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

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

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

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

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

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

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

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

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,

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

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

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

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

$$R(w, x) = \exp(-\theta \,\|\, w - x \,\|), \tag{12}$$

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 ($n \to \infty$) 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 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

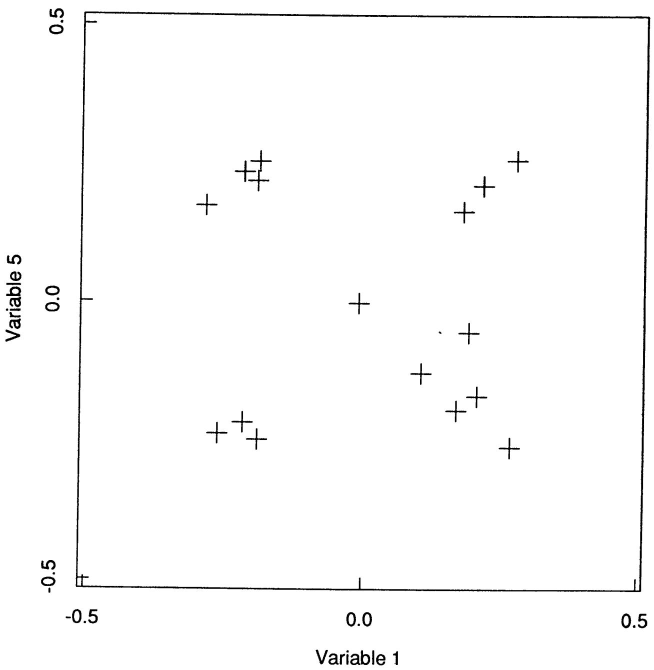

FIG. 1. *Projection of the experimental design onto the coordinates of two input variables.*

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-

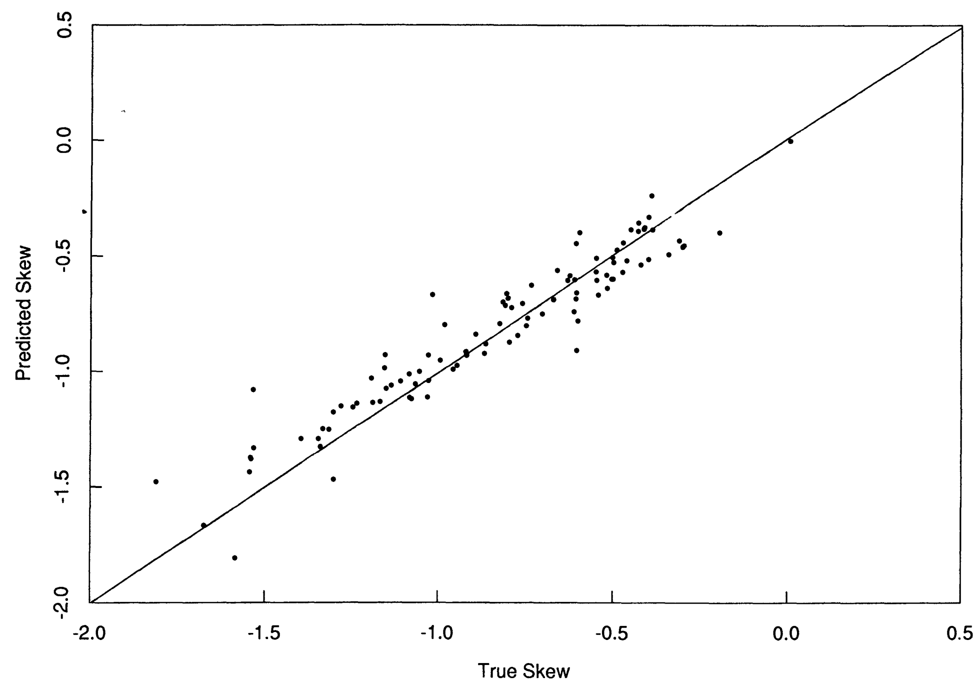

FIG. 2. *Q − Q plot of the standardized residuals against standard normal quantiles.*

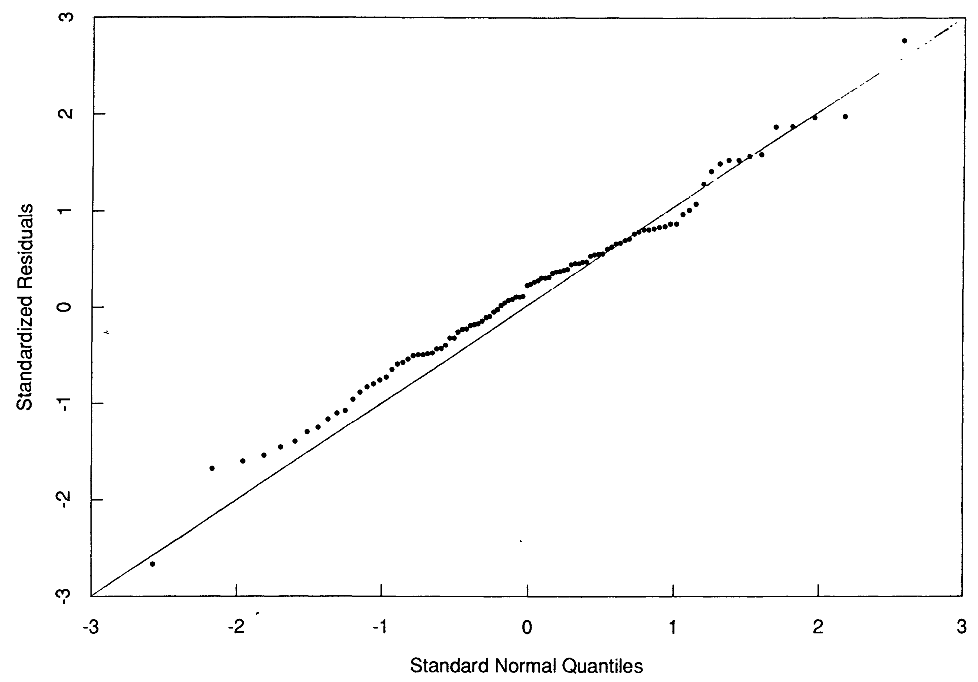

FIG. 3. *Predicted clock skew plotted against true clock skew at 100 random sites.*

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], Curran, 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 $\mathrm{MSE}[\hat{y}(x)]$ 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 $\mathrm{MSE}[\hat{y}(x)]$ 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 $9 \le n \le 25$, $p = 1.6$ 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 $\mathrm{IMSE}(S_\theta)/\mathrm{IMSE}(S)$, 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 nonsmooth 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 $\mathrm{Var}[Y(x)] = \sigma^2 + \sigma_\varepsilon^2$, 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.