Biometrika (2002), **89**, 3, *pp.* 709–718

Printed in Great Britain

An efficient design for model discrimination and parameter estimation in linear models

By ATANU BISWAS

Applied Statistics Unit, Indian Statistical Institute, 203 Barrackpore Trunk Road,
Calcutta 700035, India
atanu@isical.ac.in

AND PROBAL CHAUDHURI

Theoretical Statistics and Mathematics Unit, Indian Statistical Institute,
203 Barrackpore Trunk Road, Calcutta 700035, India
probal©isical.ac.in

Summary

We consider experimental designs in a regression set-up where the unknown regression function belongs to a known family of nested linear models. The objective of our design is to select the correct model from the family of nested models as well as to estimate efficiently the parameters associated with that model. We show that our proposed design is able to choose the true model with probability tending to one as the number of trials grows to infinity. We also establish that our selected design converges to the optimal design distribution for the true linear model ensuring asymptotic efficiency of least squares estimators of model parameters.

Some key words: Adaptive sequential design; Consistent model selection; Nested models; Optimal design; Stepwise F-test.

1. Introduction

Experimental designs for discriminating among several competing regression models have received a fair amount of attention in the existing literature; see for example Anderson (1962), Hunter & Reiner (1965), Box & Hill (1967), Hunter & Mezaki (1967), Pazman & Fedorov (1968), Froment & Mezaki (1970), Meeter et al. (1970), Atkinson (1972), Atkinson & Cox (1974), Atkinson & Fedorov (1975a,b), Hill (1978), Atkinson (1981), Spruill (1990), Dette (1994, 1995), Dette & Röder (1997), Dette & Haller (1998) and Pukelsheim (1993, Ch. 11). Efficient parameter estimation is also important. In the case of a family of competing nested linear models, one is usually interested in selecting the correct model with the smallest number of unknown parameters as well as estimating efficiently the unknown parameters associated with the true model. Given a specific linear model, efficient estimation is guaranteed if we choose an optimal design, such as a D-optimal design, an A-optimal design or an E-optimal design (Pukelsheim, 1993, Ch. 11). However, the best design for discriminating among several competing linear models may be quite different from the design that is optimal for estimating all the parameters in the unknown true model. The resulting design may even be singular for some of these models;

see also Example 3.4 in Dette (1995) and the discussion therein. Furthermore, many optimal design criteria for model discrimination depend on the specific ordering of competing models and which one of the models is the true one (Atkinson & Fedorov, 1975a,b).

In this paper we study an adaptive sequential design that will simultaneously achieve three objectives, namely selection of the correct model with the least number of unknown parameters, efficient estimation of all the parameters in that model and generation of design points so as to converge to the optimal design for that model; see also Hill et al. (1968). We restrict attention to linear models as even in this case a complete solution for our problem is not available in the existing literature. Furthermore, optimal designs for linear regression models do not depend on the unknown model parameters; this has led to an elegant and feasible solution for the problem.

2. Description of models and the adaptive sequential design

We assume that the outcome Y of the experiment conducted at X, chosen by the experimenter, satisfies the regression model

$$Y=f(X)+\varepsilon,$$

where the X is either a real variable or a vector of real variables, the unobserved residual ε has a $N(0, \sigma^2)$ distribution, and$f(.)=g_j(\beta_j, .)$ for some $1 \leqslant j \leqslant k$. Here $\{g_j(\beta_j, .), 1 \leqslant j \leqslant k\}$ is a specified family of k regression models that are linear in the parameter β_j and nested such that $g_{j-1}(\beta_{j-1}, .)$ is a special case of $g_j(\beta_j, .)$ in the sense that $g_{j-1}(\beta_{j-1}, .)$ can be obtained from $g_j(\beta_j, .)$ by assigning some specific values to some of the coordinates of the parameter vector β_j. Consequently, $g_{j-1}(\beta_{j-1}, .)$ can be viewed as a constrained version of the model $g_j(\beta_j, .)$ with linear constraints, and the dimension of β_{j-1} will be less than that of β_j.

Suppose that we can carry out at most N experiments. Let the optimal design for the jth model consist of distinct design points $\{x_{j1}, \ldots, x_{jr_j}\}$ with x_{ju} having weight w_{ju} for $1 \leqslant u \leqslant r_j$ and $j=1, 2, \ldots, k$; here $w_{ju}>0$ and $\sum_{u=1}^{r_j} w_{ju}=1$. We assume that each of the k optimal design distributions and each of these k models are such that the corresponding information matrix is nonsingular with all its eigenvalues positive. This is true for many standard optimality criteria. Out of N experiments, the first m_0 design points $X_{0,1}, \ldots, X_{0,m_0}$ are generated from the uniform mixture of k optimal design distributions corresponding to k competing models each having the same weight. This can be done in many different ways. For instance, we may select one of the k models by simple random sampling, and then generate a design point according to the optimal design distribution corresponding to the selected model. This two-stage procedure may be repeated m_0 times to generate the design points $X_{0,1}, \ldots, X_{0,m_0}$, which form the initial design distribution prior to obtaining any data. This design is sequentially updated at various stages as data become available.

