Available online at www.sciencedirect.com

SCIENCE DIRECT®

Journal of Statistical Planning and Inference 136 (2006) 1103–1119

journal of statistical planning and inference

www.elsevier.com/locate/jspi

Comparison of designs for computer experiments☆

Dizza Bursztyn[a], David M. Steinberg[b,*]

[a]*Ashkelon College, Ashkelon, Israel*

[b]*Department of Statistics and Operations Research, Raymond and Beverly Sackler Faculty of Exact Sciences, Tel Aviv University, Tel-Aviv 69978, Israel*

Received 19 June 2003; accepted 17 August 2004
Available online 8 October 2004

Abstract

For many complex processes laboratory experimentation is too expensive or too time-consuming to be carried out. A practical alternative is to simulate these phenomena by a computer code. This article considers the choice of an experimental design for computer experiments. We illustrate some drawbacks to criteria that have been proposed and suggest an alternative, based on the Bayesian interpretation of the alias matrix in Draper and Guttman (Ann. Inst. Statist. Math. 44 (1992) 659). Then we compare different design criteria by studying how they rate a variety of candidate designs for computer experiments such as Latin hypercube plans, U-designs, lattice designs and rotation designs.
© 2004 Elsevier B.V. All rights reserved.

Keywords: Alias matrix; Entropy criterion; Latin hypercube designs; Lattice designs; Maximin designs; Mean squared error criterion; Random field regression; Rotation designs

1. Introduction

Computer simulators have replaced laboratory experiments in the study of many complex processes. The major improvements in computing power have made this a cost-effective experimental technology. Computer experiments typically involve complex systems with numerous input variables. Computer experiments are deterministic: replicate observations from running the code with the same inputs will be identical. As such, standard approaches to the design and analysis of experiments are not necessarily appropriate for computer

☆ This research was supported by a grant from the Israeli Science Foundation.

* Corresponding author. Tel.: +972 3 640 8043; fax: +972 3 640 9357.
E-mail addresses: dizzal@bezeqint.net (D. Bursztyn), dms@post.tau.ac.il (D.M. Steinberg).

0378-3758/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2004.08.007

experiments. In modeling data from a computer experiment, there is no need to be concerned with reducing variance, only bias due to model inadequacy. At the design stage, concepts like blocking and randomization are irrelevant and there is no immediate way to apply standard optimality criteria that are functionals of the covariance matrix. See Sacks et al. (1989) and Kennedy and O'Hagan (2001) for good general discussions of statistical problems in computer experiments.

In this article we consider criteria for comparing experimental layouts for computer experiments. We present some drawbacks to criteria that have been proposed in recent years and suggest an alternative criterion based on the Bayesian interpretation of the alias matrix in Draper and Guttman (1992).

The article is organized as follows. In Section 2, we describe the random field regression model, which serves as the basis for many of the current design criteria. In Section 3, we survey design criteria that have appeared in other articles and point out some problems in using them. In Section 4, we propose a new criterion for design of computer experiments. Section 5 compares the criteria with respect to their treatment of sample size, projected designs and replicate design points. Section 6 describes a comparison of the criteria, using a variety of designs, including two-level factorials, Latin hypercubes, U-designs, lattice designs and rotation designs. A summary of the results is given in Section 7.

2. Random field regression models

Let $y(\mathbf{x})$ denote the output of the simulator that results from the p-dimensional input $\mathbf{x}=(x_1,\ldots,x_p)$. A popular modeling approach is to treat $y(\mathbf{x})$ as a realization of a random field, $Y(\mathbf{x})$, that includes a regression model (Sacks et al., 1989; Welch et al., 1992; Bates et al., 1996)

$$Y(\mathbf{x})=\sum_{j=0}^{k}\beta_j f_j(\mathbf{x})+Z(\mathbf{x}). \tag{1}$$

Notice that $Z(\mathbf{x})$ represents the systematic departure from the linear model $\sum\beta_j f_j(\mathbf{x})$. The random field regression model assumes that $Z(\mathbf{x})$ is Gaussian with zero mean, constant variance σ^2, and with a correlation structure $R(\mathbf{x}_1,\mathbf{x}_2)$ between p-dimensional input vectors $\mathbf{x}_1$ and $\mathbf{x}_2$. The fact that there is no measurement error in computer experiments is reflected by requiring that $R(\mathbf{x}_1,\mathbf{x}_2)$ tend to 1 as the Euclidean distance between $\mathbf{x}_1$ and $\mathbf{x}_2$ tends to 0. The particular form studied by Sacks et al. (1989) is

$$R(\mathbf{x}_1,\mathbf{x}_2)=\mathrm{corr}(Y(\mathbf{x}_1),Y(\mathbf{x}_2))=\prod_{j=1}^{p}\exp(-\lambda_j|x_{1j}-x_{2j}|^{\alpha_j}) \tag{2}$$

with $\lambda_j\geqslant 0$ and $0<\alpha_j\leqslant 2$. Denote the covariance function by Γ, so that $\Gamma(\mathbf{x}_1,\mathbf{x}_2)=\sigma^2 R(\mathbf{x}_1,\mathbf{x}_2)$.

For these stochastic models, the output $Y(\mathbf{x})$ of the simulator is estimated by the best linear unbiased predictor (BLUP) of the random field (Robinson, 1991). Given a design

$S = \{\boldsymbol{s}_1, \ldots, \boldsymbol{s}_n\}$ and data $\boldsymbol{Y}_S = [Y(\boldsymbol{s}_1), \ldots, Y(\boldsymbol{s}_n)]'$ the BLUP of $Y(\mathbf{x})$ is

$$\hat{\mathbf{Y}}(\mathbf{x}) = \mathbf{f}'(\mathbf{x})\hat{\boldsymbol{\beta}} + \mathbf{r}'(\mathbf{x})\boldsymbol{R}^{-1}(\mathbf{Y}_s - F\hat{\boldsymbol{\beta}}), \tag{3}$$

where $\hat{\boldsymbol{\beta}} = (\boldsymbol{F}'\boldsymbol{R}^{-1}\boldsymbol{F})^{-1}\boldsymbol{F}'\boldsymbol{R}^{-1}\mathbf{Y}_S$ is the generalized least-squares estimator of β, and we use the following notation:

$$\mathbf{f}(\mathbf{x}) = [\boldsymbol{f}_0(\mathbf{x}), \ldots, \boldsymbol{f}_k(\mathbf{x})]',$$

$$\boldsymbol{F} = \begin{pmatrix} \mathbf{f}'(\boldsymbol{s}_1) \\ \cdot \\ \cdot \\ \cdot \\ \mathbf{f}'(\boldsymbol{s}_n) \end{pmatrix} \quad \text{is the } n \times (k+1) \text{ expanded regression matrix,}$$

