REFERENCES

Fast Computation of Designs Robust to Parameter Uncertainty for Nonlinear Settings

Christopher M. GOTWALT and Bradley A. JONES

SAS Institute
Cary, NC 27513-2414
(*chris.gotwalt@jmp.com*; *bradley.jones@jmp.com*)

David M. STEINBERG

Department of Statistics and Operations Research
Tel-Aviv University
Ramat Aviv 69978, Israel
(*dms@post.tau.ac.il*)

Experimental design in nonlinear settings is complicated by the fact that the efficiency of a design depends on the unknown parameter values. Thus good designs need to be efficient over a range of likely parameter values. Bayesian design criteria provide a natural framework for achieving such robustness, by averaging local design criteria over a prior distribution on the parameters. A major drawback to the use of such criteria is the heavy computational burden that they impose. We present a clever quadrature scheme that greatly improves the feasibility of using Bayesian design criteria. We illustrate the method on some designed experiments.

KEY WORDS: Bayesian design; D-optimal design; Generalized linear model; Mysovskikh quadrature; Nonlinear model.

1. INTRODUCTION

The efficiency of an experimental design in a nonlinear setting depends on the unknown values of the parameters (Box and Lucas 1959). Yet it is these very values that the experiment seeks to determine. Were the parameters known, there would be no need to experiment, but someone must specify parameter values to find a good design. Bayesian methods provide a principled way to address this "chicken and egg" problem by introducing uncertainty about the parameters into the problem formulation. The resulting designs are reasonably efficient over the entire region of likely parameter vectors and thus are robust to parameter uncertainty (Chaloner and Verdinelli 1995).

The use of Bayesian design criteria has been limited by the burden of computing suitable criteria, which requires integrations over the parameter space. This is especially true for factorial designs, as the computational complexity of most quadrature schemes grows exponentially in the number of parameters. Monte Carlo integration also may require very large samples.

In this paper we present a fast and accurate quadrature scheme for computing Bayesian design criteria with respect to a prior distribution. For normal priors, the scheme exploits a reparameterization and decomposition of the integral into a radial integral with respect to a χ^2_p distribution and an integral over the surface of the unit hypersphere in $\mathbb{R}^p$. The radial integral is approximated using generalized Gauss–Laguerre quadrature with an abscissa at 0 added, whereas the integral over the hypersphere uses a method due to Mysovskikh (1980) that integrates all multivariate quintics exactly, while using only $O(p^2)$ evaluations of the integrand.

Our quadrature method can be applied to experimental design for nonlinear models and for generalized linear models (GLMs). The nonlinearity enters each of these problems differently, but both problems have the same fundamental difficulty in evaluating Bayesian design criteria.

In Section 2 we present a complete description of the design problem for nonlinear models and for GLMs, including some motivating applications. We describe the quadrature method in Section 3. We also discuss some interesting relationships of the sets of quadrature points generated by the method to experimental designs for response surfaces. In Section 4 we briefly describe how we incorporate the quadrature scheme in an algorithm for finding robust optimal designs. In Section 5 we present applications of the method to several concrete problems. We close with a discussion and some concluding remarks in Section 6.

2. BAYESIAN DESIGN FOR NONLINEAR SETTINGS

2.1 Generalized Linear Models

Specification of a GLM for a response Y involves the following features:

- The response Y has an exponential family distribution G;
- $\mathrm{E}\{Y(\mathbf{x})\} = \mu(\mathbf{x})$;
- $h[\mu(\mathbf{x})] = \mathbf{f}^T(\mathbf{x})\boldsymbol{\theta}$, where h is the link function for the model and $\mathbf{f}^T(\mathbf{x})\boldsymbol{\theta}$ is the linear predictor.

The information matrix for a design d with observations at $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is

$$I(\boldsymbol{\theta}; d) = \sum_{i=1}^{n} \mathbf{f}(\mathbf{x}_i)\mathbf{f}^T(\mathbf{x}_i)w_i, \tag{1}$$

where $w_i = 1/\{\mathrm{Var}(\mu_i)[h'(\mu_i)]^2\}$ is a weight associated with the ith observation. Thus the information depends on the parameter vector through the weights. The first term in the weight is the variance of the ith observation, as a function of its expectation, and the second term is related to the link function.

As an example, consider the crystallography experiment discussed by Woods et al. (2006). The goal of the experiment was to assess the effects of four factors on the formation of a new product. The outcome was binary: success or failure of the process. The experiment was designed to fit a first-order logistic regression model. For this experiment,

TECHNOMETRICS, FEBRUARY 2009, VOL. 51, NO. 1
DOI 10.1198/TECH.2009.0009

- $Y(\mathbf{x}) \sim \mathrm{B}(1, \mu(\mathbf{x}))$;
- $\mathrm{E}\{Y(\mathbf{x})\} = \mu(\mathbf{x})$;
- $h[\mu(\mathbf{x})] = \log[\mu(\mathbf{x})/(1 - \mu(\mathbf{x}))] = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \theta_4 x_4$.

The variance function is $\mathrm{Var}(\mu) = \mu(1-\mu)$, $h'(\mu) = 1/[\mu(1-\mu)]$, and the weights are $w_i = \mu(\mathbf{x}_i)[1 - \mu(\mathbf{x}_i)]$. This calculation shows that little information is gained by taking observations at sites $\mathbf{x}$ where response (nonresponse) is almost certain. (See Khuri et al. 2006 for a review of experimental design results for GLMs and Woods et al. 2006 and Dror and Steinberg 2006 for two recent approaches to the design of factorial experiments.)

2.2 Nonlinear Models

Nonlinear models link a response variable Y to a set of explanatory variables $\mathbf{x}$ by a model of the form

$$Y(\mathbf{x}) = \eta(\mathbf{x}; \boldsymbol{\theta}) + \varepsilon, \tag{2}$$

where $\eta(\mathbf{x}; \boldsymbol{\theta})$ is a nonlinear function of the parameter vector $\boldsymbol{\theta}$. We assume here that $\varepsilon \sim \mathrm{N}(0, \sigma^2)$, although our methods also apply in problems in which there is known or modeled dependence of the error variance on the explanatory variables. We also assume statistically independent errors.

