GRAPHICAL UNDERSTANDING OF HIGHER ORDER KERNELS

by

J.S. Marron

September 1992

Mimeo Series #2082

DEPARTMENT OF STATISTICS
Chapel Hill, North Carolina

Graphical understanding of higher order kernels

J. S. Marron*
Department of Statistics
University of North Carolina
Chapel Hill, N. C. 27599-3260

September 9, 1992

Abstract

Higher order kernels have been suggested for use in nonparametric curve estimation. The usual motivation is that they often give faster asymptotic rates of convergence. This paper provides a visual derivation of higher order kernels. This gives new insight into how they work. Furthermore, this makes it clear, in a visual sense, when they may be expected to work, and why they frequently do not work.

1 Introduction

Kernel methods provide an appealing framework for nonparametric curve estimation, including the estimation of both densities and regression functions. Major reasons for this appeal are their simplicity, and their interpretability. These properties come from the fact that they are easily understood local averages.

An interesting proposal for enhancing the performance of nonparametric curve estimators, is through the use of "higher order kernels" (formally defined in section 3). This idea has been around for many years. It dates back

*Research supported by NSF Grant DMS-92023135

at least to some of the earlier work on spectral density estimation in time series. It is still quite commonly discussed in the mathematical statistics literature.

The motivation that is usually given is based solely on the calculation of the asymptotic rate of convergence (see section 3), which is typically faster for higher order kernels. However from a data analytic viewpoint, there is a price that must be paid for the use of higher order kernels: they involve local averaging with *negative* weights. This drawback is substantial, because such a local average loses much in terms of interpretability. Much of that simple "moving average" intuition is lost, through difficulties in comprehending the effect of negative weights. Perhaps for this reason, higher order kernels are nearly never used in practice. An important question is "do these methods provide enough improvement that they should be used more in practice?". The answer must depend on the sample size, because there is usually *some* sample size large enough for the higher order kernels to be far superior, in most senses. However [2, Marron and Wand] have shown that the sample sizes needed, for the asymptotic benefits to be important, are all too often very large indeed. They conclude that in most situations, the benefits in terms of performance of higher order kernels are not sufficient to outweigh the loss of interpretability.

The goals of this paper are:

Goal I Provide *visual* insight into why higher order kernels work, in particular why they have a faster asymptotic rate of convergence.

Goal II Show graphically why very large sample sizes are required before the asymptotically predicted superiority of higher order kernels is realized. Furthermore provide a tool for understanding, in a given situation, whether or not use of a higher order kernel will be significantly beneficial.

Section 2 addresses Goal I. The ideas there are then combined with the usual asymptotic approach to higher order kernels in section 3. These ideas are extended to address Goal II in section 4. The main lesson is that higher order kernels are more effective than nonnegative ones when the underlying curve being estimated is effectively approximated by local parabolas, on neighborhoods of appropriate width.

Section 5 discusses extension to higher order kernels. Possibilities for practical application of these ideas, are briefly discussed in Section 6.

2 Visual Motivation of Higher Order Kernels

The main idea is demonstrated in Figure 1. The solid curve in both parts represents a curve to be estimated. For simplest understanding, suppose it is a regression curve. It is desired to estimate the curve, using data which consist of the curve plus some mean zero, independent, identically distributed "noise", at a grid of equally spaced points. A simulated representation of this setting is given in Figure 1a.

[put figure 1 about here]

FIGURE 1: EXAMPLE ILLUSTRATING THE VISUAL MOTIVATION FOR HIGHER ORDER KERNELS, IN NONPARAMETRIC REGRESSION. FOR EACH KERNEL ESTIMATE, THE WEIGHTS ARE REPRESENTED BY THE CURVES AT THE BOTTOM, CENTERED AT $x = 0.5$.

A simple means of constructing a kernel estimate, of the solid curve, is to take a "moving weighted average" of the observations, with weights proportional to the height of the appropriately centered "kernel function". For estimation at the point $x = 0.5$, this function is shown at the bottom of the plot. See for example [1, Chu and Marron] for discussion of other versions of kernel regression estimation, and for access to the literature.

