UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS

School of Mathematics

Bayesian Experimental Design for Model Discrimination

by

Andrew David Rose

Thesis for the Degree of Doctor of Philosophy

June, 2008

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS

SCHOOL OF MATHEMATICS

Doctor of Philosophy

BAYESIAN EXPERIMENTAL DESIGN FOR MODEL DISCRIMINATION

by Andrew David Rose

This thesis is concerned with the situation in which there are several competing linear models for describing the dependence of a response on a set of explanatory variables. The aim of the thesis is to produce methodology by which experimental designs may be selected that allow discrimination between the possible models and enable a model to be chosen that is as close as possible to the 'true' model. A Bayesian decision theoretic framework will be used for model selection and choosing an experimental design. The Bayesian approach allows prior information from previous experimentation to be used in the selection of a design and provides a means by which a model may be selected from a set of competing models.

In this thesis, the Penalised Model Discrepancy (PMD) criterion for selecting an experimental design is introduced. The criterion is first applied to the situation of screening experiments, where little prior information is available. Good designs under the PMD criterion are found for several model spaces which may be used for screening experiments. The MD, HD and F criteria are existing Bayesian criteria from the literature for selecting experimental designs for model discrimination; a comparison between these and the PMD is made via examples and a simulation study. The sensitivity of the PMD criterion to the choice of hyperparameters of the prior distribution is also investigated. The PMD criterion is then applied to the selection of follow-up runs after an initial experiment, using examples from the literature. For one example, a comparison is again made to the F, MD and HD criteria. For another example, the effect of the choice of initial design on the follow-up runs selected is investigated. Follow-up runs for a tribology experiment carried out in the School of Engineering Sciences at the University of Southampton were chosen using the PMD criterion. The results and analysis of this experiment are presented, as well as details of how the follow-up runs were chosen.

In some situations, especially when interaction terms are considered, the space of possible models can become very large. As a consequence, evaluating the PMD objective function can become very computationally expensive. Methods for reducing the computational burden of evaluating the PMD objective function are investigated, and used to select good designs for large model spaces. Methodology for improving the accuracy of the evaluation of the HD and MD objective functions for large model spaces is also given.

Contents

List of Figures

List of Tables

Acknowledgements

Thanks go to my supervisors, Professor Susan Lewis and Dr David Woods for all their help, patience and enthusiasm during the course of my studies. I would also like to thank Professor Jon Forster for his assistance.

Thanks to Dr Ramkumar Penchaliah, Dr Terry Harvey and Professor Robert Wood of the Engineering Materials and Surface Engineering group at the University of Southampton for supplying the tribology example used in this thesis.

I would like to thank all the other statistics research students, especially Roger Gill and Jeff Samuel, for keeping me entertained throughout my time at Southampton.

Finally, I would like to thank my family for all their support and encouragement over the course of my PhD.

Chapter 1

Introduction and Overview

In this chapter the concepts of the linear model, design of experiments and Bayesian inference are introduced. We then review some of the existing literature on Bayesian experimental design, in particular, approaches to the problem of model discrimination. The types of models and prior distributions for their parameters to be used throughout this thesis are described. Finally, the overall aims and specific objectives of the work are stated, and an overview of the thesis is given.

1.1 The Linear Model and Bayesian Inference

1.1.1 The Linear Model

Suppose that we wish to model the dependence of a $n \times 1$ response vector $\mathbf{Y}$ on a set of p explanatory variables. Let $\mathbf{X}$ be an $n \times p$ matrix, where $\mathbf{X}_{ij}$ is the value of the jth explanatory variable at the ith data point. Then a linear model for the dependence of $\mathbf{Y}$ on $\mathbf{X}$ is given by:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\beta}$ is a p-dimensional vector of regression parameters, and $\boldsymbol{\epsilon} \overset{iid}{\sim} N(0, \sigma^2)$ is an n-dimensional vector of random error terms. The mean responses given by the model at the n data points are held in the vector $\mathbf{X}\boldsymbol{\beta}$. The differences between the means and

observed responses is modelled by the terms in $\boldsymbol{\epsilon}$.

