What does Optimal Bandwidth Selection Mean for Nonparametric Regression Estimation?

by

J. S. Marron

University of North Carolina, Chapel Hill

November 4, 1986

ABSTRACT

Bandwidth selection in nonparametric regression estimation is shown, through an example, to be a smoothing problem. Recent developments concerning optimal bandwidth selection are presented and discussed. The theoretical part of these results include asymptotic optimality, together with a quantification of the rate of convergence to asymptotic optimality. The serious consequences of the very slow rate of convergence are illustrated through the same example. The issue of what "optimal bandwidth selection" means is considered.

Key words and phrases: cross-validation, nonparametric estimation, smoothing parameter.

1980 AMS subject classifications: primary 62G20, secondary 62G05.

Research partially supported by NSF Grant DMS-8400602.

1. Introduction

The underlying ideas in curve estimation by smoothing are easily understood by considering the example of nonparametric regression or scatterplot smoothing. Figure 1a shows a simulated regression setting, where at 75 equally spaced points, zero mean Gaussian noise has been added to a curve. The goal is to try to use the observations to recover the curve.

[Put Figure 1 here]

A simple way to try to do this is to take a moving average (defined precisely in Section 2). An example of this is given in Figure 1b. The dashed curve shows a weighted moving average of the observations where the weights correspond to the curve at the bottom of the picture. An unpleasant feature of this estimate is that it is too wiggly. The reason for this is that too few observations are being used in each average. The Central Limit Theorem says that when an average is taken over too few observations, the resulting value will be relatively far from the mean, sometimes too large and sometimes too small.

A means of decreasing the variance in an average is to increase the number of observations. This is accomplished in the present context by increasing the window width of the local average. Figure 1c demonstrates the beneficial effect of doing this in the current example.

While expanding the window width has a beneficial effect in terms of cutting down sample variance, there is a price to be paid in terms of bias. This is because observations which are far away from the point under consideration have a mean which is substantially different from

the desired value. Figure 1d demonstrates this effect. Note that the resulting estimate is lower at the peak and higher at the valley than it should be (to eliminate boundary problems, a circular design, where the right edge may be thought of as being connected to the left edge, is used here).

This is the essence of the smoothing problem. On one hand, the window width should be large enough to eliminate artifacts of the particular realization of the errors. While on the other hand, the window width should not be so large as to smooth away features of the underlying curve.

The example in Figure 1 makes this problem look substantially easier than it really is. This example was chosen from 100 different realizations of the same setting, because it is the one (out of 100) where the estimation problem is the easiest. Estimation in the case of another example in the same set up is shown in Figure 2. This hints at the magnitude of the problems involved in smoothing, because it is not so clear which of the two estimates is "best". The estimate in Figure 2a is closer to the right answer in terms of squared error, but the estimate in Figure 2b has smoothed away the "bump" on the right side and hence seems to do a better job of capturing the qualitative aspects of the underlying curve.

[Put Figure 2 here]

Figure 3 shows the realization for which the estimation of m is the hardest, out of the 100 available. Note that regardless of the value of the window width, the moving average is never very close to the

underlying curve. This should not be too disappointing, because even for the much simpler problem of estimating say a Gaussian mean, if one looks at 100 realizations, there will certainly be some where all reasonable estimates of the mean will be far from the true mean. An inspection of the scatterplot reveals that this set of data just does not reflect the underlying curve.

[Put Figure 3 here]

2. Mathematical Quantification

A convenient mathematical framework for the analysis of the smoothing problem is the following. Suppose that, for i = 1,...,n,

$$Y_i = m(x_i) + \epsilon_i,$$

where m is called the regression function, and the ϵ_i are independent mean zero errors. If the x_i are equally spaced, then a simple version of the moving average estimator of m(x) is

$$\hat{m}_h(x) = n^{-1} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x-X_i}{h}\right) Y_i,$$

where K is called the "kernel function" and h is called the "bandwidth" or "smoothing parameter". Note that $\hat{m}_h$ is an approximate weighted average when it is assumed that $\int K = 1$ (because then the sum of the weights is a Riemann sum for this integral).

In the particular example shown in Figures 1, 2 and 3, n = 75, $x_i = i/n$, $m(x) = x^3(1-x)^3$, the ϵ_i are $N(0,0.011)^2$, and

$$K(x) = \frac{15}{8}(1-4x^2)^2 1_{[-0.5,0.5]}(x).$$

Smoothing folklore says that the main problem in smoothing is

choice of h. The effect of h on the resulting estimate is very apparent from Figures 1, 2 and 3. It is easy to see from the pictures that the choice of the function K is far less crucial, so that will not be discussed here.

The traditional approach to the bandwidth problem is to quantify it by considering an error criterion such as

$$ASE(h) = n^{-1} \sum_{i=1}^{n} [\hat{m}_h(x_i) - m(x_i)]^2,$$

or

$$MASE(h) = E(ASE(h)).$$

In this setup, two candidates for "optimal bandwidth" are $\hat{h}_0$ and h_0, the minimizers of ASE and MASE respectively. An asymptotic analysis of MASE (see, for example, Rosenblatt 1969), yields simple representations for the variance and bias terms. This representation provides an elegant quantification of the smoothing problem because the variance term dominates when h is too small (ie. there is too much sample noise as in Figure 1b), and the bias term dominates for h too big (ie. features of the underlying curve are smoothed away as in Figure 1d).