A well known feature of such curve estimates is that they have a tendency to "miss peaks and valleys". This is visually apparent in Figure 1a. At locations near the peak, for example $x = 0.5$, the estimate is too low. This is because the weighted average includes points whose mean is much lower than the mean at points near the peak. This effect is well understood mathematically in terms of "bias" (quantified in section 3). Figure 1b visually analyzes the problem, by showing the expected value of the estimator in Figure 1a, which is again just the same moving average of the solid underlying regression curve.

One possibility for attempting to overcome this problem is to consider replacing this weighted average, by a similar one, which puts slightly more weight near the peak of the kernel, and which puts a corresponding amount of *negative* weight in the tails. Note that the observations with negative weight have a tendency to reverse the effect of biasing the curve downward. The dotted set of weights, and resulting moving average of the solid curve, show the effect of this operation, for a very carefully chosen version. Note the dotted curve is much closer than the dashed curve to the true underlying solid one.

The main hurdle to implementing this idea, is to settle the delicate issue of how much negative weight to use. This requires some mathematical analysis, which is done in section 3. An interesting and perhaps surprising conclusion is that the answer is nearly independent of the shape of the underlying curve.

See [3, Schucany and Sommers] for a different motivation of higher order kernels, based on the generalized jackknife.

3 Mathematics of Higher Order Kernels

The main points here are most simply presented in terms of kernel density estimation, so that is the context of this section. The main ideas (and even the final answer in terms of what generates higher order kernels) are the same for both density and regression estimation. Furthermore, Figure 1b is the same for both density and regression estimation.

See figures 2.4 and 2.5 of [4, Silverman] for a good visual introduction to kernel density estimation. A mathematical formulation of this problem is to consider using $X_1, ..., X_n$, which are a random sample from the probability density $f(x)$, to try to recover the curve $f(x)$. The kernel approach to this problem, i.e. the "kernel density estimator" is given by

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

where $K_h(\cdot) = \frac{1}{h} K\left(\frac{\cdot}{h}\right)$, for a bandwidth h and some symmetric "kernel function" K, which satisfies $\int K = 1$ (here and below, and integral sign with no limits is understood to mean $\int_{-\infty}^{\infty}$). The intuitively natural choice of K would be a probability density itself. However considerations of the type

given in section 2 motivate the consideration of kernel functions which take on negative values.

As indicated in section 2, the motivation for using kernels with negative weights appears in the expected value of the estimator, which is given here by

$$E\hat{f}_h(x) = n^{-1}\sum_{i=1}^{n} EK_h(x - X_i) = \int K_h(x-t)f(t)dt = K_h * f(x)$$

where $*$ denotes convolution (recall this is simply a continuous version of the moving weighted average). The dashed and dotted curves in figure 1b are convolutions of this type, because the expected estimator is essentially a convolution of the kernel and the true curve in both density and regression estimation.

The bias in $\hat{f}_h(x)$ will thus be exactly canceled at a point x when $K_h * f(x) = f(x)$. Now negative weights of the type shown in figure 1b, can usually be constructed to make this happen exactly, but they will be heavily dependent on both x and the true underlying f. A more practically useful solution can be developed by some simple asymptotics, as $h \to 0$. In this case, assuming f has sufficiently many derivatives at x, the expected value admits the formal Taylor representation (needed mathematical verification is easily provided, for example as in [4, Silverman])

$$E\hat{f}_h(x) = K_h * f(x) = \int K(u) f(x - hu)\, du$$

$$\sim \int K(u) \left[\sum_{j=0}^{\infty} (j!)^{-1} (-hu)^j f^{(j)}(x)\right] du$$

$$\sim f(x) + \sum_{\ell=1}^{\infty} ((2\ell)!)^{-1} h^{2\ell} f^{(2\ell)}(x) \int u^{2\ell} K(u) du$$

where the last representation uses the symmetry of K. Note that the first, $\ell = 1$, term of the bias can be eliminated by the simple condition $\int u^2 K(u)du = 0$. This provides the solution to the problem raised in section 2: how much negative weight to use. The solution is practically useful because it is independent both of x and of f.

