Nonparametric Statistics, Vol. 9, pp. 155–164
Reprints available directly from the publisher
Photocopying permitted by license only

SEQUENTIAL DESIGN FOR RESPONSE CURVE ESTIMATION

DONGRYEON PARK[a,]* and JULIAN J. FARAWAY[b]

[a]*Department of Statistics, Hanshin University, Korea, 447-791;*
[b]*Department of Statistics, University of Michigan, Ann Arbor, MI 48109*

(Received 12 December 1997; In final form 20 January 1998)

The problem of sequential design for a nonparametric regression with binary data is considered. The aim of the statistical analysis is the estimation of a quantal response curve p. An adaptive method is developed that proposes the location of the next best design point on the basis of past observations. The behavior of this estimator is discussed and its small sample properties are investigated using a simulation study.

Keywords: Nonparametric regression; dose-response curve; van der Corput sequence

1. INTRODUCTION

Suppose the outcome of an experiment is dichotomous – success or failure and that the probability of success is a function of the stimulus level at which the experiment is carried out. In these experiments, we assume that the reaction Y_i of the ith subject at stimulus level x_i $(i = 1, \ldots, n)$ is an independent Bernoulli random variable with parameter $p(x_i)$, $i = 1, \ldots, n$. The specification of the stimulus levels x_i forms the design of the experiment. We assume that p is continuous and strictly monotone increasing. We want to estimate the curve p. Much prior work has been devoted to the estimation of quantiles of p, for example the so-called ED50 level. However, in situations where the parametric form of p is unknown, we may want to estimate the whole curve p or substantial portions of it.

* Corresponding author.

Sometimes, we have a fixed sample size available and must decide on the location of all the design points $x_1, x_2, \ldots,$ in advance. If one uses a kernel-based estimate of p, then Müller and Schmitt (1988) describe the asymptotically optimal design density. You would need some prior knowledge of p to construct such a design in practice. Failing that, 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. In this paper, we present a new sequential design for the estimation of the response curve for the nonparametric regression approach. The advantage of the sequential design is that significantly greater precision may be had for the same sample size or fewer measurements may be required to obtain some specified accuracy. When the experimental runs are very expensive, the saving of a few runs by an efficient design outweights the extra effort required in designing and running a sequential experiment. In Section 2, we describe our algorithm and discuss how our design converges to the optimal design based on knowledge of the true p. In Section 3, we present a simulation study that illustrates how well our method works with small samples. In Section 4, we conclude.

2. SEQUENTIAL DESIGN FOR CURVE ESTIMATION

2.1. The Estimator

We will assume that $p \in C^2([0,1])$, and that p is strictly monotone increasing. We use the kernel-based estimate proposed in Müller and Schmitt (1988) which is defined as follows:

$$\hat{p}(x) = \frac{1}{b}\sum_{i=1}^{n}\int_{s_{i-1}}^{s_i} K\left(\frac{x-u}{b}\right) du\, Y_i, \tag{1}$$

where b is a sequence of a positive bandwidths depending on n such that $b \to 0$, $nb \to \infty$ as $n \to \infty$ and where K is a continuous kernel function satisfying: $\int K(u)du = 1$, $\int K(u)udu = 0$, $\int K(u)u^2du > 0$. The kernel is assumed to have a compact support, $[-1, 1]$ and is required to satisfy $K \in \mathrm{Lip}([-1, 1])$. $s_0 = 0$, $s_n = 1$ and $s_i = (x_i + x_{i+1})/2$ for $1 \leq i \leq (n-1)$.

2.2. Sequential Design Algorithm

A strictly positive design density f on [0, 1] satisfying $f \in \mathrm{Lip}([0, 1])$, uniquely determines the design points $x_1, \ldots, x_n$ by

$$\int_0^{x_i} f(t)dt = \frac{i-1}{n-1}.$$

Suppose that the design points, x_i, must be given in advance of the experiment. Müller and Schmitt (1988) derive the optimal design density f^* minimizing the asymptotic integrated mean squared error for p using the optimal bandwidth b as

$$f^*(x) = \frac{\sqrt{p(x)(1-p(x))}}{\int_0^1 \sqrt{p(y)(1-p(y))}dy}.$$

Let $x_1^*, \ldots, x_n^*$ be the design points based on this optimal design. In general, all the points would change if the sample size, n, were changed so there is an implicit dependence on n and so the design is only suitable when n is fixed in advance. Furthermore, the design is based on the asymptotic IMSE so it is only asymptotically optimal.

