TECHNOMETRICS, FEBRUARY 1994, VOL. 36, NO. 1

A Simple Bayesian Modification of *D*-Optimal Designs to Reduce Dependence on an Assumed Model

William DuMouchel
Department of Biostatistics
Columbia University
New York, NY 10032

Bradley Jones
The Math Works, Inc.
Natick, MA 01760

D-optimal and other computer-generated experimental designs have been criticized for being too dependent on an assumed statistical model. To address this criticism, we introduce the notion of empirical models that have both primary and potential terms. Combining this idea with the Bayesian paradigm, this article proposes a modification of the *D*-optimal approach that preserves the flexibility and ease of use of algorithmic designs while being more resistant to the biases caused by an incorrect model. These designs provide a Bayesian justification for resolution IV designs. Several theoretical examples and a practical example from the literature demonstrate the advantages of the proposed method.

KEY WORDS: Computer-generated designs; Experimental design; Minimum-bias designs; Mixture designs.

1. INTRODUCTION

1.1 Overview of the Problem and Its Solution

Assume the usual linear model $Y = X\beta + \varepsilon$, $\varepsilon \sim N(0, \sigma^2 \mathbf{I})$, where Y is an $n \times 1$ column vector of responses, X is an $n \times p$ model matrix formed from predictor terms, and β and ε are $p \times 1$ and $n \times 1$ vectors of regression coefficients and errors, respectively. Let b be the least squares estimate of β. Then $b = (X'X)^{-1} X'Y$, $\text{var}(b) = (X'X)^{-1} \sigma^2$. A *D*-optimal design maximizes $\det(X'X)$. The idea is to choose X in such a way that the uncertainty about β is small after observing Y, and increasing the determinant of $X'X$ generally reduces the error variances of the coefficients, which are proportional to $(X'X)^{-1}$.

In this definition, the design matrix X has one column for each term in the assumed model. This model provides an approximation of the behavior of the experimental system. It is natural to ask how good that approximation is. But, in maximizing $\det(X'X)$, all of the experimental effort is being expended on the precise estimation of the assumed model. So the *D*-optimal design makes no explicit provision for diagnosing inadequacies in the model.

In practice, experimenters often add centerpoints and other checkpoints to both factorial and optimal designs to help determine whether the assumed model is adequate. Although this is good practice, it is also ad hoc. This article proposes to improve on this ad hoc practice while supplying a theoretical foundation and an easy-to-apply algorithm for choosing a design robust to the modeling assumptions.

The purpose of checkpoints in a design is to provide a detection mechanism for higher-order effects than are contained in the assumed model. Call these higher-order terms *potential terms*. The assumed model consists of the *primary terms*. Typically, the sample size is not large enough to estimate all of the primary and potential terms simultaneously. The problem is to develop an approach that allows the precise estimation of all of the primary terms while providing omnibus detectability (and some estimability) for the potential terms.

In Section 2 ("The Bayesian Model"), we combine the primary and potential terms into one design matrix. This matrix generally has more columns than rows due to the addition of the columns for the potential terms. The classical approach becomes intractable at this point due to the singularity of $X'X$. The Bayesian formulation of the design problem, however, including the use of prior information about the potential terms, circumvents this difficulty. A key aspect of our solution is the scaling and centering of the terms in the model to permit the use of a standard prior distribution that has worked well in all prob-

lems considered so far. In Sections 3 and 4, we present a series of examples covering a wide variety of situations. Finally, in Section 5, we conclude by comparing this and other approaches to creating bias-resistant designs.

1.2 Computer-Generated Designs

D-optimal designs have the property that they minimize the volume of confidence ellipsoids for parameters. Many other numerical criteria have been proposed as alternative metrics for the worth of a design, the goal being to reduce the problem of choosing a design to a maximization or minimization problem. Most other criteria also can be represented as functionals of the matrix $(X'X)^{-1}$. The purpose here is not to elevate any one criteria over the others, since the criticism that *D*-optimal designs are too dependent on an assumed model relating the response to the factors applies to all such algorithmically defined designs. See St. John and Draper (1975), Hahn, Meeker, and Feder (1976), Silvey (1980), and Steinberg and Hunter (1986) for general discussions of optimizing criteria.

To evaluate a design for *D* optimality (or a similar criterion), you must specify a model, a region of the factors, and a sample size. The numerical computation of *D*-optimal designs begins with a starting design and improves it iteratively. The numerical search is difficult for large problems because most search algorithms are prone to converge to a local optimum (Welch 1982; Wu and Wynn 1978). It is our experience that the value of the determinant at the false optimum is typically quite close to that at the global optimum. If the factors in the experiment are quantitative, then the search space is technically a continuum, but it is common to discretize the set of possible candidate experimental runs. See Galil and Kiefer (1980), Mitchell (1974), Cook and Nachtsheim (1980, 1989), and Meyer and Nachtsheim (1991) for more complete discussions of the numerical search problem.

1.3 Advantages of Computer-Generated Designs

The chief advantage of an algorithmic approach like *D* optimality is its flexibility. For example, one might want to define a quadratic response surface model involving four factors but only include a subset of the 10 second-order terms in the model. Arbitrary block-size restrictions could be applied to such a model, or a categorical factor with an arbitrary number of levels could be included in the model. The *D*-optimal algorithm can tailor a design to any such model. There is also no restriction on sample size so long as $n \geq p$ so that a design can make best use of whatever resources are available rather than, for example, requiring that n be a power of 2. See also Johnson and Nachtsheim (1983) and Snee (1985).

Perhaps the best-known advantage of the algorithmic approach to design is its ability to handle noncubic and nonspherical design spaces. These are especially common in the designs for exploring mixtures, in which the experimental factors are proportions that are often each constrained within narrow ranges and also must sum to a constant. See Section 4 for an extended example.

1.4 Criticisms of the Algorithmic Approach

Criticisms of the *D*-optimal and other such algorithmic design methods usually center on the dependence of the design on an assumed model, as discussed, for example, by Box and Draper (1987). Properties like orthogonality are claimed to be desirable per se, without regard to a specific model equation. For example, a resolution IV design (Box, Hunter, and Hunter 1978; Montgomery 1991) has orthogonality properties that protect main-effect estimates against bias due to interactions. Criteria that focus on minimizing the standard errors of coefficient estimates do not allow for the necessity of checking that the model is correct and so may not include centerpoints when investigating a first-order model or, in the extreme, may have just p distinct runs with no degrees of freedom for lack of fit.