$R = \{R(\boldsymbol{s}_i, \boldsymbol{s}_j)\}$, $1 \leqslant i \leqslant n$; $1 \leqslant j \leqslant n$, is the $n \times n$ correlation matrix of $(Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n))$, and $\mathbf{r}(\mathbf{x}) = [R(\boldsymbol{s}_1, \mathbf{x}), \ldots, R(\boldsymbol{s}_n, \mathbf{x})]'$ is the vector of correlations between the Z's at the design sites and at the estimation input $\mathbf{x}$. Notice that $Y(\boldsymbol{s}_i) = \hat{Y}(\boldsymbol{s}_i)$ for $i = 1, \ldots, n$.

Alternatively $Y(\mathbf{x})$ in (1) can be regarded as a Bayesian prior on the true response function of the simulator, with the βs either specified a priori or given a prior distribution (Sacks et al., 1989; Steinberg, 1990; Bursztyn and Steinberg, 2002). The random field $Z(\mathbf{x})$ is viewed as a prior distribution reflecting uncertainty about the true response function. In the Bayesian approach the predictor of the response is the posterior mean of $Y(\mathbf{x})$ and, in the case that $Z(\mathbf{x})$ is Gaussian and improper uniform priors are assigned to the βs, it is exactly the BLUP presented above; see Sacks et al. (1989) and Morris et al. (1993).

The parameters in the correlation function can be estimated by maximum likelihood or cross-validation (Sacks et al., 1989; Currin et al., 1988, 1991; Welch et al., 1992).

3. Experimental design criteria

A good design for a computer experiment should facilitate accurate prediction. Standard design criteria like A- or D-optimality, which are based on the covariance matrix of $\hat{\boldsymbol{\beta}}$, are not suitable for computer experiments because there is no random error. We describe here several criteria that have been proposed for computer experiments. The first three criteria are meaningful only in the context of a random field regression model and the fourth criterion was derived by considering implications of such a model.

See Bates et al. (1996) for a thorough discussion of different criteria for design comparison.

1. *The integrated mean squared error* (IMSE) *criterion*: A good design should minimize IMSE, defined as

$$\text{IMSE} = \int E\{Y(\mathbf{x}) - \hat{Y}(\mathbf{x})\}^2 \, d\mathbf{x}. \tag{4}$$

The expectation is taken with respect to the random field. Following the Bayesian interpretation, the BLUP $\hat{Y}(\mathbf{x})$ is the posterior mean of $Y(\mathbf{x})$, so the IMSE is in fact $\int \text{Var}(Y(\mathbf{x}))$, where $\mathrm{Var}(Y(\mathbf{x}))$ is the posterior variance of $Y(\mathbf{x})$ given Y_S. A number of articles have explored this criterion, including Sacks and Ylvisaker (1966, 1968, 1970), Steinberg (1985), and Sacks et al. (1989).

2. *The entropy criterion*: Lindley (1956) proposed use of the change in entropy before and after collecting data as a measure of the information provided by an experiment. The basic form of the entropy criterion for computer experiments is

$$E\{\Delta H(Y)\}, \tag{5}$$

where $H(Y)$ is the entropy of the random field and $\Delta H(Y)$ is the reduction in entropy after observing Y_S. A good design should maximize the expected reduction in entropy. See Shewry and Wynn (1987), Currin et al. (1991), Bates et al. (1996), and Koehler and Owen (1996).

For a random field, the expected reduction in entropy is equal to the entropy of the field at the n design sites

$$E\{\Delta H(Y)\} = H(Y_S) = (n/2)[1 + \ln(2\pi)] + 0.5\ \ln[\det(\Gamma)], \tag{6}$$

where Γ is the covariance matrix of $\boldsymbol{Y}_S$ (Shewry and Wynn, 1987). This result has been applied in several different ways to the random field regression model described in Section 2, generating a number of related design criteria.

The most obvious application of Eq. (6) is to simply use the entropy of $(Z(\boldsymbol{s}_1), \ldots, Z(\boldsymbol{s}_n))$ (e.g. Currin et al., 1991). Equivalently, one obtains as a design goal to maximize

$$E_1 = [\det(\Gamma)]^{1/n}. \tag{7}$$

In the stationary case, one can obviously use the correlation matrix R in Eq. (7) rather than the covariance matrix Γ. As noted by Bates et al. (1996), the E_1 criterion is appropriate when the fixed regression part of the model is limited to a constant, as recommended by Welch et al. (1992). Bates et al. (1996) and Koehler and Owen (1996) also consider limiting forms when the coefficients of the fixed regression terms are assigned a prior distribution, $\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{b}, \tau^2\Sigma)$, so that $\Gamma = \tau^2 F\Sigma F' + \sigma^2 R$. Rules for determinants of sums of matrices can be used to show that

$$\det(\Gamma) = \sigma^{2n} \det(\tau^2\Sigma) \det(R) \det(F'R^{-1}F + \tau^{-2}\Sigma^{-1}). \tag{8}$$

The first two terms on the right-hand side of (8) are independent of the design, so the last two can be taken as the criterion. Two special cases using the last two terms of (8) have received the most attention. First, a non-informative prior distribution for the regression coefficients is obtained in the limit as $\tau^2 \to \infty$. The corresponding design criterion is

$$\det(R) \det(F'R^{-1}F). \tag{9}$$

Second, the fixed regression model may be limited to a single term for an overall mean, so that matrix F is just a column of 1s. This model then leads to the criterion

$$E_2 = \det(R)|R^{-1}|, \tag{10}$$

where || is defined to be the sum of the entries of a matrix.

For the remainder of this article we will focus on the E_1 criterion.

3. Johnson et al. (1990) showed that, using a correlation structure of form (2) with $\alpha = 2$ for all factors, the computation of $\det(\Gamma)$ is dominated by those design points that are closest to one another. They found that efficient designs for the determinant criterion can also be found by maximizing the criterion

$$\boldsymbol{E}_3 = \min_{i<j} \|\boldsymbol{x}_i - \boldsymbol{x}_j\|, \tag{11}$$

where $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|$ is the p-dimensional Euclidean distance between design points and the minimization is over all pairs of design points. Optimal designs with respect to this criterion are called *maximin* designs.

