Statistical Science
1989, Vol. 4, No. 4, 409–435

Design and Analysis of Computer Experiments

Jerome Sacks, William J. Welch, Toby J. Mitchell and Henry P. Wynn

Abstract. Many scientific phenomena are now investigated by complex computer models or codes. A computer experiment is a number of runs of the code with various inputs. A feature of many computer experiments is that the output is deterministic—rerunning the code with the same inputs gives identical observations. Often, the codes are computationally expensive to run, and a common objective of an experiment is to fit a cheaper predictor of the output to the data. Our approach is to model the deterministic output as the realization of a stochastic process, thereby providing a statistical basis for designing experiments (choosing the inputs) for efficient prediction. With this model, estimates of uncertainty of predictions are also available. Recent work in this area is reviewed, a number of applications are discussed, and we demonstrate our methodology with an example.

Key words and phrases: Experimental design, computer-aided design, kriging, response surface, spatial statistics.

1. INTRODUCTION

Computer modeling is having a profound effect on scientific research. Many processes are so complex that physical experimentation is too time consuming or too expensive; or, as in the case of weather modeling, physical experiments may simply be impossible. As a result, experimenters have increasingly turned to mathematical models to simulate these complex systems. Advances in computational power have allowed both greater complexity and more extensive use of such models. Virtually every area of science and technology is affected. Our direct experience has been with applications in combustion, VLSI-circuit design, controlled-nuclear-fusion devices, plant ecology, and thermal-energy storage, but this is only a small sample.

Computer models (or codes) often have high-dimensional inputs, which can be scalars or functions. The output may also be multivariate. In particular, it is common for the output to be a time-dependent function from which a number of summary responses are extracted. For simplicity here, we shall assume that interest is focused on a relatively small set of scalar inputs, x, and on a single scalar response, y. Making a number of runs at various input configurations is what we call a computer experiment. The design problem is the choice of inputs for efficient analysis of the data.

The computer models we address in this article are deterministic; replicate observations from running the code with the same inputs will be identical. It is this lack of random error that makes computer experiments different from physical experiments, calling for distinct techniques.

In the next section we describe some applications. An understanding of the scientific background and objectives will be helpful in Section 3 where the role of statistics in modeling deterministic systems is discussed. This organization also parallels our research program, which has largely responded to a number of examples. Our statistical model, adopted from kriging in the spatial statistics literature and described in Section 4, treats the response as if it were a realization of a stochastic process. This provides a statistical basis for computing an efficient predictor of the response

Jerome Sacks is Professor and Head, Department of Statistics, University of Illinois, Champaign, Illinois 61820. William J. Welch is Associate Professor, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1. Toby J. Mitchell is Senior Research Staff Member, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831-8083. Henry P. Wynn is Dean, School of Mathematics, Actuarial Science and Statistics, City University, London EC1V 0HB, England.

at untried inputs and allows estimates of uncertainty of predictions. Within this framework, Section 5 discusses design criteria and algorithms for construction of designs. Applying these methods to an electronic-circuit simulator in Section 6 demonstrates what is already possible. On the other hand, one of the purposes of this paper is to highlight open problems and questions. Some of these are discussed and summarized in Section 7.

2. EXAMPLES AND OBJECTIVES

Kee, Grcar, Smooke and Miller (1985) described a fluid-dynamics model for flames which solves a complex set of partial differential equations. In an ongoing study with M. Frenklach and H. Wang, the input vector is taken to be five rate constants controlling five of the chemical reactions, and the response is the flame velocity. The numerous other inputs to the code are set at standard values based on knowledge of the chemistry. The ultimate objective here is to tune the computer model, that is find rate constants yielding a flame velocity that matches physical data. A physical analog of this experiment is impossible, because the rate constants are indeed *constants* and cannot be manipulated in reality. The need for careful design of the inputs is underlined by the fact that a single run of the code takes up to 20 minutes on a Cray X-MP.

Following Frenklach and Rabinowitz's work, Sacks, Schiller and Welch (1989) discussed examples of methane combustion based on the solution of a large system of (ordinary) differential equations arising from chemical kinetics. Although the objective is similar to that of the above flame example, the system of equations is simpler and the numerical complexity is less, allowing statistical design and analysis for a larger set of inputs.

Another important application area is quality improvement of integrated circuits. This can involve simulation of both the manufacturing process and the circuit. For example, Nassif, Strojwas and Director (1984) described the FABRICS II simulator and applied it to the processing of a ring oscillator. In these applications, the inputs are circuit parameters such as nominal transistor sizes and/or process parameters such as reagent doses, and the response might be a circuit delay time. Often, process variability is incorporated in these models by Monte Carlo sampling of a noise distribution (e.g., Singhal and Pinel, 1981). Conditional on the noise inputs, however, the simulator is deterministic. The usual objective is to find settings of the engineering or process parameters that make the response insensitive to noise, as emphasized in recent years by Taguchi (1986) and others.

Following Taguchi, the input variables x can often be divided into control factors x_{con} and noise factors x_{noise}. In a circuit-simulator example studied by Welch, Yu, Kang and Sacks (1988), the control factors were transistor dimensions and the noise factors corresponded to manufacturing-process variability. The response y was a measure of the asynchronization of two clocks, ideally zero. Generally, given a loss function $L(y)$, a "parameter design" problem can be formalized as minimizing expected loss

$$\int L[y(x_{\text{con}}, x_{\text{noise}})]\,d\Gamma(x_{\text{noise}})$$

over x_{con}. Here $\Gamma(x_{\text{noise}})$ is an assumed distribution of the noises. In the example, $L(y)$ was y^2 and Γ was approximated by a uniform distribution on five noise combinations to represent typical and extreme noise conditions.

Another example is a thermal-energy storage model, TWOLAYER, created by A. Solomon and colleagues at Oak Ridge National Laboratory. This simulates heat transfer into and out of a wall containing two layers of phase-change materials. Currin, Mitchell, Morris and Ylvisaker (1988) described a simple experiment with melting temperature and layer thickness as inputs. The response was a heat-storage-utility index, and the main objective was to determine configurations of the input parameters yielding high values of the index. The computational time for a single run, normally several minutes on a Cray X-MP, was reduced by Currin, Mitchell, Morris and Ylvisaker (1988) for the purposes of their experiment by requiring only a coarse solution to the differential equations of the computer model.

These examples illustrate that the computer experimenter, like the physical experimenter, can have many purposes in mind. We see three primary objectives:

- Predict the response at untried inputs.
- Optimize a functional of the response.
- Tune the computer code to physical data.

These objectives prompt basic statistical questions:

- *The design problem: At which input "sites"* $S = \{s_1, \cdots, s_n\}$ *should data* $y(s_1), \cdots, y(s_n)$ *be collected?*
- *The analysis problem: How should the data be used to meet the objective?*

In this article we concentrate on the prediction objective, as it is plausibly the most basic. If a sufficiently precise predictor can be found, the experimenter then has a cheap surrogate for the simulator. "What if" questions can be explored, optimization can be performed on the predictor, etc.

3. THE ROLE OF STATISTICS

These deterministic computer experiments differ substantially from the physical experiments performed by agricultural and biological scientists of the early 20th century. Their experiments had substantial random error due to variability in the experimental units. Relatively simple models were often successful. The remarkable methodology for design of experiments introduced by Fisher (1935) and the associated analysis of variance is a systematic way of separating important treatment effects from the background noise (as well as from each other). Fisher's stress on blocking, replication and randomization in these experiments reduced the effect of random error, provided valid estimates of uncertainty, and preserved the simplicity of the models.

The above deterministic examples also differ from codes in the simulation literature (e.g., Kleijnen, 1987), which incorporate substantial random error through random number generators. It has been natural, therefore, to design and analyze such stochastic simulation experiments using standard techniques for physical experiments.

Apparently, McKay, Conover and Beckman (1979) were the first to explicitly consider experimental design for deterministic computer codes. They introduced Latin hypercube sampling, an extension of stratified sampling which ensures that each of the input variables has all portions of its range represented. Latin hypercubes are computationally cheap to generate and can cope with many input variables. These designs are aimed at an objective different from those we discussed in Section 2: namely, how a known distribution of the inputs propagates through to the output distribution. (Of course, conditional on the inputs, the output is still deterministic.) For this purpose, Iman and Helton (1988) compared Latin hypercube sampling with Monte Carlo sampling of a response surface replacement for the computer model. The response surface was fitted by least squares to data from a fractional-factorial design. They found in a number of examples that the response surface could not adequately represent the complex output of the computer code but could be useful for ranking the importance of the input variables. Because Latin hypercube sampling exercises the code over the entire range of each input variable, it can also be a systematic way of discovering scientifically surprising behavior, as noted in Iman and Helton (1988).

In the absence of independent random errors, the rationale for least-squares fitting of a response surface is not clear. Of course, least squares can be viewed as curve fitting and not necessarily employing or relying on the assumption that the departures (differences between the response and the regression model) behave like white noise. The usual problem of choosing the regression model is compounded if the response is complex. Moreover, the fit will not generally interpolate the observed data (where the function is known exactly) unless there are as many estimable coefficients in the regression as there are runs.

Despite some similarities to physical experiments, then, the lack of random (or replication) error leads to important distinctions. In deterministic computer experiments:

- The adequacy of a response-surface model fitted to the observed data is determined solely by systematic bias.
- The absence of random error allows the complexity of the computer model to emerge.
- Usual measures of uncertainty derived from least-squares residuals have no obvious statistical meaning. Though deterministic measures of uncertainty are available (e.g., $\max|\hat{y}(x) - y(x)|$ over x and a class of y's), they may be very difficult to compute.
- Classical notions of experimental unit, blocking, replication and randomization are irrelevant.

While the pioneering work of Box and Draper (1959) has relevance to the first of these points, it is unclear that current methodologies for the design and analysis of physical experiments [e.g., Box and Draper, 1987; Box, Hunter and Hunter, 1978; Fisher (1935); and Kiefer (1985)] are ideal for complex, deterministic computer models. Lest the reader wonder whether statistics has *any* role here, we assert that:

- The selection of inputs at which to run a computer code is still an experimental design problem.
- Statistical principles and attitudes to data analysis are helpful however the data are generated.
- There is uncertainty associated with predictions from fitted models, and the quantification of uncertainty is a statistical problem.
- Modeling a computer code as if it were a realization of a stochastic process, the approach taken below, gives a basis for the quantification of uncertainty and a statistical framework for design and analysis.

4. MODELING AND PREDICTION

This section discusses models for computer experiments and efficient prediction. Experimental design for this predictor is the subject of the next section.

The model we adopt here treats the deterministic response $y(x)$ as a realization of a random function (stochastic process), $Y(x)$, that includes a regression model,

$$(1) \qquad Y(x) = \sum_{j=1}^{k} \beta_j f_j(x) + Z(x).$$

The random process $Z(\cdot)$ is assumed to have mean zero and covariance

$$V(w, x) = \sigma^2 R(w, x)$$

between $Z(w)$ and $Z(x)$, where σ^2 is the process variance and $R(w, x)$ is the correlation. One rationale is that departures of the complex response from the simple regression model, though deterministic, may resemble a sample path of a (suitably chosen) stochastic process $Z(\cdot)$. Alternatively, $Y(\cdot)$ in (1) may be regarded as a Bayesian prior on the true response functions, with the β's either specified a priori or given a prior distribution.

The use of a stochastic process as a prior on a class of functions has a long history. Diaconis (1988) gave an interesting account of early uses (back to H. Poincaré in the 19th century) in one-dimensional interpolation and integration. Suldin (1959, 1960) used Brownian motion and integrals of Brownian motion to develop quadrature formulas in one dimension. Sacks and Ylvisaker (1970) independently considered the same problem for a wider class of processes, and the Brownian motion model has re-emerged in Smale (1985). Corresponding efforts in d dimensions began in Ylvisaker (1975). See Ylvisaker (1987) for a more recent discussion. Sacks and Ylvisaker (1985) used models of the form (1) with added independent measurement error for one-dimensional (physical) experimental design and analysis. Sacks, Schiller and Welch (1989) employed such models (without measurement error) for prediction in computer experiments with multi-dimensional inputs.

One method of analysis for such models is known as kriging (Matheron, 1963). Given a design $S = \{s_1, \cdots, s_n\}$ and data $y_S = \{y(s_1), \cdots, y(s_n)\}'$, consider the linear predictor

$$\hat{y}(x) = c'(x) y_S$$

of $y(x)$ at an untried x. Taking a classical frequentist stance, we can replace y_S by the corresponding random quantity $Y_s = [Y(s_1), \cdots, Y(s_n)]'$, treat $\hat{y}(x)$ as random, and compute the mean squared error of this predictor averaged over the random process. The best linear unbiased predictor (BLUP) is obtained by choosing the $n \times 1$ vector $c(x)$ to minimize

$$\text{(2)} \qquad \text{MSE}[\hat{y}(x)] = E[c'(x)Y_S - Y(x)]^2$$

subject to the unbiasedness constraint

$$\text{(3)} \qquad E[c'(x)Y_S] = E[Y(x)].$$

Alternatively, a Bayesian approach would predict $y(x)$ by

$$\text{(4)} \qquad \hat{y}(x) = E[Y(x) \mid y_S],$$

the posterior mean. The frequentist and Bayesian viewpoints will generally lead to different methods and results, except in the special case of a Gaussian process for $Z(\cdot)$ and improper uniform priors on the β's. It is an old result that the BLUP in the Gaussian case is the limit of the Bayes predictor as the prior variances on the β's tend to infinity (e.g., Parzen, 1963, Section 6).

Kimeldorf and Wahba (1970) investigated classes of prior processes for which the Bayes estimate (4) is a smoothing spline. Blight and Ott (1975) used a stochastic process as a Bayesian prior on the departure function for one-dimensional x. Steinberg (1985) and Young (1977) mitigated the effects of model inadequacy by representing $y(x)$ as a polynomial of arbitrarily-high or infinite degree and assigning a Bayesian prior to the coefficients. O'Hagan (1978, Section 3) formulated a general Bayesian approach, in which the prior on $y(x)$ is a general multidimensional Gaussian process. For a more detailed discussion of the Bayesian viewpoint applied to computer experiments see Currin, Mitchell, Morris and Ylvisaker (1988).

In this article, we shall focus mainly on the kriging predictor, partly for ties with methodology in use in other areas and partly to simplify the exposition. Moreover, the use of Gaussian spatial processes provides a bridge to the Bayesian viewpoint. Where the Bayesian view provides additional insight, however, it will be mentioned.