This design could be used where some initial estimate of the response curve is available such as in a two-stage experiment but when the observations are collected sequentially, it is possible to do better than this. We propose the following sequential design algorithm to make the actual design density as close as possible to the optimal design density.

Sequential Design Algorithm

1. Start with m initial design points, $\{x_i\}_1^m$ with corresponding binary observations $\{Y_i\}_1^m$ assumed to be generated by $P(Y_i = 1) = p(x_i)$ with true response p unknown.
2. Estimate the cumulative distribution function of the optimal design density, $\hat{F}^*(x)$ based on the current data, so

$$\hat{F}^*(x) = \int_0^x \frac{\sqrt{\hat{p}(t)(1-\hat{p}(t))}}{\int_0^1 \sqrt{\hat{p}(y)(1-\hat{p}(y))}dy}dt.$$

   where $\hat{p}$ is obtained using (1).

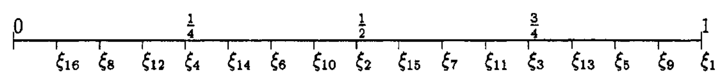
3. Choose the next design point, x_i as $\hat{F}^{*-1}(\xi_i)$ where $\xi_1 = 1$, $\xi_2 = 1/2$, and for ξ_i, $i \geq 3$

$$\xi_{2^{j-1}+k} = \begin{cases} \xi_{2^{j-2}+k} + \frac{1}{2^j} & \text{if } 1 \leq k \leq 2^{j-2} \\ \xi_k - \frac{1}{2^j} & \text{if } 2^{j-2}+1 \leq k \leq 2^{j-1} \end{cases}$$

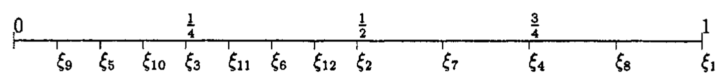
where $j \geq 2$.

4. Repeat Steps 2 and 3.

To show the pattern, here are the locations of the first 16 points



At each stage, we are bisecting the largest interval, but even so, there is usually a choice of intervals and we must choose carefully. We would like the ξ_i's to be as close to uniform as possible. For example, suppose we choose the leftmost largest interval, then the first 12 points would be



In the above figure, 8 out of 12 points are less than or equal to 1/2. This unbalanced choice is far from uniform. Note that even as the sample size increases, the design retains this awkward property.

Our arrangement of the ξ_i's looks strange, but to show that the estimator based on the sequential design has the same IMSE as that based on the optimal design, we require that $\{\xi_1, \ldots, \xi_n\}$ should be almost evenly spaced for all n. Let

$$Q_n = \max_{1 \leq i \leq n} |\xi_{(i)} - \eta_i|.$$

where $\eta_i \in [(i-1)/n,\ i/n]$ and the subscript parentheses indicate sorting in ascending order.

The ξ_i's we have chosen are effectively equivalent to a van der Corput sequence with base 2. Such sequences have been used in numerical integration problems – see Niederreiter (1978). From this literature we know that $Q_n \leq (\log_2 n/3 + 1)/n$. This is the best rate of convergence that can be achieved although slightly better constants may be obtained using special irrational numbers although there are difficulties implementing these on computers which is why we have used a van der Corput sequence. We need only an $O(n^{-4/5})$ rate which is more than satisfied by this sequence. Note that a purely random sequence would achieve only an $O(n^{-1/2})$ rate and thus would fail.

It is very difficult to get theoretical results with the estimator based on the sequential design as defined above because of the complex dependence structure. We assume that the optimal design density function is given, so the sequential design point x_i is chosen by $x_i = F^{*-1}(\xi_i)$. We also modify the estimator for the sequential design. Let

$$\hat{p}(x) = \frac{1}{b}\sum_{i=1}^{n} \int_{s_{i-1}^*}^{s_i^*} K\left(\frac{x-u}{b}\right) du\, Y_i, \tag{2}$$

where

$$s_i^* = \frac{x_i^* + x_{i+1}^*}{2}, \quad x_i^* = F^{*-1}\left(\frac{i-1}{n-1}\right).$$

Thus we "cheat" by using some knowledge of the true p in the sequential design based estimator. For the sequential design based estimator of (2) to have the same asymptotic IMSE as the optimal design, the sequential design points x_i have to be relatively close to their respective intervals $[s_{i-1}^*, s_i^*]$. Of course, if $x_i \in [s_{i-1}^*, s_i^*]$, $\forall i$, there would be no difficulty but this is not true in general and in fact it does not seem possible to design a sequence $\xi_1, \xi_2, \ldots$ so that this can be true in general. Nevertheless, the properties of our sequential design guarantees that they will be somewhat close in the general case.