Suppose that an n-run experiment d takes observations at the points $\mathbf{x}_1, \ldots, \mathbf{x}_n$. The resulting Fisher information matrix for $\boldsymbol{\theta}$ is

$$I(\boldsymbol{\theta}; d) = (1/\sigma^2) \sum_{i=1}^{n} \mathbf{g}(\mathbf{x}_i; \boldsymbol{\theta}) \mathbf{g}^T(\mathbf{x}_i; \boldsymbol{\theta}), \tag{3}$$

where $\mathbf{g}(\mathbf{x}_i; \boldsymbol{\theta}) = \nabla \eta(\mathbf{x}_i; \boldsymbol{\theta})$ is the gradient of the response function with respect to the elements of $\boldsymbol{\theta}$. The dependence of the information on $\boldsymbol{\theta}$ stems from the fact that the partial derivatives are functions of $\boldsymbol{\theta}$ in a nonlinear model.

Nonlinear statistical models often are needed to describe complex relationships in engineering and the natural sciences. Compartmental models, widely used in biology to describe the flow of materials in an organism, are a good example. Button (1979) used a compartmental model to study the concentration of a drug in the bloodstream of a horse. The drug is absorbed into the bloodstream and then gradually decays. The resulting model for the concentration $Y(t)$ in the bloodstream at time t is

$$Y(t) = \theta_3[\exp(-\theta_1 t) - \exp(-\theta_2 t)] + \varepsilon. \tag{4}$$

The parameter vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^T$ must be estimated from experimental data. Here θ_1 and θ_2 are rate constants describing drug transport between the compartments, and θ_3 is the initial concentration before it is absorbed into the blood. The experimental design problem is to select a set of times at which to draw blood samples and measure the concentration.

The first systematic treatment of experimental design for nonlinear models was that of Box and Lucas (1959), who applied the D-optimality criterion to find designs that lead to precise estimates of the model parameters. Useful reviews of optimal design for nonlinear models have been provided by Ford, Kitsos, and Titterington (1989) and Atkinson, Donev, and Tobias (2007, chap. 17).

2.3 Local D–Optimality

Here we focus on the D-optimality criterion for experimental designs. For nonlinear models, Chernoff (1953) proposed the idea of local D-optimality: find the design d^* that maximizes $|I(\boldsymbol{\theta}_g; d)|$ over all possible designs, where $|A|$ denotes the determinant of the matrix A and $\boldsymbol{\theta}_g$ is a particular choice for $\boldsymbol{\theta}$. The typical design strategy would be to guess the value of $\boldsymbol{\theta}$ by $\boldsymbol{\theta}_g$ and then find the design that is most efficient when $\boldsymbol{\theta} = \boldsymbol{\theta}_g$.

Local D-optimal designs may not be good choices in practice, because they often are very inefficient if the initial guess is poor. These designs are useful for evaluating a proposed design d, because we can compare how well it does, for a given parameter vector $\boldsymbol{\theta}$, compared with the best design possible. We adopt the standard measure of design efficiency

$$Eff(d; \boldsymbol{\theta}) = \left(|I(\boldsymbol{\theta}; d)|/|I(\boldsymbol{\theta}; d^*)|\right)^{1/p}, \tag{5}$$

where d^* is the local D-optimal design for $\boldsymbol{\theta}$ and p is the number of parameters in the model. We compute local D-optimal designs with a coordinate exchange algorithm, taking the best design found in 1000 random starts.

2.4 Bayesian D–Optimality

A natural way to achieve designs that are robust with respect to uncertainty in $\boldsymbol{\theta}$ is to assign it a prior distribution, $\pi(\boldsymbol{\theta})$, rather than a single value. The local D-optimality functional can be averaged in various ways with respect to the prior to produce an overall design criterion. The criterion that we adopt here is the expectation of the logarithm of the determinant of the information matrix,

$$\phi(d) = \int \log\left(|I(\boldsymbol{\theta}; d)|\right) d\pi(\boldsymbol{\theta}). \tag{6}$$

This criterion was developed by Chaloner and Larntz (1989) in the context of single-factor logistic regression and was applied by Woods et al. (2006) to multifactor experiments with GLM responses. Chaloner and Verdinelli (1995) reported that $\phi(d)$ can be justified as an approximation to maximizing utility when the latter is measured by the Shannon information of the posterior distribution of $\boldsymbol{\theta}$.

The primary difficulty encountered when using $\phi(d)$ as a design criterion is the need to accurately compute the expectation. Typically, this requires the evaluation of a multidimensional integral. Furthermore, algorithms for finding optimal designs need to evaluate the integral many times. Thus such methods as Monte Carlo integration are painfully slow for most problems. For example, Woods et al. (2006, pp. 287–288) used a crude numerical approximation, averaging over just nine points to evaluate a five-dimensional integral. This approach does speed computation, but at the cost of some loss of design efficiency.

3. INTEGRATION OF THE OPTIMALITY CRITERIA

In this section we describe a numerical integration technique for use with normal prior distributions that requires only $O(p^2)$ evaluations, where p is the number of parameters being integrated. We also describe a transformation that can be applied to integrate optimality criteria with respect to nonnormal priors.

3.1 Normal Priors

Our base integration method for normal priors is similar in spirit to the randomized radial-spherical integration rule of Monahan and Genz (1997). As in the Monahan and Genz approach, here the integral is reparameterized into a radial component and $p-1$ spherical surface components, where the surface dimensions are integrated using a randomly rotated Mysovskikh (1980) extended simplex quadrature for integrating over the surface of a sphere with respect to a uniform prior. Our approach differs from that of Monahan and Genz (1997) in that the radial integral is performed using generalized Gauss–Laguerre quadrature, with a point added at the prior mean.

