Optimal Designs for Model Selection

Derek R. Bingham and Hugh A. Chipman
University of Michigan and University of Waterloo

Abstract:

An important use of experimental designs is in *screening*, in which experimenters seek to identify significant effects (both main effects and potentially interactions) from a large set of candidate effects. This paper goes further than identification of effects, introducing a design criterion that seeks to maximize the ability to discriminate between models. The criterion is based on the Hellinger distance between predictive distributions under competing models, and motivated by Meyer, Steinberg and Box (1996). A bound for the criterion is obtained, greatly improving interpretability. The set of all possible models to compare is huge, and not all models are equally plausible. This challenge is addressed via a Bayesian approach. This approach uses prior distributions on the space of models, indicating preference for intuitively appealing models, such as those with few effects, more low order than high order effects, and inheritance structure between active main effects and interactions. Techniques for evaluating the criterion and searching for optimal designs are presented. The effectiveness of the criterion is illustrated via a number of examples, which consider regular and non-regular designs, robust designs, and scenarios with partial prior knowledge of which effects are significant.

Key Words: Bayesian design, complex aliasing, effect sparsity, effect hierarchy, effect heredity, Hellinger distance, model discrimination

1 Introduction

Screening designs are frequently used by experimenters to help understand the impact of a large number of factors in relatively few trials. By varying factors of interest over specified level settings and performing trials, experimenters gain insight into which factors or effects are important. The choice of the best experimental design is among the first and most fundamental issues facing an experimenter.

In many situations, however, designs are chosen because of their ability to estimate certain effects rather than for screening properties. This article considers techniques that directly address this problem, seeking designs which will provide maximal information for model selection, while at the same time allowing for optimal estimation. An important innovation in this approach is the ability to incorporate into the design optimality criterion a flexible means of specifying preference for plausible classes of models. Before outlining our approach, we review past work in experimental design and how it relates to screening.

Common choices of experiment plans are the 2^{q-k} regular fraction factorial (FF) designs (designs with an associated defining contrast sub-group), where q factors are investigated in

2^{q-k} trials. FF designs are most often ranked by the minimum aberration (MA) criterion (Fries and Hunter, 1980), which sequentially minimizes the elements of the word-length pattern. There are two underlying assumptions which motivate the use of the MA criterion:

A1. *Effect Sparsity*: The number of important effects is relatively small.

A2. *Effect Hierarchy*: Lower order effects are more likely to be important than higher order effect and effects of the same order are equally important.

An alternative class of designs are non-regular fractional factorials (NFF) such as Plackett-Burman (1945) designs. These designs have typically been used as main effects plans because of the complex aliasing structure between main effects and two-factor interactions (2fi's). Recently, methods have been proposed for analyzing complex aliasing designs that entertain models with both main effects and 2fi's (Hamada and Wu, 1992; Chipman, Hamada and Wu, 1997). These iterative approaches rely on an additional assumption to help sort through the complex aliasing structure:

A3. *Effect Inheritance*: An interaction is more likely to be important if one or more of its parent factors are also important.

In light of this additional assumption, Hamada and Wu (1992) viewed the complex aliasing of NFF's as an advantage because NFF's give the opportunity to identify promising interactions as well as main effects. NFF's can be rank-ordered by the maximum estimation capacity (EC) criterion (Sun, 1997; Cheng and Mukerjee, 1998; Chen, Steinberg and Sun, 1999), which computes the proportion of models containing all main effects and a fixed number of 2fi's that are estimable. It has been noted that NFF designs are frequently better suited for screening because they provide more information for a broader class of models and thus have higher EC than some competing regular fractional factorial designs. The difficulty with ranking design by EC is that the criterion does not recognize the effect sparsity principle when there are many factors. Typically models that are smaller than those counted by EC are preferable, and, as a consequence EC can underestimate the merits of a design. This is important because some models are more likely than others. If the estimability of some models must be sacrificed in order to achieve run-size economy, one would likely prefer to sacrifice less likely and less parsimonious models.

This belief that estimation of some models is more important than others is not easy to incorporate into criteria such as MA and EC in some practical applications. For instance, in robust parameter design, experimenters are particularly interested in estimating interactions between control and noise factors. In these cases, the design criterion should be adjusted so that the experiment maximizes the information about such interactions (e.g., see Bingham and Sitter, 2001; Bingham and Li, 2002; Wu and Zhu, 2001). Another example is when experimenters have prior knowledge about the significance of some of the effects (Franklin and Bailey, 1977). In these situations, the best experiment design is frequently not the best in terms of MA or EC.

In this article, we consider a different kind of optimality, which seeks designs that best facilitate discrimination between competing models. Two key components of this approach

are an optimality criterion that explicitly measures the ability of the design to discriminate between competing models, and a flexible framework for specifying which models are most important.

The measurement of the ability to discriminate among models is accomplished via a Bayesian approach, which has similarities to the model discrimination criterion (MD) (Meyer, Steinberg and Box, 1996, henceforth MSB) which is based on the Kullback-Leibler information. We propose a different criterion, based on the Hellinger distance between predictive densities, to help distinguish between competing models. The Hellinger distance is preferable to the Kullback-Leibler information in this setting because it requires half the computational expense and is bounded. The ability to easily calculate an upper bound greatly improves the interpretability of the Hellinger distance. Also, the Hellinger distance has appealing properties that allows experimenters to use it as a basis for choosing an appropriate run size.

The second key component of our approach is the ability to flexibly specify which models are important. This is accomplished by specifying a prior distribution on the set of all models under consideration. This prior incorporates assumptions A1-A3, and assigns a weight (i.e. prior probability) to each model. This allows for a smoothly varying way of specifying preferences for models. This approach was briefly explored in Chipman and Hamada (1996), and offers additional flexibility to the approach of MSB. In contrast, criteria like MA and EC take a much more discrete view of the model space, counting numbers of estimable effects. For example, consider a 6 factor experiment estimating only main effects and two-way interactions. Traditional criteria such as MA would place a higher priority on main effects than interactions, implicitly assuming that the model $A+B+C+D+E+F$ is more plausible than the model $A+B+AB$.

