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-

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

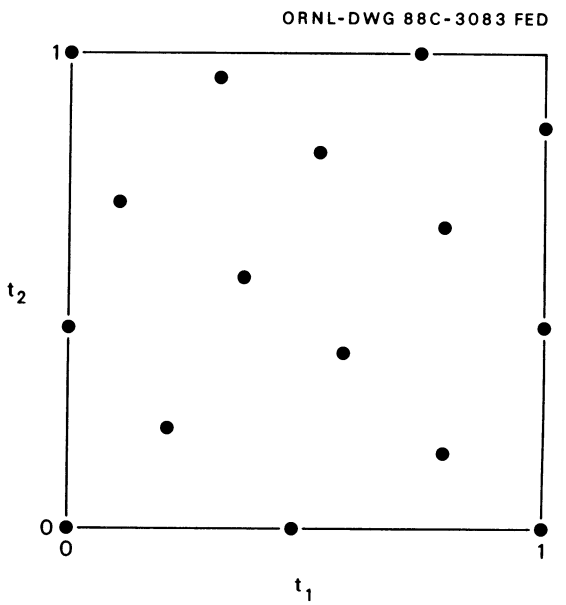

FIG. 1. *Algorithmic-generated design: 16 runs, 2 inputs. Source: Figure 9, Currin, Mitchell, Morris and Ylvisaker* (1988).

consider would be a 2^{6-1} fraction at corners of the design space, which have coordinates of $\pm\frac{1}{2}$. 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 $\pm\frac{1}{6}$ 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.