After independent observations $Y_{0,1}, \ldots, Y_{0,m_0}$ are obtained from the first m_0 experiments, we carry out some statistical tests in a stepwise manner until one specific model is selected. This can be done as follows. Consider hypotheses

$$H_j\colon g_{j-1}(\beta_{j-1}, .), \text{ is the true model}$$

against

$$K_j\colon H_j \text{ is false and } g_j(\beta_j, .) \text{ is the true model,}$$

for $j=k, k-1, \ldots, 2$. Tests are carried out in reverse order starting from H_k. Let H_j be the first null hypothesis to be rejected. If none of $H_k, H_{k-1}, \ldots, H_2$ is rejected, we select the model $g_1(\beta_1, .)$. The hypothesis H_j is rejected with level α_0^j if $T_0^j < c_0^j$, where T_0^j is the F-statistic for linear models given by

$$T_0^j = \frac{\min_{\beta_{j-1}} \sum_{s=1}^{m_0} \{Y_{0,s} - g_{j-1}(\beta_{j-1}, X_{0,s})\}^2 - \min_{\beta_j} \sum_{s=1}^{m_0} \{Y_{0,s} - g_j(\beta_j, X_{0,s})\}^2}{\min_{\beta_j} \sum_{s=1}^{m_0} \{Y_{0,s} - g_j(\beta_j, X_{0,s})\}^2}$$

$$\times \frac{m_0 - d_j}{d_j - d_{j-1}},$$

and c_0^j is the corresponding cut-off point at level α_0^j. Here d_j is the dimension of the regression model $g_j(\beta_j, .)$, which is same as the dimension of the parameter vector β_j; note that, since our models are nested, we have $d_j > d_{j-1}$. At the next stage, m_1 design points $X_{1,1}, \ldots, X_{1,m_1}$ are chosen from a design distribution which will be a non-uniform mixture of all the optimal design distributions corresponding to k models. We may again adopt a two-stage sampling procedure, first choosing one of the k models by random sampling in such a way that the models rejected by the above stepwise method have weights $1/(k+1)$, and the selected model has weight $2/(k+1)$. Then a design point is generated according to the relevant optimal design. This procedure is repeated m_1 times to generate the design points $X_{1,1}, \ldots, X_{1,m_1}$, which then lead to observations $Y_{1,1}, \ldots, Y_{1,m_1}$.

The above process is repeated to obtain $N = m_0 + m_1 + \ldots + m_n$ design points $X_{r,s}$ and observations $Y_{r,s}$, where $1 \leqslant s \leqslant m_r$ and $0 \leqslant r \leqslant n$. Here $X_{r,s}$ is generated from a mixture of k different optimal design distributions, with the optimal design corresponding to the jth model $g_j(\beta_j, .)$ given weight $(1 + h_{r,j})/(k+r)$, where $h_{r,j}$ is the number of times the jth model has been selected on the basis of stepwise hypothesis testing carried out r times, that is first using m_0 data points, then using $m_0 + m_1$ data points and so on. When $m_0 + m_1 + \ldots + m_r$ data points are obtained, tests are carried out in reverse order starting from H_k, and the hypothesis H_j will be rejected with some specified level α_r^j if $T_r^j < c_r^j$. Here T_r^j is an F-statistic given by

$$T_r^j = \frac{\sum_{i=0}^{r} [\min_{\beta_{j-1}} \sum_{s=1}^{m_i} \{Y_{i,s} - g_{j-1}(\beta_{j-1}, X_{i,s})\}^2 - \min_{\beta_j} \sum_{s=1}^{m_i} \{Y_{i,s} - g_j(\beta_j, X_{i,s})\}^2]}{\sum_{i=0}^{r} [\min_{\beta_j} \sum_{s=1}^{m_i} \{Y_{i,s} - g_j(\beta_j, X_{i,s})\}^2]}$$

$$\times \frac{\sum_{i=0}^{r} (m_i - d_j)}{(r+1)(d_j - d_{j-1})},$$

and c_r^j is the corresponding cut-off point at level α_r^j. If H_j is the first null hypothesis rejected, we select model $g_j(\beta_j, .)$. If none of the hypotheses $H_k, H_{k-1}, \ldots, H_2$ is rejected, we select the model $g_1(\beta_1, .)$.