Beliefs such as a preference for a small model with inheritance (e.g. $A+B+AB$) can lead to surprising results. In Section 4, we show that if a design criterion representing such a belief is used, a non-regular design may be preferable, even in contexts where a regular fractional factorial design is available. Other situations arise in which there is prior preference for certain models, such as the robust design problem and having prior knowledge of some effect significance, as discussed earlier in this section.

The article is organized as follows: In Section 2, the design criterion is introduced, along with the Bayesian specification, including priors on the model space. Section 3 discusses methods to identify optimal designs, including some computational techniques unique to this problem. The approach is illustrated with an example in Section 4, involving a situation in which a regular design exists. Section 5 discusses a bound for the criterion, and illustrates how it can be effectively used to choose an appropriate number of runs. Further examples are given in Section 6, including non-regular fractional factorial designs, screening designs in which some effects are a priori more likely, and robust design experiments. Section 7 concludes with a discussion.

2 Models, priors, and model discrimination criteria

In this section, we introduce the design criterion and the associated Bayesian formulation. Section 2.1 gives the design criterion based on the Hellinger distance. Section 2.2 introduces the Bayesian formulation of the linear regression model and the associated model selection problem, including priors representing assumptions A1-A3.

2.1 The design criterion

We begin by outlining the design and analysis setting that motivates our methodology. Our aim is to identify the "optimal" design with q two-level factors in n trials for estimating the parameters of the linear model:

$$Y = X\beta + \epsilon, \tag{1}$$

where Y is the $N \times 1$ vector of observations, X is the model matrix, β is the vector of factorial effects and the intercept, and ϵ is the vector of iid $N(0, \sigma^2)$ random errors.

The model selection problem amounts to identifying a subset of predictors as active, and in this setting there are typically more parameters to estimate than unique treatments. The possible models will be labeled as $M_1, M_2, \ldots, M_K$. We defer discussion of priors for M_i and (β_i, σ) until Section 2.2.

To evaluate a design's ability to distinguish between two models (say, M_i and M_j), we use the Hellinger distance between predictive densities:

$$H(f_i, f_j) = \int (f_i^{1/2} - f_j^{1/2})^2 dY = 2 - 2\int (f_i\ f_j)^{1/2} dY. \tag{2}$$

It is easiest to discriminate between models if they make different predictions, hence HD is a larger-the-better criterion. The experimental design enters this criterion through f_i and f_j, since the predictive densities are evaluated at the values of the factors specified in the design. The fact that the Hellinger distance is bounded between 0 and 2 will be used later in Section 5 to establish bounds on the corresponding design criterion.

Why is disagreement between predictive densities a good design criterion? From a philosophy of science perspective, the different models can be thought of as different scientific hypotheses or theories. One of the most effective ways to decide between competing theories is to find situations in which they predict different states of nature, and observe which state actually occurs. This philosophy motivates (2): a good design is one which will yield data that will cause different models to predict differently.

Interest lies in comparing all possible models and thus the Hellinger distance for all pairwise comparisons should be computed. We propose the HD criterion, which is a weighted average of the Hellinger distances, under all possible pairs of models, using the product of probabilities of the two models $P(M_i)P(M_j)$ as weights.

$$HD = \sum_{i<j} P(M_i)P(M_j)H(f_i, f_j). \tag{3}$$

All models are not equally likely. The weighting of the Hellinger distances serves to put priority on distinguishing the most probable models. MSB use a related approach with different model priors to average a Kullback-Leibler information measure between predictive densities over all pairs of models.

2.2 Priors

The HD criterion (3) uses predictive densities, which implies a Bayesian formulation of the problem. In this section, priors are specified in two stages: a prior on models (e.g. the M_i), and a prior on (β_i, σ) conditional on M_i.

We first introduce the priors on (β_i, σ) given model M_i. Details of the derivation of the Hellinger distance are given in the Appendix. A proper prior for σ^2 such as an inverted gamma is selected in most applications, and we recommend doing so here to guarantee that the Hellinger distance is bounded. The coefficient vector β_i and the associated matrix of regressors X_i are indexed by i. Let r_i be the number of columns in X_i (i.e., the number of effects in model M_i plus one one additional column for the intercept). We use conventional independent normal priors for the parameters of the regression model i. Thus $\pi(\beta_i|\sigma^2) \sim MVN(0, \sigma^2\Gamma_i)$, where

$$\Gamma_i = \gamma^2 \begin{pmatrix} c & 0 \\ 0 & I_{r_i-1} \end{pmatrix}. \tag{4}$$

We choose $c = 1,000,000$ so that the prior on the intercept is has mean 0 and large variance. In all calculations presented we take $\gamma = 2$. MSB suggest this is a reasonably uninformative value in the context of starting designs. More careful choice of γ is required for follow-up designs, which are not considered in this article.

This prior formulation means the Hellinger distance (2) can be written as

$$H(f_j, f_j) = 2 - \frac{2}{\left|\frac{1}{2}\left(\Sigma_i^{-1/2}\Sigma_j^{1/2} + \Sigma_i^{1/2}\Sigma_j^{-1/2}\right)\right|^{1/2}} \tag{5}$$

where

$$\Sigma_i = (I_n + X_i'\Gamma_i X_i). \tag{6}$$

Some intuition about the distance measure can be gained by considering the situation when there are only two competing models and only one trial to be conducted. In this instance, the Hellinger distance will be greatest for a design where there is little uncertainty about one model and large uncertainty about the other. The criterion amounts choosing trials where the average relative uncertainty between models is largest.

