A Bayesian Approach to the Design and Analysis of Fractionated Experiments

V. Roshan Joseph
School of Industrial and Systems Engineering
Georgia Institute of Technology
Atlanta, GA 30332-0205, USA
roshan@isye.gatech.edu

Abstract

Specifying a prior distribution for the large number of parameters in the statistical model is a critical step in a Bayesian approach to the design and analysis of experiments. This article shows that the prior distribution can be induced from a functional prior on the underlying transfer function. The functional prior requires the specification of only a few hyper-parameters and therefore, can be easily implemented in practice. The usefulness of the approach is demonstrated through the analysis of some experiments. The article also proposes a new class of design criteria and establishes their connections with the minimum aberration criterion.

KEY WORDS: Gaussian process; Minimum aberration criterion; Optimal designs; Prior information; Quality engineering.

Submitted to *Technometrics* for publication in
August, 2004

1. INTRODUCTION

Suppose an experimenter wishes to perform an experiment with p factors each at two levels. Then, there are a total of 2^p experiments, which can be prohibitively large to conduct due to the huge cost and time involved in the experiment even with a moderately large p, say 4 to 10. In fact, it is very common to have more than 10 factors in industrial experiments . In such cases, the only option one has is to perform a fraction of the full factorial experiments (for real examples, see, Taguchi 1987, Wu and Hamada 2000, and Montgomery 2004).

Using a 2^p full factorial experiment one can estimate, all the main effects, 2-factor interactions (2fi),..., p-factor interaction (pfi). There are a total of 2^p effects including the mean. Suppose the experimenter performed n experiments, where $n < 2^p$, then one can only estimate n aliases of the 2^p effects, at least in the frequentist sense. For example, consider a 2^{3-1}_{III} experiment with three factors A, B, and C. If the design generator is $C = AB$, then from this experiment, only the four aliases $I = ABC, A = BC, B = AC$, and $C = AB$ can be estimated. Now to make conclusions, some empirical principles such as the effect hierarchy principle (see, e.g, Hamada and Wu 1992) are invoked into the data analysis. The *effect hierarchy principle* states that lower order effects are more important than higher order effects; and the effects of the same order are equally important. Thus some lower order effects are estimated by assuming the higher order effects to be negligible. In the above example, we may thus estimate the intercept and the three main effects by assuming the 2fis and 3fi to be negligible. When ambiguity about the effects are still unresolved, follow-up experiments such as using a fold-over deign etc. are recommended in the literature (see, e.g, Meyer, Steinberg, and Box 1996). Note that the estimation is difficult because we have a higher number of parameters (2^p) to estimate than the size of the data (n). Although the above estimation problem can be avoided using Bayesian methods, the conclusions will depend on the prior distribution for the parameters as well as the information in the experiment. If the prior is incorrect or the information is insufficient, then the conclusions can go wrong. Nevertheless, the Bayesian approach gives a nice framework for analyzing fractionated experiments, the potential of which does not seem to have been fully exploited in the literature.

A review on Bayesian experimental design was given by Chaloner and Verdinelli (1995). One of the difficulties associated with the Bayesian approach is in the specification of the prior distribution. For example, consider a 10-factor experiment. Suppose we decided to use a multivariate normal distribution for the parameters, then we need to specify $2^{10} = 1024$ mean hyper-parameters and $1024 \times 1025/2 = 524800$ variance-covariance hyper-parameters. This is definitely a challenging task. Although some simplifying assumptions are made in the literature for practical implementation, they are mostly done in a subjective way. Interestingly, the approach in computer experiments is different and does not have the above problem. The idea was to use a functional prior on the underlying transfer function (see, e.g, Currin et al. 1991). But since interpolating functions are desirable in computer experiments, kriging models are preferred instead of the linear models. In this article we will show that by imposing such a functional prior on the transfer function, we can induce a prior distribution for the parameters in the linear model. The functional prior requires the specification of only a few hyper-parameters and therefore it avoids the difficulty of huge prior specification. This is the crucial idea behind this article. The use of linear models makes this work different from that of computer experiments. Mitchell, Morris, and Ylvisaker (1995) has studied stationary processes over the factorial points in a 2-level design (see also a follow-up work by Kerr 2001). Their work focuses on some of the properties of the prior processes and related design issues, whereas we use it as a tool to avoid the estimation problems in fractionated experiments. We also provide a general framework for developing the prior distribution in experiments with any number of levels. It is important to mention some of the other Bayesian work in the literature such as that of Box and Meyer (1993), Chipman, Hamada, and Wu (1997), and Chipman (1998). Apart from the prior specification for the model parameters, their work differs on another aspect. They use a hierarchical Bayesian model with a prior on the model space, which helps in identifying the best subset model that fits the data well. In this work, we do not use such a prior. Instead, a Bayesian version of the forward selection strategy is proposed to identify the best model. This strategy may not do an exhaustive search in the model space, but is computationally easier to implement.

An important problem for the experimenter is the designing of the experiment, i.e. choosing n runs from the set of 2^p runs. Finding the “best” possible design is a very complicated combinatorial optimization problem, but some simple techniques are proposed for designs such as 2^{p-k} fractional factorial designs. One can write down the wordlength pattern of a design (Wu and Hamada 2000) and use criteria such as maximum resolution (Box and Hunter 1961) or minimum aberration (Fries and Hunter 1980) for selecting the best design. Minimum aberration (MA) is probably the most popular criterion for design selection because of its many desirable properties. See, for examples, Chen (1992), Cheng and Mukerjee (1998) and Cheng, Steinberg, and Sun (1999). The above two criteria are strictly based on the hierarchy principle and there are many problems associated with it. For example, consider the two 2^{9-4}_{IV} designs given in Wu and Hamada (2000, Appendix 4A). Their word length patterns are $W(D_1) = (0,0,0,6,8,0,0,1,0)$ and $W(D_2) = (0,0,0,7,7,0,0,0,1)$. According to MA criterion D_1 is better because it has one lower 4-letter word than D_2. But note that D_2 has one less 5-letter word and no 8-letter word. Therefore should we prefer D_2 to D_1? According to another popular criterion known as maximum number of clear 2fis (MaxC2), D_2 is better than D_1 (Wu and Wu 2002). There is no general consensus among researchers on which criterion is the best. This article proposes a new class of criteria based on the Bayesian A-optimal criterion, a special case of which is equivalent to a weighted average of the word length pattern. This connection of Bayesian designs with MA designs will surely appeal to the frequentists. Moreover, the proposed criterion can also be used for evaluating any type of designs including nonregular designs (designs that do not have a defining contrast subgroup).

The article is organized as follows. In Section 2, a general methodology for developing a prior distribution for the model parameters is proposed. The estimation of the model parameters and the hyper-parameters is discussed in Section 3. A Bayesian version of the forward selection strategy is proposed in Section 4. The new approach is illustrated using two experiments in Section 5. The optimal design for the 2-level experiments is discussed in Section 6 and some concluding remarks and future research directions are given in Section 7.

2. PRIOR DISTRIBUTION

In this section we will propose a methodology for developing a prior distribution for the parameters in a linear model. The results will then be simplified for the case of two-level experiments.

2.1 General Methodology

Let Y be the response and $\{x_1, \cdots, x_p\}$ the set of factors. The factor x_i takes m_i values (levels) in the experiment and let $\mathcal{X}$ denotes the experimental region containing all the level combinations of the p factors. The experimental design is a subset of $\mathcal{X}$. For example, if the levels of x_i are $1, 2, \cdots, m_i$, then the experimental design is a set of points in $\mathcal{X} = \{1, 2, \cdots, m_1\} \times \cdots \times \{1, 2, \cdots, m_p\}$. We assume that $Y = f(\boldsymbol{x}) + e$, where e is the error caused by the unobserved noise factors and measurement noise. Assume $e \sim N(0, \sigma^2)$. The true transfer function f is unknown to the experimenter and can be highly nonlinear. We will put a prior on this function. Let f be a realization from a Gaussian process (GP) with mean μ_0 and covariance function $\sigma_0^2 \psi$. The covariance function is defined as $cov\{f(\boldsymbol{x}_1), f(\boldsymbol{x}_2)\} = \sigma_0^2 \psi(\boldsymbol{x}_1, \boldsymbol{x}_2)$, where $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are any two points in the experimental region. The covariance function must be a positive semi-definite function with $\psi(\boldsymbol{x}, \boldsymbol{x}) = 1$. Such functional priors using GPs are widely used in the modeling of deterministic functions (see, e.g, Santner, Williams, and Notz 2003). Thus we have the model

$$Y = f(\boldsymbol{x}) + e, e \sim N(0, \sigma^2), f(\boldsymbol{x}) \sim GP(\mu_0, \sigma_0^2 \psi). \tag{1}$$

It is possible to use a more elaborate mean function than a constant μ_0. This will be introduced in a later section to facilitate variable selection.

