NASA TECHNICAL MEMORANDUM

NASA TM X-68108

NASA TM X-68108

KULLBACK-LEIBLER INFORMATION FUNCTION AND THE SEQUENTIAL SELECTION OF EXPERIMENTS TO DISCRIMINATE AMONG SEVERAL LINEAR MODELS

by Steven M. Sidik
Lewis Research Center
Cleveland, Ohio

TECHNICAL PAPER proposed for presentation at
Annual Meeting of the American Statistical Association
Montreal, Canada, August 14-17, 1972

ABSTRACT

Assume the error variance of the process, prior probabilities of the models being correct, and prior multivariate normal distributions on the parameters of the models are specified.

A rule for termination of sampling is proposed. Upon termination, the model with the largest posterior probability is chosen as correct. If sampling is not terminated, posterior probabilities of the models and posterior distributions of the parameters are computed. The next experiment chosen is that which maximizes the expected Kullback-Leibler information function. Monte-Carlo simulation experiments were performed to investigate large and small sample behavior of this sequential adaptive procedure.

E-6891

KULLBACK-LEIBLER INFORMATION FUNCTION AND THE SEQUENTIAL SELECTION OF EXPERIMENTS TO DISCRIMINATE AMONG SEVERAL LINEAR MODELS

by Steven M. Sidik

Lewis Research Center

SUMMARY

Assume that a finite set of potential linear models relating several controlled variables to an observed variable is postulated and that exactly one of these models is the true model. The problem is to sequentially design most informative experiments so that the correct model can be determined with as little experimentation as possible. We assume that the error variance of the process is known. In addition, we assume the statistician possesses prior information which can be expressed as the prior probability that each of the proposed models is indeed the correct model and prior multivariate normal distributions on the parameters of each of the postulated model equations. After each stage of sampling, the prior distributions and the observed data values are used to compute posterior probabilities of the models being the true one and posterior distributions on the parameters of the models. Then sampling is terminated if either a prespecified number of observations has been taken or if any of the posterior probabilities of the models achieves a prespecified value. Upon termination of sampling, the model with the largest posterior probability is chosen to be the correct model. If sampling is not to be terminated, the next experiment chosen is that one in the set of allowable values of the controlled variables which maximizes the expected Kullback-Leibler information function based upon the current posterior probabilities and distributions.

E-6891

An analytical study of this procedure is too complex and difficult to adequately achieve. Hence, a number of Monte-Carlo simulation experiments were performed to obtain information about the performance of this adaptive design procedure. Two basic types of Monte-Carlo experiments were performed. In the first, one of the models was chosen to be used to generate the random observations using known fixed values for the parameters. Then a large number of observations were taken using the Kullback-Leibler information functions as a criterion to choose the sequence of experiments. If was found the posterior probability of the chosen model relatively rapidly approaches the value of 1.0 and then fluctuates near 1.0. The posterior mean of the parameters of the correct model also rapidly approaches the known fixed values used to generate the observations. In the second type of experiment, one of the models was chosen to be used to generate the random observations. Then for various combinations of the maximum number of observations, stopping criterion, prior distributions of the parameters, and error variance of the process, a large number of repetitions of the sequential design procedure were executed. Then the observed probability of correct selection and average sample number were calculated based upon the number of times the procedure chose the correct model and the number of observations taken until termination.

INTRODUCTION

The general linear model has become one of the most useful statistical tools available to the modern scientific experimenter. There have been many books and papers written about techniques for choosing the appropriate or "best" linear model to fit to a set of data already

collected. In general, these have been methods of hypothesis testing to determine which of a set of specified terms in a model equation may be dropped from the model. Much work has also been done with regard to the problem of designing best or optimal experiments to estimate the parameters of specified model equations.

In this report we study a sequential adaptive experimental design procedure for a related problem. (This paper is a summary of material from Sidik (ref. 1).) Assume that a finite set of potential linear models relating a finite set of controlled variables to an observed variable is postulated and that exactly one of these models is correct. The problem is to sequentially design most informative experiments so that the correct model equation can be determined with as little experimentation as possible. We also assume that the error variance of the process is known. In addition, we assume that the statistician possesses prior information which can be expressed by the prior probability that each of the proposed models is indeed the correct model and prior multivariate normal distributions on the parameters of the various models. We then derive an adaptive procedure for designing the successive experiments using the Kullback-Leibler information function to maximize the anticipated information for discriminating among the models. That is, after each stage of sampling, the prior distributions and the observed values are used to compute posterior probabilities of the postulated models being correct and posterior distributions on the parameters of the models. Then if sampling is not to be terminated, the next experiment chosen is that which maximizes the expected Kullback-Leibler information based on the current posterior probabilities and distributions. Sampling is terminated

whenever either a prespecified number of observations is finally taken or whenever any of the posterior probabilities of the models achieves a prespecified value. Upon termination of sampling, the model with the largest posterior probability is chosen to be the correct model.

An analytical study of this procedure is too complex and difficult to adequately achieve. Hence, a number of Monte-Carlo simulation experiments were performed to obtain information about the performance of this adaptive design procedure. Two basic types of Monte-Carlo experiments were performed. In the first, one of the models was chosen to be used to generate the random observations using known fixed values for the parameters. Then a large number of observations were taken using the Kullback-Leibler information as a criterion to adaptively choose the sequence of experiments. It was found the posterior probability of the chosen model relatively rapidly approaches the value of 1.0 and then fluctuates near 1.0. The posterior mean of the parameters of the correct model also rapidly approach the known fixed values used to generate the observations. In the second type of experiment, one of the models was chosen to be used to generate the random observations using known fixed values for the parameters. Then for various combinations of the maximum number of observations, probability stopping criterion, assumed prior distributions of the parameters, and error variance of the process, a large number of repetitions of the sequential design procedure were executed. Then a probability of correct selection and average sample number were calculated based upon the number of times the procedure chose the correct model and the number of observations taken until termination.

Lindley (ref. 2) was one of the first to consider the general idea of applying information concepts to the problems of statistical inference. He modified the concept of entropy and developed a number of interesting general results on the amount of information in an experiment about the parameters of the distribution of a random variable.

Stone (ref. 3) was one of the first to consider information concepts as applied to designing and comparing regression experiments. He used a Bayesian framework, but the problem he considers is that of parameter estimation rather than that of model selection.

Another early and more relevant paper is that of Chernoff (ref. 4) who applied the Kullback-Leibler information function to the sequential design of experiments when the cost of experimenting is small. His results are valid for the case of two terminal decisions and a finite number of experiments and states of nature. These results have been generalized by Albert (ref. 5) to an infinite number of states of nature and by Bessler (ref. 6) to an infinite number of experiments and k terminal actions. Kiefer and Sacks (ref. 7) have also provided some extensions.

Hunter and Reiner (ref. 8) considered a sequential design procedure for discriminating between two model equations. Their procedure chooses the experimental conditions which, based upon maximum likelihood estimates of the parameters from the data already collected, separate the expected values of the observed variable under the two models by as much as possible.

Box and Hill (ref. 9) discussed the use of the Kullback-Leibler information function, deriving it from considerations involving the entropy

function. They consider the use of the K-L information function to sequentially discriminate among several mechanistic (nonlinear) model equations. Besides the fact that they consider nonlinear models, their approach is different from that considered here in the sense that although they do assume prior probabilities on the proposed models, and compute posterior probabilities from the observations, they assume the parameters of the model equations are known constants.

Meeter, Pirie, and Blot (ref. 10) have done a number of computer simulations comparing the methods of Chernoff and of Box and Hill. They found that the Box-Hill procedure performed quite well on the examples in comparison to Chernoff's procedure. It is interesting to note that Chernoff seems to be the only one of these authors who defined an explicit rule for terminating sampling. Although Chernoff's procedure is known to be asymptotically optimal, it is also known to require very large sample sizes.

STRUCTURE OF THE LINEAR MODELS

In the theory of the general linear statistical model, we are concerned with problems involving model equations relating K controlled variables (x_k; k = 1, . . . ,K) to an observed variable (y). The form of the model equation is required to be linear in the unknown parameters β_k. If n observations are made upon y we let x_{ik} denote the value of x_k at which the i^{th} observation is made. Thus, for the n observations the model may conveniently be written as

$$\vec{y} = M\vec{\beta} + \vec{\varepsilon} \tag{1}$$

where

$$\vec{y}' = (y_1, y_2, \ldots, y_n)$$

$$M = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & & x_{2K} \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ \cdot & \cdot & & \cdot \\ x_{n1} & x_{n2} & \cdots & x_{nK} \end{bmatrix}$$

$$\vec{\beta}' = (\beta_1, \beta_2, \ldots, \beta_K)$$

$$\vec{\varepsilon}' = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)$$

where the variances of ε_i are finite and the ε_i are uncorrelated. The matrix M is called the design matrix for the experiment consisting of the n observations. The problem of experimental design is that of choosing the x_{ik} values in some "optimal" manner.

In certain situtations in practice the experimenter can postulate several possible models involving different variables which correspond to several possible mechanistic or empirically based theories. They may lead to the various models containing different sets of x_k. There may be some overlapping of the x_k among the models or there may be none.

There are then two problems requiring solution. The first is that of choosing experiment designs which will enable the experimenter to decide which of the potential models is the correct one. Then, having chosen the model, the second problem is to estimate the parameters. The second

problem has many solutions using a variety of standard techniques. This report concerns itself with a method of designing experiments to provide information for choosing the appropriate model equation.

We assume there are L different competing model equations. These models may be combined into one large possible model equation and then the L hypothetical models are equivalent to there being L hypotheses restricting certain sets of parameters of the large model to be a priori zero. For example, we might have two controlled variables x_1 and x_2. And suppose the model equations postulated are:

$$H_1: \quad y = \beta_1^{(1)} x_1 + \varepsilon$$

$$H_2: \quad y = \beta_2^{(2)} x_2 + \varepsilon$$

$$H_3: \quad y = \beta_1^{(3)} x_1 + \beta_2^{(3)} x_2 + \varepsilon$$

where $\beta_k^{(\ell)}$ denotes the coefficient of controlled variable k in model equation ℓ. A distinction must be made between the parameters in different models because although, for example, $\beta_k^{(j)}$ and $\beta_k^{(i)}$ are coefficients of the variable x_k, their distributions need not be the same. This notation is clumsy, however, and if we implicitly accept the fact that the distributions of the $\beta_k^{(\ell)}$ depend upon the model, we may more simply rewrite the models as

$$H_1: \quad y = \beta_1 x_1 + \varepsilon$$

$$H_2: \quad y = \beta_2 x_2 + \varepsilon$$

$$H_3: \quad y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

We say that models 1 and 2 are nested within model 3. (We will find in the following work that the performance of the adaptive procedure and the behavior of the posterior distributions are quite dependent on the structure of the nesting of models.) This is equivalent to writing one model as $y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon = \vec{X}'\vec{B} + \varepsilon$ and hypothesizing

$$H_1: \ \beta_2 = 0$$

$$H_2: \ \beta_1 = 0$$

$$H_3: \ \beta_1 \neq 0, \beta_2 \neq 0$$

In this sense it is seen that the words model and hypothesis are interchangeable and will be used interchangeably in the remainder of this report. The notation we adopt is that H_ℓ claims

$$\vec{y} = M_\ell \vec{\alpha}_\ell + \vec{\varepsilon}$$

where $\vec{\alpha}_\ell$ is the appropriate $k_\ell \times 1$ vector of β's from $\vec{B}$ which appear in model ℓ, and M_ℓ is the appropriate matrix of x's.

We now precisely state the three basic distributional assumptions about the parameters and random variables of the models:

(1) The vector $\vec{\varepsilon}$ follows a multivariate normal distribution with mean $\vec{0}$ and precision matrix T. This is denoted by $\vec{\varepsilon} \sim N(\vec{0},T)$. The precision matrix is the inverse of the covariance matrix of the distribution and we assume it is known. Since T must be positive definite definite symmetric, we need only consider the special case where $T = \tau I$ since linear transformation of the y reduces all other cases to this one. Note, that this implies τ is known.