The flexibility in choice of sample size may be viewed as a mixed blessing because a naive user of the *D*-optimal approach might choose a sample size arbitrarily and miss out on the benefits of a balanced design. For example, fractional factorial designs are usually *D*-optimal for their sample size and a particular model, but you have to choose the right sample size to get them from a *D*-optimality algorithm. Another problem is that research on algorithmic approaches to design in cases with multiple variance components is just beginning, as in the work of Smith and Verdinelli (1980).

There have been some attempts to retain the flexibility of the *D*-optimal approach while avoiding the criticisms mentioned previously. Box and Draper (1959, 1987) discussed minimum-bias designs and showed that a design that matches certain moments of the design space has desirable properties. Galil and Kiefer (1977) compared these designs to *D*-optimal designs. Cook and Nachtsheim (1982) developed designs that have high efficiency when several families of models are being considered. Welch (1983) described an experimental design algorithm based on a mean squared error criterion. See also O'Hagan (1978). This article presents a Bayesian modification of *D* optimality that allows the experimenter to "hedge bets" about an assumed model.

2. THE BAYESIAN MODEL

2.1 Consider Two Classes of Terms in the Design Model

Suppose that in addition to the p *primary* terms that you really want to fit, there are q *potential* terms that are just possibly important. Let $X = (X_{pri}|X_{pot})$, where X has $p + q$ columns in all. Unfortunately, the sample size n is usually not large enough to estimate all $p + q$ coefficients, because, typically, $p < n < p + q$. This is a generalization of the situation in which a fractional factorial resolution IV design works well, assuming that X_{pri} represents the main-effects model and X_{pot} represents the two-factor interactions. Even though the two-factor interactions cannot all be estimated, the estimates of the main effects are relatively unaffected by their presence, and careful analysis of the data can often detect and narrow down the search for active interactions.

Since $n < p + q$, it is necessary to add prior information to avoid a singular estimation problem.

2.2 Assumptions on the Scaling of Factors and Responses

To choose an appropriate prior distribution, it helps to have certain conventions about the scaling of the responses and terms. The purpose of this scaling is to be able to interpret the effect of a potential term consistently as equal to its coefficient, in units of the residual-error standard deviation, where "effect" means a predicted change in the response that cannot be accounted for by the primary terms. In our experience, a preliminary scaling and centering is necessary to permit the use of a standard prior distribution on the coefficients to work well over diverse contexts. Without loss of generality, assume that the residual-error variance of the response = 1, that each nonconstant primary term varies from -1 to 1, that each potential term is approximately uncorrelated with all primary terms, and that the range of each potential term is unity. That is, $\sigma^2 = 1$, and for each nonconstant primary term $\max\{X_{pri}\} = 1$, $\min\{X_{pri}\} = -1$, and for each potential term $\max\{X_{pot}\} - \min\{X_{pot}\} = 1$, and for each pair of primary and potential terms

$$\sum_{\text{candidates}} X_{pot}X_{pri} = 0,$$

where the max, min, and sum are taken over the set of candidate points for the design.

As an illustration of these conventions, assume that $x \in \{-1, -.5, 0, .5, 1\}$, the primary terms are $(1, x)$, and the potential terms are x^2 and x^3. Then the potential terms need to be centered and scaled as $z_1 = x^2 - .5$ and $z_2 = (x^3 - .85x)/.6$. In practice, this scaling can be achieved by performing a regression of the potential terms on the primary terms, using the candidate points (or any other desired reference distribution) to compute α and Z, where $\alpha = (X_{pri}'X_{pri})^{-1}X_{pri}'X_{pot}$, $R = X_{pot} - X_{pri}\alpha$, and $Z = R/(\max\{R\} - \min\{R\})$, and where the max, min, and determination of α use the candidate set of points. (The preceding max and min are taken separately for each column of R.) Then the definition of X becomes $(X_{pri}|Z)$ instead of $(X_{pri}|X_{pot})$. Note that α is the least squares regression coefficient of X_{pot} on X_{pri} and R is the residual from this regression. To use design-of-experiments terminology, α is the alias matrix measuring how confounded X_{pot} is with X_{pri} over the candidate set. Replacing X_{pot} by R (or Z) eliminates the aliasing over the candidates, and this orthogonality makes the interpretation of prior distributions easier.

2.3 Prior Distribution of Coefficients

Since the primary terms are likely to be active and no particular directions of their effects are assumed, the coefficients of primary terms are specified to have a diffuse prior distribution—that is, an arbitrary prior mean and a prior variance tending to infinity.

On the other hand, the potential terms are by definition thought to be unlikely to have large effects. Therefore, all coefficients of potential terms are specified to have a prior mean of 0 and a finite variance—the $N(0, \tau^2\mathbf{I})$ distribution is convenient. The value $\tau = 1$ is a suitable default choice, implying that the effect of any potential term is not expected to be much larger than the residual standard error, assumed equal to 1 by our scaling assumption. In general, a D-optimal design is invariant with respect to a linear transformation of the terms of the model. The use of this informative prior distribution for the coefficients in which the coefficients of potential terms are assumed independent with a common variance, however, requires a careful definition of these terms, as previously described, on how to scale and center them.

As a rationale for choosing τ as low as 1, one might argue that any effects of potential terms larger than the residual standard deviation would have been noticed during previous observation of the system and so should be included as primary terms, not potential terms.

2.4 Posterior Distribution of Coefficients

Let K be the $(p + q) \times (p + q)$ diagonal matrix whose first p diagonal elements are equal to 0 and whose last q diagonal elements are 1. Then, since $\sigma = 1$, $Y|\beta \sim N(X\beta, \mathbf{I})$, and, by Bayes's rule for conjugate normal distributions (Box and Tiao 1973; Lee 1989), $\beta|Y \sim N(b, [X'X + K/\tau^2]^{-1})$, where $b = [X'X + K/\tau^2]^{-1}X'Y$.