4. The alias sum of squares criterion

We propose here an alternative design criterion based on the alias matrix for a simple approximating model. The basic philosophy behind our criterion is that a good design for a computer experiment should be efficient for factor screening and should also have the flexibility to entertain more complex models in those factors that are active. We use a spectrum of potential high-degree polynomial terms to reflect, at the design stage, the requirement of modeling flexibility.

4.1. An approximate regression model

Suppose we begin the analysis of data from a computer experiment by using OLS to fit a first-order regression model,

$$\boldsymbol{y}_i = \beta_0 + \sum_{j=1}^{p} \beta_j \boldsymbol{x}_{ij} = \mathbf{f}(\mathbf{x}_i)\boldsymbol{\beta}, \tag{12}$$

which is a natural first step in factor screening. The notation here is identical to that in Section 2, but with the proviso that **f(x)** in Section 2 referred to a general regression model whereas here we take the specific case of a first-order model. We realize that the first-order model will not provide a perfect description of the output data. So we assume, as in Box and Draper (1959, 1963), that adding extra terms to the first-order model will give us a nearly exact representation,

$$\boldsymbol{y}_i = \beta_0 + \sum_{j=1}^{p} \beta_j \boldsymbol{x}_{ij} + \sum_{j=p+1} \beta_j \boldsymbol{f}_j(\boldsymbol{x}) = \mathbf{f}'(\mathbf{x}_i)\boldsymbol{\beta} + \mathbf{f}_2'(\mathbf{x}_i)\boldsymbol{\beta}_2. \tag{13}$$

The full set of experimental data can then be written in matrix form as

$$\boldsymbol{y} = \begin{bmatrix} \mathbf{f}'(\mathbf{x}_1) \\ \vdots \\ \mathbf{f}'(\mathbf{x}_n) \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} \mathbf{f}_2'(\mathbf{x}_1) \\ \vdots \\ \mathbf{f}_2'(\mathbf{x}_n) \end{bmatrix} \boldsymbol{\beta}_2 = X\boldsymbol{\beta} + X_2\boldsymbol{\beta}_2. \tag{14}$$

Here $X_2\boldsymbol{\beta}_2$ represents the extra terms not included in the original first-order regression model.

We will assume throughout that all the regression functions are an orthonormal set with respect to some weight function $w(\mathbf{x})$; i.e.

$$\int f_u(\mathbf{x}) f_v(\mathbf{x}) w(\mathbf{x})\,\mathrm{d}\mathbf{x} = \boldsymbol{\delta}_{uv}, \tag{15}$$

where $\boldsymbol{\delta}_{uv} = 1$ if $u = v$ and $= 0$ otherwise. We consider here only the case of a constant weight function and set the constant so that $f_u(\mathbf{x}) = x_u$, $u = 1, \ldots, p$. The integration extends over the entire domain of the input factors, which we take to be $[-1, 1]^p$. For the intercept term in the model, we use Eq. (15) to achieve orthonormality rather than taking it to be 1.

4.2. The alias matrix

We then find that the least-squares estimator for $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (X'X)^{-1}X'\mathbf{y} = \boldsymbol{\beta} + (X'X)^{-1}X'X_2\boldsymbol{\beta}_2 = \boldsymbol{\beta} + A\boldsymbol{\beta}_2, \tag{16}$$

where A is known as the alias matrix. The alias matrix tells us how the estimates of the constant and the first-order effects are biased by the extra terms that are included in the full model but are not in our simple approximation. We can learn about the effectiveness of the design by examining the entries of the alias matrix. Box and Draper (1959) initiated the use of the alias matrix as a guide to selecting a design and showed that the bias above is eliminated by a design whose moments equal the moments of the regression functions themselves with respect to a constant weight function. In that article and a sequel (Box and Draper, 1963) they derived response surface designs for minimizing mean squared error when the tentative model is a first- or second-degree polynomial and the extra terms are those one degree higher.

4.3. Bias as posterior variance

Our approach follows that of Draper and Guttman (1992), who suggested a clever method for turning the bias from terms not included in the model into variance, thereby making it possible to use design criteria based on the variance. Draper and Guttman (1992) assumed that $\boldsymbol{\beta}_2$ is random and has a normal distribution with mean 0 and covariance matrix $\sigma^2_\beta \boldsymbol{I}$. With these assumptions, and using the lack of random error in computer experiments, it can be easily shown that

$$\mathrm{Var}(\hat{\beta}) = \sigma^2_\beta AA', \tag{17}$$

where A is again the alias matrix. The matrix AA' thus measures the extent to which the design allows higher-order bias to affect the simple approximation.

It is now possible to apply standard variance-based design criteria and we suggest using here the A-optimality criterion for $\mathrm{Var}(\hat{\boldsymbol{\beta}})$,

$$A = \mathrm{tr}(\mathrm{Var}(\hat{\boldsymbol{\beta}})). \tag{18}$$

Trivially $\mathrm{tr}(\mathrm{Var}(\hat{\boldsymbol{\beta}})) = \sum_{i,j} a^2_{i,j}$, so we are led to the alias sum of squares criterion

$$A = \sum_{i,j} a^2_{ij}, \tag{19}$$

where a_{ij} denotes the components of the alias matrix. Mitchell (1974) proposed a similar criterion, which he called a "confounding index", for assessing the degree to which first-order terms in a regression model are biased by all possible two-factor interactions.

The Bayesian regression model here is actually a special case of the random field regression model described in Section 2. Assigning normal prior distributions to the coefficients in $\boldsymbol{\beta}_2$ makes the extra terms in Eq. (13) into a random field. See Steinberg and Bursztyn (2004) for details.

4.4. *Defining the model components*

The alias sum of squares criterion can easily be extended to include a larger base model and additional monomials, or other regression functions, among the "extra terms" in Eq. (13). The choice of terms for these models is a design decision that should depend on both the anticipated functional dependence of the output on the input factors and the size of the design. If linear effects are expected to dominate, then a linear base model with extra terms up to third-degree monomials and a moderate runs-to-factors ratio should be appropriate. If strongly nonlinear effects are anticipated, then it will be desirable to increase the number of runs and to expand the base model and the extra terms. We have used two rough rules of thumb in setting the number of runs and the models. First, the sample size must be large enough so that the base model is not singular. Second, the sample size must be small enough that no design meets the zero bias conditions derived by Box and Draper (1959); otherwise designs with repeated points could be rated as highly efficient. In defining the extra terms, we have typically chosen as extra terms monomials of at least two degrees higher than those in the base model. For modest design sizes (say up to 3–4 times as large as the number of factors), we believe that good results can be obtained by using a first-order model with extra terms up to third order. We have not examined in detail how the choice of base model and extra terms affects the designs and this could be a useful topic for further research.