The above asymptotic representation for the expected value also shows two other things. First, when using a higher order kernel, one may expect a

faster rate of convergence (this does indeed happen). Second, the improvement can be extended by making more of the moments of K vanish.

This last point motivates the definition of a "kernel of order k", as a symmetric function K which satisfies

$$\int x^{\ell} K(x)\, dx = \begin{cases} 1 & \ell = 0 \\ 0 & \ell = 1, ..., k-1 \\ C > 0 & \ell = k \end{cases} .$$

Note that the higher the kernel order, the faster the asymptotic rate of convergence. However it is intuitively clear from the above asymptotic representation, and the ideas in section 2 that these kernels can be viewed as successively finer and finer adjustments to the bias problem. Hence it is sensible to expect the effects to be smaller and smaller for higher kernel orders, and also to take larger and larger sample sizes before significant improvements are realized. These points have been observed in [2, Marron and Wand].

The dashed kernel function used in Figure 1a is Gaussian in shape. The higher order kernel used in figure 1b is the Gaussian type higher order kernel described in [5, Wand and Schucany].

4 The Failure of Higher Order Kernels

This section provides visual insight as to why higher order kernels often, for moderate sample sizes, do not yield the performance promised visually in figure 1, and asymptotically in section 3. This time the pictures are for density estimation, but again the Gaussian and Gaussian based higher order kernels are used.

[put figure 2 about here]

FIGURE 2: TRIMODAL NORMAL MIXTURE, WITH EXPECTED VALUE OF THE SECOND ORDER AND FOURTH ORDER KERNEL DENSITY ESTIMATES, FOR KERNELS SHOWN AT THE BOTTOM.

Figure 2 shows an underlying true density which is a symmetric mixture of three normal distributions. As expected, at the point $x = 0.5$, the second order, nonnegative, kernel estimate is biased downward. However, now the

higher order kernel does not provide much improvement at this point. The reason is visually apparent: where the higher order kernel takes on negative values, the true density is actually *higher* (at some places), than at the peak at $x = 0.5$. This tends to substantially dampen the effect of the negative weights that was so helpful in Figure 1.

From a mathematical point of view, the problem is caused by the above Taylor expansion *not* well representing what occurs in this picture. In particular, the true density function is *not* well approximated by a parabola over the region of interest, especially including the locations where the kernel is negative. This parabolic approximation is essential for the condition $\int u^2 K(u) du = 0$ to give the visual cancellation illustrated in Figure 1b.

Note that for the valleys, near $x = 0.74$ and $x = 0.26$, this same effect is even worse. Here the expected value of the higher order kernel estimate is even worse than the second order estimate.

The expansions in section 3 indicate that the performance of the higher order kernels, relative to the second order, should improve with smaller h. This effect is demonstrated in Figures 2b and 2c. In Figure 2b, the performance is much improved at $x = 0.5$, with the higher order kernel giving far superior bias performance. This fits well with the "local parabolic approximation" insight above However, at the peaks near $x = .15$ and $x = .85$ the improvement of the higher order kernel are quite marginal. The reason is again visually clear: the density is not well approximated by a parabola in the needed regions, especially in the region where the higher order kernel is putting its negative weight.

In Figure 2c, the true density, at all points, is now well approximated by a parabola over windows of the radius needed by this much narrower kernel. As expected, the asymptotics now provide an effective description of the situation, with the higher order kernel giving far superior performance at all locations.

Plots have been made to see how well these ideas can be used to explain the often poor moderate sample size performance of higher order kernels, for all of the test densities of [2, Marron and Wand]. These are too many to include, and the main ideas are roughly similar, so only Figure 3 is included here.