Product correlation functions are by far the most commonly used type of correlation functions (Currin et al. 1991). It is given by

$$\psi(\boldsymbol{x}_1, \boldsymbol{x}_2) = \prod_{i=1}^{p} \psi_i(x_{1i}, x_{2i}), \tag{2}$$

where $\boldsymbol{x}_j = (x_{j1}, \cdots, x_{jp})'$ for $j = 1, 2$. We assume $\psi_i(x_{1i}, x_{2i}) \in [0, 1]$ for all i, which is satisfied for all the correlation functions used in computer experiments. The most popular choice is the exponential correlation function given by $\psi_i(x_{1i}, x_{2i}) = \exp(-\theta_i|x_{1i} - x_{2i}|^{\alpha_i})$, where $0 < \alpha_i \leq 2$ and $0 < \theta_i < \infty$ for all i. This correlation function is appropriate for a quantitative factor. Qualitative factors such as different types of machines, methods, etc. also arise in experiments. It is not meaningful to consider the transfer function to be a realization from a continuous stochastic process in the space of the qualitative factors. But fortunately we only need its distribution over some points in $\mathcal{X}$ and therefore discrete factors can also be included into the above setup. For qualitative factors we may have to assign equal correlation between any two levels, whereas for the quantitative factors it can be a decreasing function of the distance between the levels. Note that the existing factorial design theory treats all the factors as qualitative, which is one of its drawbacks (Cheng and Ye 2004).

We will approximate $f(\boldsymbol{x})$ with a linear model. Each variable x_i can be represented using $m_i - 1$ dummy variables. Let $z_1, \cdots, z_{m_1-1}$ be the dummy variables for x_1; $z_{m_1}, \cdots, z_{m_1+m_2-2}$ the dummy variables for x_2; $\cdots$; and $z_{m_1+\cdots+m_{p-1}-p+2}, \cdots, z_{m_1+\cdots+m_p-p}$ the dummy variables for x_p. There are many coding systems that are popular in regression analysis for defining the dummy variables such as orthogonal polynomial coding, Helmert coding, treatment coding, etc. We may select a coding system to get a nice interpretation for the parameters. Now define the variables in the full linear model as follows. Let $u_0 = 1, u_1 = z_1, \cdots, u_{m_1+\cdots+m_p-p+1} = z_1 z_{m_1}, \cdots, u_{q-1} = z_{m_1} \cdots z_{m_1+\cdots+m_p-p}$, where $q = m_1 m_2 \cdots m_p$. Note that we do not need the interactions among the dummy variables within an x variable. For example, consider an experiment with two factors each at three levels denoted by 1, 2, and 3. Using orthogonal polynomial coding (Wu and Hamada 2000, chapter 2)

$$z_1 = x_1 - 2, z_2 = 3(x_1 - 2)^2 - 2, z_3 = x_2 - 2, \text{ and } z_4 = 3(x_2 - 2)^2 - 2.$$

Now $u_1 = z_1, u_2 = z_2, u_3 = z_3, u_4 = z_4, u_5 = z_1 z_3, u_6 = z_1 z_4, u_7 = z_2 z_3$, and $u_8 = z_2 z_4$. These eight variables correspond to the linear main effect of x_1, quadratic main effect of x_1, $\cdots$, quadratic × quadratic interaction effect of x_1 and x_2 respectively.

Let $f(\boldsymbol{x}) = \mu_0 + \sum_{i=0}^{q-1} \beta_i u_i + \delta(\boldsymbol{x})$. Because the f values can be exactly reproduced by

$\mu_0 + \sum_{i=0}^{q-1} \beta_i u_i$ in $\mathcal{X}$, set $\delta(\boldsymbol{x}) = 0$ for $\boldsymbol{x} \in \mathcal{X}$. Thus we have

$$f(\boldsymbol{x}) = \mu_0 + \sum_{i=0}^{q-1} \beta_i u_i, \ \boldsymbol{x} \in \mathcal{X}. \tag{3}$$

A full factorial design consists of all the points in $\mathcal{X}$. Based on the values taken by each factor in the full factorial design, we can compute the values of $u_0, u_1, \cdots, u_{q-1}$. Denote this $q \times q$ matrix by U_p. We have, $\boldsymbol{f} = \mu_0 \mathbf{1} + U_p \boldsymbol{\beta}$, where $\boldsymbol{f}$ denotes the vector of function values for the full factorial design, $\boldsymbol{\beta} = (\beta_0, \cdots, \beta_{q-1})'$, and $\mathbf{1}$ is a column of 1's. Note that $\boldsymbol{f}$ has a multivariate normal distribution with $E(\boldsymbol{f}) = \mu_0 \mathbf{1}$ and $var(\boldsymbol{f}) = \sigma_0^2 \Psi_p$, where Ψ_p is the correlation matrix for the points in the full factorial design. Using this we can induce a prior distribution for all the q parameters in the model. We obtain $E(\boldsymbol{\beta}) = E\{U_p^{-1}(\boldsymbol{f} - \mu_0 \mathbf{1})\} = \mathbf{0}$ and $var(\boldsymbol{\beta}) = var\{U_p^{-1}(\boldsymbol{f} - \mu_0 \mathbf{1})\} = \sigma_0^2 U_p^{-1} \Psi_p (U_p^{-1})'$, where $\mathbf{0}$ is a column of 0's. In addition, $\boldsymbol{\beta}$ has a multivariate normal distribution. Thus we have the following result.

Theorem 1: *Under (1) and (3),*

$$\boldsymbol{\beta} \sim N\left(\mathbf{0},\ \sigma_0^2 U_p^{-1} \Psi_p (U_p^{-1})'\right).$$

The above prior is different from that used in the literature on Bayesian linear models (see, e.g, Chipman, George, and McCulloch 2001). In the literature, the variance-covariance matrix is given by $\tau^2 \Sigma$, where τ^2 is a constant and Σ is usually taken as either I_q or $(U_D' U_D)^{-1}$. Here I_q is the q-dimensional identity matrix and U_D is the model matrix corresponding to the experimental design. The second choice for Σ can not be used in our problem because in fractionated experiments the number of rows in U_D is less than q and therefore $(U_D' U_D)^{-1}$ does not exist. The first choice of I_q assigns equal variances to all effects, which may not always be good as we show in the next section.

The result of Theorem 1 is general and can be used for any type of factors with any number of levels. But as the number of factors and/or levels increase, the variance-covariance matrix becomes huge and difficult to handle. Therefore some simplifications are essential for its practical implementation. We will now simplify the result for the case of factors experimented at two levels.

2.2 Two-Level Experiments

Consider the case of a 2-level experiment where the two levels are coded as -1 and $+1$. The rest of the article concentrates on 2-level experiments. Details of three and higher level experiments will appear elsewhere. Here $\mathcal{X} = \{-1, 1\}^p$. For two levels, we do not need any dummy variables. Therefore, let $u_0 = 1$, $u_1 = x_1$, $\cdots$, $u_p = x_p$, $u_{p+1} = x_1 x_2$, $\cdots$, and $u_{2^p-1} = x_1 x_2 \cdots x_p$. Thus the linear model we would like to fit is $Y = \mu_0 + \sum_{i=0}^{2^p-1} \beta_i u_i + e$, where $\beta_1, \cdots, \beta_p$ are the main effects, $\beta_{p+1}, \cdots, \beta_{p+\binom{p}{2}}$ the 2fis, $\cdots$, and β_{2^p-1} the pfi. We will obtain a prior distribution for these 2^p parameters using Theorem 1. Assume the prior process in (1) to be stationary for all p. Then, for the product correlation function in (2), $\psi_i(x_{1i}, x_{2i})$ depends only on $|x_{1i} - x_{2i}|$. Also, in most practical cases we will take all the ψ_i's to be the same, say ψ_0. Therefore, let $\psi_i(x_{1i}, x_{2i}) = \psi_0(|x_{1i} - x_{2i}|)$. Thus, consider the correlation function of the form

$$\psi(\boldsymbol{x}_1, \boldsymbol{x}_2) = \prod_{i=1}^{p} \psi_0(|x_{1i} - x_{2i}|). \tag{4}$$

Let

$$r = \frac{1 - \psi_0(2)}{1 + \psi_0(2)} \text{ and } \tau^2 = \frac{\sigma_0^2}{(1 + r)^p}.$$

Then, we have the following result. The proof is given in the Appendix.

Proposition 1. *For 2-level experiments with the correlation function in (4),*

$$\begin{aligned}
\beta_0 &\sim N(0, \tau^2) \\
\beta_i &\sim N(0, \tau^2 r), \quad i = 1, \cdots, p \\
\beta_i &\sim N(0, \tau^2 r^2), \quad i = p+1, \cdots, p + \tbinom{p}{2} \\
&\vdots \\
\beta_{2^p-1} &\sim N(0, \tau^2 r^p),
\end{aligned}$$

and they are independent.

Note that the main effects, 2fis, $\cdots$, pfi are independent and their variances decrease geometrically at a rate r. The above result can be deduced from a more general result obtained by Mitchell, Morris, and Ylvisaker (1995), but unlike here their motivation was not to develop a prior distribution to facilitate Bayesian estimation in fractionated experiments. Moreover, the description and proof given here based on linear model theory are more transparent and straightforward. The importance of this result is that we need to specify only two hyper-parameters τ^2 and r for obtaining the prior distribution of 2^p parameters. The result also gives a justification for the effect hierarchy principle, which is fundamental to the frequentist analysis. Since $0 \leq r \leq 1$, the variance of a lower order effect is more than that of a higher order effect, and therefore it is more probable to be of larger magnitude than the higher order effect, thus justifying the hierarchy principle.