While this type of quantification does a lot to aid in understanding the smoothing problem, it is not very useful for practical choice of h because the resulting answer is a function of things that are harder to estimate than the curve m itself (things like the second derivative of m enter in).

3. Automatic bandwidth selection

Most data driven bandwidth selectors use the fact that the regression function m(x) is the best predictor, in the mean square sense, of a new Y value corresponding to a given x value. The first way in which one might try to use this is by choosing h to minimize the prediction error,

$$p(h) = n^{-1} \sum_{i=1}^{n} [Y_i - \hat{m}_h(x_i)]^2.$$

This method has the the problem that p(h) has a trivial minimum at the "no smoothing point" (the value of h where $\hat{m}_h(x_i) = Y_i$). Furthermore, p(h) has a bias towards h too small along the entire range of possible h's, which can be seen either analytically by taking expected values, or intuitively by observing that p(h) is using the same set of data both to construct the estimator and to assess it.

A means of overcoming this problem, which is based on the intuitive explanation of it, is the idea of cross-validation, where one takes $\hat{h}$ to be the minimizer of

$$CV(h) = n^{-1} \sum_{i=1}^{n} [Y_i - \hat{m}_{h,i-}(x_i)]^2,$$

where $\hat{m}_{h,i-}$ is a version of $\hat{m}_h$ which does not make use of Y_i (see Priestley and Chao 1972 for a method of doing this).
This method of smoothing parameter selection was first proposed by Clarke (1975) and Wahba and Wold (1975).

A number of recent papers, see for example Rice (1984), Härdle and Marron (1985a,b) and Burman and Chen (1985) have established some fairly promising theoretical results about CV and $\hat{h}$ which can be grouped into a category called "asymptotic optimality". These typically are of the

form, under some sort of regularity condition,

$$\frac{ASE(\hat{h})}{ASE(h_0)} \rightarrow 1, \tag{3.1}$$

or

$$\hat{h}/\hat{h}_0 \rightarrow 1, \tag{3.2}$$

in some mode of convergence.

It is important to keep in mind that results of the form (3.1) and (3.2) are asymptotic in character. A question which should always be asked about such results is: what do they mean for the particular data set at hand? In particular, have the effects described by the asymptotics "taken over" for, say a sample of size 75?

One means of addressing this issue is to look at some examples. Figure 4 shows another realization from the same 100 simulations that Figures 1, 2, and 3 came from. The dashed and dotted curves show the kernel estimators with window widths h = .25 (which was $\hat{h}$ the minimizer of CV(h)), and h = .68 (which was h_0 the minimizer of ASE(h)), respectively. While results of the type (3.1) and (3.2) say that $\hat{h}$ and $\hat{h}_0$, and also the resulting curves, should be approximately the same, it is readily apparent that they are not. Of course this example was one of the worst of 100 (although there are a few even worse), but most of the realizations exhibited substantial behavior that is not in keeping with one might expect from (3.1) and (3.2).

[Put Figure 4 here]

So we have at least one example of a setting where the asymptotic results are doing a very poor job of describing what actually happens.

Of course we could try to understand this further by looking at some more simulated examples, but simulation studies have the drawback that their results are always dependent on the particular example being studied, so no general conclusions can be drawn.

4. Rates of Convergence to Asymptotic Optimality

A method of understanding the discrepancy between the theory and the simulations discussed above is to compute the rates of convergence in (3.1) and (3.2). Härdle, Hall, and Marron (1987) (see Rice 1984 for a related result) have shown that, essentially under the assumptions that K is nonnegative and m'' is continuous,

$$(4.1) \qquad n^{1/10}\left(\frac{\hat{h} - \hat{h}_0}{h_0}\right) \sim N(0,\sigma^2),$$

$$(4.2) \qquad n[ASE(\hat{h}) - ASE(\hat{h}_0)] \sim C\cdot\chi^2_1,$$

in distribution, as $n \to \infty$. Note that the relative rate of convergence $n^{-1/10}$ is excruciatingly slow. This says that, looking across different samples from the same setup, there will typically be a great deal of difference between $\hat{h}$ and $\hat{h}_0$.

Results (4.1) and (4.2) make $\hat{h}$ look like a pretty poor choice of bandwidth. However, there is a pair of companion results which makes it look much better:

$$(4.3) \qquad n^{1/10}\left(\frac{h_0 - \hat{h}_0}{h_0}\right) \sim N(0,\sigma_0^2),$$

$$(4.4) \qquad n[ASE(h_0) - ASE(\hat{h}_0)] \sim C_0\cdot\chi^2_1,$$

Observe that, although $\hat{h}$ and $\hat{h}_0$ may be expected to be a long way apart,

the difference between the two notions of "optimal bandwidth", $\hat{h}_0$ and h_0, may also be expected to be of the same large order.