(2) For each $\ell = 1, \ldots, L$ the prior distribution of $\vec{\alpha}_\ell$ is

$$\vec{\alpha}_\ell \sim N(\vec{\mu}_{\ell,0}, \Psi_{\ell,0})$$

where $\vec{\mu}_{\ell,0}$ and $\Psi_{\ell,0}$ are known.

(3) The prior probability that the ℓ^{th} model is the correct model equation is assumed specified and denoted by $\theta_{\ell,0}$. We assume one and only one of the models is correct and hence that $\sum_{\ell=1}^{L} \theta_{\ell,0} = 1.0$.

We now describe the space A of allowable experiments in more detail. If the number of elements of $\vec{X}$ is K, then a choice of experiment $a \varepsilon A$ is composed of the number of J of observations to take and J vectors from some subset of Euclidean k-space. The J vectors specify the values of the controlled variables x_{ik}. At the j^{th} experiment or j^{th} stage of experimenting the particular choice from A is denoted a_j.

PREREQUISITE DISTRIBUTION THEORY

In the remainder of this report, much use will be made of the distribution of the observed variable, the posterior probabilities of the models, and the posterior distributions of the parameters of the model equations. We present only the notations for these distributions and define the appropriate probability density functions. The distributions are developed in reference 1 and can also be derived from results in DeGroot (ref. 11) and Raiffa and Schlaifer (ref. 12).

Let $f_\ell(\vec{y}_{j+1} | a_{j+1}, \vec{\alpha}_\ell)$ denote the density function of the vector $\vec{y}_{j+1}$ under H_ℓ when the parameter values are given by $\vec{\alpha}_\ell$ at stage $j + 1$ of sampling and the experiment is a_{j+1}. Let the probability density function of $\vec{\alpha}_\ell$ after j stages of sampling be denoted $\xi_{\ell,j}(\vec{\alpha})$.

This is a preposterior density since it serves as the posterior density of $\vec{\alpha}_\ell$ after j stages of sampling and the prior density of $\vec{\alpha}_\ell$ before the $j+1^{st}$ stage of sampling occurs.

Lemma 1: After j stages of sampling, $\vec{\alpha}_\ell$ follows a multivariate normal distribution with mean vector $\vec{\mu}_{\ell,j}$ and precision matrix $\Psi_{\ell,j}$. That is, after j stages of sampling,

$$\vec{\alpha}_\ell \sim N(\vec{\mu}_{\ell,j}, \Psi_{\ell,j})$$

where

$$\begin{aligned}\Psi_{\ell,j} &= \Psi_{\ell,j-1} + M'_{\ell,j} T M_{\ell,j} \\ &= \Psi_{\ell,0} + \sum_{i=1}^{j} M'_{\ell,i} T M_{\ell,i} \end{aligned} \tag{2}$$

and

$$\begin{aligned}\vec{\mu}_{\ell,j} &= \Psi_{\ell,j}^{-1}(M'_{\ell,j} T \vec{y}_j + \Psi_{\ell,j-1}\vec{\mu}_{\ell,j-1}) \\ &= \Psi_{\ell,j}^{-1}\left(\sum_{i=1}^{j} M'_{\ell,i} T \vec{y}_i + \Psi_{\ell,0}\vec{\mu}_{\ell,0}\right) \end{aligned} \tag{3}$$

and where $M_{\ell,i}$ denotes the design matrix specified by a_i under H_ℓ.

(For proof see ref. (1).)

We now turn to determining the distribution of $\vec{y}_{j+1}$. This is done in two stages. First, we do not know which of the models is in fact the correct one. Then for any given model, we do not know the value of $\vec{\alpha}_\ell$. Let $f_\ell(\vec{y}_{j+1} | a_{j+1}, \vec{\alpha})$ denote the distribution of $\vec{y}_{j+1}$ under H_ℓ when experiment $a_{j+1} \varepsilon A$ is performed and $\vec{\alpha}_\ell$ is specified. Since we do not know $\vec{\alpha}_\ell$ we must average this distribution over all $\vec{\alpha}_\ell$. Let

$f_\ell(\vec{y}_{j+1}|a_{j+1})$ denote the mixture of the densities $f_\ell(\vec{y}_{j+1}|a_{j+1},\vec{\alpha})$ with respect to the marginal posterior of $\vec{\alpha}_\ell$.

Lemma 2: The conditional distribution of $\vec{y}_j$ given H_ℓ and a_j is a multivariate normal distribution with mean vector $\vec{s}_{\ell,j}$ and precision matrix $R_{\ell,j}$ where

$$R_{\ell,j} = T\left[I - M_{\ell,j}(M'_{\ell,j}TM_{\ell,j} + \Psi_{\ell,j-1})^{-1}M'_{\ell,j}T\right] \tag{4}$$

$$\vec{s}_{\ell,j} = R^{-1}_{\ell,j}TM_{\ell,j}(M'_{\ell,j}TM_{\ell,j} + \Psi_{\ell,j-1})^{-1}\Psi_{\ell,j-1}\vec{\mu}_{\ell,j-1} \tag{5}$$

(For proof see ref. 1.)

Since the true model is unknown we now compute the mixture of the distributions of Lemma 2 with respect to the probabilities $\theta_{\ell,j-1}$ as

$$f(y_j|a_j) = \sum_{\ell=1}^{L} \theta_{\ell,j-1} f_\ell(\vec{y}_j|a_j) \tag{6}$$

To compute the posterior probability of each model being correct after the observation $\vec{y}_{j+1}$ is obtained, we apply Bayes theorem directly to get

$$\theta_{\ell,j+1} = \frac{f_\ell(\vec{y}_{j+1}|a_{j+1})\theta_{\ell,j}}{\sum_{k=1}^{L} f_k(\vec{y}_{j+1}|a_{j+1})\theta_{k,j}} \tag{7}$$

ENTROPY FUNCTIONS AND THE KULLBACK-LEIBLER INFORMATION FUNCTION

When comparing a number of experiments to determine which is the optimal one to perform, one must define optimal. In this report, that experiment which yields the largest expected K-L information is defined

as the optimal experiment. In particular, let $I(w,a)$ denote the expected K-L information as a function of the experiment a and the current state w (i.e., the current values of the θ_ℓ, $\vec{\mu}_\ell$, and Ψ_ℓ) of the process. This function will be specified explicitly later. In this section, we first describe how the K-L information arises from attempting to reduce the entropy of the probabilities of the models. We then develop an expression for $I(w,a)$ and finally discuss the operational meaning of the use of $I(w,a)$ from a heutristic point of view.

Development of the K-L Information Function

The problem under consideration here is that we must choose one of a set of postulated model equations. For each model we have the posterior probability $\theta_{\ell,j}$ that it is the correct one. We would like to choose experiments which cause the posterior probability of the correct model to increase most rapidly. An indirect method of accomplishing this is to choose experiments which most rapidly decrease the entropy of the set of probabilities $\theta_{\ell,j}$. The entropy is defined as

$$\mathcal{E}(w) = -\sum_{\ell=1}^{L} \theta_{\ell,j} \ln(\theta_{\ell,j})$$

It can be verified that the entropy attains a maximum when all the probabilities are equal and attains a minimum when any one of the probabilities is one and the rest are zero.

Box and Hill (ref. 9) proposed the use of the expected decrease between the entropy at the current stage of sampling and the anticipated entropy at the next stage of sampling as the criterion for selection of

experiments. They found, however, that the entropy function is quite intractable analytically and applied a well-known inequality to show the expected K-L information function provides an upper bound on the reduction of entropy. Let $\theta_i(\vec{y}|w,a)$ denote the posterior probability of model i if the value $\vec{y}$ is observed when the state was w. Let $w(\vec{y})$ denote the state of the process after observing the value $\vec{y}$ when it was in state w. Then the anticipated entropy is given by

$$E\{\mathcal{G}[w(\vec{y}),a]\} = -\int\left\{\sum_{\ell=1}^{L}\theta_\ell(\vec{y}|w,a)\ln\left[\theta_\ell(\vec{y}|w,a)\right]\right\} f(\vec{y}|w,a)d\vec{y}$$

Thus, if the current state of the sampling process is w, and the experiment $a\in A$ is performed, the expected decrease in entropy, $R(w,a)$, is defined as

$$\begin{aligned}
R(w,a) &= \mathcal{G}(w) - E\{\mathcal{G}[w(\vec{y}),a]\} \\
&= -\sum_{i=1}^{L}\theta_i\ln(\theta_i) + \int\left\{\sum_{i=1}^{L}\theta_i(\vec{y}|w,a)\ln\left[\theta_i(\vec{y}|w,a)\right]\right\} \cdot \\
&\qquad \left[\sum_{k=1}^{L}\theta_k f_k(\vec{y}|w,a)\right] d\vec{y} \\
&= -\sum_{i=1}^{L}\theta_i\ln(\theta_i) + \int\sum_{\ell=1}^{L}\theta_\ell f_\ell(\vec{y}|w,a)\ln\left[\frac{\theta_\ell f_\ell(\vec{y}|w,a)}{\sum_{k=1}^{L}\theta_k f_k(\vec{y}|w,a)}\right] d\vec{y} \\
&\le \int\sum_{\ell=1}^{L}\theta_\ell\left\{\sum_{i=1}^{L}\theta_i f_\ell(\vec{y}|w,a)\ln\left[\frac{f_\ell(\vec{y}|w,a)}{f_i(\vec{y}|w,a)}\right]\right\} d\vec{y} \qquad (8)
\end{aligned}$$

by application of the following inequality (Kullback (ref. 13), p. 15)

$$\sum_{i=1}^{L} \theta_i f_\ell(\vec{y}|w,a) \ln\left[\frac{f_\ell(\vec{y}|w,a)}{f_i(\vec{y}|w,a)}\right] \geq f_\ell(\vec{y}|w,a) \ln\left[\frac{f_\ell(\vec{y}|w,a)}{\sum_{k=1}^{L} \theta_k f_k(\vec{y}|w,a)}\right]$$

Let

$$I(w,a,i,j) = \int f_i(\vec{y}|w,a) \ln\left[\frac{f_i(\vec{y}|w,a)}{f_j(\vec{y}|w,a)}\right] d\vec{y} \tag{9}$$

We note $I(w,a,i,j)$ is defined as the expected amount of information in the observations from experiment a for discriminating against H_j in favor of H_i. Let $\mathcal{I}(w,a)$ denote the matrix whose i,j element is $I(w,a,i,j)$. Then the inequality (8) may be written as

$$R(w,a) \leq \vec{\theta}' \mathcal{I}(w,a) \vec{\theta} = I(w,a) \tag{10}$$

Meeter et al (ref. 10) proposed the following heuristic argument in favor of using $I(w,a)$. If one knew that H_i were indeed the correct hypothesis and wished to maximize the information about all H_k for $k \neq i$, then it would be natural to maximize

$$\sum_{k \neq i} \theta_k I(w,a,i,k)$$

But since H_i is assumed correct only with probability θ_i, it is equally natural to multiply the foregoing expression by θ_i and sum over i to obtain the anticipated information. But in doing this, one does end up with $I(w,a)$.

Evaluation of K-L Information Function

From Lemma 2 we have (if $\vec{y}$ is Jx1) that the density of $\vec{y}$ under H_ℓ is given by

$$f_\ell(\vec{y}|a) = (2\pi)^{-J/2}|R_\ell|^{1/2} e^{-(\vec{y}-\vec{s}_\ell)'R_\ell(\vec{y}-\vec{s}_\ell)/2}$$

Hence

$$\frac{f_m(\vec{y}|a)}{f_n(\vec{y}|a)} = |R_m|^{1/2}|R_n|^{-1/2} \frac{e^{-(\vec{y}-\vec{s}_m)'R_m(\vec{y}-\vec{s}_m)/2}}{e^{-(\vec{y}-\vec{s}_n)'R_n(\vec{y}-\vec{s}_n)/2}}$$

Moreover

$$\ln\left[\frac{f_m(\vec{y}|a)}{f_n(\vec{y}|a)}\right] = \frac{1}{2}(\ln|R_m| - \ln|R_n|) - \frac{1}{2}(\vec{y}-\vec{s}_m)'R_m(\vec{y}-\vec{s}_m) + \frac{1}{2}(\vec{y}-\vec{s}_n)'R_n(\vec{y}-\vec{s}_n) \qquad (11)$$

and

$$I(w,a,m,n) = \int \ln\left[\frac{f_m(\vec{y}|a)}{f_n(\vec{y}|a)}\right] f_m(\vec{y}|a)\,d\vec{y}$$

$$= E\left\{\ln\left[\frac{f_m(\vec{y}|a)}{f_n(\vec{y}|a)}\right]\right\} \qquad (12)$$

where the expectation is taken under the assumption $\vec{y} \sim N(\vec{s}_m, R_m)$. Note that $I(w,a,m,m) = 0.0$ for $m = 1, \ldots, L$.

