma.

1 LEMMA: *Let $\mathcal{H} = \{x : p(x) > (12t)^{-1}\}$ and let p be bounded and uniformly continuous. Then*

$$(a) \quad \begin{aligned} &\sup\nolimits_{\mathcal{H}} |h_k(x) - \tfrac{k}{n}\min((2p(x))^{-1}, t)| = o(\tfrac{k}{n}) \quad \textit{eventually}, \quad \textit{a.s.} \\ &\sup\nolimits_{\mathcal{H}^c} |h_k(x) - \tfrac{k}{n}t| = 0 \quad \textit{eventually}, \quad \textit{a.s.} \end{aligned}$$

If, in addition, p has a continuous first derivative then
(b)

$$\sup_{\mathcal{H}} |2p(x)(s_k(x) - r_k(x)) - (P_n - P)B(x, s_k(x))| = O(\log n(\frac{k^{1/2}}{n^{1+\delta}} + \frac{k^{1/4}}{n})) \ \textit{a.s.}\Box$$

The error terms in Lemma 1 are uniform in k, for $n^\delta \leq k \leq n^{1-\delta}$. From Lemma 1.a follows the uniform continuity of the density estimates.

2 LEMMA: *If p is bounded and uniformly continuous then*

$$\begin{aligned} \sup_x |\hat{p}_h - p| &\rightarrow 0 \text{ a.s.} \\ \sup_x |\tilde{p}_h - p| &\rightarrow 0 \text{ a.s.} \\ \sup_x |\hat{p}_{\bar{h}} - p| &\rightarrow 0 \text{ a.s. } \Box \end{aligned}$$

The result of Lemma 1.a can also be employed to bound the integrated square error:

$$ISE(\hat{p}_h) = \int (\hat{p}_h - p)^2 = O_p(\frac{1}{k} + (\frac{k}{n})^4).$$

The k^{-1} term bounds the 'variance': $\int(\hat{p}_h - \bar{p}_h)^2$, while the second term bounds the 'bias squared': $\int(\bar{p}_h - p)^2$. Lemma 3 below states these results. They also hold for the smoothed estimator $\hat{p}_{\bar{h}}$.

3 LEMMA: *If p is bounded, and Lipschitz then*

(*a*)
$$\int(\hat{p}_h - \bar{p}_h)^2 = O_p(\frac{1}{k}).$$

$$\int(\tilde{p}_h - \bar{\bar{p}}_h)^2 = O_p(\frac{1}{k}).$$

If, in addition, p is twice differentiable then

(*b*)
$$\int(\bar{p}_h - p)^2 = O_p((\frac{k}{n})^4)$$

$$\int(\bar{\bar{p}}_h - p)^2 = O_p((\frac{k}{n})^4 + \frac{1}{k}) \qquad \square$$

The following lemma gives conditions under which the cross-validatory choice of k for $\tilde{p}_k$ the BMP estimator is asymptotically optimal. It requires the finer approximation of r_k found in Lemma 1.b. The two main conditions are that p has a continuous first derivative and that the set of x such that $p(x) = (2t)^{-1}$ has Lebesgue measure 0. The differentiability condition can be weakened to a Lipschitz condition on p without invalidating the conclusion of the lemma. For ease of proof the authors use the stricter requirement. The second condition avoids flat spots in the density that occur exactly at the truncation level. It could be removed if a smooth function were to replace the truncation operation. That is,

$$h_k(x) = r_k(x)\varphi(\frac{kt}{nr_k(x)})$$

where

$$\varphi(y) = \begin{array}{ll} y & 0 < y < 1 \\ 1 & y \geq 1. \end{array}$$

The function φ could be replaced by a smoother, differentiable function to avoid the second requirement. Again, for simplicity of argument this more general case is not considered here.

4 LEMMA: *Suppose p is bounded and continuous with a continuous first derivative. Also suppose $\{x : p(x) = (2t)^{-1}\}$ has Lebesgue measure 0. If k_{CV} is chosen to minimize:*

$$CVISE(k) = \int \tilde{p}_h^2 - \frac{2}{n(n-1)} \sum_{i \neq j} \frac{1}{h_k(X_i)} \omega(\frac{X_i - X_j}{h_k(X_i)})$$

then

$$\frac{ISE(k_{CV})}{ISE(k_{ISE})} \longrightarrow 1 \textit{ in probability,}$$

where k_{ISE} minimizes $ISE(k) = \int(\tilde{p}_h - p)^2$. □

A similar result holds for our estimators $\hat{p}_h$ and $\hat{p}_{\bar{h}}$ provided either p has a second derivative, or a positive constant γ can be found such that

$$M_1(\frac{k}{n})^\gamma \leq \int (\bar{p}_h - p)^2 \leq M_2(\frac{k}{n})^\gamma$$

for $M_1, M_2 > 0$. We do not present its proof here.

The proofs of these four lemmas appear in the next section.

5. Proofs

The proofs of all four lemmas employ rates of convergence for empirical processes. We find the following adaptation of Theorem 2.4 of Pollard (1986) most useful.

LEMMA 5: *Let P be a probability distribution on $\mathbb{R}^d$. Let $\{\mathcal{F}_n\}$ be a sequence of permissible collections of uniformly bounded, real-valued functions on $\mathbb{R}^d$ such that given $\epsilon > 0$ there exists a subclass $\mathcal{F}_n^*$ of $\mathcal{F}_n$ with*

$$\textit{cardinality } (\mathcal{F}_n^*) \leq A\epsilon^{-V}$$

and for each $f \in \mathcal{F}_n$ there exists $f^* \in \mathcal{F}_n^*$ such that $Q|f - f^*| < \epsilon$, for all probability measures Q on $\mathbb{R}^d$. (Note the constants A, V and the uniform bound on the functions do not depend on n.) Then, for $0 < \alpha \leq \frac{1}{2}$,

$$\sup_{n^\delta \leq k \leq n^{1-\delta}, \mathcal{F}_n} \frac{k^\alpha |P_n f - Pf|}{P_n|f| + P|f| + \frac{k^{2\alpha}}{n}} = O(\sqrt{\log n}) \;\; a.s. \square$$

The term 'permissible' refers to measurability conditions. See Pollard (Appendix C, 1984) for a discussion of the concept. This lemma will be applied to a variety of collections $\mathcal{F}_n$. The condition on the metric entropy of $\mathcal{F}_n$ is called the Euclidean property (Nolan and Pollard 1987). Section 5 of Pollard (1989) presents many examples and ways to confirm the Euclidean property. In particular, the collection of indicator functions for sets has this property if the collection of sets themselves form a Vapnik-Červonenkis class of sets. The collection of balls in $\mathbb{R}^d$ is a Vapnik-Červenenkis class, and so its indicators meet the conditions of Lemma 5. So does the collection of indicators for annuli $\mathcal{A}$, because the annulus is formed by the intersection of one ball with the complement of another ball. Four other collections of functions will be used in the proofs of the lemmas, and therefore must meet the Euclidean property. In the definitions below take $g_k(x) = \min(s_k(x), \frac{k}{n}t)$:

(1) $\mathcal{W}_n = \{w_{k,x} : w_{k,x}(y) = w(\frac{x-y}{g_k(x)})\}$

(2)(a) $\mathcal{V}_n = \{v_{k,x} : v_{k,x}(y) = w'(\frac{x-y}{g_k(x)})(\frac{x-y}{g_k(x)})\}$

(b) $\mathcal{U}_n = \{u_{k,x} : u_{k,x}(y) = w''(\frac{x-y}{g_k(x)})(\frac{x-y}{g_k(x)})^2\}$

(3) $\Gamma_n = \{\gamma_k : \gamma_k(x,y) = P^z\{|y-z| \leq g_k(z)\}\frac{k}{n}g_k^{-2}(z)[w_{k,x}(z) + v_{k,x}(z)]\}$

For the present, we take it for granted that these collections of functions are Euclidean, and so meet the conditions required of Lemma 5. The Euclidean property is established in the Appendix along with the proof of Lemma 5.

PROOF OF LEMMA 1.a. Apply Lemma 5 for $\alpha = 1/2$ to the collection $\mathcal{B}$ of all balls.

$$\sup_{k,\mathcal{B}} \frac{\sqrt{k}|P_n(B) - P(B)|}{P_nB + PB + k/n} = O(\sqrt{\log n}) \quad a.s.$$

This implies that $PB(x, r_k(x)) \leq 2k/n$ for all x and k, eventually, almost surely. Use this upper bound to show

$$\sup_{x,k} k^{-1/2}|P_nB(x,r_k(x)) - PB(x,r_k(x))| = O(\sqrt{\log n}/n) \quad a.s. \tag{4}$$

Equation (4) is the basis of Lemma 1a. First it is used to establish

$$\sup_{\mathcal{H}^c} |h_k(x) - \frac{k}{n}t| = 0 \qquad eventually,\ a.s.$$

where $\mathcal{H} = \{x : p(x) > (12t)^{-1}\}$. On $\mathcal{H}^c$, the facts that

$$\min((2p(x))^{-1}, t) = t$$

and that p is uniformly continuous imply either

$$\sup_{y \in B(x, r_k(x))} p(y) < \frac{1}{4t}$$

or, for some $C_t > 0$,

$$r_k(x) \geq C_t \ .$$

The latter possibility implies $h_k(x) = tk/n$ for n large. The first possibility also implies $h_k(x) = tk/n$ almost surely, eventually, because

$$\frac{1}{2t} r_k(x) \geq PB(x, r_k(x)) > \frac{k}{2n} \quad \text{all } k \text{ and } x, \text{ eventually, a.s.}$$

The second inequality follows from 4.

