Ann. Inst. Statist. Math.
29 (1977), Part A, 301-306

ON THE EXISTENCE OF SEARCH DESIGNS WITH CONTINUOUS FACTORS

J. N. SRIVASTAVA AND S. GHOSH*

(Received July 30, 1976; revised Dec. 27, 1976)

Abstract

Consider the search linear model defined as follows. Let $\boldsymbol{y}(N\times 1)$ be a vector of N observations such that

$$(1) \qquad \mathrm{E}(\boldsymbol{y})=A_1\boldsymbol{\xi}_1+A_2\boldsymbol{\xi}_2\,, \qquad \mathrm{V}(\boldsymbol{y})=\sigma^2 I_N$$

where σ^2 may or may not be known, $A_1(N\times\nu_1)$ and $A_2(N\times\nu_2)$ are known matrices, $\boldsymbol{\xi}_1(\nu_1\times 1)$ is unknown and $\boldsymbol{\xi}_2(\nu_2\times 1)$ is partly known in the following sense. We know that at most k elements of $\boldsymbol{\xi}_2$ are non zero but we do not know particularly which these nonzero elements are. The problem is to make inferences about the elements of $\boldsymbol{\xi}_1$ and, furthermore, to search the nonzero elements of $\boldsymbol{\xi}_2$ and make inferences about them. We want $\boldsymbol{y}$ to be such that the above problem can be resolved with certainty when $\sigma^2=0$; the underlying design corresponding to $\boldsymbol{y}$ is then called a search design. It has been shown in earlier work that for a search design, we must have $N\geqq\nu_1+2k$. In this paper, we consider the special case of search linear models, when the object of the experiment is to fit an appropriate response surface. We establish a basic result, namely, that when the true response surface is representable by a polynomial, then search designs exist for which $N=\nu_1+2k$, irrespective of the value of ν_2.

1. Introduction

The model (1) was introduced in Srivastava [3]. The case $\sigma^2=0$ is called the "noiseless case". In statistical problems, we always have $\sigma^2>0$. The noiseless case is, however, important in search theory because the difficulties in the noiseless case are also present when $\sigma^2>0$. We now recall the following theorem from Srivastava [3].

THEOREM 1. *Consider the model* (1) *with* $\sigma^2=0$. *A necessary and*

* This work was supported by NSF Grant No. MPS73-05086 A01.

sufficient condition that the search and inference problem can be completely solved (in this noiseless case), is that for every $(N\times 2k)$ *submatrix* A_{20} *of* A_2, *we have*

$$\text{Rank}\,(A_1: A_{20})=\nu_1+2k\,. \tag{2}$$

By "completely solved", we mean that we will be able to search the nonzero elements of $\boldsymbol{\xi}_2$ without any error, and obtain estimators (with zero variance) of the nonzero elements of $\boldsymbol{\xi}_2$ and of the elements of $\boldsymbol{\xi}_1$. A set of observations $\boldsymbol{y}$ for which condition (2) is satisfied, is said to be based on a "search design".

Clearly, even though the basic design used be a search design, the probability of correct search and estimation in the noisy case ($\sigma^2>0$), will depend upon the size of noise present. This feature is, of course, present in all statistical problems.

Also, it is clear that search linear models would fit real life situation better than ordinary general linear models. The reason is that in an ordinary linear model, a set of parameters like $\boldsymbol{\xi}_2$ gets ignored (particularly, if ν_2 is large). Thus, in most situations, the ordinary linear model gives a biased fit. Earlier authors have considered the "minimization" or "balancing" of bias. (See, for example, Box and Draper [1], Hedayat, Raktoe and Federer [2].) The search linear models offer the opportunity for "searching" the bias and "correcting" for it.

In the theory of factorial experiments with continuous factors, we approximate the response surface by some linear combinations of certain (known) functions of factor levels. A natural question to ask is concerning the validity of such approximations. Of course, such approximations are never *totally* wrong, a fact verified on the basis of the large amount of empirical evidence available from various sciences, and also partly justified from theoretical considerations.