[put figure 3 about here]

FIGURE 3: FIGURE 3A IS THE OPTIMAL MEAN INTEGRATED SQUARED ERROR, AS A FUNCTION OF SAMPLE SIZE. OTHERS SHOW THE UNDERLYING DENSITY, OVERLAYED WITH THE FOURTH ORDER KERNEL FUNCTIONS FOR THE OPTIMAL BANDWIDTHS AT SELECTED SAMPLE SIZES, REPRESENTED IN FIGURE 3A.

Figure 3a shows how the Mean Integrated Squared Error (for the kernel of order k):

$$MISE_k(h) = E \int \left(\hat{f}_h - f\right)$$

evaluated at the optimal bandwidth h_{MISE_k} (the bandwidth to minimize $MISE_k$, with ties broken arbitrarily), goes down with the sample size. These are neither asymptotic, nor simulation based approximations, but instead are calculated exactly, using the ideas in [2]. The target density is the asymmetric claw density #12 from that paper (a Normal mixture, with parameters given there), shown here as the solid curve in Figures 3b, 3c and 3d. The lessons of $MISE$ are slightly different from those studied above, in that this provides a single number which summarizes (in an integral sense, over all locations) the expected squared error. Figure 3a shows that for sample sizes roughly up to $n = 70$, the fourth order kernel is superior (in the sense of minimum $MISE$), while for about $70 \leq n \leq 10,000$, the nonnegative second order kernel is superior, with fourth order kernel again becoming dominant for the very large (at least in the opinion of many statisticians) $n > 10,000$.

Perhaps the most interesting question arising from Figure 1a is: "why is the second order kernel better for the middle sample sizes?" This is answered in Figure 1c, where deeper visual insight is given for the particular choice $n = 79$ (chosen to minimize the ratio $MISE_2/MISE_4$). The reason that this is the n where $MISE_4$ is at it worst, relative to $MISE_2$, is easily understood from the lessons above: the underlying density is not well approximated by a parabola (on the needed neighborhoods) at any of the peaks. This counts heavily against the fourth order kernel, since $MISE$ is squared error, which feels the largest deviations, which are usually at the peaks, most strongly. For larger values of n, the optimal bandwidth h_{MISE_4} decreases, so there is good parabolic approximation for the fatter peaks, but the skinnier peaks are enough to keep $MISE_2 < MISE_4$, until quite large values of n.

Figure 3d shows that for $n = 100,000$, the optimal bandwidth h_{MISE_4} is so small that there is good parabolic approximation everywhere, except at the thinnest peak. Clearly for larger values of n the asymptotics will then "kick in", and the fourth order kernel will be clearly dominant.

Figure 3b is of interest, because it shows that with only $n = 10$ observations, it is optimal in the $MISE_4$ sense to ignore all of the spikes, and simply try to capture the "gross global structure" of the underlying density. The reason the fourth order kernel is slightly superior in this situation is that with such large bandwidths, all kernel estimators spread substantial amounts of probability mass outside the interval $[-3, 3]$, but the negative weights in the tail of the fourth order kernel have a tendency to mitigate this effect somewhat.

5 Extension to Higher Orders

The above discussion focuses on the "first" higher order kernel, the one of order $k = 4$. It is clear that one may similarly motivate kernels of successively higher order, by looking further and further into the tails of the kernel, and adding successively smaller bits alternating between negative and positive. As noted in section 3, for any fixed sample size, the relative magnitudes of benefits will diminish with increasing k.

A natural question is: "Are there situations where $MISE_6 \ll MISE_4 \approx MISE_2$?" In view of lessons above, we expect $MISE_4 \approx MISE_2$ to indicate that the underlying density is poorly approximated by a parabola on neighborhoods with radius h_{MISE_4}. To have $MISE_6 \ll MISE_4$, we need the underlying density to be well approximated by quartic polynomials on neighborhoods of radius h_{MISE_6}. While this might occur at a few points (e.g. at $x = 0.5$ in Figure 2), it seems very unlikely to happen in enough locations to give the above relationship between the $MISE$'s. This also fits in with the usual asymptotic idea that if the first order asymptotics are not properly descriptive, it is too much to ask that the second order asymptotics help much.