Our requirement to normalize the higher-order regression functions is important in order to justify the assumption that the coefficients of the additional regression functions are of roughly the same order of magnitude, as reflected in their common variance. Alternatively, one might want to adopt a prior in which the variances decrease with increasing order of the regression function, as in Steinberg (1985). In that case, the normalization is again important to assure that the prior variances relate to coefficients of comparable regression functions.

4.5. *The A-criterion and accurate predictions*

In this section, we show that the *A*-criterion (in Eq. (19)) can also be derived using the idea of integrated mean squared error. Consider again the first-order model given in Eq. (12) and the "correct" model given in Eq. (13). The least-squares estimator is given in Eq. (16).

We would like to assess how the extra terms affect estimates of the response from the first-order model. The estimated response at an arbitrary input $\mathbf{x}$ is

$$\hat{y}(\mathbf{x}) = \boldsymbol{f}_1'(\mathbf{x})\hat{\boldsymbol{\beta}} = \boldsymbol{f}_1'(\mathbf{x})[\boldsymbol{\beta} + A\boldsymbol{\beta}_2] = \boldsymbol{f}_1'(\mathbf{x})\boldsymbol{\beta} + \boldsymbol{f}_1'(\mathbf{x})A\boldsymbol{\beta}_2.$$

Thus

$$y(\mathbf{x}) - \hat{y}(\mathbf{x}) = [\boldsymbol{f}_2'(\mathbf{x}) - \boldsymbol{f}_1'(\mathbf{x})A]\boldsymbol{\beta}_2 = \boldsymbol{v}'(\mathbf{x})\boldsymbol{\beta}_2.$$

In order to evaluate the model and the design we can look at the integrated mean squared error, again using $w(\mathbf{x})$ as a weight function.

$$\begin{aligned}\text{IMSE} &= \int [y(\mathbf{x}) - \hat{y}(\mathbf{x})]^2 \boldsymbol{w}(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &= \int \boldsymbol{\beta}_2'\boldsymbol{v}(\mathbf{x})\boldsymbol{v}'(\mathbf{x})\boldsymbol{\beta}_2\boldsymbol{w}(\mathbf{x})\,\mathrm{d}\mathbf{x} = \boldsymbol{\beta}_2'\left[\int \boldsymbol{v}(\mathbf{x})\boldsymbol{v}'(\mathbf{x})\boldsymbol{w}(\mathbf{x})\,\mathrm{d}\mathbf{x}\right]\boldsymbol{\beta}_2.\end{aligned}$$

Notice that

$$\boldsymbol{\beta}_2'\left[\int \boldsymbol{v}(\mathbf{x})\boldsymbol{v}'(\mathbf{x})\boldsymbol{w}(\mathbf{x})\,\mathrm{d}\mathbf{x}\right]\boldsymbol{\beta}_2 = \mathrm{tr}\left(\left[\int \boldsymbol{v}(\mathbf{x})\boldsymbol{v}'(\mathbf{x})\mathbf{w}(\mathbf{x})\,\mathrm{d}\mathbf{x}\right]\boldsymbol{\beta}_2\boldsymbol{\beta}_2'\right)$$

and

$$\begin{aligned}&\int \boldsymbol{v}(\mathbf{x})\boldsymbol{v}'(\mathbf{x})\boldsymbol{w}(\mathbf{x})\,\mathrm{d}\,\mathbf{x} \\ &\quad= \int (\boldsymbol{f}_2(\mathbf{x}) - \boldsymbol{A}'\boldsymbol{f}_1(\mathbf{x}))(\boldsymbol{f}_2'(\mathbf{x}) - \boldsymbol{f}_1'(\mathbf{x})\boldsymbol{A})\boldsymbol{w}(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &\quad= \int (\boldsymbol{f}_2(\mathbf{x})\boldsymbol{f}_2'(\mathbf{x}))\boldsymbol{w}(\mathbf{x})\,\mathrm{d}\mathbf{x} - \boldsymbol{A}'\int (\boldsymbol{f}_1(\mathbf{x})\boldsymbol{f}_2'(\mathbf{x}))\boldsymbol{w}(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &\qquad - \int (\boldsymbol{f}_2(\mathbf{x})\boldsymbol{f}_1'(\mathbf{x}))\boldsymbol{w}(\mathbf{x})\mathrm{d}\mathbf{x}\boldsymbol{A} + \boldsymbol{A}'\int (\boldsymbol{f}_1(\mathbf{x})\boldsymbol{f}_1'(\mathbf{x}))\boldsymbol{w}(\mathbf{x})\,\mathrm{d}\mathbf{x}\boldsymbol{A} \\ &\quad= I + A'A.\end{aligned}$$

The final equality results from our orthogonality requirement, which implies that

$$\int \boldsymbol{f}_2(\mathbf{x})\boldsymbol{f}_2'(\mathbf{x})\,\mathrm{d}\mathbf{x} = \boldsymbol{I}_u,$$

$$\int \boldsymbol{f}_1(\mathbf{x})\boldsymbol{f}_2'(\mathbf{x})\,\mathrm{d}\mathbf{x} = 0,$$

$$\int \boldsymbol{f}_1(\mathbf{x})\boldsymbol{f}_1'(\mathbf{x})\,\mathrm{d}\mathbf{x} = \boldsymbol{I}_r.$$

Again we find that the quality of the design is related to the matrix A.

5. Sample size and projections

5.1. Comparisons across sample size

A useful experimental design criterion should be able to compare designs of different sizes. The entropy and determinant based criteria for computer experiments have some problems in this regard, as does the minimum distance criterion.

We noted in Section 3 that the entropy criterion used is often to maximize $[\det(R)]^{1/n}$. For most random field models that have been proposed, the correlations are non-negative and, for $n \geqslant 2$, $(\det(R))^{1/n} < 1$. However, trivially $\det(R) = 1$ for a one-point design. The conclusion that adding data leads to a worse design is obviously unreasonable. The problem of comparing sample sizes also affects large designs. Typically, $\det(R)$ is a decreasing function of the sample size.