It can be shown (ref. 1) that

$$I(w,a,m,n) = \frac{1}{2}\left\{\ln|R_m| - \ln|R_n|\right\} - \frac{1}{2}J + \frac{1}{2}\,tr(R_n R_m^{-1}) + \frac{1}{2}(\vec{s}_m-\vec{s}_n)'R_n(\vec{s}_m-\vec{s}_n) \qquad (13)$$

And

$$I(w,a,m,n) + I(w,a,n,m) = -J + \frac{1}{2}\left[\text{tr}\left(R_n R_m^{-1}\right) + \text{tr}\left(R_m R_n^{-1}\right)\right]$$

$$+ \frac{1}{2}\left[(\vec{s}_m - \vec{s}_n)' R_n (\vec{s}_m - \vec{s}_n) + (\vec{s}_n - \vec{s}_m)' R_m (\vec{s}_n - \vec{s}_m)\right]$$

$$= -J + \frac{1}{2}\left[\text{tr}\left(R_n R_m^{-1}\right) + \text{tr}\left(R_m R_n^{-1}\right)\right]$$

$$+ \frac{1}{2}\left[(\vec{s}_m - \vec{s}_n)'(R_m + R_n)(\vec{s}_m - \vec{s}_n)\right] \qquad (14)$$

Equation (14) is given in slightly different form in Kullback (1968, p. 190). Thus,

$$I(w,a) = \sum_{n=2}^{L}\sum_{m=1}^{n-1} \theta_n \theta_m [I(w,a,m,n) + I(w,a,n,m)]$$

$$= \sum_{n=2}^{L}\sum_{m=1}^{n-1} \theta_n \theta_m \left\{-J + \frac{1}{2}\left[\text{tr}\left(R_n R_m^{-1}\right) + \text{tr}\left(R_m R_n^{-1}\right)\right]\right.$$

$$\left. + \frac{1}{2}\left[(\vec{s}_m - \vec{s}_n)'(R_m + R_n)(\vec{s}_m - \vec{s}_n)\right]\right\}$$

$$= -J \sum_{n=2}^{L}\sum_{m=1}^{n-1} \theta_m \theta_n + \frac{1}{2}\sum_{n=1}^{L} \theta_n \text{tr}\left[\left(\sum_{m\neq n} \theta_m R_m\right) R_n^{-1}\right]$$

$$+ \frac{1}{2}\sum_{n=2}^{L}\sum_{m=1}^{n-1} \theta_n \theta_m (\vec{s}_m - \vec{s}_n)'(R_m + R_n)(\vec{s}_m - \vec{s}_n) \qquad (15)$$

The last form of this equation appears to be the most convenient for computing purposes.

Intuitive Analysis

Looking at the computing form of equation (15) it can be seen that there are three terms. The first term is $-J \sum_{n=2}^{L} \sum_{m=1}^{n} \theta_m \theta_n$. The value of this term does not depend upon a and hence has no effect upon the choice of a. From this consideration we note that computing the value of this term would not be beneficial if only one more stage of experimentation is available.

The third term of the sum is a weighted sum of the quadratic forms

$$(\vec{s}_m - \vec{s}_n)'(R_m + R_n)(\vec{s}_m - \vec{s}_n)$$

Thus, this term is a separating function in the sense that these quadratic forms will be maximized when the pairs of expected values of $\vec{y}$ under the various hypotheses are as far apart as possible in comparison to the precisions of $\vec{y}$. If the precisions of R_m and R_n are large then $\vec{s}_m$ and $\vec{s}_n$ do not need to be far apart to provide much information whereas if these precisions are small then the expected values $\vec{s}_m$ and $\vec{s}_n$ must be further apart to provide the same information. The weighting factors are the products $\theta_n \theta_m$. Thus, when θ_n and θ_m are both small, $\theta_n \theta_m$ is very small and the information due to the separation of $\vec{s}_n$ and $\vec{s}_m$ is discounted somewhat. If θ_n and θ_m are large then the information due to separation of $\vec{s}_n$ and $\vec{s}_m$ is given more importance. Thus, this third term causes experiments to be chosen which separate the expected values of $\vec{y}$ under the respective hypotheses which are still in serious contention for being chosen.

It is interesting to note that some authors (Hunter and Reiner (ref. 8), e.g.) have proposed criteria for selection of experiments involving only distances between expected values. In a later paper, Box and Hill (ref. 9) proposed that the distances as such are not important, but the distances weighted by some function of the variability about the expected values are important. It is seen here that the expected K-L information function does just that.

The second term in equation (15) is $\frac{1}{2}\sum_{n=1}^{L}\theta_n \mathrm{tr}\left[\left(\sum_{m\neq n}\theta_m R_m\right)R_n^{-1}\right]$. This can be thought of as a weighted sum of ratios of precisions. If only one y value is to be observed, this component becomes

$$\frac{1}{2}\sum_{n=1}^{L}\theta_n \frac{\sum_{m\neq n}\theta_m R_m}{R_n} \tag{16}$$

It would be interesting to see when this term is maximized. Upon taking partial derivatives of equation (16), setting to zero, and simplifying, one arrives at the following set of simultaneous nonlinear equations.

$$\sum_{k=1}^{L}\theta_i\left(\frac{R_k^2 - R_i^2}{R_i}\right) = 0 \qquad i = 1, \ldots, L$$

It can be immediately seen that one solution to this system is $R_1 = R_2 = \ldots = R_L$. This solution implies that the experiments should tend to give the same precision for the expected value of $\vec{y}$ under each hypothesis. This term is not considered any further here.

In summary, it can be seen that the expected K-L information function in this case is basically a rather simple separating function. One

would be hard pressed to construct a much simpler separating function which has more intuitive appeal.

THE SEQUENTIAL DECISION PROCEDURE

Three components are required for a sequential adaptive decision procedure; (1) a rule which determines if sampling should be terminated or continued, (2) a rule which specifies the experiment to be performed given the current state of the system, and (3) a rule which selects the model equation which will be claimed to be true when sampling is terminated.

Experiment Selection Rule

The procedure adopted for this paper is the so-called myopic procedure. This rule simply chooses as the next experiment that one which maximizes the anticipated K-L information for the next stage only.

We assume that an upper limit, J_{MAX}, to the number of observations is specified. This number may be infinite. An allocation of the observations to the stages of sampling is described by a $J_{MAX}x1$ vector $\vec{n}$, where n_i gives the number of observations at stage i. The question arises as to how the observations should be allocated. That is, should all J_{MAX} be taken at once, strictly one-at-a-time, or in different sized groups. As the first step in answering this, let A_j denote the set of experiments in A which specify that exactly j observations should be taken. For any given state w, let $a_j^*(w)$ denote the element of A_j such that

$$I[w,a_j^*(w)] = \sup_{a\epsilon A_j} I(w,a)$$

Lemma 3: For any w, and i,j such that $i > j$ we have $I[w,a_i^*(w)] \geq I[w,a_j^*(w)]$.

Proof: We introduce the following notation. Let $y_k(a_i^*)$, $k = 1, \ldots,i$ denote the random variables observed under $a_i^*(w)$ and $y_k(a_j^*)$, $k = 1, \ldots,j$ denote the random variables observed under $a_j^*(w)$. Define another experiment $\tilde{a}_i \epsilon A_i$ by choosing the first j observations according to a_j^* and the remaining $i - j$ observations according to the last $i - j$ of a_i^*. This leads to the random variables

$$\tilde{y}_k(\tilde{a}_i) = \begin{cases} y_k(a_j^*) & k = 1, \ldots,j \\ y_k(a_i^*) & k = j + 1, \ldots,i \end{cases}$$

Because $I(w,a,m,n)$ is positive definite and is additive for independent observations

$$I(w,\tilde{a}_i,m,n) \geq I(w,a_j^*,m,n)$$

Thus

$$I(w,\tilde{a}_i) = \vec{\theta}'[I(w,\tilde{a}_i,m,n)]\vec{\theta} \geq \vec{\theta}'[I(w,a_j^*,m,n)]\vec{\theta} = I(w,a_j^*)$$

But by definition $I(w,a_i^*) \geq I(w,\tilde{a}_i)$ and hence

$$I(w,a_i^*) \geq I(w,a_j^*)$$

Q.E.D.

The lemma simply proves that the optimal experiment with more observations will be expected to provide at least as much information as the optimal one with fewer observations. In determining an allocation one

should also consider the cost of experimenting. In particular, if we assume that each observation has a constant cost associated with it, then it is reasonable to choose the experiment which maximizes

$$\frac{1}{j} I(w,a_j) \qquad j = 1, \ldots, J_{MAX}$$

Thus, prior to stage k let $m = \sum_{i=1}^{k-1} n_i$ and assume $m < J_{MAX}$. The optimal experiment is the element $a^* \varepsilon A$ which for the current state w_{k-1} yields

$$j = 1, \ldots, J_{MAX} - m \left\{ \max_{a \varepsilon A_j} \frac{1}{j} I(w_{k-1},a) \right\}$$

If sampling has not been terminated by the rules developed in the next section then we stop when $\Sigma n_i = J_{MAX}$ and select the model according to the rules in the next section.

Stopping and Model Selection Rules

We now discuss the problems of determining which of the postulated models is the true one and determining when the results of the experiments are sufficiently informative to stop sampling and make the choice.

Box and Hill (ref. 9) suggested that for their procedure, experimenting be terminated whenever one model is clearly superior to the others. This is obviously a reasonable statement but it is in need of formal definition before it can be used as a stopping and selection rule.

(1) Stopping rule: Let θ_{min} be some specified value $1/L < \theta_{min} \leq 1.0$. This value is the probability stopping criterion. Let J_{MAX} denote the maximum number of observations permitted. Then terminate

sampling whenever either $\underset{i=1,L}{\text{MAX}}\{\theta_i\} \geq \theta_{min}$ or J_{MAX} observations have been taken, whichever occurs first.

(2) Model selection rule: Upon termination choose the correct model to be $H_{\ell*}$ where $\theta_{\ell*} = \underset{i=1,L}{\text{MAX}}\{\theta_i\}$.

SOME COMPUTER SIMULATION RESULTS

General Simulation Procedure

The sequential procedure proposed consisted of (1) an experiment termination rule, (2) an experiment selection rule, and (3) a model selection rule. Because of the mathematical complexity of the distributions involved (in particular, see eq. (6)) it was not feasible to analytically examine how well these rules work. The general procedure by which the Monte Carlo simulation technique was used to study performance is outlined in the following algorithm. (The FORTRAN computer program is included in ref. 1.)