For example, first and second degree polynomials are widely used in chemical and other industrial experiments for fitting response surfaces. The great popularity of this approach to explain the experimental data suggests that such polynomials do give a fit which is certainly not extremely bad. The reason is that if the approximation of a response surface by such a polynomial was in general very poor, people would not be employing them so often to try to explain their data. The partial theoretical justification arises from the fact that in a large variety of situations, the variation in natural phenomena is sufficiently smooth. In other words, in many natural phenomena, the response surface is adequately representable by the first two or three terms in its Taylor expansion. This is particularly so if all the experimental points are quite close to some center point, this center point being the point around with the Taylor expansion is considered.

Such a situation may arise in practice, for example, when the center point corresponds to a combination of factor levels at which some industrial article is presently manufactured. The experimental points then correspond to suggested "small" changes in the levels of these factors.

In spite of the above, it is also true that in most experiments, though the fit provided by a low degree polynomial does not go wildly wrong, it still is considerably improvable. This is particularly so since in almost every experiment there do occur a few functions of factor levels whose contributions to response were assumed negligible, but which in fact are non-negligible, and which are also difficult to pinpoint in advance. The use of search linear models is, therefore, very much called for.

This paper presents a basic result regarding existence of designs with the minimum number of runs for response surface experiments when a search linear model is used. The problems of obtaining search designs, and optimality criteria for the same will be considered in separate communications. Some basic developments on optimal search designs will be found in Srivastava [5].

2. Search linear models with continuous factors

Consider a factorial experiment with m continuous factors. Treatment combinations will be denoted by $\boldsymbol{x}'=(x_1, x_2, \cdots, x_m)$ where $x_i \in [a_i, b_i]$, $i=1, 2, \cdots, m$, and where a_i and b_i are real numbers such that $a_i<b_i$. Thus, $\boldsymbol{x}$ belongs to the m-dimensional closed interval $[\boldsymbol{a}, \boldsymbol{b}]=[a_1, b_1]\times[a_2, b_2]\times\cdots\times[a_m, b_m]$. Let $f_1, f_2, \cdots, f_\nu$ be known functions on $[\boldsymbol{a}, \boldsymbol{b}]$ with the following properties:

(i) $f_1, \cdots, f_\nu$ are continuous and real valued on $[\boldsymbol{a}, \boldsymbol{b}]$,

(ii) $f_1, \cdots, f_\nu$ are linearly independent, and

(iii) $f_1, \cdots, f_\nu$ are distinct from zero.

For each $\boldsymbol{x} \in [\boldsymbol{a}, \boldsymbol{b}]$, an experiment is performed, whose outcome is an observation $y(\boldsymbol{x})$. Consider the search linear model (1) for this special case of continuous factors. For N treatment combinations $\boldsymbol{x}_1(m\times 1), \cdots, \boldsymbol{x}_N(m\times 1)$, and the corresponding observations $y(\boldsymbol{x}_1), \cdots, y(\boldsymbol{x}_N)$, in the model (1), we then have

$$(3) \qquad \boldsymbol{y}'=(y(\boldsymbol{x}_1), \cdots, y(\boldsymbol{x}_N)), \qquad \boldsymbol{\xi}_1'=(\beta_1, \cdots, \beta_{\nu_1}),$$
$$\boldsymbol{\xi}_2'=(\beta_{\nu_1+1}, \cdots, \beta_\nu), \qquad \nu=\nu_1+\nu_2,$$

and

$$(4) \qquad A_1=\begin{bmatrix} f_1(\boldsymbol{x}_1) & \cdots & f_{\nu_1}(\boldsymbol{x}_1) \\ & \cdots & \\ f_1(\boldsymbol{x}_N) & \cdots & f_{\nu_1}(\boldsymbol{x}_N) \end{bmatrix}, \qquad A_2=\begin{bmatrix} f_{\nu_1+1}(\boldsymbol{x}_1) & \cdots & f_\nu(\boldsymbol{x}_1) \\ & \cdots & \\ f_{\nu_1+1}(\boldsymbol{x}_N) & \cdots & f_\nu(\boldsymbol{x}_N) \end{bmatrix}.$$

The following lemma, whose proof is obvious, gives the reason for assuming $f_1, \cdots, f_\nu$ to be linearly independent.