To give some technical details connected with implementing the BLUP of the response at an untried input we use the notation

$$f(x) = [f_1(x), \cdots, f_k(x)]'$$

for the k functions in the regression,

$$F = \begin{pmatrix} f'(s_1) \\ \vdots \\ f'(s_n) \end{pmatrix}$$

for the $n \times k$ expanded design matrix,

$$R = \{R(s_i, s_j)\}, \quad 1 \le i \le n;\ 1 \le j \le n,$$

for the $n \times n$ matrix of stochastic-process correlations between Z's at the design sites, and

$$r(x) = [R(s_1, x), \cdots, R(s_n, x)]'$$

for the vector of correlations between the Z's at the design sites and an untried input x. With these definitions, the MSE (2) is

$$\text{(5)} \qquad \sigma^2[1 + c'(x)Rc(x) - 2c'(x)r(x)],$$

and the unbiasedness constraint (3) is $F'c(x) = f(x)$. Introducing Lagrange multipliers $\lambda(x)$ for the constrained minimization of the MSE, the coefficient $c(x)$ of the BLUP must satisfy

$$\text{(6)} \qquad \begin{pmatrix} 0 & F' \\ F & R \end{pmatrix} \begin{pmatrix} \lambda(x) \\ c(x) \end{pmatrix} = \begin{pmatrix} f(x) \\ r(x) \end{pmatrix}.$$

Then, by inverting the partitioned matrix, the BLUP can be written as

$$(7) \qquad \hat{y}(x) = f'(x)\hat{\beta} + r'(x)R^{-1}(Y_S - F\hat{\beta}),$$

where $\hat{\beta} = (F'R^{-1}F)^{-1}F'R^{-1}Y_S$ is the usual generalized least-squares estimate of β. Under the model, the two terms on the right of (7) are uncorrelated, and the second can be interpreted as a smooth of the residuals. Therefore, the fit can be viewed as two stages: obtain the generalized least-squares predictor and then interpolate the residuals as if there were no regression model.

A convenient representation for the MSE (2) is obtained by substituting (6) in (5) to give

$$(8) \qquad \begin{aligned} &\mathrm{MSE}[\hat{y}(x)] \\ &= \sigma^2\left[1 - (f'(x)\ r'(x))\begin{pmatrix} 0 & F' \\ F & R \end{pmatrix}^{-1}\begin{pmatrix} f(x) \\ r(x) \end{pmatrix}\right]. \end{aligned}$$

Equations (7) and (8) are also the limiting posterior mean and variance of $Y(x)$ when a diffuse prior is placed on the β's.

Of course, the correlation $R(w, x)$ has to be specified to compute any of these quantities. It should reflect the characteristics of the output of the computer code. For a smooth response a covariance function with some derivatives would be preferred, whereas an irregular response might call for a function with no derivatives.

A natural class is the stationary family $R(w, x) = R(w - x)$, presuming that any anticipated nonstationary behavior can be modeled via the regression component. Within this family we restrict attention to correlations $R(w, x) = \prod R_j(w_j - x_j)$, which are products of one-dimensional correlations. Of special interest to us are those of the form

$$(9) \qquad R(w, x) = \prod \exp(-\theta_j \,|\, w_j - x_j \,|^p),$$

where $0 < p \leq 2$. (We can also permit p to vary with j.) The case $p = 1$ is the product of Ornstein-Uhlenbeck processes; these are continuous but otherwise not very smooth. The case $p = 2$ gives a process with infinitely differentiable paths (mean square sense) and is useful when the response is analytic.

An alternative correlation function, related to (9) with $p = 1$, is the product of linear correlation functions,

$$(10) \qquad R(w, x) = \prod(1 - \theta_j \,|\, w_j - x_j \,|)_+.$$

The predicted response $\hat{y}(x)$ using this correlation is a linear spline. From a one-dimensional correlation function $R_j(x_j, w_j)$, a smoothed correlation can be obtained by integrating,

$$\tilde{R}_j(w_j, x_j) = \int^{w_j}\int^{x_j} R_j(u, v)\, du\, dv.$$

Such correlations are not stationary. However, as shown by Mitchell, Morris and Ylvisaker (1988), stationary versions can be produced by a modified technique. In particular, the cubic correlation on the unit cube

$$(11) \qquad \prod[1 - a_j(w_j - x_j)^2 + b_j \,|\, w_j - x_j \,|^3],$$

for certain choices of a_j and b_j, is the stationary version of integrating (10) and produces cubic spline predictors.

The product form of the correlations is especially convenient for some of our computations. This rules out correlations like

$$(12) \qquad R(w, x) = \exp(-\theta \,\|\, w - x \,\|),$$

where $\|\cdot\|$ is Euclidean distance in d dimensions, but we are optimistic that the product families already provide enough flexibility for adequate prediction in most cases.

Given the family of correlations, there still remains the question of selecting or estimating the parameters of the family [θ_j and p in (9) say]. In Currin, Mitchell, Morris and Ylvisaker (1988) and Sacks, Schiller and Welch (1989), we have found that cross validation and maximum likelihood estimation (MLE) are useful at the analysis stage (i.e., after data have been collected) and in data-adaptive sequential design (see Section 5).

Assuming a Gaussian process, the likelihood is a function of the β's in the regression model, the process variance σ^2, and the correlation parameters. Given the correlation parameters, the MLE of the β's is the generalized least-squares estimate, and the MLE of σ^2 is

$$\hat{\sigma}^2 = \frac{1}{n}(y_S - F\hat{\beta})'R^{-1}(y_S - F\hat{\beta}).$$

With these definitions of $\hat{\beta}$ and $\hat{\sigma}^2$, the problem is to minimize $(\det R)^{1/n}\hat{\sigma}^2$, which is a function of only the correlation parameters and the data.

5. EXPERIMENTAL DESIGN

5.1. Introduction

The design of deterministic computer experiments has been partly addressed in the literature. For example, Sacks and Ylvisaker (1984, 1985), Welch (1983) and references mentioned therein have considered nonparametric systematic departures from regression models. Random error is also included, but the resulting sampling-variance contribution to mean squared error can be set to zero, and these approaches have helped shape our formulation. For the most part, however, the designs used for fitting predictors have been those developed for physical experiments. Such designs typically have appealing features of symmetry and are often optimal in one or more senses in settings which include random noise. Their appropriateness for computer experiments, however, is by no means clear. Latin hypercube sampling, discussed in Section 3, is aimed at objectives different from those we have in mind.

There has also been some work in design for numerical integration, where function evaluations can be viewed as a computationally cheap computer experiment. Much is known about design for one-dimensional quadrature. In particular, Sacks and Ylvisaker (1970) constructed good designs (finite n) from asymptotically ($n \rightarrow \infty$) optimal designs. These methods, however, do not carry over to $d > 1$ dimensions (see Ylvisaker, 1975). Similarly, in the numerical analysis literature (Davis and Rabinowitz, 1984) results for $d = 1$ offer little guide to $d > 1$.

5.2. Design Criteria

For a fixed number of runs, n, and for specified correlation structure R, we need a criterion for choosing a design that predicts the response well at untried inputs in the experimental region $\mathscr{X}$. Here, we consider functionals of the MSE matrix or kernel

$$M = \{E[Y(w) - \hat{y}(w)][Y(x) - \hat{y}(x)]\}$$

for all w and x in $\mathscr{X}$. The diagonal elements are the $\mathrm{MSE}[\hat{y}(x)]$ given in (8). In the Bayes case when the β's in (1) are known constants, M is just the posterior covariance matrix of the process. When the β's have prior variances that tend to infinity, M is the limiting posterior covariance matrix of $Y(\cdot)$. We now list various criteria based on M.

Integrated Mean Squared Error (IMSE). The IMSE criterion chooses the design S to minimize

$$\int_{\mathscr{X}} \mathrm{MSE}[\hat{y}(x)]\phi(x)\,dx$$

for a given weight function $\phi(x)$. From (8) the IMSE can be written as

$$\sigma^2\left\{1 - \mathrm{trace}\left[\begin{pmatrix} 0 & F' \\ F & R \end{pmatrix}^{-1} \cdot \int \begin{pmatrix} f(x)f'(x) & f(x)r'(x) \\ r(x)f'(x) & r(x)r'(x) \end{pmatrix} \phi(x)\,dx\right]\right\}. \tag{13}$$

These integrals simplify to products of one-dimensional integrals if $\mathscr{X}$ is rectangular and the elements of $f(x)$ and $r(x)$ are products of functions of a single input factor. Thus, polynomial regression models and product correlations can be numerically convenient.

The IMSE criterion is essentially the trace of M (suitably normalized). We assume that $\phi(x)$ is uniform, though other weights cause no real difficulty. This criterion has proved to be effective in terms of *actual* squared error of prediction in test examples reported by Sacks, Schiller and Welch (1989).

Maximum Mean Squared Error (MMSE). Instead of integrating the MSE of prediction, MMSE is a minimax criterion which chooses the design to minimize

$$\max_{x \in \mathscr{X}} \mathrm{MSE}[\hat{y}(x)].$$

Sacks and Schiller (1988) compared IMSE and MMSE for discrete regions. For continuous regions, however, this criterion is computationally complex. It involves a d-dimensional optimization of a function with numerous local optima at every iteration of a given design-optimization algorithm.

Entropy. A criterion advanced by Lindley (1956) in his work on Bayesian design is the minimization of the expected posterior entropy. Shewry and Wynn (1987, 1988) applied it to spatial sampling, and Currin, Mitchell, Morris and Ylvisaker (1988) applied it to the design of computer experiments. It quantifies the "amount of information" in an experiment. In the present setting, if the experimental region $\mathscr{X}$ is discrete, the entropy criterion chooses the design S to minimize $E(-\log g)$, where g is the conditional density of $Y(\cdot)$ on $\bar{S} = \mathscr{X} - S$ given Y_S. Using a classical decomposition of entropy, Shewry and Wynn (1987) showed that minimizing the expected posterior entropy on $\bar{S}$ is equivalent to maximizing the *prior* entropy on S. When $Y(\cdot)$ is Gaussian, this is the same as choosing S to maximize the determinant of V_S, the covariance matrix for $Y(\cdot)$ on S. Straightforward algebra also shows that, in the limiting Bayes case as the prior variances of the β's tend to infinity, maximization of $\det V_S$ is equivalent to maximizing $\det R \cdot \det(F'R^{-1}F)$. If the β's are regarded as fixed (as in Currin, Mitchell, Morris and Ylvisaker, 1988, for the case of a constant prior mean), the last determinant disappears and the entropy criterion reduces to maximization of $\det R$.

5.3. Algorithms

There is no way to implement the ideas set forth above without a method of constructing designs. The utility of D-optimal designs for standard analysis of variance and regression problems with independent experimental errors has only been realized by the development of accessible algorithms (Fedorov, 1972; Mitchell 1974; Welch, 1985; and Wynn, 1970).

Because standard designs can be inefficient or even inappropriate for deterministic computer codes, the need for computer software is even greater. Of course, efficiency has to be weighed against computational cost and convenience. Computer models like the flame code in Section 2, which themselves are expensive to run on supercomputers, justify the cost of supercomputing in constructing good designs. It is these models we have in mind here. Less effort would be warranted to design for a code that runs on a workstation, say, and so there is also a need for cheap, less sophisticated algorithms.

We now describe some of the algorithms we have used. They can be classified as single-stage methods, sequential methods without adaption to the data, and sequential methods with adaption.

Single-stage design fixes n in advance, and all n design sites are simultaneously optimized according to one (or perhaps a combination) of the above criteria. In addition to standard optimization routines, such as quasi-Newton, a number of exchange algorithms have been tried, primarily when the experimental region is a large, finite grid. At each iteration, an exchange replaces a site in the design by a site that improves the criterion. Currin, Mitchell, Morris and Ylvisaker (1988) adapted Mitchell's (1974) DETMAX excursion algorithm for the entropy criterion. The exchange algorithms used by Shewry and Wynn (1987) exchange sites by adding a random candidate site to the design and deleting the worst site. When the design is close to a (possibly local) optimum, the random candidates are restricted to neighborhoods of the current sites. A simulated annealing algorithm was found useful by Sacks and Schiller (1988) in problems with a small, finite experimental region. For larger problems, the time taken for the annealing process to converge to the optimum was far too long. Simulated annealing algorithms typically require many exchanges and are therefore feasible only when exchanges are cheap. Unfortunately, in our context each exchange may require substantial linear algebra. For continuous regions, we currently prefer standard optimization routines, at least for $n \times d < 100$.

Sequentially designing one site at a time reduces the computational burden from a single $n \times d$-dimensional optimization to a sequence of d-dimensional optimizations. Unlike physical experiments, sequential schemes for computer experiments are no more difficult to organize than a single stage. The design can also adapt to information gathered about the regression model and $R(w, x)$. Furthermore, there is the option of allowing n to be determined as data accumulate, stopping the algorithm as soon as there is sufficient information. Fully sequential design is, therefore, the most natural for computer experiments; unfortunately, it is also the most difficult to treat theoretically.

A sequential design algorithm devised for the IMSE criterion, though *ad hoc*, avoids some pitfalls (see Section 7) encountered in using simple one step look ahead schemes. It starts by dividing the experimental region into a number of subregions or boxes. Each new point is added by computing the contribution to the current IMSE from each box, finding the box with the largest contribution, and adding a point *in that box* that most reduces the contribution *in that box.* The example of the next section exercises this algorithm.

6. CIRCUIT-SIMULATOR EXAMPLE

To illustrate what is already possible, we take a circuit-simulator code similar to that considered by Welch, Yu, Kang and Sacks (1988) and mentioned in Section 2, but differing in the circuit topology. Again, the response is a clock asynchronization or "skew," and we consider six transistor widths as inputs. To avoid getting sidetracked by issues specific to quality control, we do not consider the noise factors here (they are kept fixed at average levels), nor do we perform any circuit-design optimization. We only consider the problem of predicting the clock skew as a function of the six input widths.

The experimental region of interest for the six widths is rectangular, which we transform to the unit cube $[-\frac{1}{2}, \frac{1}{2}]^6$. We assume the model

$$Y(x) = \beta + Z(x), \tag{14}$$

where $Z(\cdot)$ has a correlation function given by (9). This model is selected for various reasons. The regression component includes only the constant β partly because our previous experience in other examples has indicated that this simplification does not affect predictive performance. Moreover, engineering experience does not suggest strong trend over the region of interest. The circuit-simulator clock skew is believed to behave smoothly as a function of the transistor widths; by putting $p = 2$ in (9), a smooth correlation function for $Z(\cdot)$ is obtained. (This initial major assumption of smoothness is revised later by estimating p.) A similar model also gives good predictions when applied to the data in Welch, Yu, Kang and Sacks (1988).