1. Input:

$\vec{\mu}_{\ell,0}$	the prior means of the parameters of the models
$\Psi_{\ell,0}$	the prior precision matrices of the parameters of the models
$\theta_{\ell,0}$	the prior probabilities of the models being correct
N	the number of simulations
θ_{min}	probability stopping criterion
J_{MAX}	maximum number of observations
ℓ^*	the model which generates the observed variable (simulates choice by nature)

$\vec{\mu}^*$ values of the parameters of the true model (simulates choice by nature)

2. $n \leftarrow 0$
3. $PCS \leftarrow 0$
4. $N_j \leftarrow 0$ (for $j = 1, J_{MAX}$)
5. $j \leftarrow 0$
6. $j \leftarrow j + 1$
7. Determine optimal $a \varepsilon A$ as described in the section entitled "Experiment Selection Rule". Denote as a^* and let M_a^* denote design matrix for model ℓ^* when a^* is chosen. (All simulations in this report consider strictly one-at-a-time sampling for simplicity.)
8. $y_j \leftarrow M_a^* \vec{\mu}^*$
9. Generate a pseudo-random observation ε_j from a $N(0,\tau)$ distribution. (Simulates action by nature)
10. $y_j \leftarrow y_j + \varepsilon_j$
11. For $\ell = 1, \ldots, L$ compute $\theta_{\ell,j}$, $\Psi_{\ell,j}$, and $\vec{\mu}_{\ell,j}$ from y_j and $\theta_{\ell,j-1}$, $\Psi_{\ell,j-1}$, and $\vec{\mu}_{\ell,j-1}$ as described in the section entitled "Prerequisite Distribution Theory".
12. Find k such that $\theta_{k,j} = \underset{i}{MAX}\{\theta_{i,j}\}$
13. If $j \geq J_{MAX}$ or $\theta_{k,j} \geq \theta_{min}$ go to 14. Otherwise go to 6.
14. $N_j \leftarrow N_j + 1$
15. If $k = \ell^*$; $PCS \leftarrow PCS + 1$
16. $n \leftarrow n + 1$
17. If $n \geq N$ go to 18. Otherwise go to 5.

18. PCS ← PCS/N

19. ASN $\leftarrow \left(\sum_{i=1}^{J_{MAX}} iN_i \right) \Big/ N$

20. Stop

Upon stopping, the value of PCS is the observed probability of correctly choosing ℓ^* as the true model for the prior distributions specified when in fact the true model is given by H_{ℓ^*} and the true value of the parameters is given by $\vec{\mu}^*$. ASN gives the average sample number upon termination.

The above algorithm can be easily used for either large sample or small sample studies. For example, for large sample studies set $\theta_{min} = 1.0$ and J_{MAX} to some large number, say 100 or 500. For small sample studies set $\theta_{min} < 1.0$, J_{MAX} to some small number, and N to some larger number, say 500 or 1000. The following studies are some of the more interesting results from reference 1.

Large Sample Studies

In this section we examine the large sample properties of the posterior probabilities of the models and the posterior means of the parameter distributions. Two sets of nested polynomial models are studied. The posterior probabilities of each model, the posterior means of the parameter distributions, and the proportion of times each of the allowable values of the independent variable is chosen as optimal are tabulated for simulations of 100 and 500 observations.

The two sets of nested polynomial models have the following general form:

$$H_\ell: \quad y = \sum_{j=0}^{\ell-1} \beta_j x^j + \varepsilon, \quad \ell = 1,L$$

Two values of L are studied, and for each of these choices, two choices of $H_{\ell*}$ are made. The values of τ, $\theta_{\ell,0}$, and $\Psi_{\ell,0}$ are specified as

$$\tau = 100.0$$

$$\Psi_{\ell,0} = I$$

$$\theta_{\ell,0} = \frac{1}{L}$$

for all simulations. The values of $\vec{\mu}_{\ell,0}$ are tabulated at the tops of figures 1 and 2 and the resulting functions are graphed on the interval $x\varepsilon[-1,+1]$ at the bottoms of the respective figures. The value of τ represents a quantity known by the statistician and nature while $\Psi_{\ell,0}$, $\vec{\mu}_{\ell,0}$ and $\theta_{\ell,0}$ represent the statisticians prior information. For $L = 4$, the two choices of $H_{\ell*}$ are H_2 and H_3. For $L = 6$, the two choices of $H_{\ell*}$ are H_3 and H_5. For simplicity, the actual values of $\vec{\mu}_{\ell*,0}$ were chosen to be the values of the parameters used to generate the data for each of the four cases. That is, $\vec{\mu}_{\ell*,0} = \vec{\mu}^*$.

For these simulations, the definition of A was arbitrarily taken to be

$$A = \left\{x = -1 + \frac{2i}{9}: \quad i = 0, \ldots, 9\right\}$$

Note that sampling is strictly one observation per stage.

The simulation results are summarized in table 1 and given in further detail in tables 2 through 9. For each choice of L and ℓ^*, five simulations of 100 observations and five simulations of 500 observations were

performed. For the simulations of 500 observations, reporting the results for the first 100 observations thus gives results for a total of 10 simulations with 100 observations. For these simulations, the sample paths of the $\theta_{\ell,j}$ were printed out and the choice of $a^{(i)}$ at each stage were printed. The posterior means of the parameter distributions were printed only after the last stage. Tables 2, 4, 6, and 8 give the posterior probabilities after 100 observations and the first 100 out of 500 observations. The proportions p_i of using $a^{(i)}$ are also given. Tables 3, 5, 7, and 9 give the same information for the 500 observation simulations.

Figures 3 and 4 present typical sample paths for the posterior probability of the correct model. In figure 3, the value of $\theta_{2,j}$ is plotted for the first 250 observations of the third simulation for $L = 4$ and $\ell^* = 2$. In figure 4, the value of $\theta_{3,j}$ is plotted for the first 250 observations of the first simulation for $L = 4$ and $\ell^* = 3$. These figures illustrate the typical behavior of $\theta_{\ell^*,j}$. It fairly rapidly rises to a value of about 0.85 to 0.95 and then slowly and erratically oscillates. As discussed in reference 1, this osciallation is suspected to be because of the nested nature of the model equations.

For $L = 4$, consideration of tables 2 to 5 show that the Euclidean distance of $\vec{\mu}_{\ell,j}$ from the vector $\begin{pmatrix} \vec{\mu}^* \\ \vec{0} \end{pmatrix}$ decreases with j for values of ℓ greater than ℓ^*. This is in accord with conclusions in chapter 3 of reference 1. For $L = 6$ and $\ell^* = 3$ we again see the same behavior as evidenced by tables 6 and 7. However, for $\ell^* = 5$, an entirely different situation arises. To understand this we should note

that the model used to generate the sequential observations is

$$y = 0.5\,x + 0.1\,x^4 + \varepsilon$$

This function can be very closely approximated by a model of the form

$$y = ax + bx^2 + \varepsilon$$

over the range of x values considered. And in fact we note from tables 8 and 9 that there is a marked preference for choosing the lower degree model as indicated by $\theta_{3,j}$ becoming close to 1.0. It is also interesting to note the behavior of $\vec{\mu}_{\ell,j}$ for $\ell > 3$. We do <u>not</u>, in general, see that $\vec{\mu}_{\ell,j} \to \begin{pmatrix} \vec{\mu}_{3,j} \\ \vec{0} \end{pmatrix}$ as might be expected when H_3 is so close to being true, except for the case of $\ell = 4$. For $\vec{\mu}_5$ we note that the average posterior mean of the coefficient of x^3 is quite close to zero and the <u>sum</u> of the posterior means of the coefficients of x^2 and x^4 is quite close to 0.1. For $\vec{\mu}_6$ we note that the sums of the posterior means of the coefficients of x^2 and x^4 are close to 0.1 and the sum of the posterior means of the coefficients of x, x^3, and x^5 is close to 0.5. From these simulation studies it is not clear whether this behavior is simply because 500 observations is not a sufficiently large number to discriminate well between such nearly equivalent functions or if this behavior will persist no matter how large the number of observations.

We now turn to a discussion of the observed proportions of times the $a^{(i)}$ were chosen as the optimal experiments. From tables 2 and 3 which present the results of $L = 4$ and $\ell^* = 2$ we see that the largest p_i are for p_0, p_4, p_5, and p_9. These correspond to $x = -1$, $x = -1/9$, $x = +1/9$, and $x = +1$. Because of the discretization of the interval

(-1,+1) we might assume that the asymptotically most informative experiments were $x = -1$, $x - 0$, and $x = +1$. From tables 4 and 5 we see the largest p_i are p_0, p_2, p_7, and p_9 corresponding to $x = -1$, $x = -5/9$, $x = +5/9$, and $x = +1$. The relationship of these proportions and x points to the experimental designs which are optimal from other considerations might be interesting. For example, Kiefer and Wolfowitz (ref. 14) consider optimal designs for regression problems of a somewhat different nature. The comparison of the current results with such other works is currently being pursued but will not be reported at this time.

Small Sample Performance Studies

In this section we examine the performance of the proposed sequential procedure as measured by the PCS and ASN values. Two studies are presented of the problem of discriminating among the three models

$$H_1: \quad y = \beta_1 x_1 + \varepsilon$$

$$H_2: \quad y = \beta_2 x_2 + \varepsilon$$

$$H_3: \quad y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

The first study assumes H_3 is true and the second study assumes H_2 is true. The experiment space A is defined as

$$A = \{(x_1,x_2): \quad x_i = \pm 1; \text{ one-at-a-time sampling}\}$$

Study One - H_3 Assumed True

We study discriminating among

$$H_1: \quad y = \beta_1 x_1 + \varepsilon$$

$$H_2: \quad y = \beta_2 x_2 + \varepsilon$$

$$H_3: \quad y = \beta_1 x_1 + \beta_2 x_2 + \varepsilon$$

$$A = \{(x_1, x_2): \quad x_i = \pm 1; \text{ one-at-a-time sampling}\}$$

where

$$\Psi_{\ell,0} = I \qquad \theta_{\ell,0} = \frac{1}{3}$$

$$\vec{\mu}_{1,0} = (1.0) \qquad \vec{\mu}_{2,0} = (1.0)$$

and

$$\vec{\mu}^* = \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix}$$

Then a number of simulation experiments were performed for each combination of:

$$\tau = 0.50, 1.0, 2.0$$

$$\theta_{min} = 0.70, 0.80, 0.90$$

$$J_{MAX} = 8, 16$$

and

$$\vec{\mu}_{3,0} = \begin{pmatrix} 0.0 \\ 0.0 \end{pmatrix}, \begin{pmatrix} 0.50 \\ 0.50 \end{pmatrix}, \begin{pmatrix} 1.0 \\ 1.0 \end{pmatrix}, \begin{pmatrix} 1.5 \\ 1.5 \end{pmatrix}$$

The experiments for $J_{MAX} = 8$ used 1500 simulations and for $J_{MAX} = 16$ used 1000 simulations.

The choice of prior means deserves some comment. Figure 5 illustrates the points in (β_1,β_2) coordinate space corresponding to the prior means. The points corresponding to $\vec{\mu}_{1,0}$ and $\vec{\mu}_{2,0}$ are as close to $\vec{\mu}^*$ as possible since $\vec{\mu}_{1,0}$ is restricted to the horizontal axis and $\vec{\mu}_{2,0}$ to the vertical. The four choices for $\vec{\mu}_{3,0}$ then span a range about $\vec{\mu}^*$ and hence the resulting PCS and ASN values will indicate the importance of mis-specified prior means.

Tables 10 and 11 present the observed PCS and ASN values for the combinations of θ_{min}, τ, and $\vec{\mu}_{3,0}$.

In general, the results are about what should be expected. The PCS increases with τ and ASN decreases with τ. PCS increases as $\vec{\mu}_{3,0}$ gets closer to $\vec{\mu}^*$. We also note that in most cases, PCS increases with θ_{min} for fixed τ and $\vec{\mu}_{3,0}$. For some values of τ, however, the value of PCS increases and then decreases as θ_{min} increases. This is particulary apparent when $\vec{\mu}_{3,0} = \vec{\mu}^*$. There does not seem to be any ready explanation for this.

Study two - H_2 Assumed True

A much less extensive study of this case was made than the case of H_3 assumed true. The same model equations were postulated and we assume

$$\left.\begin{aligned} \Psi_{\ell,0} &= I \\ \theta_{\ell,0} &= \frac{1}{3} \end{aligned}\right\} \quad \ell = 1, \ldots, L$$

$$\vec{\mu}_{1,0} = (0.0)$$

$$\vec{\mu}_{3,0} = (0.0,\ 1.0)'$$

$$\vec{\mu}^* = (1.0)$$

The values of τ, θ_{min}, and $\vec{\mu}_{2,0}$ which were simulated are tabulated in table 12 along with the simulation results. Figure 6 illustrates the prior means. Only one level of J_{MAX} (=8) was considered. Also, only 500 simulations were performed for each of these cases. The results are generally the same as for H_3 true.