2.5 Bayes *D*-Optimal: Minimize Determinant of Posterior Covariance Matrix

The Bayesian analog to a *D*-optimal design in the case that prior information is available is the design that numerically maximizes the determinant of $(X'X + K/\tau^2)$ with respect to X. A major advantage of this approach is that most *D*-optimal search algorithms need very little modification to solve this problem. In the Detmax or K-Exchange algorithms (Cook and Nachtsheim 1989; Mitchell 1974) the value of $(X'X)^{-1}$ for the starting design is replaced by $(X'X + K/\tau^2)^{-1}$, and otherwise the algorithms, which are based on updating the matrix $(X'X)^{-1}$ and its determinant as candidate points are switched in and out of the designs, are unchanged.

This Bayesian definition of *D* optimality is standard in the Bayesian design literature; see, for example, Lindley (1956), Chaloner (1984), and Pilz (1991). We are not aware of previous proposals to use a proper prior distribution routinely to attack singular design problems and to make the *D*-optimal approach less model-dependent. See, however, Steinberg (1985) and Draper and Guttman (1992), who used similar models to help decide how to scale classical designs for the purpose of reducing prediction mean squared error.

2.6 Analysis Strategy

A fully Bayesian analysis would use the posterior distribution of all $p + q$ elements of β as given previously. The posterior mean $b = b(\tau)$ is interpreted as a Bayesian shrinkage estimator of the coefficients or merely as a ridge regression estimator. See Efron and Morris (1973), Oman (1984), and Vuchkov (1977) for discussion of such estimators. One can examine the ridge trace—namely, $b(\tau)$ as τ varies—to evaluate the sensitivity of the parameter estimates to the prior distribution. Another, more classical, approach just uses K/τ^2 as a device for getting a design less dependent on the choice of primary terms but then fits the model containing primary terms only by least squares. The effect of potential terms can then be investigated by forward stepwise selection or other methods of regression diagnostics.

3. THEORETICAL EXAMPLES

These four examples represent simple situations in which the naive *D*-optimal approach does poorly but the proposed Bayesian modification works well. The next section contains a more realistic application based on a case study from the literature.

3.1 Example 1: Two Quantitative Factors

Suppose that A and B are two quantitative factors and an interaction model is assumed for an experiment in which $n = 5$. The design points are (a_i, b_i), $i = 1$–5, with a discretized 5×5 set of candidate points: $a_i, b_i \in \{-1, -.5, 0, .5, 1\}$, as in Figure 1. As shown in the left part of the figure, the *D*-optimal design concentrates on the four cornerpoints, replicating one of them. When the two pure squares are the potential terms in the Bayesian model (with $\tau = 1$), the design is as shown on the right, providing a Bayesian rationale for including a centerpoint.

To review the calculations in this example, the ith row of $X_{(5\times 6)}$ is $(1, a_i, b_i, a_ib_i, a_i^2 - .5, b_i^2 - .5)$. The first $p = 4$ columns of X are primary terms. The last $q = 2$ columns of X are potential terms, so $p = 4 < n = 5 < p + q = 6$. Then we choose (a_i, b_i), $i = 1$–5, to maximize the determinant of $(X'X + K)$, where

$$K = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}.$$

This selection of a centerpoint occurs when the default value of $\tau = 1$ is used. If K is replaced by K/τ^2 in the preceding calculation, the centerpoint is included whenever $\tau > .61$, but the design is concentrated on the corners whenever $\tau < .61$ (to two digits). There is an abrupt change in the Bayesian *D*-optimal solution at about $\tau = .612$ so that only the five candidate points at a corner or at the center are ever in the design.

3.2 Example 2: Two Constrained Factors

The second example is a modification of the first, in which the candidate set is restricted to the 15 points of a 5×5 triangular region, $n = 9$, and we assume that the primary model of interest is the full quadratic model so that $p = 6$. As depicted in the left part of Figure 2, the *D*-optimal design replicates each of the

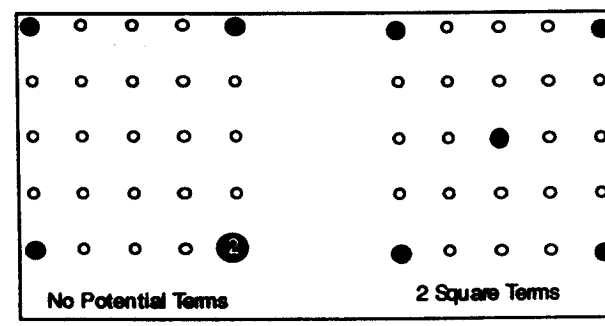

Figure 1. D-Optimal Design for Interaction Model (left) and Bayesian Version (right) if n = 5 and Potential Terms are (A^2, B^2).

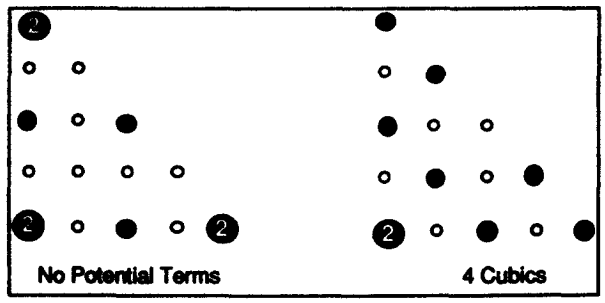

Figure 2. D-Optimal Design for the Quadratic Model (left) and Bayesian Version (right) When Potential Terms Are the $q = 4$ Cubic Terms (A^3, A^2B, AB^2, B^3).

three cornerpoints (actually, replicating any three of the six points is a *D*-optimal solution) so that only six unique points are run and there are no degrees of freedom available for testing goodness of fit. The right side of Figure 2 shows the Bayesian design when the $q = 4$ cubic potential terms are used. In this case only one point is replicated so that the 3 df for residuals from the quadratic are partitioned into 1 for pure error and 2 for lack of fit. This provides a Bayesian rationale for "checkpoints." The design at the right of Figure 2 is chosen by the Bayesian *D*-optimal algorithm whenever $.71 < \tau < 1.06$. The design is concentrated on six points for $\tau < .43$, goes to seven points for $.43 < \tau < .71$, and has no replication for $\tau > 1.06$. As you increase τ from 0, your declared faith in the adequacy of the quadratic model steadily decreases and the number of design points in the optimal design increases accordingly. The proposed default value, $\tau = 1$, provides a reasonable compromise design in this example.