1.1.2 Bayesian Inference

In classical inference, the parameters Θ associated with a statistical model $f(\mathbf{y}|\boldsymbol{\theta})$ are treated as having a fixed, albeit unknown, value. Once data are obtained, $\boldsymbol{\theta}$ may be estimated (by using maximum likelihood estimation, for example), and point hypotheses about the parameters may be tested. In the Bayesian framework, $\boldsymbol{\theta}$ is instead treated as a random variable. Before data are available, a prior distribution $f(\boldsymbol{\theta})$ represents the current beliefs and knowledge about the parameters. Upon observation of the data $\mathbf{y}$, this distribution is updated to a posterior distribution $f(\boldsymbol{\theta}|\mathbf{y})$ via Bayes theorem:

$$f(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta})}{\int\limits_{\Theta} f(\boldsymbol{\theta})f(\mathbf{y}|\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{1.1}$$

For parameters taking discrete values, the integral in equation (1.1) is replaced by a summation.

1.1.3 Bayesian Inference for Linear Models

In this thesis we will consider the situation where there are $M \geq 2$ competing linear models of the form

$$m_i : \mathbf{Y} = \boldsymbol{X_i}\boldsymbol{\beta}_i + \boldsymbol{\epsilon_i}, \quad i = 1, \ldots, M, \tag{1.2}$$

where $\boldsymbol{\beta}_i$ is a p_i-dimensional vector of regression parameters, $\mathbf{X}_i$ is the $n \times p_i$ model matrix for model m_i and $\varepsilon_i \overset{iid}{\sim} N(0, \sigma_i^2)$ is a n-dimensional vector of random error terms. In a Bayesian framework, each model is given a prior probability $P(m_i)$ and prior distributions are assigned to the model parameters $\boldsymbol{\beta}_i$ and σ_i^2. Bayes' theorem is used to obtain posterior model probabilities and parameter distributions after data $\mathbf{y}$ have been observed.

A commonly used prior for the joint distribution of $\boldsymbol{\beta}_i$ and σ_i^2 is a Normal Inverse-Gamma distribution. This is a conjugate distribution for the linear model described in Section 1.1.1. Details of this distribution may be found in O'Hagan and Forster (2004), chapter 11. We assign an inverse-gamma prior distribution to σ_i^2, which

does not depend on i. That is,

$$f(\sigma_i^2) = \frac{(a/2)^{\frac{d}{2}}}{\Gamma(d/2)}(\sigma_i^2)^{-(\frac{d+2}{2})}\exp(-a/(2\sigma_i^2)), \tag{1.3}$$

or $\frac{1}{\sigma_i^2} \sim \text{Gamma}(\frac{a}{2}, \frac{d}{2})$, where a and d are hyperparameters. Given model m_i and the value of σ_i^2, the conditional density of $\boldsymbol{\beta}_i$ is $N(\boldsymbol{\mu}_i, \sigma_i^2\mathbf{V}_i)$, where $\boldsymbol{\mu}_i$ and $\mathbf{V}_i$ are the p_i-dimensional prior mean and the $p_i \times p_i$ prior variance-covariance matrix, respectively, for model m_i. The posterior distributions are of the same family:

$$\frac{1}{\sigma_i^2}|\mathbf{Y}, m_i \sim \text{Gamma}\left(\frac{a_i^*}{2}, \frac{d^*}{2}\right), \tag{1.4}$$

$$\boldsymbol{\beta}_i|\sigma_i^2, \mathbf{Y}, m_i \sim N(\boldsymbol{\mu}_i^*, \sigma\mathbf{V}_i^*) \tag{1.5}$$

where

$$\begin{aligned} \mathbf{V}_i^* &= (\mathbf{V}_i^{-1} + \mathbf{X}_i'\mathbf{X}_i)^{-1} \\ \boldsymbol{\mu}_i^* &= \mathbf{V}_i^*(\mathbf{V}_i^{-1}\boldsymbol{\mu}_i + \mathbf{X}_i'\mathbf{y}) \\ a_i^* &= a + \boldsymbol{\mu}_i'\mathbf{V}_i^{-1}\boldsymbol{\mu}_i + \mathbf{y}'\mathbf{y} - (\boldsymbol{\mu}_i^*)'(\mathbf{V}_i^*)^{-1}\boldsymbol{\mu}_i^* \\ d^* &= d + n. \end{aligned} \tag{1.6}$$