The prior distribution on the model space is constructed via simplifying assumptions, such as independence of the activity of main effects (Box and Meyer 1986, 1993), and independence of the activity of higher order terms conditional on lower order terms (Chipman 1996, and Chipman, Hamada, and Wu 1997). These simplifications correspond to the three model building assumptions (A1-A3) used for screening experiments. We work within these assumptions to construct the prior on each model.

First, the prior probability that an effect is active should be relatively small to obey the effect sparsity assumption. Second, to follow the effect hierarchy assumption, the prior probability that a main effect is active (denoted p_A for factor A) should be at least as large as the prior probability that an interaction term is in the model. To represent the effect inheritance assumption, the probability that an interaction is included in the model will depend on the presence of the parent main effects. That is, let $p_{AB,0} \leq p_{AB,1} \leq p_{AB,2}$ denote the conditional probabilities that an AB interaction is active, given 0, 1 or 2 of main effects A and B are active. This hierarchy can be extended to higher order interactions, but we confine our presentation to main effects and 2fi's (Chipman (1996) discusses the choice of priors for higher order interactions).

The choice of prior probability that an effect is active should coincide with the experimenter's interpretation of the effect sparsity principle. If a 12-run design with 6 factors is to be performed, clearly one would not anticipate 11 active effects (main effects and 2fi's). Indeed, this is implicit in the willingness to perform a 12-run design. Consequently, the probability that an effect is active should be relatively small.

Our choice of prior distribution is flexible and can be calibrated to reflect the experimenter's view of A1-A3. Let p denote the prior probability that a main effect is active. For 2fi's, we use the following prior specification,

$$p_{AB,i} = \begin{cases} 0.01p & \text{if} \quad i = 0 \\ 0.5p & \text{if} \quad i = 1 \\ p & \text{if} \quad i = 2. \end{cases} \tag{7}$$

A more general case is considered in the appendix, along with calculations that show that the expected number of main effects under (7) is

$$E(\#\text{ effects}) = pq + pq(q-1)\left\{.005 + .49p + .005p^2\right\}. \tag{8}$$

Here q is the number of factors considered in the experiment. The first term in (8) is the expected number of main effects, and second gives the expected number of 2fi's.

The choice of p is now made so that expected number of active effects under the prior matches that of the experimenter's prior belief. For a specified number of effects expected to be active, (8) can easily be solved for p. This is a particularly attractive feature of the methodology since it explicitly builds in the experimenter's prior belief about the size of the model. In most situations, it is easier for an experimenter to express belief about the number of anticipated effects rather than a probability associated with an effect.

MSB consider a more extreme prior than (7): An interaction can only be active if both parents are active, in which case it must be active. That is, $p_{AB,0} = p_{AB,1} = 0$ and $p_{AB,2} = 1$. In effect, interactions are *forced* into the model, meaning that the models $\{A, B\}$, $\{A, AB\}$, $\{B, AB\}$ would receive zero prior mass, and only $\{A, B, AB\}$ would receive nonzero mass. Under the more flexible formulation with $p_{AB,2} < 1$, the mass that MSB allocate to $\{A, B, AB\}$ is split between that model and the other three listed.

While somewhat restrictive, the prior suggested by MSB significantly simplifies computation. The number of potential models is dramatically reduced. If main effects and all possible

two way interactions are considered, p main effects would lead to $\binom{p}{2} = (p^2 - p)/2$ interactions, and a total of $K = 2^{(p^2+p)/2}$ models. This is important when evaluating HD, where the constituent elements of the summation are for every possible pair of models. With 8 factors, 128 models would have positive mass under MSB, while the total model space would contain 69 billion models in our formulation. In the next section we propose a solution to this problem.

3 Searching for optimal designs

The search for promising designs involves two challenges:

- Evaluation of the HD criterion.
- An effective search algorithm for HD-optimal designs.

In many cases, to evaluate HD, we cannot calculate (3) for all pairs of models because the model space is too large. Instead, we attempt to evaluate the largest terms in (3), by discarding models that have small prior probability $P(M_i)$. By replacing an average over all pairs of models with an average over the most probable models, the evaluation of HD becomes tractable. Raftery, Madigan, and Hoeting (1997) take a similar approach to model averaging, sampling from a probability distribution on models and discarding all but the most probable.

The strategy outlined above is implemented as follows: To identify models with high probability, simulate a large number of models from the prior distribution. For those models, evaluate their exact prior probability, and keep the most probable models. Factorization of the prior facilitates such simulation, with main effects first being drawn from independent Bernoulli distributions, followed by higher order terms which depend on the main effects.

Since many design optimization algorithms will involve multiple runs from different start points, there will be many evaluations of HD. We have found it effective to use a small set of models for preliminary searches and then re-evaluate HD for all promising designs using a larger set of models.

The design optimization algorithm used here is an exchange algorithm similar to that of Wynn (1972). Since we only consider 2-level factors, all designs considered have candidate values of ± 1 for all effects. Our exchange algorithm begins with a random n-run design, with design points chosen with replacement from the 2^q full factorial design, and considers adding a single run. After evaluating HD for all $n + 1$ run designs that contain the current n runs, the design with the best improvement in HD is selected. A similar process is then repeated, identifying the one run whose removal causes the smallest reduction in HD. Iterations alternate between addition and deletion of runs until no further changes are produced. There is no guarantee that this approach will converge to the HD-optimal design, therefore multiple random designs are used as start points to generate a variety of promising designs.

4 Example: 6 Factors in 16-Runs

An example where a resolution IV FF design exists is first considered. This is done to see how an HD-optimal design compares to the most obvious choice for an experimenter. Suppose an experiment with 6 factors in 16 trials is to be performed. The MA 2^{6-2} FF design with 6 factors (labeled $A-F$) has defining contrast sub-group:

$$I = ABCE = ABDF = CDEF. \tag{9}$$

To find the HD-optimal design, the prior probability that a main effect is active must first be chosen. The prior specification on the model space is obtained by setting (7) equal to the expected number of effects. If the expected number of active effects is 5, then (7) implies that $p = 0.410$.