Now turn to the region $\mathcal{H}$. We use (4) again, this time with a Taylor series expansion of $PB(x, r_k(x))$, to show that $r_k(x)$ is uniformly close to $k/n2p(x)$. Truncation at t only decreases the distance between $r_k(x)$ and its approximation. Consider x in $\mathcal{H}$.

$$\frac{1}{2p(x)}|(P_n - P)B(x, r_k(x))| = |\frac{k}{n2p(x)} - r_k(x) + \frac{r_k(x)}{2p(x)} \int_{\{|y| \leq 1\}} (p(x) - p(x + r_k(x)y))dy|.$$

Apply (4) again to get

$$\sup_{x \in \mathcal{H}} r_k(x) = O(\frac{k}{n}) \ \textit{a.s.},$$

which gives the desired rate of convergence. □

PROOF OF LEMMA 1.b. The difference of the indicators $B(x, r) - B(x, s)$ is the signed indicator for an annulus. Call it $A(x, r, s)$. Note $A(x, r, s) = -A(x, s, r)$. Then, because $\frac{k}{n} = PB(x, s_k(x)) = P_n B(x, r_k(x))$:

$$PA(x, s_k(x), r_k(x)) = (P_n - P)A(x, r_k(x), s_k(x)) + (P_n - P)B(x, s_k(x)).$$

Alternatively, use the differentiability condition on p and the fact that $p(x) \geq (12t)^{-1}$ on $\mathcal{H}$ to express the expected value on the left above as:

$$PA(x, s_k(x), r_k(x)) = (s_k(x) - r_k(x))(2p(x) + O(\frac{k}{n})) \quad \text{almost surely.}$$

Combine these two equalities.

$$\sup_{\mathcal{H}} |2p(x)(s_k(x) - r_k(x)) - (P_n - P)B(x, s_k(x))| \leq$$

$$\sup_{\mathcal{H}} |(P_n - P)A(x, r_k(x), s_k(x)| + O(\sqrt{k \log n}/n^{1+\delta}) \quad a.s.$$

The order term on the right hand side of the inequality follows from the bound

$$\sup_{x,k} k^{-1/2}|(P_n - P)B(x, s_k(x))| = O(\sqrt{\log n}/n) \quad a.s$$

To complete the proof apply Lemma 5 to the collection of annuli $\mathcal{A}$ for $\alpha = 1/4$.

$$\sup_{\mathcal{A}} \frac{k^{1/4}|P_n(A) - P(A)|}{P_n A + PA + \frac{\sqrt{k}}{n}} = O(\sqrt{\log n}) \quad a.s.$$

Now for $A(x, r_k(x), s_k(x))$ we have, uniformly in k:

$$P_n|A(x, r_k(x), s_k(x))| = |(P_n - P)B(x, s_k(x))| = O(\sqrt{k \log n}/n)$$

$$P|A(x, r_k(x), s_k(x))| = |(P_n - P)B(x, r_k(x))| = O(\sqrt{k \log n}/n)$$

Then

$$\sup_{x,k} k^{-1/4}|(P_n - P)A(x, r_k(x), s_k(x))| = O(\log n/n) \quad a.s.$$

This establishes part b of Lemma 1. □

PROOF OF LEMMA 2: The first of the three conclusions is proved here; the other two follow by similar arguments. Break the difference $\hat{p}_h - p$ into two parts: $\hat{p}_h - \bar{p}_h$ and $\bar{p}_h - p$. Treat the two parts separately. The typical change of variables yields

$$\bar{p}_h(x) - p(x) = \int w(z)[p(x + zh_k(x)) - p(x)]dz \,. \tag{5}$$

Lemma 1.a implies $\sup_x h_k(x) \rightarrow 0$ almost surely. This fact and the uniform continuity of p imply the difference inside the square brackets tends to 0, almost surely. As for the variance term, let $\eta_k(x) = \frac{k}{n}\min((2p(x))^{-1}, t)$. Then

$$\begin{aligned} |\hat{p}_h(x) - \bar{p}_h(x)| &= h_k(x)^{-1}|(P_n - P)^y w(\tfrac{x-y}{h_k(x)})| \\ &= \eta_k(x)^{-1}|(P_n - P)^y w(\tfrac{x-y}{\eta_k(x)})| \ + \ e_k(x) \end{aligned} \tag{6}$$

On $\mathcal{L}$ the error $e_k(x)$ committed in substituting $\eta_k(x)$ for $h_k(x)$ is almost surely 0; on $\mathcal{H}$ the error can be bounded by using a corollary of Lemma 1.a:

$$\sup_{\mathcal{H}} |\frac{\eta_k(x) - h_k(x)}{\eta_k(x)}| = o(1) \quad a.s.$$

The compact support of w and differentiability of w imply

$$\sup_{\mathcal{H}} e_k(x) = o(1) \quad a.s.$$

To complete the proof apply Lemma 5 to the class of functions

$$\{w_{k,x}(y) : \ w_{k,x}(y) = w\left(\frac{x-y}{\eta_k(x)}\right)\}$$

to show the first term in the right hand side of (6) is negligible. □

PROOF OF LEMMA 3: A refinement of (5) and (6) from the proof of Lemma 2 leads to the proof of this lemma. For the integrated squared bias, twice differentiability of p and symmetry of ω updates (5) for some constant $C > 0$ and $0 < t(x) < t$ to:

$$\int (\bar{p}_h(x) - p(x))^2 \leq C(\frac{k}{n})^4 \int_{\mathcal{H}} (\int_0^1 \frac{1}{2} z^2 \omega(z) p''(x + \frac{k}{n} z t(x)) dz)^2 dx \ .$$

For the variance, look more closely at the error $e_k(x)$ in (6). The additional requirement that p is Lipschitz and the upper bound: $k/n \leq n^{-\delta}$, mean that for some positive α,

$$\sup_{\mathcal{H}} |(\eta_k(x) - h_k(x))/\eta_k(x)| = o(n^{-\alpha}) \quad a.s.$$

Apply Lemma 5 to the collections $\mathcal{W}_n$ and $\mathcal{V}_n$ with g_k replaced by η_k. Then the following equalities hold almost surely.

$$\begin{aligned}
\sup_{\mathcal{H}} \sqrt{k} e_k(x) &= o(n^{-\alpha}) \sup_{\mathcal{H}} |\sqrt{k}(P_n - P)^y \eta_k(x)^{-1} w_{k,x}(y)| \\
&+ o(n^{-\alpha}) \sup_{\mathcal{H}} |\sqrt{k}(P_n - P)^y \eta_k^{-1}(x) v_{k,x}(y)| \\
&= o(n^{-\alpha}) \sup_{\mathcal{H}} |\frac{n}{\sqrt{k}}(P_n - P)^y w_{k,x}(y)| + o(n^{-\alpha}) |\frac{n}{\sqrt{k}}(P_n - P)^y v_{k,x}(y)| \\
&= o(n^{-\alpha}) O(\log n)
\end{aligned}$$

The bound on e_k implies

$$\int k e_k^2 = \int_{\mathcal{H}} k e_k^2 + \int_{\mathcal{L}} k e_k^2 = O_p(1) .$$

Then for $\int(\hat{p}_\eta - \bar{p}_\eta)^2$, an upper bound on its expected value is, for $p_0 = \sup_x p(x)$,

$$\begin{aligned}
&\frac{1}{n} \int \int \eta_k(x)^{-2} \omega^2 (\frac{x-y}{\eta_k(x)}) dx dP(y) \\
&\leq \frac{2 p_0 t}{n} \int (\frac{k}{n})^{-1} \omega^2(z) dz dP(y)
\end{aligned}$$

With Markov's inequality, the conclusion of Lemma 3 for $\hat{p}$ is established. We now state how this proof can be modified for $\tilde{p}$. For the bias, the typical change of variables (5) yields a ratio of $h(x)/h(x + zh(x))$, which can then be crudely bounded using Lemma 1. As for the variance term, replace r_k by s_k, rather than $k/n2p(x)$, in (6). Care must be taken with the error term. We refer the reader to the proof of Lemma 4, which uses very similar techniques. □

PROOF OF LEMMA 4: Define:

$$\begin{aligned}
L_n(h) &= \int (\tilde{p}_h - p)^2 \\
M_n(h) &= \int \tilde{p}_h^2 - \frac{2}{n(n-1)} \sum_{i \neq j} \frac{1}{h(X_i)} \omega(\frac{X_i - X_j}{h(X_i)}) + \int p^2 \\
\bar{L}_n(h) &= \frac{1}{h^2} \sum_{i=1}^{n} \int \frac{1}{h(X_i)^2} \omega^2 (\frac{x - X_i}{h(X_i)}) dx + \int (\bar{\tilde{p}}_h - p)^2 .
\end{aligned}$$

Similarly define $L_n(g)$, $M_n(g)$ and $\bar{L}_n(g)$, where $g_k(x) = \min(s_k(x), \frac{k}{n}t)$. Stone (1984), Burman (1985), Marron (1985), Nolan and Pollard (1987) and others all show that in the fixed bandwidth case, the scale parameter σ_M that minimizes $M_n(\cdot)$ does almost as well as the σ_L that minimizes $L_n(\cdot)$, in the sense that

$$L_n(\sigma_M)/L_n(\sigma_L) \to 1 \quad \text{in prob.}$$

Most proofs of the asymptotic optimality are obtained by comparing L_n and M_n with the expected value of L_n. Typically the result follows from:

$$\sup_k \frac{|L_n(h) - M_n(h) + Z_n|}{\bar{L}_n(h)} = o_p(1) \tag{7}$$

$$\sup_k \frac{|L_n(h) - \bar{L}_n(h)|}{\bar{L}_n(h)} = o_p(1), \tag{8}$$

where Z_n is a random variable that does not depend on h or k. Rather than establish (7) and (8) directly, we establish their counterparts:

$$\sup_k \frac{|L_n(g) - M_n(g) + Z_n|}{\bar{L}_n(g)} = o_p(1) \tag{9}$$

$$\sup_k \frac{|L_n(g) - \bar{L}_n(g)|}{\bar{L}_n(g)} = o_p(1). \tag{10}$$

Then (7) and (8) follow from:

$$\sup_k \frac{|L_n(h) - M_n(h) - L_n(g) + M_n(g)|}{\frac{1}{k} + \|\bar{p}_g - p\|^2} = o_p(1) \tag{11}$$

$$\sup_k \frac{|L_n(h) - \bar{L}_n(h) - L_n(g) + \bar{L}_n(g)|}{\frac{1}{k} + \|\bar{p}_g - p\|^2} = o_p(1) \tag{12}$$

$$\sup_k \frac{\frac{1}{k} + \|\bar{p}_g - p\|^2}{\frac{1}{k} + \|\bar{p}_h - p\|^2} = o_p(1). \tag{13}$$

To prove (9) and (10) the methods of proof in Nolan and Pollard (1987) carry over completely, because (1)is a Euclidean class of functions (see Appendix). The proofs of (12) and (13) closely follow that of (11); we present only the proof of (11) here.

First note that Lemma 1b implies

$$\text{(14)} \qquad \sup_{k,\mathcal{H}} \frac{|h_k - g_k|}{\sqrt{k}} = O(\sqrt{\log n/n}) \quad \text{eventually, a.s.}$$

$$\sup_{k,\mathcal{L}} |h_k - g_k| = 0 \quad \text{eventually, a.s.}$$

The proof uses repeated applications of Lemma 5 to the collections $\mathcal{W}_n$, $\mathcal{V}_n$, and $\mathcal{U}_n$ with $\alpha = 1/2$. For $\mathcal{W}_n$, Lemma 5 implies

$$\text{(15)} \qquad \sup_{x,k} \sqrt{k}\frac{|(P_n - P)^y w_{k,x}|}{g_k(x)} = O(\sqrt{\log n}) \quad a.s.$$

Similar rates apply to $\mathcal{V}_n$ and $\mathcal{U}_n$.