Model Probabilities

For $i = 1, \ldots, M$ a probability that model m_i is the true model is assigned. This prior probability is denoted by $P(m_i)$. These probabilities may be updated to posterior probabilities using Bayes' theorem:

$$P(m_i|\mathbf{y}) = \frac{P(m_i)f(\mathbf{y}|m_i)}{\sum_j P(m_j)f(\mathbf{y}|m_j)} \tag{1.7}$$

where, under (1.4) and (1.5), the marginal likelihood of the data $\mathbf{y}$ given model m_i is

$$f(\mathbf{y}|m_i) = \frac{|\mathbf{V}_i^*|^{1/2}a^{d/2}\Gamma(d^*/2)}{|\mathbf{V}_i|^{1/2}(a_i^*)^{d^*/2}\Gamma(d/2)\pi^{n/2}}. \tag{1.8}$$

These model probabilities can be used to calculate the marginal posterior probability that a factor is active. The definition of an 'active' factor or effect differs across the literature, see, for example, Box and Meyer (1986). In this thesis an active effect is defined as being one of a non-negligible magnitude; an active factor is one that is involved in one or more active effects. The marginal posterior probability that factor j is active is given by $\sum_{i:j\in m_i} P(m_i|\mathbf{y})$.

Choice of Hyperparameters

The values used for the hyperparameters $a, d, \boldsymbol{\mu}_i$ and $\mathbf{V}_i$ should reflect prior beliefs about the parameters $\boldsymbol{\beta}_i$ and σ_i. If little prior information is available, an noninformative prior should be used. If data are available from a previous experiment, as in the examples in Chapters 4 and 5, then a natural method of finding a prior distribution for use in future work is to start with an noninformative prior, then update it to a posterior distribution given the data to produce a more informative prior for subsequent experiments.
Using $\boldsymbol{\mu}_i = 0$ is a neutral choice of prior mean for $\boldsymbol{\beta}$, and reflects the fact that we do not know the direction of the effects, so view each component of $\boldsymbol{\beta}$ as equally likely to be positive or negative. If we have no prior information about correlations between the distributions of the components of $\boldsymbol{\beta}_i$, we should make $\mathbf{V}_i$ a diagonal matrix, as the parameters are viewed as independent *a priori*. As we do not know the magnitude of $\boldsymbol{\beta}_i$, we might assume that we should make the numbers on the diagonal of $\mathbf{V}_i$ as large as possible, corresponding to a 'flat' prior on $\boldsymbol{\beta}$. However, if we do this, smaller models will be favoured too strongly. Suppose $\mathbf{V}_i = \lambda \mathbf{I}_{p_i}$ where p_i is the number of terms in model m_i. Then, as $\lambda \to \infty$, $\mathbf{V}_i^* \to (\mathbf{X}_i'\mathbf{X}_i)^{-1}$ and $|\mathbf{V}_i| = \lambda^{p_i} \to \infty$. Hence using (1.8), $f(\mathbf{y}|m_i)/f(\mathbf{y}|m_j) \to 0$ if $p_i < p_j$. So for large λ, the posterior probability will automatically be greatest for the smallest models. This effect is an example of Lindley's paradox (Jeffreys, 1939), which shows that a diffuse prior can lead to a Bayesian analysis favouring a smaller model when a classical likelihood ratio test would have rejected this in favour of a larger model. To stop this from happening, we can either use $\lambda_i = \lambda_0^{1/p_i}$ for some constant λ_0, or use $\lambda = 1$ for all models. As the intercept term is present in all models, it is possible to put a large prior variance on this term without causing problems with the relative probabilities of models with different numbers of terms. We also need

to select a and d, the hyperparameters of the inverse-gamma prior distribution of σ_i^2. The mean of the distribution in (1.3) is $a/(d-2)$ for $d > 2$ and its variance is $2a^2/\left((d-2)^2(d-4)\right)$ for $d > 4$. Hence, if we want a prior distribution with finite mean and variance for σ^2 we should use $d > 4$ and choose a and d to give a reasonable mean and variance for the distribution of σ^2. The effects of changing a and d are discussed in Section 3.3.4. Alternatively, we can use an improper prior distribution for σ^2. For example, using $a = 0, d = 0$ gives the improper prior distribution $f(\sigma^2) \propto \sigma^{-2}$.