A maximum of 40 models were used to evaluate HD in the exchange algorithm. Fifty random restarts of the exchange algorithm were used, and the resultant 50 designs were then re-evaluated using 400 models.

The HD-optimal design is shown in Appendix 1. The HD-optimal design is not the same as the MA FF design. Indeed, it is not a regular FF design. So why are the designs different? Quite simply, the design criteria are measuring different things. MA FF designs emphasize first the estimability of all main effects free of aliasing with other main effects and interactions, followed by the estimability 2fi's. Here the MA design results in some 2fi's being completely aliased (e.g., $AC = BE$). HD-optimality, on the other hand, emphasizes the estimability of models and several effects are partially aliased, but none are fully aliased.

The HD-optimal design is, in fact, an orthogonal array, and therefore there is no aliasing between main effects. There is partial aliasing between main effects and some 2fi's. Hall (1961) found that there are exactly 5 non-isomorphic 16-run orthogonal arrays. The first design is the regular FF design. The remaining four designs can be found in Wu and Hamada (2000, page 333) and are labeled I-IV. The HD-optimal design is the same as selecting columns 1, 4, 7, 9, 12 and 14 of Halls design II.

As noted by Hamada and Wu (1992), the partial aliasing of NFF's can have advantages over regular FF designs. For instance, in this example the MA resolution IV design is the traditional choice. However, the model $Y = A + B + AC$ is indistinguishable from the model $Y = A + B + BE$ because the effects AC and BE are fully aliased. However, the model $Y = A + B + C + D + E + F$ is estimable using the MA FF design since the design is resolution IV. Thus, implicitly the latter model is viewed as more likely than the former two models under the MA criterion.

Under the HD criterion, $Y = A + B + AC$ and $Y = A + B + BE$ will typically have more prior mass (both have prior mass of 4.82×10^{-4}) and thus be viewed as more likely than $Y = A + B + C + D + E + F$ (prior mass of 7.11×10^{-7}) because the criterion penalizes large models and instead puts more mass on small, more parsimonious sub-sets of effects. The HD-optimal design is able to estimate each of the three models and is best at distinguishing between the models in terms of Hellinger distance under the prior in (7).

5 A bound and a graphical representation of the HD criterion

Evaluation of the suitability of an experiment design is an important step in design selection. For example, trade-offs between run size and estimability are often necessary. The choice of run size is facilitated by a bounded design criterion. In this section we explore a bound for HD and its value in determining an appropriate run size.

The Hellinger distance is bounded above by 2, and ideally each model is perfectly distinguishable in practice. So, ideally Hellinger distance between all predictive densities will be 2. In practice, this will not happen for finite sample sizes in this setting, but helps establish a useful upper bound on the HD criterion.

$$HD \leq 2 \sum_{i<j} P(M_i)P(M_j). \tag{10}$$

This bound still depends on the model probabilities, and also the number of models included in the summation. An upper bound is given in the following theorem.

Theorem: For all possible probability distributions on the model space,

$$HD < 1.$$

Proof: Suppose that there are K models. The probability distribution that will maximize (10) will put equal mass $1/K$ on each of the K models. Then

$$HD \leq 2 \sum_{i<j} P(M_i)P(M_j) \leq 2 \sum_{i<j} \frac{1}{K}\frac{1}{K} = 2 \times \binom{K}{2} K^{-2} = 2K(K-1)K^{-2}/2 < 1 \tag{11}$$

Jones and DuMouchel (1996) stress the importance of interpretability of a design optimality criterion, and express an opinion that the MD criterion of MSB is difficult to interpret. The upper bound established in the Theorem helps shed some light on the overall quality of the design. In practice, we can use the upper either use 1 as an upper bound or compute $2 \sum_{i<j} P(M_i)P(M_j)$ directly from the approximated model space to help put an HD-value into perspective. We discuss this issue further after introducing a useful plot.

The bound (10) does not depend on sample size n, and one would expect that as run size increases, HD approaches the bound. To explore this, we suggest identifying HD-optimal designs for a variety of run sizes n, and examining a plot of HD against n. The rate at which HD approaches the upper bound will help trade off run size economy against model discrimination ability.

To illustrate this approach, we return to the example in Section 4, with 6 main effects and $\binom{6}{2} = 15$ 2fi's arising from 6 factors. HD-optimal designs for $n = 1$ to 32 trials were found. For the 32 run case, there exists a resolution VI FF design ($I = ABCDEF$) which can estimate all main effects and 2fi's. The HD values for the HD-optimal designs are plotted versus the run size in Figure 1. This example has a HD upper bound $2 \sum_{i<j} P(M_i)P(M_j) =$

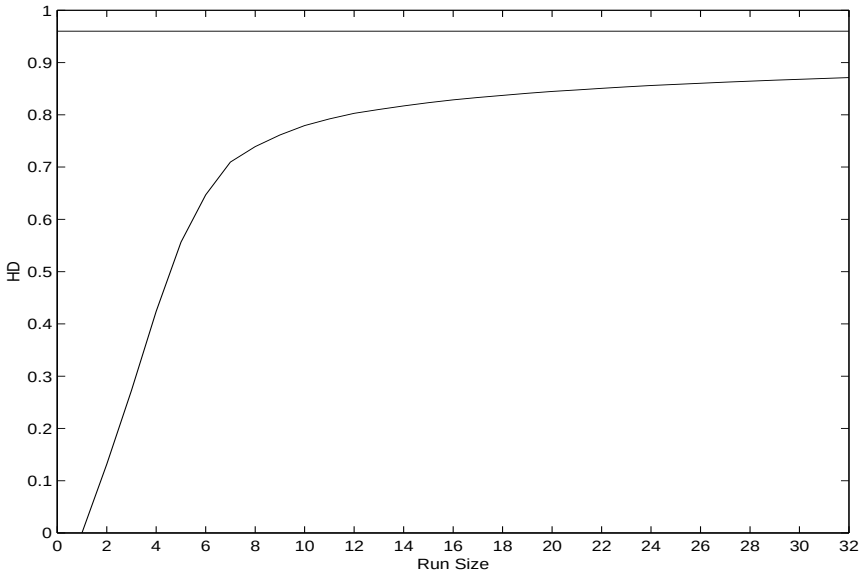

Figure 1: Optimal HD Values Versus Run-Size, for 6 factors with $p = 0.410$. The upper bound of 0.969 for HD is also plotted.