To prove (11) reexpress the numerator as

$$\text{(16)} \qquad 2(P - P_n)^x \otimes P_n^y \nu_k(x,y,y)$$

$$-\frac{2}{n^2(n-1)} \sum_{i \neq j} \nu_k(X_i, X_j, X_j)$$

$$-\frac{2}{n} P_n^x w(0)[h_k(x)^{-1} - g_k(x)^{-1}]$$

where $\nu_k(x,y,z) = h_k(z)^{-1} w(\frac{x-y}{h_k(z)}) - g_k(z)^{-1} w(\frac{x-y}{g_k(z)})$. Use (14) to show the third term is $o_p(k^{-3/2})$. It converges to 0 in probability when normalized by the denominator in (11). The assumption of derivatives for ω allow a Taylor series expansion of $\nu_k(x,y,y)$:

$$g_k(y)^{-2}[\omega_{k,x}(y) + v_{k,x}(y)]\,.$$

From (14) and the expectation of the Taylor series expansion, the second term in (16) is $o_p(n^{-1})$. Rewrite the first term in (16) as:

$$\text{(17)} \qquad (P_n - P)^x \otimes P_n^y [\frac{h_k(y) - g_k(y)}{g_k(y)} \frac{w_{k,x}(y) + v_{k.x}(y)}{g_k(y)}] + o_p(\frac{1}{k})$$

The $o_p(1/k)$ is uniform in k; it bounds the second term in the Taylor series expansion of w_k uniformly in y. To see this apply (14), (15) and the versions of (15) for $\mathcal{U}_n$ and $\mathcal{V}_n$ to

$$(\frac{h_k(y) - g_k(y)}{g_k(y)})^2|(P_n - P)^x \frac{2w_{k,y}(x) + 5v_{k,y}(x) + u_{k,y}(x)}{g_k(y)}|.$$

This leaves only the first term in (17).

To bound this term requires the full strength of Lemma 1.b. If $h_k(y) = g_k(y) = \frac{k}{n}t$ then the contribution from (17) is exactly 0. This is the case, almost surely, on

$$\mathcal{L}_n = \{y : p(y) \leq \frac{1}{2t} - \frac{1}{(\log n)^4}\}.$$

(Note that $\mathcal{L}_n$ does not approximate $\mathcal{L}$ but $\{x : p(x) \geq 1/2t\}$.) The equality above follows from the differentiability of p and an argument similar to the proof that $h_k(y) = \frac{k}{n}t$ on $\mathcal{L}$ in Lemma 1.a. Also, on the 'complementary' set $\mathcal{H}_n = \{x : p(x) \geq \frac{1}{2}t^{-1} + \log^{-4} n\}$ the functions h_k and g_k can be replaced by r_k and s_k, respectively. On the intermediate region $\mathcal{M}_n$, use (15) to bound (17), normalized by $\sqrt{k}$, by

$$\sqrt{k}\log n \; P_n^y\{y \in \mathcal{M}_n\}|(P_n - P)^x \frac{w_{k,x}(y) + v_{k,x}(y)}{g_k(y)}| = o_p(1)$$

Here is where the condition $\int\{x : p(x) = \frac{1}{2}t^{-1}\}dx = 0$ is employed.

Now we need only concern ourselves with the region $\mathcal{H}_n$. On $\mathcal{H}_n$ bound (17) by,

$$|(P_n - P)^x \otimes P_n^y[(P_n - P)^z \mathcal{H}_n\{z \in B(y, s_k(y))\}(p(y)s_k(y))^{-2}(w_{k,x}(y) + v_{k,x}(y))]|$$

$$+O_p(\frac{1}{\sqrt{k}n^\delta} + \frac{1}{k^{3/4}}) \log n \sup_y |(P_n - P)^x(w_{k,x}(y) + v_{k,x}(y))s_k(y)^{-1}| \,.$$

Here, the indicator function of the set $\mathcal{H}_n$ is identified as $\mathcal{H}_n$ as well. Again, by (15) the second term when normalized by k converges to 0 uniformly in k. Finally it remains to show:

$$\sup_k \frac{|(P_n - P)^x \otimes (P_n - P)^z[P^y\{z \in B(y, s_k(y))\}\frac{k/n}{p(y)s_k(y)}\mathcal{H}_n \frac{w_{k,x}(y)+v_{k,x}(y)}{s_k(y)}]|}{1/n + k/n \int |\bar{p}_g - p|^2} = o_p(1).$$

The error incurred by replacing the expectation P_n^y by P^y can be ignored. Here we have a degenerate U-statistic process indexed by the collection of functions Γ_n. As in the proofs of (9) and (10) in Nolan and Pollard (1987), Theorem 9 provides the desired rate of convergence if $\sup_{x,z} |\gamma_k(x,z)| < 0$; $\sup_x P^z|\gamma_k(x,z)| + P^z|\gamma_k(z,x)| \leq ck/n$; and the functions in Γ_n meet the Euclidean property. The first two conditions are easily met. The last is justified in the appendix. This completes the proof of Lemma 4. □

Acknowledgement: The authors thank E. Chow for programming the simulations.

References

[1] I.S. Abramson. On bandwidth variation in kernel estimates - a square root law. *Ann. Statist.*, 10:1217-1233, 1982.

[2] I.S. Abramson. Adaptive density flattening - a metric distortion principle for combating bias in nearest-neighbor estimates. *Ann. Statist.*, 12:880-886, 1984.

[3] L. Breiman, W. Meisel, and E. Purcell. Variable kernel estimates of probability densities. *Technometrics*, 19:135-144, 1977.

[4] P. Burman. A data dependent approach to density estimation. *Z. Wahr. Verw. Gebiete.*,69:609-628, 1985.

[5] L. Devroye and T.J. Wagner. The strong uniform consistency of nearest-neighbor density estimates. *Ann. Statist.*, 5:536-540, 1977.

[6] L. Devroye and C.S. Penrod. The strong uniform convergence of multivariate variable kernel estimates of probability densities. 1982. Technical Report, Applied Research Laboratories, University of Texas, Austin.

[7] P. Hall and J.S. Marron. Variable window width kernel estimates of probability densities. 1989. Unpublished manuscript.

[8] D.O. Loftsgaarden and C.P. Quesenberry. A nonparametric estimate of a multivariate density function. *Ann. Math. Statist.*, 36:1049-1051, 1965.

[9] Y.P. Mack and P.K. Bhattacharya. Weak convergence of k-nn density and regression estimators with varying k and applications. *Ann. Statist.*, 15:976-994, 1987.

[10] Y.P. Mack and H.G. Muller. Adaptive nonparametric estimation of a multivariate regression function. *J. Mult. Ann.*, 23:169-182, 1987.

[11] Y.P. Mack and M. Rosenblatt. Multivariate k-nearest-neighbor density estimates. *J. Mult. Ann.*, 9:1-15, 1979.

[12] J.S. Marron. An asymptotically efficient solution to the bandwidth problem of kernel density estimation. *Ann. Statist.*, 13:1011-1023, 1985.

[13] H.G. Muller and Stadtmuller U. Variable bandwidth kernel estimators of regression curves. *Ann. Statist.*, 15:182-201, 1987.

[14] D. Nolan and J.S. Marron. Uniform consistency of automatic and location-adaptive delta-sequence estimators. *Probab. Th. Rel. Fields*, 80:619-632, 1989.

[15] D. Nolan and D. Pollard. U-Processes: rates of convergence. *Ann. Statist.*, 15:780-799, 1987.

[16] D. Pollard. *Convergence of Stochastic Processes.* Springer, New York, 1984

[17] D. Pollard. Rates of uniform almost-sure convergence for empirical processes indexed by unbounded classes of functions. 1986. Unpublished manuscript.

[18] D. Pollard. Asymptotics via empirical processes. *Statist. Sci.,* 4:341-366, 1989.

[19] B.W. Silverman. *Density Estimation for Statistics and Data Analysis.* Chapman Hall, 1986.

[20] C. Stone. An asymptotically optimal window selection rule for kernel density estimates. *Ann. Statist.*, 12:1285-1297, 1984.

Appendix:

In the appendix we establish the Euclidean properties of the classes $\mathcal{W}_n$, $\mathcal{U}_n$, $\mathcal{V}_n$, Γ_n. To do so, we need a few properties of Euclidean classes of functions. Throughout this section we assume the collections of functions are uniformly bounded by the constant 1. The following properties are taken from Nolan and Pollard (1987).

(1) If $\{\mathcal{F}\}$ and $\{\mathcal{G}\}$ are Euclidean then $\mathcal{F}+\mathcal{G}$ is Euclidean, where

$$\mathcal{F}+\mathcal{G} = \{f+g : f \in \mathcal{F} \text{ and } g \in \mathcal{G}\}.$$

(2) If $\mathcal{F}$ and $\mathcal{G}$ are Euclidean then $\mathcal{F}\cdot\mathcal{G}$ is Euclidean, where

$$\mathcal{F}\cdot\mathcal{G} = \{fg : f \in \mathcal{F} \text{ and } g \in \mathcal{G}\}.$$

(3) If $\mathcal{G}$ is a finite-dimensional vector space of real functions then the collection of sets of the form $\{g > 0\}$ is a Vapnik Cervonenkis class.

(4) If the collection sets on $\mathbb{R}^{d+1}$ defined by the graph $(f) = \{(x,s) \in \mathbb{R}^{d+1}$: $0 < s < f(x)$ or $0 > s > f(x)\}$ for $f \in \mathcal{F}$, is a Vapnik Cervonenkis class of sets, then $\mathcal{F}$ is a Euclidean class of functions.

First we show that $\mathcal{W}_n$ in 5.1 is Euclidean. As in Nolan and Pollard (1987), (1), (3) and (4) and the fact that ω is of bounded variation imply $\mathcal{W}_n$ is Euclidean with Euclidean constants that do not depend on n. Briefly, the property of bounded variation implies ω can be expressed as the sum of two monotone functions G and H. By (1), if $\{G_{k.x}\}$ and $\{H_{k,x}\}$ are Euclidean then so is $\mathcal{W}_n$. Consider the graphs of the functions $G_{k,x}$:

From (3) and (4) we find that $\{G_{k,x}\}$ is a Euclidean class of functions and the Euclidean constants do not depend on n. A similar argument implies $\{H_{k,x}\}$ is Euclidean. As for $\mathcal{V}_n$ and $\mathcal{U}_n$, with property (2) plus the fact that

$$\{\frac{x-y}{g_k(x)}\{|x-y| \leq g_k(x)\}\ , x \in \mathcal{R}, n^{\delta} \leq k \leq n^{1-\delta}\ \}$$

is Euclidean, they are handled by an argument similar to that for $\mathcal{W}_n$.

For Γ_n, consider the simpler functions

$$\gamma_k(x,y) \ = \ P^z\{y \in B(z, \frac{k}{n2p(z)}\}\frac{2p(z)}{k/n}w(\frac{(x-z)2p(z)}{k/n})\{p(z) > \frac{1}{2}t\}$$

Then

$$|\gamma_k - p| \ = \ p \qquad\qquad for\ |x-y| > \frac{k}{n}t$$

and by differentiability of ω and p

$$|\gamma_k - p| \; < \; C \sup_{|u|<n^{-\delta}t} |p(x-u) - p(x)| \leq cn^{-\delta} \qquad \textit{otherwise.}$$

Therefore for $\epsilon < n^{-\delta/2}$ no approximation is needed and for $\epsilon \geq n^{-\delta/2}$ approximate Γ_n by the single function p. This argument can be adapted to hold when $s_k(z)$ replaces $k/(2np(z))$.