The result in Proposition 1 can be generalized using the more general product correlation function given in (2). Under the stationary assumption, the only change will be in the variance structure. Let $\psi_i(x_{1i}, x_{2i}) = \psi_i(|x_{1i} - x_{2i}|)$, $r_i = \{1 - \psi_i(2)\}/\{1 + \psi_i(2)\}$ for $i = 1, \cdots, p$, and $\tau^2 = \sigma_0^2\{\prod_{i=1}^p (1 + r_i)\}^{-1}$. Let β_{ij} denote the 2fi between factors i and j, $\cdots$, and $\beta_{12\cdots p}$ the pfi. Then,

$$var(\beta_0) = \tau^2, \;\; var(\beta_i) = \tau^2 r_i, \;\; var(\beta_{ij}) = \tau^2 r_i r_j, \;\; \cdots, var(\beta_{12\cdots p}) = \tau^2 \prod_{k=1}^{p} r_k.$$

This result is useful if we know some effects are more important than others. If a particular main effect is considered to be more important, say for the i^{th} factor, then it should be assigned a higher value of r_i relative to the others. Note that this will make all the interaction effects involving this factor to be more important. This can be considered a weak justification to the effect heredity principle proposed in Hamada and Wu (1992) and used by Chipman, Hamada, and Wu (1997) (Effect heredity principle states that the chances of an interaction effect being important is less if none of its parent effects are important). For simplicity, the rest of the article focuses on the case of equal r_i's, which can easily be extended to deal with unequal r_i's.

3. ESTIMATION

Let D be the design matrix, which has n rows and p columns corresponding to the p experimental factors. Let $\boldsymbol{y} = (y_1, \cdots, y_n)'$ be the response values obtained from the experiment. For the moment, assume that we have an unreplicated experiment. Therefore the rows of D are unique. We want to fit the model $Y = \mu_0 + \boldsymbol{u}'\boldsymbol{\beta} + e$, where $e \sim N(0, \sigma^2)$, $\boldsymbol{u} = (u_0, u_1, \cdots, u_{2^p-1})'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_{2^p-1})'$. Assume that the e's are independent and σ^2 is known. Let U_D be the model matrix generated from D, i.e., U_D is an $n \times 2^p$ matrix in which the rows correspond to the rows of D and columns correspond to the u variables. Using Proposition 1, we have $E(\boldsymbol{\beta}) = \mathbf{0}$ and $var(\boldsymbol{\beta}) = \tau^2 R$, where $R = diag(1, r, \cdots, r^p)$. Thus we have the following Bayesian model

$$\boldsymbol{y}|\boldsymbol{\beta} \sim N(\mu_0\mathbf{1} + U_D\boldsymbol{\beta}, \sigma^2 I_n) \text{ and } \boldsymbol{\beta} \sim N(\mathbf{0}, \tau^2 R),$$

where I_n is the identity matrix. The posterior distribution of $\boldsymbol{\beta}$ given the data is

$$\boldsymbol{\beta}|\boldsymbol{y} \sim N\left(RU_D'(U_DRU_D' + \frac{\sigma^2}{\tau^2}I_n)^{-1}(\boldsymbol{y} - \mu_0\mathbf{1}), \tau^2 R - \tau^2 RU_D'(U_DRU_D' + \frac{\sigma^2}{\tau^2}I_n)^{-1}U_DR\right). \tag{5}$$

The above form of the posterior is different from the usual one used in the literature (see equation (19)). The reason for preferring the above form is explained in the Appendix. Thus the estimates (posterior means) for the 2^p parameters are given by

$$\hat{\boldsymbol{\beta}} = RU_D'(U_DRU_D' + \frac{\sigma^2}{\tau^2}I_n)^{-1}(\boldsymbol{y} - \mu_0\mathbf{1}). \tag{6}$$

When p is very large, the matrices U_D and R become huge and difficult to handle. In such cases a computationally easier form can be obtained as follows. Since $\boldsymbol{f} = U_p\boldsymbol{\beta}$, we have $\boldsymbol{f}_D = U_D\boldsymbol{\beta}$, where $\boldsymbol{f}_D$ is the vector of function values corresponding to the design matrix D. Therefore $var(\boldsymbol{f}_D) = \sigma_0^2\Psi_D = \tau^2 U_DRU_D'$, where Ψ_D is the matrix of correlations corresponding to D obtained using (4). Therefore $U_DRU_D' = (1+r)^p\Psi_D$. Thus we obtain

$$\hat{\boldsymbol{\beta}} = \frac{1}{(1+r)^p}RU_D'(\Psi_D + \frac{\sigma^2}{\sigma_0^2}I_n)^{-1}(\boldsymbol{y} - \mu_0\mathbf{1}).$$

Let $\boldsymbol{\beta}_s$ be the parameters we are mainly interested (say the main effects and two-factor interactions) and U_s the corresponding model matrix. Then

$$\hat{\boldsymbol{\beta}}_s = \frac{1}{(1+r)^p}R_sU_s'(\Psi_D + \frac{\sigma^2}{\sigma_0^2}I_n)^{-1}(\boldsymbol{y} - \mu_0\mathbf{1}),$$

where R_s is obtained from R corresponding to the subset $\boldsymbol{\beta}_s$. The above expression does not contain any huge matrices and therefore can be easily computed. The correlation matrix can be directly constructed as follows. The ij^{th} element of Ψ_D is

$$(\Psi_D)_{ij} = \prod_{k=1}^{p} \psi_0(|x_{ik} - x_{jk}|) = \psi_0(2)^{h_{ij}} = \left(\frac{1-r}{1+r}\right)^{h_{ij}},$$

where $(x_{i1}, \cdots, x_{ip})$ and $(x_{j1}, \cdots, x_{jp})$ are the i^{th} and j^{th} rows of the design matrix D and $h_{ij} = \sum_{k=1}^{p} |x_{ik} - x_{jk}|/2$ is the number of times the values in these two rows differ (Note that $\psi_0(0) = 1$).

There are three hyper parameters μ_0, τ^2 (or σ_0^2), and r in the Bayesian model. They can be estimated from the data using empirical Bayes method. By integrating out $\boldsymbol{\beta}$, we obtain

$$\boldsymbol{y} \sim N\left(\mu_0 \mathbf{1}, \tau^2 U_D R U_D' + \sigma^2 I_n\right) = N\left(\mu_0 \mathbf{1}, \sigma_0^2 \Psi_D + \sigma^2 I_n\right). \tag{7}$$

The log-likelihood is given by,

$$l = constant - \frac{1}{2} \log det(\sigma_0^2 \Psi_D + \sigma^2 I_n) - \frac{1}{2}(\boldsymbol{y} - \mu_0 \mathbf{1})'(\sigma_0^2 \Psi_D + \sigma^2 I_n)^{-1}(\boldsymbol{y} - \mu_0 \mathbf{1}).$$

The empirical Bayes estimates can be obtained by maximizing the log-likelihood. Thus

$$(\hat{\mu}_0, \hat{\sigma}_0^2, \hat{r}) = arg \max_{\mu_0, \sigma_0^2, r} \; l \tag{8}$$

Now consider the case of unknown σ^2. A good estimate of σ^2 can be obtained if we have replicates. Let N be the total number of replicates in the experiment and D_N the design matrix obtained by repeating the rows in D as many times as the replicates. If we put a flat prior on σ^2, then the estimate of σ^2 (posterior mode) is given by

$$\hat{\sigma}^2 = \frac{1}{N}(\boldsymbol{y} - U_{D_N}\hat{\boldsymbol{\beta}})'(\boldsymbol{y} - U_{D_N}\hat{\boldsymbol{\beta}}). \tag{9}$$

(If prior information on σ^2 is available, then one can consider using the usual inverted gamma conjugate prior after inserting σ^2 into the prior variance of $\boldsymbol{\beta}$, see, e.g, Chipman, George, and McCulloch 2001). Equations (6), (8) and (9) need to be solved iteratively to obtain the final estimates. Note that D should be replaced by D_N in those equations. The procedure can be greatly simplified if we have balanced replication, which is the case in most well designed experiments. Suppose there are m replicates in each run and s_i^2 is the sample variance for the i^{th} run. Then, $\hat{\sigma}^2 = \sum_{i=1}^{n} s_i^2/n$ is a good estimate of σ^2. Now the estimation of $\boldsymbol{\beta}$, μ_0, σ_0^2, and r can be done from (6) and (8) by using the sample means instead of $\boldsymbol{y}$ and with σ^2 replaced by σ^2/m.

4. VARIABLE SELECTION

By the effect sparsity principle, not all of the variables are required to get a good prediction and only a few may be important. Therefore most often one would like to identify a parsimonious model containing only the important variables. This leads to the problem of variable selection, which is a very important problem in regression analysis. The problem is more complex in experiments because of the huge number of candidate variables (2^p). In a 2-level experiment, there are a total of 2^{2^p} possible models. For example, in an experiment with 10 factors there are a total of 2^{1024} models. It is almost impossible to search through all of them to find the best model. This immediately rules out techniques such as all-subsets regression. Whereas procedures like forward selection or step-wise regression can be easily used. Another feasible approach is to use the stochastic search variable selection (SSVS) procedure introduced by George and McCulloch (1993). It is a Bayesian procedure in which a prior probability is assigned for each model and then a model with the highest posterior probability given the data is identified through Gibbs sampling. See Chipman, George, and McCulloch (2001) for a review of several Bayesian variable selection techniques. Chipman (1996) introduced hierarchical priors that incorporates the effect heredity principle into variable selection. Chipman, Hamada, and Wu (1997) and Chipman (1998) used those priors together with the SSVS procedure to obtain a very useful variable selection strategy for experiments.