DISCUSSION OF RESULTS

We now make some general observations concerning the results of the simulation experiments.

First, consider the large sample results. In the context of the fact that sequential procedures are primarily developed in the hope that reliable decisions can be made with small samples rather than large samples, these results are not of primary importance. It is interesting and informative to know, however, that the procedures are consistent. Since the study of limiting posterior distributions resulting from sequentially chosen experiments is known to be an extremely difficult and delicate problem, simulation experiments may be helpful by indicating to researchers what large sample behavior is likely to be true.

In the problems studied in reference 1 it seems quite likely that when non-nested models are encountered, the posterior probability of the true hypothesis has a limiting value of unity. For those non-nested models, the posterior mean of the parameters of the true hypothesis seemed to converge to the values of the unknown parameters generating the data.

When nested models are encountered, however, the results are not as enlightening. It appears that if the posterior probability of the correct hypothesis does not achieve a limit of unity, it at least attains a large

value (in the range of 0.85 to 0.95) and then randomly fluctuates about that value. There is indication that the conjecture of Box and Hill that for certain nested models there is a distinct preference by the sequential procedure to choose the model with the smaller number of parameters is true. For instance, the polynomial study $L = 6$, $\ell^* = 5$ indicates that if a model with more parameters is true but can be approximated closely by one with fewer parameters, there is a preference for the smaller model.

This point raises another question which is especially important as regards nested models. Although these simulation results appear to support the observation of Box and Hill, that the posterior probability of the correct model rather rapidly gets close to unity (in the range of 0.85 to 0.95), it is not clear that it ever does attain unity. In fact, it is quite possible that eventually θ_{ℓ^*} fluctuates about some value which may be a function of L, the numbers of parameters in the models, and the space A. This would have quite a bit to do with the choice of θ_{min}. Too large a value would cause excessive (perhaps even infinite) sample sizes.

In examining the small sample performance simulation experiments, it is seen that PCS drops off fairly rapidly as the distance of the prior mean of the correct model from the true values of the parameters increases. This supports the conjecture of Chernoff and Meeter et al. that there may often be "initial bungling". It should be noted, however, that in all cases studied, the prior means of the competing models were all set to be as close to the true model parameter values as could be done. Thus, in a sense, these experiments can be considered to be presenting the most

unfavorable situation possible to the sequential procedure. In actual application it might be more reasonable to assume that the prior distributions of all the models are mis-specified to the same extent. This problem of "initial bungling" should also indicate that the statistician should have the prior precision matrices of the parameter distributions be as vague as the prior information permits.

One approach studied by Kiefer and Sacks (ref. 7) was to plan small initial experiments as a basis for gaining information to plan a large second experiment. An alternative not studied in this report, but which seems worthy of investigation, would be to set a lower limit, say J_{MIN}, as the minimum number of observations taken before a stopping rule is applied. The sequential procedure would use the same rule as developed for selection of experiments but large posterior probabilities on the models would be ignored until a sufficient number of observations are taken to avoid the consequences of initial bungling. This also makes sense from the point of view of obtaining parameter estimates. Surely an experimenter would not be content to terminate sampling with two or three observations even if the resulting probabilities are overwhelmingly in favor of one hypothesis unless he had extremely good prior information.

APPENDIX

LIST OF SYMBOLS

A	the space of allowable experiments
A_j	the space of allowable experiments requiring exactly j observations
ASN	average sample number
a	element of A
$a^{(i)}$	the i^{th} experiment A
a_j	experiment in A performed at the j^{th} stage of sampling
a^*	optimal experiment in A
$\vec{B}$	vector of parameters appearing in combined model equations
$E\{X\}$	expectation of the random variable X
$E\{X\|Y\}$	conditional expectation of the random variable X given the value of Y
$\mathcal{E}(w)$	entropy of the probabilities at state w
$\mathcal{E}[w(\vec{y}),a]$	entropy of the posterior probabilities if system is in state w and the value $\vec{y}$ is observed
$f_\ell(\vec{y}\|a,\vec{\alpha})$	density function of $\vec{y}$ under model ℓ when $\vec{\alpha}$ is given and experiment a is to be performed
$f_\ell(\vec{y}\|a)$	marginal density function of $\vec{y}$ under model ℓ when experiment a is to be performed
$f_\ell(\vec{y}_{j+1}\|a,\vec{\alpha})$	density function of $\vec{y}_{j+1}$ under model ℓ when a and $\vec{\alpha}$ are given
$f_\ell(\vec{y}_{j+1}\|a)$	marginal density function of $\vec{y}_{j+1}$ under model ℓ when a is given

H_ℓ	denotes hypothesis ℓ about the form of the model equation
I	identity matrix
$I(w,a)$	expected information in experiment a when state of system is w
$I(w,a,i,j)$	expected information for discriminating in favor of H_i against H_j in experiment a when state of system is w
$\mathcal{I}(w,a)$	matrix of $I(w,a,i,j)$
J_{MAX}	upper limit on total number of observations
K	number of controlled variables
ℓ	subscript denoting model equation ($\ell = 1, \ldots, L$)
L	number of model equations or hypotheses postulated
ℓ^*	true model equation subscript
M	design matrix
M_ℓ	design matrix for model ℓ
$M_{\ell,j}$	design matrix for model ℓ under experiment a_j
$N(\vec{\mu},T)$	normal distribution with mean vector $\vec{\mu}$ and precision matrix T
N	number of simulations for Monte-Carlo study
n	counter on simulations performed in algorithm
N_j	counter or number of times sampling terminates with j observations in algorithm
$\vec{n}$	vector of n_i
n_i	number of observations taken at stage i
PCS	probability of correct selection
p_i	proportion of times $a^{(i)}$ performed in a sequence of experiments

$R_{\ell,j}$ precision matrix of distribution of y_j under model ℓ

$R(w,a)$ expected reduction in entropy if experiment a is performed and state is w

$\vec{s}_{\ell,j}$ mean vector of distribution of y_j under model ℓ

T precision matrix of distribution of ε

w state of sampling system defined by values of $\theta_i, \vec{\mu}_\ell, \Psi_\ell$

$\vec{X}$ vector of x_i

$x_{i,k}$ value of x_k at i^{th} observation

$y, \vec{y}$ observed variable

$\vec{\alpha}_\ell$ vector of parameters in model equation ℓ

β_k coefficient of x_k

$\beta_k^{(\ell)}$ coefficient of x_k in model ℓ

$\vec{\varepsilon}$ vector of observation errors

θ_ℓ probability model ℓ is correct

$\theta_{\ell,j}$ posterior probability that model ℓ is correct after j stages of sampling

θ_{min} probability stopping criterion

$\vec{\mu}_\ell$ mean vector of distribution of parameters in model ℓ

μ^* values of parameters of the true model

$\vec{\mu}_{\ell,j}$ mean vector of distribution of parameters in model ℓ after j stages of sampling

$\xi_{\ell,j}(\vec{\alpha})$ density function of parameters in model ℓ after j stages of sampling

τ	precision of distribution of ε
$\Psi_{\ell,j}$	precision matrix of distribution of parameters of model ℓ after j stages of sampling
$\vec{0}$	vector of zeros
$\lvert\ \rvert$	determinant of a matrix
$\sim$	distributed as

APPENDIX C

TABLES

TABLE 1. - SUMMARY OF SIMULATION RESULTS PRESENTED IN TABLES 2 THROUGH 9

Model number	Parameter	L = 4, $\ell^* = 2$		L = 4, $\ell^* = 3$		L = 6, $\ell^* = 3$		L = 6, $\ell^* = 5$	
		100 obs	500 obs	100 obs	500 obs	100 obs	500 obs	100 obs	500 obs
1	θ_1	0	0	0	0	0	0	0	0
2	θ_2	.966	.983	0	0	0	0	.010	0
3	θ_3	.032	.016	.922	.962	.860	.877	.902	.941
4	θ_4	.002	.001	.078	.038	.109	.107	.062	.023
5	θ_5					.022	.012	.017	.029
6	θ_6					.009	.004	.009	.006
1	β_0	0.0971	0.0891	0.1322	0.1377	0.1116	0.1357	0.0278	-0.0240
2	β_0	0.1010	0.0951	0.1313	0.1384	0.1253	0.1286	0.0316	0.0382
	β_1	.4981	.5025	.2485	.2522	.2411	.2544	.5114	.4977
3	β_0	0.1013	0.0941	-0.0070	-0.0067	-0.0038	-0.0015	-0.0249	-0.0099
	β_1	.4981	.5025	.2478	.2525	.2493	.2505	.5086	.5032
	β_2	-.0007	.0021	.2578	.2634	.2157	.2544	.1179	.0981
4	β_0	0.1010	0.0943	-0.0070	-0.0335	-0.0035	-0.0015	-0.0237	-0.0097
	β_1	.5029	.4915	.2497	.2416	.2572	.2429	.5168	.5046
	β_2	-.0004	.0018	.2579	.2634	.2645	.2544	.1168	.0976
	β_3	-.0076	.0112	-.0026	.0146	-.0089	.0102	-.0083	-.0017
5	β_0					0.0040	0.0015	-0.0157	0.0001
	β_1					.2564	.2440	.5120	.5009
	β_2					.2222	.2577	-.0020	.0227
	β_3					.0032	.0088	-.0048	.0024
	β_4					.0360	-.0015	.0908	.0730
6	β_0					0.0030	-0.0037	-0.0180	-0.0021
	β_1					.2390	.2559	.5057	.4638
	β_2					.2291	.2619	.0603	.0376
	β_3					.0596	-.0373	.0438	.1340
	β_4					.0301	-.0050	.0553	.0598
	β_5					-.0534	.0345	-.0413	-.0969

The column headings give the values of L and ℓ^* and the number of observations. The row headings present the parameters whose average posterior values are given. The probabilities listed for 100 observations are the averages after five simulations of 100 observations and the values after the first 100 observations of the 500 observation simulations. The averages of the posterior parameter means are based only upon the five full simulations of 100 and 500 observations, respectively. The posterior probabilities for 500 observations are based upon five simulations of 500 observations each.

TABLE 2. - L = 4, $\ell^* = 2$

Model	Param	After 100 observations					After first 100 of 500 observations				
1	θ_1	0	0	0	0	0	0	0	0	0	0
2	θ_2	.973	.979	.974	.976	.975	.976	.976	.931	.977	.923
3	θ_3	.025	.019	.024	.023	.024	.023	.023	.063	.022	.071
4	θ_4	.002	.001	.002	.001	.002	.001	.001	.006	.001	.006
1	β_0	0.0795	0.1017	0.0906	0.1271	0.0865	*	*	*	*	*
2	β_0	0.1187	0.1017	0.0753	0.1032	0.1059	*	*	*	*	*
	β_1	.5192	.5022	.4935	.4892	.4865					
3	β_0	0.1263	0.1021	0.0682	0.1067	0.1033	*	*	*	*	*
	β_1	.5191	.5021	.4935	.4894	.4866					
	β_2	-.0152	-.0008	.0141	-.0069	.0055					
4	β_0	0.1239	0.1022	0.0688	0.1059	0.1041	*	*	*	*	*
	β_1	.4858	.5044	.4771	.5186	.5288					
	β_2	-.0126	-.0010	.0133	-.0051	.0032					
	β_3	.0351	-.0024	.0170	-.0350	-.0527					
	p_0	0.25	0.23	0.23	0.17	0.19	0.18	0.23	0.27	0.21	0.25
	p_1	0	0	0	0	0	0	0	0	0	0
	p_2	.05	.05	.03	.17	.20	.24	.06	.04	.02	.01
	p_3	.02	.03	.02	.07	0	.02	.01	.01	.03	0
	p_4	.35	.20	.14	.01	.17	.04	.26	.43	.11	.25
	p_5	.06	.20	.27	.14	.06	.12	.14	.01	.19	.24
	p_6	.01	0	.06	.09	.02	.03	.01	.01	.17	0
	p_7	.05	.05	0	.13	.21	.19	.06	.02	.01	.01
	p_8	0	0	0	0	0	0	0	0	0	0
	p_9	.21	.24	.25	.22	.15	.19	.23	.21	.26	.24

*Not recorded.