PROOF OF LEMMA 5: This proof closely follows the proof of Theorem 2.1, Pollard (1986). To begin, assume the f in $\mathcal{F}_n$ are nonnegative and the envelope F is bounded by 1. The result for general f follows by considering $f\{f > 0\}$ and $f\{f \leq 0\}$ separately. Let P_n' represent the distribution based on a second sample $X_1', ..., X_n'$ from P, independent of the first sample. Define for each $k \in [n^{\delta}, n^{1-\delta}]$ and each $f \in \mathcal{F}_n$,

$$\begin{aligned} A_n(k,f) &= \{|P_nf - Pf| > \epsilon_{n,k}(P_nf + Pf + \gamma_{n,k})\} \\ B_n(k,f) &= \{|P_n'f - Pf| \leq \frac{1}{3}\epsilon_{n,k}(Pf + \gamma_{n,k})\} \end{aligned}$$

where

$$\begin{aligned} \epsilon_{n,k} &= M\sqrt{\log n}/k^{\alpha} \\ \gamma_{n.k} &= k^{2\alpha}/n \end{aligned}$$

Our goal is to show

$$\sum_n \mathbf{P}(\bigcup_{k,f} A_n(k,f)) < \infty \; .$$

For each n, bound the individual summand:

$$(1) \qquad \mathbf{P}(\bigcup_{k,f} A_n(k,f)) \leq 2\mathbf{P}(\bigcup_{k,f} A_n(k,f) \cap B_n(k,f))$$

This inequality follows from the independence of $\{A_n\}$ and $\{B_n\}$ and the fact that there exists an n_o such that for $n \geq n_o$, $\mathbf{P}B_n(k,f) \geq \frac{1}{2}$. The lower bound of $\frac{1}{2}$ for $\mathbf{P}B_n(k,f)$ follows from Chebychev's inequality applied to $B_n(k,f)^c$. The inequality (1) is a classical result for countable $\mathcal{F}_n$ (Loeve 1977, 18.1.A). Pollard (1986, section 6) extends this result to uncountable $\mathcal{F}_n$ that are permissible (Pollard 1984, Appendix C).

On the event $A_n(k,f) \cap B_n(k,f)$

$$\begin{aligned} |P_n f - P_n' f| &\geq \epsilon_{n,k}[P_n f + \frac{2}{3} P f + \frac{2}{3}\gamma_{n,k}] \\ &\geq \frac{1}{6}\epsilon_{n,k}[P_n f + \gamma_{n,k} + P_n' f + \gamma_{n,k}] \end{aligned}$$

This inequality implies
(2)

$$\mathbf{P}(\bigcup_{k,f} A_n(k,f)) \leq 2\mathbf{P}\{\exists f \in \mathcal{F}_n, k \in [n^\delta, n^{1-\delta}] : |P_n f - P_n' f| \geq \frac{1}{6}\epsilon_{n,k}[P_n f + P_n' f + 2\gamma_{n,k}]\}$$

Introduce a third sample, a sample of sign variables $\{\sigma_i\}$ where $\sigma_i = \pm 1$ with probability $\frac{1}{2}$, independent of the first two samples. Then the right hand side of (2) equals:

$$2\mathbf{P}\{\exists f, k : |\frac{1}{n}\sum_i \sigma_i(f(X_i) - f(X_i'))| > \frac{1}{6}\epsilon_{n,k}[P_n f + P_n' f + 2\gamma_{n,k}]\}$$

$$(3) \quad \leq 4\mathbf{P}[\mathbf{P}\{\exists f, k : |\frac{1}{n}\sum_i \sigma_i f(X_i)| > \frac{1}{1}2\epsilon_{n,k}(P_n f + \gamma_{n,k})|X_1, ..., X_n\}]$$

Next approximate $\mathcal{F}_n$ within $\frac{1}{2}4\epsilon_{n,n^\delta}\gamma_{n,n^\delta}$ by $\mathcal{F}_n^*$. According to the main condition of the lemma, for some constant C,

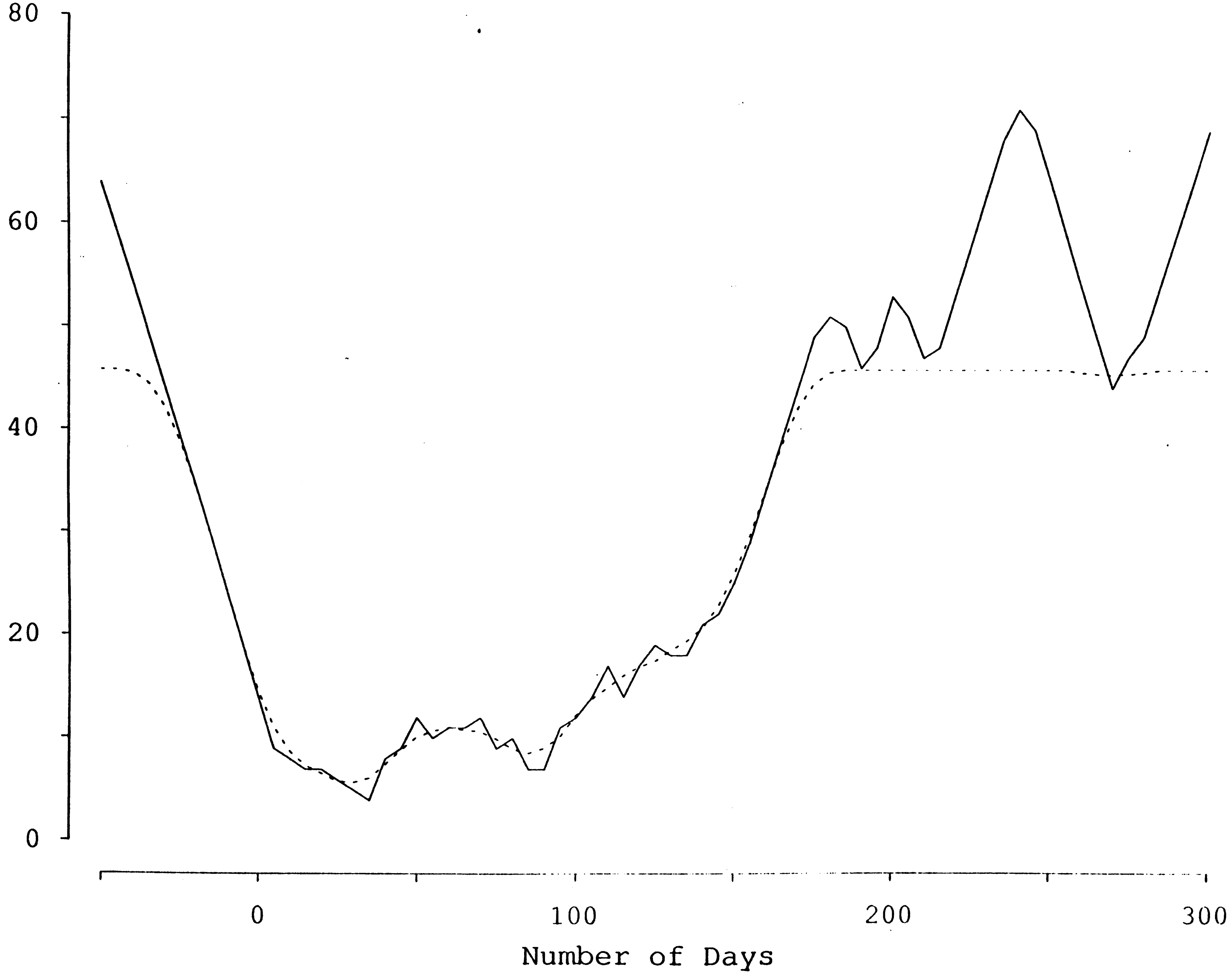
$$cardinality(\mathcal{F}_n^*) \leq C(n^{1-\delta})^V \ .$$

This approximation provides an upper bound on the right hand side of (3):

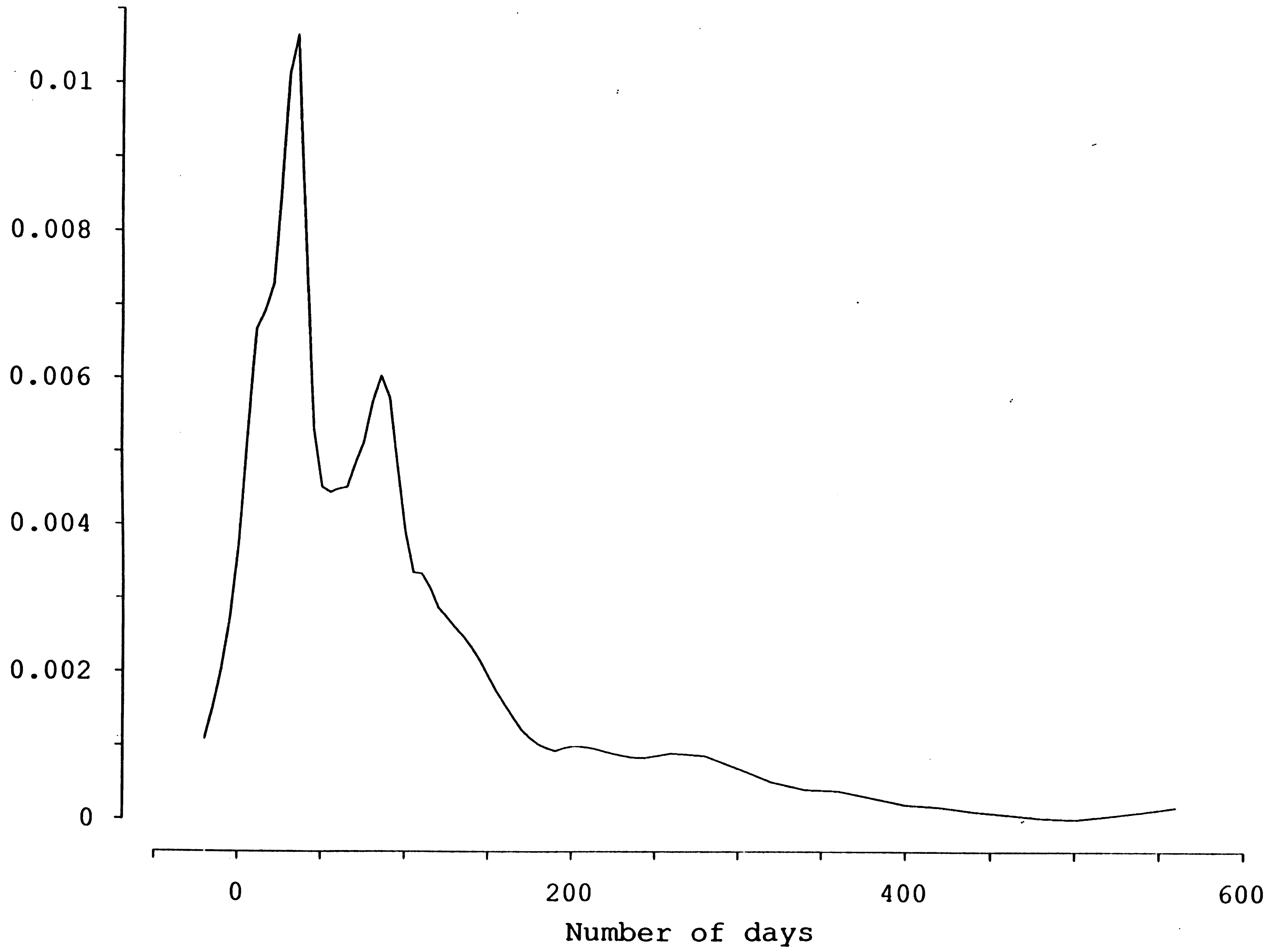
$$\begin{aligned} & 4C(n^{1-\delta})^{V+1}\mathbf{P} \max_{k,\mathcal{F}_n^*} \mathbf{P}\{|\frac{1}{n}\sum_i \sigma_i f(X_i)| > \frac{1}{2}4\epsilon_{n,k}(P_n f + \gamma_{n.k})|X_1, ..., X_n\} \\ \leq \ & 8C(n^{1-\delta})^{V+1}\mathbf{P} \max_{k,\mathcal{F}_n^*} \exp \frac{-1}{256}\epsilon_{n,k}^2 (P_n f + \gamma_{n.k})^2 / \frac{1}{n} P_n f^2] \\ \leq \ & 8C(n^{1-\delta})^{V+1} \exp - M^2 \log n / 576] \end{aligned}$$

The first inequality is due to a conditional application of Hoeffding's inequality to the centered, bounded random variables $\{\sigma_i f(X_i)\}$. If M is sufficiently large, the final upper bound has a finite sum in n. This concludes the proof of Lemma 5.

