SEQUENTIAL DESIGN FOR THE NONPARAMETRIC REGRESSION OF CURVES AND SURFACES.

Julian J. Faraway
Department of Statistics, University of Michigan.

Appearing in *Computing Science and Statistics* **22**, 104-110.

ABSTRACT

The problem of sequentially selecting the design points for a nonparametric regression is considered. It is desired to estimate some unknown function over some known region. Observations may be taken at a design point chosen on the basis of past observations. An adaptive method is described which both selects the local amount of smoothing required for the regression estimate and proposes the location of the next best design point. The method developed has a wide application in the estimation of surfaces where no parametric model is known or appropriate.

1. INTRODUCTION

Consider the problem of estimating some unknown function $r(x)$ for $x \in \Omega \subset R^k$. We may take measurements at points $x_1,\ x_2, ...$, and observe $Y_1,\ Y_2, ...$, where

$$Y_i \ = \ r(x_i) \ + \ e_i \, for \ \ i = 1, 2, ... \ ,$$

where e_i are mean zero, independent errors. This is a problem of very wide interest. Response surface methodology tackles it by parameterizing r in some way and then proceeds to the related design and estimation problems. See Box & Draper(1987) or Ripley (1981) for an introduction to these techniques.

In some cases, we may know or have a good idea about the functional form of r and then a parametric method would apply. However, we may know little about r other than that it is reasonably smooth. In this case, a nonparametric method of estimating r may be appropriate.

Sometimes, we have a fixed sample size available and must decide on the location of all the design points
$x_1,\ x_2, ...,$ in advance . For the parametric case this problem has been considered in detail. See Draper N. (1988) for an introduction and Myers R., Khuri A. & Carter Jr. W. (1989) for a review and related references. Much less has been done for the nonparametric case. If one uses a kernel-based estimate of r then Müller (1984) describes the asymptotically optimal design density in the one-dimensional case.

However, it is often possible to observe the results of the measurements sequentially so that we may decide on the position of the next design point on the basis of the previous observations. The advantage is that a significantly greater precision may be had for the same sample size or fewer measurements may be required to obtain some specified accuracy.

We shall use the kernel-based method of nonparametric regression. Müller(1984) paper indicates, that for fixed kernel estimation (the same bandwidth is used globally), that at least asymptotically, the optimal density for the design points is uniform for any $r(x)$ (assuming i.i.d. errors and ignoring edge effects). However, Müller also shows that if local smoothing is used, then asymptotically, the optimal design density is some function of the roughness of $r(x)$. This indicates that, if there is to be any hope of the sequential design outperforming the evenly spaced design in the long run, local smoothing must be employed.

We utilize an adaptive procedure for selecting the local bandwidths, which is based on a paper by Müller and Stadtmüller (1987). At each stage this procedure can be used to estimate r and hence the asymptotically optimal design density. The next design point is chosen so as to make the density of the actual design points as close as possible to the estimated optimal design density. See also preliminary work in Faraway & Rothman (1989).

In sections 2 & 3, we describe the method and a simulation for the one-dimensional case. In sections 4 & 5, we do the same for higher dimensions and in section 6 we discuss some extensions to the ideas hitherto presented.

2. METHOD FOR THE ESTIMATION OF CURVES

We will describe the methods for the simplest one-dimensional case and indicate the extensions later.

Let $x_1, ..., x_n \in [0, 1]$ be design points and let $e_1, ..., e_n$ be independent and identically distributed with unknown distribution F that has zero mean and variance $\sigma^2 < \infty$. We observe $Y_1, ..., Y_n$ where

$$Y_i \ = \ r(x_i) \ + \ e_i \ \ for \ \ \ i = 1, ..., n.$$

Let us assume that r is twice continuously differentiable and has periodicity 1 (enabling us to ignore edge effects for the time being). Consider variable kernel regression estimates of the form

$$\hat{r}_n(x) \ = \ \sum_{i=1}^{n} h(x)^{-1} w_i K((x - x_i)/h(x)) Y_i$$

where K is a positive symmetric kernel function, $h(x)$ is the bandwidth at x and w_i is a weight corresponding to the density of the design points at x_i. For evenly spaced design points $w_i = 1/n$, but in other cases we used $w_i = (x_{(i+1)} - x_{(i-1)})/2$. This is a sensible modification of the Priestley-Chao style estimator for unevenly spaced design points. Let $MSE(x) \ = \ (\hat{r}_n(x) \ - \ r(x))^2$ and $IMSE \ = \ \int_0^1 MSE(x)dx$. Our objective is to minimize the IMSE. Appropriate selection of $h(x)$ is essential to the estimation of $r(x)$.