Since f^* is strictly positive on [0, 1], F^{*-1} has a bounded derivative on [0, 1] so for some $\gamma_i \in [s_{i-1}^*, s_i^*]$ and constant M,

$$|x_i - \gamma_i| = |F^{*-1}(\xi_i) - F^{*-1}(\eta_i)| \leq M|\xi_i - \eta_i| = O(n^{-\delta})$$

where $4/5 < \delta \leq 1$ and $\eta_i \in [(i-1)/n, i/n]$.

This is sufficient to work through the derivation of the IMSE as in Müller and Schmitt (1988) where this fact is crucial to the bias and the variance calculation. In this manner, we show that the estimator based on the sequential design has the same IMSE as that based on the optimal design.

3. SIMULATION STUDY

We need to show that, for small samples, the sequential design approaches the performance of the optimal design and works better than the equally spaced design that might be used in a non-sequential experiment. Our sequential procedure used below uses no knowledge of the true response and is thus a fair test of its ability.

One practical difficulty is that the kernel estimate (1) is not necessarily monotone even if the true response curve is monotone increasing as is usually assumed. To fix this, we used the "Pool Adjacent Violators Algorithm" as a monotonizing transformation. (See Barlow, Bartholomew, Bremner and Brunk (1972)).

An appropriate choice of the bandwidth b is also very important. For the evenly spaced design, Müller and Schmitt (1988) adapted the Rice criterion (1984) for response curve estimation. Since our sequential design is not evenly spaced, we made a further change to allow for this. We chose b to minimize

$$\hat{R}(b) = \sum_{i=1}^{n} w_i(y_i - \hat{p}(x_i))^2 + 2\hat{\sigma}^2 \sum_{i=1}^{n} w_i \left(\frac{1}{b} \int_{s_{i-1}}^{s_i} K\left(\frac{x_i - u}{b}\right) du \right)$$

where $w_i = s_i - s_{i-1}$ and $\hat{\sigma}^2 = (1/2(n-1))\Sigma_{i=2}^{n}(Y_i - Y_{i-1})^2$.

We used the Epanechnikov kernel for K. Although the estimation of p on [0, 1] is the objective, we restricted the selection of b to [.25, .75] to avoid edge effects in the choice.

We compared the sequential design to the fixed design where the design points are evenly spaced and to the optimal design where the points are chosen by

$$\int_0^{x_i^*} f^*(t)dt = \frac{i-1}{n-1}$$

where f^* is the optimal design density described in Section 2.2.

For the "true" response curves in our experiment, we used the probit model $p_1(x) = \Phi((x - 0.5)/0.1)$, the normal mixture model $p_2(x) = 0.5\Phi((x - 0.4)/0.05) + 0.5\Phi((x - 0.6)/0.05)$, and the Weibull model $p_3(x) = 1 - \exp(-(x/0.5)^4)$. For each model, $N = 400$ Monte Carlo runs were made.

The sequential procedure needs some initial points before it can work–we used 20 evenly spaced points. We then followed our sequential design procedure up to a sample size of $n = 100$.

In Figures 1–3, the Monte Carlo IMSE of each design is plotted against the sample size for each model. The standard errors are estimated to be at most 3% and are not significant factor in interpreting the results. The reader may be curious why the curves for the evenly spaced and optimal designs in particular are so rough.