Suppose that the prior distribution on the parameter, $\boldsymbol{\theta}$, is a p-dimensional Normal($\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$), where $\boldsymbol{\Sigma}$ is a $p \times p$ positive definite covariance matrix. Then the design optimality criterion in Equation (6) can be written as

$$\phi(d) = \int_{\mathbb{R}^p} \log|\mathbf{I}(\boldsymbol{\theta}, d)|(2\pi)^{-p/2} \times |\boldsymbol{\Sigma}|^{-1/2} e^{-1/2(\boldsymbol{\theta}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{\theta}-\boldsymbol{\mu})}\, d\boldsymbol{\theta} \quad (7)$$

by performing two changes of variable, first to a standard normal distribution, and then reparameterizing that into a radius-squared component, τ, and a $(p-1)$-dimensional component, $\mathbf{v}$, that is integrated over the surface of the unit sphere in $\mathbb{R}^p$, S_{p-1}.

We now use two changes of variable to reexpress the foregoing formula as a radial-spherical integral. The first change of variable is to a p-dimensional standard normal distribution. The second change of variable then gives a radius-squared component, τ, and a $(p-1)$-dimensional component, $\mathbf{v}$, that is integrated over the surface of the unit sphere in $\mathbb{R}^p$, S_{p-1}. Letting $\mathbf{L}$ represent the lower triangular Cholesky root of $\boldsymbol{\Sigma}$, $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{L}^T$, we get

$$\phi(d) = \int_0^\infty \int_{S_{p-1}} \log|\mathbf{I}(\boldsymbol{\mu} + \mathbf{L}\sqrt{\tau}\mathbf{v}, d)| \times \frac{\tau^{p/2-1}}{\Gamma(p/2)2^{p/2}} e^{-\tau/2}\, d\tau\, d\mathbf{v}. \quad (8)$$

The outer radial integral can be numerically integrated with a modification of generalized Gauss–Laguerre quadrature with abscissas and weights, $\{\tau_i, w_{Ri}\}_{i=0}^{n_R}$, with $r_0 = 0$, and where n_R is the number of radial abscissas. The inner spherical integral can be approximated using the Mysovskikh extended simplex rule with abscissas and weights $\{\mathbf{v}_j, w_{Sk}\}_{k=0}^{n_S}$, where n_S is the number of points in the extended simplex rule. Thus one could use the approximation

$$\phi(d) \approx w_{R_0} \log|\mathbf{I}(\boldsymbol{\mu}, d)| + \sum_{i=1}^{n_R}\sum_{k=1}^{n_S} w_{R_i} w_{S_k} \log|\mathbf{I}(\boldsymbol{\mu} + \mathbf{L}\sqrt{\tau_i}\mathbf{v}_k, d)|. \quad (9)$$

We go one step further in practice, however. Monahan and Genz (1997) recommended randomly rotating the extended simplexes as a bias reduction measure. Following their lead, we also allow for more than one randomly rotated extended simplex per sphere, and average over the extended simplexes. For each nonzero radius, we create n_Q random orthogonal matrices, $\{\mathbf{Q}_{ij}\}_{j=1}^{n_Q}$, and average the value of the integrand over these orthogonal rotations. Thus the integral approximation that we propose can be written as

$$\phi(d) \approx w_{R_0} \log|\mathbf{I}(\boldsymbol{\mu}, d)| + \sum_{i=1}^{n_R}\sum_{j=1}^{n_Q}\sum_{k=1}^{n_S} \frac{w_{R_i} w_{S_k}}{n_Q} \times \log|\mathbf{I}(\boldsymbol{\mu} + \mathbf{L}\sqrt{\tau_i}\mathbf{Q}_{ij}\mathbf{v}_k, d)|. \quad (10)$$

If the number of radii, n_R, is two, then the radial integral approximation has order five, as does the Mysovskikh spherical rule. Thus the overall integral approximation will integrate perfectly any multivariate quintic. For this reason, we recommend using two radii for the outer integral.

We now describe how to form the Mysovskikh extended simplex points, the generalized Gauss–Laguerre abscissas, and their corresponding weights. The extended simplex method begins by creating a $p+1$ vertex-centered simplex in $\mathbb{R}^p$,

$$\mathbf{v}_{ij} = \begin{cases} -\sqrt{\dfrac{p+1}{p(p-j+2)(p-j+1)}}, & j < i \\ \sqrt{\dfrac{(p+1)(p-i+1)}{p(p-i+2)}}, & j = i \\ 0, & j > i. \end{cases} \quad (11)$$

The construction proceeds by creating the midpoints of the simplex vertices, and is completed by adding the negatives of all vertices and midpoints.

The extended simplex weights take on only two values. The simplex vertex points and their negatives receive a weight of $p(7-p)/2(p+1)^2(p+2)$, and the midpoints and their negatives receive a weight of $2(p-1)^2/p(p+1)^2(p+2)$. The algorithm generates $(p+1)(p+2)$ abscissas, but in smaller dimensions, some points may be generated more than once; for example, if $p = 3$, then there are only 14 unique abscissas, rather than 20. The corresponding weights of the abscissa are then summed.

To approximate the radial integral, we use generalized Gauss–Laguerre quadrature, with a point added at 0. This added point at 0 is quite beneficial. With only one function evaluation, we are able to improve the order of the radial integral approximation by as much as adding another nonzero radius and extended simplex, which would require $O(p^2)$ function evaluations. Using the approach of Cassity (1965), the nonzero abscissas, τ_i, are the roots of the generalized Laguerre polynomial with parameter $p/2$, each multiplied by two. This scaling is because the radial integral is essentially with respect to a Gamma$(p/2, 2)$ density. It is important to note that adding the zero point causes the generalized Laguerre shape parameter to be incremented by one. Without the evaluation at 0, the abscissas would be twice the roots of the generalized Laguerre polynomial with parameter $p/2-1$. Cassity (1965) also provided an explicit formula for the quadrature weights. A specialized algorithm for finding the roots of the generalized Laguerre polynomials was given by Press et al. (1992), but when forming the coefficients of the orthogonal polynomial, any polynomial root finding algorithm may be used.