LEMMA 1. *If in our model $f_1, \cdots, f_\nu$ are linearly dependent, then the parameters β's will be confounded.*

3. Existence of search designs

Let $\boldsymbol{x}_1(m\times 1)$, $\boldsymbol{x}_2(m\times 1), \cdots, \boldsymbol{x}_t(m\times 1)$ be a set of t distinct points in $[\boldsymbol{a}, \boldsymbol{b}]$. Let

$$T(t\times m)=\begin{bmatrix} \boldsymbol{x}_1' \\ \cdot \\ \cdot \\ \cdot \\ \boldsymbol{x}_t' \end{bmatrix}=\begin{bmatrix} x_{11} & \cdots & x_{1m} \\ & \cdots & \\ & \cdots & \\ & \cdots & \\ x_{t1} & \cdots & x_{tm} \end{bmatrix} \tag{5}$$

and

$$A(t\times\nu)=\begin{bmatrix} f_1(\boldsymbol{x}_1) & f_2(\boldsymbol{x}_1) & \cdots & f_\nu(\boldsymbol{x}_1) \\ & & \cdots & \\ f_1(\boldsymbol{x}_t) & f_2(\boldsymbol{x}_t) & \cdots & f_\nu(\boldsymbol{x}_t) \end{bmatrix}, \qquad t\leqq\nu, \tag{6}$$

be the treatment and the model matrices respectively. Note that there is a correspondence between the treatment matrix T and the point $(x_{11}, \cdots, x_{1m}, \cdots, x_{t1}, \cdots, x_{tm})$ in the mt-dimensional closed interval R, where $R=[\boldsymbol{a}, \boldsymbol{b}]\times\cdots\times[\boldsymbol{a}, \boldsymbol{b}]$.

LEMMA 2. *For any submatrix $A_0(t\times t)$ of $A(t\times\nu)$, the condition*

$$\det A_0=0 \text{ for all } (x_{11}, \cdots, x_{tm}), \qquad x_{ij}\in[a_j, b_j], \tag{7}$$

implies that $f_1, \cdots, f_\nu$ are linearly dependent.

PROOF. Let

$$A_0=\begin{bmatrix} f_{i_1}(\boldsymbol{x}_1) & \cdots & f_{i_t}(\boldsymbol{x}_1) \\ & \cdots & \\ f_{i_1}(\boldsymbol{x}_t) & \cdots & f_{i_t}(\boldsymbol{x}_t) \end{bmatrix}. \tag{8}$$

There exist constants $\lambda_1, \cdots, \lambda_t$ not all zero such that

$$\lambda_1 f_{i_1}(\boldsymbol{x})+\cdots+\lambda_t f_{i_t}(\boldsymbol{x})=0\,, \tag{9}$$

for all $\boldsymbol{x}\in[\boldsymbol{a}, \boldsymbol{b}]$. This implies that $f_{i_1}, \cdots, f_{i_t}$ are linearly dependent. This completes the proof.

The following well known result will be needed later.

LEMMA 3. *If $\phi(x_1, \cdots, x_n)$ is a polynomial of finite degree in $x_1, \cdots, x_n$, in which not all coefficients are zero, then the Lebesgue measure of* *the set of points satisfying* $\phi(x_1, \cdots, x_n)=0$ *in n-dimensional space is zero.*

A matrix $A(t\times\nu)$, $t\leqq\nu$, is said to have the property P_t if every set of t columns of A are linearly independent.