The values of the posterior probabilities and parameter means after ten simulations, of 100 observations each, of the sequential selection procedure. The last five columns are data from the first 100 observations of the 500 observation simulations tabulated in table 3. The posterior means were not recorded for these cases. Also listed are the proportions p_i of the times each $a^{(i)}$ was chosen as the optimal experiment.

TABLE 3. - L = 4, $\ell^* = 2$

Model	Param	After 500 observations				
1	θ_1	0	0	0	0	0
2	θ_2	.991	.985	.990	.991	.957
3	θ_3	.009	.015	.009	.009	.040
4	θ_4	0	0	0	0	.003
1	β_0	0.0875	0.0599	0.0905	0.0757	0.1317
2	β_0	0.0921	0.0984	0.0999	0.0942	0.0909
	β_1	.4964	.5010	.5028	.5014	.5108
3	β_0	0.0923	0.1032	0.0984	0.0937	0.0827
	β_1	.4964	.5010	.5028	.5014	.5108
	β_2	-.0005	-.0096	.0029	.0012	.0163

TABLE 3. - Continued.

Model	Param	After 500 observations				
4	β_0	0.0923	0.1022	0.0985	0.0935	0.0805
	β_1	.4948	.4886	.5043	.4882	.4817
	β_2	-.0004	-.0086	.0028	.0014	.0139
	β_3	.0016	.0126	-.0017	.0141	.0293
	p_0	0.234	0.264	0.236	0.236	0.228
	p_1	0	0	0	0	.002
	p_2	.050	.010	.056	.044	0
	p_3	.008	0	.004	.008	0
	p_4	.280	.422	.302	.306	.076
	p_5	.110	.060	.070	.088	.418
	p_6	.018	0	.026	.076	.006
	p_7	.072	.010	.088	.038	.002
	p_8	0	0	0	0	0
	p_9	.228	.226	.218	.204	.268

The values of the posterior probabilities and parameter means after 5 simulations, of 500 observations each, of the sequential selection procedure. Also listed are the proportions p_i of the times each $a^{(i)}$ was chosen as the optimal experiment.

TABLE 4. - L = 4, $\ell^* = 3$

Model	Faram	After 100 observations					After first 100 of 500 observations				
1	θ_1	0	0	0	0	0	0	0	0	0	0
2	θ_2	0	0	0	0	0	0	0	0	0	0
3	θ_3	.788	.828	.967	.966	.966	.962	.916	.966	.941	.920
4	θ_4	.212	.172	.033	.034	.034	.038	.084	.034	.059	.080
1	β_0	0.1140	0.1548	0.1260	0.1292	0.1368	*	*	*	*	*
2	β_0	0.1183	0.1515	0.1255	0.1286	0.1324	*	*	*	*	*
	β_1	.2452	.2533	.2385	.2575	.2482					
3	β_0	-0.0043	-0.0089	-0.0159	0.0104	-0.0162	*	*	*	*	*
	β_1	.2467	.2510	.2389	.2552	.2470					
	β_2	.2263	.2939	.2683	.2216	.2788					
4	β_0	-0.0045	-0.0091	-0.0158	0.0105	-0.0162	*	*	*	*	*
	β_1	.1838	.3095	.2395	.2633	.2523					
	β_2	.2268	.2945	.2681	.2215	.2788					
	β_3	.0833	-.0779	-.0009	-.0107	-.0070					
	p_0	0.18	0.17	0.17	0.17	0.17	0.18	0.17	0.17	0.17	0.17
	p_1	0	0	0	0	0	0	0	0	0	0
	p_2	.32	.31	.30	.30	.29	.32	.30	.30	.31	.30
	p_3	0	.01	0	.02	0	0	0	0	.01	.02
	p_4	0	0	.03	0	.02	0	.03	.02	0	0
	p_5	.03	.01	.02	.03	.04	.02	.02	.02	.01	.05
	p_6	0	0	.01	.02	0	.01	0	0	0	.02
	p_7	.30	.32	.30	.28	.30	.30	.31	.31	.32	.28
	p_8	0	0	0	0	0	0	0	0	0	0
	p_9	.17	.18	.17	.18	.18	.17	.17	.18	.18	.16

The values of the posterior probabilities and parameter means after 10 simulations, of 100 observations each, of the sequential selection procedure. The last 5 columns are data from the first 100 observations of the 500 observation simulations tabulated in table 5. The posterior means for these 5 cases were not tabulated. Also listed are the proportions p_i of the times each $a^{(i)}$ was chosen as the optimal experiment.

TABLE 5. - L = 4, $\ell^* = 3$

Model	Param	After 500 observations				
1	θ_1	0	0	0	0	0
2	θ_2	0	0	0	0	0
3	θ_3	.953	.982	.953	.969	.954
4	θ_4	.047	.018	.047	.031	.046
1	β_0	0.1368	0.1385	0.1383	0.1394	0.1354
2	β_0	0.1376	0.1398	0.1391	0.1387	0.1369
	β_1	.2561	.2463	.2608	.2550	.2426
3	β_0	-0.0227	0.0085	-0.0069	-0.0028	-0.0096
	β_1	.2566	.2469	.2611	.2546	.2434
	β_2	.2906	.2392	.2648	.2548	.2677

TABLE 5. - Continued.

Model	Param	After 500 observations				
4	β_0	-0.0227	0.0085	-0.0069	-0.0028	-0.0096
	β_1	.2355	.2385	.2399	.2715	.2225
	β_2	.2905	.2392	.2649	.2548	.2678
	β_3	.0281	.0112	.0281	-.0223	.0277
	p_0	0.178	0.178	0.178	0.178	0.178
	p_1	0	0	0	0	0
	p_2	.322	.320	.320	.318	.318
	p_3	0	0	0	.002	.004
	p_4	0	.006	.004	0	0
	p_5	.004	.004	.004	.002	.010
	p_6	.002	0	0	0	.004
	p_7	.318	.318	.318	.320	.312
	p_8	0	0	0	0	0
	p_9	.176	.174	.176	.180	.174

The values of the posterior probabilities and parameter means after 5 simulations, of 500 observations each, of the sequential selection procedure. Also listed are the proportions p_i of the times each $a^{(i)}$ was chosen as the optimal experiment.

TABLE 6. - L = 6, $\ell^* = 3$

Model	Param	After 100 observations					After first 100 of 500 observations				
1	θ_1	0	0	0	0	0	0	0	0	0	0
2	θ_2	0	0	0	0	0	0	0	0	0	0
3	θ_3	.9554	.9564	.7630	.9098	.9432	.2733	.9534	.9471	.9449	.9554
4	θ_4	.0378	.0365	.1756	.0738	.0469	.5534	.0400	.0444	.0457	.0378
5	θ_5	.0052	.0055	.0451	.0130	.0075	.1168	.0050	.0070	.0075	.0052
6	θ_6	.0016	.0017	.0162	.0034	.0024	.0564	.0016	.0015	.0019	.0016
1	β_0	0.1567	0.1026	0.0891	0.0839	0.1259	*	*	*	*	*
2	β_0	0.1298	0.1197	0.1192	0.1118	0.1462	*	*	*	*	*
	β_1	.2696	.2258	.2151	.2422	.2527					
3	β_0	-0.0114	0.0098	-0.0332	0.0004	0.0156	*	*	*	*	*
	β_1	.2564	.2354	.2396	.2542	.2607					
	β_2	.2882	.2282	.3290	.2329	.2463					
4	β_0	-0.0110	0.0098	-0.0317	-0.0008	0.0161	*	*	*	*	*
	β_1	.2421	.2427	.3033	.2117	.2864					
	β_2	.2877	.2282	.3268	.2346	.2453					
	β_3	.0189	-.0099	-.0846	.0562	-.0341					
5	β_0	-0.0056	0.0200	-0.0082	-0.0175	0.0314	*	*	*	*	*
	β_1	.2458	.2403	.2870	.2232	.2855					
	β_2	.2562	.1692	.1788	.3427	.1641					
	β_3	.0148	-.0080	-.0676	.0432	.0336					
	β_4	.0271	.0512	.1288	-.0966	.0694					
6	β_0	-0.0061	0.0209	-0.0130	-0.0166	0.0300	*	*	*	*	*
	β_1	.2429	.1945	.3250	.2050	.2278					
	β_2	.2583	.1767	.1921	.3418	.1767					
	β_3	.0265	.1913	-.2368	.1214	.1958					
	β_4	.0254	.0432	.1203	-.0968	.0584					
	β_5	-.0089	-.1555	.1320	-.0609	-.1739					
	p_0	0.13	0.15	0.18	0.17	0.17	0.18	0.10	0.11	0.17	0.13
	p_1	.02	.08	.03	.06	.10	.01	.04	.11	.12	.02
	p_2	.17	.24	.32	.27	.22	.32	.12	.12	.20	.17
	p_3	.08	.03	0	0	.02	0	.05	.02	.02	.08
	p_4	.06	.03	.02	.07	.03	.01	.07	.14	.07	.06
	p_5	.03	.03	.07	.03	.05	0	.11	.02	.04	.03
	p_6	0	.12	.14	.14	.02	.06	.02	.06	.07	0
	p_7	.29	.17	.11	.09	.21	.26	.16	.15	.15	.29
	p_8	.05	.02	.01	.05	.04	0	.16	.12	.04	.05
	p_9	.17	.13	.12	.12	.14	.16	.17	.15	.12	.17

*Not recorded.

The values of the posterior probabilities and parameter means after 10 simulations, of 100 observations each, of the sequential selection procedure. The last 5 columns are data from the first 100 observations of the 500 observation simulations tabulated in table 7. The posterior means were not recorded for these 5 cases. Also listed are the proportions of the times each $a^{(i)}$ was chosen as the optimal experiment.

TABLE 7. - L = 6, $\underline{\ell}^* = 3$

Model	Param	After 500 observations				
1	θ_1	0	0	0	0	0
2	θ_2	0	0	0	0	0
3	θ_3	.6046	.9812	.9722	.9746	.8526
4	θ_4	.3388	.0175	.0257	.0230	.1316
5	θ_5	.0425	.0009	.0018	.0021	.0125
6	θ_6	.0141	.0003	.0003	.0003	.0032
1	β_0	0.1321	0.1378	0.1325	0.1349	0.1413
2	β_0	0.1356	0.1278	0.1247	0.1168	0.1383
	β_1	.2434	.2616	.2541	.2549	.2581
3	β_0	-0.0026	0.0067	-0.0066	-0.0021	-0.0029
	β_1	.2446	.2556	.2486	.2466	.2571
	β_2	.2542	.2336	.2605	.2615	.2622
4	β_0	-0.0027	0.0067	-0.0067	-0.0017	-0.0030
	β_1	.2076	.2515	.2340	.2341	.2872
	β_2	.2544	.2335	.2608	.2609	.2623
	β_3	.0491	.0055	.0194	.0167	-.0399
5	β_0	-0.0113	0.0077	0.0111	0.0063	-0.0062
	β_1	.2081	.2518	.2332	.2394	.2873
	β_2	.2926	.2279	.2845	.2061	.2774
	β_3	.0486	.0052	.0203	.0101	-.0401
	β_4	-.0300	.0049	-.0200	.0496	-.0121
6	β_0	-0.0118	0.0047	-0.0100	0.0061	-0.0073
	β_1	.2164	.3128	.2432	.2386	.2682
	β_2	.2949	.2455	.2801	.2070	.2819
	β_3	.0130	-.2307	-.0225	.0134	.0405
	β_4	-.0317	-.0098	-.0168	.0490	-.0155
	β_5	.0270	.1773	.0328	-.0026	-.0618
	p_0	0.178	0.136	0.148	0.116	0.168
	p_1	.004	.064	.028	.070	.012
	p_2	.314	.206	.198	.086	.282
	p_3	.004	.036	.092	.112	.018
	p_4	.010	.014	.038	.110	.014
	p_5	0	.092	.014	.024	.016
	p_6	.022	.004	.012	.022	.002
	p_7	.296	.190	.280	.288	.304
	p_8	0	.104	.026	.014	.010
	p_9	.172	.154	.164	.158	.174

The values of the posterior probabilities and parameter means after five simulations, of 500 observations each, of the sequential selection procedure. Also listed are the proportions of the times each $a^{(i)}$ was chosen as the optimal experiment.

