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 $\mathrm{Corr}(Y_{++}, Y_{+-}) = \mathrm{Corr}(Y_{+-}, Y_{--}) = e^{-\theta} = \rho$, then $E(W_+) = E(W_-) = 0$, and $\mathrm{Corr}(W_+, W_-)$ 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 $\sqrt{\mathrm{IMSE}}$ 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, $\sqrt{\mathrm{IMSE}}$ 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

	$\sqrt{\mathrm{IMSE}}$	
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 $\sqrt{\text{IMSE}}$ 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 $-½$. $\sqrt{\text{IMSE}}$ 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.

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

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-