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

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

PARAMETER ESTIMATION

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

ADDITIONAL REFERENCES

BECKER, R., CHAMBERS, J. and WILKS, A. (1988). *The New S Language*. Wadsworth and Brooks/Cole, Pacific Grove, Calif.

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

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

SHARIFZADEH, S., KOEHLER, J., OWEN, A. and SHOTT, J. (1989). Using simulators to model transmitted variability in IC manufacturing. *IEEE Trans. Manuf. Sci.* To appear.

Comment

Anthony O'Hagan

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

The authors comment that the conditioning of R deteriorates with n. This is a serious problem when searching for designs, because R is ill-conditioned over a great part of the design space, namely wherever two coordinates are sufficiently close in value. The problem is much worse for $p = 2$ than for $p = 1$. However, good designs invariably arise in that part of the design space where R is relatively well-conditioned. It may

be possible to take account of this in the search algorithm, to both speed the search and evade numerical problems on the way. Nevertheless, large n must always present problems.

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

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

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

representing non-independence between x_i and x_j.

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

ADDITIONAL REFERENCES

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

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

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

Comment

Michael L. Stein

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

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

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

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

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

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

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