Smoothed & Truncated Nearest-Neighbor Distance (dashed line)
15th Nearest-Neighbor Distance (solid line)

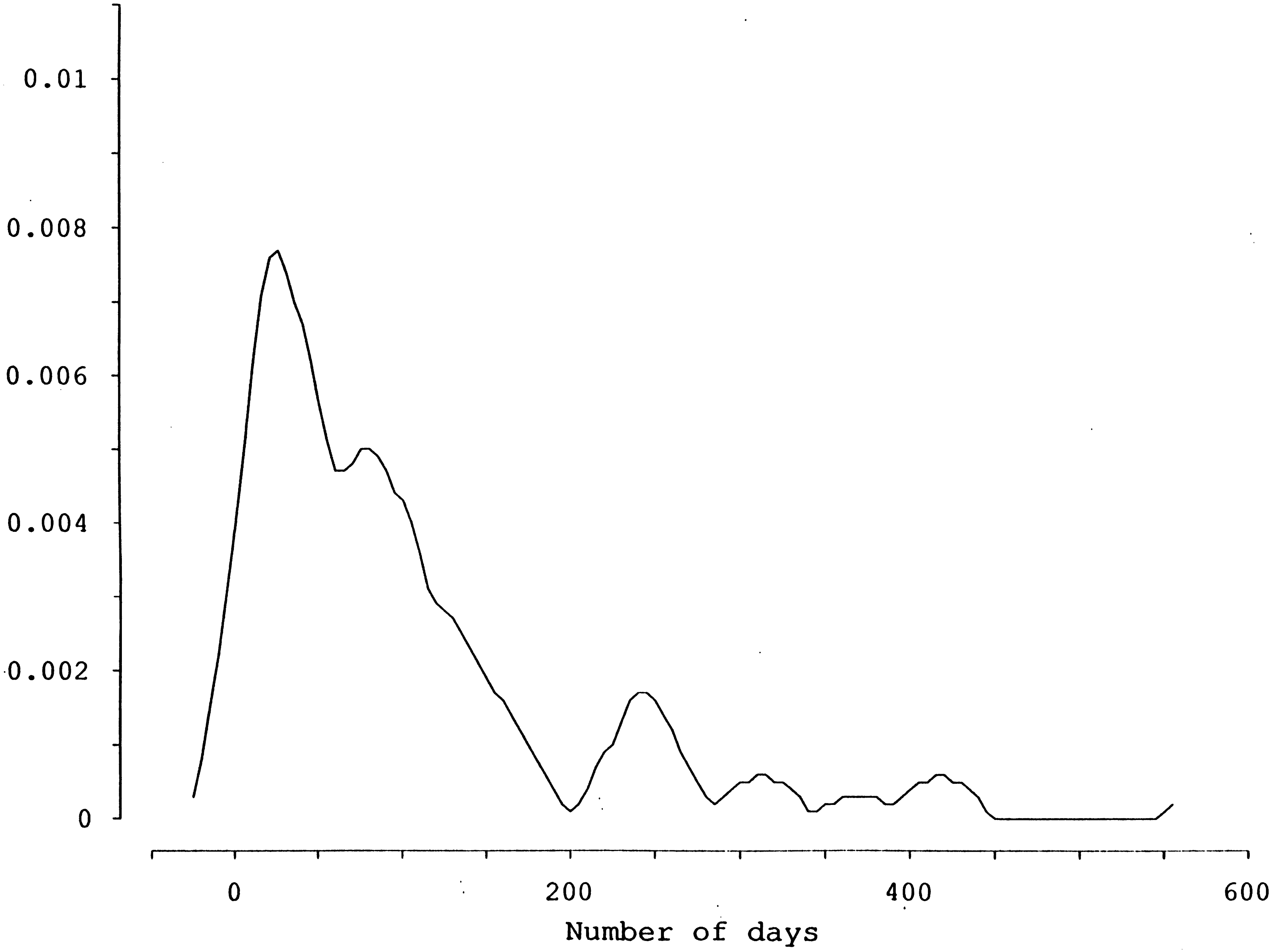


Smoothed a[illegible] uncated Nearest-Neighbor Estimate: k=15 t=250



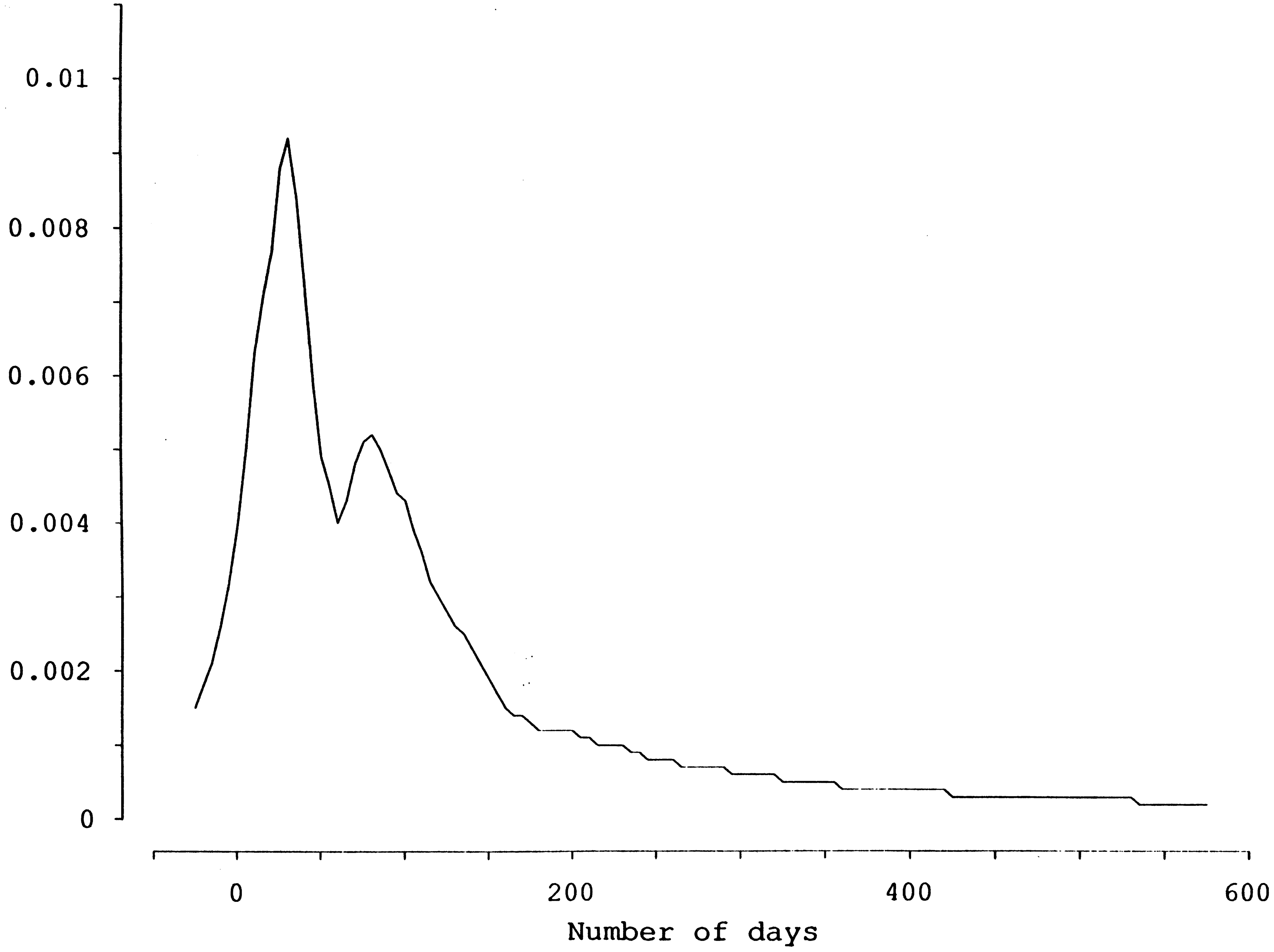
2b

Kernel estimate: h=30