The sample size problem is not resolved by using the expected change in entropy, rather than the determinant, as the criterion. The expected change in entropy includes a direct reduction proportional to the sample size, $(n/2)[1 + \ln(2\pi)]$. However, this reduction is independent of the other parameters in the model and can be offset by the change in the determinant, again implying that smaller samples are preferable to larger ones. In particular, note that the second term in the change in entropy is proportional to σ^{2n} and so is highly sensitive to the assumed value of σ.

The minimum distance between any two design points is also a monotone decreasing function of the sample size. Thus a naive comparison of minimum distances for designs of differing sizes will be biased in favor of the smaller design. It is not clear how one might adjust the criterion to properly reflect sample size.

The alias sum of squares criterion can provide a direct measure for comparing designs of different sizes. The analysis in Section 4.5 showed that the criterion can be related to the ability to accurately predict the response function, averaged across the domain of the experimental factors. These prediction accuracies will typically improve as the sample size is increased. However, poor choice of design sites might actually increase the bias, indicating that a larger design is indeed inferior.

5.2. Replicate observations

In computer experiments, replicate observations at the same input values generate identical outputs. Most of the suggested designs for computer experiments, such as Latin hypercubes and lattice designs, simply avoid replicate input settings. Design criteria should also reflect the undesirability of replicates. The “effective” design for a computer experiment is really the design consisting of the distinct points with replicates discarded. One might naturally want to require that a design criterion produce the same assessment when restricted to the distinct sites.

The entropy criterion E_1 and the minimum distance criterion will equal 0 when there are replicates and so correctly label such designs as undesirable. However, these criteria make no distinction among designs with replicate points. A natural alternative is to evaluate these criteria and to fit the BLUP using only the distinct design sites. For design comparison, we are again faced with the sample size comparison problem discussed in Section 5.1.

The alias sum of squares criterion also does not rate designs on the basis of their distinct sites. In general, the assessment will depend on which sites have been repeated. There is one special case where the alias sum of squares criterion is identical for replicates and that occurs when an entire design is replicated an equal number of times. In that case, it is easy to show that the replication does not modify the design assessment.

5.3. Projections

Some of the input factors in a computer experiment may prove to have negligible effects on the outputs, a property often described as *factor sparsity*. In that case, the effective design is the projection of the original design onto the subset of *active factors*. Typically there is no advance knowledge as to which factors will be active, so designs should perform well under arbitrary projections. For example, projection can generate replication. We will see in the next section that the entropy and minimum distance criteria sometimes favor designs whose points are at the extremes of the input space and that do have replicate points under projection.

A simple way to include projection efficiency in comparing designs is to take a global criterion (like the determinant or the entropy) and to evaluate it for projections along with the full set of input factors. This approach still leaves some questions, such as how to summarize performance across the various projections and how much weight should be given to projections as opposed to the full factor design. There are also some computational issues, as the number of projections will be very large for experiments with many factors. Ideally, one would prefer a criterion that automatically produces good projected designs, without having to explicitly define and assess them.

6. Design comparisons

In this section we study the performance of five design criteria by examining their assessment of several candidate classes of designs for computer experiments. As one of the classes is the standard two-level fractional factorials, we consider only run sizes that are powers of 2. In all cases the design region is $[-1, 1]^p$, where p is the number of factors. The particular settings we examine are designs for (1) five factors in 16 runs, (2) nine factors in 32 runs and (3) 21 factors in 64 runs. We present assessments of both the full designs and of projections onto subsets of select sizes. For the five-factor designs, we average the design criteria over all possible projections onto three of the five factors. For the nine-factor designs, we average over all possible projections onto three or five factors. For the 21 factor designs, we average over projections onto five or eight factors. Due to the large number of projections in this case, we chose 30 random projections of each size.

6.1. Design criteria

The design criteria we compared are

1. The alias sum of squares criterion, when a first-order model is fitted and there are extra terms for all second-order effects, pure cubics and pure quartics.

2. The entropy criterion $\mathrm{Det} = [\det(R)]^{1/n}$. We considered four cases of the Sacks et al. (1989) covariance function: $\lambda = 0.05, 0.5$ and $\alpha = 1, 2$. The combinations are denoted by Det(1): $\lambda = 0.05$ and $\alpha = 2$, Det(2): $\lambda = 0.5$ and $\alpha = 2$, Det(3): $\lambda = 0.05$ and $\alpha = 1$, Det(4): $\lambda = 0.5$ and $\alpha = 1$.
3. The minimum distance criterion of Johnson et al. (1990), $\mathrm{Dist} = \{\min_{i<j} \|x_i - x_j\|\}$.
4. The integrated mean squared error criterion $\mathrm{IMSE} = \int E\{Y(\mathbf{x}) - \hat{Y}(\mathbf{x})\}^2 \, d\mathbf{x}$ for random field models. The same covariance functions as noted in point (2) for the entropy criterion were considered and results are denoted by $\mathrm{IMSE}(1) \ldots, \mathrm{IMSE}(4)$. We evaluated the IMSE by averaging the posterior variance at a random sample of 5000 points (for five-factor designs) or 15,000 points (for nine-factor designs). The IMSE criterion requires much more computer time than the other criteria. So we did not evaluate this criterion for the 21-factor designs or for projections of the smaller designs.

6.2. *Classes of designs*

We included the following classes of experimental designs in our comparison.

Latin Hypercube Designs (McKay et al., 1979): We generated Latin hypercube plans with the values for each factor chosen at random from within n equal width bins. LHC designs are easily computed and projections onto subsets of input factors are also LHC designs. LHC designs include random permutation of the values for each factor. All our results for LHC designs were computed by averaging over 100 realizations of the corresponding LHC design.

U-Designs (Owen, 1992; Tang, 1993): U-designs are a special class of Latin hypercube designs in the p-dimensional unit cube that use orthogonal arrays (OAs) (Hedayat et al., 2000) to obtain nearly uniform projections jointly for two or more input variables. These designs divide each factor axis into n bins, as in LHC designs, but also group the bins into a small number (say 2–5) of coarser bins. The OA works at the coarse level and guarantees equal numbers of design points in each coarse bin formed by crossing two or more factors. Within the coarse bins, the points are still spread out in accord with the rules for a LHC.