For fitting the quadratic model, the average variance of coefficients is about 10% lower for the design on the left of Figure 2 than for the one on the right. Interestingly, however, the average variance of prediction over the 15 candidate points is 5% lower for the Bayesian design, even assuming that the quadratic model is correct. Of course, the Bayesian design is much better if some cubic terms are active.

3.3 Example 3: Four Quantitative Factors

In this example consider four quantitative factors, denoted *A*, *B*, *C*, and *D*, respectively, and take a 3^4-candidate set $= \{-1, 0, 1\}^4$. Suppose that the sample size $n = 9$ and the $p = 5$ primary terms constitute the first-order model. A *D*-optimal design and the Bayesian designs resulting from three sets of potential terms (always using $\tau = 1$) are shown in Table 1. The leftmost design in Table 1 shows a *D*-optimal design for the first-order model, which has no factor settings at 0 and which does not include the resolution IV eight-run fractional factorial design. There are many different *D*-optimal designs for this problem. The one shown is a resolution III design with one point replicated (runs 4 and 5). The second design in Table 1 shows the Bayesian design when the potential terms are chosen to be the $q = 4$ squares of the factors. The Latin square design sometimes called L9 results. (In this case $n = p + q$ so that the *D*-optimal design for the nine-term model can be computed, resulting in the same design.) The third design in Table 1 shows the Bayesian design when the potential terms are defined as the $q = 6$ two-factor interactions, without the square terms. It consists of a resolution IV half fraction of the 2^4 design, plus one point (run 5 in Table 1) from the other half fraction. Finally, the rightmost design in Table 1 shows the result of including both the square terms and the interactions as potential terms ($q = 10$). It consists of the eight-run resolution IV fractional factorial design plus an overall centerpoint. In all three cases, the Bayesian design would be preferred to the original *D*-optimal design by most statisticians.

Table 1. Designs for Four Quantitative Factors (A, B, C, D) for n = 9 and a First-Order Primary Model for Different Sets of Potential Terms

Designs	A B C D	A B C D	A B C D	A B C D
1	+ + + +	+ + + +	+ + + +	+ + + +
2	+ + − +	+ 0 0 −	+ + − −	+ + − −
3	+ − + −	+ − − 0	+ − + −	+ − + −
4	+ − − −	0 + − −	+ − − +	+ − − +
5	+ − − −	0 0 + 0	[+ + + −]	0 0 0 0
6	− + + −	0 − 0 +	− + + −	− + + −
7	− + − −	− + 0 0	− + − +	− + − +
8	− − + +	− 0 − +	− − + +	− − + +
9	− − − +	− − + −	− − − −	− − − −
	No pot. (2-Level)	Squares (L9)	Interactions (FF + #5)	Both (FF + Ctrpt)

3.4 Example 4: Eight Two-Level Factors

In a screening situation, suppose that there are 8 two-level factors and $n = 16$. The primary terms include the first-order model ($p = 9$). The resolution IV 2^{8-4} fractional factorial design is an excellent design for this problem. This design is also D-optimal for the first-order model, but there are many other orthogonal resolution III designs that have the same determinant of $X'X$ as the resolution IV design. Even worse for a person using the D-optimal approach, a computer search algorithm can easily become trapped in a local optimum and end up with a nonorthogonal design that is not even truly D-optimal when trying to solve this problem. More than one commercially available experimental design program has failed to find a global optimum for this D-optimal problem, and none can be expected to find the resolution IV design consistently.

When the $q = 28$ two-factor interactions are used as potential terms and the default value $\tau = 1$ is used, the Bayesian algorithm consistently converges to the resolution IV 2^{8-4} fractional factorial design. In this example, not only does the Bayesian model provide a theoretical justification for the resolution IV design, but the Bayesian computations provided a successful computational path as well. Of course, for this problem there are far simpler ways to compute the 2^{8-4} design, but the point here is that the Bayesian modification corrected a serious defect of the D-optimal method that may occur in more complicated situations that have no simple classical design solution.

To further explore the behavior of the Bayesian algorithm, we tried two more examples within the two-level fractional factorial domain, each with six factors. In the 16-run 2^{6-2} situation, with $p = 7$ first-order terms and $q = 15$ second-order terms, the algorithm consistently converged to a resolution IV design when $.25 < \tau < .45$ but not when $\tau = 1$. In the 32-run 2^{6-1} situation, with $p = 22$ first- and second-order terms and $q = 20$ third-order terms, the algorithm consistently converged to a resolution VI design when $\tau = 1$. Further research is needed to explore the correspondence between Bayesian solutions and achieving orthogonality between primary and potential terms, which must depend on τ.

4. A PRACTICAL EXAMPLE: CONSTRAINED MIXTURE DESIGN

Snee (1981) described an experiment to determine how the research octane of various blends of gasoline depends on the component proportions of the blend. There are five mixture components:

Component	Range
Butane (B)	0–.15
Isopentane (I)	0–.30
Reformate (R)	0–.35
Cat Cracked (C)	0–.60
Alkylate (A)	0–.60

In addition to the natural mixture requirement, $B + I + R + C + A = 1$, Snee (1981) stated that the range of experimentation was required to satisfy all of the following constraints: $B + I \leq .30$, $C + A \leq .70$, and $97 \leq 101.8\,B + 99.6\,I + 112.4\,R + 94.2\,C + 99.8\,A \leq 101$. The experiment was a 25-run D-optimal design based on a full quadratic model with 15 terms, with the addition of a 26th run, the centroid of the design region. The model fit the data well, and after eliminating six of the second-order terms that had small coefficients, the author found the model (Snee 1981, table 4) shown in Table 2. There is no constant term because the Scheffé parameterization was being used. See, for example, Cornell (1990) for more information on the special considerations that apply to the design and analysis of mixture experiments.