Note that the above procedure leads to a dependent set of data points $(x_{r,s}, Y_{r,s})$, for $1 \leqslant s \leqslant m_r$ and $0 \leqslant r \leqslant n$. However, the conditional distribution of $Y_{r,s}$ given all the observations and design points obtained prior to obtaining $Y_{r,s}$, depends only on $X_{r,s}$. This is discussed in more detail in § 4, where we develop some theoretical results that follow from this fact; see also results on the product form of the likelihood based on dependent data generated by adaptive sequential designs for nonlinear experiments in Chaudhuri & Mykland (1993, 1995). The form of T_r^j is computationally convenient. As the observations from the m_r trials in the rth stage of the experiment become available, we only need to compute the sums of squares based on these m_r observations and add them to the sums of squares obtained in the preceding $(r-1)$ stages.

Let us note here that many earlier authors, see for example Atkinson & Fedorov (1975a,b), have used sequential procedures that are based only on the ordering of the residual sum of squares for different regression models for choosing the best model at any stage. However, this is not very useful with our nested models as a richer model will always yield a smaller residual sum of squares. The stepwise F-test used in our sequential scheme is a device for dealing with such nested families; see also Andrews (1971).

At all stages, the design points are selected from a mixture of all optimal design distributions. As a consequence, our scheme provides a protection for each of the competing models and their associated optimal designs in small sample situations. Models that are selected more often will have more weight in subsequent stages, along lines similar to the bandit-model approach used in clinical trials for selecting the best of a set of competing treatments (Berry & Fristedt, 1985; Hardwick, 1995).

3. Some illustrative examples

3·1. *Example* 1

We consider the case $k=2$, with the following two competing models.

Model 1: $Y=\alpha+\beta X+\varepsilon$,
Model 2: $Y=\alpha+\beta X+\gamma X^2+\varepsilon$.

Suppose that X lies in the interval $[0, 1]$. The D-optimal design for Model 1 uses design points at 0 and 1 with the same weight $\frac{1}{2}$, and the D-optimal for Model 2 uses design points at 0, $\frac{1}{2}$ and 1 with the same weight $\frac{1}{3}$. The first m_0 design points are generated from a distribution that puts weights $\frac{5}{12}$, $\frac{1}{6}$ and $\frac{5}{12}$ at the points 0, $\frac{1}{2}$ and 1, which is a uniform mixture of the two optimal designs. We then compute T_0, the test statistic for testing $H:\gamma=0$ against $K:\gamma\neq 0$ from the available data. If $T_0>c_0$, Model 2 is the winner at the initial stage. Otherwise Model 1 is selected. Then the design distribution is updated, and the selected model will have $\frac{2}{3}$ weight in the new mixture distribution. At the second stage, a new set of m_1 design points are chosen from this new mixture distribution. We carry out the testing procedure after each stage. After the rth stage, with $m_0+\ldots+m_r$ samples, we compute the test statistic T_r and decide for Model 1 or Model 2. If, up to the rth stage and up to the $(m_0+\ldots+m_r)$th trial, Model 1 is selected s times and Model 2 is selected $(r+1-s)$ times, the design distribution will be a mixture of the two optimal design distributions with weights $(1+s)/(3+r)$ and $(r+2-s)/(3+r)$. In other words, the three design points 0, $\frac{1}{2}$ and 1 will have weights

$$\frac{2r+s+7}{6(3+r)},\quad \frac{r+2-s}{3(3+r)},\quad \frac{2r+s+7}{6(3+r)}$$

respectively. Define an indicator variable Z_{r+1} such that it takes values 1 or 0 according as T_{r+1} is greater than or less than c_{r+1}. Then it is clear that

$$\begin{aligned}\mathrm{pr}(X_{r+1,i}=0\,|\,\text{past data}) &= \frac{1}{r+3}\left[\frac{1}{3}\left(1+\sum_{t=0}^{r} Z_t\right)+\frac{1}{2}\left\{1+\sum_{t=0}^{r}(1-\pi_t)\right\}\right]\\ &= \frac{1}{2}-\frac{1}{6(r+3)}\left(1+\sum_{t=0}^{r} Z_t\right)\\ &= \mathrm{pr}(X_{r+1,i}=1\,|\,\text{past data}),\end{aligned}$$

$$\mathrm{pr}(X_{r+1,i}=\tfrac{1}{2}\,|\,\text{past data})=\frac{1}{3(r+3)}\left(1+\sum_{t=0}^{r} Z_t\right).$$

After taking expectations, we obtain the unconditional probabilities as

$$\mathrm{pr}(X_{r+1,i}=0)=\mathrm{pr}(X_{r+1,i}=1)=\frac{1}{2}-\frac{1}{6(r+3)}\left(1+\sum_{t=0}^{r} Z_t\right),$$

$$\mathrm{pr}(X_{r+1,i}=\tfrac{1}{2})=\frac{1}{3(r+3)}\left(1+\sum_{t=0}^{r} \pi_t\right),$$

where $\pi_t=\mathrm{pr}(T_t>c_t)$.