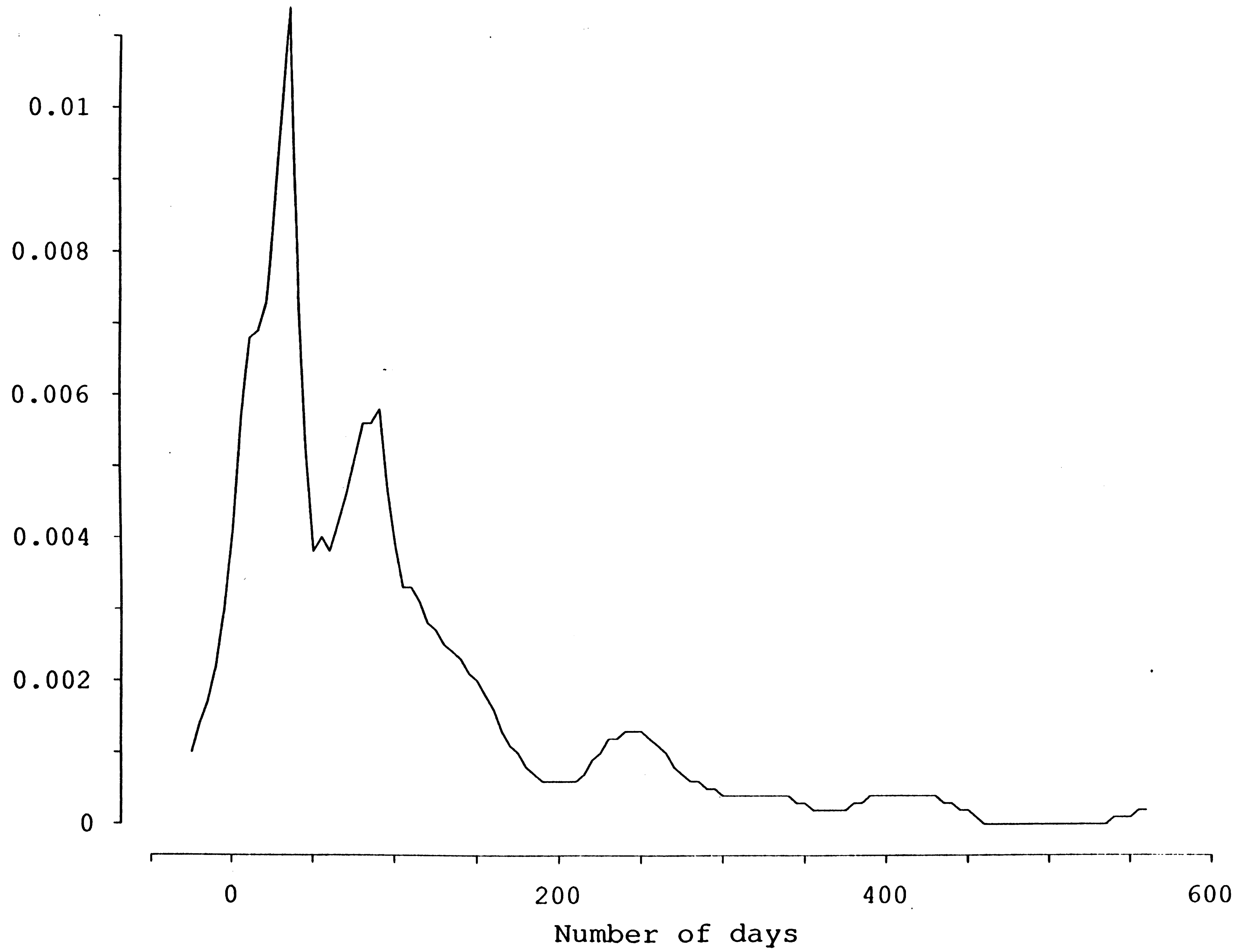
2c

22nd Nearest-N   bor Estimate



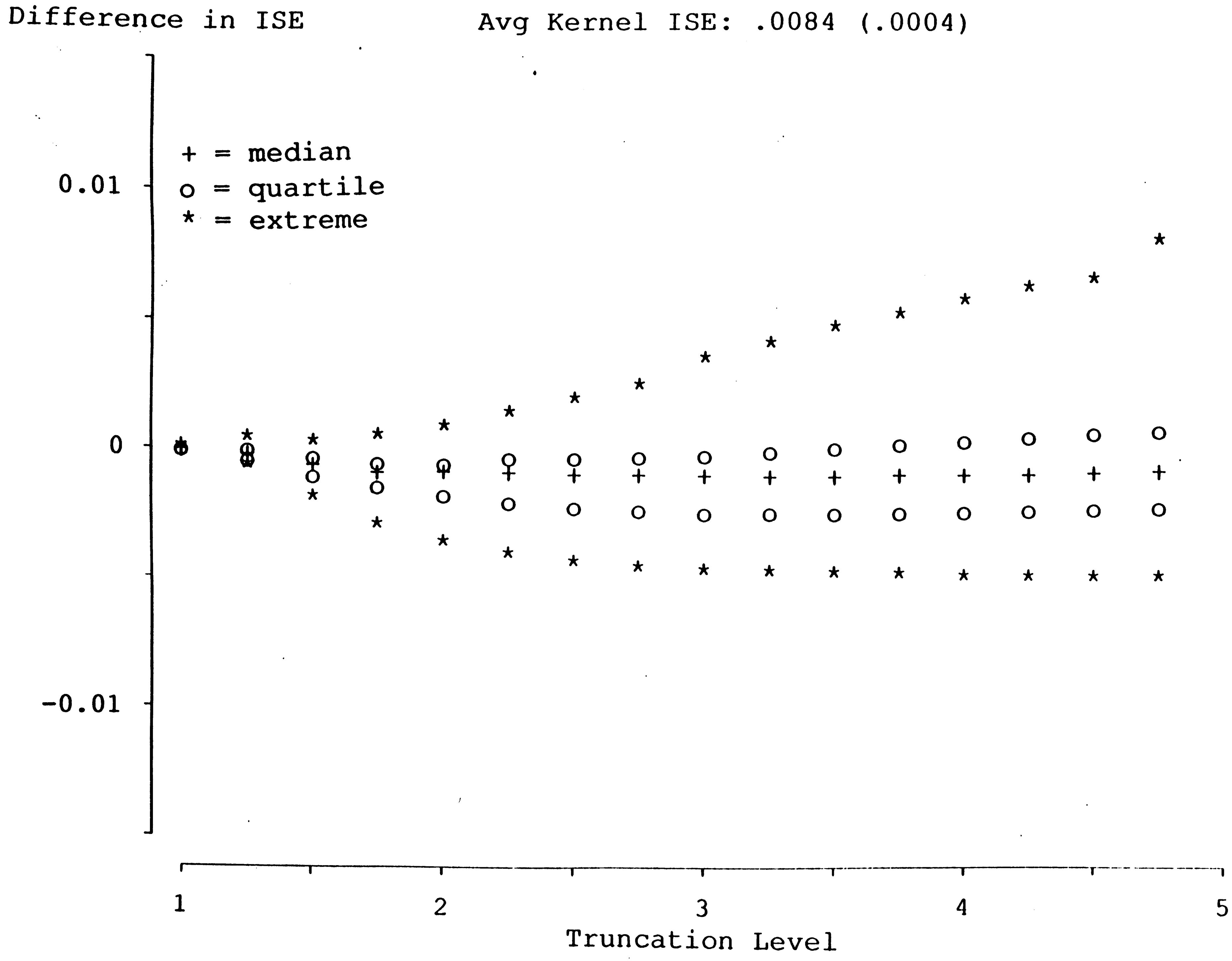
(d

Truncated Nearest-Neighl ;timate: k=15 t=250



3a.

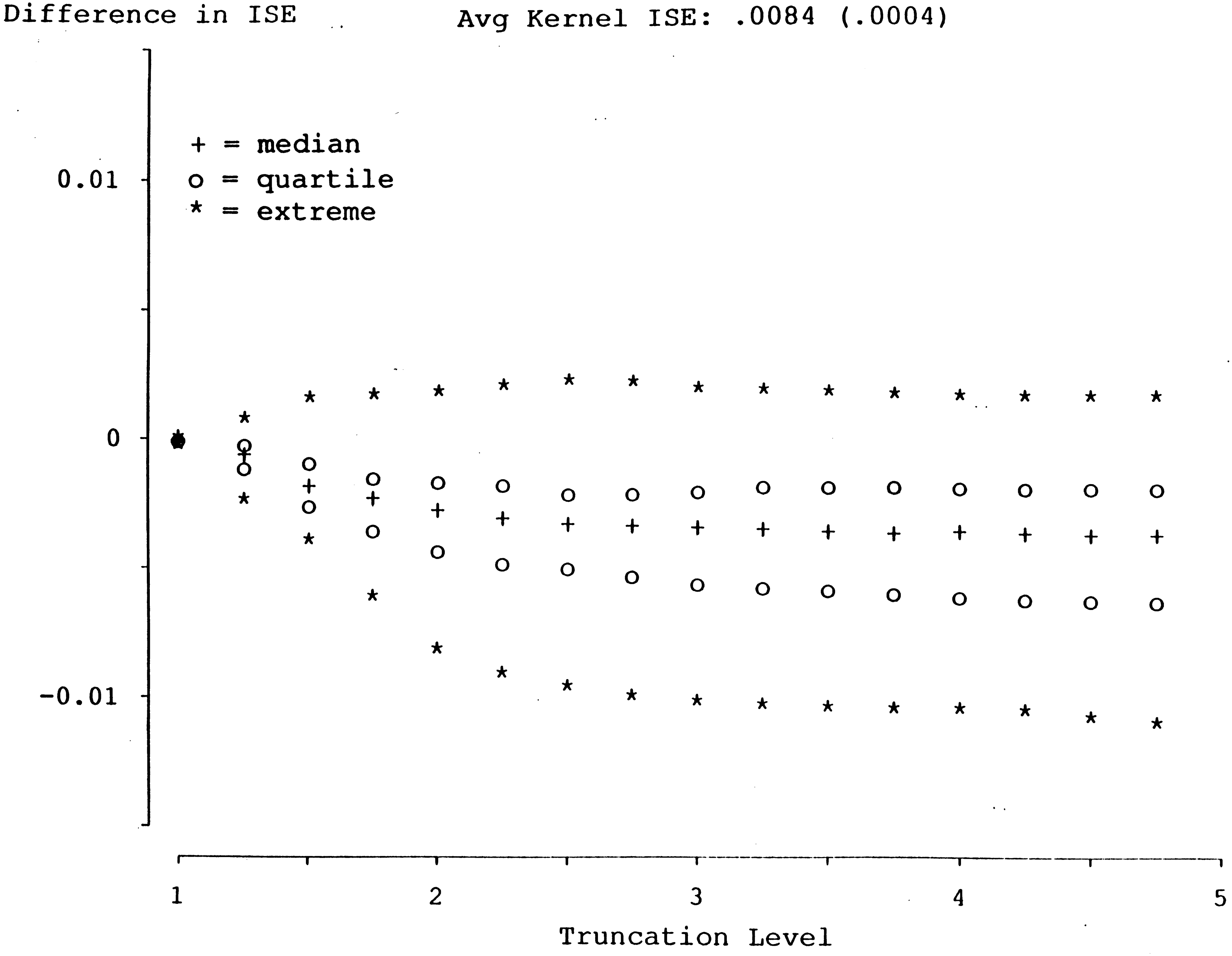
Trunca NN vs Kernel for Double Exponential
Avg Kernel ISE: .0084 (.0004)