These ideas have been verified by looking at appropriate analogs of Figure 3a, for all of the test densities in [2, Marron and Wand]. The situation $MISE_6 \ll MISE_4 \approx MISE_2$ never occurred for any of these.

6 Practical Application?

In section 4 it was made clear that higher order kernels are better than non-negative ones when the underlying function to be estimated is well approximated by a parabola, on neighborhoods with radius roughly the effective window width of the higher order kernel. Can this principle be used in practice? In other words, given a set of data, and a nonparametric estimate, can one tell whether there will be a significant benefit from using a higher order kernel?

A simple approach would be to construct a plot analogous to Figure 1a (except the true underlying curve can not be present!), showing the estimate and also giving a good idea of the effective window width by showing the kernel function at the bottom. In situations where the estimate has features (e.g. peaks or valleys) of nearly the same width as the kernel, then it seems clear that there can not be much improvement from the use of a higher order kernel. However, when it appears that local parabolas fit well (to the estimate!) in all neighborhoods of a size somewhat larger than the kernel width (since "appropriate bandwidths" for a higher order kernel are typically larger, see [2, Marron and Wand]), then there is the potential for improvement through the use of higher order kernels. Unfortunately one can never be sure about the improvement, because sharp features may be present, but simply smoothed away and not visible in the given estimate.

References

[1] Chu, C. K. and J. S. Marron (1992) Choosing a kernel regression estimator, *Statistical Science*, 6, 404-436.

[2] Marron, J. S. and M. P. Wand (1992) Exact mean integrated squared error, *Annals of Statistics*, 20, 712-736.

[3] Schucany, W. R. and J. P. Sommers (1977) Improvement of kernel-type density estimators. *Journal of the American Statistical Association*, 72, 420-423.

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

[5] Wand, M. P. and W. R. Schucany (1990) Gaussian-based kernels, *Canadian Journal of Statistics*, 18, 197-204.