Lattice Designs (Fang et al., 1994, 2000): Lattice sets were motivated by the desire to find good sets of evaluation points for numerical computation of multi-dimensional integrals. Such sets can also be useful as factorial designs for computer experiments. For our comparisons, we used lattice designs that are generated as follows:

$$x_{ij} = \left\{ \frac{2ih_j - 1}{2n} \right\}, \quad i = 1, \ldots, n, \quad j = 1, \ldots, p,$$

where $\{x\}$ is the fractional part of x, $(n; h_1, \ldots, h_p)$ is a vector of integers satisfying $1 \leqslant h_j < n$, $h_i \neq h_j$ for $i \neq j$, and $p < n$, and $(h_1, \ldots, h_p) = (1, a, a^2, \ldots, a^{p-1})$, as suggested by Korobov (1959). For some run and factor sizes of interest here, we were not able to find a lattice design with this type of generating vector that is not singular for the first-order model. So we used 17 runs (with $a = 3$), 31 (with $a = 3$) and 61 runs (with $a = 2$).

Rotation designs (Beattie and Lin, 1997; Bursztyn and Steinberg, 2001, 2002): These designs are generated by rotating standard fractional factorial designs. Take an orthogonal starting design D and rotate it to obtain a new design matrix $D_R = DR$, where R is any

Table 1
Values of the design criteria for $n = 16$ and $p = 5$

Criteria	LH	LD	UD	RD3	RD4	FF
A	11.64	7.58	8.55	9.28	4.47	81.67
Det(1)	0.0376	0.0337	0.0384	0.0987	0.0329	0.2877
Det(2)	0.5900	0.6397	0.6142	0.8394	0.6386	0.9984
Det(3)	0.1478	0.1491	0.1648	0.1943	0.1543	0.1197
Det(4)	0.8204	0.8329	0.8440	0.9133	0.8412	0.9372
Dist	0.6403	1.0189	0.7416	0.7997	1.0593	2.8284
IMSE(1)	0.0068	0.0067	0.0072	0.0077	0.0075	0.0313
IMSE(2)	0.360	0.324	0.360	0.369	0.324	0.749
IMSE(3)	0.0682	0.0630	0.0600	0.0666	0.0648	0.157
IMSE(4)	0.659	0.638	0.644	0.664	0.641	0.881

The experimental designs are (1) Latin hypercube—LH (2) Lattice design—LD (3) U-design—UD (4) Rotation design with rotations of three factors using two angles 25° and 80°—RD3 (5) Rotation design with rotations of four factors using two angles 25° and 80°—RD4 (6) Fractional factorial design—FF.

Table 2
Values of the design criteria for $n = 32$ and $p = 9$

Criteria	LH	LD	UD	RD3	RD4	RD4	FF
					(1)	(2)	
A	24.73	15.03	16.37	17.65	9.29	7.41	147.00
Det(1)	0.0791	0.0799	0.0798	0.0935	0.0655	0.0721	0.4069
Det(2)	0.8483	0.8882	0.8691	0.8920	0.8480	0.8891	0.9998
Det(3)	0.2096	0.2299	0.2326	0.1879	0.2014	0.2099	0.1648
Det(4)	0.9524	0.9644	0.9633	0.9478	0.9453	0.9572	0.9859
Dist	1.0196	1.4267	0.9553	0.9352	0.8519	1.3589	2.8284
IMSE(1)	0.0282	0.0271	0.0275	0.0282	0.0261	0.0259	0.0942
IMSE(2)	0.732	0.725	0.723	0.739	0.701	0.695	0.996
IMSE(3)	0.125	0.119	0.111	0.127	0.126	0.122	0.268
IMSE(4)	0.909	0.905	0.902	0.913	0.910	0.905	1.000

The experimental designs are (1) Latin hypercube—LH (2) Lattice design—LD (3) U-design—UD (4) Rotation design with rotations of three factors using four angles 25°, 55°, 80°, 125°—RD3 (5) Rotation design with rotations of four factors—RD4. RD4(1) is using two angles 25° and 80° and RD4(2) six angles 15°, 25°, 35°, 45°, 55°, 65° (6) Fractional factorial design—FF.

orthonormal matrix. Then scale the points in D_R so that all lie in $[-1, 1]^p$. One can choose R so that each factor has many levels which makes it possible to detect and to estimate many possible higher-order effects. For the rotation designs we used Bursztyn and Steinberg's (2002) method for rotations of three factors and Bursztyn and Steinberg's (2001) method for rotations of four factors. In both cases we used two rotation angles, 25° and 80°.

6.3. *Design comparison*

The results of our assessments are presented in Tables 1–6. We summarize them below.

1. The alias sum of squares criterion tends to favor the rotation designs and the lattice designs. The good performance of these designs holds both for the full design and for the

Table 3
Values of the design criteria for $n = 64$ and $p = 21$

Criteria	LH	LD	UD	RD3	RD4	FF
A	135.36	49.58	112.39	80.99	31.36	343.00
Det(1)	0.2893	0.2932	0.2679	0.3162	0.1628	0.8440
Det(2)	0.9982	0.9987	0.9944	0.9883	0.9846	1.0000
Det(3)	0.4342	0.4450	0.4643	0.4051	0.3643	0.5176
Det(4)	0.9998	0.9999	0.9999	0.9990	0.9987	1.0000
Dist	2.0463	2.4732	1.8256	1.3899	1.9881	4.8990

The experimental designs are (1) Latin hypercube—LH (2) Lattice design—LD (3) U-design—UD (4) Rotation design with rotations of three factors with four angles 20°, 55°, 80°, 125°—RD3 (5) Rotation design with rotations of four factors using two angles 25° and 80°—RD4 (6) Fractional factorial design—FF.

Table 4

Criteria	LH	LD	UD	RD3	RD4	FF
(a) *Values of the design criteria for projections of three factors out of five, for a* 16 *run computer experiment*						
A	2.59	1.92	1.64	2.62	1.59	49.00
Det(1)	0.0032	0.0018	0.0041	0.0028	0.0030	0
Det(2)	0.1912	0.1782	0.2384	0.1475	0.2041	0
Det(3)	0.0761	0.0748	0.0815	0.0452	0.0774	0
Det(4)	0.5604	0.5639	0.5946	0.3423	0.5652	0
Dist	0.3193	0.6264	0.4358	0.1489	0.5205	0
(b) *Values of the design criteria for projections of four factors out of five, for a* 16 *run computer experiment*						
A	5.86	4.18	4.17	5.35	2.85	65.33
Det(1)	0.0161	0.0121	0.0187	0.0191	0.0131	0.1087
Det(2)	0.4118	0.4272	0.4622	0.4136	0.4305	0.9637
Det(3)	0.1118	0.1112	0.1224	0.0891	0.1149	0.0329
Det(4)	0.7148	0.7244	0.7473	0.5834	0.7289	0.7476
Dist	0.4901	0.8169	0.6291	0.3976	0.8007	2.0000

projected design and for all the design sizes. The Latin hypercube designs were rated as somewhat less successful for the full designs but were comparable to the best designs for all the projections. The two-level fractional factorials were clearly inferior in all cases studied.