Table 2. Reduced Model for Octane

Term	Coefficient	Standard error
B	155.1	23.6
I	97.7	.7
R	108.6	.8
C	95.0	.4
A	101.4	.4
B*I	−44.6	27.3
B*R	−77.0	27.7
B*C	−67.6	28.7
B*A	−60.0	28.5

NOTE: Residual standard deviation = .30.

4.1 Scenario for This Example

For the purpose of this example, we assume that the coefficients and residual standard deviation given previously are the true values, and we consider what might happen if an experimenter, incorrectly, assumes that a first-order model is appropriate and plans a smaller 12-run experiment. Indeed, Snee (1981, p. 125) remarked, "The statistical significance of these [interaction] coefficients was unexpected." To fit a first-order model ($p = 5$), a sample size of $n = 12$ is reasonable. For example, if the full simplex were available to work within, running two replicates at the centroid and at each of the five vertices would be a standard option. In this example, the many constraints on the factors make an algorithmic design necessary. First, four different design strategies are presented, after which they are compared with respect to several different criteria.

4.2 First-Order Strategy

The *first-order strategy* is based on a *D*-optimal search over the constrained design region using the first-order model ($p = 5$) with $n = 12$. The candidate set for the search algorithm consists of the 28 extreme vertices of the design region (listed by Snee 1981). Table 3A shows the design based on this strategy. The runs have been sorted by columns from left to right to make them easier to examine. Note that only two values of butane are represented, and there are just eight distinct runs because four runs are repeated once each.

4.3 First-Order-With-Centerpoint Strategy

The *first-order-with-centerpoint strategy* is similar to the first-order strategy except that 2 of the 12 design points are first specified as the centroid of the design region, here defined as the average of the extreme vertices. In particular, this strategy first includes two replicates at the centroid and then chooses 10 more runs from the candidate set of extreme vertices so as to maximize the determinant of $X'X$ based on all 12 runs and the first-order model. The resulting design is given in Table 3B. This design has nine distinct runs, since besides the centroid listed in rows 7–8 of Table 3B, there are two other pairs of replicates (rows 1–2 and 11–12).

4.4 Bayesian Strategy

Third, the *Bayesian strategy* takes the first-order model ($p = 5$) for the primary terms and all the second-order effects ($q = 10$) as the potential terms. [To satisfy the scaling conventions mentioned previously—in particular to have the primary terms range from -1 to 1—it is necessary to switch to a slack variable model, where $X_{\text{pri}} = (1, B/.075, -1, I/.15 - 1, R/.175 - 1, C/.3 - 1)$ and X_{pot} is the set of squares and cross-products of all but the constant term of X_{pri}. X_{pot} is then centered and scaled to produce Z, as described in Sec. 2. After the design is computed, it is reexpressed in the original units.] The candidate set for the search algorithm consists of 770 points—the extreme vertices, the midpoints of edges and of planes for constraint boundaries, the centroid, and a lattice grid in which each proportion is a multiple of .05 within the constrained region. The candidate set was computed using the RS/Discover software (BBN Software Products 1992), which uses the algorithm described by Piepel (1988). The resulting design (assuming $\tau = 1$), shown in Table 3C, has no repeated runs.

4.5 Omniscient Strategy

The fourth and final strategy we call the *omniscient strategy*, which is based on the (unlikely) premise that the experimenter knows exactly which second-order terms are active and then computes the 12-run *D*-optimal design that is best for the true model. In this case, there are $p = 9$ terms in the "true" model,

Table 3. Computer-Generated 12-Run Designs For Gasoline Blending

Run	But	Iso	Ref	Cat	Alk
A. D-optimal design based on first-order strategy					
1	.000	.000	.350	.600	.050
2	.000	.000	.350	.600	.050
3	.000	.300	.000	.100	.600
4	.000	.300	.049	.600	.051
5	.000	.300	.100	.000	.600
6	.000	.300	.285	.415	.000
7	.150	.034	.116	.100	.600
8	.150	.034	.116	.100	.600
9	.150	.127	.023	.600	.100
10	.150	.127	.023	.600	.100
11	.150	.150	.266	.434	.000
12	.150	.150	.266	.434	.000
B. D-optimal design based on first-order-with-centerpoint strategy					
1	.000	.000	.350	.600	.050
2	.000	.000	.350	.600	.050
3	.000	.300	.000	.100	.600
4	.000	.300	.049	.600	.051
5	.000	.300	.100	.000	.600
6	.000	.300	.285	.415	.000
7	.068	.121	.175	.444	.192
8	.068	.121	.175	.444	.192
9	.150	.034	.116	.100	.600
10	.150	.127	.023	.600	.100
11	.150	.150	.266	.434	.000
12	.150	.150	.266	.434	.000
C. Modified D-optimal design based on Bayesian strategy					
1	.000	.000	.350	.600	.050
2	.000	.100	.200	.250	.450
3	.000	.300	.049	.600	.051
4	.000	.300	.100	.000	.600
5	.000	.300	.285	.415	.000
6	.050	.250	.000	.100	.600
7	.075	.225	.276	.424	.000
8	.100	.000	.200	.600	.100
9	.150	.000	.150	.177	.523
10	.150	.000	.311	.539	.000
11	.150	.150	.000	.548	.152
12	.150	.150	.082	.018	.600
D. D-optimal design based on omniscient strategy					
1	.000	.000	.300	.461	.239
2	.000	.300	.049	.600	.051
3	.000	.300	.100	.000	.600
4	.000	.300	.285	.415	.000
5	.075	.000	.225	.600	.100
6	.075	.225	.000	.100	.600
7	.075	.225	.276	.424	.000
8	.150	.000	.150	.177	.523
9	.150	.000	.311	.539	.000
10	.150	.150	.023	.600	.077
11	.150	.150	.082	.018	.600
12	.150	.150	.266	.434	.000

given at the beginning of this section. The candidate set consists of the same 770 runs as for the Bayesian strategy. The resulting design, shown in Table 3D, also has no repeated runs.