In this article, we do not put a prior on the model space. Instead, a Bayesian version of the forward selection strategy is proposed to do the variable selection. Indeed this strategy may not do a thorough search in the model space as in the SSVS procedure, but is computationally much simpler. Here forward selection is preferred over backward elimination because the number of candidate variables is much larger than the expected number of significant effects. Therefore a backward elimination strategy will be much more time consuming to implement than the forward selection strategy. Analysis of many real experiments shows that the proposed strategy works well in practice.

Suppose that k variables are "important". By this we mean that most of the variation in the response can be explained by using these k variables. Let $v_1, \cdots, v_k \in \{u_1, \cdots, u_{2^p-1}\}$ denote the k variables. For example v_1 could be x_3, v_2 could be x_1x_2, etc. Now write the transfer function as

$$f(\boldsymbol{x}) = \sum_{i=0}^{k} \mu_i v_i + \epsilon(\boldsymbol{x}), \quad \epsilon(\boldsymbol{x}) \sim GP(0, \sigma_k^2 \psi), \tag{10}$$

which is an extension of the model (1) used in Section 2. Here the prior mean of the transfer function is expanded to include all the important variables. The variability around the prior mean (σ_k^2) is expected to decrease as more variables are included in the prior mean. For the moment, assume that the correlation function (ψ) is known and does not change with k. Note that before the analysis, we do not know anything about the v variables. We will identify them one-by-one through a forward selection strategy. In 2-level experiments, approximate the transfer function by

$$f(\boldsymbol{x}) = \sum_{i=0}^{k} \mu_i v_i + \sum_{i=0}^{2^p-1} \beta_i u_i, \ \boldsymbol{x} \in \mathcal{X}.$$

It is easy to show that the distribution of β_i's induced from (10) is the same as that in Proposition 1, with τ^2 replaced by $\tau_k^2 = \sigma_k^2/(1+r)^p$. The model parameters and the hyper-parameters can be estimated as in the previous section.

To keep the exposition simple, we will explain the forward selection strategy using the case of unreplicated experiments. In this case, we cannot obtain an unbiased estimate of σ^2 and therefore assuming it to be small compared to the variation of the response in the experiment, we let $\sigma^2 = 0$. The parameters can be estimated as follows. Let V_k be the model matrix corresponding to $\boldsymbol{\mu}_k = (\mu_0, \cdots, \mu_k)'$. Then the estimate of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = \frac{1}{(1+r)^p} R U_D' \Psi_D^{-1} (\boldsymbol{y} - V_k \boldsymbol{\mu}_k), \tag{11}$$

and the posterior variance of $\boldsymbol{\beta}$ is given by

$$var(\boldsymbol{\beta}|\boldsymbol{y}) = \tau_k^2 \left(R - \frac{1}{(1+r)^p} R U_D' \Psi_D^{-1} U_D R \right). \tag{12}$$

To start the forward variable selection, set $k = 0$. Now we can estimate $\boldsymbol{\beta}$ from (11). The first entering variable v_1 can then be chosen as the one with the largest coefficient of $\hat{\beta}_i$. More correctly, since the posterior variance of the β_i's are different, we may instead use their standardized version. This can be justified as follows. Let $\hat{\sigma}^2_{\beta_i}$ be the posterior variance of β_i, which is the $(i+1)^{\text{th}}$ diagonal element of $var(\boldsymbol{\beta}|\boldsymbol{y})$. Then the $(1-\alpha)$ highest posterior density credible interval of β_i is given by $\hat{\beta}_i \pm \Phi^{-1}(1-\alpha/2)\hat{\sigma}_{\beta_i}$, where Φ is the standard normal distribution function. The credible interval will not contain 0 if $|\hat{\beta}_i/\hat{\sigma}_{\beta_i}| > \Phi^{-1}(1-\alpha/2)$. Therefore we can compute the ratio

$$t_i = \frac{\hat{\beta}_i}{\hat{\sigma}_{\beta_i}}, \tag{13}$$

and find the entering variable corresponding to the largest $|t_i|$.

The hyper-parameters at each step in the forward variable selection procedure can be estimated using empirical Bayes methods as described in the previous section. For $\sigma^2 = 0$, we obtain

$$\hat{\boldsymbol{\mu}}_k = (V_k' \Psi_D^{-1} V_k)^{-1} V_k' \Psi_D^{-1} \boldsymbol{y} \tag{14}$$

and

$$\hat{\sigma}_k^2 = \frac{1}{n}(\boldsymbol{y} - V_k \hat{\boldsymbol{\mu}}_k)' \Psi_D^{-1} (\boldsymbol{y} - V_k \hat{\boldsymbol{\mu}}_k). \tag{15}$$

Now consider the case with unknown correlation function. In general, the correlation function can change with k. Therefore, we should estimate r at each step of the forward selection procedure. Substituting for $\hat{\boldsymbol{\mu}}_k$ and $\hat{\sigma}_k^2$ in the integrated likelihood and maximizing, we obtain

$$\hat{r}^{(k)} = arg \min_{r \in [0,1]} \ n \log \hat{\sigma}_k^2 + \log det(\Psi_D). \tag{16}$$

In highly fractionated 2-level experiments, there may not be enough information to precisely estimate $r^{(k)}$. Then penalized likelihood estimation could be employed to get a stable estimate (Li and Sudjianto, 2005). Moreover, in such cases, we may avoid estimating $r^{(k)}$ at each step by taking $\hat{r}^{(k)} = \hat{r}^{(0)}$ for all k. This significantly reduces the computations. In fact, the strategy becomes simpler than that of the frequentist approach when $n << 2^p$. This is because at the i^{th} step of the frequentist variable selection procedure, $2^p - i$ linear regressions need to be performed to identify the entering variable. Whereas in the Bayesian procedure all the $2^p - i$ coefficients can be estimated simultaneously and therefore the entering variable can be easily identified.

To summarize the variable selection procedure, start with $k = 0$. First estimate $r^{(0)}$ from (16), then $\boldsymbol{\mu}_0$ and σ_0^2 from (14) and (15) respectively. Now compute $\hat{\boldsymbol{\beta}}$ and $var(\boldsymbol{\beta}|\boldsymbol{y})$ from (11) and (12), and identify v_1 by finding the largest $|t_i|$. Repeat the above procedure with $k = 1, 2, \cdots$ until the improvement in the fit is negligible. The fit of the model can be assessed by defining the R^2 measure in the usual way:

$$R_k^2 = 1 - \frac{\sum_{i=1}^n \{y_i - \hat{y}_k(\boldsymbol{x}_i)\}^2}{\sum_{i=1}^n (y_i - \hat{\mu}_0)^2}, \tag{17}$$

where $\hat{y}_k(\boldsymbol{x}) = \sum_{i=0}^k \hat{\mu}_i v_i$ and $\hat{\boldsymbol{\mu}}_k = (\hat{\mu}_0, \cdots, \hat{\mu}_k)'$. The R_k^2 can be plotted against k, and a model can be selected when the increase in R_k^2 becomes small. This approach is very similar to that used in regression analysis (see, e.g, Neter et al. 1996). We will explain the procedure with some examples in the next section.

5. EXAMPLES

Example 1: Consider the experiment reported by Hunter, Hodi, and Eager (1982) for improving the fatigue life of weld-repaired castings. There were seven factors and a 12-run Plackett-Burman design was used for the experiment. The design matrix (the first seven columns) and the data are given in Table 1. Plackett-Burman designs have a complex aliasing structure and so traditionally it was used only to analyze the main effects, ignoring all interactions. The half-normal plot (Daniel 1959) of the main effects is shown in Figure 1a. We see that only the factor F seems to have an effect on the response. The R^2 for the model with this effect is 45%. If we also include the main effect of D, then the R^2 becomes 59%.