When the number of stages n, and hence the number of trials N, grows to infinity, we would like to have π_n tending to 0 under the hypothesis H and to 1 under the hypothesis K for a shrinking sequence of significance levels used in our stepwise tests for model selection. This will ensure that, under H, $\mathrm{pr}(X_{n,i}=x)\to\frac{1}{2}$ for $x=0, 1$, and $\mathrm{pr}(X_{n,i}=\frac{1}{2})\to 0$; and, under K, $\mathrm{pr}(X_{n,i}=x)\to\frac{1}{3}$ for $x=0, \frac{1}{2}, 1$.

3·2. *Example* 2

We now consider some models involving two covariates z_1 and z_2. We again consider the case $k=2$, with the following two competing models.

Model 1: $Y=\alpha z_1+\beta z_2+\varepsilon$,
Model 2: $Y=\alpha z_1+\beta z_2+\gamma z_1 z_2+\varepsilon$.

Suppose that both z_1 and z_2 can take only two values 0 and 1. Here the design point is a vector $X=(z_1, z_2)$, and the three possible design points are $(1, 0)$, $(0, 1)$ and $(1, 1)$. We consider the E-optimal design which puts weights $\frac{1}{2}$ on each of the points $(1, 0)$ and $(0, 1)$ for efficient estimation of the parameters of Model 1, and puts weights $\frac{1}{3}$ on each of the points $(1, 0)$, $(0, 1)$ and $(1, 1)$ for Model 2. As before, the first m_0 design points are generated from the uniform mixture of the two optimal designs which puts weights $\frac{5}{12}$, $\frac{1}{6}$ and $\frac{5}{12}$ at the points $(1, 0)$, $(0, 1)$ and $(1, 1)$. The test for $H:\gamma=0$ against $K:\gamma\neq 0$ is to be carried out at this and the subsequent stages, and the design distribution is updated accordingly. If, up to the rth stage, Model 1 is selected s times and Model 2 is selected $(r+1-s)$ times, then, as in Example 1, we have

$$\begin{aligned}\mathrm{pr}\{X_{r+1,i}=(1, 0)\,|\,\text{past data}\} &= \frac{1}{2}-\frac{1}{6(r+3)}\left(1+\sum_{t=0}^{r} Z_t\right)\\ &= \mathrm{pr}\{X_{r+1,i}=(0, 1)\,|\,\text{past data}\},\\ \mathrm{pr}\{X_{r+1,i}=(1, 1)\,|\,\text{past data}\} &= \frac{1}{3(r+3)}\left(1+\sum_{t=0}^{r} Z_t\right),\end{aligned}$$

with the definitions of the Z_t's as before. Consequently the unconditional probabilities can be written as

$$\mathrm{pr}\{X_{r+1,i}=(1, 0)\}=\mathrm{pr}\{X_{r+1,i}=(0, 1)\}=\frac{1}{2}-\frac{1}{6(r+3)}\left(1+\sum_{t=0}^{r} \pi_t\right),$$

$$\mathrm{pr}\{X_{r+1,i}=(1, 1)\}=\frac{1}{3(r+3)}\left(1+\sum_{t=0}^{r} \pi_t\right).$$

Once again, when the number of stages n grows to infinity, we would like to have π_n

tending to 0 under the hypothesis H and to 1 under the hypothesis K for a shrinking sequence of significance levels used in our stepwise tests for model selection.

4. Main results and their implications

As we have already pointed out, the response variable $Y_{r,s}$ is dependent on the previous design points and responses only through $X_{r,s}$, in the sense that once $X_{r,s}$ is chosen the response $Y_{r,s}$ depends only on it. As in a non-adaptive regression scenario, we can restrict ourselves to inference that is conditional on given design points, and this leads to the following useful theorem.