It is important to note that results (4.1) - (4.4) are also asymptotic in character, so we should again worry about whether the effects that they describe have taken over for reasonable sample sizes. The simplest thing one should expect from (4.1) is that sometimes $\hat{h}$ will be relatively large and $\hat{h}_0$ will be relatively small, while at other times the opposite should occur. This phenomenon may be seen in Figures 4 and 5. Recall Figure 4 shows the kernel estimates as dashed and dotted curves with $\hat{h}_0 = .28$, and $\hat{h} = .7$ respectively, for one realization of the data. Figure 5 shows $\hat{h} = .2$ as the dotted curve and $\hat{h}_0 = .8$ as the dashed curve, for another realization. Summary statistics which show that the limiting distributions fit quite well in the present example are given in Härdle, Hall, and Marron (1987).

[Put Figure 5 here]

While the rate of convergence given in (4.1) is terribly slow, the fact that this same rate appears in (4.3) makes one feel that it might be the best possible rate. It is conjectured that this is the case, in a minimax sense analogous to that used by Farrell (1972), Stone (1980, 1982) and a number of others. This can be formulated as: given any bandwidth selector $\hat{h}^*$ (ie. <u>any</u> function of the data),

$$\lim_{\epsilon \to 0} \liminf_{n \to \infty} \sup_{f \in \mathcal{F}} P_f\left[\left|\frac{\hat{h} - \hat{h}_0}{h_0}\right| > \epsilon n^{-1/10}\right] = 1, \tag{4.5}$$

where

$$\mathcal{F} = \{\text{densities } f: |f''(x)| \leq B\},$$

for a constant B. For an indication of how such a result could be established, see the related work in density estimation by Hall and Marron (1987).

For some interesting extensions to the cases of higher order kernels and to higher dimensions, along with additional discussion, see Marron (1986).

5. Some Conclusions

The first consequence of (4.1) - (4.4) is that cross-validation is optimizing the error criteria ASE and MASE only in a very loose sense. However this should not be taken to be a weak point of cross-validation, because (4.5) indicates in a strong sense that nothing else will be able to do a much better job of optimizing at least ASE.

There are two possible ways to interpret this fact. The first is to forget about trying to estimate curves by smoothing, because there is no way to effectively optimize ASE. This is not reasonable because there has been and will continue to be much effective data analysis done by smoothing (see for example the collection of examples in Chapter 6 of Silverman 1986). The second interpretation of these results is that we should no longer think of smoothing in terms of trying to optimize error criteria such as ASE and MASE.

It is a small step from not thinking of smoothing in terms of optimizing ASE and MASE, to no longer thinking about using smoothing to recover the regression curve m(x). Then how should we think about smoothing methods? It seems we should think instead about discovering

attributes of the particular set of data at hand. Note that sometimes this will mean recovering the original curve, recall Figure 1c, while sometimes not, recall Figure 3.

If we accept the ideas in the above paragraph, then the fact that cross-validation is not really optimizing error criteria does not matter. But the question which is then vital to the continued use of cross-validation is: what is it doing?

6. What does Cross-Validation "feel"?

Careful consideration of the 100 realizations in the example discussed above have given rise to the following heuristic idea. First note that choosing a bandwidth is the same as choosing a curve from the family of estimates indexed by h. The curve chosen by cross-validation seems to be one for which there is no more detail than one can expect to see with that amount of smoothing. But once a curve has no more detail than could be expected, all smoother ones will share that property. What is interesting about the one chosen by cross-validation is that it seems to correspond to the smallest bandwidth for which this is the case.

To help understand this consider Figure 6. Note that the underlying curve is no longer shown here because we are now not thinking of recovering this curve, but instead about what makes cross-validation choose a given curve from the family of possibilities. Figure 6a shows the estimate with h = .01, for a realization where this bandwidth was selected by cross-validation. Figures 6b, c and d show three other

realizations, where cross-validation chose curves which are much smoother. All four curves have roughly the same amount of oscillation, which is to be expected with this amount of smoothing, but what is interesting about Figure 6a, is that it oscillates in a very smooth, regular fashion, while the other three curves are very ragged. Figure 6a has only details which can be resolved through this much smoothing while the others have more detail (ie. "rough edges") than one would expect to see through a window of this width.

[Put Figure 6 here]

Another example of this is given in Figures 7a and b, where estimates from the same realization are shown for the bandwidths h = .28 and .7 respectively. Figure 7b was the curve chosen by cross-validation. Observe that this choice does not look nearly as bad now as it did in Figure 5 (Figure 7 shows the same realization and estimators as shown there). On the other hand Figure 7a, which shows the minimizer of ASE looks substantially less believeable, because it shows features one can not expect to resolve through a window of the given width.

[Put Figure 7 here]

To see what is meant by the <u>smallest</u> bandwidth where the curve has no more detail than one can expect to recover, consider Figures 8a and b, which show another realization, with $\hat{h} = .55$ and $\hat{h}_0 = .8$ respectively. Both curves have the same amount of detail, so it seems quite reasonable, in the absence of other information (such as the fact that the second curve minimizes ASE), to prefer the first curve.

[Put Figure 8 here]

The realizations in the pictures above were chosen because, out of the 100 available, they most clearly make the points intended. Figures 9 and 10 show the realizations which seem to cast the most doubt on the idea presented at the beginning of this section.