Now consider the new approach. Here, there are no replicates and no extra information is available about the σ^2. Therefore, we let $\sigma^2 = 0$. Now for step 0 of the forward selection strategy, we first estimate the hyper-parameters. We obtain

$$\hat{r}^{(0)} = arg \min_{r \in [0,1]} \; n \log \hat{\sigma}_0^2 + \log det(\Psi_D) = 0.63,$$

Table 1: Design Matrix and Data, Cast Fatigue Experiment

Run	A	B	C	D	E	F	G	8	9	10	11	Y
1	+	+	−	+	+	+	−	−	−	+	−	6.058
2	+	−	+	+	+	−	−	−	+	−	+	4.733
3	−	+	+	+	−	−	−	+	−	+	+	4.625
4	+	+	+	−	−	−	+	−	+	+	−	5.899
5	+	+	−	−	−	+	−	+	+	−	+	7.000
6	+	−	−	−	+	−	+	+	−	+	+	5.752
7	−	−	−	+	−	+	+	−	+	+	+	5.682
8	−	−	+	−	+	+	−	+	+	+	−	6.607
9	−	+	−	+	+	−	+	+	+	−	−	5.818
10	+	−	+	+	−	+	+	+	−	−	−	5.917
11	−	+	+	−	+	+	+	−	−	−	+	5.863
12	−	−	−	−	−	−	−	−	−	−	−	4.809

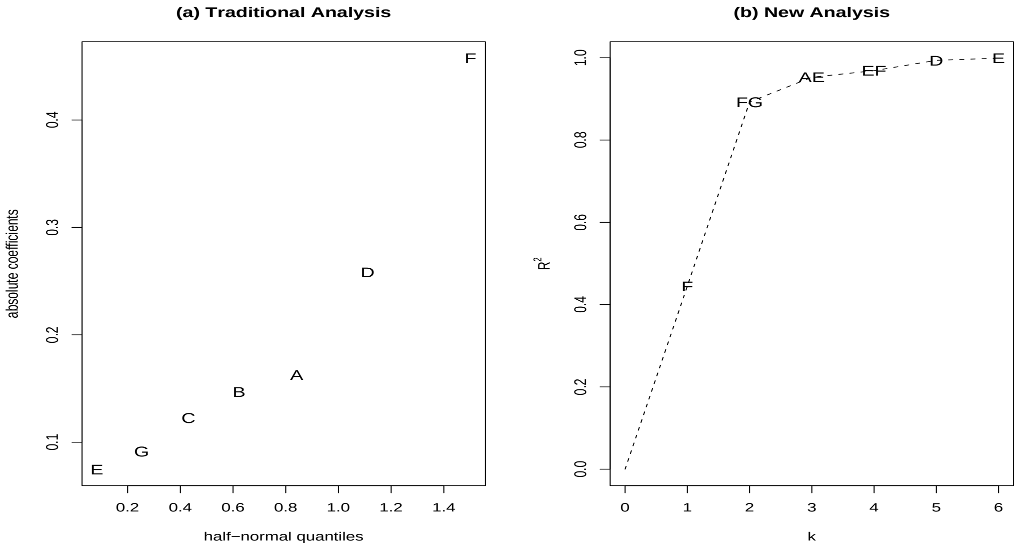

Figure 1: Analysis of Cast Fatigue Experiment

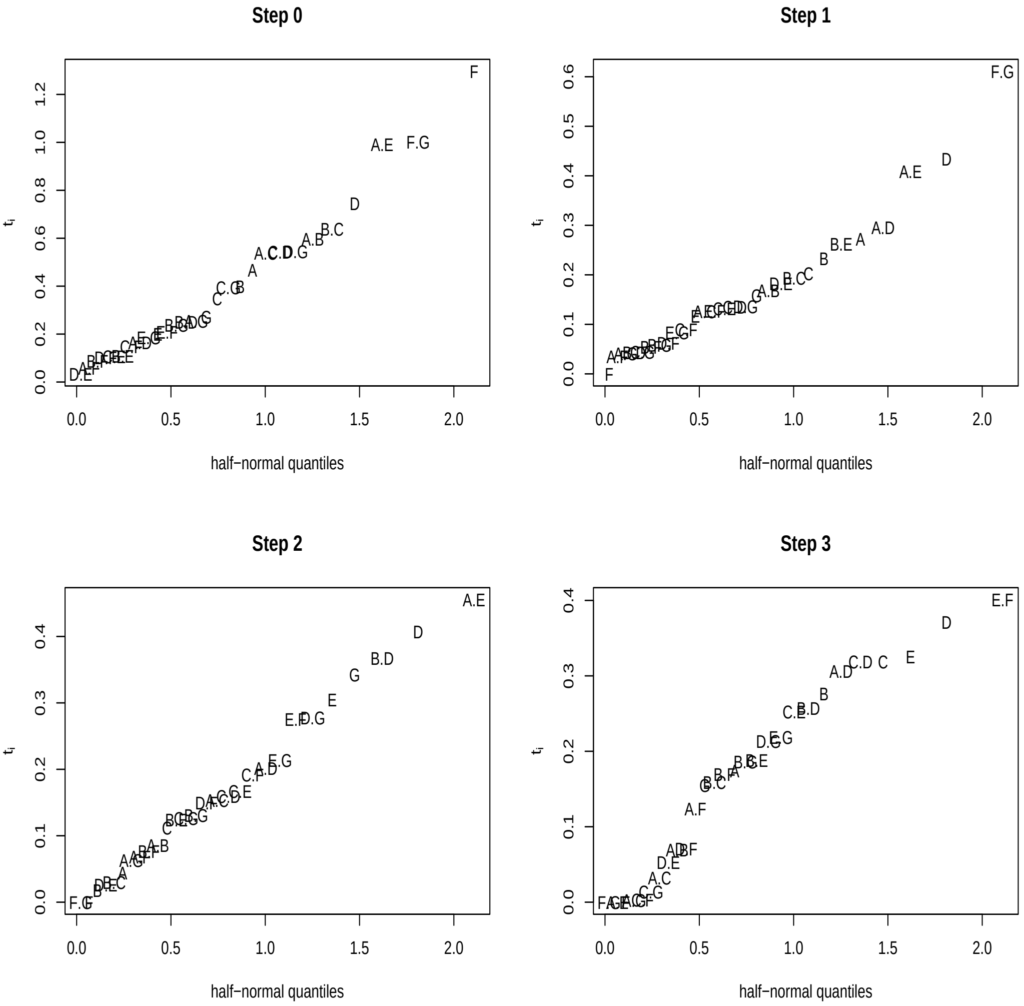

Figure 2: Forward Variable Selection in Cast Fatigue Experiment

$$\hat{\mu}_0 = (V_0'\Psi_D^{-1}V_0)^{-1}V_0'\Psi_D^{-1}\boldsymbol{y} = \frac{\mathbf{1}'\Psi_D^{-1}\boldsymbol{y}}{\mathbf{1}'\Psi_D^{-1}\mathbf{1}} = 5.73,$$

and

$$\hat{\sigma}_0^2 = \frac{1}{n}(\boldsymbol{y} - \hat{\mu}_0\mathbf{1})'\Psi_D^{-1}(\boldsymbol{y} - \hat{\mu}_0\mathbf{1}) = 0.47.$$

Suppose we are interested only in the main effects and 2fis. The t_i's for them are calculated from (13) and are plotted in the first panel of Figure 2. We see that the main effect of F is large. Therefore in step 0, we select factor F and move to step 1. Note that to identify the largest effect, we do not need a half-normal plot. It is plotted only for illustration. For step 1, the prior mean of the transfer function is $\mu_0 + \mu_1 v_1$, where $v_1 = x_6$. Thus V_1 is an $n \times 2$

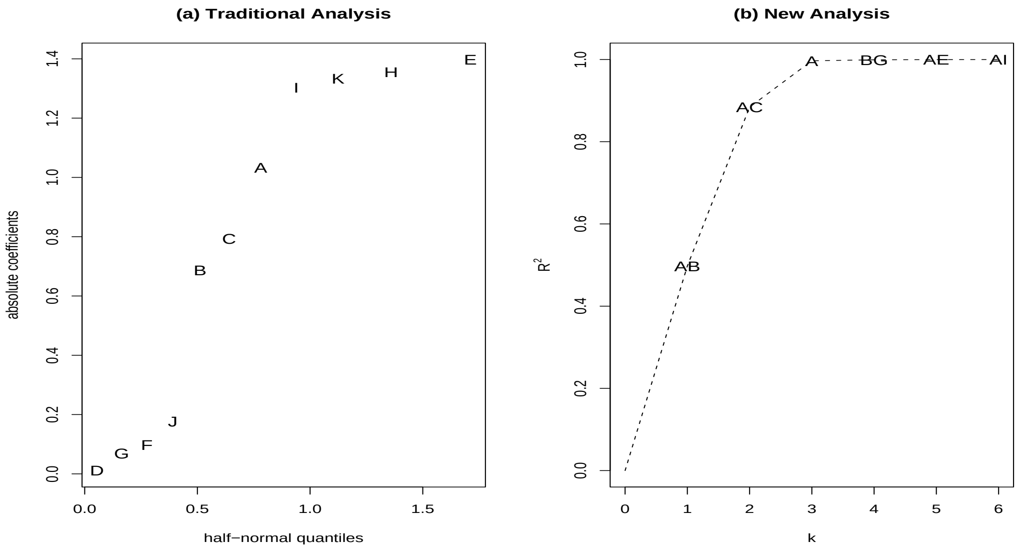

Figure 3: Analysis of Example 2, the true model is $Y = A + 2AB + 2AC + \epsilon$.

matrix whose first column is $\mathbf{1}$ and the second column is the values of x_6 in the design. We obtain $\hat{r}^{(1)} = 1$, $\hat{\boldsymbol{\mu}}_1 = (5.73, 0.46)'$, and $\hat{\sigma}_1^2 = 0.26$. The half-normal plot of the t_i's is given in the second panel of Figure 2. We see that the two-factor interaction FG is large. Thus choose $v_2 = x_6 x_7$ and proceed to step 2. Continuing similarly, the effects AE, EF, $\cdots$ are selected (see Figure 2) The R_k^2 can be calculated at each step using (17) and is plotted in Figure 1b. We see that only the two effects F and FG seem to be important (The importance of the above 2fi was first identified by Hamada and Wu (1992) using a frequentist forward selection strategy). A model with the above two effects gives an $R^2 = 89\%$, which is a substantial improvement over the model with F and D. This clearly shows the superiority of the new analysis over the traditional analysis.