Müller and Stadtmüller's (1987) method of selecting $h(x)$ can be summarized as follows:

1) Use the Mallow criterion to get a global bandwidth $\hat{h}$. The IMSE for given h is estimated by

$$n^{-1} \sum_{i=1}^{n} (Y_i - \hat{r}_n)^2 - \hat{\sigma}^2 + \frac{2\hat{\sigma}^2}{n\hat{h}} K(0)$$

where $\hat{\sigma}^2$ is any consistent estimator of σ^2. Minimizing this over h gives a good global bandwidth choice. See Rice (1984) for details.

2) Estimate r'' using another, smoother, kernel.

3) Obtain the local bandwidths from

$$\hat{h}(x) \ = \ \hat{h} \left[\frac{\int_0^1 \hat{r}\prime\prime(t)^2 dt}{\hat{r}\prime\prime(x)^2} \right]^{1/5}$$

This formulation is based on the asymptotic expansions of the IMSE and the MSE. Some trimming of $\hat{h}(x)$ will be necessary where $\hat{r}\prime\prime(x)$ is small or zero.

The optimal design is known to be proportional to $r\prime\prime^{2/9}$ from Müller(1984). We use the following the sequential design algorithm to make the actual design density correspond as closely as possible to the estimated (asymptotically) optimal design density.

1) Estimate the optimal design density as $\hat{r}''^{2/9}$, appropriately normalized.

2) Calculate the quantiles of this density as
$\tilde{x}_1, ..., \tilde{x}_{n+1}$, where $\tilde{x}_{n+1} = 1$.

3) Choose the design point x_{n+1} to minimize
$\sum_{i=1}^{n+1}(x_{(i)} - \tilde{x}_i)^2$ where $x_{(i)}$ indicates that the x_i's have been sorted.

The adaptive bandwidth selection method will not work for very small sample sizes since it is difficult to estimate the second derivative. Therefore, it is necessary to place some number of equally spaced design points n_0 before the procedure may be initiated. Choice of n_0 is not crucial, but 20 seemed to work well enough in the simulations.

The reader may have some apprehension as to the time required to calculate the optimal position of the next design point. If this were too slow, it would limit the practical application of the procedure in time-critical situations. However, it takes only a few seconds on a SUN 3/60 in the simulations described below.

3. SIMULATION FOR CURVES

A simulation was carried out in order to ascertain the potential practicability of the method. We used the following test functions:

i) $r(x) = 4\sin(2\pi x)$.

ii) $r(x) = \sin(4\pi x)$.

iii) $r(x) = N(0.25, 0.02) + N(0.75, 0.02)$ where N(.,.) is the standard normal density.

iv) $r(x) = \begin{cases} 20(0.1 - |0.5 - x|) & \text{if } |0.5 - x| < 0.1 \\ 0 & \text{otherwise} \end{cases}$.

The first three functions exhibit varying amounts of roughness and the fourth violates the smoothness assumptions of the method. We compare the sequential design to the fixed design where the points are evenly spaced and to the optimal design where the points are chosen to have density proportional to $r''^{2/9}$. $r(x)$ must be evaluated on a grid of points which should be as fine as may be afforded. We use 200 points. We choose $\sigma = 0.1(max r(x) - min r(x))$ and make 400 replications. We use the Epanechnikov kernel for K. We start with a sample size of 20 and follow the procedure up to a sample size of 100. Note that iv) violates the assumptions made for the procedure and so it is not possible to calculate the optimal design density. In figure 1, the average IMSE is plotted against sample size. The evenly spaced design is given by the dotted line, the sequential design by the solid line and the optimal design by the dashed line. The standard errors are estimated to be at most 3% and are not significant factor in interpreting the results.