Theorem 1. *We assume that the information matrix corresponding to the model $g_j(\beta_j, .)$ based on the design points $X_{i,s}$ ($i \leqslant s \leqslant m_i$) is nonsingular for each $0 \leqslant i \leqslant r$. Then, under the null hypothesis H_j, the conditional distribution of the test statistic T_r^j given the covariates $X_{i,s}$, for $1 \leqslant s \leqslant m_i$ and $0 \leqslant i \leqslant r$, for testing H_j against K_j will be F with $(r+1)(d_j - d_{j-1})$ and $\sum_{i=0}^{r}(m_i - d_j)$ degrees of freedom, just as in the case of independent observations taken in $r+1$ groups.*

Note that the F-statistic that we consider here is different from the conventional F-statistic based on independent data points arising in the non-sequential case; the exact finite-sample distribution of the conventional F-statistic is not tractable when dependent data are generated by our adaptive sequential scheme. Even its asymptotic distribution for such dependent data is unknown to us, and it seems rather hard to obtain. Of course, in a non-sequential situation involving independent data, the standard F-test enjoys many optimal power properties, but it is not clear how its power compares with that of our F-test in adaptive sequential problems. The only claim that can be made about the standard F-statistic in this case is that it remains the relevant likelihood ratio statistic, because of the product form of the likelihood. Asymptotic results about its behaviour may therefore be available by using some of the results of Chaudhuri & Mykland (1993, 1995) and Hu (1998), for example. Since our subsequent results depend critically on the exact distribution of our new F-statistic, we have not investigated the standard F-statistic further in this paper.

For $r = 0, 1, 2, \ldots, n$ and for $j = k, k-1, \ldots, 2$, define a set of indicator variables $Z_{j,r}$ such that

$$Z_{j,r} = \begin{cases} 1 & \text{if } T_r^u \leqslant c_r^u \text{ for } u = k, \ldots, j+1 \text{ and } T_r^j > c_r^j, \\ 0 & \text{otherwise.} \end{cases}$$

Note that the above automatically implies that $Z_{l,r} = 0$ for all $l \neq j$. We also define another indicator variable $Z_{1,r}$, which takes value 1 if and only if $T_r^u \leqslant c_r^u$ for all $u = k, \ldots, 2$. Suppose that x is a design point common to optimal designs associated with b models indexed by $j_1, \ldots, j_b$ with corresponding weights $w_{j_1}, \ldots, w_{j_b}$. For $0 \leqslant r \leqslant n-1$, define

$$B_r(x) = \sum_{v=1}^{b} w_{j_v}\left(1 + \sum_{t=0}^{r} Z_{j_v,t}\right).$$

Then the conditional probability that a specified trial will be conducted at the design point x given the past data is

$$\text{pr}(X_{r,s} = x \mid \text{past data}) = \sum_{v=1}^{b} \frac{w_{j_v}}{k+r}\left(1 + \sum_{t=0}^{r-1} Z_{j_v,t}\right) = \frac{B_{r-1}(x)}{k+r}, \qquad (1)$$

where $1 \leqslant r \leqslant n$ and $1 \leqslant s \leqslant m_r$. Writing $\pi_{j,r} = E(Z_{j,r}) = \mathrm{pr}(Z_{j,r} = 1)$, we have

$$\mathrm{pr}(X_{r,s} = x) = \sum_{v=1}^{b} \frac{w_{j_v}}{k+r}\left(1 + \sum_{t=0}^{r-1} \pi_{j_v,t}\right). \tag{2}$$

The design should be such that, if the number of experiments is large, the adaptive sequential procedure will lead to the convergence of our mixture design to the optimal design corresponding to the true model. This will ensure most efficient estimation of model parameters. For our scheme, this is guaranteed by the following theorem.

THEOREM 2. *Let x be an optimal design point associated with the jth model with weight w, where $1 \leqslant j \leqslant k$. Assume that* $\max_{0 \leqslant r \leqslant n-1} \max_{1 \leqslant j \leqslant k} \alpha_r^j \to 0$ *and* $\mu_n = \min_{0 \leqslant r \leqslant n} m_r \to \infty$ *at an appropriate rate as n, and consequently N, goes to infinity. Then, for all $p_n \leqslant r \leqslant n$,* $\mathrm{pr}(X_{r,s} = x)$ *converges to w whenever the jth model is true and p_n is a sequence of positive integers such that $p_n < n$ and $p_n \to \infty$. In other words, if the jth model is true, the limiting proportion of times the design point x is chosen in the experiment with N trials is w. We will also achieve complete consistency in model selection in the sense that ultimately the correct model will be selected with probability tending to one, and the probability that an incorrect model will be selected tends to zero.*

It is natural to investigate the performance of the usual least squares estimator of the parameter vector corresponding to the correct model when data are generated by our adaptive sequential scheme. Clearly, its finite-sample distribution will be quite complicated, because of the dependent nature of the data. Nevertheless, the following theorem ensures the asymptotic efficiency of the least squares estimator in the correct model.