4.6 Prediction Errors Using the First-Order Model

These designs are first compared based on how well the true, nonlinear response surface can be estimated by a first-order model based on the data from each design. Suppose that $\mu(x)$ is the true mean of the response (octane level) when the component proportions are the row vector x and that $\mu(X)$ is the column vector of true means for the runs of an experiment with design matrix X, where μ is computed using the supposed true model found by Snee (1981). Then the bias and prediction variance of the first-order model at x are bias $= \mu(x) - x(X'X)^{-1}X'\mu(X)$ and variance $= x(X'X)^{-1}x'\sigma^2$. The first two columns of Table 4 show the average squared bias and the average variance for the 770 values of x that make up the candidate set, described previously, that was used to generate the Bayes and omniscient designs, assuming that the coefficients of the model and the residual standard deviation are as given at the beginning of this section. As expected, the design based on the first-order strategy has the greatest bias and the lowest variance of prediction. The average squared bias of the first-order design is so large, however, that it completely overwhelms the comparatively minor reduction in average variance that it achieves. The design based on the Bayesian strategy has the lowest average squared bias and the lowest mean squared prediction error (sum of bias^2 and variance) of all four designs. The square roots of the mean squared error of predictions for the four designs are .34, .31, .27, and .27, respectively. The Bayes and omniscient strategies produce roughly equivalent designs by this criteria, and the design that includes two centerpoints is about midway between them and the first-order strategy design.

4.7 Detecting Lack of Fit

Next we compare how well data from each of the designs can be expected to provide evidence that the first-order model is incorrect. Although there are many possible data-analytic approaches to detecting model inadequacies, only one method will be examined here—the computation of an F statistic for lack of fit, based on fitting the true model including the four interaction terms and comparing it with the fit to the first-order model. When the true model is not first-order, the F statistic has a noncentral F distribution (Bickel and Doksum 1977), with noncentrality parameter equal to the sum of the squared biases for runs in the design, divided by σ^2:

$$\begin{aligned}\text{noncentrality} &= [\mu(X) - X(X'X)^{-1}X'\mu(X)]'[\mu(X) \\ &\quad - X(X'X)^{-1}X'\mu(X)]/\sigma^2 \\ &= \mu(X)'[I - X(X'X)^{-1}X']\mu(X)/\sigma^2.\end{aligned}$$

The attained significance or P value of the F test is a random quantity, but an idea of how significant the lack of fit will typically appear can be obtained by assuming that the numerator and denominator of the F statistic comparing the two models are equal to their expected values. In that case, the value of the F statistic will be $F = 1 + \text{noncentrality}/\text{df}_{\text{numerator}}$. The degrees of freedom for the F test vary, depending on which of the four designs are being used. For the designs from the Bayes and omniscient strategies, the nine-term true model can be fit, so, since $n = 12$ and the first-order model has $p = 5$, the F test has 4 df for lack of fit (numerator) and 3 df for error (denominator). For the purpose of fitting the nine-term interactions model, the first-order design and the first-order-with-centerpoints design are singular;

Table 4. Comparison of the Designs Based on Their Ability to Predict Using a First-Order Model, to Detect Lack of Fit to That Model, and to Estimate the Quadratic Model With a Bayesian Shrinkage Estimator

	First-order model (least sq.)		F test for lack of fit		Full quad. (shrinkage est.)	
Design	Ave. bias^2	Ave. variance	Noncentrality	Significance*	Ave. bias^2	Ave. variance
First-order	.0944	.0218	1.04	.378	.0935	.0248
Centerpoint	.0690	.0255	3.78	.224	.0510	.0339
Bayes	.0461	.0248	12.03	.142	.0063	.0372
Omniscient	.0471	.0255	9.06	.179	.0154	.0369

*P value if $F = 1 +$ noncentrality/$\text{df}_{\text{numerator}}$. The F statistic has $\text{df}_{\text{numerator}} = 3$, $\text{df}_{\text{denominator}} = 4$, for the first-order and centerpoint designs; $\text{df}_{\text{numerator}} = 4$, $\text{df}_{\text{denominator}} = 3$ for the Bayes and omniscient designs.

the 12 × 9 model matrix is of rank 8. This is obvious by examining the first-order design in Table 3A because there are only eight unique runs. But it is also true for the design in Table 3B: The value of B for each run can be expressed as a linear combination of the terms B^*I, B^*R, B^*C, and B^*A. Because of this, the F statistic has just 3 df for lack of fit (numerator) and 4 df for error (denominator). The computation of the F statistic in these cases would be made by comparing the first-order model to the interactions model in which one of the four interaction terms has been left out.

Columns 3 and 4 of Table 4 show the computed values of the noncentrality parameter and the tail area of F for each of the four designs being considered. Since all four of the significance values in column 4 are greater than .1, we see that none of the designs are very powerful for detecting the presence of the interaction terms. The sample size is just too small (or alternatively there is just too much variability). The designs based on the first-order model, however, especially the one without centerpoints, are clearly much less powerful than the other two. If the first-order model were to be rejected whenever $P < .2$, for example, then the first two designs would typically lead to acceptance, but the other two designs would typically not. If the residual standard deviation were .2 instead of .3, then the noncentralities would each be multiplied by $(.3/.2)^2 = 2.25$, and the four P values in column 4 would be .29, .11, .06, and .08, respectively. It is interesting that even though the omniscient design is D-optimal for the true model it is not quite as good as the Bayes design at distinguishing the true model from the first-order model using the F test.