Note that in the first three cases, the sequential design outperforms the evenly spaced design. As can be seen, the evenly spaced design requires a substantially larger sample to attain the same degree

Figure 1

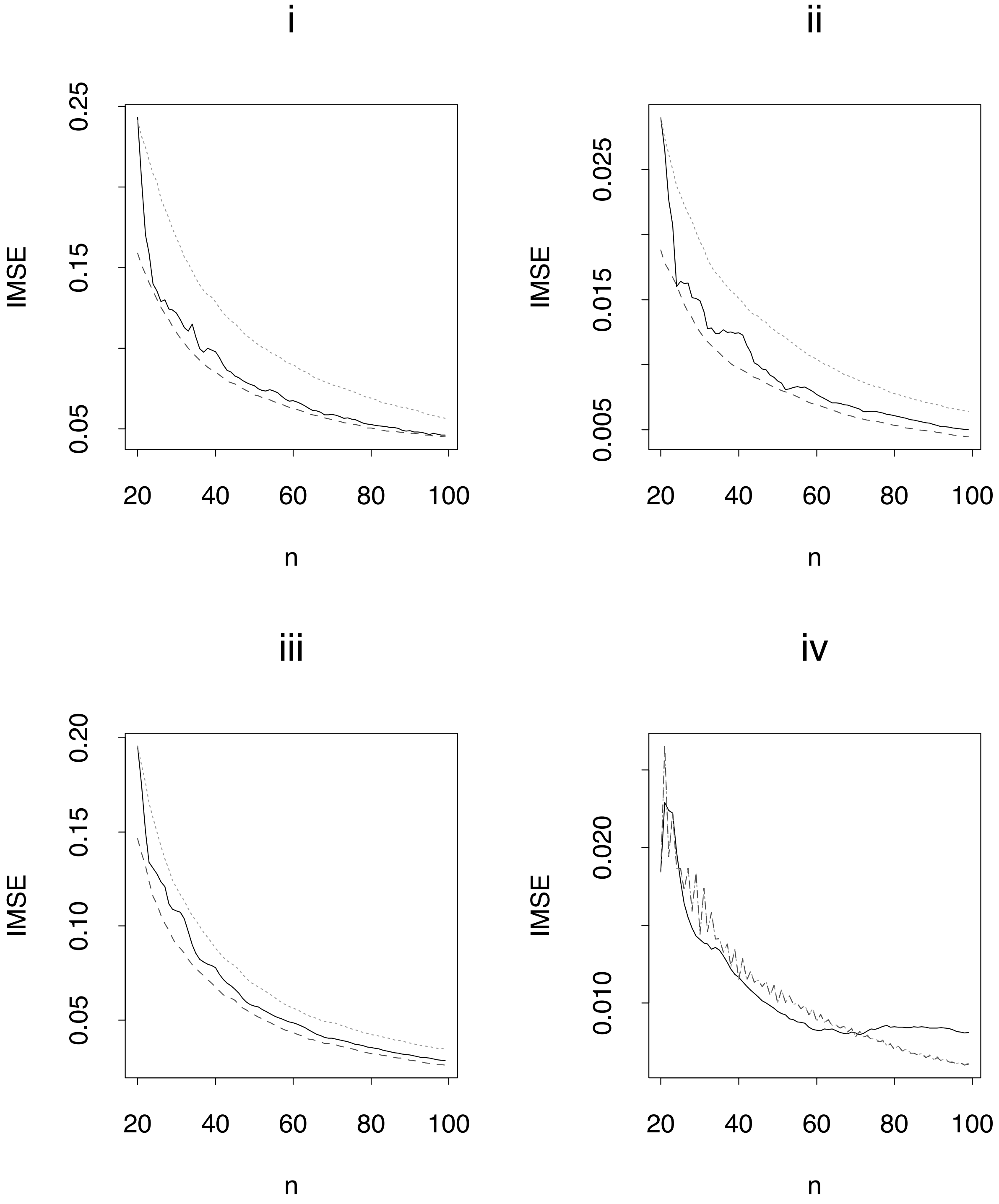


Figure 1: Curve simulation results: Evenly space is dotted, sequential is solid and optimal is dashed

of accuracy as the sequential design. Furthermore, the sequential design becomes increasingly closer to the optimal design as the sample size increases, thus validating the sequential design algorithm we have proposed. In the first three cases, they become close very quickly indicating that the use of the sequential design procedure may improve the accuracy of a previously evenly spaced design with just a few extra observations. Case iv) was included to test the procedure when the assumptions are violated. As can be seen, no loss (or gain) is incurred as a result of using the sequential procedure.