2. The Integrated mean squared error criterion gives the two-level fractional factorials consistently bad performance ratings. The other designs are given similar ratings. The best design depends on the particular covariance model, but there is not much difference among the non-factorial designs.

3. The entropy criterion typically favors the 2^{k-p} designs when all factors are considered. When $\alpha = 2$ and λ is small, there is a sharp preference for the 2^{k-p} designs. When λ is large, the correlations fall off quickly and all the designs have determinants close to 1. The only full designs in which the 2^{k-p} is not the winner is when $\alpha = 1$ and λ is small. For the five-factor, 16 run and the nine-factor, 32 run settings, the two-level design is slightly worse than all the others (which are comparable), but with 21 factors and 64 runs, the two-level

Table 5

Criteria	LH	LD	UD	RD3	RD4	RD4	FF
					(1)	(2)	
(a) *Values of the design criteria for projections of three factors out of nine, for a* 32 *run computer experiment*							
A	1.05	0.92	0.58	1.19	1.52	1.02	49.00
Det(1)	0.0001	0.0000	0.0001	0.0000	0.0000	0.0000	0
Det(2)	0.0473	0.0509	0.0587	0.0120	0.0329	0.0405	0
Det(3)	0.0427	0.0453	0.0442	0.0135	0.0369	0.0389	0
Det(4)	0.4166	0.4416	0.4387	0.1449	0.3723	0.3940	0
Dist	0.1769	0.4889	0.2407	0.0260	0.1699	0.1942	0
(b) *Values of the design criteria for projections of five factors out of nine, for a* 32 *run computer experiment*							
A	4.38	2.72	3.37	4.14	3.41	2.50	81.67
Det(1)	0.0081	0.0097	0.0079	0.0071	0.0065	0.0075	0.0376
Det(2)	0.3844	0.4373	0.4259	0.3432	0.3311	0.3768	0.5759
Det(3)	0.0943	0.1002	0.1012	0.0682	0.0857	0.0899	0.0084
Det(4)	0.7195	0.7491	0.7482	0.6027	0.6829	0.7095	0.4193
Dist	0.4586	0.5877	0.8645	0.2044	0.4663	0.5735	1.2063

Table 6

Criteria	LH	LD	UD	RD3	RD4	FF
(a) *Values of the design criteria for projections of five factors out of* 21*, for a* 64 *run computer experiment*						
A	2.07	1.86	1.42	1.89	3.39	81.67
Det(1)	0.0011	0.0009	0.0013	0.0007	0.0003	0
Det(2)	0.2018	0.2147	0.2252	0.1534	0.0796	0
Det(3)	0.0596	0.0607	0.0614	0.0347	0.0472	0
Det(4)	0.6130	0.6298	0.6328	0.4831	0.5068	0
Dist	0.3228	0.7067	0.3981	0.0952	0.2742	0
(b) *Values of the design criteria for projections of eight factors out of* 21*, for a* 64 *run computer experiment*						
A	6.94	5.59	5.82	6.56	6.74	130.67
Det(1)	0.0201	0.0192	0.0213	0.0199	0.0072	0.1491
Det(2)	0.6448	0.6712	0.6718	0.6325	0.3950	0.9550
Det(3)	0.1267	0.1276	0.1326	0.0938	0.1006	0.0436
Det(4)	0.8759	0.8842	0.8906	0.8312	0.7844	0.8456
Dist	0.6864	1.1185	0.8536	0.4745	0.5643	2.0438

fractional factorial again has the highest determinant. Similar results hold for the projected designs when there are enough factors that the 2^{k-p} projections do not have replicates. When there are replicates, the 2^{k-p} projections get 0 ratings and the best design is usually the U-design.

4. The minimum distance criterion also favors the 2^{k-p} designs when all factors are considered. The lattice designs are at or near the top among the other classes, but are a distant second to the two-level factorials. For the projected designs, as with the entropy

criterion, the 2^{k-p} projections continue to get the best ratings when there are no replicates, but are ruled out when there are replicates. In those cases, the lattice design is typically rated as much better than the others.

5. It is not surprising that the entropy and minimum distance criteria favor the two-level factorials for the settings that we have examined. The determinant in the entropy criterion is bounded from above by 1 and would achieve that bound if the responses at the design points were all independent (i.e. had 0 correlation). In general, spreading the points far apart decreases all the correlations, so the entropy criterion and the minimum distance criterion will both be improved, in general, by pushing the design points to the extremes of the design space.

6. There is good overall agreement between the alias sum of squares and IMSE criteria with respect to the best and worst designs in our tables. We also examined how closely these criteria agreed with one another for the 100 randomly generated LHC designs. Although these designs cover a much narrower range of criteria values than the designs in our tables, we found high correlations (0.65–0.85) between the *A*- and IMSE-criteria values. These correlations are similar to those found between values of the IMSE-criteria for different covariance function parameters. For the parameter values and designs that we studied, the correlations among the IMSE-criteria ranged from 0.60 to 0.88.

7. Discussion

The choice of experimental points is an important issue in planning an efficient computer experiment. Various authors have suggested intuitive goals for good designs, including "good coverage", ability to fit complex models, many levels for each factor, and good projection properties. At the same time, a number of different mathematical criteria have been put forth for comparing designs. We have proposed here a new criterion, based on the alias matrix for a simple model, and compared it to three criteria that have been advocated by other researchers, the integrated mean squared error, the design entropy and minimum distance criteria. We have compared these criteria with respect to ability to guide the choice of sample size and to produce good designs with respect to lack of replication and projective properties. We also compared several classes of designs using the criteria..

The results of our comparison show that the entropy and minimum distance criteria both tend to favor regular fractional factorial designs when the factor space is large. Fractional factorials do not match up well to the intuitive goals listed above. All the points are in corners of the design space, which does not correspond well to notions of good coverage. Each factor has only two levels so that nonlinear dependence on a single factor cannot be modeled. Projections onto small subsets typically contain replicate points, which have no value in computer experiments. The reason that fractional factorials do well is that the criteria favor designs in which the points are distant from one another. With a small number of factors, inter-point distances can only be made large by spreading out the design points in the interior of the region. However, a high-dimensional design space provides a large number of distant corners and inter-point distances are maximized by designs with all the points in those corners. With the entropy criterion, this phenomenon is especially strong when the exponent in the correlation function is 2. We have seen that, to some extent, this problem can be overcome by considering projections, and not just the full design space, in assessing design options. Certainly that should be a part of any design inquiry.