Figure 1a

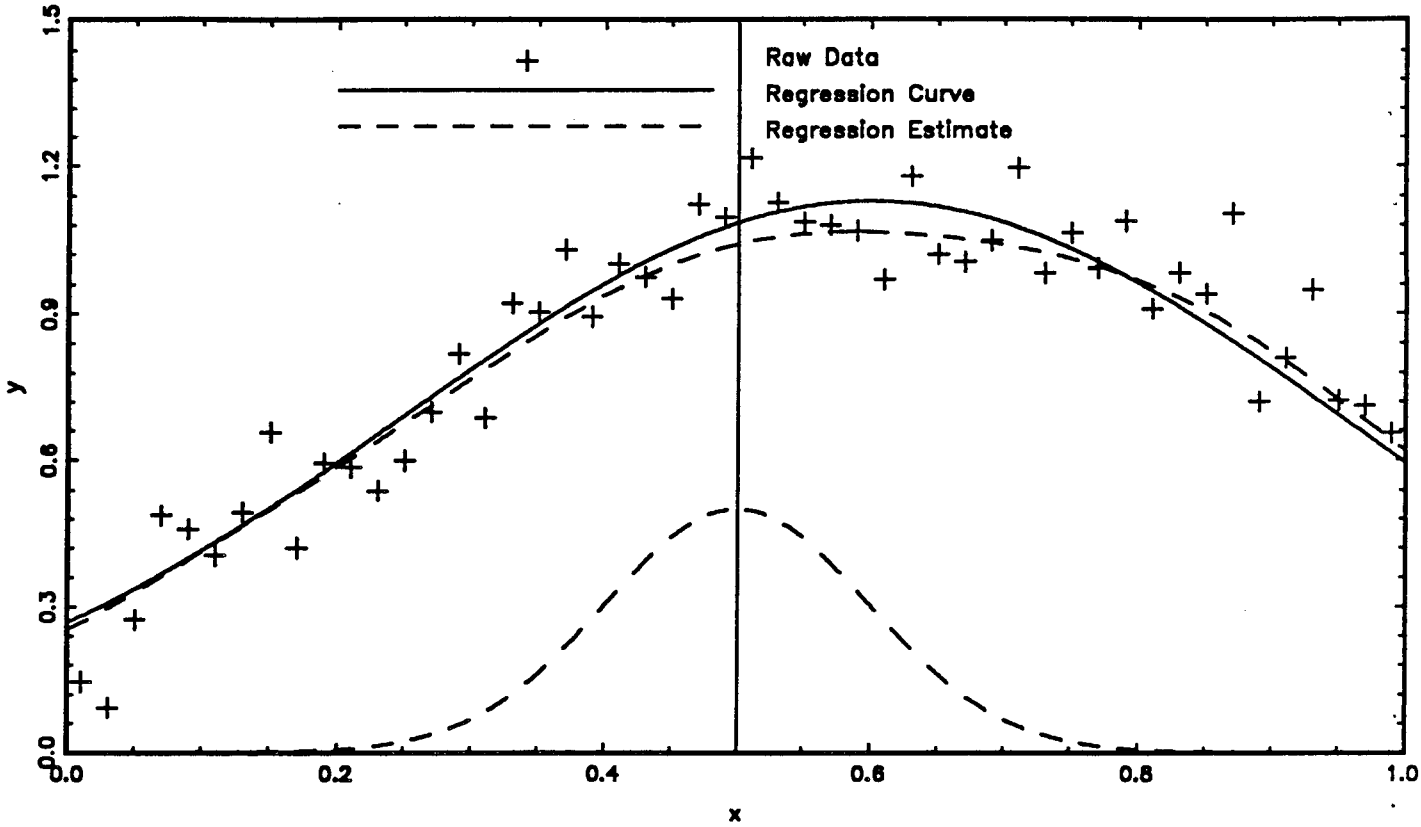

Figure 1b

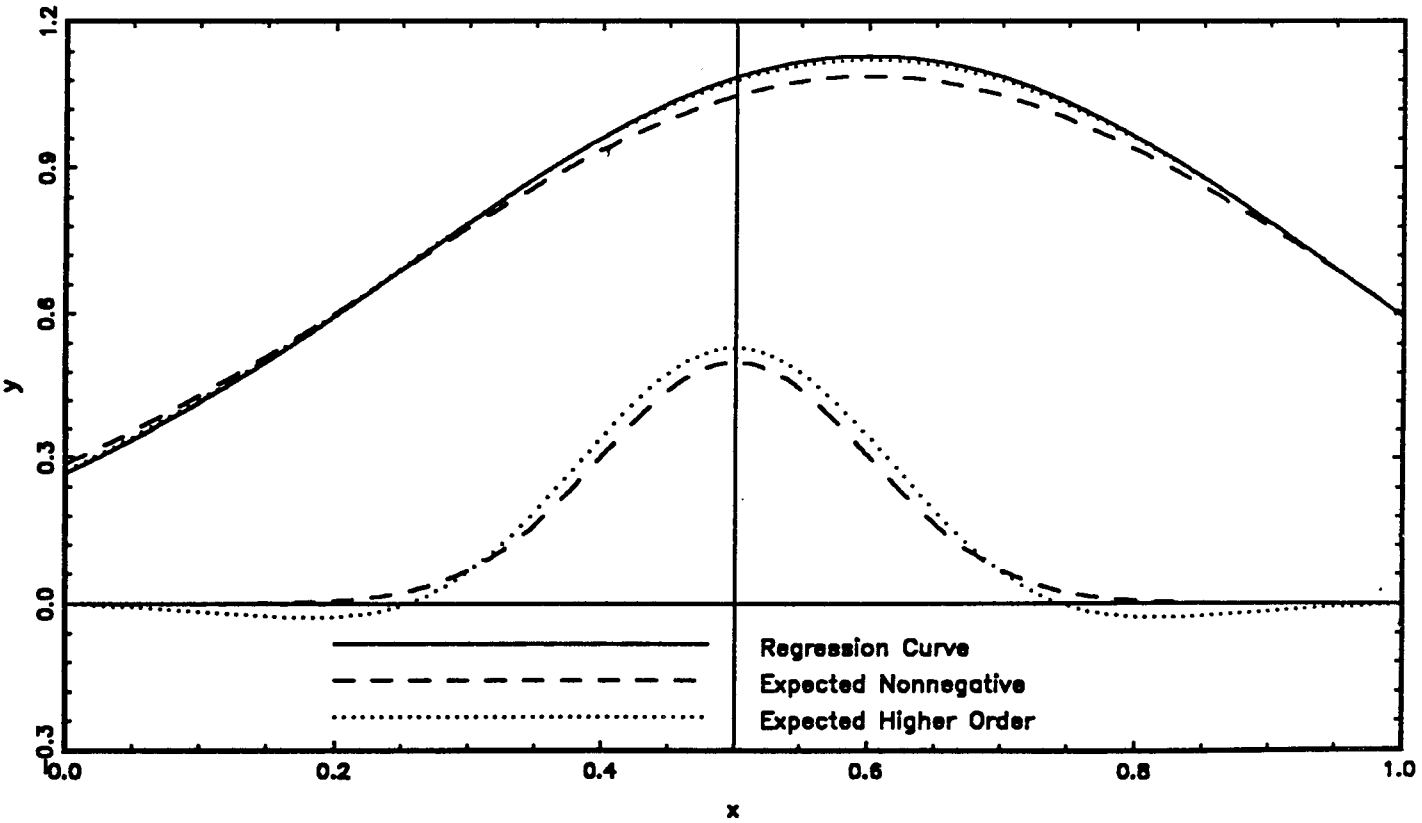