THEOREM 3. *Assume all the conditions of the preceding two theorems. Let $g_j(\beta_j, .) = \langle \beta_j, G_j(.) \rangle$ be the true linear regression model, where β_j and $G_j(.)$ are d_j-dimensional vectors and $\langle ., . \rangle$ denotes the usual Euclidean inner product of vectors. Consider all N data points $(X_{r,s}, Y_{r,s})$ for $1 \leqslant s \leqslant m_r$ and $0 \leqslant r \leqslant n$, and let b_j denote the least squares estimator of β_j based on these observations. Let*

$$I_{j,N} = \sum_{r=0}^{n} \sum_{s=1}^{m_r} \{G_j(X_{r,s})\}^{\mathrm{T}} \{G_j(X_{r,s})\}$$

be the $d_j \times d_j$ information matrix associated with the jth model and these data points. Then, if $N^{-1} \sum_{r=0}^{p_n} m_r \to 0$ as $n \to \infty$, we have that $N^{-1} I_{j,N} \to I_j^*$, in probability, where $I_j^* = \sum_{u=1}^{r_j} \{G_j(x_{ju})\}^{\mathrm{T}} \{G_j(x_{ju})\} w_{ju}$ is the optimal information matrix associated with the jth model. Furthermore, $N^{1/2}(b_j - \beta_j)$ asymptotically normally distributed with zero mean and $(I_j^*)^{-1}$ as the covariance matrix.

5. CONCLUDING REMARKS AND DISCUSSION

The choice of the value of m_r and the m_r design points at the rth stage of the experiment is an important issue. It should always be such that the resulting information matrix associated with the largest model, i.e. the kth model, based on those m_r design points is nonsingular. This is necessary for Theorem 1, which guarantees proper F-distributions for the test statistics used in model selection. We have described in § 2 a two-stage sampling scheme for generating the design points $X_{r1}, \ldots, X_{rm_r}$. The procedure used here assumes the knowledge of the optimal design distributions for all the competing models, and our strategy is to build the optimal design by adaptively changing the mixture of all those optimal design distributions. For a finite sample, the resulting optimal mixture might include many support points with low weights especially when there are several nested models in the family, some involving quite a few parameters. To obtain meaningful results in this situation, we need sufficiently large m_r values and accordingly a sufficiently large value of N.

The adaptive updating of the mixture of competing optimal designs depends on the choice of the significance levels α_r^j used in our stepwise F-test for model selection. However, this dependence is only through the values of the mixing proportions in the mixture distribution. The α-values have no effect on the support points of the designs, which are completely determined by the nested family of models at the outset. We have indicated in the proof of Theorem 2 some appropriate asymptotic orders for the α-values.

It would be appropriate to discuss the difference between our approach and some of the existing methodologies available in the literature for similar and related problems. Optimal designs for nested models were studied by Anderson (1962). Spruill (1990) considered similar problems, determining the optimal approximate design with respect to a maximin criterion which maximises the local power of the F-tests. Dette (1994, 1995) considered the problem of finding optimal designs for the degree of a polynomial regression. More recently Dette & Röder (1997) considered optimal discrimination designs for multi-factor experiments and Dette & Haller (1998) considered optimal designs for identification of the order of a Fourier regression. All these authors considered optimal design solely for model discrimination; the problem of efficient estimation of parameters was not considered simultaneously. For instance, the optimal discriminating design considered by Dette (1995, p. 1254) for model discrimination between a linear and a quadratic model on the interval $[-1, 1]$ puts a very small mass at 0 and divides the remaining mass equally between the points -1 and 1. However, for efficient parameter estimation, the optimal design puts equal weights at -1, 0 and 1 for the quadratic model and equal weights at -1 and 1 for the linear model.

ACKNOWLEDGEMENT

The authors are grateful to the referees and the editor for their careful reading and some valuable suggestions, which led to a substantial improvement over an earlier version of the paper.

APPENDIX

Proofs

Proof of Theorem 1. In view of the discussion preceding the statement of the theorem, it is clear that the variables $Y_{r,s}$, for $1 \leqslant s \leqslant m_r$, given the variables $Y_{i,s}$, for $1 \leqslant s \leqslant m_i$ and $0 \leqslant i \leqslant r-1$, and the design points $X_{i,s}$, for $1 \leqslant s \leqslant m_i$ and $0 \leqslant i \leqslant r$, are conditionally independently and normally distributed, and the conditional distribution of $Y_{r,s}$ depends on the past data only through $X_{r,s}$. Consequently, under H_j and given the same response variables and the design points, the sum of squares

$$\min_{b_j} \sigma^{-2} \sum_{s=1}^{m_r} \{Y_{r,s} - g_j(\beta_j, X_{r,s})\}^2$$

and the difference of sums of squares

$$\min_{b_{j-1}} \sigma^{-2} \sum_{s=1}^{m_r} \{Y_{r,s} - g_{j-1}(\beta_{j-1}, X_{r,s})\}^2 - \min_{b_j} \sigma^{-2} \sum_{s=1}^{m_r} \{Y_{r,s} - g_j(\beta_j, X_{r,s})\}^2$$