TABLE 8. - L = 6, $\ell^* = 5$

Model	Param	After 100 observations					After first 100 of 500 observations				
1	θ_1	0	0	0	0	0	0	0	0	0	0
2	θ_2	0	0	0	.021	.042	0	0	.011	.003	.026
3	θ_3	.945	.942	.956	.895	.848	.943	.877	.852	.904	.857
4	θ_4	.043	.043	.038	.063	.042	.048	.106	.089	.070	.076
5	θ_5	.007	.012	.005	.015	.033	.007	.013	.035	.018	.028
6	θ_6	.004	.003	.001	.006	.035	.002	.004	.013	.006	.013
1	β_0	0.1502	-0.0299	0.0231	-0.0079	0.0034	*	*	*	*	*
2	β_0	0.0356	0.0189	0.0288	0.0316	0.0431	*	*	*	*	*
	β_1	.5159	.5101	.5123	.5079	.5106					
3	β_0	-0.0348	-0.0413	-0.0412	-0.0098	0.0026	*	*	*	*	*
	β_1	.5040	.5077	.5096	.5084	.5133					
	β_2	.1467	.1265	.1478	.0837	.0850					
4	β_0	-0.0333	-0.0396	-0.0414	-0.0070	0.0026	*	*	*	*	*
	β_1	.4874	.5288	.5019	.5513	.5146					
	β_2	.1450	.1252	.1478	.0810	.0849					
	β_3	.0217	-.0261	.0099	-.0456	-.0016					
5	β_0	-0.0232	-0.0242	-0.0431	-0.0041	0.0161	*	*	*	*	*
	β_1	.4985	.5201	.5039	.5516	.4861					
	β_2	.0642	-.0063	.1616	.0184	-.1260					
	β_3	.0101	-.0185	.0074	-.0463	.0235					
	β_4	.0774	.1223	-.0131	.0605	.2068					
6	β_0	-0.0298	-0.0234	-0.0435	-0.0067	0.0135	*	*	*	*	*
	β_1	.6005	.5398	.5233	.5143	.3504					
	β_2	.1187	-.0120	.1655	.0517	-.0226					
	β_3	-.3886	-.1048	-.0708	.1797	.6034					
	β_4	.0292	.1270	-.0167	.0301	.1071					
	β_5	.3037	.0673	.0599	-.1893	-.4482					
	p_0	0.10	0.16	0.14	0.24	0.18	0.18	0.14	0.17	0.18	0.22
	p_1	.05	.06	.07	.02	.13	.25	.04	.02	.01	.01
	p_2	.08	.20	.15	.03	.03	.06	.17	.04	.25	.02
	p_3	.04	.05	.07	.07	.02	.01	.06	.01	.03	0
	p_4	.08	.13	.08	.26	.18	.16	.06	.10	.10	.05
	p_5	.08	.08	.10	.12	.19	.07	.17	.25	.06	.41
	p_6	.08	.04	.06	.02	.03	.03	0	0	.14	.01
	p_7	.11	.10	.09	.03	.05	.09	.19	.19	.05	.02
	p_8	.22	.02	.10	0	.01	.06	.03	.01	.01	0
	p_9	.16	.16	.14	.21	.18	.09	.14	.21	.17	.26

*Not recorded.

The values of the posterior probabilities and parameter means after 10 simulations, of 100 observations each, of the sequential selection procedure. The last 5 columns are data from the first 100 observations of the 500 observation simulations tabulated in table 9. Also listed are the proportions of the times each $a^{(i)}$ was chosen as the optimal experiment.

TABLE 9. - L = 6, $\ell^* = 5$

Model	Param	After 500 observations				
1	θ_1	0	0	0	0	0
2	θ_2	0	0	0	0	0
3	θ_3	.974	.882	.899	.976	.976
4	θ_4	.020	.024	.029	.021	.022
5	θ_5	.003	.075	.062	.002	.002
6	θ_6	.002	.020	.009	0	0
1	β_0	-0.1051	0.0255	-0.0926	0.0130	0.0390
2	β_0	0.0290	0.0373	0.0436	0.0351	0.0458
	β_1	.4860	.5032	.4903	.5038	.5053
3	β_0	-0.0181	-0.0137	0.0018	-0.0258	0.0065
	β_1	.5016	.5008	.5019	.5056	.5061
	β_2	.1075	.1046	.0803	.1189	.0791
4	β_0	-0.0179	-0.0142	0.0027	-0.0257	0.0066
	β_1	.5035	.4859	.5201	.5177	.4956
	β_2	.1071	.1050	.0782	.1187	.0790
	β_3	-.0025	.0198	-.0237	-.0160	.0138
5	β_0	-0.0100	0.0012	0.0159	-0.0202	0.0136
	β_1	.4959	.4940	.5016	.5169	.4962
	β_2	.0413	-.0152	-.0328	.0897	.0306
	β_3	.0061	.0081	0	-.0151	.0129
	β_4	.0653	.1185	.1116	.0241	.0455
6	β_0	-0.0166	-0.0044	0.0157	-0.0186	0.0135
	β_1	.4232	.4352	.4617	.5039	.4949
	β_2	.0891	.0094	-.0250	.0835	.0309
	β_3	.2801	.2087	.1232	.0400	.0179
	β_4	.0224	.0996	.1030	.0288	.0452
	β_5	-.2070	-.1452	-.0854	-.0430	-.0039
	p_0	0.158	0.106	0.208	0.174	0.132
	p_1	.252	.190	.314	.010	.128
	p_2	.108	.054	.016	.306	.138
	p_3	.040	.052	.014	.006	.008
	p_4	.170	.184	.100	.020	.104
	p_5	.022	.034	.114	.030	.088
	p_6	.056	0	0	.072	.002
	p_7	.074	.134	.052	.208	.172
	p_8	.036	.118	.110	.022	.092
	p_9	.084	.128	.072	.152	.132

The values of the posterior probabilities and parameter means after five simulations, of 500 observations each, of the sequential selection procedure. Also listed are the proportions of the times each $a^{(i)}$ was chosen as the optimal experiment.

TABLE 10. - SMALL SAMPLE STUDY ONE

$[\ell^* = 3, \vec{\mu}^* = (i), J_{MAX} = 8]$

θ_{min}	τ	$\vec{\mu}_{3,0}$	PCS	ASN
0.70	0.5	(0, 0)	0.133	6.36
.70	.5	(0.5, 0.5)	.458	7.15
.70	.5	(1.0, 1.0)	.544	6.82
.70	.5	(1.5, 1.5)	.446	5.89
.80	.5	(0, 0)	.173	7.50
.80	.5	(0.5, 0.5)	.468	7.78
.80	.5	(1.0, 1.0)	.531	7.52
.80	.5	(1.5, 1.5)	.460	7.24
.90	.5	(0, 0)	.229	7.98
.90	.5	(0.5, 0.5)	.479	7.92
.90	.5	(1.0, 1.0)	.513	7.81
.90	.5	(1.5, 1.5)	.439	7.76
.70	1.0	(0, 0)	.397	5.49
.70	1.0	(0.5, 0.5)	.673	5.88
.70	1.0	(1.0, 1.0)	.737	5.29
.70	1.0	(1.5, 1.5)	.621	4.84
.80	1.0	(0, 0)	.558	6.90
.80	1.0	(0.5, 0.5)	.755	6.94
.80	1.0	(1.0, 1.0)	.771	6.50
.80	1.0	(1.5, 1.5)	.700	6.22

TABLE 10. - Continued.

θ_{min}	τ	$\vec{\mu}_{3,0}$	PCS	ASN
0.90	1.0	(0, 0)	0.605	7.80
.90	1.0	(0.5, 0.5)	.765	7.45
.90	1.0	(1.0, 1.0)	.777	7.15
.90	1.0	(1.5, 1.5)	.689	7.12
.70	2.0	(0, 0)	.699	4.24
.70	2.0	(0.5, 0.5)	.871	4.03
.70	2.0	(1.0, 1,0)	.877	3.62
.70	2.0	(1.5, 1.5)	.723	3.48
.80	2.0	(0, 0)	.868	5.45
.80	2.0	(0.5, 0.5)	.962	4.99
.80	2.0	(1.0, 1.0)	.970	4.63
.80	2.0	(1.5, 1.5)	.872	4.61
.90	2.0	(0, 0)	.944	6.46
.90	2.0	(0.5, 0.5)	.967	5.66
.90	2.0	(1.0, 1.0)	.969	5.48
.90	2.0	(1.5, 1.5)	.939	5.80

*Not recorded.

Resulting PCS and ASN values for $J_{MAX} = 8$ and the combinations of θ_{min}, τ, and $\vec{\mu}_{3,0}$. Results are based upon 1500 simulations of the procedure for each combination.

TABLE 11. - SMALL SAMPLE STUDY ONE

$[\ell^* = 3, \vec{\mu}^* = (i), J_{MAX} = 16]$

θ_{min}	τ	$\vec{\mu}_{3,0}$	PCS	ASN
0.70	0.5	(0, 0)	0.354	9.48
.70	.5	(0.5, 0.5)	.665	10.7
.70	.5	(1.0, 1.0)	.723	9.63
.70	.5	(1.5, 1.5)	.555	7.38
.80	.5	(0, 0)	.508	13.6
.80	.5	(0.5, 0.5)	.761	13.3
.80	.5	(1.0, 1.0)	.806	12.3
.80	.5	(1.5, 1.5)	.661	11.8
.90	.5	(0, 0)	.574	15.5
.90	.5	(0.5, 0.5)	.752	14.6
.90	.5	(1.0, 1.0)	.800	13.9
.90	.5	(1.5, 1.5)	.710	13.8
.70	1.0	(0, 0)	.548	6.53
.70	1.0	(0.5, 0.5)	.821	6.82
.70	1.0	(1.0, 1.0)	.825	6.09
.70	1.0	(1.5, 1.5)	.637	5.36
.80	1.0	(0, 0)	.808	9.48
.80	1.0	(0.5, 0.5)	.971	9.30
.80	1.0	(1.0, 1.0)	.961	8.16
.80	1.0	(1.5, 1.5)	.865	7.86

TABLE 11. - Continued

θ_{min}	τ	$\vec{\mu}_{3,0}$	PCS	ASN
0.90	1.0	(0, 0)	0.927	12.1
.90	1.0	(0.5, 0.5)	.973	10.8
.90	1.0	(1.0, 1.0)	.964	10.1
.90	1.0	(1.5, 1.5)	.958	10.6
.70	2.0	(0, 0)	.700	4.25
.70	2.0	(0.5, 0.5)	.878	4.17
.70	2.0	(1.0, 1.0)	.855	3.59
.70	2.0	(1.5, 1.5)	.714	3.51
.80	2.0	(0, 0)	.911	5.67
.80	2.0	(0.5, 0.5)	.990	5.12
.80	2.0	(1.0, 1.0)	.988	4.84
.80	2.0	(1.5, 1.5)	.894	4.71
.90	2.0	(0, 0)	.996	7.13
.90	2.0	(0.5, 0.5)	1.00	6.25
.90	2.0	(1.0, 1.0)	1.00	5.94
.90	2.0	(1.5, 1.5)	.995	6.20

Resulting PCS and ASN values for $J_{MAX} = 16$ and the combinations of θ_{min}, τ, and $\vec{\mu}_{3,0}$. Results based upon 1000 simulations.