3.2 Nonnormal Priors

Although the numerical integration methodology detailed in the previous section is specialized to normally distributed priors, it can easily be adapted to a wide class of nonnormal priors, provided that the prior distributions on the parameters are independent. Suppose that there are p parameters, θ_i, each with its own prior density, $g_i(\theta_i)$. Assume that for each i, over the support of θ_i, g_i is differentiable and nonzero off of a set of Lebesgue measure 0, and let G_i represent the cumulative distribution function (cdf) for θ_i.

The Bayesian D-optimality criterion for this situation can be written as

$$\phi(d) = \int \log(I(\theta_1, \ldots, \theta_p; d)) \prod_{i=1}^{p} g_i(\theta_i)\, d\theta_i. \quad (12)$$

Let z_i, where $i = 1, 2, \ldots, p$ be a random sample from a Normal(0, 1) distribution. Then $G^{-1}(\Phi(z_i))$ has the same distribution as θ_i. Because the determinant of the Jacobian for the transformation from $\mathbf{z}$ to $\boldsymbol{\theta}$ is $|J(\mathbf{z})| = \prod_{i=1}^{p} \Phi(z_i)/g_i(G_i^{-1}(\Phi(z_i)))$, a convenient cancellation occurs in the change of variable from $\boldsymbol{\theta}$ to $\mathbf{z}$,

$$\phi(d) = \int \log\bigl(I\bigl(G_1^{-1}(\Phi(z_1)), \ldots, G_p^{-1}(\Phi(z_p)); d\bigr)\bigr) \times \prod_{i=1}^{p} \phi(z_i)\, dz_i. \quad (13)$$

This suggests an extension of our method to nonnormal priors. Let $\mathbf{G}^{-1}(\boldsymbol{\Phi}(\mathbf{z}))$ represent the vector mapping of $\mathbf{z}$ to the $\boldsymbol{\theta}$. The following expression can be viewed as a plausible approximation to the optimality criterion:

$$\phi(d) \approx w_{R_0} \log\bigl|\mathbf{I}\bigl(\mathbf{G}^{-1}(\boldsymbol{\Phi}(\mathbf{0})), d\bigr)\bigr| + \sum_{i=1}^{n_R} \sum_{j=1}^{n_Q} \sum_{k=1}^{n_S} \frac{w_{R_i} w_{S_k}}{n_Q} \log\bigl|\mathbf{I}\bigl(\mathbf{G}^{-1}(\boldsymbol{\Phi}(\sqrt{\tau_i}\mathbf{Q}_{ij}\mathbf{v}_k)), d\bigr)\bigr|.$$

We recommend using a larger number of radii and random rotations for nonnormal priors, because the integral approximation likely will not be as accurate as that for normal priors.

3.3 Integration of Test Functions

Here we present two simple examples to illustrate the accuracy of our quadrature algorithm for normal prior distributions. For both examples, we assume that $Z_1, \ldots, Z_p$ are independent standard normal random variables.

In the first example, we wish to approximate $\mathrm{E}\{\sum Z_i^2\}$. The algorithm gives the exact answer of p. Suppose that we were to approach this problem by Monte Carlo integration, sampling random normal vectors, computing their sums of squares, and estimating the integral by the average. As $\sum Z_i^2 \sim \chi_p^2$, we know that a Monte Carlo sample of size N will be unbiased for the correct answer with a standard error of $\sqrt{2p/N}$. With $p = 8$, a sample of 2500 points is needed to achieve a standard error that is 1% of the correct answer.

In the second example, we approximate $\mathrm{E}\{\exp[-\sum_{i=1}^{n} Z_i^2]\}$, for $n = 1, 2, \ldots, 8$. Table 1 gives the results using $n_R = 2, 4$, and 8 spheres, as well as the exact result, $3^{-n/2}$.

Table 1. Integral approximation as a function of the number of spheres used in the quadrature

	Number of spheres			
n	2	4	8	Exact
1	0.60403965	0.5790283	0.57735685	0.57735027
2	0.38259399	0.33704331	0.33335192	0.33333333
3	0.25573645	0.19795682	0.19248448	0.19245009
4	0.18040391	0.11786078	0.11116252	0.11111111
5	0.13362496	0.07151613	0.06421703	0.06415003
6	0.10312165	0.04447346	0.03711623	0.03703704
7	0.0822472	0.02848914	0.02147027	0.02138334
8	0.06732053	0.01887063	0.0124357	0.01234568

3.4 Relation to Experimental Designs

The spherical quadrature points in the scheme of Mysovskikh (1980) are closely related to the simplex-sum designs of Box and Behnken (1960a). Those p-factor designs are built up from a simplex of $p + 1$ points centered at the origin. Along with the simplex points, they include the sums of pairs, triplets, and so on of the vertices. Because the sum of all of the simplex vertices is the origin, summing any p vertices gives the negative of the remaining vertex, as in the quadrature points. Similarly, adding any $p - 1$ vertices gives the negative of the sum of the remaining two vertices.

Some well-known designs occur when $p + 1$ is a multiple of 4. In that case, the rows of a saturated Plackett and Burman (1946) design form a simplex that can be used to generate the quadrature points. For $p = 3$ dimensions, take the points of the 2^{3-1} design as the simplex. As Box and Behnken (1960a) pointed out, the simplex-sum design is equivalent to the 3-factor central composite design, with edge midpoints giving the axial points. For the quadrature rule, the axial points will be cospherical with the factorial points. The quadrature scheme assigns a weight of 3/40 to each cube point and a weight of 3/45 to each axial point. Note that the axial points occur twice in the quadrature scheme (once from the original simplex and again from the folded over simplex), and so have double the standard midpoint weight.