[Put Figures 9 and 10 here]

Figures 9a and b show estimates with $h = .53$ and $\hat{h} = .56$ respectively. It takes some careful examination to find detail that should not be there in Figure 9a. Figure 9c shows an overlay of these two curves, which reveals a very slight "bump" on the lower right hand slope of Figure 9a, which could perhaps be considered to be more detail than could be resolved through a window of this width (keep in mind that much greater sensitivity is required at larger bandwidths.).

Figure 10a, b and c show another realization with $h = .31$, .52 and .77 respectively. You may want to test your understanding of the idea presented in this section by guessing which of the three curves was chosen by cross-validation. The hardest choice is probably between curves b and c so these two are overlaid in Figure 10d. The answer is given at the end of the list of references.

7. Final Hopes

First it is hoped that this idea, or (more realistically) something similar, is correct. If so, the usefulness of cross-validation as a data analytic tool will be greatly enhanced. In fact, if this is right, a case could be made for the bandwidth chosen by cross-validation as

being more reasonable than the minimizer of error criteria such as ASE or MASE.

The second hope is that some version of this idea can be quantified. It is fun to look at pictures, but no firm conclusions can be reached until it is understood, on a mathematical level what is happening in cross-validation. Also if some quantification can be found, the usefulness of cross-validation should be even more enhanced, and means of improving it will probably become apparent.

References

Burman, P. and Chen, K. W. (1985). "Nonparametric estimation of a regression function," unpublished manuscript.

Clark, R. M. (1975), "A calibration curve for radio carbon dates," *Antiquity*, 49, 251-266.

Farrell, R. H. (1972), "On the best obtainable rates of convergence in estimation of a density function at a point," *Annals of Mathematical Statistics*, 43, 170-180.

Härdle, W., Hall, P. and Marron, J. S. (1987), "How far are automatically chosen regression smoothers from their optimum?," to appear with discussion in *Journal of the American Statistical Association*.

Härdle, W. and Marron, J. S. (1985a), "Optimal bandwidth selection in nonparametric regression function estimation," *Annals of Statistics*, 12, 1465-1481.

Härdle, W. and Marron, J. S. (1985b), "Asymptotic nonequivalence of some bandwidth selectors in nonparametric regression," *Biometrika*, 72, 481-484.

Hall, P. and Marron, J. S. (1987), "The amount of noise inherent in bandwidth selection for a kernel density estimator, " *Annals of Statistics*, to appear.

Marron. J. S. (1986), "Will the art of smoothing ever become a science?", *Function estimates*, (J. S. Marron, ed.) American Mathematical Society Series: Contemporary Mathematics, 9, 169-178.

Priestley, M. B. and Chao, M. T. (1972), "Non-parametric function fitting," *Journal of the Royal Statistical Society*, series B, 34, 385-392.

Rice, J. (1984), "Bandwidth choice for nonparametric regression," *Annals of Statistics*, 12, 1215-1230.

Rosenblatt, M. (1969), "Conditional probability density and regression estimates", in *Multivariate Analysis*, P. R. Krishaiah, ed., 25-31.

Silverman, B. W. (1986), *Density estimation for statistics and data analysis*, Chapman and Hall.

Stone, C. J. (1980), "Optimal convergence rates for nonparametric

estimators," *Annals of Statistics*, 8, 1348-1360.

Stone, C. J. (1982), "Optimal global rates of convergence of nonparametric regression," *Annals of Statistics*, 10, 1040-1053.

Wahba, G. and Wold, S. (1975), "A completely automatic french curve: fitting spline functions by cross-validation," *Communications in Statistics*, 4, 1-17.

Figure 10: Curve c was chosen by cross-validation.

Captions

Figure 1: Data set number 84, scatterplots together with regression function as solid curves, and with moving average estimators as dashed curves (relative weights shown at bottom) with window widths, (b) h = .2, (c) h = .51, (d) h = .9.

Figure 2: Data set number 34, scatterplots together with regression function as solid curves, and with moving average estimators as dashed curves (relative weights shown at bottom) with window widths, (a) h = .44, (c) h = .64.

Figure 3: Data set number 58, scatterplots together with regression function as solid curves, and with moving average estimators as dashed curves (relative weights shown at bottom) with window widths, (a) h = .2, (b) h = .35, (c) h = .66.

Figure 4: Data set number 68, regression function as solid curve, together with kernel estimators with bandwidths $\hat{h}$ = .25 (dashed curve) and $\hat{h}_0$ = .68 (dotted curve).

Figure 5: Data set number 10, regression function as solid curve, together with kernel estimators with bandwidths $\hat{h}_0$ = .28 (dashed curve) and $\hat{h}$ = .7 (dotted curve).

Figure 6: Kernel estimators with bandwidths all h = .1 for the datasets (a) 1, (b) 2, (c) 3 and (d) 4.

Figure 7: Data set number 68, kernel estimators with bandwidths (a) $\hat{h}$ = .25 and (b) $\hat{h}_0$ = .7.

Figure 8: Data set number 61, kernel estimators with bandwidths (a) $\hat{h}$ = .55 and (b) $\hat{h}_0$ = .8.

Figure 9: Data set number 51, kernel estimators with bandwidths (a) h = .53, (b) h = .66 and (c) overlay of a and b.