Partly based on our experience with the earlier problem, we allow a total of 32 runs of the simulator for the experimental design. Choosing a single-stage design would mean specifying $\theta_1, \cdots, \theta_6$ and carrying out a 192-variable (6×32) optimization of the design-point coordinates. To reduce the computational burden and to allow adjustment of the model in midstream, we select a first-stage design of 16 points by setting $\theta_1 = \cdots = \theta_6 = 2$ for efficiency-robustness in the sense of Sacks, Schiller and Welch (1989) (described further in Section 7). Optimizing the IMSE over $6 \times 16 = 96$ coordinates using a quasi-Newton library routine takes about 11 minutes on a Cray X-MP. The design, given in the first 16 rows of Table 1, is probably only locally optimal. The

TABLE 1
Experimental design and clock skews for the circuit-simulator example

Run	Experimental design						Skew
1	0.21	−0.26	0.23	−0.21	−0.17	−0.27	−0.972
2	−0.19	0.18	0.22	0.21	0.25	0.28	−0.620
3	−0.19	−0.08	−0.28	−0.28	−0.25	−0.18	−0.711
4	0.19	−0.25	0.28	0.28	−0.06	0.19	−1.040
5	−0.28	0.25	−0.22	−0.21	0.17	0.19	−0.532
6	−0.22	0.21	0.17	0.16	−0.22	−0.22	−0.799
7	−0.22	−0.12	0.27	−0.25	0.23	−0.11	−0.940
8	0.11	0.23	−0.27	0.24	−0.13	0.22	−0.416
9	−0.19	−0.19	−0.19	0.24	0.22	−0.17	−0.500
10	0.17	0.21	0.19	−0.24	−0.20	0.19	−1.293
11	−0.26	−0.24	0.01	0.01	−0.24	0.26	−1.152
12	0.18	0.25	−0.21	−0.21	0.16	−0.28	−0.161
13	0.28	0.18	0.21	0.20	0.25	−0.18	−0.496
14	0.27	−0.18	−0.23	0.21	−0.26	−0.20	−0.612
15	−0.01	0.00	0.00	0.00	0.00	0.01	−0.604
16	0.22	−0.22	−0.17	−0.16	0.21	0.22	−0.897
17	0.10	−0.30	−0.32	−0.38	0.33	−0.30	−0.342
18	0.01	0.31	0.35	0.45	−0.36	0.41	−1.199
19	−0.32	0.45	−0.47	0.44	0.36	−0.28	−0.083
20	−0.27	0.37	0.33	−0.33	0.37	0.30	−1.048
21	−0.41	0.38	−0.32	−0.29	−0.47	0.37	−1.088
22	0.14	0.38	0.36	−0.40	−0.46	−0.49	−0.804
23	−0.15	−0.30	−0.28	0.28	0.29	0.26	−0.444
24	−0.24	−0.36	0.38	0.30	0.35	−0.37	−0.799
25	−0.46	−0.39	0.29	−0.37	−0.46	0.34	−1.918
26	0.17	0.36	−0.26	0.29	−0.41	−0.40	−0.535
27	0.23	−0.20	0.26	0.34	−0.45	−0.27	−1.242
28	0.31	−0.32	−0.25	−0.31	−0.19	0.29	−1.129
29	−0.01	−0.33	0.34	−0.43	0.47	0.37	−1.214
30	0.20	−0.37	−0.36	0.46	−0.45	0.39	−1.049
31	0.21	0.31	0.32	−0.20	0.45	−0.46	−0.135
32	−0.21	0.29	−0.27	0.20	0.40	0.41	−0.256

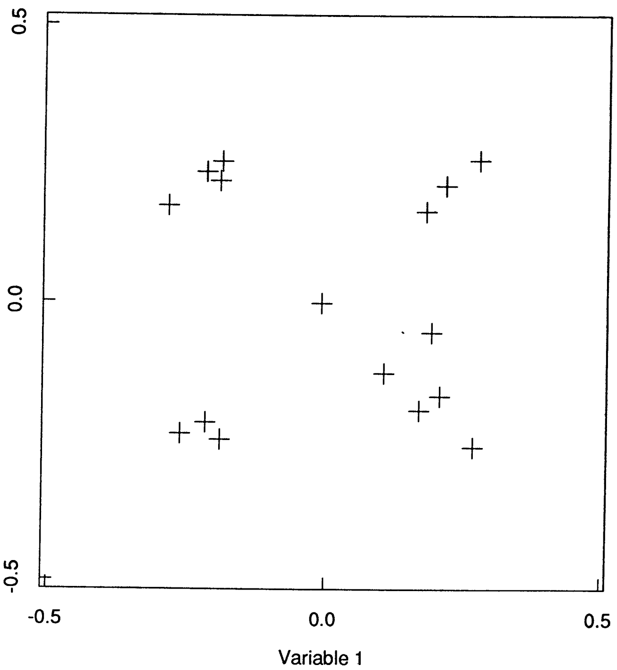

FIG. 1. *Projection of the experimental design onto the coordinates of two input variables.*

projection onto two of the six input coordinates in Figure 1 shows that the design is well away from the boundary, very likely a feature of the IMSE criterion with the constant regression model.

With the data from running the simulator at these 16 points, the MLE of p is 2 (the upper constraint) and those of $\theta_1, \cdots, \theta_6$ are .00, .39, .42, .53, 1.97 and .46. These values are now used in the generation of the second-stage design by the sequential strategy outlined in Section 5. The experimental region is broken into 32 boxes by dividing each of the last five input ranges in half. The first variable is not used to define these boxes as $\hat{\theta}_1 \approx 0$, suggesting that the response is fairly constant (highly correlated) over this factor, though it is still included in the second-stage design. The second set of 16 points, generated one at a time, is given in the second half of Table 1. These points are less concentrated in the center of the design region than the first-stage design, about which we have some misgivings. The MLE of p recomputed from all 32 observations is 1.54, indicating a less-smooth surface than initially thought. The MLEs of $\theta_1, \cdots, \theta_6$ are .00, .06, .19, .34, .14 and .32, again suggesting that the first factor is irrelevant.

To investigate the effectiveness of the BLUP based on this design, we can compare the true responses from the simulator at 100 random points $r_1, \cdots, r_{100}$ in the experimental region with predictions from the BLUP. (We chose a computationally cheap circuit-simulator code to allow this evaluation.) One summary statistic is the empirical integrated squared error

$$\frac{1}{100} \sum [\hat{y}(r_i) - y(r_i)]^2,$$

which equals $(.122)^2$ (relative to a data range of about 2). The maximum absolute discrepancy between the true clock skew and the BLUP over these 100 points is .458. For comparison, a quadratic response surface with 28 unknown coefficients fitted by least squares to the data from our design gives an empirical integrated squared error of $(.674)^2$ and a maximum absolute error of 1.71. This illustrates the potential danger in extrapolating polynomial models, but part of the poor performance may be due to our design, which is not intended for this sort of analysis.

It is also interesting to see whether the MSE (8) of the BLUP is a meaningful indicator of uncertainty in prediction. From the MSEs at the 100 random points (again based on the 32-point MLEs), one can compute standardized residuals $[y(r_i) - \hat{y}(r_i)]/\{\text{MSE}[\hat{y}(r_i)]\}^{1/2}$. The Q–Q plot in Figure 2 shows that these standard-

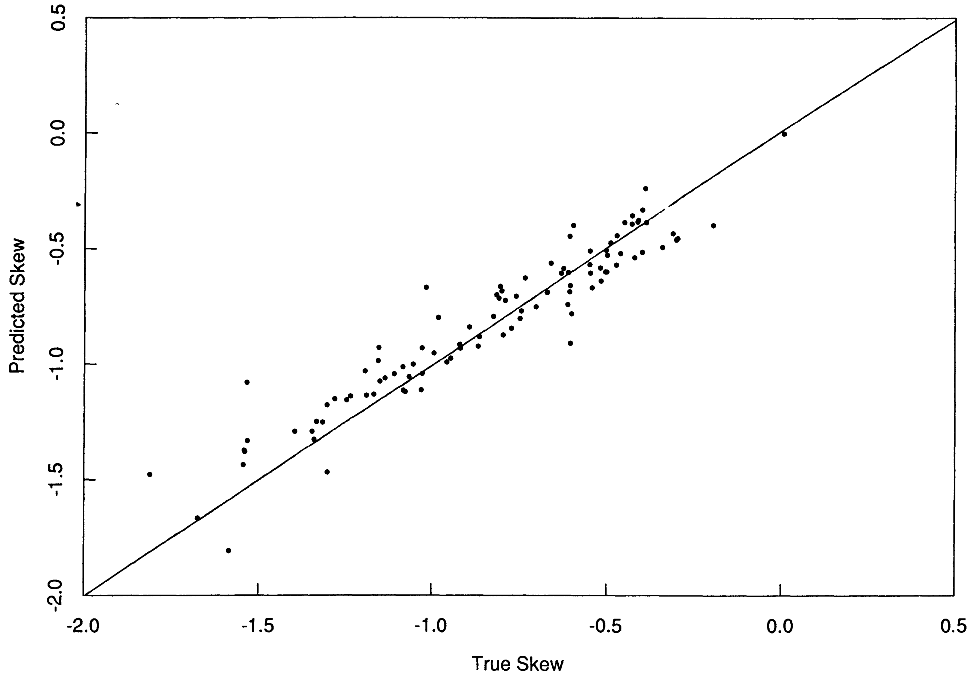

FIG. 2. *Q − Q plot of the standardized residuals against standard normal quantiles.*

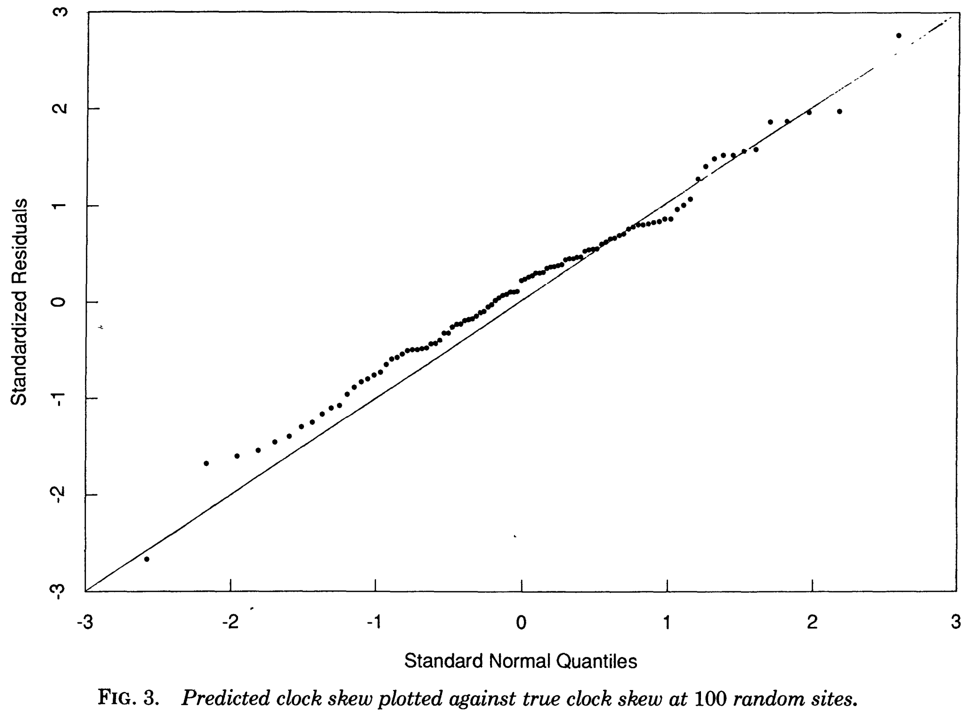

FIG. 3. *Predicted clock skew plotted against true clock skew at* 100 *random sites.*

ized residuals are approximately standard normal, suggesting a central-limit-theorem effect. Also, the slope of the plot is fairly close to 1, indicating that the MSEs do, indeed, provide a valid estimate of error in this example. The plot of $\hat{y}(r_i)$ against $y(r_i)$ in Figure 3 also shows that the poorest predictions tend to be where there are large negative skews. Possibly, the computer code is erratic at such extreme clock skews and harder to predict.

For insight into the relative effects of the six inputs, the response can be decomposed into an average, main effects for each input, two-input interactions and higher-order interactions. Define the average of $y(x)$ over the experimental region by

$$\mu_0 = \int y(x) \prod_{h=1}^{6} dx_h,$$

the main effect of input x_i (averaged over the other inputs) by

$$\mu_i(x_i) = \int y(x) \prod_{h \neq i} dx_h - \mu_0,$$

the interaction effect of x_i and x_j by

$$\mu_{ij}(x_i, x_j) = \int y(x) \prod_{h \neq i,j} dx_h - \mu_i(x_i) - \mu_j(x_j) - \mu_0,$$

and so on for higher-order interactions. These effects are estimated by replacing $y(x)$ by $\hat{y}(x)$. In the current example, visual inspection of the estimated effects up to two-input interactions suggests that the average, the main effects for factors 2–6, and the interaction of x_4 and x_6 are the important effects. The predictor

$$\hat{\mu}_0 + \hat{\mu}_2(x_2) + \cdots + \hat{\mu}_6(x_6) + \hat{\mu}_{46}(x_4, x_6)$$

gives an empirical squared error of $(.128)^2$, supporting this interpretation.

Using a different design criterion (entropy), algorithm (adaptation of DETMAX), and correlation function [(9) with $p = 1$ at the first stage and (11) at the second stage], Curin, Mitchell, Morris and Ylvisaker (1988) arrived at a design concentrated on the boundary of the experimental region. When used to predict at the same 100 random points, they reported an empirical integrated squared error of $(.163)^2$ and maximum absolute error of .369 over the same 100 random points. Thus, the predictions from this alternative approach are worse on average than the design produced by the IMSE criterion, but the maximum error is better.

7. DISCUSSION

We now summarize a number of open statistical problems that we have discussed only briefly so far and some alternative approaches.

7.1. Simulator Complexity

Almost all of the simulation codes we have worked with are differential-equations solvers. Many of the numerical and other difficulties we have encountered with these codes have implications for the statistical design and analysis.