Example 2: Consider an example given by Hamada and Wu (1992). The experiment uses the same 12-run Plackett-Burman design with 11 factors (denoted as A, B, $\cdots$, K). The data is generated from the model $Y = A + 2AB + 2AC + \epsilon$, where $\epsilon \sim N(0, 0.25^2)$, but we will analyze the experiment as though the true model is unknown. The half-normal plot from the traditional analysis is shown in Figure 3. This analysis shows that the main effects of E, H, K, and I are significant. Thus it completely misidentifies the true model. Now consider the new analysis. The R^2-plot is given in the second panel of Figure 3. We see that the new approach correctly identifies AB, AC, and A as the most significant effects. Hamada and Wu (1992) used the above example to show that their forward selection strategy can sometimes fail. Their primary method works well when only a few interactions are significant and are smaller than the main effects, which is not the case here. This example shows that the proposed procedure can perform better than their frequentist procedure.

6. OPTIMAL DESIGN

The objective is to choose the "best" n points from the set of 2^p points in $\{-1,1\}^p$. One good criterion is to choose the n points such that the prediction error in $\{-1,1\}^p$ is as small as possible. Let U_p denotes the full factorial model matrix. Now, we want to find a D such that $E\{(\boldsymbol{f}-\mu_0\mathbf{1}-U_p\hat{\boldsymbol{\beta}})'(\boldsymbol{f}-\mu_0\mathbf{1}-U_p\hat{\boldsymbol{\beta}})\}$ is a minimum, where $\boldsymbol{f}$ is the vector of function values in $\{-1,1\}^p$ and $\hat{\boldsymbol{\beta}}=E(\boldsymbol{\beta}|\boldsymbol{y})$ for which the expression is given in (6). But since $\boldsymbol{f}=\mu_0\mathbf{1}+U_p\boldsymbol{\beta}$, the above criterion reduces to minimizing $2^pE[tr\{var(\boldsymbol{\beta}|\boldsymbol{y})\}]$. Note that by virtue of the normal distribution, $var(\boldsymbol{\beta}|\boldsymbol{y})$ does not depend on the data. Therefore, define

$$\begin{aligned} A(D,r,\lambda) &= \tau^{-2}\sum_{i=0}^{2^p-1} var(\beta_i|\boldsymbol{y}) \\ &= tr\{R-RU_D'(U_DRU_D'+\lambda I_n)^{-1}U_DR\}, \end{aligned}$$

where $\lambda=\sigma^2/\tau^2$. The Bayesian A-optimal criterion is to minimize $A(D,r,\lambda)$. For details on different Bayesian optimality criteria, see Chaloner and Verdinelli (1995) and the references therein. For a given r and λ, we can minimize $A(D,r,\lambda)$ with respect to D to obtain the optimal design. Because λ is not known before the experiment, we need to choose a value of it for designing the experiment. If we assume that the factor effects are large compared to the error variance, then we can neglect the contribution of the error term, which is equivalent to setting $\lambda=0$. Usually, the objective function in the Bayesian A-optimal design criterion is defined as $tr\{(U_D'U_D+\lambda R^{-1})^{-1}\}$. Note that $A(D,r,0)$ cannot be obtained as a special case of the above expression, because it is defined only when $\lambda>0$ (see the Appendix). Therefore, even though $A(D,r,0)$ is important, it has not received much attention in the literature.

The objective function $A(D, r, \lambda)$ gives equal importance to all the parameters in the model. But in some cases we may want to give more importance to certain parameters. Therefore, we may define a more general objective function $\sum_{i=0}^{2^p-1} w_i var(\beta_i|\boldsymbol{y})$, where w_i's are some pre-specified weights. Consider some special cases of it. Define,

$$A_0(D, r, \lambda) = \tau^{-2} var(\beta_0|\boldsymbol{y}), A_1(D, r, \lambda) = \tau^{-2} \sum_{i=1}^{p} var(\beta_i|\boldsymbol{y}), \cdots, A_p(D, r, \lambda) = \tau^{-2} var(\beta_{2^p-1}|\boldsymbol{y}).$$

Thus $A(D, r, \lambda) = \sum_{i=0}^{p} A_i(D, r, \lambda)$. Here $A_0(D, r, \lambda)$ focuses on minimizing the variance of β_0, $A_1(D, r, \lambda)$ on the main effects, and so on. This leads to a new class of design criteria. When n is small, we may not be interested in estimating all the 2^p parameters in the model. Suppose we are interested only in the main effects and 2fis. Then, our design criterion should be to minimize $A_0(D, r, \lambda) + A_1(D, r, \lambda) + A_2(D, r, \lambda)$. The criterion can again be modified depending on the objective of the experiment. If the objective is to obtain a good prediction model, then the above criterion can be used. But if the objective is optimization, then we do not need to obtain a precise estimate of the intercept and therefore, we should minimize $A_1(D, r, \lambda) + A_2(D, r, \lambda)$, whereas if screening is the objective, then $A_1(D, r, \lambda)$ should be minimized.

The above criteria are highly nonlinear, non-convex functions of binary variables. To optimize them, general purpose algorithms such as genetic algorithms can be used (Hamada et al. 2001). Alternatively, one may try to derive more efficient algorithms by exploiting the structure of the objective function. This will not be discussed here. Instead, we will study some popular designs using the above criteria. Consider a 2^{p-k} fractional factorial design. Toman (1994), Mitchell, Morris, and Ylvisaker (1995) and Kerr (1999, 2001) discuss some closely related issues. Here we present some new results and obtain additional insights into the 2^{p-k} designs. A 2^{p-k} design has 2^k words in the defining contrast subgroup, including a column of 1's denoted as $\boldsymbol{I}$. From this defining contrast subgroup, we can obtain 2^{p-k} independent aliasing relationships. For example, for a 2_{IV}^{6-2} design with generators $\mathbf{5} = \mathbf{123}$ and $\mathbf{6} = \mathbf{124}$, the defining contrast subgroup is $\boldsymbol{I} = \mathbf{1235} = \mathbf{1246} = \mathbf{3456}$. This can be used to obtain the other $2^4 - 1 = 15$ independent aliasing relationships of the design, such as $\mathbf{1} = \mathbf{235} = \mathbf{246} = \mathbf{13456}$. Let $J_0(D)$ denote the indices of the effects in the defining contrast subgroup and $J_j(D)$, $j = 1, \cdots, 2^{p-k}$, that of the other aliasing relationships. For example, $J_1(D)$ contains the indices of the effects **1**, **235**, **246**, and **13456**. Then for a positive fraction (all design generators have positive signs) of a 2^{p-k} design , we obtain for an $i \in J_j(D)$

$$\beta_i | \boldsymbol{y} \sim N\left(\frac{R_{ii}}{\sum_{i \in J_j(D)} R_{ii} + \lambda/2^{p-k}}\, l_j,\ \tau^2 R_{ii} - \frac{\tau^2 R_{ii}^2}{\sum_{i \in J_j(D)} R_{ii} + \lambda/2^{p-k}} \right), \tag{18}$$

where $l_j = (U_{p-k}^{-1}(\boldsymbol{y} - \mu_0 \mathbf{1}))_j$ is the estimate of the j^{th} aliasing relationship. The proof is given in the Appendix. See Kerr (2001) for a related result.

Let $W(D) = \{N_1(D), N_2(D), \cdots, N_p(D)\}$ be the wordlength pattern of design D, where $N_i(D)$ denotes the number of words of length i in the defining contrast subgroup of the design D. Noting that $N_0(D) = 1$ and $J_0(D)$ gives the indices of the effects in the defining contrast subgroup, from (18) we obtain

$$A_0(D, r, \lambda) = 1 - \frac{1}{1 + \sum_{i=1}^{p} r^i N_i(D) + \lambda/2^{p-k}}.$$

Thus, Minimizing $A_0(D, r, \lambda)$ with respect to D is equivalent of minimizing $\sum_{i=1}^{p} r^i N_i(D)$. This shows that the design that minimizes $A_0(D, r, \lambda)$ can be obtained by minimizing a weighted average of $N_1(D), \cdots, N_p(D)$ with lesser weights on the higher order effects. On the other hand, a minimum aberration (MA) design can be obtained by sequentially minimizing $N_1(D), \cdots, N_p(D)$. The sequential minimization imposes a stronger requirement of the effect hierarchy principle. Because Proposition 1 gives a justification of the hierarchy principle and gives a quantification of the importance of each effect, minimizing a weighted average of the word length pattern is a more reasonable criterion.

As an example consider the following two 2_{IV}^{9-4} designs discussed in the introduction. $D_1 : \mathbf{6} = \mathbf{123}, \mathbf{7} = \mathbf{124}, \mathbf{8} = \mathbf{125}, \mathbf{9} = \mathbf{1345}$ and $D_2 : \mathbf{6} = \mathbf{123}, \mathbf{7} = \mathbf{124}, \mathbf{8} = \mathbf{134}, \mathbf{9} = \mathbf{2345}$. The first one is an MA design, whereas the second one is the maximum number of clear 2fis (MaxC2) design. Their word length patterns are $W(D_1) = (0, 0, 0, 6, 8, 0, 0, 1, 0)$ and $W(D_2) = (0, 0, 0, 7, 7, 0, 0, 0, 1)$. It is easy to see that $6r^4 + 8r^5 + r^8 < 7r^4 + 7r^5 + r^9$ for all $r \in (0, 1)$ and therefore D_1 is uniformly better than D_2 in terms of minimizing $A_0(D, r, \lambda)$. Thus, in this case the MA design is uniformly better than the MaxC2 design. Theoretically it is possible to have a word length pattern where this is not true, but it is not clear that such a design really exists! We can obtain the following result fairly easily. The proof is given in the Appendix.