Figure 10: Data set number 71, kernel estimators with bandwidths (a) h = .31, (b) h = .52, (c) h = .77 and (d) overlay of b and c.

Figure 1a

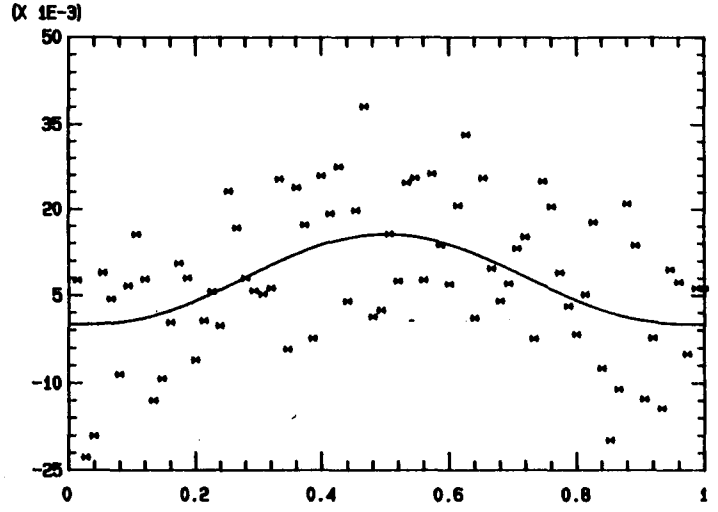


Figure 1b

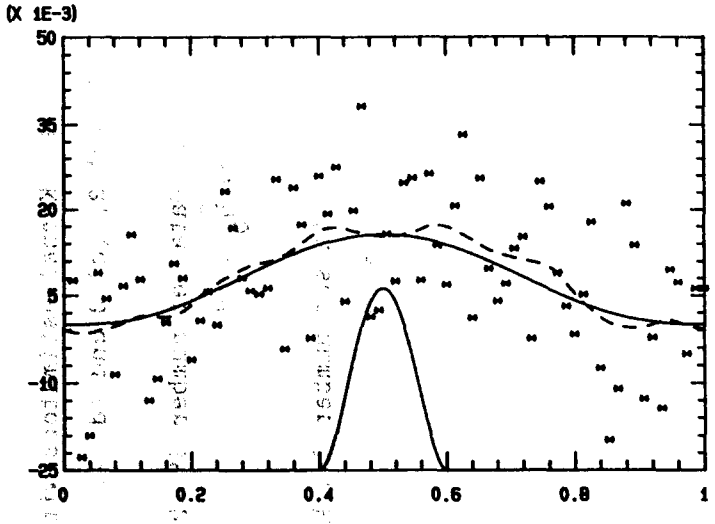


Figure 1c

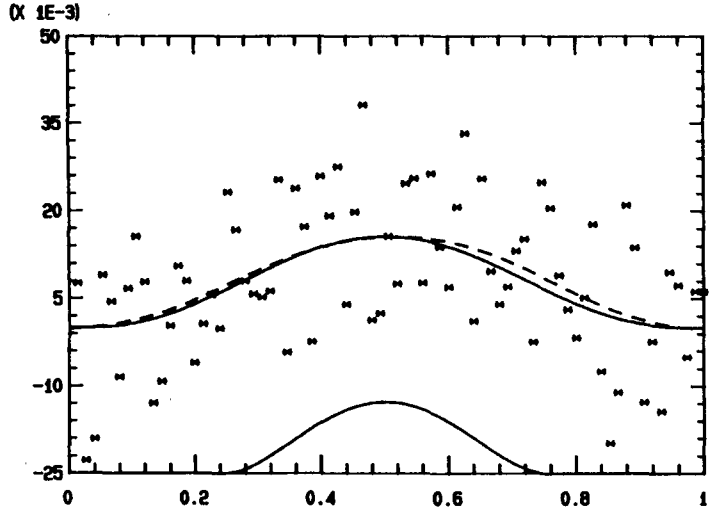


Figure 1d

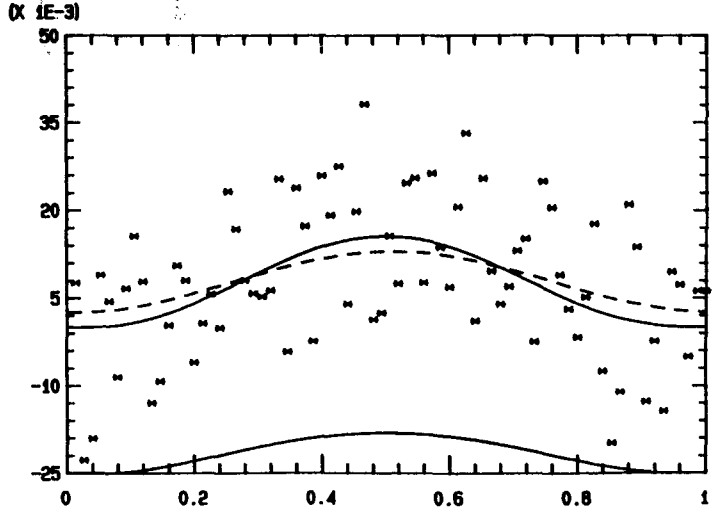


Figure 2a

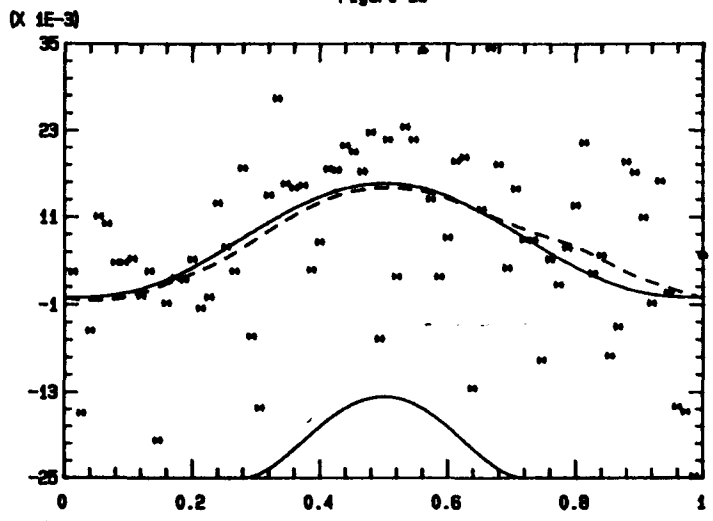
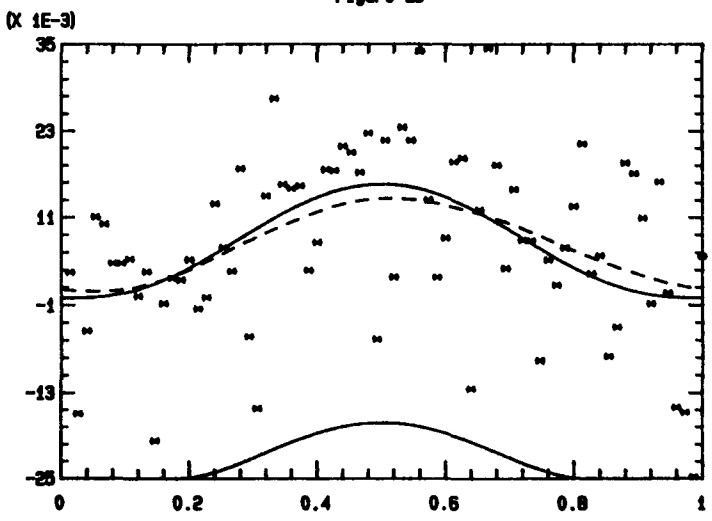