4.8 Prediction Errors Using the Full Quadratic Model

Since the full quadratic model has $p + q = 15$ terms, a design with $n = 12$ runs cannot obtain unique least squares estimates of all the coefficients. The prior distribution used to obtain the Bayes design, however, can be used to generate a posterior distribution for the coefficients using the data from any of the designs, and predictions can be based on the expectation of the posterior distribution. Let XZ denote the 12 × 15 model matrix of the full quadratic slack variable model, scaled as defined previously. Let xz denote a row vector of the 15 terms in this model for an arbitrary point x in the design region. Let K be a diagonal matrix with 0 in the first five diagonal positions, and 1 in the last 10 positions. Then the prediction and its bias and variance at x [conditional on the true values of $\mu(x)$] are

$$\text{prediction} = xz(XZ'XZ + K)^{-1}XZ'Y,$$

$$\text{bias} = \mu(x) - xz(XZ'XZ + K)^{-1}XZ'\mu(X),$$

and

$$\text{variance} = xz(XZ'XZ + K)^{-1}XZ'XZ(XZ'XZ + K)^{-1}xz'\sigma^2.$$

The last two columns of Table 4 show the average, over the candidate set of 770 points, of the squared bias and the variance of prediction when this Bayesian shrinkage estimator is used to make predictions for each of the four designs. Here the Bayesian design performs best because it is theoretically tailored for this situation. Note that the first-order design does no better with the quadratic shrinkage estimate than with the first-order least squares estimate. The centerpoint design is also not able to take much advantage of the shrinkage estimator's ability to fit the interaction terms in the true model. The omniscient design does much better than the first two designs but not as well as the Bayes design. The square roots of the average mean squared errors of predictions for the four designs are .34, .29, .21, and .23, respectively. The omniscient design could attain a bias of 0 and a smaller variance if the true model is fitted to the data by least squares, but that would involve knowing the correct model in advance.

4.9 Summary of the Practical Example

This example takes a real-life situation in which interactions were present but not anticipated by the researchers and compares the proposed Bayesian strategy to other alternative strategies, assuming that the researchers had not designed for the full quadratic model, as did Snee (1981). A D-optimal design based solely on the first-order model gives very poor predictions, averaged over the entire design region, even if a sophisticated shrinkage estimate is used, and does not give any indication of lack of fit. Including two centerpoints in the D-optimal design based on a first-order model improves the bias and power to detect lack of fit somewhat. The Bayesian design has greatly improved bias, mean squared prediction error, and lack-of-fit detection properties, doing at least as well in these respects, for this example, as does the D-optimal design based on the true model with interactions.

These criteria for comparing designs are all based on frequentist statistical theory. We can expect that the Bayesian design would fare comparatively even better if measured by Bayesian criteria that use the prior distribution to average over the parameter space.

5. CONCLUSION

This article has argued that there are many situations in which an algorithmic approach to the design problem makes sense and that the flexibility afforded by the approach will be even more in demand as more experimenters turn to computer programs rather than statistical consultants for design assistance. See Nachtsheim (1987) and Koch, Morris, Nachtsheim, and Welch (1993) for discussions of commercial software for design of experiments. A computer user wants a solution to the problem as posed, and a computer program is less likely than an expert statistical consultant to be able to suggest the best alteration of the posed problem that fits a classical design paradigm.

This situation creates an increasing need to have D-optimal and similar designs be less dependent on the assumed model and more able to produce designs that allow for model checking. The Bayesian modification of the D-optimal algorithm proposed here is one such approach. It has several desirable properties. First, it is reasonably automatic—there is a need to specify a set of terms that are potentially active, but usually this will be a set of next-higher-order terms in a polynomial model. The default value of $\tau = 1$ for the standard deviation of the prior distribution seems to work well, thanks to the scaling conventions discussed in Section 2. Second, the procedure is easy to carry out because the computations involved are no more difficult (in Ex. 4 of Sec. 3 they seemed to be more numerically stable) than the standard D-optimal algorithm. This is important, since D-optimal algorithms are widely available and used, making it easy to implement the proposed algorithm as a minor modification of current software. When n and the size of the candidate set are fixed, the number of computations required per iteration of a D-optimal design algorithm are approximately proportional to the number of terms in the model (Galil and Kiefer 1980). Thus the Bayesian calculations will use about $(1 + q/p)$ times the computer time of the unmodified D-optimal search because in our experience the number of iterations until convergence of the Bayesian algorithm averages no greater than for the D-optimal one. In Example 4 of Section 3, where there are $p = 9$ primary terms and $q = 28$ potential terms, only a few seconds of extra computer time are required.

Compared with other proposed algorithmic approaches to creating designs that are model robust, the proposed method is simple, powerful, and flexible. For example, the linear-optimal designs of Cook and Nachtsheim (1982) require that the design not be singular with respect to all models under consideration, but the proposed Bayesian method, inspired by resolution IV fractional factorial designs, is designed for situations in which $n < p + q$. As shown in Example 3 of Section 3, the method is flexible in that different choices of sets of potential terms can lead to different designs tailored to the specific needs of a problem. The minimum mean squared error designs of Welch are derived to be powerful in detecting a certain type of model inadequacy: $|\mu(x) - x\beta| = \pm\delta$ for every x in the candidate set. The choice of δ in the Welch procedure corresponds to the choice of τ in the Bayesian procedure. In situations in which potential lack of fit is expected to be a smooth function of x, the Bayesian design may be preferred. Galil and Kiefer (1977) showed that the minimum-bias designs of Box and Draper (1959) are sometimes not as powerful as D-optimal designs, even when measured by the criterion of mean squared error. Box and Draper (1959) suggested expanding the range of their minimum-bias designs to improve their variance properties. In a similar vein, Steinberg (1985) and Draper and Guttman (1992) advocated scaling designs so that there is some compromise between concerns for bias and for variance. The latter two articles used models similar to our Bayesian model but for the purpose of fixing the scale of the design inside a larger operability region. In general, however, rescaling a D-optimal design will neither reduce the correlations between primary and potential terms nor improve the ability to detect and estimate the effects of potential terms.

The Bayesian designs have the advantage of being specifically adapted to a powerful and easily implemented analysis method because the Bayesian posterior mean of β is a shrinkage-regression estimator with good frequentist statistical properties, as exemplified by the example of Section 4. Smith and Campbell (1980) criticized the use of ridge-regression estimators in general but advocated their use in cases like the present in which the scaling and centering of the terms and the choice of K are based on a Bayesian model and prior knowledge of the process being studied.

Finally, the Bayesian model itself is not unreasonable for many situations in which the primary model is plausible but not known exactly, so the resulting Bayesian design and analysis solutions can be expected to have good properties, as shown in the theoretical examples of Section 3 and in the practical example of Section 4. The proposed method has proven to be a valuable tool for finding computer-generated experimental designs when the experimenter wants to "hedge bets" about an assumed model. We encourage further research to find more examples and theoretical properties of these designs.

ACKNOWLEDGMENTS

We thank James Lucas for comments on an early version of this article and the associate editor for helpful advice on presentation.