1.2 Design of Factorial Experiments

Experimental design to investigate several variables involves selecting a set of combinations of values, or levels, of factors at which to perform experimental runs. When the factors are qualitative or are limited to a fixed number of levels, factorial or fractional factorial designs are often used. If f factors are investigated in the experiment, with factor i having k_i levels, then a full factorial design consists of all $\prod_{i=1}^{f} k_i$ combinations of the factor levels. A fractional factorial design uses some subset of the available runs from a full factorial design. A regular fractional factorial design is one that is generated by a defining relation, which can be interpreted as a set of equations that must be satisfied by the factor levels in every row of the design. For factors with two levels, the fractional factorial design generated by a defining relation consists of all possible combinations of factor levels ± 1 that satisfy the defining relation. In these designs, the correlations between factorial effects (main effects and interactions) are either 0 (uncorrelated) or 1 (total aliasing). There is no partial aliasing between factorial effects; see, for example, Box, Hunter and Hunter (2005), Chapter 6.

Example For five factors, $A \ldots E$, each at two levels, denoted by -1, 1, a quarter (2^{5-2}) regular fraction is defined by $I = ABCD = ABE$. This fraction has the 8 runs shown in Table 1.1.

In this example, the mean is aliased with the four-factor $ABCD$ interaction, and the three-factor ABE interaction. The main effects are totally aliased with two and three factor interactions, for example A with BE and BCD. Some two-factor interactions are also aliased together, for example AB with CD. The resolution of a design is the length

Table 1.1: The 2^{5-2} design defined by $I = ABCD = ABE$.

A	B	C	D	E
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

of the shortest word in its defining relation. For example, the design shown in Table 1.1 has resolution III because the shortest word in the defining relation (ABE) contains three factors.

Regular factorial and fractional factorial designs of resolution III or higher are examples of main effects orthogonal designs. These are designs in which the columns containing the values of the factors are all orthogonal to each other. An advantage of using a main effects orthogonal design is that the least squares estimators of all the main effects are uncorrelated.

Irregular fractions have no defining relation and may allow estimation of main effects and lower order interactions in fewer runs than a regular fractional factorial design. The most well known are the designs of Plackett and Burman (1946) which are used for estimating main effects when interactions are believed to be negligible. For 2-level factors, Plackett and Burman designs may be constructed for any number of runs that is a multiple of four, unlike full and fractional factorial designs, for which the number of runs must be a power of two. A complete catalogue of two-level main effects orthogonal designs in 12, 16 and 20 runs was provided by Sun, Li and Ye (2002), and was found by computer search.

1.3 Design for Model Discrimination

Consider the situation in which there are M competing linear models of the form given in equation (1.2). The problem is how to choose, by experiment, a model m^* from the set of models $\mathcal{M}$ under consideration which provides the 'best' approximation to the observed response. Throughout this thesis, this best model is called the 'true' or

'correct' model.