- A single run of the code may be computationally expensive, for example the 20-minute run time for the flame code (see Section 2), obviously calling for efficient design and analysis.
- The coarse solution to the TWOLAYER code (see Section 2) is a step-like function that may not mirror important features of the accurate solution. An accurate solution is expensive.
- The mathematical model itself may be a poor approximation to reality. For example, the simple, deterministic function used by Taguchi (1986, Chapter 6) for parameter design of a Wheatstone bridge generates negative electrical resistances over part of the region of experimentation. Such aberrant data are misleading and can degrade the analysis. In complex settings, computer-model deficiencies are not so easy to identify. In this article we have largely ignored the problem of validating codes against reality. Rather we have focused on prediction of the computer code itself. Of course, a predicted response that is surprising may help to identify defects in the code.
- The inputs may be of high dimension. This interacts with the first difficulty. If the data are expensive, scientists and statisticians are fully aware of the difficulty in obtaining adequate information about many factors, and screening to reduce dimension is necessary. Thus, expensive data (few runs) and low dimension go together. Cheap data, however, allow many runs, so many factors can be investigated and often are.

7.2. Estimation of Model Parameters

Because the correlation matrix of the data, R, is $n \times n$, the maximum-likelihood computations outlined in Section 4 can be formidable. Vecchia (1988) approximated the likelihood by writing it as a product of conditional densities and conditioning on only a small number of nearest sites. The approximation is cheaper to compute but may retain most of the information.

Properties of the MLE are not well understood and are under study. Mardia and Marshall's (1984) asymptotic results on consistency are not applicable if the region for x is bounded. Their Monte Carlo studies of small-sample behavior indicated substantial variability in the estimates. The validities of the BLUP and measures of uncertainty calculated by substituting MLEs of the correlation parameters therefore appear questionable, but our experience is that even crude MLEs can lead to useful predictions and quantification of uncertainty. Stein (1988) showed that under special circumstances the BLUP can be not only consistent but asymptotically efficient even when the correlation function is misspecified, provided the misspecification leads to a "compatible" Gaussian measure.

7.3. Design Algorithms

All algorithms we have tried for single-stage design are impeded by a number of computational obstacles.

- The optimization is over $n \times d$ design-site coordinates. Though symmetries in the optimal designs are sometimes present, we have not found ways to exploit them to reduce the dimension of the optimization. Since there can be numerous local optima, several tries are necessary.
- Evaluating a "trial" design at each iteration of an optimization algorithm typically involves the solution of a set of at least n linear equations, for example (13) to compute the IMSE. (The vectorizing architectures of computers like the Cray X-MP we have used are ideal for this type of linear algebra, however.)
- The correlation matrix R in (13) (and in other criteria) can be poorly conditioned, and naive rules for cheaply updating the solution from one iteration to the next may lead to numerical errors. For a given correlation function, the conditioning of R becomes worse as n increases.

Thus, the design criteria of Section 5.2 require particularly careful numerical analysis. The computation of D-optimal or other efficient designs for experiments with independent errors shares some of these difficulties, but to a far lesser degree.

As discussed in Section 5.3, sequential design is computationally cheaper and allows adaption to the data. Simple (myopic) sequential strategies of adding the next point to minimize the value of the new design criterion do not work well, however, at least for the IMSE and MMSE criteria. There is a tendency for design sites to eventually "pile up." This may seem counter-intuitive but consider the following example. With the MMSE criterion, take $d = 1$ and $\mathscr{X} = [-1/2, 1/2]$. Suppose model (1) has no regression component, and let $Z(\cdot)$ have correlation function $\exp[-(w - x)^2]$. Let the first site, s_1, be placed at zero. If the second site, s_2, is to the left of zero, a straightforward calculation of MSE$[\hat{y}(x)]$ from (8) shows that the maximum MSE $[\hat{y}(x)]$ occurs at $x = 1/2$, and the maximum decreases as s_2 tends to zero. Exact replication does not occur—the limiting design enables $y(0)$ and $y'(0)$ to be evaluated—but this is inefficient relative to the best two-site design. In several dimensions, we have observed that the first few design sites do not pile up in this way, but the same phenomenon eventually occurs. This is not a problem for the entropy criterion, because it places each new design site where the current MSE$[\hat{y}(x)]$ is maximized, thereby avoiding the neighborhoods of existing design sites.

We described in Section 5.3 a modified sequential algorithm for the IMSE criterion which overcomes this problem by dividing the experimental region. To test the efficiency and running time of this algorithm, we constructed various designs with $9 \leq n \leq 25$, $p = 1.6$ or 2 in correlation (9), $d = 2, 3$, or 4 dimensions, and constant (β) or first order $(\beta_0 + \sum x_j\beta_j)$ regressions. The sequential algorithm required only about 20–30% of the CPU time of a full optimization of all n design sites. Further computational gains would be possible by updating, rather than recomputing, the IMSE as each new site is introduced. Clearly, any sequential scheme without adaption to the data has to be less efficient than an optimal one-stage scheme. Nonetheless, some comparisons show that the efficiency of the designs constructed by the sequential algorithm just described ranges from 40–90%. The lower efficiencies tend to arise when small IMSEs are compared; that is, when n is large, d is small and the regression has just the constant term. Adapting the correlation structure to the data (e.g., by MLE) could lead to sequential methods which outperform one-stage algorithms, especially if the data indicate that some inputs are more important than others.

7.4. Efficiency-Robustness of Designs

Assumptions have to be made about the model for $Y(\cdot)$ and the design criterion. It is natural to ask a number of questions about the efficiency of a design if assumptions change.

- *How sensitive are optimal designs to the choice of correlation structure?*
- *What effect does the regression part of the model have on design?*
- *How do designs chosen by one criterion perform with respect to other criteria?*
- *Are there sub-optimal designs which are robust to choice of criterion?*
- *How important is optimality in this setting?*
- *Are there cheap-to-construct alternatives that perform reasonably well?*

Answers to these questions are limited to a large extent because of the difficulty in computing optimal designs; at the moment we can only refer to some fragmentary, anecdotal results.

Sacks, Schiller and Welch (1989) investigated the effect of the correlation function on the efficiency of the design and predictor. Their study was limited to the effect of the correlation parameter θ within the family (9) with $p = 2$. They computed IMSE-optimal designs for various values of θ. For a given "true" θ, the efficiency of one of these designs, S, relative to the optimal design S_θ was defined to be IMSE(S_θ)/IMSE(S), and there will be some worst-case value of

θ, which minimizes this efficiency. The design that maximizes the worst-case efficiency was deemed to be robust to θ. A further complication is that when evaluating IMSE(S), the BLUP can be based on the true θ or that assumed when generating the design. If the data will be extensive enough to estimate the correlation structure, the true θ may be appropriate, otherwise the assumed θ is retained at the prediction stage. Sacks, Schiller and Welch (1989) considered both cases. Typically, designs for "moderately small" θ resulted. This approach requires computing a number of optimal designs and is limited to problems with $n \times d < 100$, say. For larger problems these efficiency-robust designs can be used, however, to start a sequential scheme.

Currin, Mitchell, Morris and Ylvisaker (1988) implicitly considered robustness of efficiency of the entropy criterion to the correlation structure, although they made no study. In several examples, they designed using (9) with $p = 1$ and θ very large. The intuition was that this prior represents hard-to-predict (low correlation) functions, whereas any reasonable design would deal adequately with easier functions. [There is a connection between designs produced by the entropy criterion as correlations become smaller and those from maximizing the minimum distance between the design sites (Johnson, Moore and Ylvisaker, 1988).] A measure of efficiency based on *differences* in MSEs would lead to a choice of a low-correlation prior, whereas the contrary findings of Sacks, Schiller and Welch (1989) were based on *relative* efficiency.

Sacks and Schiller (1988) investigated the effect of qualitatively different correlation functions—(9) with $p = 2$ versus (12)—on robustness of efficiency. They used MMSE as the criterion, had no regression model and designed on a grid. The θ's of the two correlation functions were chosen to match correlations between Z's at nearest neighbor grid points. This study showed that designs optimal by the MMSE criterion for one correlation were over 80% efficiency for the other (the entries in their Table 3.1 need to be re-ordered). In contrast, we have found that, in predicting two-dimensional integrals, good designs for correlation (9) with $p = 1$ behave poorly in terms of relative efficiency when $p = 2$.

Whether or not a design has robustness of efficiency with respect to alternative correlation functions, the properties of the BLUP will be seriously affected. In particular, higher correlations dramatically increase the apparent precision of prediction. Fortunately, using the data to estimate correlation parameters may lead to effective prediction and reliable estimates of uncertainty (as in the example of Section 6).

The role of the regression model is not yet clear, but it seems to be less important than in design for traditional models with "white noise" errors. Systematic departure from the regression model just becomes part of $Z(\cdot)$, and the BLUP is always an interpolator. In the circuit-simulator experiment, for example, our regression model included only a constant term, yet the predictor appears to follow the true surface, which is clearly not constant, reasonably well. In Example 2 of Sacks, Schiller and Welch (1989) a special class of designs was employed for a methane-combustion code, and it was noted that the effect of the regression model was negligible at the prediction stage. The BLUP was able to adapt to the absence or presence of regression terms: a smaller regression model is compensated for by a covariance function with larger estimated correlations. This phenomenon has some theoretical justification in ongoing work with Y. B. Lim and W. J. Studden on the asymptotic behavior of designs and predictors as the correlation gets large in (9) with $p = 2$.

Sacks and Schiller (1988) found that the entropy and MMSE criteria produce very different designs. The example of Section 6 indicates strong differences between designs from the entropy and IMSE criteria. The entropy criterion tends to push the design sites away from one another, so for small n the optimal design lies on the boundary of the experimental region. As n increases, some interior sites appear—the higher the dimension, the larger n has to be for this to occur. Attraction to the boundary seems not to be a feature of the IMSE and MMSE criteria. In fact, the first 16 runs in Table 1, chosen nonsequentially by IMSE, are well in from the boundary. These remarks are concerned only with the appearance of the designs; we know of no comprehensive investigations of efficiency robustness with respect to the entropy, IMSE, and MMSE criteria. It may turn out that new criteria are necessary, possibly incorporating robustness explicitly.

7.5. Some Alternative Approaches

There are some close connections between the experimental designs produced by the IMSE criterion and previous approaches aimed at minimizing the impact of systematic error in physical experiments. The primary design criterion of Box and Draper (1959, 1963) is also an integrated mean squared error, including components from squared bias and error variance. The variance component turned out to be unimportant for design in the sense that "all-bias" designs that minimize the bias component do fairly well even when the variance component is substantial. Despite modeling the systematic departures by higher-order polynomials rather than a stochastic process, these all-bias designs are qualitatively similar to those from our use of the IMSE criterion, with design points away from the boundaries of the region of interest. It is plausible that they may be competitive for computer experiments, but the numerical burdens are again extensive.

We have some doubts about transferring least-squares fitting of response surfaces to computer experiments, however. Comparisons can be made by computing the root average squared error or maximum absolute error from test data. In the circuit-simulator example of Section 6, the least-squares quadratic fit is only about 18% and 27% efficient by these criteria relative to the fit from model (14). In this comparison the design constructed for the IMSE criterion was used for both fits. Sacks, Schiller and Welch (1989) reported an example where the least-squares fit to data from a standard factorial design with design points at the boundary of the region of interest had similarly low efficiency.

Our methods are interpolation schemes and could be compared to methods in the numerical analysis literature. The correlation functions (10) and (11) lead to linear and cubic splines. In one dimension, the correlation (9) with $p = 2$ is related to Lagrangian interpolation when θ is small. There is little information in the literature about the construction of good designs for higher-dimensional interpolation.

In the presence of systematic rather than random error, a good experimental design tends to fill out the design space rather than being concentrated on the boundary. Low-discrepancy sequences such as Halton (1960) sequences for numerical integration of non-smooth functions have this "space filling" property (as do Latin hypercube designs). Also, the use made of discrepancy criteria and error bounds based on maximum or average bias are closer in spirit to the approach of this paper than to the randomization bounds of classical Monte Carlo (see Niederreiter, 1978). The efficiencies of these easy-to-generate designs for the objective of prediction should be investigated, especially for very large experimental designs, where criterion optimization may be infeasible.

7.6. Kriging and Spatial Design

In the kriging and spatial statistics literature, the random process $Z(\cdot)$ is often modeled using the variogram $E[Z(w) - Z(x)]^2$ rather than the covariance function. Analogous computational formulas for the BLUP, etc. follow. The variogram permits a wider class of processes, but we are not certain that the added flexibility is needed in our applications. Estimation of the variogram has been studied by several authors; see Cressie (1988) for a recent review.

The data to which spatial methods are applied usually have a two- or three-dimensional x space. They sometimes appear to have measurement error or may be more erratic than responses from computer codes. Geostatistical models used often incorporate a so-called "nugget effect" for erratic local behavior. While we have not addressed such models, it is worth noting that correlation functions of the form (9) with $0 < p \leq 1$ may be useful for modeling such erratic data.

It is not obvious that methods of estimating the variogram extend well from low-dimensional spatial coordinates to the typically high-dimensional inputs of computer experiments. Similarly, results like those in Yfantis, Flatman and Behar (1987) on the properties of regular-grid designs, while interesting for two-dimensional x, are not apparently relevant for computer experiments.

Though we have stressed that deterministic observations are the unique feature of computer experiments, the methodology can be extended to settings where systematic and random error are both important. The covariance function can be adapted so that $\mathrm{Var}[Y(x)] = \sigma^2 + \sigma_\epsilon^2$, where σ_ϵ^2 is the variance of the measurement error. (In kriging applications, this can be difficult to distinguish from the nugget effect.) Thus, these approaches should also be useful for physical experiments.

8. CONCLUSIONS

Many scientists feel that statistics is irrelevant to their problems, even for physical experimentation. Their experiments, they claim, have little random variation but are plagued by possibly large systematic biases. These criticisms are not unfounded. There is little easily implemented methodology that addresses systematic error, and the reality might appear even starker for computer experiments with no measurement error. Predictions are nonetheless made with uncertainty, a statistical problem. The stochastic models we have applied to computer experiments quantify uncertainty about the response where it is unobserved and provide a framework for efficient design and analysis, which has been useful in a number of applications.

ACKNOWLEGMENTS

This research program grew out of NSF-funded Workshops on Efficient Data Collection in 1984 and 1985. The authors were supported by AFOSR, NSF Grants DMS-86-09819 and DMS-87-03802, Office of Naval Research Contract N00014-85-K-0357, and the Applied Mathematical Sciences Program of the U.S. Department of Energy through Contract DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc. Extensive use was made of the Cray X-MP at the National Center for Supercomputing Applications, University of Illinois, partially supported by Cray

Research, Inc. We also acknowledge the contributions of Chona Bernardo, Robert Buck, Carla Currin, Max Morris, Donald Ylvisaker and Tat Kwan Yu and the improvements suggested by the referees.