For $p = 7$, the vertices of a 2^{7-4} design provide a simplex. The midpoints of the edges of this simplex, and the folded simplex, are precisely the noncenter points of the seven-factor design derived by Box and Behnken (1960b) using a construction based on combining balanced incomplete block designs with two-level factorials. Box and Behnken (1960a) found the same design with the simplex-sum approach. They showed that good designs with smaller sample size sometimes could be found by including only sums of certain order. The case of $p = 7$ exemplifies that strategy, with only sums of two or six vertices included. Thus this simplex-sum design does not include the original vertices or their negatives. Interestingly, the quadrature scheme here gives 0 weight to the original vertices and their negatives, so that only the points of the Box–Behnken design are used.

Crosier (1991) used the vertex midpoint technique to create new three-level response surface designs for 11 and 15 factors. Our quadrature points include the points in Crosier's designs along with the simplex points, which now will have positive

weight. Also note that the quadrature points are all equidistant from the origin, so the factor levels for the factorial (simplex) points will be smaller than those for the vertex midpoints, each of which have $(p+1)/2$ factors set at 0 and the remaining $(p-1)/2$ factors set to the same absolute value.

Morris (2000) proposed the class of augmented pairs designs. These designs were constructed by taking an initial set of points and then adding the midpoint of every edge joining two points. The scheme is nearly identical to that for the quadrature points; however, Morris did not require the initial set to be a simplex, did not project the midpoints onto the sphere circumscribing the initial points, and did not fold over his designs. Morris intentionally set up his initial points with all factors at two levels, which guarantees that the final design has all factors at three levels. When $p+1$ is not a multiple of 4, these initial designs do not generate a simplex.

Doehlert (1970) and Crosier (1993) developed additional classes of designs that exploit a simplex of points with the origin as one of the vertices rather than at the center of the simplex. Those designs include the vertices of the simplex and all possible differences of two points, rather than sums of points.

4. A CANDIDATE SET–FREE ALGORITHM FOR CALCULATING THE OPTIMAL DESIGN

To maximize $\phi(d)$, we use a fairly general but simple-minded approach similar in spirit to the coordinate exchange algorithm of Meyer and Nachtsheim (1995). The algorithm begins by generating integration abscissas. We use the same set of abscissas throughout the remainder of the algorithm for evaluating the objective function, to improve the accuracy in comparing any two candidate designs. We then generate random starting designs until a nonsingular starting design has been found. We proceed to an iterative stage in which each iteration consists of a pass through all the runs. For each run, we loop through the factors, optimizing them one at a time. If the factor is categorical, then we evaluate the design at all levels of that factor and retain the level that had the most desirable value of the objective function. For numerical factors, we carry out a full continuous optimization using the method of Brent (1973) after initializing the factor to a random value within its experimental range. This approach circumvents the need to set up a grid of candidate levels for numerical factors. The gridding technique is problematic; it can possibly lead to inefficient designs if the grid is too coarse, yet can make the optimization very slow if the grid is too fine.

For many models, such as nonlinear regression models and GLMs, a change to a row of the design matrix leads to a low-rank change of the information matrix. In these special cases we may simply integrate the log of the multiplicative changes in the determinants with respect to the prior. Because this makes it unnecessary to rebuild and reinvert the whole information matrix at each of the parameter values, the optimization algorithm is much faster than it would be otherwise. The formula for evaluating these multiplicative changes is a natural generalization of that of Meyer and Nachtsheim (1995) and has been given by Fedorov (1972, p. 161).

After each iteration, we compute the total change in $\phi(d)$ since the end of the last iteration. If this change in objective function is small, say $<1e{-}8$, then we terminate the iterations. Once the iterations are completed, a final cleanup step is performed. Similar to the findings of Dror and Steinberg (2006), the designs found by our algorithm frequently consist of clusters of points. These clusters often would be replicates if the design were calculated with perfect precision. We collapse these clusters by performing one iteration of a row-exchange algorithm using the design as its own candidate set. In practice, this has two benefits. It increases the efficiency of the design and also provides an important practical benefit, facilitating execution of the resulting experiment by reducing the number of levels of the continuous factors that must be run, often significantly so.

5. EXAMPLES

5.1 Logistic Regression

We return to the experiment described by Woods et al. (2006), which assessed the effects of four factors on formation of a new product using a first-order logistic regression model. Woods et al. (2006) sought designs that optimize $\phi(d)$ and considered three uniform prior distributions for the regression parameters. Dror and Steinberg (2006) also examined this the problem and applied a different design algorithm based on clustering the points in a collection of locally D-optimal designs.

We apply our method to this same logistic regression example, computing both 16-run and 48-run designs. We use the prior labelled β_3 in table 1 of Woods et al. (2006), which assigns independent uniform distributions with rather large uncertainty to all of the regression coefficients. Our 16-run design, presented in Table 2, took less than 10 seconds to compute on a Pentium 4 computer with 2 GB of RAM and a 1.6-GHz processor. We assessed the quality of a design as was done previously, by computing its local D-efficiency with respect to each point in a set of coefficient vectors sampled from the prior. The median efficiency for our design was 44.8%, and the 10th percentile was 31.5%. The comparable figures for the Woods design are 42.3% and 26.3%. Our design had higher efficiency for 68.3%

Table 2. Our 16-run design for the crystallography experiment

Run	x_1	x_2	x_3	x_4
1	0.7381	−0.4912	−1	0.7268
2	1	−0.5506	0.8877	1
3	0.7425	0.0812	0.2066	−0.9529
4	0.5571	−1	0.6348	0.9904
5	−1	−0.0203	−0.9275	−0.4018
6	−0.0967	−0.2200	−1	−1
7	0.2960	−0.5898	0.4767	−1
8	−0.8731	1	0.7421	−1
9	−0.8575	0.4135	1	1
10	0.4973	−1	−1	1
11	−0.2031	1	0.6207	0.9616
12	1	−1	−0.3794	−0.6266
13	−0.9325	0.3328	−1	1
14	−0.6186	0.8939	1	1
15	0.7306	−0.1954	1	−1
16	−1	0.8394	−0.9012	−0.3907