0.969, which was computed using 400 of the most probable models generated from the prior. The HD optimal 16-run design is close to this bound with HD=0.832.

This plot is an extremely useful tool in evaluating the design and we recommend its routine creation. The are two distinguishing features. First, as the run size increases from 1, the HD criterion increases quickly. With each new trial, there are more models with high prior probability that that can be estimated and distinguished. Second, after about $n = 12$ the rate of increase of the curve decreases. At this point, almost all models are estimable and the gains are due largely to increases in power to distinguish between models. Thus, in addition to providing a means to evaluate a design relative to larger designs, the plot also aids in run-size selection. Indeed, this plot suggests that we might instead consider performing a 12-run design instead.

Considering that there are $6 + 15 = 21$ effects, it may seem surprising that such small designs seem adequate. This feature emphasizes the fact that the criterion seeks to discriminate between *relevant models*, rather than estimate all possible effects. Indeed, the greatest gains in HD occur before the expected number of effects (set to be 5 here).

For comparison, we construct a similar plot for the MD criterion of MSB. Figure 2 indicates that this criterion increases almost linearly with n, regardless of the sample size. Unlike the Hellinger distance, the Kullback-Leibler information is not bounded and thus we would not expect to observe the general pattern observed in Figure 1. There appears to be less information in the MD criterion to aid in selecting a run size than in the HD criterion.

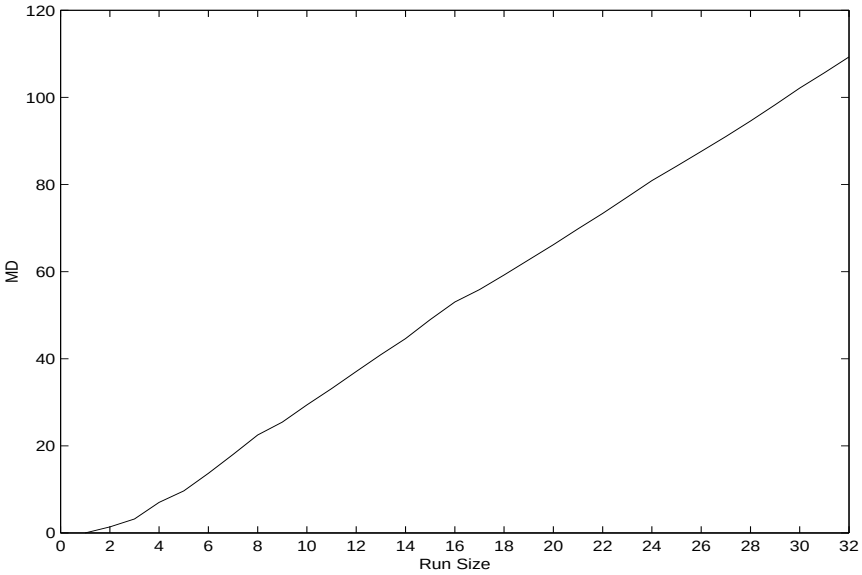

Figure 2: Optimal MD Values Versus Run-Size ($p = 0.410$)

6 More Examples

In this section we demonstrate the methodology using additional examples. We find HD-optimal designs in some common situations and finish with two specialized cases that illustrate the flexibility of the proposed approach. The models considered contain main effects and 2fi's only.

6.1 5 Factors in 12 Runs

In this example, a 12-run experiment with 5 factors is considered. As before, the prior probability that a main effect is active must first be chosen. If the expected number of active effects is 4.0, then (7) implies that $p = 0.429$. Again, a maximum of 40 models were used to evaluate HD in the exchange algorithm, 50 random restarts of the exchange algorithm were used, and the resultant candidate designs were re-evaluated using 400 models.

The HD-optimal design is shown in Appendix 2. It turns out that the HD-optimal design is the same as the 12-run Plackett-Burman design. The HD-optimal design has HD=0.830, and $2\sum_{i<j} P(M_i)P(M_j) = 0.977$ for the larger approximated model space. Thus we are fairly close to the upper bound.

To evaluate the goodness of this design relative to other HD-optimal under the prior, HD-optimal designs for n=1 to 16 trials have been found. The HD values for the HD-optimal designs are plotted versus the run size in Figure 3, indicating that 12 trials should be adequate for screening.

In practice this procedure does not guarantee that an orthogonal array will provide

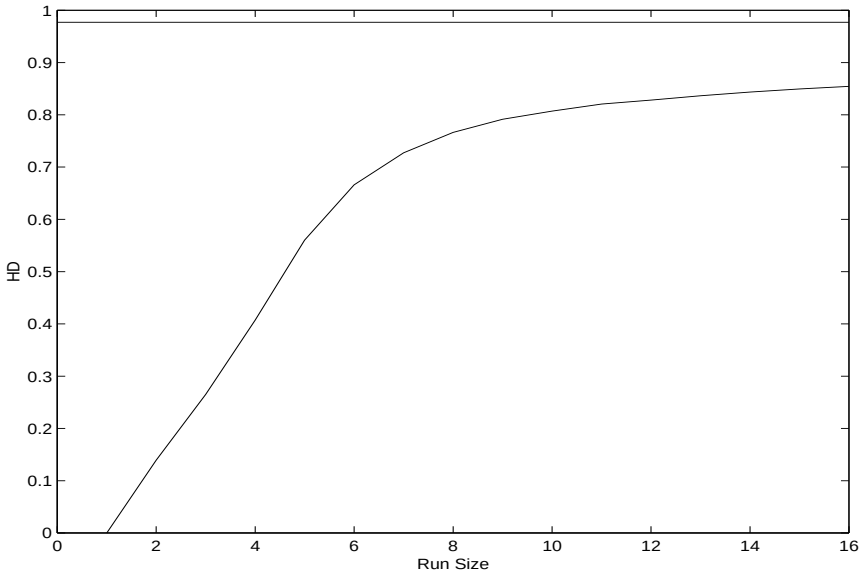

Figure 3: Optimal HD Values Versus Run-Size for a 5-factor design with $p = 0.429$. The upper bound of 0.977 for HD is also plotted.