TABLE 12. - SMALL SAMPLE STUDY TWO

$[\ell^* = 2, \vec{\mu}^* = (i), J_{MAX} = 8]$

θ_{min}	τ	$\vec{\mu}_{2,0}$	PCS	ASN
0.70	0.5	(1.0)	0.760	7.86
.80	.5	(1.0)	.734	7.98
.90	.5	(1.0)	.740	7.98
.70	1.0	(.5)	.828	7.63
.70	1.0	(1.0)	.882	7.20
.70	1.0	(1.5)	.800	6.86
.80	1.0	(1.0)	.872	7.98
.90	1.0	(.5)	.880	7.97
.90	1.0	(1.0)	.898	7.98
.90	1.0	(1.5)	.832	7.99
.70	2.0	(1.0)	.900	5.13
.80	2.0	(1.0)	.936	7.89
.90	2.0	(1.0)	.934	7.98

The PCS and ASN values resulting from 500 simulations of the sequential procedure for each of the tabulated combinations of θ_{min}, τ, and $\vec{\mu}_{2,0}$.

REFERENCES

1. Sidik, S. M. (1972). Kullback-Leibler Information Function and the Sequential Selection of Experiments to Discriminate Among Several Linear Models. Ph.D. Thesis, Case Western Reserve University.

2. Lindley, D. V. (1956). On the Measure of the Information Provided by an Experiment. Ann. Math. Statist. 27 986-1005.

3. Stone, M. (1959). Application of a Measure of Information to the Design and Comparison of Regression Experiments. Ann. Math. Statist. 30 55-70.

4. Chernoff, H. (1959). Sequential Design of Experiments. Ann. Math. Statist. 30 755-770.

5. Albert, A. E. (1961). The Sequential Design of Experiments for Infinitely Many States of Nature. Ann. Math. Statist. 32 774-799.

6. Bessler, S. (1960). Theory and Application of the Sequential Design of Experiments, k-actions and Infinitely Many Experiments. Part I - Theory. Part II - Applications. Technical reports No. 55 and No. 56. Applied Mathematics and Statistics Laboratories, Stanford University.

7. Kiefer, J. and Sacks, J. (1963). Asymptotically Optimum Sequential Inference and Design. Ann. Math. Statist. 36 705-750.

8. Hunter, W. G. and Reiner, A. M. (1965). Designs for Discriminating Between Two Rival Models. Technometrics 7 307-324.

9. Box, G. E. P. and Hill, W. J. (1967). Discrimination Among Mechanistic Models. Technometrics 9 57-72.

10. Meeter, D., Pirie, W., and Blot, W. (1970). A Comparison of Two Model-Discriminating Criteria. Technometrics 12 457-470.

11. DeGroot, M. H. (1970). Optimal Statistical Decisions. McGraw-Hill, New York.

12. Raiffa, H. and Schlaifer, R. (1961). Applied Statistical Decision Theory. Harvard Univ. Graduate School of Business Administration.

13. Kullback, S. (1968). Information Theory and Statistics. Dover Publications, New York.

14. Kiefer, J. and Wolfowitz, J. (1959). Optimum Designs in Regression Problems. Ann. Math. Statist. 30 271-294.

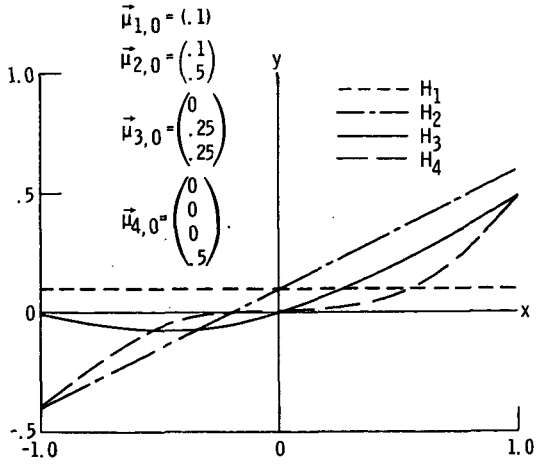

Figure 1. - Tabulations of the prior means of the parameters and graphs of the resulting functions over the interval $[-1, +1]$ for large sample polynomial study one.

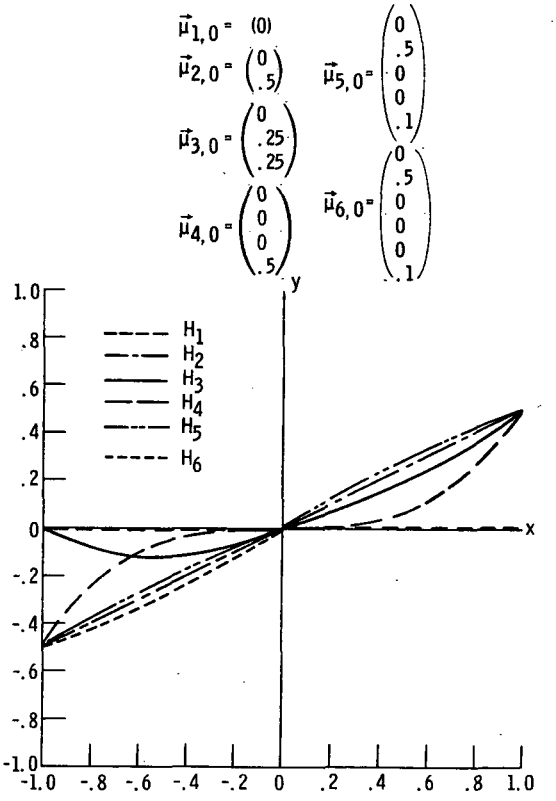

Figure 2. - Tabulations of the prior means of the parameters and graphs of the resulting functions over the interval $[-1, +1]$ for large sample polynomial study two.

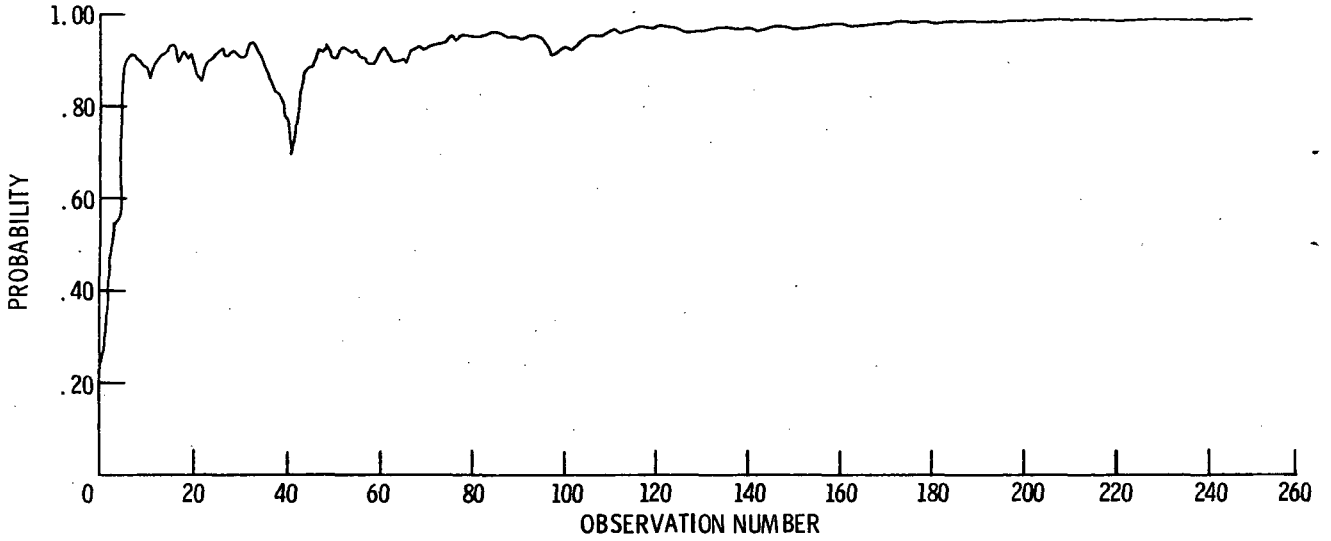

Figure 3. - The sample path of $\theta_{2,j}$ for L = 4, $\ell^* = 2$ for the first 250 observations of simulation no. 3. A well behaved path for nested models.

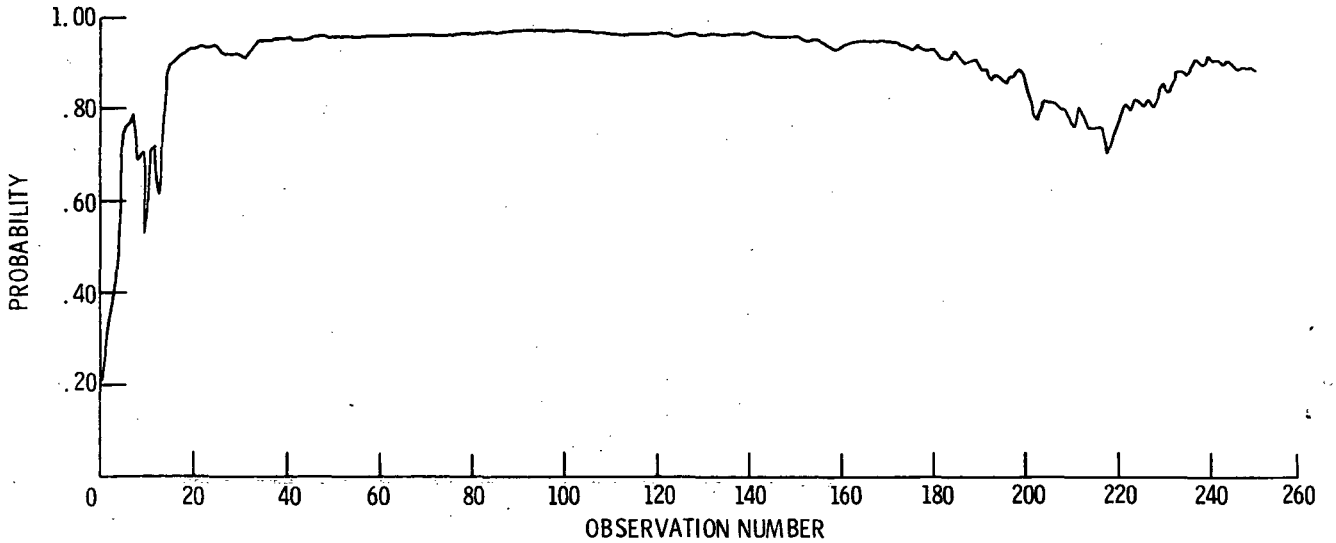

Figure 4. - The sample path of $\theta_{3,j}$ for L = 4, $\ell^* = 3$ for the first 250 observations of simulation no. 1. A typical path for nested models.

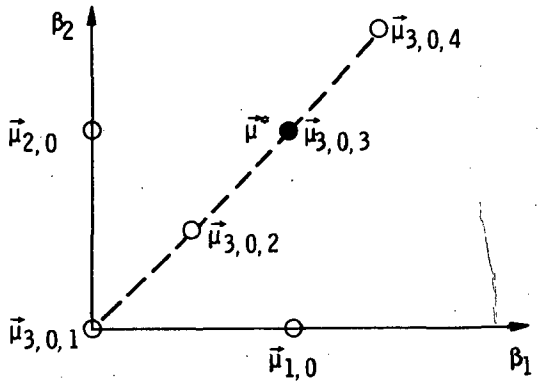

Figure 5. - Illustration of prior means for performance simulation experiment one.

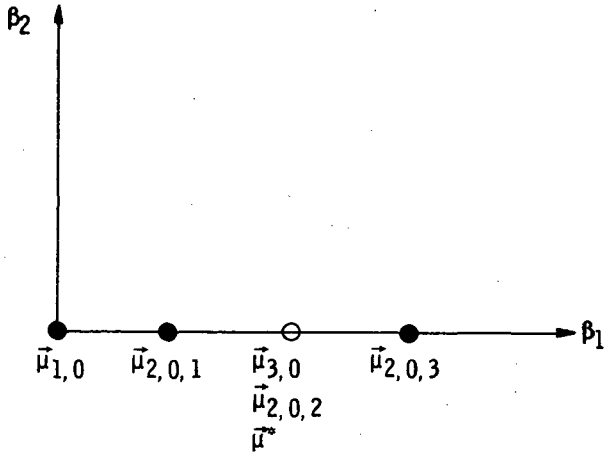

Figure 6. - Illustration of prior means for small sample performance simulation experiment two.