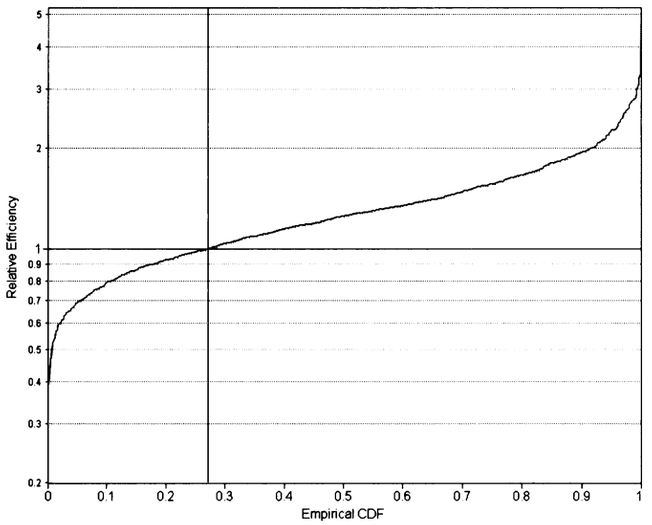

Figure 1. The empirical cdf of the relative efficiencies.

of the vectors and in some cases was more than twice as efficient.

The actual plan in the crystallography experiment was to perform 48 runs. Woods et al. (2006) recommended running 3 replicates of their 16-run design. With our computational methods, we were able to evaluate and optimize a good approximation to the $\phi(d)$ criterion for a full 48-run experiment. We again computed efficiencies relative to the 48-run local optimal design for a sample of coefficient vectors sampled from the prior. Our design had a median efficiency of 38.6% and a 10th percentile of 28.0%, whereas the Woods et al. design, replicated three times, had a median efficiency of 36.2% and a 10th percentile of 27.2%. Directly comparing the two designs shows that the relative efficiency of our design had a median of 109.1% and exceeded 1.44 for >10% of the vectors. Figure 1 shows the empirical cdf of the relative efficiencies of our design versus that of Woods et al. Our design is also more efficient than the 48-run experiment of Dror and Steinberg (2006). Over a sample of 1000 parameter vectors from the prior, our design had a median relative efficiency of 104% and was more efficient for 60.8% of the sampled vectors.

5.2 Compartmental Model

Here we consider the experimental design problem for the compartmental model presented in Equation (4). This problem also was studied by Atkinson et al. (1993). To allow comparison of results, we adopt the same prior distribution that they used: $\theta_1 \sim \mathrm{U}(0.01884, 0.09884)$, $\theta_2 \sim \mathrm{U}(0.298, 8.298)$, and θ_3 assumed to be equal to 21.8. The priors on θ_1 and θ_2 reflect parameter values actually estimated from Button's (1979) experimental data. For θ_3, it is easy to show that the design efficiency does not depend on the value of this parameter, and thus there is no need to assign it a prior distribution at the design stage.

We apply our approach to derive an 18-run Bayesian D-optimal design, matching the sample size actually used by Button (1979). The design is shown in Table 3, along with that of an 18-run approximation to the continuous Bayesian D-optimal design found by Atkinson et al. (1993), using the same support points and setting the number of runs to be as close as possible to 18 times the optimal weights. Our design has four support points, compared with the five support points in the design of Atkinson et al.

Table 3. Designs for the compartmental model, with 18 runs, produced by our algorithm and by rounding the continuous design of Atkinson et al. (1993)

Run	Our design	Atkinson et al.
1	0.2030	0.2034
2	0.2030	0.2034
3	0.2030	0.2034
4	0.2030	0.2034
5	0.2030	0.2034
6	0.9717	1.1967
7	0.9717	1.1967
8	0.9717	1.1967
9	0.9717	1.1967
10	3.3439	2.8323
11	3.3439	2.8323
12	3.3439	7.8229
13	20.1917	20.1899
14	20.1917	20.1899
15	20.1917	20.1899
16	20.1917	20.1899
17	20.1917	20.1899
18	20.1917	20.1899

Before comparing the designs, we note that Atkinson et al. (1993) did not provide exact details on how they computed the average log determinant criterion. Our design was found by integrating over just a small set of points in the parameter space. Furthermore, we chose to use uniform priors, following Atkinson et al., even though our method works more efficiently with normal priors. Comparisons of the designs are based on the local efficiencies of both designs and on their relative efficiencies to one another, at a large number of parameter vectors sampled from the prior. For our design, the median efficiency is 83.0% and the 10th percentile is 67.9%, compared with 84.3% and 63.3% for the design of Atkinson et al. The relative efficiency of our design compared with the design of Atkinson et al. is close to 1 throughout most of the parameter space, with a median of 99.6% and 10th and 90th percentiles of 96.4% and 106.9%. The lowest efficiencies for both designs are seen for small values of θ_2, but the degradation is more severe for our design. As a result, some relative efficiencies fall below 80%.

5.3 Augmented Logistic Regression

Here we show how our methods can be used to efficiently add runs to an experiment. We carried out 89 experimental runs to study a computer-guided metal-working process. We have coded the three input factors to the interval $[-1, 1]$. One important outcome was whether or not wrinkling was observed in the finished piece. Figure 2 plots the wrinkling outcome against the two most influential factors, revealing a roughly elliptical region in which wrinkling did not occur. A quadratic logistic regression model with 10 parameters fits the data well. There are highly significant second-order effects for both strong factors, but considerable uncertainty remains; for example, the linear and quadratic terms in factor E have coefficients of -8.87 and 14.22, with standard errors of 2.29 and 3.93.