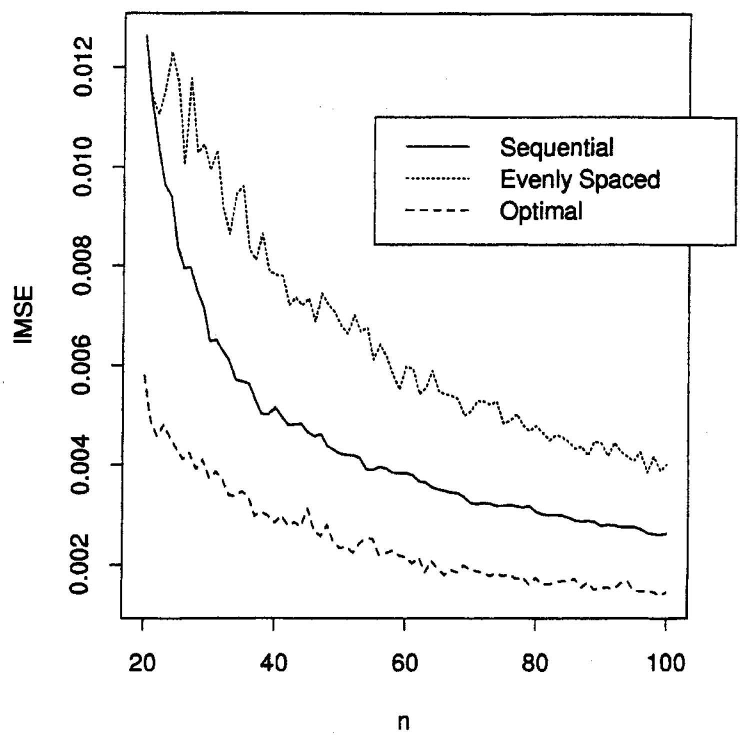


FIGURE 1 The average of estimated IMSE for the probit model, $p_1(x)$.

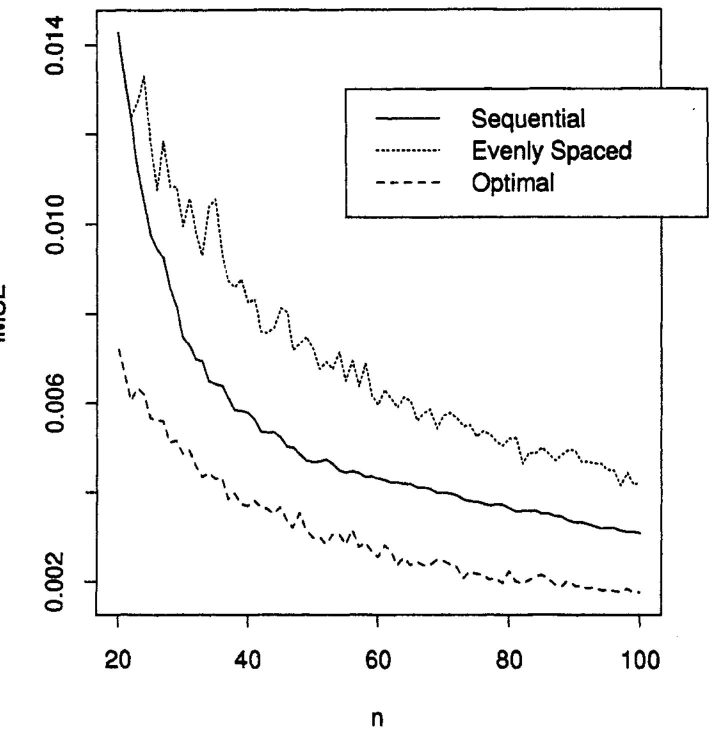


FIGURE 2 The average of estimated IMSE for the normal mixture model, $p_2(x)$.

When one point is added, these two designs change completely so there is a lack of continuity in n. Furthermore, the bandwidth selection procedure must be repeated, further adding to the "roughness".

In each model, the sequential design outperforms the evenly spaced design. In other words, the evenly spaced design requires a substantially larger sample to attain the same degree of accuracy as the sequential design. Table I shows how many samples are required to attain the same degree of accuracy for each design. In some cases, we can save more than half the experimental runs by using the sequential design rather than the evenly spaced design. Of course, more computational work is required for the sequential design but when the experimental runs are very expensive, this saving of runs outweights the extra effort.