4. METHOD FOR SURFACES

The sequentially-based estimation of curves is instructive but most applications will be in the estimation of surfaces, especially in two dimensions. This involves essentially the same idea, but with some complications introduced by higher dimensions. Consider estimation of $r(x)$ over $\Omega \subset \Re^p, p = 2, 3$

An extension of the Gasser-Müller(1979) style estimate to higher dimensions would be:

$$\hat{r}(x) = h^{-p} \sum_{i=1}^{n} \int_{A_i} K((x-s)/h) Y_i ds$$

where $x_i \in A_i$ and $\bigcup_i A_i = \Omega$ and $A_i \cap A_j = \emptyset \; \forall i \neq j$ and where K is a positive spherically symmetric kernel function.

However, it will expensive to compute this estimate since it requires integration over the regions A_i, so I have used a modification of this estimator by approximating the integral thus:

$$\hat{r}(x) = h^{-p} \sum_{i=1}^{n} w_i K((x-x_i)/h) Y_i$$

where w_i = area of A_i and where $x \in A_i \Leftrightarrow \; i$ minimizes $|x - x_j|$ over j.

We may compute the A_i by constructing the Voronoi tesselation on the points $x_i \; for \; i = 1, ..., n$.

Asymptotic results for this estimator may be calculated in the usual manner with the use of integral approximations and Taylor expansions. We must also impose some regularity conditions on r and on d, the asymptotic density of the design points. For a more rigorous exposition of the properties of kernel-based estimates of surfaces, see Ahmad & Lin (1984), Mack & Müller(1987) and Müller(1988).

The asymptotic bias is given by

$$BIAS(x) = E\hat{r}(x) - r(x) = \frac{h^2}{2} Q(r)(x),$$

where

$$Q(r)(x) = \int z^T \nabla^2 r(x) z K(z) dz$$

and where $\nabla^2 r(x)$ is the Hessian matrix of mixed second partials of r at x. The restriction $p = 2, 3$ is necessary for integral approximation used in the bias calculation. The restriction that $p < 4$ is hardly severe since the method would be impractical in higher dimensions without a very large amount of data.

The asymptotic variance is given by

$$VAR(x) = E(\hat{r}(x) - E\hat{r}(x))^2 = \frac{\sigma^2}{nh^p d(x)} \int K^2(x) dx.$$

And so the asymptotic MSE is given by

$$MSE(x) = VAR(x) + BIAS^2(x),$$

which is minimized by

$$h(x) = (\frac{p\sigma^2 K_2}{nd(x)Q^2(r)(x)})^{1/(p+4)},$$

where $K_2 = \int K^2(x)dx$.

If we use the variable kernel estimator:

$$\hat{r}(x) = h(x)^{-p} \sum_{i=1}^{n} w_i K((x - x_i)/h(x))Y_i$$

with $h(x)$ defined as above we find that

$$MSE(x) \propto Q^2(r)(x)^{p/(p+4)} d(x)^{-4/(p+4)}$$

We may now compute the the asymptotically optimal design. We must be careful about boundary effects. We may modify the kernel in the neighborhood of the boundary or assume that r is periodic on Ω. Using the optimal fixed bandwidth:

$$d(x) \sim Uniform$$

again, but using variable bandwidths:

$$IMSE \propto \int_\Omega Q^2(r)(x)^{p/(p+4)} d(x)^{-4/(p+4)} dx.$$

Minimizing IMSE over d s.t. $\int_\Omega d(x)dx = 1$ gives the asymptotically optimal design density

$$d(x) \propto Q^2(r)(x)^{p/(p+8)}.$$

So variable bandwidths must again be used to obtain any advantage over an evenly spaced design.