Figure 3a

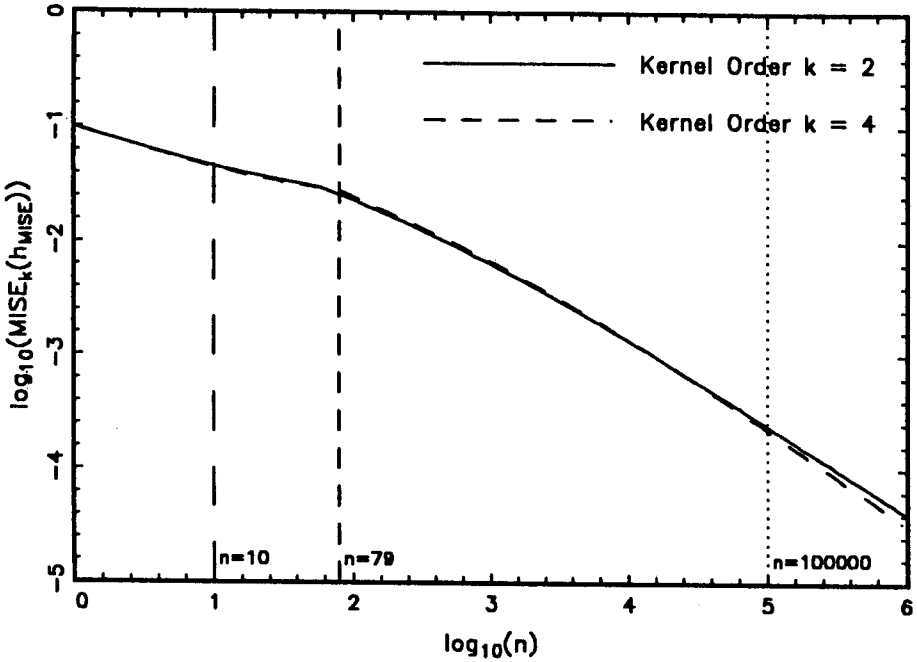

Figure 3b