The entropy and mean squared error criteria discussed in this paper are motivated by a random field regression model whereas our *A*-criterion relies on a simple low-degree polynomial approximation, with some of the terms part of a fitted model and others considered as bias. Although these two modeling approaches appear quite different from one another, they actually have much in common. Steinberg and Bursztyn (2004) show that the random field model can be viewed as a Bayesian regression model with proper priors assigned to all but a small subset of the coefficients. The "extra terms" in our Eq. (13) then form the basis for the random field.

We believe that the simple criterion that we have proposed, based on the alias matrix relative to a first-order model, is a simple and effective solution for assessing and comparing different designs. In our study, it showed good consistency between how full designs were rated and how their projections were rated. The criterion gave consistently low marks to the fractional factorial. The criterion favored designs that have little confounding between high- and first-order terms. Such designs will also be the most able to accommodate high-order terms in the predictor. We found good overall agreement between the *A*-criterion and the IMSE-criterion, both with respect to the widely differing designs in our tables and with respect to the much narrower range of criterion values in our randomly generated LHC designs. The *A*-criterion has a major computational advantage over IMSE and this may be especially important in attempting to optimize within a class of designs.

References

Bates, R.A., Buck, R.J., Riccomagno, E., Wynn, H.P., 1996. Experimental design and observation for large systems. J. Roy. Statist. Soc. B 58, 77–94 Discussion 95–111.

Beattie, S.D., Lin, D.K.J., 1997. Rotated factorial design for computer experiments. Proceedings of SPES, American Statistical Association.

Box, G.E.P., Draper, N.R., 1959. A basis for the selection of a response surface design. J. Amer. Statist. Assoc. 54, 622–654.

Box, G.E.P., Draper, N.R., 1963. The choice of a second order rotatable design. Biometrika 50, 335–352.

Bursztyn, D., Steinberg, D.M., 2001. Rotation designs for experiments in high bias situations. J. Statist. Plann. Inference 97, 399–414.

Bursztyn, D., Steinberg, D.M., 2002. Rotation designs: orthogonal first-order designs with higher order projectivity. J. Appl. Stochastic Models Bus. Ind. 18, 197–206.

Currin, C., Mitchell, T., Morris, M., Ylvisaker, D., 1988. A Bayesian approach to the design and analysis of computer experiments. ORNL-6498, available from National Technical Information Service.

Currin, C., Mitchell, T., Morris, M., Ylvisaker, D., 1991. Bayesian prediction of deterministic functions, with applications to the design and analysis of computer experiments. J. Amer. Statist. Assoc. 86, 953–963.

Draper, N.R., Guttman, I., 1992. Treating bias as variance for experimental design purposes. Ann. Inst. Statist. Math. 44, 659–671.

Fang, K., Wang, Y., Bentler, P.M., 1994. Some applications of number-theoretic methods in statistics. Statist. Sci. 9, 416–428.

Fang, K.T., Lin, D.K.J., Winker, P., Zhang, Y., 2000. Uniform design: theory and application. Technometrics 42, 237–248.

Hedayat, A.S., Sloane, N.A.S., Stufken, J., 2000. Orthogonal Arrays, Springer, Berlin.

Johnson, M.E., Moore, I.M., Ylvisaker, D., 1990. Minimax and maximin distance designs. J. Statist. Plann. Inference 26, 131–148.

Kennedy, M.C., O'Hagan, A., 2001. Bayesian calibration of computer models. J. Roy. Statist. Soc. B 63, 425–450 Discussion 450–464.

Koehler, J.R., Owen, A.B., 1996. Computer experiments. In: Ghosh, S., Rao, C.R. (Eds.), Handbook of Statistics. Elsevier, Amsterdam, pp. 261–308.

Korobov, N.M., 1959. The approximate computation of multiple integrals. Dokl. Akad. Nauk SSSR 124, 1207–1210.

Lindley, D.V., 1956. On a measure of the information provided by an experiment. Ann. Math. Statist. 27, 986–1005.

McKay, M.D., Beckman, R.J., Conover, W.J., 1979. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21, 239–245.

Mitchell, T.J., 1974. Computer construction of D-optimal first-order designs. Technometrics 16, 211–220.

Morris, M.D., Mitchell, T.J., Ylvisaker, D., 1993. Bayesian design and analysis of computer experiments: use of derivatives in surface prediction. Technometrics 35, 243–255.

Owen, A.B., 1992. Orthogonal arrays for computer experiments, integration and visualization. Statist. Sinica 2, 439–452.

Robinson, G.K., 1991. That BLUP is a good thing: the estimation of random effects. Statist. Sci. 6, 15–51.

Sacks, J., Ylvisaker, D., 1966. Designs for regression problem with correlated errors. Ann. Math. Statist. 37, 66–89.

Sacks, J., Ylvisaker, D., 1968. Designs for regression problem with correlated errors; Many parameters. Ann. Math. Statist. 39, 49–69.

Sacks, J., Ylvisaker, D., 1970. Designs for regression problem with correlated errors III. Ann. Math. Statist. 41, 2057–2074.

Sacks, J., Welch, W.J., Mitchell, T.J., Wynn, H.P., 1989. Design and analysis of computer experiments. Statist. Sci. 4, 409–435.

Shewry, M.C., Wynn, H.P., 1987. Maximum entropy sampling. J. Appl. Statist. 14, 165–170.

Steinberg, D.M., 1985. Model robust response surface designs: scaling two-level factorials. Biometrika 72, 513–526.

Steinberg, D.M., 1990. A Bayesian approach to flexible modeling of multivariable response functions. J. Multivariate Anal. 34, 157–172.

Steinberg, D.M., Bursztyn, D., 2004. Some data analytic tools for understanding random field regression models. Technometrics.

Tang, B., 1993. Orthogonal array-based Latin hypercubes. J. Amer. Statist. Assoc. 88, 1392–1397.

Welch, W.J., Buck, R.J., Sacks, J., Wynn, H.P., Mitchell, T.J., Morris, M.D., 1992. Screening, predicting, and computer experiments. Technometrics 34, 15–25.