Several criteria have been proposed for selecting designs for the purpose of discriminating between a set of possible models to describe the relationship between a response and a set of explanatory variables. In this subsection we briefly describe the main non-Bayesian approaches to this problem. A review of Bayesian criteria will be given in Section 1.3.1. The model discrimination capability of the search designs of Srivastava (1975) is measured by a criterion called resolving power. For discriminating between $M = 2$ models, the T-optimality criterion was introduced by Atkinson and Federov (1975a). This criterion is based on maximising the residual sum of squares for the alternative model, given the known 'true' model. Atkinson and Federov (1975b) generalised this criterion to allow several competing models. A different approach using criteria based on the Subspace Angle and Expected Prediction Difference, was recently introduced by Jones, Li, Nachtsheim and Ye (2007). These criteria aimed to select designs where differences between the fitted values from the M models are as large as possible. The subspace angle of a pair of models is a measure of the closeness between the spaces spanned by their model matrices for a given design. To select a design, Jones *et al.* (2007) defined two criteria based on the subspace angle: either to maximise the minimum out of the subspace angles between pairs of models in $\mathcal{M}$, or to maximise the average subspace angle over all pairs of models in $\mathcal{M}$. The expected prediction difference between two models is the expectation of the squared magnitude of the difference between the fitted values of the two models, where the response is normalised to lie on the unit sphere. As with the subspace angle, Jones *et al.* (2007) defined two criteria - the minimum and average expected prediction difference over all pairs of possible models. Agboto, Li and Nachtsheim (2006) evaluated the designs in the catalogue of Sun *et al.* (2002) under a variety of model discrimination criteria, including the expected prediction difference.

1.3.1 Bayesian Experimental Design for Model Discrimination

A comprehensive review of Bayesian experimental design was given by Chaloner and Verdinelli (1995). In a Bayesian framework, experimental design may be approached by using decision theory. There are two decisions to be made: firstly, the choice of design to

use and secondly an inferential decision based on the outcome of the experiment, such as the choice of a model or estimators of the model parameters. Lindley (1972) followed this approach, and suggested selecting a design d to maximise the expected utility of the final inferential decision δ chosen from set of possible decisions Δ. That is, to choose d, from the set of designs under consideration $\mathcal{D}$, to maximise

$$U(d) = \int_{\mathcal{Y}} \max_{\delta \in \Delta} \int_{\Theta} U(\delta, \boldsymbol{\theta}, d, \mathbf{y}) f(\theta|\mathbf{y}, d) f(\mathbf{y}|d) d\boldsymbol{\theta} d\mathbf{y}, \tag{1.9}$$

where $\boldsymbol{\theta}$ is the vector of unknown model parameters. Using utility functions based on the accuracy of parameter estimates or the gain in Shannon information on the parameters leads to the Bayesian A- and D- optimality criteria respectively (see Chaloner and Verdinelli (1995) for references). A more model-robust approach was implemented by DuMouchel and Jones (1994) using Bayesian D-optimality and a particular choice of prior where a distinction is made between terms which must be estimated in the model (primary) and those which may possibly need to be included (potential).

There are several existing approaches to producing designs for the purpose of model discrimination in a Bayesian framework. Box and Hill (1967) introduced the D (for discrimination) criterion, based on maximising the expected Kullback-Leibler distance (see, for example, Chaloner and Verdinelli (1995)) between the predictive densities of pairs of competing models. We refer to this criterion as MD, following Meyer, Steinberg and Box (1996), who used this criterion for the selection of follow-up designs.

The MD criterion selects a design that maximises

$$MD = \sum_{0 \le i \ne j \le m} P(m_i) P(m_j) \int f(\mathbf{Y}|m_i) \log \left(\frac{f(\mathbf{Y}|m_i)}{f(\mathbf{Y}|m_j)} \right) d\mathbf{Y}. \tag{1.10}$$

The objective function (1.10) is the expected Kullback–Leibler distance between the prior predicted densities of models m_i and m_j. For a normal inverse-gamma prior distribution, the prior predictive distribution under model m_i of the response in future experiments with model matrix $\mathbf{X}_i$ is $\mathbf{Y} \sim N(\hat{\mathbf{Y}}_i, a_i \Sigma_i)$ where $\hat{\mathbf{Y}}_i = \mathbf{X}_i \boldsymbol{\mu}_i$, and $\Sigma_i = \mathbf{I} + \mathbf{X}_i \mathbf{V}_i \mathbf{X}_i'$ (see O'Hagan and Forster 2004). Meyer *et al.* (1996) showed that, for a normal inverse-gamma prior, the MD objective function can be written as

$$MD = \frac{1}{2}\sum_{0\le i\neq j\le m} P(m_i)P(m_j) \times \left\{-n + \mathrm{tr}\left(\boldsymbol{\Sigma}_j^{-1}\boldsymbol{\Sigma}_i\right) + d\left[(\hat{\mathbf{Y}}_i - \hat{\mathbf{Y}}_j)'\boldsymbol{\Sigma}_j^{-1}(\hat{\mathbf{Y}}_i - \hat{\mathbf{Y}}_j)/a_i\right]\right\}. \tag{1.11}$$