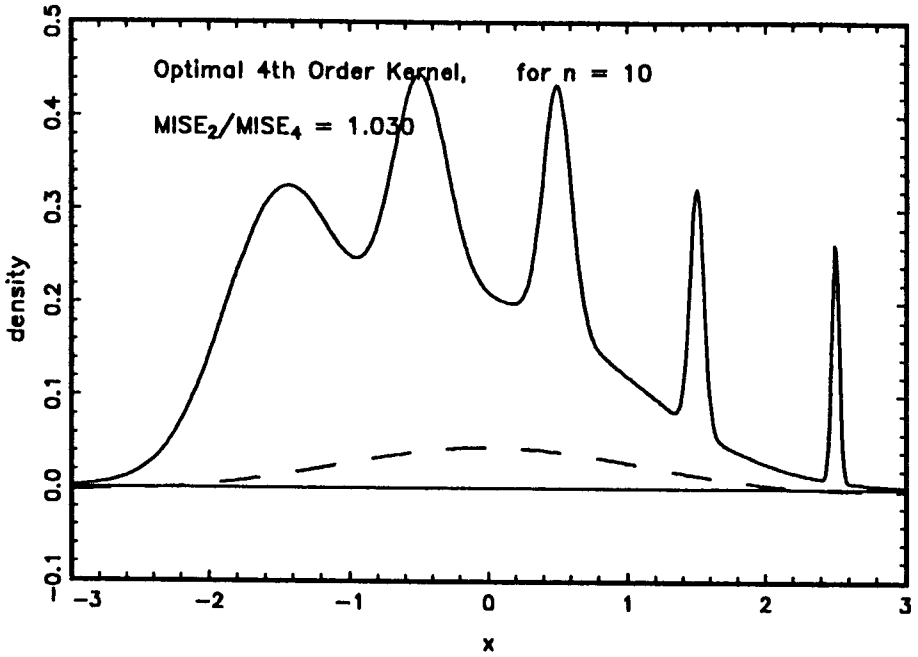

Figure 3c

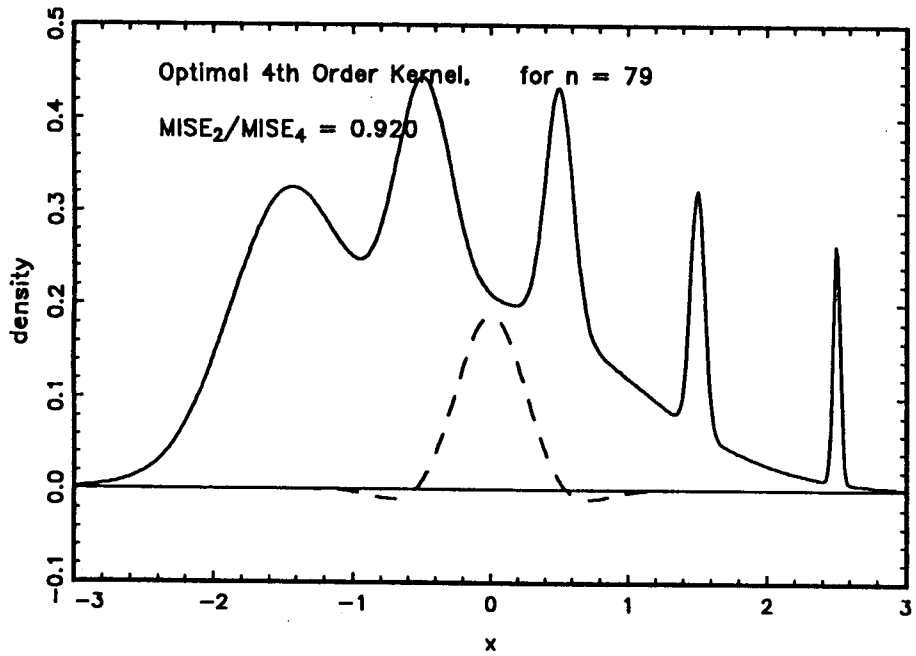

Figure 3d