PROPOSITION 2: *There exists an* $r_0 \in (0,1)$ *such that for all* $r \in (0, r_0)$ *and for all* λ*, an MA design minimizes* $A_0(D, r, \lambda)$.

The criterion $A_0(D, r, \lambda)$ considers only the posterior variance of β_0. Because optimization is the most common objective in industrial experiments, a more fair comparison of the designs can be obtained using $A_{12}(D, r, \lambda) = A_1(D, r, \lambda) + A_2(D, r, \lambda)$. Consider the case with $\lambda = 0$ and let $A_{12}(D, r) = A_{12}(D, r, 0)$. It can be easily calculated as

$$A_{12}(D, r) = tr(R_s - \frac{1}{(1+r)^p} R_s U_s' \Psi_D^{-1} U_s R_s),$$

where U_s is the model matrix that includes only the main effects and 2fis; and R_s is the corresponding sub-matrix of R. Numerically one can show that $A_{12}(D_1, r) > A_{12}(D_2, r)$ when $r \leq 0.1145$ and $A_{12}(D_1, r) < A_{12}(D_2, r)$ when $r > 0.1145$. Thus, in the above example, the MA design is preferred over the MaxC2 design only when $r > 0.1145$. But since an MA design is better than the MaxC2 design for a larger and more meaningful range of r, in practice we should choose the MA design unless there is a compelling reason to use a small value of r. A more careful analysis shows that $A_1(D_1, r, 0) < A_1(D_2, r, 0)$ for all r, but this is not true in the case of $A_2(D, r, 0)$. This is expected because the MaxC2 design has a lesser amount of contamination on the 2fis by the other 2fis & main effects and so $A_2(D, r, 0)$ can be smaller. The optimality of MA design with respect to $A_1(D, r, 0)$ is also not surprising, because Cheng, Steinberg, and Sun (1999) and Tang and Deng (1999) have shown that an MA design sequentially minimizes the number of interactions confounded with the main effects.

We note that the MA criterion can be used only with the regular fractional factorial designs, whereas the criterion $A(D, r, \lambda)$ is very general and can be used with any kinds of designs. It will be interesting to compare $A(D, r, \lambda)$ with a criterion for nonregular designs such as the minimum G_2-aberration criterion proposed by Tang and Deng (1999).

7. CONCLUSIONS

Specifying a prior distribution for the sheer number of parameters in the linear model is a difficult and important step in the design and analysis of fractionated experiments. We have shown that by using a functional prior on the underlying transfer function, a prior distribution can be induced for all the model parameters. Since the functional prior requires the specification of only a few hyper-parameters, the approach can be easily implemented in practice. We have given the details of estimation of the model parameters and the hyper-parameters. A Bayesian version of the forward selection strategy is proposed for variable selection. The usefulness of the new approach is demonstrated using the analysis of the cast fatigue experiment and another simulation example. We have also proposed a new class of design criteria and gave a Bayesian justification to the popular minimum aberration criterion.

Functional priors using Gaussian processes are widely used in the design and analysis of computer experiments. This article uses this idea in the design and analysis of physical experiments, which led to many new developments. It is possible to use some of the new findings, such as the prior specification for the model parameters, in the analysis of computer experiments. Extensions of these ideas thus, will lead to a *unified approach* to both computer and physical experiments.

This article focuses on 2-level experiments. Although the result of Theorem 1 is very general and can be used with any number of levels, it is not as appealing as that of Proposition 1 and does not give much insight into the properties of the model parameters. Ingenious selection of the coding system may improve the interpretation. This has to be worked out case by case by taking three levels, four levels, etc. Distinction between qualitative and quantitative factors is also to be made when dealing with higher level designs. This is part of the future research work. The design of experiments for three levels is to be given special attention because of its importance in industrial experiments (see, e.g, Taguchi 1987 and Wu and Hamada 2000). The methods for design of experiments for higher levels are not as widely accepted as those for two level experiments and therefore, there is great potential for improvement over the existing methods.

ACKNOWLEDGEMENTS

The author would like to thank the Associate Editor and the referees for their valuable comments and suggestions. The research was supported by the US National Science Foundation CAREER grant DMI-0448774.

APPENDIX: PROOFS

Proof of Proposition 1

Let U_i be the full factorial model matrix of the first i factors including all interaction columns and a column of 1's. Generate it using the following recursive relationship

$$U_i = \begin{pmatrix} U_{i-1} & -U_{i-1} \\ U_{i-1} & U_{i-1} \end{pmatrix},$$

for $i = 1, \cdots, p$ and with $U_0 = 1$. Thus U_p is a $2^p \times 2^p$ matrix, where the p^{th} factor appears only in the last 2^{p-1} columns.

Let Ψ_p be the matrix of correlations obtained from (4) corresponding to the factor levels in the U_p matrix. By the above construction of the full factorial matrix, the p^{th} factor takes -1 for the first 2^{p-1} rows and 1 for the remaining 2^{p-1} rows. Therefore the correlation between any two points within the first 2^{p-1} rows or the last 2^{p-1} rows is the same as that of the correlation with $p-1$ factors. Whereas the correlation between a point in the first 2^{p-1} rows and a point in the last 2^{p-1} rows will differ by a factor of $\psi_0(2)$ (because of the product correlation structure). Thus we have

$$\Psi_p = \begin{pmatrix} \Psi_{p-1} & \psi_0(2)\Psi_{p-1} \\ \psi_0(2)\Psi_{p-1} & \Psi_{p-1} \end{pmatrix}.$$

Simple matrix algebra gives

$$var(\boldsymbol{\beta}) = \frac{\sigma_0^2}{2^{2p}} U_p' \Psi_p U_p = \frac{\sigma_0^2}{2^{2p-1}} \{1 + \psi_0(2)\} \begin{pmatrix} U_{p-1}' \Psi_{p-1} U_{p-1} & 0 \\ 0 & r U_{p-1}' \Psi_{p-1} U_{p-1} \end{pmatrix}.$$

Because the p^{th} factor is arbitrary and since the last 2^{p-1} columns contain effects of order one higher than the first 2^{p-1} columns, we conclude that the variance decrease geometrically at the rate of r and that the effects are uncorrelated. Noting that $U_0'\Psi_0 U_0 = 1$, we obtain $var(\beta_0) = \sigma_0^2/2^p\{1+\psi_0(2)\}^p = \sigma_0^2/(1+r)^p$. Denote it as τ^2. Now, the result follows from Theorem 1. $\diamondsuit$

Proof and Discussion of Equation (5)

We have $\boldsymbol{y} = \mu_0\mathbf{1} + U_D\boldsymbol{\beta} + \boldsymbol{e}$, where $\boldsymbol{e} = (e_1, \cdots, e_n)'$. Since $\boldsymbol{e}$ is independent of $\boldsymbol{\beta}$, we obtain $cov(\boldsymbol{\beta}, \boldsymbol{y}) = \tau^2 R U_D'$ and $var(\boldsymbol{y}) = \tau^2 U_D R U_D' + \sigma^2 I_n$. Also, $E(\boldsymbol{\beta}) = \mathbf{0}$, $var(\boldsymbol{\beta}) = \tau^2 R$, and $E(\boldsymbol{y}) = \mu_0\mathbf{1}$. Now (5) can be obtained using the formula for the conditional distribution of normal variates (see, e.g, Santner, Williams, and Notz 2003, page 211).

When $\sigma^2/\tau^2 > 0$, using Woodbury's formula (see Harville 1997, page 424) (5) can be simplified to

$$\boldsymbol{\beta}|\boldsymbol{y} \sim N\left((U_D'U_D + \frac{\sigma^2}{\tau^2}R^{-1})^{-1}U_D'(\boldsymbol{y} - \mu_0\mathbf{1}), \sigma^2(U_D'U_D + \frac{\sigma^2}{\tau^2}R^{-1})^{-1}\right), \tag{19}$$

which is the most commonly used form in the literature. We prefer to use the form in (5) due to the following reason. Consider the case with $\sigma^2/\tau^2 = 0$. This can happen if $\sigma^2 = 0$ or if $\tau^2 = \infty$. The latter leads to the case of a noninformative prior for $\boldsymbol{\beta}$, which should not be used because we are dealing with the problem of estimating a higher number of parameters than the number of observations. Therefore we only need to consider the case of $\sigma^2 = 0$. If $rank(U_D) = n$, then $(U_D R U_D')^{-1}$ exists and from (5) we obtain

$$\boldsymbol{\beta}|\boldsymbol{y} \sim N\left(RU_D'(U_D R U_D')^{-1}(\boldsymbol{y} - \mu_0\mathbf{1}), \tau^2 R - \tau^2 R U_D'(U_D R U_D')^{-1}U_D R\right). \tag{20}$$