both have conditional χ^2 distributions with $m_r - d_j$ and $d_j - d_{j-1}$ degrees of freedom respectively, and they are conditionally independently distributed. Since these conditional distributions do not depend on the conditioning variables, the same will be true of the unconditional distributions. Furthermore, these sum of squares and difference of sums of squares are independently distributed, and they are independent of the past data on which conditioning was done. This can be used repeatedly to verify that the numerator and denominator of the statistic T_r^j have independent χ^2 distributions and consequently T_r^j has an F-distribution with $(r+1)(d_j - d_{j-1})$ and $\sum_{i=0}^{r}(m_i - d_j)$ degrees of freedom. □

Proof of Theorem 2. Let the α_r^j's be such that

$$\max_{0 \leqslant r \leqslant n-1} \max_{1 \leqslant j \leqslant k} \alpha_r^j \to 0, \quad \max_{0 \leqslant r \leqslant n-1} \max_{1 \leqslant j \leqslant k} \mu_n^{-1} F\left\{\alpha_r^j, (r+1)(d_j - d_{j-1}), \sum_{i=1}^{r} (m_i - d_j)\right\} \to 0$$

as $n \to \infty$, where $c_r^j = F\{\alpha_r^j, (r+1)(d_j - d_{j-1}), \sum_{i=1}^{r}(m_i - d_j)\}$ is the $100(1-\alpha_r^j)$th percentile point of the F-distribution with $(r+1)(d_j - d_{j-1})$ and $\sum_{i=1}^{r}(m_i - d_j)$ degrees of freedom. Note that it is always possible to choose m_r's and α_r^j's to satisfy these requirements. For instance, given a value of n, we can choose m_r's so that μ_n is of the order of $\log n$ and c_r^j's are of the order of $\sqrt{(\log n)}$ as $n \to \infty$. Clearly, if the jth model, corresponding to the alternative hypothesis K_j, is not true, then under the null hypothesis H_j we have

$$\pi_{j,r} = \mathrm{pr}_{H_j}(Z_{j,r} = 1) \leqslant \mathrm{pr}_{H_j}(T_r^j < c_r^j) = \alpha_r^j \to 0,$$

for all $0 \leqslant r \leqslant n-1$ as $n \to \infty$.

Let us now consider the case when the jth model, corresponding to the alternative hypothesis K_j, is true but K_l is false for all $l > j$; that is H_l is true for all $l>j$. Since the optimal design for each of the k linear models is such that the associated optimal information matrix has all of its eigenvalues positive, and since we choose that $\mu_n \to \infty$ as $n \to \infty$, the least squares estimator of β_j based on m_r observations obtained in the rth stage of the experiment is consistent for each $0 \leqslant r \leqslant n$. Furthermore, T_r^j/μ_n will remain positive and bounded away from zero in probability as $n \to \infty$. Consequently, using our assumption that $c_r^j/\mu_n \to 0$, we obtain

$$\mathrm{pr}_{K_j}(T_r^j > c_r^j) \to 1,$$

as $n \to \infty$. Observe next that

$$\begin{aligned}\pi_{j,r} = \mathrm{pr}_{K_j}(Z_{j,r} = 1) &\geqslant \sum_{u=j+1}^{k} \mathrm{pr}_{K_j}(T_r^u \leqslant c_r^u) + \mathrm{pr}_{K_j}(T_r^j > c_r^j) - (k-j)\\ &= \mathrm{pr}_{K_j}(T_r^j > c_r^j) - \sum_{u=j+1}^{k} \mathrm{pr}_{K_j}(T_r^u > c_r^u) = \mathrm{pr}_{K_j}(T_r^j > c_r^j) - \sum_{u=j+1}^{k} \alpha_r^j,\end{aligned}$$

which converges to 1 as $n \to \infty$. Suppose now that the jth model is true, and x is an optimal design point for the models indexed by the set $S = \{j_1, \ldots, j_b\}$ with corresponding weights $w_{j_1}, \ldots, w_{j_b}$. Then from (1) and (2) in § 4 and using Toeplitz's lemma we get that $\mathrm{pr}(X_{r,s} = x)$ converges to w_j or 0 according as j is a member of S or not. This completes the proof. □

Proof of Theorem 3. Convergence of $N^{-1}I_{j,N}$ to I_j^* in probability follows from Theorem 2 using martingale convergence arguments that are very similar to those used in the proof of Result 3.3 in Chaudhuri & Mykland (1995). Recall now that, for normally distributed residuals in the linear regression model, the least squares estimator b_j is also the maximum likelihood estimator of β_j. Note that, in spite of the dependent nature of the data $(X_{r,s}, Y_{r,s})$, for $1 \leqslant s \leqslant m_r$ and $0 \leqslant r \leqslant n$, the likelihood remains in the product form (Chaudhuri & Mykland, 1993, 1995). Asymptotic normality of $N^{\frac{1}{2}}(b_j - \beta_j)$ with mean zero and $(I_j^*)^{-1}$ as the dispersion matrix now follows by arguments and martingale central limit results similar to those in the proof of Result 2.6 in Chaudhuri & Mykland (1995).