We must now extend the work of Müller & Stadtmüller (1987) to variable kernel estimators of surfaces. We proceed in a similar manner by first obtaining an estimate of the best global bandwidth, $\hat{h}$, using the Mallow criterion extended to work with higher dimensional estimators. Then $Q(r)(x)$ must be estimated using a sufficiently smooth kernel. We will describe the specific kernels used later. Also d must be estimated. We may then obtain

$$\hat{h}(x) \;=\; \hat{h} \left[\frac{\int_\Omega \hat{Q}^2(r)(t)dt}{\int_\Omega \frac{1}{\hat{d}(t)} dt \hat{d}(x) \hat{Q}^2(r)(x)} \right]^{1/(p+4)}$$

The sequential design idea is used again for surfaces. We may use the same bandwidth selection ideas and develop an algorithm parallel to the one described for the one dimensional case. Difficulties arise when attempting to realize this. In one dimension, we attempt to make the design points as close as possible to the quantiles of the estimated asymptotically optimal design density. We need an analog of these quantiles, $\tilde{x}_i$, in higher dimensions. We have used the following method:

1) Locate the maximum of $\tilde{d}(x)$, the estimated asymptotically optimal design density, and place one design point $\tilde{x}_i$ there.

2) Remove mass 1/(n+1) from $\tilde{d}$ at this and surrounding locations.

3) Goto 1)

Since in practice $\tilde{d}(x)$ will be computed on a grid, this method may be implemented simply.

The next design point, x_{n+1}, may then be chosen to minimize

$$\sum_{i=1}^{n+1} |\tilde{x}_i - x_{(i)}|^2$$

where $x_{(i)}$ indicates a permutation of the $x's$ to minimize this quantity. The expense of constructing the design is now substantially more than it was in the curve estimation case but it is still much less than that required for the estimation itself.

The idea of choosing x_{n+1} so that the density estimate based on $x_i \; for \; i = 1, ..., n+1$ is as close as possible to $\tilde{d}(x)$ may occur to the reader. However, this method did not work well in practice, tending to add design points in areas of maximum curvature but nowhere else. This method would also require a selection of bandwidth for the density estimate which is problematic.

It is conceivable that, in extreme circumstances, design points will not be added to an area of locally high curvature simply because it has not been detected. To avoid this difficulty, set $\tilde{d}(x) \propto max(\tilde{d}(x), C)$ where C is some constant (possible chosen to diminish with n).

5. THE SIMULATION

Again to avoid edge effects, we consider only periodic functions on (0,1]x(0,1] so that our functions are defined on the surface of a torus. We consider the following test functions :

$$i) r(x, y) = sin(2\pi x) + sin(2\pi y).$$

$$ii) r(x, y) = sin(2\pi x) * sin(2\pi y).$$

$$iii) r(x, y) = sin(4\pi x) + sin(4\pi y).$$

$$iv) r(x, y) = 0.$$

We compute the estimate on a grid of 20x20=400 points and use 400 replications. We start with n=25, which is about the minimum necessary to obtain some estimate of $Q(r)$ and end at n=100. We compare the evenly spaced design, the sequential design and the design based on knowledge of true r. We choose $\sigma = 0.05(maxr(x) - minr(x))$. Constructing an “evenly spaced” design for given n is problematic. Our initial 25 points are a 5x5 square grid. The next 25 points are constructed by adding 0.1 to each of the coordinates of the original 25 points. The next 25 points are constructed by adding 0.1 to one of the coordinates of the original 25 points and the remaining 25 are constructed by adding 0.1 to the other coordinate. For any given n, this will be somewhat sub-optimal, but it will serve as a reasonable comparison to the sequential method. The evenly spaced design was randomly translated for each replication to avoid an interaction effect with r caused by the coincidence of design points with the peaks and valleys of $r(x)$ and such like. We used the Epanechnikov kernel for the estimation of r and

$$K(x) \;=\; \begin{cases} const.(1-|x|^2)^3 & \text{if } |x|<1 \\ 0 & \text{otherwise} \end{cases}$$

for the estimation of r''. The bandwidth used for the estimation of r'' was $2.5\hat{h}$ where $\hat{h}$ is the global bandwidth chosen by the Rice criterion using the Epanechnikov kernel. Why 2.5? - asymptotically we expect the optimal bandwith for the estimation of r'' to be a constant multiple of $\hat{h}$. This constant could be determined by calculation, but note that our primary purpose is not the best estimate of r'' but of $h(x)$ and $d(x)$. For this reason the constant 2.5 was determined by simulation experiment as best across the range of surfaces considered here. The $d(x)$ necessary for the calculation of $h(x)$ was estimated using a kernel density estimate using the Epanechnikov kernel and the bandwidth $\hat{h}$.