Figure 2b



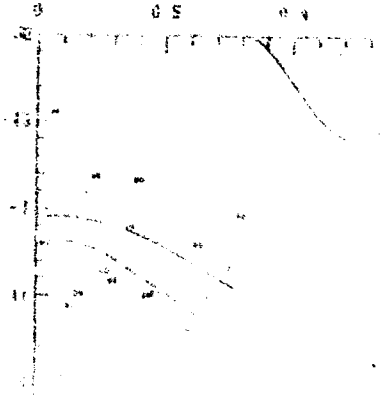


Figure 3a

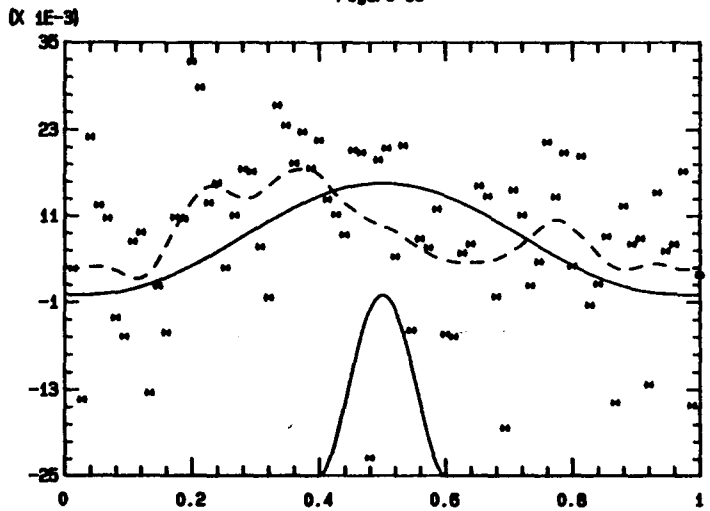


Figure 3b

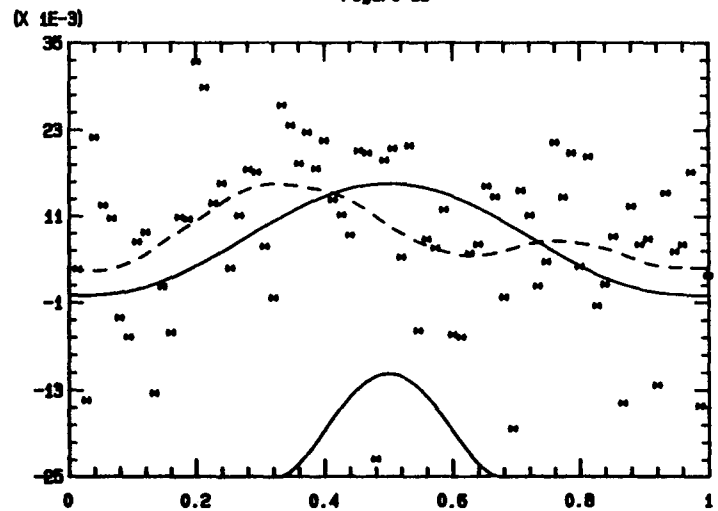


Figure 3c

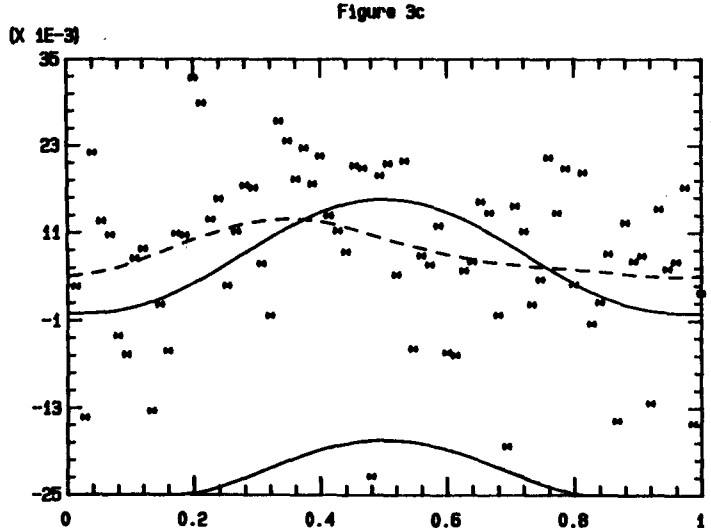