References

ANDERSON, T. W. (1962). The choice of the degree of a polynomial regression as a multiple decision problem. *Ann. Math. Statist.* **33**, 255–65.

ANDREWS, D. F. (1971). Sequentially designed experiments for screening out bad models with F tests. *Biometrika* **58**, 427–32.

ATKINSON, A. C. (1972). Planning experiments to detect inadequate regression models. *Biometrika* **59**, 275–93.

ATKINSON, A. C. (1981). A comparison of two criteria for the design of experiments for discriminating between models. *Technometrics* **23**, 301–5.

ATKINSON, A. C. & COX, D. R. (1974). Planning experiments for discriminating between models (with Discussion). *J. R. Statist. Soc.* B **36**, 321–48.

ATKINSON, A. C. & FEDOROV, V. V. (1975a). The design of experiments for discriminating between two rival models. *Biometrika* **62**, 57–70.

ATKINSON, A. C. & FEDOROV, V. V. (1975b). Optimal design: experiments for discriminating between several models. *Biometrika* **62**, 289–303.

BERRY, D. A. & FRISTEDT, B. (1985). *Bandit Problems: Sequential Allocation of Experiments.* New York: Chapman and Hall.

BOX, G. E. P. & HILL, W. J. (1967). Discrimination among mechanistic models. *Technometrics* **9**, 57–71.

CHAUDHURI, P. & MYKLAND, P. A. (1993). Nonlinear experiments: optimal design and inference based on likelihood. *J. Am. Statist. Assoc.* **88**, 538–46.

CHAUDHURI, P. & MYKLAND, P. A. (1995). On efficient designing of nonlinear experiments. *Statist. Sinica* **5**, 421–40.

DETTE, H. (1994). Discrimination designs for polynomial regression on compact intervals. *Ann. Statist.* **22**, 890–903.

DETTE, H. (1995). Optimal designs for identifying the degree of a polynomial regression. *Ann. Statist.* **23**, 1248–66.

DETTE, H. & HALLER, G. (1998). Optimal designs for the identification of the order of a Fourier regression. *Ann. Statist.* **26**, 1496–1521.

DETTE, H. & RÖDER, I. (1997). Optimal discrimination designs for multifactor experiments. *Ann. Statist.* **25**, 1161–75.

FROMENT, G. F. & MEZAKI, R. (1970). Sequential discrimination and estimation procedures for rate modeling in heterogeneous catalysis. *Chem. Eng. Sci.* **25**, 293–301.

HARDWICK, J. P. (1995). A modified bandit as an approach to ethical allocation in clinical trials. In *Adaptive Designs*, IMS Lecture Notes — Monograph Series, **25**, Ed. N. Flournoy and W. F. Rosenberger, pp. 65–89. Hayward, CA: Institute of Mathematical Statistics.

HILL, P. D. H. (1978). A review of experimental design procedures for regression model discrimination. *Technometrics* **20**, 15–21.

HILL, W. J., HUNTER, W. G. & WICHERN, D. W. (1968). A joint design criterion for the dual problem of model discrimination and parameter estimation. *Technometrics* **10**, 145–60.

HU, I. (1998). On sequential designs in nonlinear problems. *Biometrika* **85**, 496–503.

HUNTER, W. G. & MEZAKI, R. (1967). An experimental design strategy for distinguishing among rival mechanistic models — an application. *Can. J. Chem. Eng.* **45**, 247–9.

HUNTER, W. G. & REINER, A. M. (1965). Designs for discriminating between two rival models. *Technometrics* **7**, 307–23.

MEETER, D., PINE, W. & BLOT, W. (1970). A comparison of two model-discrimination criteria. *Technometrics* **12**, 457–70.

PAZMAN, A. & FEDOROV, V. V. (1968). Planning of regression and discrimination experiments on NN scattering. *Soviet J. Nuclear Phys.* **6**, 619–21.

PUKELSHEIM, F. (1993). *Optimal Design of Experiments.* New York: John Wiley.

SPRUILL, M. G. (1990). Good designs for testing the degree of a polynomial mean. *Sankhyā* B **52**, 67–74.

[*Received April* 2000. *Revised October* 2001]