optimal discrimination between the models of interest. However, that the HD-optimal design and the Plackett-Burman are the same suggests that the Plackett-Burman design, which has previously been suggested for screening, is well suited to that task. Perhaps a more important implication for this article is that *the HD criterion can be used to select designs of any desired run size, with some assurance that the designs are sensible.* Moreover, examining HD as a function of run size, it is easier to select an appropriate number of runs.

6.2 Designs with prior knowledge

Experiments are frequently performed with prior knowledge about the significance of some of the effects. That is, expert knowledge may indicate that some effects are likely to be unimportant. In this context, the regression effects can be classified into two broad classes: a *negligible set* of effects thought to be unimportant; and a *requirements set* of effects that should be estimated. Franklin and Bailey (1977) present a search algorithm for finding fractional factorial designs that estimate the requirements set assuming the effects in the negligible set are inert. To address the same problem, Wu and Chen (1992) introduce a graph aided method to identify fractional factorial designs that estimate the requirements set. Their approach begins with a MA fractional factorial and creates a set of linear graphs under the assumption that some effects are negligible.

Our approach is easily adapted to find designs that address this problem by giving relatively high prior probability to effects in the requirements set and relatively low prior probability to the effects in the negligible set. It is tempting to place a prior probability 0 on

the effects in the negligible set and of 1 on the effects in the requirements set. If this is done, there is really no need to run the algorithm since there is only 1 model in the model space, and one could elect to perform optimal design (e.g., D-optimal design) as an alternative. In addition, the design will not provide any robustness to the assumption that an effect is actually negligible. Instead the goal is to identify models which, with high probability contain effects from the requirements set, and with low probability contains effects from the negligible set.

As an illustration, consider the following example outlined in Wu and Chen (1992). An experiment is to be run with 11 factors ($A-K$), each at 2 levels. Prior knowledge indicates that the requirements set should contain all main effects and 2fi's among $A-F$. The negligible set contains all other 2fi's. A reasonable prior specification for the effects in the requirements set is to choose the probability of a significant main effect to $p_1 = 0.5$ and for the interactions,

$$p_{AB,i} = \begin{cases} 0.01p_1 & \text{if} \quad i = 0 \\ 0.5p_1 & \text{if} \quad i = 1 \\ p_1 & \text{if} \quad i = 2. \end{cases} \tag{12}$$

For effects in the negligible set, we suggest setting the probability of a significant main effect to $p_2 = 0.2$ and for the interactions,

$$p_{AB,i} = \begin{cases} 0.01p_2 & \text{if} \quad i = 0 \\ 0.5p_2 & \text{if} \quad i = 1 \\ p_2 & \text{if} \quad i = 2. \end{cases} \tag{13}$$

In this case, all main effects would have prior probability $p_1 = 0.5$, 2fi's among $A-F$ would use (12), and other interactions would use (13).

An advantage of our approach is that it identifies an optimal design and gives the best assignment of factors to columns of the design matrix. The aforementioned approaches will identify a design, but no obvious optimality criteria exist, and are restricted to only regular fractional factorial.

6.3 Robust Parameter Experiments

An important application in industrial statistics is robust parameter design where experimenters are interested in determining the levels of *control* factors that minimizes the impact of variation due to hard-to-control *noise* factors (e.g., see Wu and Hamada, 2000). Variance reduction is achieved through control factors that interact with noise factors. Consequently, interest lies in models that contain control-by-noise interactions.

Our flexible approach can be modified to find designs that emphasize estimation of models containing effects of primary interest. For instance, consider an experiment with 8 runs where there are 4 factors: 3 control factors ($A-C$) and 1 noise factor (N). To emphasize models that contain control-by-noise interactions, the prior probability that N and interactions involving N are made relatively large. As a consequence, models containing N are given higher prior probability and contribute more to the HD-value in (3).

Interactions involving only control factors are given the weak heredity prior (e.g., ($p_{AB,0} = 0.01, p_{AB,1} = .25, p_{AB,2} = .5$), and interactions involving N are given higher prior probability (e.g., $p_{AN,0} = 0.5, p_{AN,1} = .5, p_{AN,2} = .5$)). 50 random starts were used to identify the optimal designs.

The optimal design gives a fractional factorial design with $I = ABC$. Notice that design is able to estimate all control-by-noise interactions. Indeed, this is the optimal design for this configuration given in Wu and Hamada (2000) and Bingham and Sitter (2001).

7 Discussion

In many instances experimenters prefer orthogonal designs when there is one available. The methodology introduced in Section 3 does not guarantee that the HD-optimal design is an orthogonal array. Instead, the methodology formally incorporates A1-A3 into the design construction criterion. If an orthogonal array is preferred we can depart somewhat from our choice of model prior and adapt the methodology so that orthogonal arrays can be identified when they exist.

Orthogonal arrays have main effect columns that are orthogonal, and when used as experiment designs they emphasize the estimability of main effects. In our framework, this amounts to putting priority on models containing active main effects. This is achieved by placing relatively high prior probability ($p_A = 0.5$) on main effects and relatively small prior probability on interaction effects. We have found that for 2fi's, the following prior distribution, conditional on main effects, is suitable in most cases

$$p_{AB,i} = \begin{cases} 0.01 & \text{if} \quad i = 0 \\ 0.10 & \text{if} \quad i = 1 \\ 0.25 & \text{if} \quad i = 2. \end{cases}$$

Selecting relatively small prior probability for 2fi's implies that models with mostly main effects and a few 2fi's will impact the optimality criterion and thus the choice of optimal design.