Figure 7a

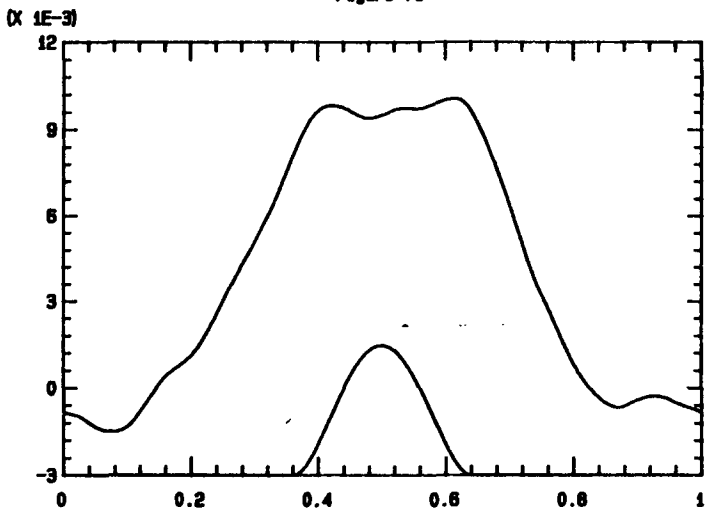


Figure 7b

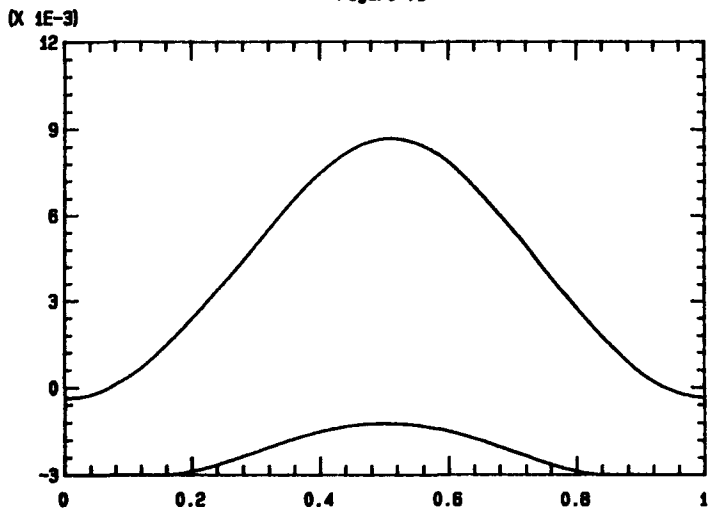


Figure 8a

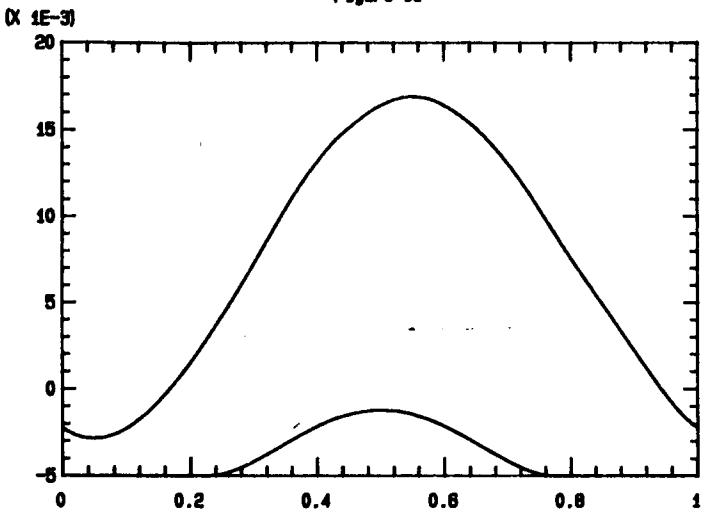


Figure 8b