REFERENCES

Blight, B. J. N. and Ott, L. (1975). A Bayesian approach to model inadequacy for polynomial regression. *Biometrika* **62** 79–88.

Box, G. E. P. and Draper, N. R. (1959). A basis for the selection of a response surface design. *J. Amer. Statist. Assoc.* **54** 622–654.

Box, G. E. P. and Draper, N. R. (1963). The choice of a second order rotatable design. *Biometrika* **50** 335–352.

Box, G. E. P. and Draper, N. R. (1987). *Empirical Model-Building and Response Surfaces.* Wiley, New York.

Box, G. E. P., Hunter, W. G. and Hunter, J. S. (1978). *Statistics for Experimenters.* Wiley, New York.

Cressie, N. (1988). Variogram. *Encyclopedia of Statistical Sciences* **9** 489–491. Wiley, New York.

Currin, C., Mitchell, T., Morris, M. and Ylvisaker, D. (1988). A Bayesian approach to the design and analysis of computer experiments. ORNL Technical Report 6498, available from the National Technical Information Service, Springfield, Va. 22161.

Davis, P. J. and Rabinowitz, P. (1984). *Methods of Numerical Integration*, 2nd ed. Academic, Orlando, Fla.

Diaconis, P. (1988). Bayesian numerical analysis. In *Statistical Decision Theory and Related Topics IV* (S. S. Gupta and J. O. Berger, eds.) **1** 163–175. Springer, New York.

Fedorov, V. V. (1972). *Theory of Optimal Experiments.* Academic, New York.

Fisher, R. A. (1935). *The Design of Experiments.* Oliver and Boyd, Edinburgh.

Halton, J. H. (1960). On the efficiency of certain quasi-random sequences of points in evaluating multi-dimensional integrals. *Numer. Math.* **2** 84–90.

Iman, R. L. and Helton, J. C. (1988). An investigation of uncertainty and sensitivity analysis techniques for computer models. *Risk Analysis* **8** 71–90.

Johnson, M., Moore, L. and Ylvisaker, D. (1988). Minimax and maximin distance designs. Technical Report, UCLA Statistics Series #13.

Kee, R. J., Grcar, J. F., Smooke, M. D. and Miller, J. A. (1985). A FORTRAN program for modeling steady laminar one-dimensional premixed flames. Sandia Report SAND85-8240, available from the National Technical Information Service, Springfield, Va. 22161.

Kiefer, J. C. (1985). *Jack Carl Kiefer Collected Papers* **3**: *Design of Experiments* (L. D. Brown, I. Olkin, J. Sacks, and H. P. Wynn, eds.). Springer, New York.

Kimeldorf, G. S. and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist.* **41** 495–502.

Kleijnen, J. P. C. (1987). *Statistical Tools for Simulation Practitioners.* Dekker, New York.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *Ann. Math. Statist.* **27** 986–1005.

Mardia, K. V. and Marshall, R. J. (1984). Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika* **71** 135–146.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology* **58** 1246–1266.

McKay, M. D., Conover, W. J. and Beckman, R. J. (1979). A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics* **21** 239–245.

Mitchell, T., Morris, M. and Ylvisaker, D. (1988). Existence of smoothed stationary processes on an interval. Unpublished manuscript.

Mitchell, T. J. (1974). An algorithm for the construction of "*D*-optimal" experimental designs. *Technometrics* **16** 203–210.

Nassif, S. R., Strojwas, A. J. and Director, S. W. (1984). FABRICS II: A statistically based IC fabrication process simulator. *IEEE Trans. Computer-Aided Design* **CAD-3** 40–46.

Niederreiter, H. (1978). Quasi-Monte Carlo methods and pseudo-random numbers. *Bull. Amer. Math. Soc.* **84** 957–1041.

O'Hagan, A. (1978). Curve fitting and optimal design for prediction (with discussion). *J. Roy. Statist. Soc. Ser. B* **40** 1–42.

Parzen, E. (1963). A new approach to the synthesis of optimal smoothing and prediction systems. In *Mathematical Optimization Techniques* (R. Bellman, ed.) 75–108. Univ. California Press, Berkeley.

Sacks, J. and Schiller, S. (1988). Spatial designs. In *Statistical Decision Theory and Related Topics IV* (S. S. Gupta and J. O. Berger, eds.) **2** 385–399. Springer, New York.

Sacks, J., Schiller, S. B. and Welch, W. J. (1989). Designs for computer experiments. *Technometrics* **31** 41–47.

Sacks, J. and Ylvisaker, D. (1970). Statistical designs and integral approximation. In *Proc. 12th Bien. Sem. Canad. Math. Congress* (R. Pyke, ed.) 115–136. Canadian Mathematical Congress, Montreal.

Sacks, J. and Ylvisaker, D. (1984). Some model robust designs in regression. *Ann. Statist.* **12** 1324–1348.

Sacks, J. and Ylvisaker, D. (1985). Model robust design in regression: Bayes theory. In *Proc. of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L. M. Le Cam and R. A. Olshen, eds.) **2** 667–679. Wadsworth, Monterey, Calif.

Shewry, M. C. and Wynn, H. P. (1987). Maximum entropy sampling. *J. Appl. Statist.* **14** 165–170.

Shewry, M. C. and Wynn, H. P. (1988). Maximum entropy sampling and simulation codes. *Proc. 12th World Congress on Scientific Computation, IMAC88* **2** 517–519.

Singhal, K. and Pinel, J. F. (1981). Statistical design centering and tolerancing using parametric sampling. *IEEE Trans. Circuits and Systems* **CAS-28** 692–702.

Smale, S. (1985). On the efficiency of algorithms of analysis. *Bull. Amer. Math. Soc. (N.S.)* **13** 87–121.

Stein, M. L. (1988). Asymptotically efficient prediction of a random field with a misspecified covariance function. *Ann. Statist.* **16** 55–63.

Steinberg, D. M. (1985). Model robust response surface designs: Scaling two-level factorials. *Biometrika* **72** 513–526.

Suldin, A. V. (1959). Wiener measure and its applications to approximation methods. I. *Izv. Vyssh. Uchebn. Zaved. Mat.* **6**(13) 145–158. (In Russian.)

Suldin, A. V. (1960). Wiener measure and its applications to approximation methods. II. *Izv. Vyssh. Uchebn. Zaved. Mat.* **5**(18) 165–179. (In Russian.)

Taguchi, G. (1986). *Introduction to Quality Engineering.* Asian Productivity Organization, Tokyo.

Vecchia, A. V. (1988). Estimation and model identification for continuous spatial process. *J. Roy. Statist. Soc. Ser. B* **50** 297–312.

Welch, W. J. (1983). A mean squared error criterion for the design of experiments. *Biometrika* **70** 205–213.

Welch, W. J. (1985). ACED: Algorithms for the construction of experimental designs. *Amer. Statist.* **39** 146.

Welch, W. J., Yu, T. K., Kang, S. M. and Sacks, J. (1988). Computer experiments for quality control by parameter design. Technical Report No. 4, Dept. Statistics, Univ. Illinois.

WYNN, H. P. (1970). The sequential generation of D-optimum experimental designs. *Ann. Math. Statist.* **41** 1655–1664.

YFANTIS, E. A., FLATMAN, G. T. and BEHAR, J. V. (1987). Efficiency of kriging estimation for square, triangular and hexagonal grids. *Math. Geol.* **19** 183–205.

YLVISAKER, D. (1975). Designs on random fields. In *A Survey of Statistical Design and Linear Models* (J. N. Srivastava, ed.) 593–607. North-Holland, Amsterdam.

YLVISAKER, D. (1987). Prediction and design. *Ann. Statist.* **15** 1–19.

YOUNG, A. S. (1977). A Bayesian approach to prediction using polynomials. *Biometrika* **64** 309–317.

Comment

Max D. Morris

The authors have provided an interesting and readable account of a statistical approach to the problem of approximating an unknown, deterministic computer model. The approximation of unknown functions, of at least a few arguments, has received considerable attention in other specialty areas of mathematics, but is relatively new to statistics. A statistical approach brings a unique potential for dealing with uncertainty in the problem. In particular, it can lead to measures of quality for each prediction, and a structure on which to base the design of efficient experiments. Techniques which are relevant for approximating computer models are particularly timely, because the scientific and technical professions are quickly becoming reliant upon these as research tools, and this manuscript reports some of the first serious efforts to make statistics relevant to these activities.

THE CLASSICAL APPROACH

At the end of Section 3, the authors give their basic argument for treating this problem statistically: "Modeling a computer code as if it were a realization of a stochastic process . . . gives a basis for the quantification of uncertainty . . ." Following this, Section 4 outlines their strategy which seems clearly classical (as opposed to Bayesian) in form; it is what a classical statistician would do if the computer model actually had been generated as a realization of the stochastic process. While this strategy does provide a mathematical structure for dealing with uncertainty, classical statisticians who like to motivate their analyses with fictional accounts of random sampling and hypothetical replays of an experiment may find this an uncomfortable setting. After all, unless one randomizes the experimental design, there will not be a credible frequentist probability structure in this problem. (My own usual preference for classical procedures is heavily dependent on credible frequentist models. In this problem, the Bayesian approach seems somewhat more direct to me.)

A classical statistician, in order to proceed, will need to be more pragmatic, by saying that a credible frequentist model is unnecessary so long as the method works. The first test of whether the method works is whether it produces good approximations to computer models. These authors, and others they have referenced, have assembled a body of evidence that indicates that this and similar methods have the potential to produce good approximations. The second test, which should be of particular concern to statisticians, is whether it produces good (useful, dependable, meaningful?) measures of uncertainty. Passing this second test will be important if we are to take seriously any claims of quantified prediction uncertainty or design optimality. It is encouraging that the mean square errors of prediction calculated in the example of Section 6 seem to behave as we would hope.

CHOICE OF CORRELATION FUNCTION

As the authors point out in Section 4, the hopes of the pragmatic classical statistician will be pinned on the supposition that the computational model "though deterministic, may resemble a sample path of a (suitably chosen) stochastic process . . ." So, choosing a suitable stochastic process, presumably one for which y would be a "typical" realization, becomes an issue. This is particularly true for preliminary design purposes (before data are taken from which a correlation structure can be estimated). Some guidelines for this selection process are well-known; the authors note that $p = 2$ processes produce smoother realizations than $p = 1$ processes. Also, a tentative value of θ must be chosen for preliminary design purposes; the authors use $\theta = 2$ in the example of Section 6.

When selecting a process in several dimensions, some attention should probably be paid to the degree of interaction among inputs for typical realizations.

Max D. Morris is a Research Staff Member, Mathematical Sciences Section, Oak Ridge National Laboratory, P.O. Box 2009, Oak Ridge, Tennessee 37831-8083.

The following may be useful in thinking about what the product correlation form of equation (9) and a particular value of θ imply about these interactions. Using the unit cube design space $[-\frac{1}{2}, +\frac{1}{2}]^d$, suppose we set all but two inputs (say inputs 1 and 2) to arbitrary constant values, and denote by Y_{++}, Y_{+-}, Y_{-+}, and Y_{--} the process at $(x_1, x_2) = (+\frac{1}{2}, +\frac{1}{2})$, $(+\frac{1}{2}, -\frac{1}{2})$, $(-\frac{1}{2}, +\frac{1}{2})$, and $(-\frac{1}{2}, -\frac{1}{2})$, respectively. Let W_+ and W_- be the effects of the second input at the high and low values of the first input, respectively:

$$W_+ \equiv Y_{++} - Y_{+-}$$
$$W_- \equiv Y_{-+} - Y_{--}$$

If the process is stationary, i.e., the linear model piece of equation (1) is omitted except for an intercept, with $\text{Corr}(Y_{++}, Y_{+-}) = \text{Corr}(Y_{+-}, Y_{--}) = e^{-\theta} = \rho$, then $E(W_+) = E(W_-) = 0$, and $\text{Corr}(W_+, W_-)$ is also ρ. When W_+ and W_- are of different sign, increasing input 2 increases the response at one level of input 1 and decreases it at the other, a two-factor interaction generally considered to be rather serious and hoped to be rather rare in most modeling contexts. θ values of 5.0, 0.5 and 0.05, lead to ρ values of 0.01, 0.61, and 0.95, which are associated with "prior probabilities" of about 0.50, 0.30, and 0.10, respectively, that two inputs will have this kind of interaction on any such square region in the design space. By itself this seems to suggest that, unless fairly complex interaction patterns are expected in y, small values of θ (perhaps $\frac{1}{2}$ or less) are reasonable. When the linear model portion of the authors' equation (1) is included, weaker correlations can be used without implying a prior preference for these effect-reversing interactions.

Of course, other issues must also be addressed in choosing preliminary correlation values for design purposes. In particular, using the relatively small values of θ suggested above may lead to stronger-than-desirable correlations in each dimension individually. Sacks, Schiller and Welch (1989) suggested picking a preliminary θ value based on robustness considerations, while Currin, Mitchell, Morris and Ylvisaker (1988) conservatively chose a weak correlation to limit the inference which could be drawn at one site from data observed nearby. Knowing how to pick a correlation structure, and when to change it, will be critically important steps in hardening this methodology for general use.

OPTIMAL DESIGN

In Section 7.4, the authors pose a number of important questions including: "How important is optimality in this setting? Are there cheap-to-construct alternatives that perform reasonably well?" Answers will be important in this problem, because real computer models often have more inputs (larger d) than is customary in many physical experiments, and so full-scale design optimization will be a numerical problem of large dimension. The 16-run design used in the first stage of the example of Section 6 was computed by minimizing IMSE, assuming the correlation function of equation (9) with $\theta = 2$ and $p = 2$. Construction of the design required 11 minutes of time on a Cray X-MP computer, and the resulting value of $\sqrt{\text{IMSE}}$ was 0.6347 (arbitrarily fixing $\sigma^2 = 1$). I looked at a few cheap-to-construct 16-run alternatives, including the two-level resolution 4 fractional factorial design generated by I = ABCD = CDEF, centered in the design space and scaled so that the absolute value of each element in the design matrix varied from 0.05 to 0.5 in increments of 0.05. Assuming $\theta = 2$ and $p = 2$ as the authors did, $\sqrt{\text{IMSE}}$ values for these designs are shown in Table 1. (Values are also given for $p = 1$ for comparison.) In particular, the design scaled so that each input takes values $-\frac{1}{4}$ and $+\frac{1}{4}$ is nearly as good, with respect to IMSE, as the authors' design. Further, this design produces IMSE values similar to those of the optimal design for different values of θ and p (Table 2), suggesting that cheap-to-construct near-optimal designs may share any process-robustness properties the optimal design may have. Finally, since the authors' optimal design,

TABLE 1
$\sqrt{IMSE}$ for various scalings of a 16-run fractional factorial design

	$\sqrt{\text{IMSE}}$	
Scaling*	$p = 2$	$p = 1$
0.05	0.9061	1.1976
0.10	0.8389	1.0985
0.15	0.7527	1.0426
0.20	0.6798	1.0138
0.25	0.6508	1.0021
0.30	0.6773	1.0011
0.35	0.7432	1.0059
0.40	0.8213	1.0131
0.45	0.8913	1.0200
0.50	0.9446	1.0254

I = ABCD = CDEF

* Absolute value of each element in the design matrix.