In figure 2, the average IMSE of the variable kernel estimate is plotted against sample size. The evenly spaced design is given by the solid line, the sequential design by the dotted line and the optimal design by the dashed line. The simulation standard errors are estimated to be at most 3% and are not significant factor in interpreting the results. One can see that in the first two cases, the sequential design significantly outperforms the evenly spaced design and in the third case the sequential design overtakes the evenly spaced design as the sample size increases. The sequential design approaches the optimal design, thus validating the sequential design procedure. The evenly spaced design produces a rather erratic result but this is to be expected given the fact it is not truly evenly spaced.

An examination of the design points showed convergence towards the expected design. The IMSE of the optimal design is about 91% of the IMSE of the evenly spaced design in the limit for i)-iii) (and about 98% for i)-iii) for the one-dimensional example discussed earlier). The gain in using the sequential design for small samples is much larger than this.

For very small sample sizes the global bandwidth estimator outperforms the variable kernel estimator, but this changed as the sample size increased. For the sequential design, the variable kernel estimator overtook the global bandwidth estimator at about sample size 60,80,100 for surfaces i),ii),iii) respectively.

6. DISCUSSION

Edge effects , heteroscedascity and weighting may be allowed for with a suitable modification of the estimates. Suppose the variance of the error is given by $\sigma^2(x)$ and introduce a weight function $\gamma(x)$, which may also be used to allow for edge effects.

Thus, we now wish to obtain best estimates of the weighted IMSE , $\int_\Omega MSE(x)\gamma(x)dx$.

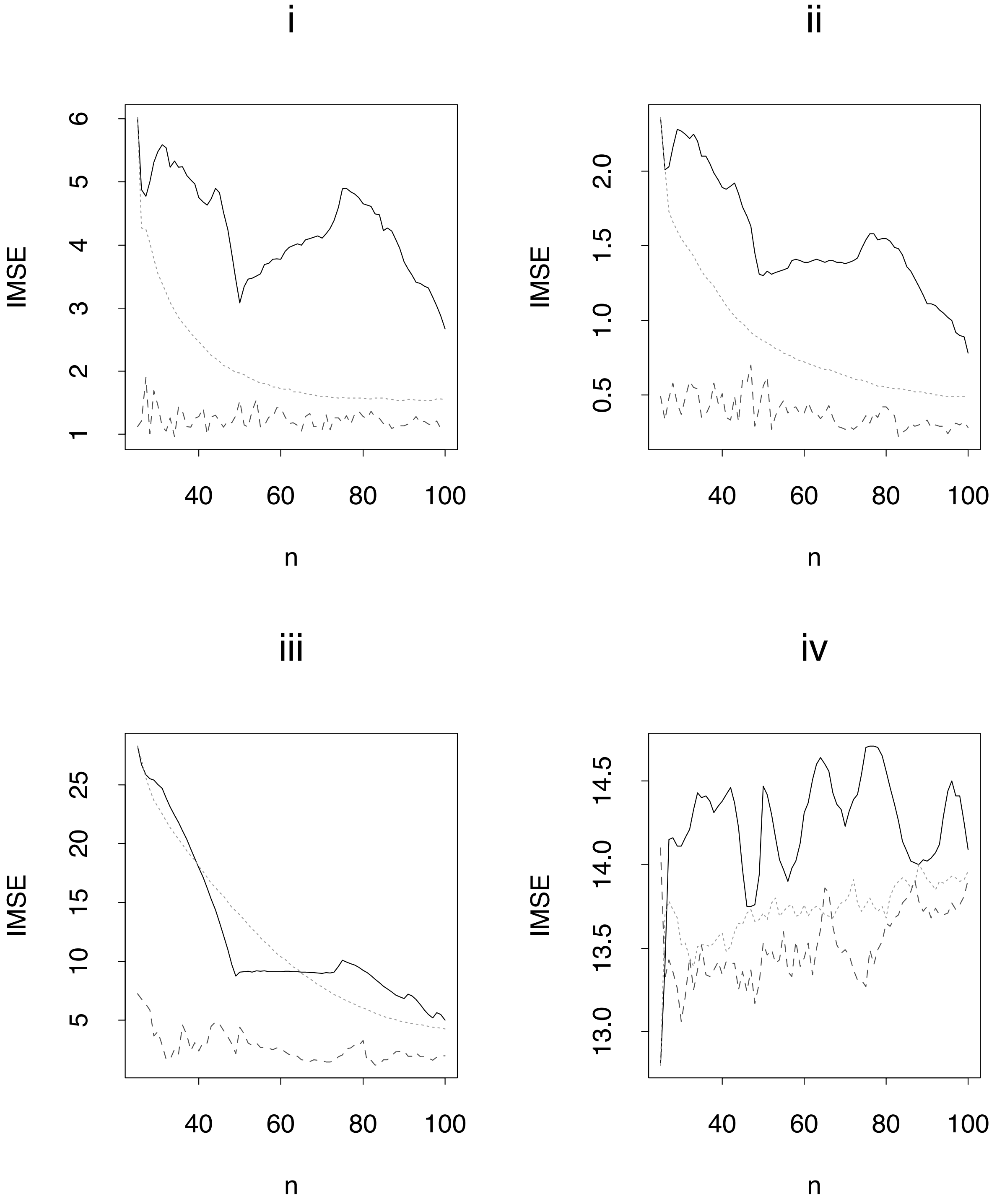
The optimal variable bandwidth choice now becomes