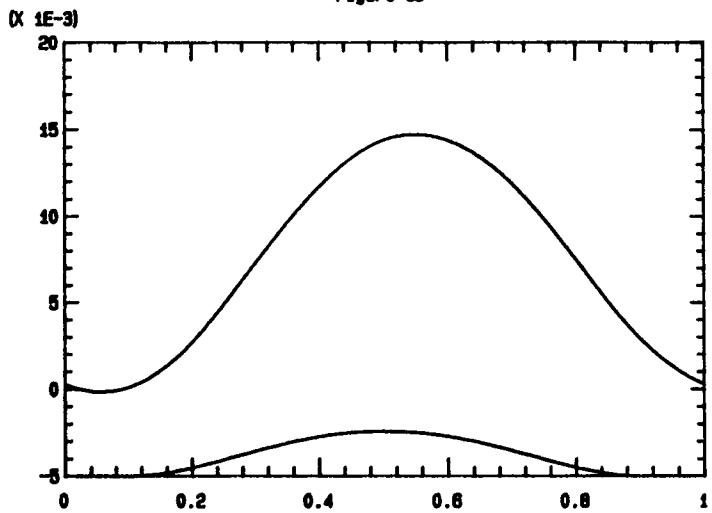


Figure 10a

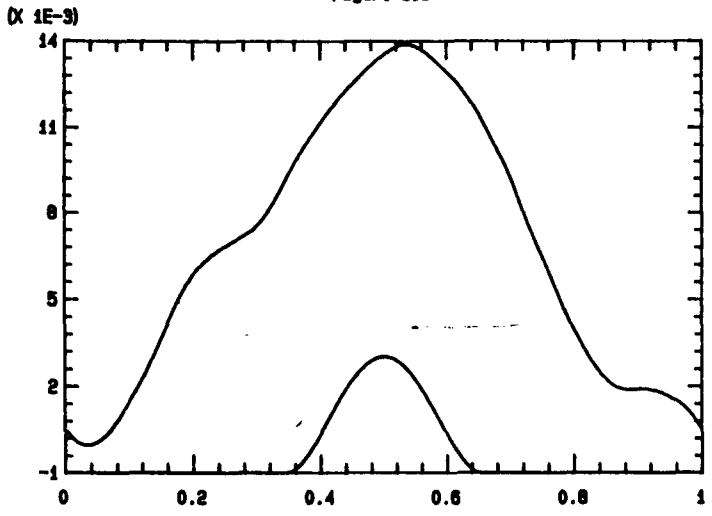


Figure 10b

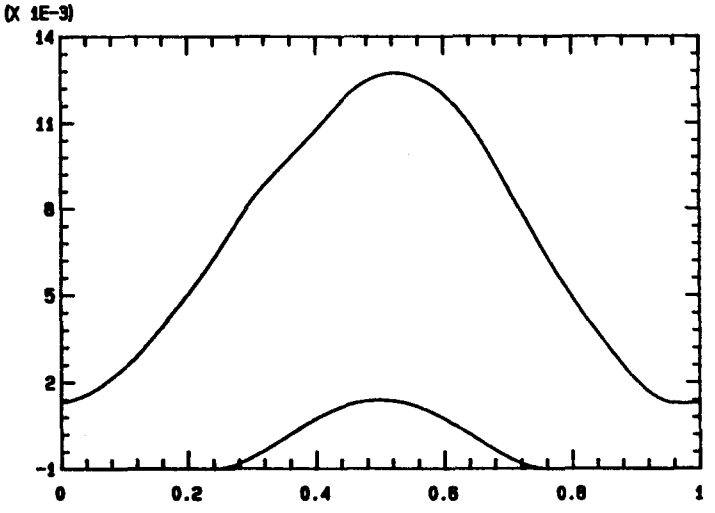


Figure 10c

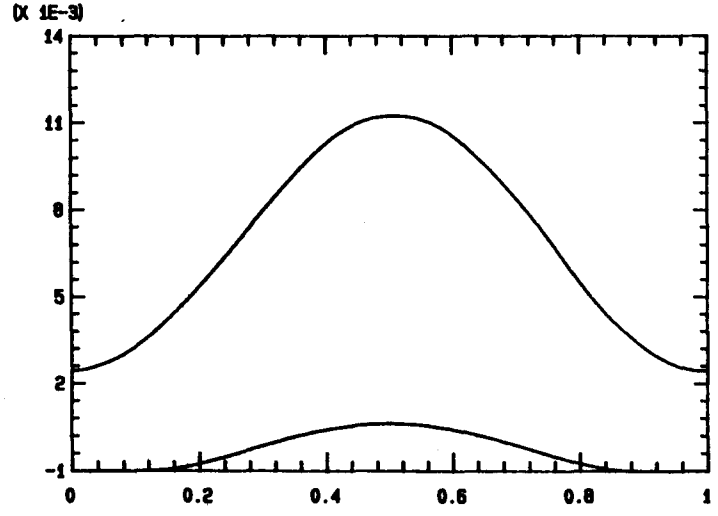


Figure 10d