TABLE 2
$\sqrt{IMSE}$ for the optimal design of Section 6 and the resolution 4 fractional factorial on $(\pm\frac{1}{4})^6$ for several values of θ, and $p = 2, 1$

	Optimal design		Fractional factorial	
θ	$p = 2$	$p = 1$	$p = 2$	$p = 1$
8.0	0.9795	1.0306	0.9798	1.0306
4.0	0.8548	1.0275	0.8601	1.0283
2.0	0.6347	0.9982	0.6508	1.0021
1.0	0.4011	0.8851	0.4239	0.8935
0.5	0.2265	0.7008	0.2470	0.7146

like the optimally scaled fractional factorial, places many input values about halfway between the center and edge of the design region, I was curious about how much of the optimality could be credited to this property alone. So I generated 100 random 16-run designs, where each element of the design matrix could be +¼ or −¼ with equal probability (the only restriction on the randomization was that no two runs could be identical), and evaluated the criterion for each of these. For $\theta = 2$ and $p = 2$, the smallest and largest values of $\sqrt{\mathrm{IMSE}}$ for these designs were 0.6743 and 0.7138, not as close to optimal as the shrunken fractional factorial, but also not too bad, and surprisingly (to me) consistent.

Of course, one example does not prove that there will always exist a cheap, simple, nearly optimal design. Also, as the authors note, it may not be so important to save 11 minutes of supercomputer time generating an optimal experimental design if the computer model itself requires even more time per run. But computing costs aside, I believe that a sizable gain in design simplicity and symmetry is often worth a small price in optimality.

Another related issue is how designs generated by different optimality criteria compare. Using the entropy criterion described in Currin, Mitchell, Morris and Ylvisaker (1988), I generated a locally optimal 16-run design for the problem of Section 6, again using $\theta = 2$ and $p = 2$. This design is almost entirely in the corners of the design space; only 4 of the 96 entries in the design matrix are other than +½ or −½. $\sqrt{\mathrm{IMSE}}$ for this design is 0.9343, which is not much different from that of the largest fractional factorial considered above. Just as in experimental design for linear models, there is no reason to believe that two "good" criteria should lead to exactly the same design. However, these two criteria are motivated by the same general goal—that of relatively good prediction of y in an overall sense—and it is somewhat disturbing to me that the results of these approaches seem so dramatically different. Somewhere along the line, I expect to learn either that the approaches are not as similar as I've assumed, or that the designs are not as different as they appear.

CONCLUSION

In summary, I think that both the approach outlined in this paper and the Bayesian alternative described by Currin, Mitchell, Morris and Ylvisaker (1988) are promising tools for approximating computer models. A number of issues, such as selection of a stochastic process and criteria against which designs may be measured, must eventually be addressed in considerably more detail. However, this paper marks an excellent beginning, and the authors are to be congratulated on a job well done.

ACKNOWLEDGMENT

This research was sponsored by the Applied Mathematical Sciences Research Program, Office of Energy Research, U.S. Department of Energy under contract DE-AC05-84OR21400 with Martin Marietta Energy Systems, Inc.

Comment

Robert G. Easterling

The authors, referred to hereafter as SWMW, are to be commended for their pioneering work in bringing statistical thinking and methods to the design and analysis of computer experiments. Critical decisions are being made and conclusions drawn based on complex computer models. Data may be lurking about, so it is natural and vitally important that statisticians get involved, and even when data are not lurking or visible, SWMW show that statistical ideas can be profitably used.

The authors address prediction in the sense of developing an interpolating function that can be used economically as a surrogate for the computer model in, e.g., finding the region in the input space that optimizes the output. But computer models are also used to make predictions in the more conventional sense of statements about a possible future outcome, such as the greenhouse effect, nuclear winter or the temperature reached in the core of a nuclear reactor in the event of a hypothesized accident. Inputs to such calculations can be based on data, such as reliability data pertaining to nuclear power plant safety systems, so the output of the computer calculation is a statistical prediction—a function, at least in part, of data. For informed decision-making, we need to be able to say something about the statistical and other uncer-

Robert G. Easterling is Supervisor of the Statistics, Computing and Human Factors Division (7223), Sandia National Laboratories, Albuquerque, New Mexico 87185.

tainty of this prediction. The calculation of standard errors and statistical prediction limits is not at all straightforward, but methods such as the jackknife and bootstrap can be useful (or all we have). These methods are computer-intensive, and we may need to apply them to a surrogate computer model, rather than the actual, so this is another possible application of SWMW's methods.

Experimental design problems also arise in this context in deciding how to collect data pertaining to the inputs in order to most efficiently control or reduce the statistical uncertainty of the computed prediction. The calculation of effects and interactions, as in SWMW's Section 6, is one way to identify influential inputs and thereby guide subsequent data collection. Though these statistical prediction and experimental design aspects of the use of computer models are beyond the scope of SWMW's paper, I mention them to bring them to the readers' attention. Important decisions are being made based on bare point estimates calculated from complex computer models, whose very complexity can endow them with unwarranted credibility and camouflage the lack of data. If we want to strengthen the data foundation of these decisions, we will have to tackle these problems.

Exercising expensive, important computer models calls for a great deal of circumspection, and SWMW exhibit care that is all too rare. In too many areas of application, the standard approach is to take a shotgun approach (Monte Carlo or Latin hypercube sampling), where the shotgun is aimed and loaded with some highly dubious probability distributions (see Downing, Gardner and Hoffman, 1985, and Easterling, 1986). Though the primary objective in these cases may be to approximate the (dubious) distribution of the output, these randomly chosen input sites are also used to fit surrogates and evaluate input effects. When resources are dear, and the objective is to learn something about the complex processes being modeled, it seems almost criminal to me to turn over the exercise of the computer model to a random number generator. We need to use all the statistical and subject-matter intelligence that can be mustered.

SWMW entrust the selection of input sites to computer optimization routines, which may be whimsical but at least are not random. They reject the use of "standard" experimental designs because they "can be inefficient or even inappropriate for deterministic computer codes." This rejection, however, seems to be based on the fact that the subsequently fitted (naive) polynomial models may not provide very good surrogates for the computer code. The fault, though, lies with the model, not necessarily with the design. I think standard designs might provide fits of kriging models, or other interpolators, that are not appreciably worse than fits obtained from SWMW's "optimal" designs, and would offer more-than-offsetting advantages.

For example, consider Figure 1, which is a 16-run computer-selected design in two inputs given in Currin, Mitchell, Morris and Ylvisaker (1988). (For whimsy, note that the computer picked three of the four corner points, one point on three of the edges, two on the other, and points that are roughly diagonal.) About the only place in this design that one can see the effect of one of the t_is while holding the other fixed is along the edges. Being able to evaluate simple effects at many points in the design space seems to me to be a valuable aid in understanding the nature of the complex function being studied.

Consequently, I would prefer a 4^2 design in this example. This design provides many clean, comparable measures of the simple effects of the two inputs, requires zero CRAY time, is geometrically appealing and, for these reasons, should be much easier to sell to the code proprietor or user (unless that person is swayed by the computer-mystique of the "optimal" design). I can only conjecture, but I expect that the resulting fitted kriging model would provide a surrogate that will perform practically as well as one fitted to the Figure 1 design.

Another design that might be considered if subject-matter knowledge suggested that the response was smoother in one direction than the other would be a 3 x 5 arrangement. Such knowledge might also suggest transformations, rotations, etc. We need to turn those black boxes into gray boxes.

Consider next SWMW's Section 6 example of 32 runs with 6 inputs. A "standard" design some might

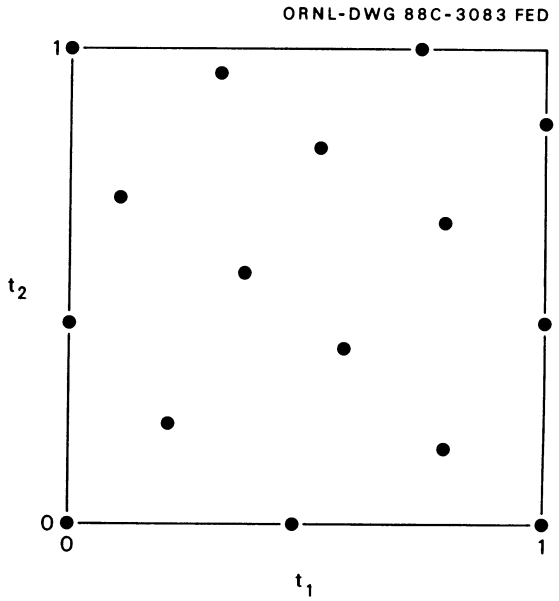

FIG. 1. *Algorithmic-generated design*: 16 *runs*, 2 *inputs*. *Source*: *Figure* 9, *Currin, Mitchell, Morris and Ylvisaker* (1988).

consider would be a 2^{6-1} fraction at corners of the design space, which have coordinates of ±½. I'm sure SWMW would reject this design as a basis for fitting their kriging model and I would too. As an alternative, based on an adaptation of standard designs, I would propose a 2^{6-2} at corners of the cube plus an interior 2^{6-2} at corners of the $(-\frac{1}{4}, \frac{1}{4})^6$ cube. Subject-matter expertise could help choose the particular fractions (and I think subject matter expertise would be better used in this way than in specifying parameters for the covariance function). One might consider some sort of optimization scheme for choosing the inner fraction, given the outer fraction. (I would think that complementary fractions ought to be used. For example, if I = ABCD = CDEF = ABEF is the selected defining contrast for the outer fraction, then I = ABCD = −CDEF = −ABEF would be one of my candidate defining contrasts for the inner.) Additionally, one might optimize the dimension of the inner cube—perhaps the corners should be at ±⅙ instead in order to more uniformly fill in the design space. Or it might help to pull the outer cube points in slightly from the corners of the design space. The optimality problems this approach suggests are of much smaller dimension than those of SWMW, so they ought to be easier to solve, if one is determined to optimize something.

I would encourage statisticians and code analysts to investigate and use adaptations of standard designs such as these I have suggested. The optimal design community sometimes says that optimality criteria shouldn't dictate a design, but rather they should provide a starting point that might lead to a more appropriate design. One doesn't have to do much nudging on the points in Figure 1 to see a 4^2 design emerging. The projection in SWMW's Figure 1 of their example's first 16 points suggests a conventional 2^2-plus-center-point design on an interior cube. After the first 16 points, the authors change their design approach with the result that the subsequent 16 points are forced out toward the edges. So I think that deep down we have similar objectives and concepts of good designs. My experience in this and other contexts is that optimality algorithms seem to be trying to get to a recognizable, reasonable design, but they're so muscle-bound they can't quite make it. Of course, once you realize this, you can skip the CRAY exercise and go directly for a reasonable design.

In both examples, the algorithmic designs are kind of "ugly"—to coin a new technical term. They look like what might emerge from an observational study or if the experimenter could not control the factor settings very well. Surely no one would deliberately design a real experiment this way, so why is it right for a computer experiment? The authors' answer is that the perfect repeatability of a computer run, as opposed to the imperfect repeatability of field or laboratory experiments, makes things, well, different. To me, though, this property negates only the utility of replication. It doesn't cancel out the attractiveness of properties such as balance, symmetry, collapsibility and comparability (of simple effects) that make factorial designs so powerful, informative and "pretty." If the objective was to fit a highly nonlinear model, then an algorithmic design might be called for. But here the model is (or can be)

$$Y = \text{constant plus correlated error},$$

so doesn't it seem right that geometric and space-filling ideas should be used? Again, let's not turn the exercise of computer codes over to a computer program until we've fully applied our statistical and subject-matter expertise.

In closing, though I am skeptical about the proposed experimental designs in the context of computer experiments, I congratulate the authors for this timely, well-written, and thought-provoking paper, and I appreciate the opportunity to help air some of the issues involved. I hope readers will be stimulated to take a statistical look at the use of computer models in their field of application.

ADDITIONAL REFERENCES

Downing, D. J., Gardner, R. H. and Hoffman, F. O. (1985). An examination of response-surface methodologies for uncertainty analysis in assessment models. *Technometrics* **27** 151–163.

Easterling, R. G. (1986). Letter to the Editor. *Technometrics* **28** 91–92.

Comment

Mark E. Johnson and Donald Ylvisaker

A search for new directions to pursue in the Design of Experiments was undertaken at workshops at Berkeley in June of 1984 and January of 1985, and at UCLA in July of 1985, but the thought that design of computational experiments might stand alone as a substantial topic for research can be dated from the January 1986 workshop at UCLA. Subsequent workshops at Urbana in May of 1987 and at Santa Fe in September of 1987 confirmed this promise. The push for (and the early organization behind) this development should be credited to Toby Mitchell.

The paper under discussion does a nice job of capturing the richness and fascination of the subject. It gives a faithful representation of trends in the choice of priors, in the choice of criteria, in the use of cross-validation and maximum likelihood estimation, and in territories for application. It seems appropriate to us that this brief comment concentrate on two general themes: how does the area relate to others which have come before, and what particular contributions might we expect from it in the future? What we take to be the important problems will be mentioned in the course of the discussion.

The philosophical approach of uncertainty measures does indeed go back a long way, as has been well documented in Section 4. Beyond the shared need for a catalog of workable and representative priors, the real problem is that of modification of the prior based on observation and somewhat beyond cross-validation or maximum likelihood over a parametric family. Some clever ideas about modification have been put forward by Mitchell but there is no general method to fall back on in this area.