THEOREM 2. *If* $f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \cdots, f_\nu(\boldsymbol{x})$ *are a set of* ν *linearly independent polynomials of finite degree in* $x_1, x_2, \cdots,$ *and* x_m, *then almost all points in the mt-dimensional closed interval of the mt variables* x_{ij} $(i=1, \cdots, t; j=1, \cdots, m)$ *are such that* $A(t\times\nu)$, *in* (6), *has the property* P_t.

PROOF. Consider the submatrix $A_0(t\times t)$ of A, as at (8). Let

$$\phi_{i_1, i_2, \cdots, i_t}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t)=\det A_0\,. \tag{10}$$

Clearly, $\phi_{i_1, i_2, \cdots, i_t}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t)$ is a polynomial of finite degree in the variables x_{ij}, $i=1, \cdots, t$, $j=1, \cdots, m$. Also, it is not identically zero, since in that case it will contradict the assumption of linear independence by Lemma 2. In other words, $\phi_{i_1, i_2, \cdots, i_t}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t)$ does not have all the coefficients zero. Therefore, by Lemma 3, the set of all points satisfying $\phi_{i_1, \cdots, i_t}(\boldsymbol{x}_1, \cdots, \boldsymbol{x}_t)=0$ in R denoted by $S(i_1, i_2, \cdots, i_t)$, has Lebesgue measure zero. For each of the $\binom{\nu}{t}$ choices of $(i_1, i_2, \cdots, i_t)$ the set $S(i_1, i_2, \cdots, i_t)$, which is a subset of R, has Lebesgue measure zero. Let $S=\bigcup_{\substack{\text{all}\\(i_1, \cdots, i_t)}} S(i_1, i_2, \cdots, i_t)$ and $\bar{S}=R-S$. Then S also has Lebesgue measure zero, and $\bar{S}$ and R have the same Lebesgue measure. Also, for each point in $\bar{S}$, the matrix A has the property P_t. This completes the proof.

It is clear that Theorem 2 is true for more general f's than just for polynomials as considered here.

From Theorems 1 and 2 it follows, that for experiments with continuous factors when the true model is representable by a polynomial, there exists seach design with (ν_1+2k) treatments. Indeed, except for a set of measure zero, all possible designs with (ν_1+2k) runs are search designs.

The significance of this result is that, irrespective of the value of ν_2, we can, in the noiseless case, solve the search and estimation problem completely, using only (ν_1+2k) treatments. Thus, in the more exact sciences where σ^2 is small, the above shows that there is a great scope of discriminating between competing models using only a small number of runs. (The "competing models" correspond to different sets of k elements of $\boldsymbol{\xi}_2$.) Note that if k is not known exactly, but an upper bound k^* is known, one could do an experiment with (ν_1+2k^*)

observations. When noise is present, the above results are still important, for they lead us to think in terms of "optimal search designs" with small number of runs, rather than of the so-called "optimal" designs, or of designs for which no special goodness property is claimed.

COLORADO STATE UNIVERSITY

REFERENCES

[1] Box, G. E. P. and Draper, N. R. (1959). A basis for the selection of a response design, *J. Amer. Statist. Ass.*, **54**, 622-654.

[2] Hedayat, A., Raktoe, B. L. and Federer, W. T. (1974). On a measure of aliasing due to fitting an incomplete model, *Ann. Statist.*, **2**, 650-660.

[3] Srivastava, J. N. (1975). Designs for searching nonnegligible effects, *A Survey of Statistical Design and Linear Models*, edited by J. N. Srivastava, North-Holland Publishing Company, Amsterdam, 507-519.

[4] Srivastava, J. N. (1976a). Some further theory of search linear models, *Contributions to Applied Statistics*, (ed., Walter Joh. Ziegler), Birkhäuser Vêrlag, Basel und Stuttgart.

[5] Srivastava, J. N. (1976b). Optimal search designs, or designs optimal under bias free optimality criteria, *Statistical Decision Theory and Related Topics* II, Edited by S. S. Gupta and D. S. Moore, 375-409.