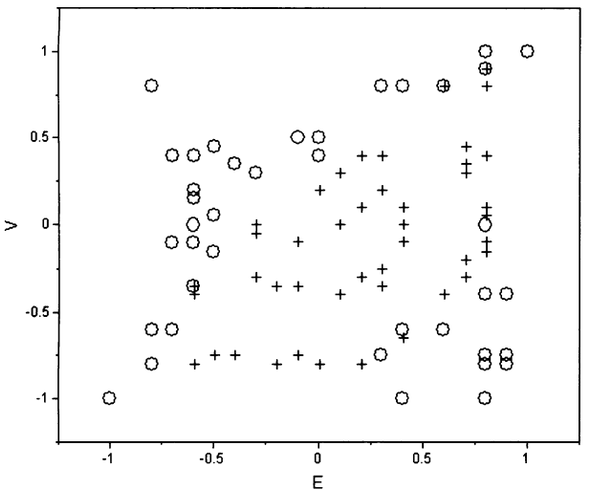

Figure 2. Results from the metal-working process. The plot shows the two most important input factors, E and V. Successful pieces are marked with +; wrinkled pieces, with ∘.

Inspection of the regression output reveals that the estimated probability was >0.9 for 31 points and <0.1 for another 26 points. These points have low weights and so make only small contributions to the information matrix.

Now suppose that resources are available to make 120 pieces altogether. How should the remaining 31 runs be chosen to gain more precise estimates of the parameters in the model for wrinkling? For this problem, we can use the analysis of the data on hand to generate a prior distribution for the model parameters. We assume a multivariate normal prior centered at the maximum likelihood estimates and use the inverse information matrix as the covariance matrix. The augmentation points are shown in Table 4, along with their weights at the prior mean. Most of the weights are >0.2.

To assess the quality of a robust augmentation design, we use a slightly different approach than that used in the two preceding examples. The measure of design efficiency given in Equation (5) compares the information matrices from two designs. In augmentation problems, those matrices share all of the runs already performed; thus the ratio of determinants often will be close to 1 and will not be informative. Instead, following Chaloner and Verdinelli (1995), we consider $\log(|I(\boldsymbol{\theta}; d)|)$ to be the utility when $\boldsymbol{\theta}$ is the parameter vector and d is the design. Now suppose that we have an initial design d_0 with n_0 runs and we augment it to a final design d with n runs. The total gain in utility from using d rather than d_0 is $\log(|I(\boldsymbol{\theta}; d)|) - \log(|I(\boldsymbol{\theta}; d_0)|)$. Part of the utility gain is due to better choices of the design points, but part also is attributable simply to the larger sample size. We prefer to remove this latter component from the utility gain, and we do so by treating d_0 as a design measure and then computing the information that we would get from a design like d_0 but with n runs. A simple calculation shows that this is achieved by adding $p\log(n/n_0)$ to the utility for d_0. Thus the gain in utility from effective design is

$$\delta(d) = \log\big(|I(\boldsymbol{\theta}; d)|\big) - \big[\log\big(|I(\boldsymbol{\theta}; d_0)|\big) + p\log(n/n_0)\big]. \quad (14)$$

Table 4. Our augmentation design with weights corresponding to the prior mean

Run	E	V	F	Weight
1	1	−0.0167557	−1	0.24305936
2	−0.0007006	0.45877387	−1	0.24460862
3	−0.6014322	−1	−1	0.24342047
4	−0.0007006	0.45877387	−1	0.24460862
5	1	−0.0167557	−1	0.24305936
6	−0.6014322	−1	−1	0.24342047
7	0.96912724	0.34225494	1	0.23535021
8	1	1	−1	0.24720689
9	−0.0600845	−1	−1	0.21796003
10	0.03117945	0.25061888	1	0.24731592
11	1	−0.017285	−1	0.24277398
12	0.03117945	0.25061888	1	0.24731592
13	0.4	−0.6	−0.92	0.17156421
14	1	1	−1	0.24720689
15	−0.0600845	−1	−1	0.21796003
16	0.03117945	0.25061888	1	0.24731592
17	−0.2236767	−1	1	0.23260494
18	−0.6014322	−1	−1	0.24342047
19	−0.0007006	0.45877387	−1	0.24460862
20	1	−0.0236703	0.15794961	0.22858891
21	0.96912724	0.34225494	1	0.23535021
22	−0.2236767	−1	1	0.23260494
23	−0.5	−0.75	0.84	0.18938314
24	1	1	−1	0.24720689
25	1	1	−1	0.24720689
26	0.8	0.8	0.68	0.20272361
27	−0.0007006	0.45877387	−1	0.24460862
28	−0.0600845	−1	−1	0.21796003
29	0.96912724	0.34225494	1	0.23535021
30	−0.2236767	−1	1	0.23260494
31	0.3	−0.75	0.76	0.22407826

A proposed robust augmentation design d can now be compared, for a fixed value of $\boldsymbol{\theta}$, with the locally optimal augmentation $d^*(\boldsymbol{\theta})$, by the information gain ratio, $\delta(d)/\delta(d^*(\boldsymbol{\theta}))$.

We computed information gain ratios for our augmentation design for 1000 parameter vectors sampled from the posterior distribution of $\boldsymbol{\theta}$. For comparison purposes, we also considered two other designs. The first of these designs was found by applying our methodology with a prior with the same means and standard deviations but with the correlations set equal to 0; the second is the locally optimal augmentation at the maximum likelihood estimates.

Some relevant quantiles from the augmentation ratios are summarized in Table 5. The table clearly shows that the best augmentation design is the one that uses the posterior distribution, complete with covariance information and the worst design is the locally optimal design at the maximum likelihood estimates. Including the correlation in the prior gave a more efficient augmentation than is achieved with independent priors for >98% of the sample. The median improvement in information gain ratio was by a factor of 2.35, and the lower quartile was 1.93.