Monte Carlo is another tool that introduces probabilistic notions for use in a deterministic world. Design can and does play a role in problems of a similar vein. Here we are thinking about settings in which individual evaluation is desired over a vast array of "objects" and, while easy to perform, is still only possible for relatively few of them. Interest might then center on the proper allocation of resources to neighborhoods which are determined by a suitable "distance" between objects, say.

In saturated or super-saturated contexts, Latin hypercube sampling and off-line control are appropriate techniques in the absence of interactions, and some Bayesian methodology seems necessary in any event. Does this paper contain enough evidence, anecdotal or otherwise, to suggest that the present research will establish its own identity, lead to catalogs of useful designs or give real guidance to someone possessed of a like problem? The answer seems to be: not yet.

Our use of the terms object and distance is premeditated. In Johnson, Moore and Ylvisaker (1988) it is shown that, in the absence of good prior knowledge, designs of a geometric type have certain robustness properties. Such robustness is associated with low correlations between observations. Coupled with the thought that few observations mean large separations (surely consistent with low correlations), certain design problems are reduced to more basic geometric ones. In effect, one goes full circle to return to a deterministic question. (Incidently, the IMSE criterion does not show up in a very favorable light in these considerations, while the MMSE criterion surfaces readily and has a natural connection with D-optimality.)

Our preference in any event is to remain free of thinking in terms of regular or stylized design spaces, such as the unit cube in d dimensions. This can be aided by limiting consideration to finitely many sites, hence finitely many designs, while not really violating the spirit of what is feasible. Structure might then be imposed with the notion of distance between sites. The problems become: what distance is appropriate and, most importantly of all, how should the distance chosen undergo modification in the face of observed data? Thus, beginning with the collection of data at distant and suitably chosen sites, where ought one to turn now for further experimentation? This point of view seemingly emphasizes design and plays down the role of analysis of posterior uncertainty. However, the answer selected might well come out of such an analysis.

In summary, we find the area of design for computational experimentation is a lively one and the present article attests well to that. Much thought is still required but, since successful applications continue to emerge, this is certainly a worthwhile enterprise.

Mark E. Johnson is Professor, School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332. Donald Ylvisaker is Director, Division of Statistics, Department of Mathematics, University of California, Los Angeles, California 90024.

Comment

A. Owen, J. Koehler and S. Sharifzadeh

We have been running computer experiments related to semiconductor process design and recently switched over to the paradigm described by the authors. We have found it to be more flexible than response surface methodology in handling deterministic responses.

The Bayesian approach suggests how to interpolate, extrapolate, assess uncertainty and construct designs. To what extent do the Bayesian answers make sense, if one does not hold the prior belief? The authors cite several works in which connections are drawn between well accepted interpolation methods and various priors and give an example in which the uncertainty assessment is accurate. It would be very interesting if the uncertainty assessments were reasonbly accurate for a large class of underlying functions. Have the authors investigated this point? We doubt that the Bayesian method will help in extrapolation (which we suspect should be avoided) and thus are worried that the optimal designs sometimes concentrate near the center of the design space.

Our main comments are directed at the design problem and at estimation of the parameters of the covariance model. Our applications have 5 to 10 input variables and a like number of outputs. The programs we use are fast enough to make it feasible to consider 50 or more runs.

Before addressing the design and estimation issues, we wish to point out that ideas from exploratory data analysis have a role to play in computer experimentation. The authors (with their coworkers) have plotted contours, trajectories and the additive main effects (mentioned in Section 6) of the response functions. We think their contributions are noteworthy and look forward to further developments. When there are many response variables, care should be taken in optimizing a functional of the responses without first considering the tradeoffs among competing goals. The approach taken in Sharifzadeh, Koehler, Owen and Shott (1989) is to evaluate the model functions at thousands of input points and to explore the resulting data set with interactive graphics, in this case S (Becker, Chambers and Wilks, 1988).

DESIGN ISSUES

In the authors' Figure 1, the design points are all quite close to the center. We share the misgivings of the authors, suspecting that this leads to a robustness problem. Extrapolation by conditional expectation depends to a far greater degree on the covariance function used than does interpolation. Thus outside of the convex hull of the data, the predicted values will depend strongly on hard-to-verify properties of the model.

We have been using low discrepancy sequences, mentioned in Section 7.5, as designs. These designs are constructed so that the empirical measure of the design points is close in a Kolmogorov-Smirnov sense to the uniform measure on the cube. These should be good designs in the case of large θ, when estimation is difficult. Johnson, Moore and Ylvisaker (1988) characterize the optimal designs in the large θ limit. Minimizing the maximum distance from a point in the cube to a design point leads to their version of G optimality and maximizing the minimum distance between two sample points leads to their version of D optimality. Low discrepancy sequences (such as Halton-Hammersley sequences) tend to have small, but not minimal, maximum distances from points in the cube.

We have found that sometimes some of the θ_j appear quite small while others are large. That is a response variable is heavily dependent on a few of the d inputs and not very sensitive at all to the others. We may not know in advance which input variables are the important ones or, more commonly, each output variable may depend most strongly on a different small set of inputs. This opens up the possibility of reducing the dimension of the problem by considering the response as a function of the most important inputs, possibly with some noise due to the other inputs. For instance, in our first experiment the thickness of a layer of SiO_2 only depended on the oxidation temperature. Unfortunately, our design (an all-bias design) only used three distinct values of the temperature in 43 runs. If our experiment had had 43 nearly equispaced temperature values, the results would have been more informative.

Low discrepancy designs have the added benefit that when projected onto a cube defined by a subset of the original variables they are still nearly uniform.

A. Owen is Assistant Professor of Statistics, J. Koehler is a Ph.D. candidate in Statistics, and S. Sharifzadeh is a Ph.D. candidate in Electrical Engineering, all at Stanford University. All three are affiliated with the Center for Integrated Systems. Their mailing address is: Department of Statistics, Stanford University, Stanford, California 94305.

Thus if the dimension can be reduced, the design in the remaining dimensions is still reasonably good. The optimal designs depicted in Johnson, Moore and Ylvisaker (1988) do not tend to project uniformly.

We prefer the sequences of Faure (1982) to the Halton-Hammersley sequences. The Halton-Hammersley sequences are usually based on the first d prime numbers, whereas Faure uses the same prime number (the smallest prime $r \geq d$) on each axis. When $n = r^k$, the Faure sequences exercise each input variable in much the same way Latin hypercube designs do. Moreover for $k \geq 2$ they exercise pairs of input variables in that, for any given pair of inputs, one can partition their domain into r^2 squares and find r^{k-2} points in each square. Similarly there are equidistribution properties for three or more axes. The equidistribution properties of the Halton-Hammersley sequences are different for each marginal subcube, depending on the associated primes. We have found that with $n = r^2$ and $r = 5$ or 7 that the Faure sequences appear to lie on planes in three dimensions. This is alleviated by replacing each digit b in the base r representation of the Faure sequence by $\sigma(b)$ where σ is a permutation of $0, \ldots, r-1$. The permutation does not alter the equidistribution properties. One can inspect three-dimensional scatterplots to make sure that a given permutation is effective.

PARAMETER ESTIMATION

We would like to mention a quick way of estimating $\theta_1 \ldots, \theta_d$ in the covariance given by the authors' equation (9) with $p = 1$. When the function $Y(x)$ is nearly additive, we can estimate the main effects using scatterplot smoothers. This corresponds to the inner loop of the ACE algorithm in Breiman and Freidman (1985). Let g_j denote the estimate of the jth main effect. A very smooth $g_j(\cdot)$ is evidence that θ_j is small and a rough $g_j(\cdot)$ suggests that θ_j is large. The roughness may be assessed by $\mathscr{R}_j = \sum_{i=1}^{m}(g_j(i/m) - g_j((i-1)/m))^2$ where the domain of g_j has been rescaled to $[0, 1]$. The expected value of $\mathscr{R}_j$ may be expressed in terms of θ_1 through θ_d, for fixed σ. The d equations in d unknowns can be solved iteratively. The likelihood can be used to choose between the answers from several different values of m. This avoids a high dimensional search for $\theta_1, \ldots, \theta_d$. The first time we tried it, we got better parameter values (as measured by likelihood) than we had found by searching. Alternatively it suggests starting values for such a search.

ADDITIONAL REFERENCES

Becker, R., Chambers, J. and Wilks, A. (1988). *The New S Language*. Wadsworth and Brooks/Cole, Pacific Grove, Calif.

Breiman, L. and Friedman, J. H. (1985). Estimating optimal transformations for multiple regression and correlation. *J. Amer. Statist. Assoc.* **80** 580–598.

Faure, H. (1982). Discrepance de suites associees a un systeme de numeration (en dimension s). *Acta Arith.* **41** 337–351.

Sharifzadeh, S., Koehler, J., Owen, A. and Shott, J. (1989). Using simulators to model transmitted variability in IC manufacturing. *IEEE Trans. Manuf. Sci.* To appear.

Comment

Anthony O'Hagan

The authors are to be congratulated on their lucid and wide-ranging review. Like others before, I have independently rediscovered many of the ideas and results presented here. I therefore sincerely hope that the greater prominence given to those ideas and results by this excellent paper will enable future researchers to start well beyond square one. I first have some comments concerning the derivation of the basic estimator (7), and I will then discuss the model and the practical implementation of the methods from my own experience.

The authors mention three derivations of (7). In a classical framework, it is the MLE if the process $Z(\cdot)$ is Gaussian, and relaxing this assumption it is the BLUP, minimizing (2). Thirdly, it is the posterior mean of $Y(x)$ in a Bayesian analysis with a Gaussian $Z(\cdot)$ and a uniform prior on β. It is first worth pointing out that with a proper multivariate normal prior $\beta \sim N(b, B)$ and known σ^2 the posterior mean of $Y(x)$ has the same form as (7), but with $\hat{\beta}$ replaced by the posterior mean of β, i.e.,

$$\tilde{\beta} = (F'R^{-1}F + \sigma^2 B^{-1})(F'R^{-1}F\hat{\beta} + \sigma^2 B^{-1}b).$$

The interpretation of (7), as comprising the fitted regression model plus smoothed residuals, still holds.

We can also dispense with normality in the Bayesian framework, using a similar device to (2). The

Anthony O'Hagan is Senior Lecturer and Chair, Department of Statistics, University of Warwick, Coventry CV4 7AL, United Kingdom.

same estimator may be derived as the Bayes Linear Estimator (BLE), which also minimizes (2), but to distinguish the different derivations it is important to recognize the conditioning. The BLUP minimizes, over the class of $\hat{y}(x)$ which are unbiased and linear in Y_s,

$$(*1) \qquad E[\{\hat{y}(x) - Y(x)\}^2 \mid \beta, \sigma^2],$$

conditioning on the parameters being mandatory in classical statistics. In contrast, we can think of the posterior mean of $Y(x)$ in general as minimizing (unconstrained) the posterior expected squared error

$$(*2) \qquad E[\{\hat{y}(x) - Y(x)\}^2 \mid Y_s].$$

When all distributions are normal, the posterior mean is (7) with $\hat{\beta}$ replaced by $\tilde{\beta}$ and happens to be linear in Y_s. The BLE minimizes

$$(*3) \qquad E[\{\hat{y}(x) - Y(x)\}^2].$$

over the class of $\hat{y}(x)$ which are linear in Y_s. Only first- and second-order moments need be specified, and the solution is again (7) with $\hat{\beta}$ replaced by $\tilde{\beta}$ and reduces strictly to (7) if $B^{-1} \longrightarrow 0$. We can consider (*3) as the expected MSE, i.e., the expectation of (*1) with respect to the prior distribution of the parameters. We can also consider it as the prior expected squared error, i.e., the expectation of (*2) with respect to the preposterior distribution of Y_s.

There are therefore two Bayesian derivations of (7), paralleling its two classical derivations, in the case of a diffuse prior distribution for β. With proper prior information about β and known σ^2, both yield the same structure as (7) but with $\tilde{\beta}$ replacing $\hat{\beta}$. With unknown σ^2, the posterior mean will no longer be linear in Y_s. The BLE solution is no longer obviously appropriate, but see the variance-modified BLE of Goldstein (1979). The BLE has also, incidentally, been rediscovered several times (see O'Hagan, 1987 and references therein).

My own work on design and analysis of error-free data has been in the context of numerical integration, where the objective has been to make inference about the integral of $Y(\cdot)$. This work is described in O'Hagan (1988). My motivation and practical experience lies in the case where $Y(\cdot)$ is an unnormalized density function over $\mathbf{R}^d$. This is because, in a Bayesian analysis of a complex problem, the posterior density is generally an intractable function and is only known up to a normalizing constant. Integrating the density to obtain this normalizing constant is therefore the first task in analyzing the posterior information. This is a very specific context, and my main comment on the model is that context is very important. My context implies that $Y(\cdot)$ is non-negative and will tail away to zero in all directions fast enough to be integrable. I therefore set $Y(x) = T(x)g(x)$, where $g(\cdot)$ is a fixed, proper density function on $\mathbf{R}^d$ and $T(\cdot)$ is now assumed to follow a model identical to (1). This is very different to assuming (1) for the original process $Y(\cdot)$. There is always prior information about the shape of $Y(\cdot)$. To some extent this is captured in the regression part of (1), but if $Y(\cdot)$ is constrained then we need a model that recognizes both the constraint and the fact that the variability of $Y(\cdot)$ must be reduced when it comes close to the constraint.

My experience with using this model, although very limited, reinforces many of the authors' comments. I simulated a wide range of posterior densities, in one dimension only, as mixtures of normal or t distributions, applied my Bayesian quadrature rules with various p, θ and polynomial regression terms, and calculated sample MSEs. Like the authors, I found that there was generally no benefit in using the regression terms, apart of course from a constant term. Since my functions were quite smooth, it is not surprising that $p = 2$ performed better than $p = 1$.

The authors remark that the apparent precision of prediction is dramatically increased by decreasing θ. I found this too and proposed a general value of $\theta = 1$ for my specific context. I did not attempt to estimate p and θ, but unknown p and θ are not easy to handle within the full Bayesian framework. The authors' maximum likelihood estimates easily translate into posterior modes, assuming uniform prior distributions for these parameters. However, merely substituting these estimates into the rest of the analysis is an approximation to the full Bayesian analysis, at best, and is bound to underestimate posterior uncertainty about $Y(\cdot)$

For design, my optimality criterion was different from any suggested by the authors. I was interested in posterior variance of the integral. Just as $\operatorname{var}(X + Y) \neq \operatorname{var}(X) + \operatorname{var}(Y)$ in general, this variance is different from integrated MSE. It takes account of posterior covariances between $Y(x)$ and $Y(w)$, which in the classical framework would be replaced by covariances between $\hat{y}(x)$ and $\hat{y}(w)$. Despite the different criterion, my experiences with optimal design were similar to the authors'. In particular, for $d = 2$, the few optimal designs I derived were quite unlike traditional quadrature rules.