This work points to several other possible areas of further research. The computational burden of evaluating HD for starting designs is due to diffuse nature of the model prior. Our calculations do not however take advantage of prior structure. It might be possible to group like models together, allowing for faster computation of HD.

We chose to only consider designs with factor levels of ± 1. Designs with real valued factors are important although harder to optimize.

Although we only consider linear regression with Gaussian errors, (3) is sufficiently general that it could be applied to other classes of models. A natural extension would be to consider generalized linear models. While in practice straightforward, some issues will arise, such as the need for approximations to posterior and predictive distributions. These and other predictive models are currently under investigation.

References

Bingham, D.R. & Li, W.. (2002), "A Class of Optimal Robust Parameter Designs", *Journal of Quality Technology*, to appear.

Bingham, D.R. & Sitter, R.R. (2001), "Optimal Split-Plot Designs for Robust Parameter Experiments", *Technometrics*, tentatively accepted.

Box, G. E. P. and Meyer, R. D. (1986), "An Analysis for Unreplicated Fractional Factorials", *Technometrics*, 28, 11–18.

Box, G. E. P. and Meyer, R. D. (1993). "Finding the Active Factors in Fractionated Screening Experiments", *Journal of Quality Technology*, 25, 94–104.

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", *JRSS B*, 61, 85-93.

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

Chipman, H. A. and Hamada, M. S. (1996), "Factor-based or Effect-based Modeling? Implications for Design", (Discussion of the paper "Follow-up Designs to Resolve Confounding in Multi-factor Experiments", by R. D. Meyer, D. M. Steinberg, and G. E. P. Box), *Technometrics*, 38, 317–320.

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.

Franklin, M. F. , and Bailey, R. A. (1977), "Selection of Defining Contrasts and Confounded Effects in Two-Level Experiments", *Applied Statistics*, 26, 321-326

Hall, Jr., M, (1961), "Hadamard Matrices of Order 16", *Research Summary*, 1, 21-26 Pasadena, CA., Jet Propulsion Laboratory.

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

Jones, B. and DuMouchel, W. (1996) "Discussion of "Follow-up Designs to Resolve Confounding in Multi-factor Experiments", by R. D. Meyer, D. M. Steinberg, and G. E. P. Box), ", *Technometrics*, 38, 323–326.

Meyer, R. D., Steinberg, D. M., and Box, G. E. P (1996) "Follow-up Designs to Resolve Confounding in Fractional Factorials", *Technometrics*, 38, 303–313.

Plackett, R. L. and Burman, J. P. (1945) "The Design of Optimum Multifactorial Experiments", *Biometrika*, 33, 305–325 and 328–332.

Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997) "Bayesian Model Averaging for Linear Regression Models". *Journal of the American Statistical Association*, 92, 179–191.

Sun, D. X. (1997) "Estimation Capacity and Related Topics in Experimental Design". Un-published PhD. Thesis, University of Waterloo.

Wu, C. F. J and Chen, Y.. (1993) "A graph-aided method for planning two-level experiments when certain interactions are important", *Technometrics*, 34, 162-175.

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

Wu, C. F. J and Zhu, Y. (2000) "Optimal selection of single arrays for parameter design experiments," *submitted.*.

Wynn, H. P. (1972) "Results in the Theory and Construction of D-Optimum Experimental Designs," *Journal of the Royal Statistical Society*, Series B., 34, 133–147 (with discussion 170-185).

Appendix

A Derivation of Hellinger Distance

We now sketch the derivation of the Hellinger distance between predictive densities. We refer to the model and priors defined in Section 2 to save space.

Let Y be the $n \times 1$ vector of independent observations from the linear model in (1). The model matrix, X_i is an $n \times r_i$ matrix with the first column corresponding to the intercept and the remaining $r_i - 1$ columns corresponding to the factorial effects in model M_i. The prior specification for the coefficients vector for model M_i is $\pi(\beta_i|\sigma^2) \sim MVN(0, \sigma^2\Gamma_i)$, where Γ is defined in (4). Therefore, the predictive distribution of Y is normal with mean 0 and variance $\sigma^2\Sigma_i$, where Σ_i is defined in (6).

Following the outline in Meyer, Steinberg and Box (1996), we proceed conditionally on σ^2 and integrate out σ^2 in the last step. Let f_i and f_j be the predictive densities of models M_i and M_j respectively. The Hellinger distance between the predictive densities is

$$H(f_i, f_j) = 2 - 2\int (f_i\ f_j)^{1/2} dY.$$

We now need to integrate $(f_i\ f_j)^{1/2}$ over the data to compute the Hellinger distance.

$$\int (f_i\ f_j)^{1/2} dY \quad = \quad \int \frac{\exp\{-\frac{1}{2}(Y'\frac{\Sigma_i^{-1}}{2\sigma^2}Y + Y'\frac{\Sigma_j^{-1}}{2\sigma^2}Y)\}}{(2\ \pi)^{n/2}|\sigma^2\Sigma_i|^{1/4}|\sigma^2\Sigma_j|^{1/4}} dY$$