3b

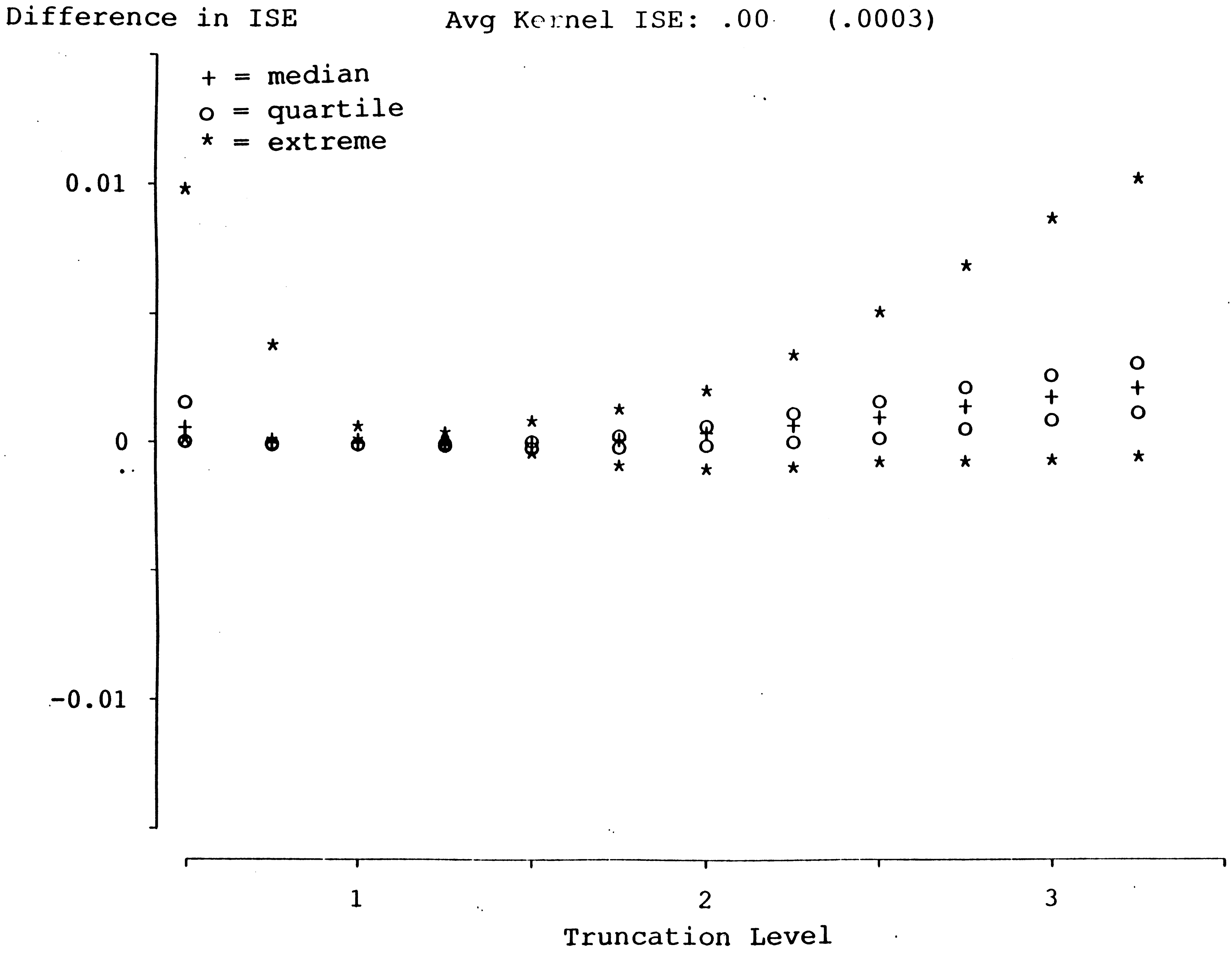
Trun(BMP vs Kernel Double Exponential

Avg Kernel ISE: .0084 (.0004)

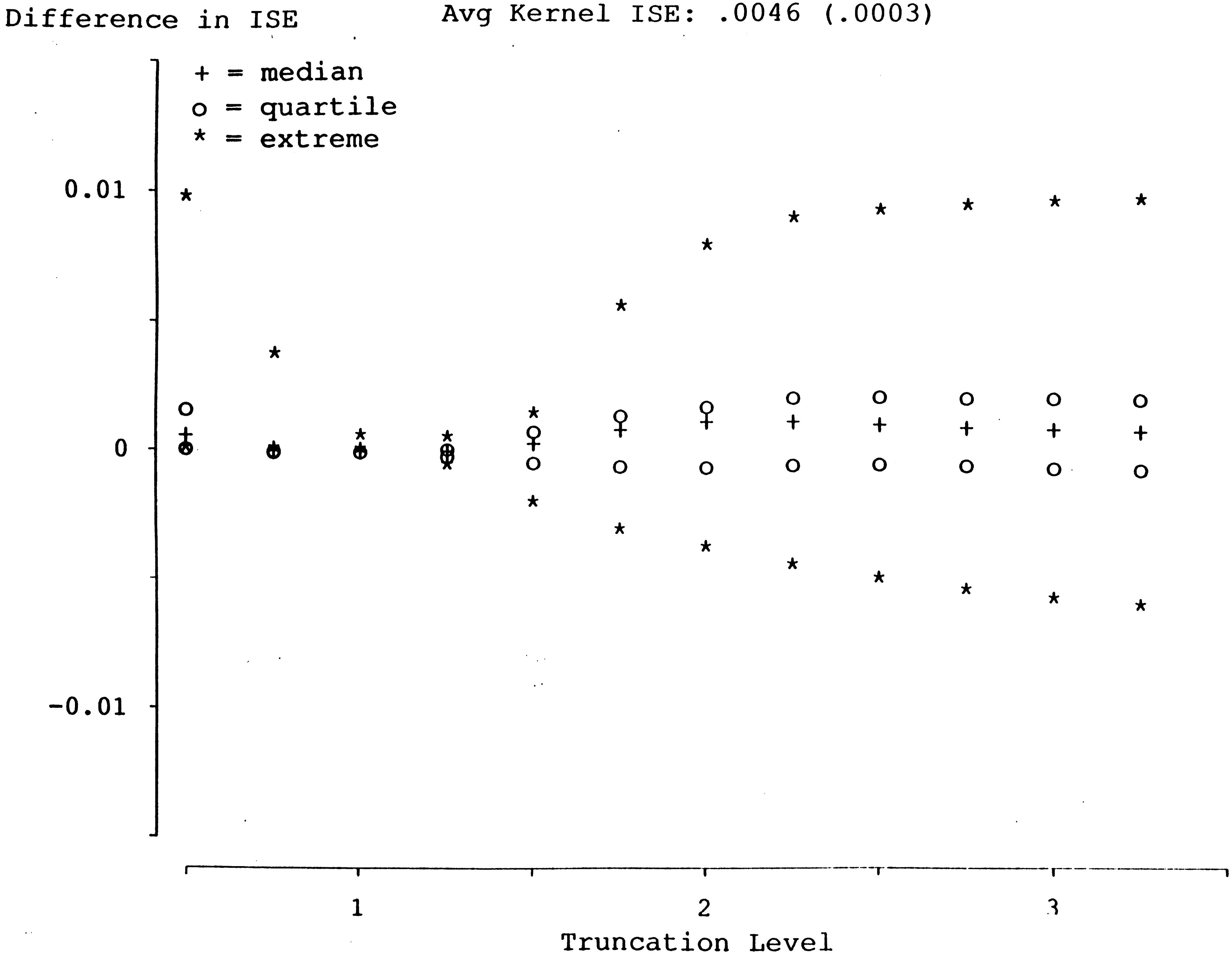


Truncated NN vs Kernel for Normal

Avg Kernel ISE: .00 (.0003)

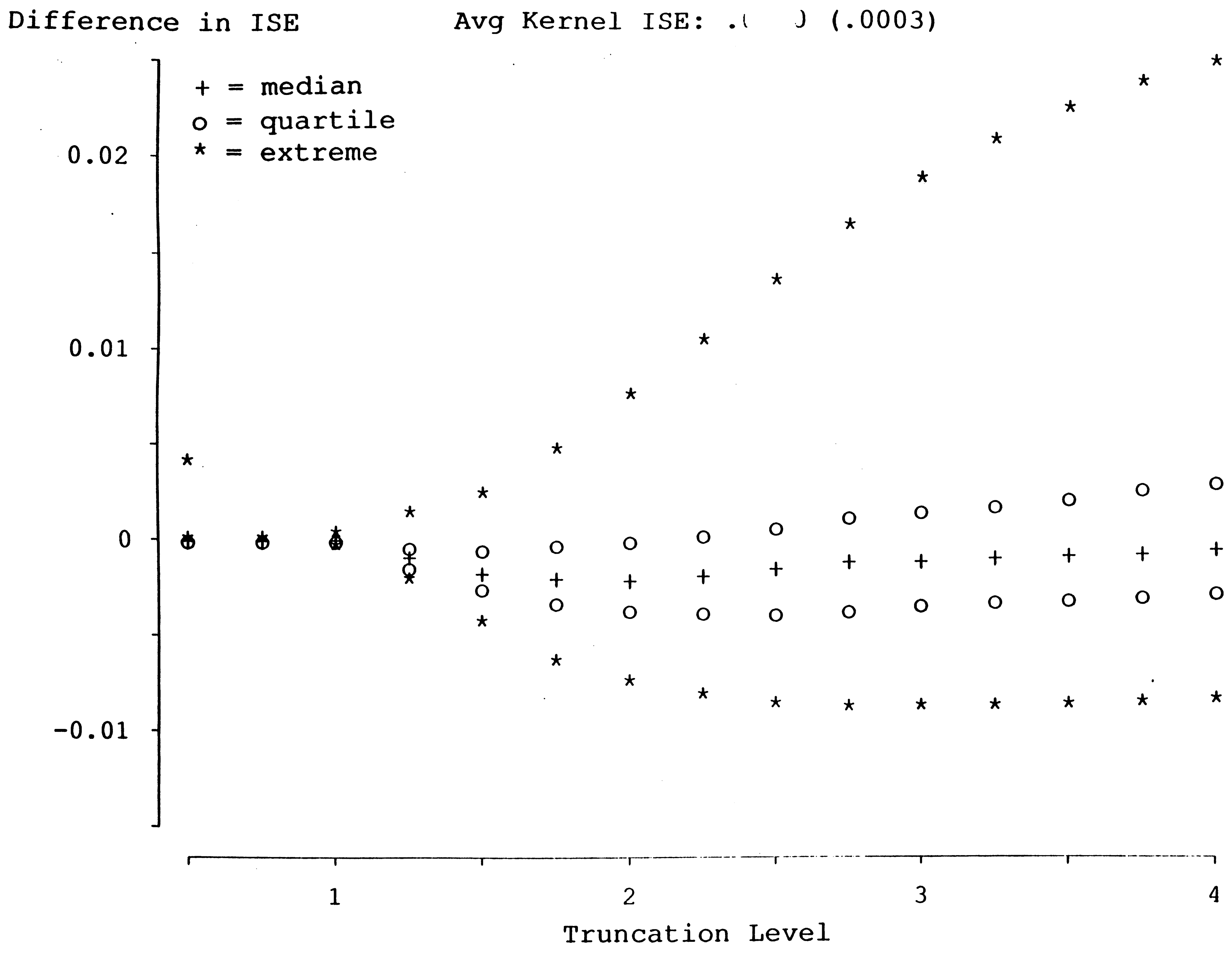


runcated BMP vs Kernel ıor Nolmal
Avg Kernel ISE: .0046 (.0003)