$$h(x) = (\frac{p\sigma^2(x)K_2}{nd(x)Q^2(r)(x)})^{1/(p+4)}$$

and the asymptotically optimal design becomes

$$d(x) \propto Q^2(r)(x)^{p/(p+8)}\sigma^2(x)^{2/(p+8)}\gamma(x)^{(p+4)/(p+8)}.$$

Figure 2


i

ii

iii

iv

Figure 2: Surface simulation results: Evenly space is dotted, sequential is solid and optimal is dashed

Of course, $\sigma^2(x)$ will have to be estimated before these modifications may be implemented. Carroll (1980) has a method for estimating the form of this heteroscedascity.

Also, we may not require the same accuracy of estimation of $r(x)$ for all x. For example, we may want more accurate estimation in regions where $r(x)$ is large. In this case, we could make the weight function γ a function of r. We would have to change γ as our estimate of r improved.

Another potential problem is dependency in the errors. It is not unlikely that, in a practical situation, there will be some dependency of observations taken at the same or nearby design points.

In some situations, it may not be possible or desirable to take measurements one at a time. The design points may need to be assigned in batches because of some physical or cost constraint or maybe because there is some expense in moving the design point some distance from it's previous position. The sequential design procedures described above may be easily modified to take account of this situation.

REFERENCES

Box G. & Draper N. (1987) "Empirical Model building and Response Surfaces" *Wiley, New York.*

Carroll R. (1980) "Adapting for heteroscedascity in linear models" *Annals of Statistics* **10** 1224-1233

Draper N. (1988) "Response Surface designs" *Encyclopaedia of Stat. Sci. Vol. 8.* Wiley.

Faraway J. & E. Rothman (1989) "Sequential design for nonparametric regression" *Technical Report #174, Dept. of Statistics, Univ of Michigan*

Gasser T. & Müller H. (1979) "Kernel Estimation of Regression functions" in *"Smoothing Techniques for Curve estimation", Springer Lecture notes* **757** 23-68

Gasser T. & Müller H. (1984) "Nonparametric estimation of regression functions and their derivatives" *Scandinavian Journal of Statistics.* **11** 171-185

Mack Y. & Müller H. (1987) "Adaptive nonparametric estimation of a multivariate regression function" *J. Multi. Anal.* **23** 169-182

Müller H. (1984) "Optimal designs for nonparametric kernel regression" *Statistics and Probability Letters.* **2** 285-290

Müller H. & Stadtmüller U. (1987) "Variable kernel bandwidth estimators of regression curves" *Annals of Statistics.* **15** 182-201

Müller H. (1988) "Nonparametric Regression Analysis of Longitudinal Data" *Springer Verlag, Berlin.*

Myers R., Khuri A. & Carter Jr. W. (1989) "Response Surface Methodology: 1966-1988" *Technometrics* **31** 137-153

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

Ripley B. (1981) "Spatial Statistics" *Wiley, New York.*