$$
\begin{aligned}
&= \int \frac{\exp\{-\frac{1}{2}Y'(\frac{\Sigma_i^{-1}}{2\sigma^2}+\frac{\Sigma_j^{-1}}{2\sigma^2})Y\}}{(2\,\pi)^{n/2}|\sigma^2\Sigma_i|^{1/4}|\sigma^2\Sigma_j|^{1/4}}dY \\
&= \frac{\left|\left(\frac{\Sigma_i^{-1}}{2\sigma^2}+\frac{\Sigma_j^{-1}}{2\sigma^2}\right)^{-1}\right|^{1/2}}{|\sigma^2\Sigma_i|^{1/4}|\sigma^2\Sigma_j|^{1/4}} \int \frac{\exp\{-\frac{1}{2}Y'(\frac{\Sigma_i^{-1}}{2\sigma^2}+\frac{\Sigma_j^{-1}}{2\sigma^2})Y\}}{(2\,\pi)^{n/2}\left|\left(\frac{\Sigma_i^{-1}}{2\sigma^2}+\frac{\Sigma_j^{-1}}{2\sigma^2}\right)^{-1}\right|^{1/2}}dY \\
&= \frac{\left|\left(\frac{\Sigma_i^{-1}}{2\sigma^2}+\frac{\Sigma_j^{-1}}{2\sigma^2}\right)^{-1}\right|^{1/2}}{|\sigma^2\Sigma_i|^{1/4}|\sigma^2\Sigma_j|^{1/4}} \\
&= \frac{1}{|\sigma^2\Sigma_i|^{1/4}|\sigma^2\Sigma_j|^{1/4}\left|\frac{\Sigma_i^{-1}}{2\sigma^2}+\frac{\Sigma_j^{-1}}{2\sigma^2}\right|^{1/2}} \\
&= \frac{1}{|\sigma^2\Sigma_i|^{1/4}\left|\frac{1}{2}\left(\frac{\Sigma_i^{-1}}{\sigma^2}+\frac{\Sigma_j^{-1}}{\sigma^2}\right)\right|^{1/2}|\sigma^2\Sigma_j|^{1/4}} \\
&= \frac{1}{\left|\frac{1}{2}\left(\Sigma_i^{-1/2}\Sigma_j^{1/2}+\Sigma_i^{1/2}\Sigma_j^{-1/2}\right)\right|^{1/2}}
\end{aligned}
$$

Lastly, we need to integrate over σ^2. Notice that σ^2 cancels out from the derivation of the Hellinger distance. Therefore, selecting any proper prior distribution (say an inverted gamma distribution) will simply result in the Hellinger distance in (2). If an improper prior is selected, then the Hellinger distance may not be bounded above by 2. Substituting the expression in the last step for $\int(f_i\ f_j)^{1/2}dY$ in the Hellinger distance, we get the expression in (5).

B Expected Number of Active Effects

Here, the expected number of active effects in a model is derived for a general form of prior (7). There are q factors and $\binom{q}{2}$ two-way interactions being considered in the model. Let p be the probability that a specific main effect is active. The probability $p_{AB,i}$ that a specific interaction (say AB) is active, given i main effect parents are active is

$$
p_{AB,i} = \begin{cases} c_1 p & \text{if} \quad i = 0 \\ c_2 p & \text{if} \quad i = 1 \\ c_3 p & \text{if} \quad i = 2. \end{cases} \tag{14}
$$

Conditional on f active main effects, the expected number of active effects (main effects plus two-way interactions) is

$$
f + \binom{f}{2} c_3 p + f(q-f)c_2 p + \binom{q-f}{2} c_1 p \tag{15}
$$

This is because $\binom{f}{2}$ interactions will have two active parents, $f(q-f)$ interactions will have one active parent, and $\binom{q-f}{2}$ interactions will have no active parents. Expanding (15) yields an expected total number of terms (including main effects) as

$$\begin{aligned} \mathrm{E}(\#\text{effects} \quad &| \quad f \text{ active main effects}) \\ &= c_1 pq(q-1)/2 + f\left[1 + \frac{p}{2}(c_1 - c_3) + pq(c_2 - c_1)\right] \\ &+ f^2 \frac{p}{2}\left[c_1 - 2c_2 + c_3\right]. \end{aligned} \tag{16}$$

Since f is Binomial with q trials and probability of success p, we have $\mathrm{E}(f) = pq$ and $\mathrm{E}(f^2) = pq(1-p+pq)$. Taking the expectation of (16) with respect to f yields

$$\begin{aligned} \mathrm{E}(\#\text{effects}) \quad &= \quad c_1 pq(q-1)/2 + pq\left[1 + \frac{p}{2}(c_1 - c_3) + pq(c_2 - c_1)\right] \\ &\quad + pq(1-p+pq)\frac{p}{2}\left[c_1 - 2c_2 + c_3\right]. \end{aligned}$$

Further simplification yields

$$\mathrm{E}(\#\text{effects}) = pq + p\binom{q}{2}\left\{c_1 + 2p(c_2 - c_1) + p^2(c_1 - 2c_2 + c_3)\right\} \tag{17}$$

For specified values of q, c_1, c_2, c_3 and an expected number of effects, this cubic in p can easily be solved for p. Note that the expected number of main effects is pq, the first term of (17).

C HD Optimal Designs

16-Run HD-Optimal Design

-1	-1	-1	-1	-1	1
-1	-1	-1	1	1	-1
-1	-1	1	-1	1	1
-1	-1	1	1	-1	-1
-1	1	-1	1	-1	-1
-1	1	-1	1	1	1
-1	1	1	-1	-1	1
-1	1	1	-1	1	-1
1	-1	-1	-1	-1	-1
1	-1	-1	-1	1	1
1	-1	1	1	-1	1
1	-1	1	1	1	-1
1	1	-1	-1	1	-1
1	1	-1	1	-1	1
1	1	1	-1	-1	-1
1	1	1	1	1	1

12-Run HD-Optimal Design

+1	+1	-1	+1	+1
-1	+1	+1	-1	+1
+1	-1	+1	+1	-1
-1	+1	-1	+1	+1
-1	-1	+1	-1	+1
-1	-1	-1	+1	-1
+1	-1	-1	-1	+1
+1	+1	-1	-1	-1
+1	+1	+1	-1	-1
-1	+1	+1	+1	-1
+1	-1	+1	+1	+1
-1	-1	-1	-1	-1