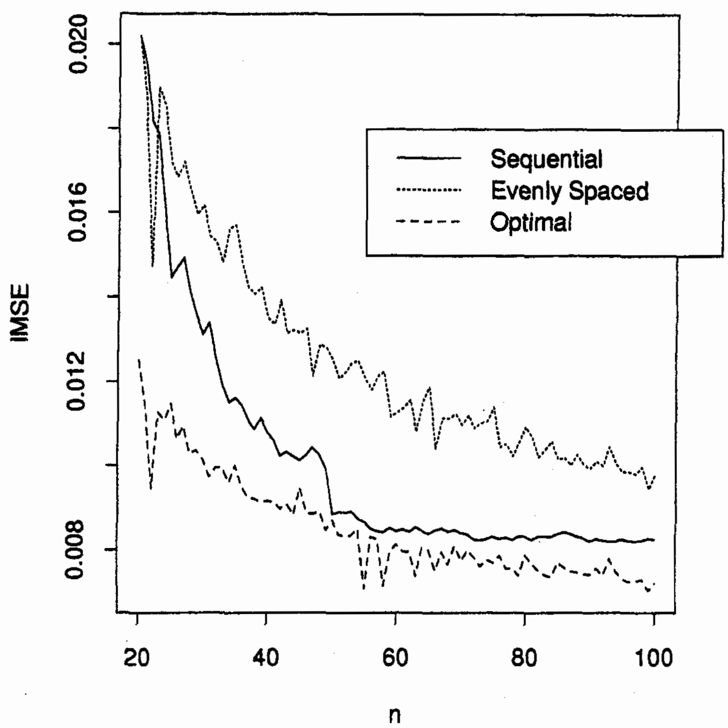


FIGURE 3 The average of estimated IMSE for the Weibull model, $p_3(x)$.

TABLE I Sample size to attain the same degree of accuracy

Model	IMSE	Opt. design	Seq. design	Even. design
	0.0057	20	34	58
$p_1(x)$	0.0038	31	54	100
	0.0025	55	100	> 100
	0.0072	20	31	44
$p_2(x)$	0.0043	36	60	100
	0.0030	54	100	> 100
	0.0125	20	32	54
$p_3(x)$	0.0099	35	49	100
	0.0082	57	100	> 100

4. DISCUSSION

We have presented a new sequential design and used a simulation study to show that it is effective. We have provided some evidence that this sequential estimator is asymptotically equivalent to the estimator based on the optimal design. We have not proved this for the practical estimator and, furthermore, we believe that such a proof would be very difficult given the dependent nature of the estimator.

We have used an estimator with a global bandwidth, but often the design points will be unequally spaced and so a local bandwidth may perform better. The use of local bandwidths will change the optimal design and may lead to better estimates.

We have focussed on the estimation of the whole curve p, but often specific percentiles of p are of special interest. The method above can be modified easily by introducing a weight function.

Acknowledgements

The authors thank the referee for a suggestion that led to a substantive improvement in this paper.

References

Barlow, R. E., Bartholomew, D. J., Bremner, J. M. and Brunk, H. D. (1972) Statistical Inference Under Order Restrictions, John Wiley, New York.

Chu, C. K. (1993) "A New Version of the Gasser–Mueller Estimator", *Nonparametric Statistics*, **3**, 187–193.

Eubank, R. L. (1988) Spline Smoothing and Nonparametric Regression, Marcel Dekker, New York.

Faraway, J. (1990) "Sequential Design for the Nonparametric Regression of Curves and Surfaces", *Proceedings of the 22nd Symposium on the Interface between Computing Science and Statistics*, pp. 104–110, Springer.

Gasser, T. and Müller, H. (1979) "Kernel Estimation of Regression Functions", In Smoothing Techniques for Curve Estimation, pp. 23–68, Springer, Heidelberg.

Gasser, T. and Müller, H. (1984) "Estimating Regression Functions and Their Derivatives by the Kernel Method", *Scandinavian Journal of Statistics*, **11**, 171–185.

Müller, H. (1984) "Optimal Designs for Nonparametric Kernel Regression", *Statistics & Probability Letters*, **2**, 285–290.

Müller, H. and Schmitt, T. (1988) "Kernel and Probit Estimates in Quantal Bioassay", *Journal of the American Statistical Association*, **83**, 750–759.

Niederreiter, H. (1978) "Quasi-Monte Carlo Methods and pseudo random numbers", *Bulletin of the American Mathematical Society*, **84**, 957–1041.

Rice, J. (1984) "Bandwidth choice for Nonparametric Regression", *The Annals of Statistics*, **12**, 1215–1230.