[*Received May 1992. Revised March 1993.*]

REFERENCES

BBN Software Products (1992), *RS/Discover Reference Manual*, Cambridge, MA: Bolt Beranek Newman.

Bickel, P. J., and Doksum, K. A. (1977), *Mathematical Statistics: Basic Ideas and Selected Topics*, San Francisco: Holden-Day.

Box, G. E. P., and Draper, N. R. (1959), "A Basis for the Selection of a Response Surface Design," *Journal of the American Statistical Association*, 54, 622–654.

——— (1987), *Empirical Model-Building and Response Surfaces*, New York: John Wiley.

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

Box, G. E. P., and Tiao, G. C. (1973), *Bayesian Inference in Statistical Analysis*, Reading, MA: Addison-Wesley.

Chaloner, K. (1984), "Optimal Bayesian Experimental Design for Linear Models," *The Annals of Statistics*, 12, 283–300; Corrigendum (1985), 13, 836.

Cook, R. D., and Nachtsheim, C. J. (1980), "Comparison of Algorithms for Constructing Exact *D*-Optimal Designs," *Technometrics*, 22, 315–324.

——— (1982), "Model Robust, Linear-Optimal Designs," *Technometrics*, 24, 49–54.

——— (1989), "Computer-Aided Blocking of Factorial and Response-Surface Designs," *Technometrics*, 31, 339–346.

Cornell, J. A. (1990), *Experiments With Mixtures—Designs, Models, and the Analysis of Mixture Data* (2nd ed.), New York: John Wiley.

Draper, N. R., and Guttman, I. (1992), "Treating Bias as Variance for Experimental Design Purposes," *Annals of Institute of Statistical Mathematics*, 44, 659–671.

Efron, B., and Morris, C. (1973), "Stein's Estimation Rule and its Competitors—an Empirical Bayes Approach," *Journal of the American Statistical Association*, 68, 117–130.

Galil, Z., and Kiefer, J. (1977), "Comparison of Box–Draper and *D*-Optimum Designs for Experiments With Mixtures," *Technometrics*, 19, 441–444.

——— (1980), "Time and Space-Saving Computer Methods, Related to Mitchell's DETMAX, for Finding *D*-Optimum Designs," *Technometrics*, 21, 301–313.

Hahn, G., Meeker, W., and Feder, P. (1976), "The Evaluation and Comparison of Experimental Designs for Fitting Regression Relationships," *Journal of Quality Technology*, 8, 140–157.

Johnson, M. E., and Nachtsheim, C. J. (1983), "Some Guidelines for Constructing Exact *D*-Optimal Designs on Convex Design Spaces," *Technometrics*, 25, 271–277.

Koch, D., Morris, W. T., Nachtsheim, C. J., and Welch, W. J. (1993), "Computer Software for Robust Product Design," Working Paper 93-11, University of Minnesota, Dept. of Operations and Management Science.

Lee, P. M. (1989), *Bayesian Statistics: An Introduction*, New York: Oxford University Press.

Lindley, D. V. (1956), "On a Measure of the Information Provided by an Experiment," *The Annals of Mathematical Statistics*, 27, 986–1005.

Meyer, R., and Nachtsheim, C. J. (1991), "The Coordinate Exchange Algorithm for Constructing Exact Optimal Experimental Designs," Working Paper 91-18, University of Minnesota, Dept. of Operations and Management Science.

Mitchell, T. J. (1974), "An Algorithm for the Construction of *D*-Optimal Experimental Designs," *Technometrics*, 16, 203–211.

Montgomery, D. (1991), *Design and Analysis of Experiments*, New York: John Wiley.

Nachtsheim, C. J. (1987), "Tools for Computer-Aided Design of Experiments," *Journal of Quality Technology*, 19, 132–160.

O'Hagan, A. (1978), "Curve Fitting and Optimal Design for Prediction" (with discussion), *Journal of the Royal Statistical Society*, Ser. B, 40, 1–41.

Oman, S. D. (1984), "A Different Empirical Bayes Interpretation of Ridge and Stein Estimators," *Journal of the Royal Statistical Society*, Ser. B, 46, 544–557.

Piepel, G. F. (1988), "Programs for Generating Extreme Vertices and Centroids of Linearly Constrained Experimental Regions," *Journal of Quality Technology*, 20, 125–139.

Pilz, J. (1991), *Bayesian Estimation and Experimental Design in Linear Regression Models*, New York: John Wiley.

St. John, R. C., and Draper, N. R. (1975), "*D*-Optimality for Regression Designs: A Review," *Technometrics*, 17, 15–23.

Silvey, S. D. (1980), *Optimal Design*, London and New York: Chapman and Hall.

Smith, A. F. M., and Verdinelli, I. (1980), "A Note on Bayes Designs for Inference Using a Hierarchical Linear Model," *Biometrika*, 67, 613–619.

Smith, G., and Campbell, F. (1980), "A Critique of Some Ridge Regression Methods" (with discussion), *Journal of the American Statistical Association*, 75, 74–103.

Snee, R. D. (1981), "Developing Blending Models for Gasoline and Other Mixtures," *Technometrics*, 23, 119–130.

——— (1985), "Computer Aided Design of Experiments—Practical Experiences," *Journal of Quality Technology*, 17, 222–236.

Steinberg, D. (1985), "Model Robust Response Surface Designs: Scaling Two-Level Factorials," *Biometrika*, 72, 513–526.

Steinberg, D., and Hunter, W. (1984), "Experimental Design: Review and Comment," *Technometrics*, 26, 71–130.

Vuchkov, I. N. (1977), "A Ridge-Type Procedure for Design of Experiments," *Biometrika*, 64, 147–150.

Welch, W. J. (1982), "Branch-and-Bound Search for Experimental Designs Based on *D*-Optimality and Other Criteria," *Technometrics*, 24, 41–48.

——— (1983), "A Mean Squared Error Criterion for the Design of Experiments," *Biometrika*, 70, 205–213.

Wu, C. F., and Wynn, H. P. (1978), "The Convergence of General Step-length Algorithms for Regular Optimum Design Criteria," *The Annals of Statistics*, 6, 1276–1285.