Whereas, since $rank(U_D'U_D) = n < 2^p$, $U_D'U_D$ is not invertible and therefore (19) cannot be used. Interestingly, even in the case of $\tau^2 = \infty$, the posterior mean in (5) exists and is equal to that in (20). More practically, (20) can be used when σ^2/σ_0^2 is small, which is a very reasonable assumption. This is because, we have $U_D R U_D' + \sigma^2/\tau^2 I_n = (1+r)^p(\Psi_p + \sigma^2/\sigma_0^2 I_n)$ which is approximately $(1+r)^p\Psi_p$ if $\sigma^2/\sigma_0^2 << 1$. This simplification cannot be done for (19). Thus, the form in (5) is more general and is therefore used in this article. $\diamondsuit$

Proof of Equation (18)

Let U_{p-k} be the full factorial model matrix for $p-k$ factors. Let $\boldsymbol{e} = (e_1, \cdots, e_n)'$. We have $\mu_0 \mathbf{1} + U_D \boldsymbol{\beta} + \boldsymbol{e} = \boldsymbol{y}$. Therefore, $Q\boldsymbol{\beta} + \boldsymbol{v} = \boldsymbol{l}$, where $Q = U_{p-k}^{-1} U_D$, $\boldsymbol{v} = U_{p-k}^{-1}\boldsymbol{e}$ and $\boldsymbol{l} = U_{p-k}^{-1}(\boldsymbol{y} - \mu_0\mathbf{1})$. Now consider the i^{th} component of $\boldsymbol{\beta}$. We want to find the conditional distribution of $\beta_i|\boldsymbol{y}$. First note that $\boldsymbol{v} \sim N(0, \sigma^2 2^{k-p} I_{2^{p-k}})$ and therefore the components of $\boldsymbol{v}$ are independent. Also $\boldsymbol{v}$ does not contain any information about $\boldsymbol{\beta}$. By Proposition 1, the components of $\boldsymbol{\beta}$ are also independent. For regular fractional factorial designs, an effect is either fully aliased or independent of the other effects and appears in only one aliasing relationship. Moreover, for positive fractions $Q_{ji} = 1$ if $i \in J_j(D)$ and 0 otherwise (see, e.g, Tang and Deng 1999). Then for an $i \in J_j(D)$, the conditional distribution of $\beta_i|Q\boldsymbol{\beta} + \boldsymbol{v} = \boldsymbol{l}$ is the same as the conditional distribution of $\beta_i|\sum_{i\in J_j(D)} \beta_i + v_j = l_j$. Now by using the formula for the conditional distribution of normal variables, we obtain

$$\begin{aligned}
E(\beta_i|\boldsymbol{y}) &= E(\beta_i|\sum_{i\in J_j(D)} \beta_i + v_j = l_j) \\
&= \tau^2 R_{ii}\{\sum_{i\in J_j(D)} \tau^2 R_{ii} + \frac{\sigma^2}{2^{p-k}}\}^{-1} l_j, \\
var(\beta_i|\boldsymbol{y}) &= var(\beta_i|\sum_{i\in J_j(D)} \beta_i + v_j = l_j) \\
&= \tau^2 R_{ii} - \tau^2 R_{ii}\{\sum_{i\in J_j(D)} \tau^2 R_{ii} + \frac{\sigma^2}{2^{p-k}}\}^{-1} \tau^2 R_{ii}.
\end{aligned}$$

It is easy to verify that for designs that are not positive fractions, the sign of the means can change depending on the sign of the design generators but not the variances. ◇

Proof of Proposition 2

$A_0(D, r, \lambda)$ can be minimized by minimizing $g(r, D) = \sum_{i=1}^{p} r^i N_i(D)$. Let D^* be an MA design. Then, for any other design D, $g(r, D^*) - g(r, D) = \sum_{i=1}^{p} r^i\{N_i(D^*) - N_i(D)\}$. Let j be the largest value of i for which $N_i(D^*) \neq N_i(D)$. Then, $g(r, D^*) - g(r, D) = r^j h(r)$, where $h(r) = \sum_{i=j}^{p} r^{i-j}\{N_i(D^*) - N_i(D)\}$. Since D^* is an MA design, $N_j(D^*) < N_j(D)$. Therefore $h(0) < 0$. Also, since $h(r)$ is a continuous function of r, there exists an $r_0 \in (0, 1)$ such that $h(r) < 0$ for all $r \in (0, r_0)$. Thus $g(r, D^*) < g(r, D)$ for all $r \in (0, r_0)$, which completes the proof. ♢

REFERENCES

Box, G. E. P. and Hunter, J. S. (1961), "The Fractional Factorial Designs," *Technometrics*, 3, 311-351 and 449-458.

Box, G. E. P. and Meyer, R. D. (1993), "Finding the Active Factors in Fractionated Screening Experiments," *Journal of Quality Technology*, 25, 94-105.

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

Chipman, H. (1996), "Bayesian Variable Selection with Related Predictors," *Canadian Journal of Statistics*, 24, 17-36.

Chipman, H., Hamada, M., and Wu, C. F. J. (1997), "A Bayesian Variable-Selection Approach for Analyzing Designed Experiments with Complex Aliasing," *Technometrics*, 39, 372-381.

Chipman, H. (1998), "Fast Model Search for Designed Experiments with Complex Aliasing," *Quality Improvement Through Statistical Methods* (Ed. Abraham, B.), Boston: Birkhäuser, 207-220.

Chipman, H., George, E. I., and McCulloch, R. E. (2001), "The Practical Implementation of Bayesian Model Selection," (with discussion), *IMS Lecture Notes - Monograph Series*, 38, 65-134.

Chen, J. (1992), "Some Results on 2^{n-k} Fractional Factorial Designs and Search for Minimum Aberration Designs," *Annals of Statistics*, 20, 2124-2141.

Cheng, C. S. and Mukerjee, R. (1998), "Regular Fractional Factorial Designs with Minimum Aberration and Maximum Estimation Capacity," *Annals of Statistics*, 26, 2289-2300.

Cheng, C. S., Steinberg, D. M., and Sun, D. X. (1999), “Minimum Aberration and Model Robustness for Two-Level Fractional Factorial Designs,” *Journal of Royal Statistical Society, Series B*, 61, 85-93.

Cheng, S. W. and Ye, K. Q. (2004), “Geometric Isomorphism and Minimum Aberration for Factorial Designs with Quantitative Factors,” *Annals of Statistics*, 32, 2168-2185.

Currin, C., Mitchell, T. J., Morris, M. D., and Ylvisaker, D. (1991), “Bayesian Prediction of Deterministic Functions, with Applications to the Design and Analysis of Computer Experiments,” *Journal of American Statistical Association*, 86, 953-963.

Daniel, C. (1959), “Use of Half-Normal Plots in Interpreting Factorial Two-Level Experiments,” *Technometrics*, 1, 311-341.

Fries, A. and Hunter, W. G. (1980), “Minimum Aberration Designs,” *Technometrics*, 22, 601-608.

Harville, D. A. (1997), *Matrix Algebra from a Statistician’s Perspective*, New York: Springer.

Hamada, M. and Wu, C. F. J. (1992), “Analysis of Designed Experiments with Complex Aliasing,” *Journal of Quality Technology*, 24, 130-137.

Hamada, M., Martz, H. F., Reese, C. S., and Wilson, A. G. (2001), “Finding Near Optimal Bayesian Experimental Designs via Genetic Algorithms,” *American Statistician*, 55, 175-181.

Hunter, G. B., Hodi, F. S., and Eager, T. W. (1982), “High Cycle Fatigue of Weld Repaired Cast Ti-6Al-4V,” *Metallurgical Transactions*, 13A, 1589-1594.

Kerr, M. K. (1999), *Stationary Processes on 2^k for Bayesian Experimental Design*, Ph.D. Thesis, University of California, Los Angeles.

Kerr, M. K. (2001), “Bayesian Optimal Fractional Factorials,” *Statistica Sinica*, 11, 605-630.

Li, R. and Sudjianto, A. (2005), "Analysis of Computer Experiments Using Penalized Likelihood in Gaussian Kriging Models," *Technometrics*, 47, 111-120.

Meyer, R. D., Steinberg, D. M., and Box, G. (1996), "Follow-up Designs to Resolve Confounding in Multifactor Experiments," (with discussion), *Technometrics*, 38, 303-332.

Mitchell, T. J., Morris, M. D., and Ylvisaker, D. (1995), "Two-Level Fractional Factorials and Bayesian Prediction," *Statistica Sinica*, 5, 559-573.

Montgomery, D. C. (2004), *Design and Analysis of Experiments*, 6th Edition, New York: Wiley.

Neter, J., Kutner, M. H., Nachtsheim, C. J., and Wasserman, W. (1996), *Applied Linear Statistical Models*, Boston: McGraw-Hill.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003), *The Design and Analysis of Computer Experiments*, New York: Springer.

Taguchi, G. (1987), *System of Experimental Design*, Vol 1 & 2, White Plains, New York: Unipub/Kraus International.

Tang, B. and Deng, L. Y. (1999), "Minimum G_2-Aberration for Nonregular Fractional Factorial Designs," *Annals of Statistics*, 27, 1914-1926.

Toman, B. (1994), "Bayes Optimal Designs for Two and Three Level Factorial Experiment," *Journal of the American Statistical Association*, 89, 937-946.

Wu, C. F. J., and Hamada, M. (2000), *Experiments: Planning, Analysis, and Parameter Design Optimization*, New York: Wiley.

Wu, H. and Wu, C. F. J. (2002), "Clear Two-Factor Interactions and Minimum Aberration," *Annals of Statistics*, 30, 1496-1511.