6. DISCUSSION

Bayesian design criteria are very attractive for nonlinear problems as a means to achieve designs that are robust to the

Table 5. Comparison of three augmentation designs in terms of quantiles of the ratio of information gained to that which would be gained by a locally optimal design, over a sample of 1000 parameter vectors from a prior distribution based on the initial data. The first two designs take priors from the initial data, the first using all of the information including correlations and the second using only means and standard deviations and then assuming independence. The third design is the locally optimal design at the maximum likelihood estimates from the initial data

	Percentiles of information gain ratios				
Prior	2.5	25	50	75	97.5
Correlated parameters	0.29	0.54	0.65	0.76	0.88
Independent parameters	0.14	0.24	0.29	0.33	0.40
Maximum likelihood estimate local	−0.12	0.04	0.11	0.17	0.26

unknown parameter values. Using these criteria poses computational challenges that have severely limited their use in practice, however. In this article we have described a simple integration method that can dramatically streamline these computations, making implementation of Bayesian criteria much more practical.

We have used several examples to illustrate that our integration scheme can achieve better accuracy than simple Monte Carlo integration while making far fewer function evaluations. We have shown that efficient, robust experimental designs can be found very quickly.

An important benefit of our procedure is that it allows the use of Bayesian design criteria in much larger problems than previously considered. As the number of parameters in the model increases, Monte Carlo samples must increase at an exponential rate to maintain accuracy. The number of function evaluations needed to use a Bayesian design criterion can be prohibitively large even with just four or five parameters (as in Woods et al. 2006). In our method, however, the number of function evaluations increases only as the square of the number of parameters, and thus we are able to extend the application of Bayesian robust design to a much wider realm of problems.

ACKNOWLEDGMENTS

The work of D.S. was carried out in part while he was on sabbatical at SAMSI.

[Received December 2007. Revised June 2008.]

REFERENCES

Atkinson, A. C., Chaloner, K., Herzberg, A. M., and Juritz, J. (1993), "Optimum Experimental Designs for Properties of a Compartmental Model," *Biometrics*, 49, 325–337.

Atkinson, A., Donev, A., and Tobias, R. (2007), *Optimum Experimental Designs, With SAS*, Oxford: Oxford University Press.

Box, G. E. P., and Behnken, D. W. (1960a), "Simplex-Sum Designs: A Class of Second Order Rotatable Designs Derivable From Those of First Order," *Annals of Mathematical Statistics*, 31, 838–864.

——— (1960b), "Some New Three-Level Designs for the Study of Quantitative Variables," *Technometrics*, 2, 455–475.

Box, G. E. P., and Lucas, H. (1959), "Design of Experiments in Non-Linear Situations," *Biometrika*, 46, 77–90.

Brent, R. P. (1973), *Algorithms for Minimization Without Derivatives*, Englewood Cliffs, NJ: Prentice Hall.

Button, C. (1979), "Pharmacokinetics, Bioavailability, and Dosage Regimens of Digoxin in the Horse and Dog," Ph.D. thesis, Texas A&M University.

Cassity, C. R. (1965), "Abscissas, Coefficients, and Error Term for the Generalized Gauss–Laguerre Quadrature Formula Using the Zero Ordinate," *Mathematics of Computation*, 19, 287–296.

Chaloner, K., and Larntz, K. (1989), "Optimal Bayesian Design Applied to Logistic Regression Experiments," *Journal of Statistical Planning and Inference*, 21, 191–208.

Chaloner, K., and Verdinelli, I. (1995), "Bayesian Experimental Design: A Review," *Statistical Science*, 10, 273–304.

Chernoff, H. (1953), "Locally Optimal Designs for Estimating Parameters," *Annals of Mathematical Statistics*, 24, 586–602.

Crosier, R. B. (1991), "Some New Three-Level Response Surface Designs," Technical Report CRDEC–TR-308, U.S. Army Chemical Research, Development and Engineering Center.

——— (1993), "Shell Designs," Technical Report ERDEC–TR-104, U.S. Army Chemical Research, Development and Engineering Center.

Doehlert, D. H. (1970), "Uniform Shell Design," *Journal of the Royal Statistical Society*, Ser. C, 19, 231–239.

Dror, H. A., and Steinberg, D. M. (2006), "Robust Experimental Design for Multivariate Generalized Linear Models," *Technometrics*, 48, 520–529.

Fedorov, V. (1972), *Theory of Optimal Experiments*, New York: Academic Press.

Ford, I., Kitsos, D., and Titterington, D. M. (1989), "Recent Advances in Nonlinear Experimental-Design," *Technometrics*, 31, 49–60.

Khuri, A., Mukherjee, B., Sinh, B., and Ghosh, M. (2006), "Design Issues for Generalized Linear Models," *Statistical Science*, 21, 376–399.

Meyer, R., and Nachtsheim, C. (1995), "The Coordinate Exchange Algorithm for Constructing Exact Optimal Experimental Designs," *Technometrics*, 37, 60–69.

Monahan, J., and Genz, A. (1997), "Spherical-Radial Integration Rules for Bayesian Computation," *Journal of the American Statistical Association*, 92, 664–674.

Morris, M. D. (2000), "A Class of Three-Level Experimental Designs for Response Surface Modeling," *Technometrics*, 42, 111–121.

Mysovskikh, I. P. (1980), "The Approximation of Multiple Integrals Using Interpolatory Cubature Formulae," in *Quantitative Approximation*, eds. R. A. Devore and K. Sherer, New York: Academic Press, pp. 217–243.

Plackett, R. L., and Burman, J. P. (1946), "The Design of Optimum Multifactorial Experiments," *Biometrika*, 33, 305–325.

Press, W., Teuklosky, S., Vetterling, W., and Flannery, B. (1992), *Numerical Recipies in C* (2nd ed.), New York: Cambridge University Press.

Woods, D. C., Lewis, S. M., Eccleston, J. A., and Russell, K. G. (2006), "Designs for Generalized Linear Models With Several Variables and Model Uncertainty," *Technometrics*, 48, 284–292.