The authors comment that the conditioning of R deteriorates with n. This is a serious problem when searching for designs, because R is ill-conditioned over a great part of the design space, namely wherever two coordinates are sufficiently close in value. The problem is much worse for $p = 2$ than for $p = 1$. However, good designs invariably arise in that part of the design space where R is relatively well-conditioned. It may be possible to take account of this in the search algorithm, to both speed the search and evade numerical problems on the way. Nevertheless, large n must always present problems.

The only comment in the paper which jars with my own experience is the reference to designing for very large θ, in the Currin, Mitchell, Morris and Ylvisaker (1988) paper. When θ is large, you cannot estimate $Z(\cdot)$ except very locally to each design point. The second part of (7), which smooths the residuals, consists of zero almost everywhere except for blips at each design point to make $y(x)$ pass through the observation. Designs for this case will be exclusively concerned with estimating the regression function and, like classical optimal design for regression, will place clusters of points at the boundaries of the design region. Such designs must be very poor when θ is in reality not large.

I was very intrigued to see the decomposition of $Y(\cdot)$ into main effects, interactions, etc. In my context where $Y(\cdot)$ is a multivariate density function, the main effects are just marginal densities. The interactions as defined, however, have no particular value. Instead I would define

$$\mu_{ij}(x_i, x_j) = \int y(x) \prod_{h\neq i,j} dx_h - \mu_i(x_i)\mu_j(x_j),$$

representing non-independence between x_i and x_j.

It should be clear from my remarks how much I have enjoyed reading this paper. The wealth of detail and the authors' breadth of knowledge make it one that I am sure to turn to repeatedly.

ADDITIONAL REFERENCES

GOLDSTEIN, M. (1979). The variance modified linear Bayes estimator. *J. Roy. Statist. Soc. Ser. B* **41** 96–100.

O'HAGAN, A. (1987). Bayes linear estimators for randomized response models. *J. Amer. Statist. Assoc.* **82** 580–585.

O'HAGAN, A. (1988). Bayesian quadrature. Warwick Statistics Research Report 159, Univ. Warwick.

Comment

Michael L. Stein

I wholeheartedly agree with the authors that statisticians can and should contribute to the design and analysis of computer experiments. Too often statisticians shy away from problems that do not fit into the standard statistical frameworks; the authors are to be congratulated for their trailblazing efforts. Furthermore, I agree that a sensible way to approach these problems is to view the output from the computer model as a realization of a stochastic process. Where I think further work is needed is in the development of appropriate stochastic models.

The model given by (9) in this article by Sacks, Welch, Mitchell and Wynn has some undesirable properties. For $0 < p < 2$, a stochastic process with this covariance function will not be mean square differentiable. As noted by the authors, for $p = 2$, the process is infinitely mean square differentiable. Not allowing processes that are differentiable but not infinitely differentiable strikes me as unnecessarily restrictive. A more flexible class of correlation functions is (Yaglom, 1987, page 139)

$$\prod \frac{1}{\Gamma(\nu)2^{\nu-1}} (\alpha_j \,|\, w_j - x_j \,|)^\nu K_\nu(\alpha_j \,|\, w_j - x_j \,|),$$

where K_ν is a modified Bessel function of order ν (Abramowitz and Stegun, 1965, page 374). A stochastic process with this covariance function will be m times mean square differentiable if and only if $\nu > m$. The α_js measure the range of the correlation: a large α_j indicates that correlations die out quickly in the x_j direction.

A problem with all of the correlation functions used by Sacks, Welch, Mitchell and Wynn is that they do not allow for the inclusion of prior knowledge such as that most of the variation in the output $y(\cdot)$ can probably be explained by main effects plus perhaps some low order interactions, which in fact occurred in the circuit simulator example they discuss. If we expected most of the variation in $y(\cdot)$ could be explained by main effects, we might want to model $Y(x)$ as

$$Y(x) = \sum Y_j(x_j) + Z(x), \tag{1}$$

Michael L. Stein is Assistant Professor, Department of Statistics, University of Chicago, 5734 University Avenue, Chicago, Illinois 60637.

where the Y_j's and Z are independent Gaussian processes with covariance functions $\sigma_j(x_j - w_j)$ and $\sigma_z(x - w)$ respectively, so that

$$\operatorname{cov}(Y(x),\ Y(w)) = \sum \sigma_j(x_j - w_j) + \sigma_z(x - w).$$

One specific parametric form of this model that might be worth exploring is

$$\begin{aligned}\operatorname{cov}(Y(x),\ Y(w)) &\\ &= \sum C_j(\alpha_j \mid w_j - x_j \mid)^\nu K_\nu(\alpha_j \mid w_j - x_j \mid)\\ &\quad + D \prod (\beta_j \mid w_j - x_j \mid)^\nu K_\nu(\beta_j \mid w_j - x_j \mid).\end{aligned}$$

A large C_j would correspond to an important main effect. The model for $Z(\cdot)$ is somewhat problematic as it allows $Z(\cdot)$ to have an additive component. Following the decomposition into main effects and interactions from Section 6 of the article by Sacks, Welch, Mitchell and Wynn, it might be more satisfying to define $Z(\cdot)$ to be a stochastic process with no additive component:

$$\begin{aligned}Z(x) = Z^*(x) - \sum_j \int Z^*(x) \prod_{h \neq j} dx_h&\\ + (d-1) \int Z^*(x)\,dx,&\end{aligned}$$

where d is the number of dimensions of x and $Z^*(x)$ is a Gaussian process with some simple covariance function. I think it would be very interesting to find optimal designs under some models of the general form given by (1). If the optimal designs are very different from those obtained by Sacks, Welch, Mitchell and Wynn for their models, that would call into question the effectiveness of their designs for processes where most of the variation can be explained by main effects.

ADDITIONAL REFERENCES

ABRAMOWITZ, M. and STEGUN, I. (1965). *Handbook of Mathematical Functions*, 9th ed. Dover, New York.

YAGLOM, A. M. (1987). *Correlation Theory of Stationary and Related Random Functions* **1**. Springer, New York.

Rejoinder

Jerome Sacks, William J. Welch, Toby J. Mitchell and Henry P. Wynn

We thank the discussants for their incisive comments, suggestions and questions. Nearly all the discussants have been key participants at the workshops mentioned by Johnson and Ylvisaker; all have been instrumental in the development of new methodologies for the design and analysis of computer experiments. Most of the comments and our responses are concerned with the choice of the experimental design and the choice of the correlation function.

We had hoped that the example of Section 6 would attract some suggestions from the discussants, and in this we are not disappointed. Morris' results on the first-stage, 16-point design are interesting—in particular, they indicate that the concentration of the design in the center of the region also occurs for the much rougher process corresponding to $p = 1$ in (9). As this is only a preliminary stage, and there is not much to be lost by using a cheaper design anyway, his scaled quarter fraction makes a lot of sense. In a seven-dimensional problem, Sacks, Schiller and Welch (1989) similarly reduced the optimization problem by restricting attention to scaled central-composite designs. Without doing the optimization or amassing experience from many problems, though, we cannot know when the relative performance of cheap designs will be satisfactory.

For all 32 runs, Easterling recommends two complementary quarter fractions. He rightly points out the advantage of not having to optimize anything, and we tried these fractions on $\{-\frac{1}{2}, \frac{1}{2}\}^6$ and $\{-\frac{1}{4}, \frac{1}{4}\}^6$. In some recent applications where data are cheap to generate, we have been using Latin hypercube designs, and for comparison we also report results for a 32-run Latin hypercube. The six factors have the same 32 equally spaced values, $-\frac{31}{64}$, $-\frac{29}{64}, \ldots, \frac{31}{64}$, but in different random orders. For both designs, the predictor is based on model (14) after re-estimating the parameters $\theta_1, \ldots, \theta_6$ and p in the correlation function (9). Table R1 shows the average squared error of prediction at the same 100 random points we used previously. For ease of comparison, the results for our original design are repeated. The complementary quarter fractions and the Latin hypercube perform similarly, with our design showing a modest advantage.

It is of interest to note that, for certain values of n and d, scaled standard designs can be optimal. For 8 points in 4 dimensions and 16 points in 5 dimensions the optimization problem is still tractable. Suitably scaled half fractions of maximum resolution are apparently optimal or very close to optimal for the IMSE criterion when the model has only a constant term for the regression, and the correlation family is (9) with $p = 2$. The scaling of the design depends on the value of θ. (We regret that we did not include more of these anecdotes.) Extrapolating these results to the six-dimensional problem at hand, we tried a half fraction (I = ABCDEF) on $\{-0.37, 0.37\}^6$, a scaling associated with a small value of θ. As seen in Table R1, this design performs much the same as the complementary quarter fractions. It is quite likely that a Faure sequence, favored by Owen, Koehler and Sharifzadeh, or a geometrical design, as in Johnson, Moore and Ylvisaker (1988), would also perform similarly.

These very different designs produce rather similar results, then. In this respect, our choice of example was less revealing than we had hoped. Whether this is a feature of the particular surface or something more general we can only guess.

In other examples optimal design provides greater benefits. The 4^2 design suggested by Easterling for the problem in Currin, Mitchell, Morris and Ylvisaker (1988) performs relatively poorly. For the cubic correlation function (11) with parameters determined by maximum likelihood, the empirical average squared error of prediction on a 21×21 grid is $(0.62)^2$ for the optimal design shown in Easterling's figure, compared to $(0.94)^2$ for the 4^2 design. Although this is just one example, it does indicate that the well balanced, symmetric design does not necessarily perform best, and the difference is not necessarily trivial. Incidentally to us, the Currin, Mitchell, Morris and Ylvisaker design seems rather elegant—beauty of a design is in the eye of the beholder.

One very important place for the use of optimal design is for less-regular regions where intuition is lacking. Ongoing work by D. Cox, J. Park and C. Singer on a computer model for a nuclear-fusion device (Tokamak) has a six-dimensional region which is determined by linear constraints. The cost of generating data is also fairly high, 3–5 minutes on a Cray 2 per run. Here, no regular, symmetric designs easily come to mind.

TABLE R1
Empirical average squared error of prediction at the 100 *random points*

Design	Average squared error
Authors'	$(.122)^2$
Complementary quarter fractions	$(.146)^2$
Latin hypercube	$(.136)^2$
Half fraction on $\{-0.37, 0.37\}^6$	$(.143)^2$
Authors' (Stein's correlation function)	$(.115)^2$

We are intrigued by O'Hagan's experience in applying similar models to the estimation of integrals. We have some unpublished results relating to quadrature in two dimensions. Low-discrepancy sequences (e.g., Halton sequences) perform well on the average for functions generated by the model (9) when $p = 1$. However, for functions generated by (9) with $p = 2$, which are much smoother, the average performance of the Halton sequences is poor relative to the optimal designs obtainable for small problems or relative to various *ad hoc* schemes for larger problems.

Morris' connections between interactions and the correlation parameters θ in (9) suggest small rather than large values of θ, a view shared by O'Hagan. Our experience with estimating these parameters in a number of examples, using models with no regression terms other than a constant, is also consistent with smaller values. This stands in contrast to Johnson and Ylvisaker's results that *designs* based on large θ have certain robustness properties. How efficient their designs are when small values of θ are appropriate, or when linear models are incorporated, is not clear.

To sum up on the choice of design, we suspect with Easterling that, providing that the design does a reasonable job of infiltrating the space, the predictor is probably more important than the design. Unfortunately, we do not have enough evidence to strengthen these suspicions, nor to make notions like infiltration more precise. Sometimes standard designs like fractional factorials do fairly well; for other problems they perform rather poorly. To know which case is true, the optimal design is necessary as a benchmark. Johnson and Ylvisaker correctly point out that we do not yet have catalogs of useful designs that can be automatically applied. Clearly, more work is needed, and advanced computations seem indispensable.

Stein suggests a more flexible class of correlation functions. We took the design and data of Table 1 and maximized the likelihood over $\alpha_1, \ldots, \alpha_6$ and $\nu = 1$, 2, 3. Predictions based on $\hat{\nu} = 1$ and $\hat{\alpha} = (.260, .255, .446, .566, .466, .934)$ at the 100 random points give the results reported in Table R1. There is some improvement and this family may be worth pursuing further. We do note, however, that optimization of the likelihood is more costly, and there may be numerical instabilities associated with computing K_ν as ν grows.

Stein's additive covariance models seem promising. Though they introduce a number of additional parameters (the C's and D), we agree they may be useful when the output is nearly additive. There may also be important design considerations.

Any help in estimating the correlation parameters is welcome, and we look forward to seeing further details of the method outlined by Owen, Koehler and Sharifzadeh. These discussants also note that smaller θ's may indicate inactive factors, with implications for dimension reduction. This is usually valid, but we have found, for example, cases in which $\hat{\theta}$ is close to zero for a variable with a strong linear effect. This is backed up by theoretical work on asymptotic behavior as $\theta \rightarrow 0$ when $p = 2$ in (9), as mentioned in Section 7.4. To avoid overinterpretation of the θ's we endorse the plotting of the main effects, interactions and so on defined in Section 6.

Black box or gray box? We could not agree more with Easterling about the need to employ subject-matter expertise. Progress in applications and new methodologies requires two-way exchange between statisticians and the scientists conducting these experiments. In our experience, as in the example of Section 6, the expert has usually reduced the number of factors to a small set, most of which are active for one response or another. To ensure that all important factors are included, however, a screening stage might be used to determine the active set empirically. In this case, as Owen, Koehler and Sharifzadeh point out, designs that project well for one or relatively few active factors will also be more useful for prediction. Overly symmetric designs like fractional factorials may have replicates when projected in this way. It might be helpful to incorporate prior information on effects sparsity into the assumed models, with implications for design.

O'Hagan sheds some more light on the Bayesian viewpoint here, to which he has made important contributions, and Morris points out some difficulties with the frequentist interpretation. In earlier drafts we did attempt to discuss these philosophical matters more fully, but we gave up due to differences amongst the authors! A full Bayesian viewpoint might offer some advantages. Unfortunately, as O'Hagan points out, unknown correlation parameters are not easy to deal with in a full Bayesian framework.

We are grateful that the discussants share and reinforce our excitement in developing this area. It is clear to us that there is much work to be done; we hope that there will be many to do it.