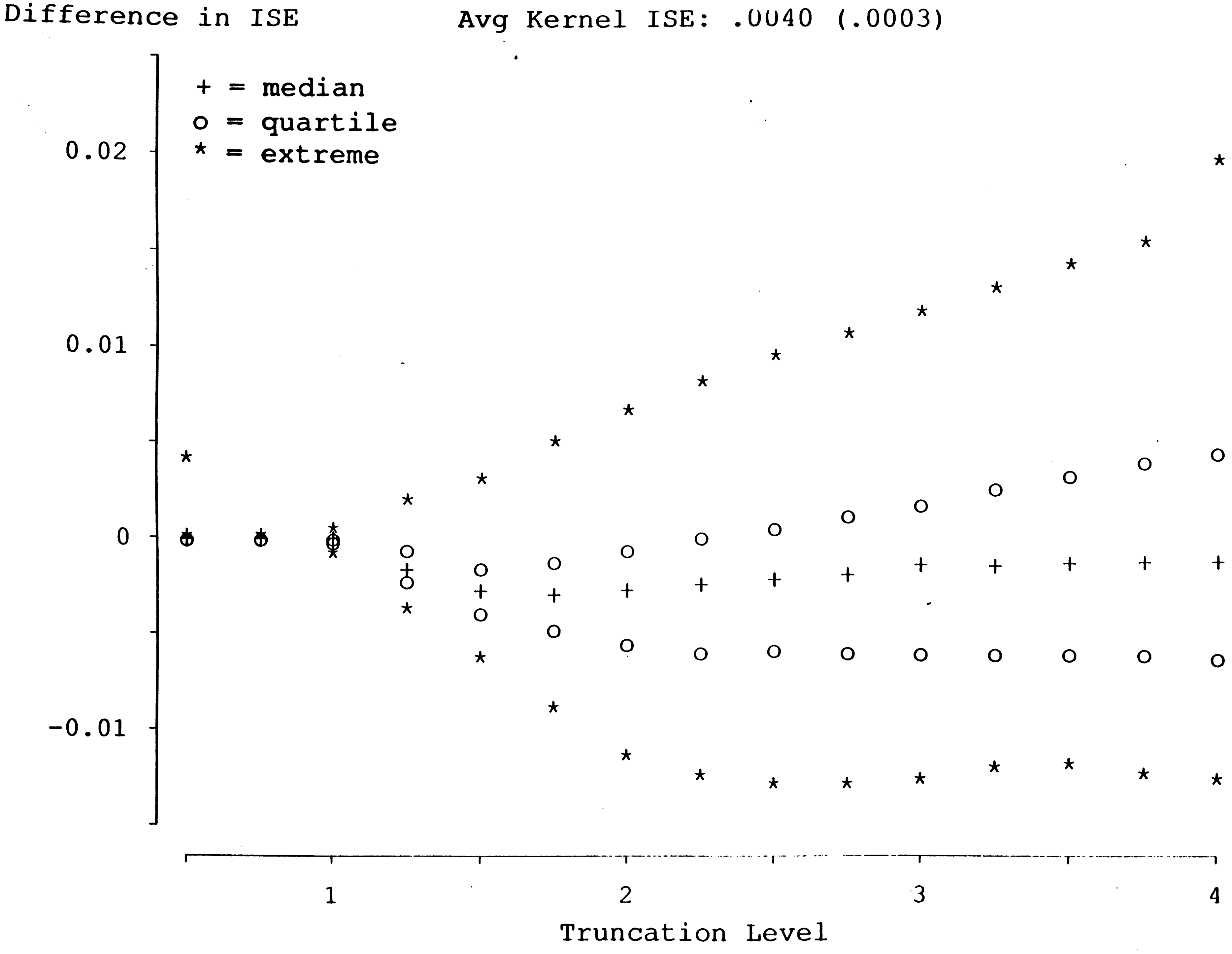
Truncat NN vs Kernel for Mixt f Double Exponentials

Avg Kernel ISE: .) (.0003)



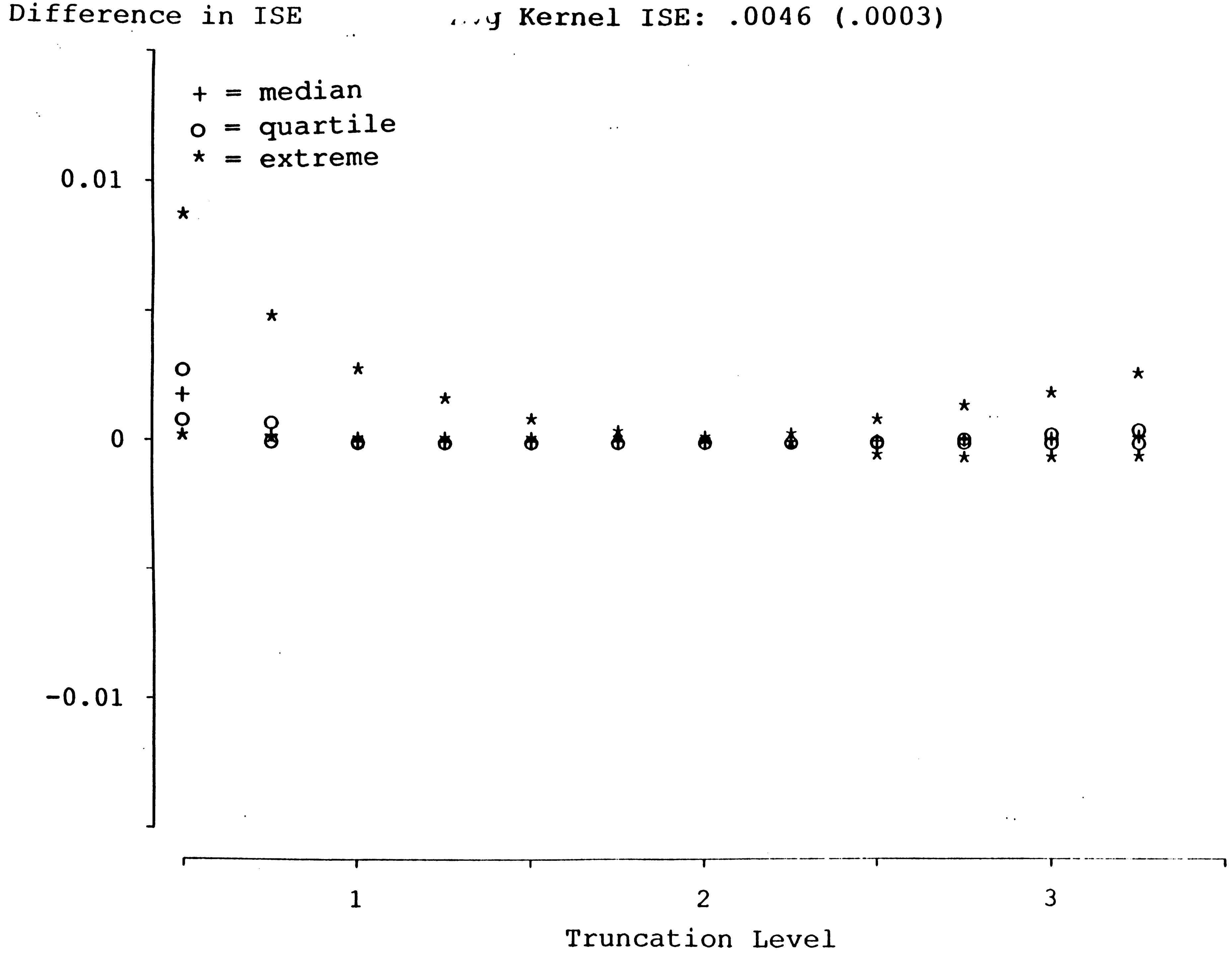
Truncated BMP vs Kernel for M of Double Exponentials

Avg Kernel ISE: .0040 (.0003)



T[illegible]at[illegible] vs Kernel for [illegible]xture of Normals

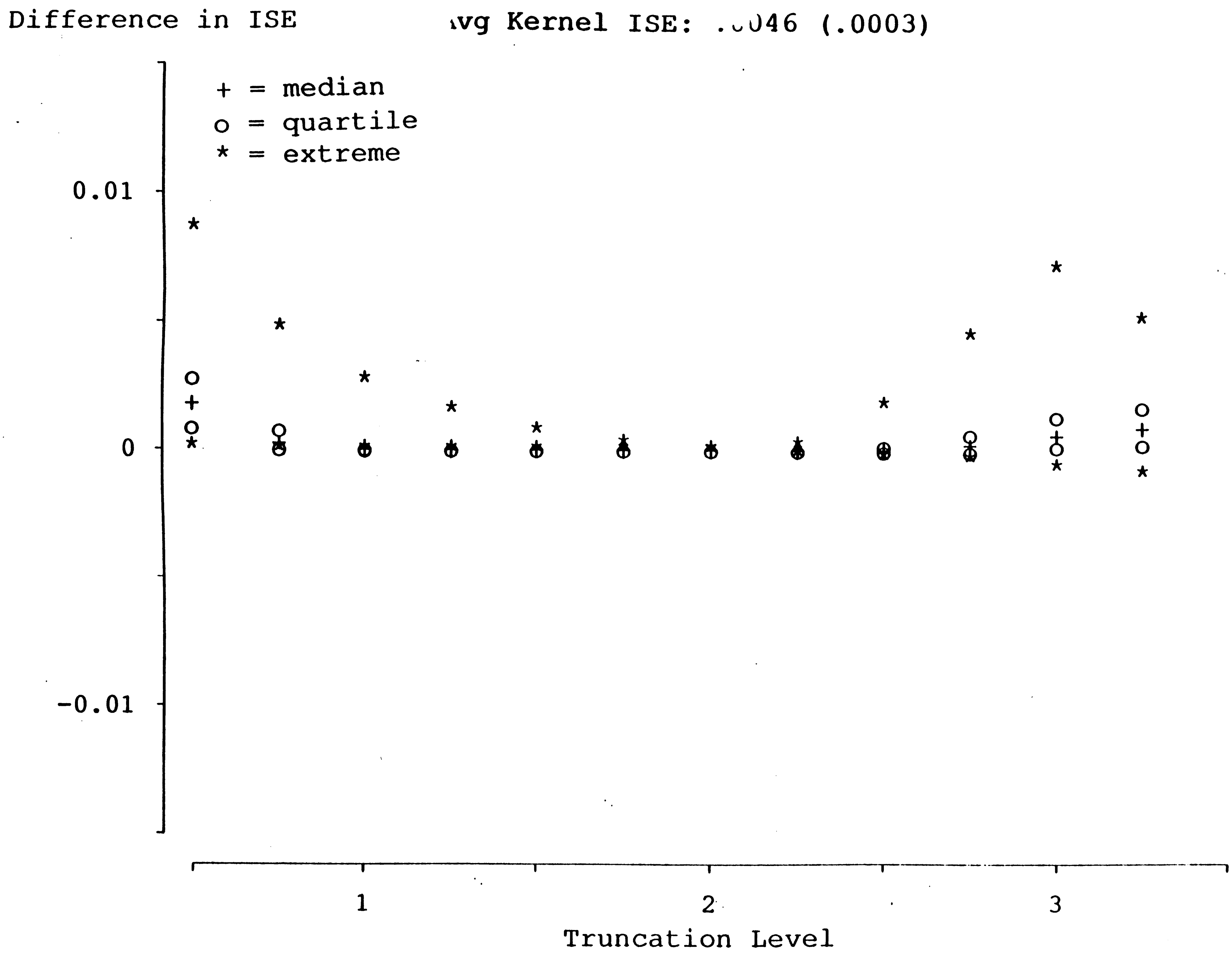
[illegible]g Kernel ISE: .0046 (.0003)



3A

Trunc BMP vs Kernel Mixture of Normals

vg Kernel ISE: ..046 (.0003)


Difference in ISE
+ = median
o = quartile
* = extreme
Truncation Level