A further development of the work of DuMouchel and Jones (1994), is the F criterion suggested by Jones and DuMouchel (1996) for use in model discrimination. It aims to find a design that maximises

$$F = |\mathbf{V}_0^{-1} + \mathbf{X}_f'\mathbf{X}_f||\mathbf{V}_0|,$$

where

$$\mathbf{V}_0 = \sum_i P(m_i)\mathbf{V}_{ei} + \sum_i P(m_i)(\boldsymbol{\mu}_{ei} - \boldsymbol{\mu}_0)(\boldsymbol{\mu}_{ei} - \boldsymbol{\mu}_0)',\ \boldsymbol{\mu}_0 = \sum_i P(m_i)\boldsymbol{\mu}_{ei}.$$

For a model m_i, the matrix $\mathbf{V}_{ei}$ is the prior variance-covariance matrix expanded to include all possible model terms, with zero entries in rows and columns that represent terms not included in the model. Likewise, $\boldsymbol{\mu}_{ei}$ is $\boldsymbol{\mu}_i$, the prior mean for model m_i, with extra zero entries for terms not in model m_i and $\mathbf{X}_f$ is the model matrix for the full model containing all possible terms. The criterion is related to the Bayesian D-optimality criterion for a 'super-model' that contains all possible terms, which is to maximise $|\mathbf{V}_f^{-1} + \mathbf{X}_f'\mathbf{X}_f|$, where $\mathbf{V}_f$ is the prior variance-covariance matrix for the model containing all possible terms.

The HD criterion of Bingham and Chipman (2007) is similar to the MD but based on Hellinger distances, which gives the advantage of an upper bound on the objective function. The HD criterion is to maximise

$$HD = \sum_{i<j} P(m_i)P(m_j)H(f_i, f_j)$$

where f_i is the prior predictive distribution of new observations under model m_i. The Hellinger distance between two densities f_i, f_j is given by

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

For a linear model, we can write $\int (f_i f_j)^{1/2} dY$ in closed form as

$$\frac{\exp\left(\left(\frac{\hat{\mathbf{Y}}_i'\boldsymbol{\Sigma}_i^{-1}\hat{\mathbf{Y}}_i}{2a_i} + \frac{\hat{\mathbf{Y}}_j'\boldsymbol{\Sigma}_j^{-1}\hat{\mathbf{Y}}_j}{2a_j}\right) - \left(\frac{\boldsymbol{\Sigma}_i^{-1}\hat{\mathbf{Y}}_i}{2a_i} + \frac{\boldsymbol{\Sigma}_j^{-1}\hat{\mathbf{Y}}_j}{2a_j}\right)' \left(\frac{\boldsymbol{\Sigma}_i^{-1}}{2a_i} + \frac{\boldsymbol{\Sigma}_j^{-1}}{2a_j}\right)^{-1} \left(\frac{\boldsymbol{\Sigma}_i^{-1}\hat{\mathbf{Y}}_i}{2a_i} + \frac{\boldsymbol{\Sigma}_j^{-1}\hat{\mathbf{Y}}_j}{2a_j}\right)\right)}{\left|\left(\frac{\boldsymbol{\Sigma}_i^{-1}}{2a_i} + \frac{\boldsymbol{\Sigma}_j^{-1}}{2a_j}\right)\right|^{1/2} |a_i\boldsymbol{\Sigma}_i|^{1/4} |a_j\boldsymbol{\Sigma}_j|^{1/4}}.$$

For a derivation of this result when the prior means are non-zero, see Appendix A.

1.4 Aims and structure of the thesis

The overall aim of this thesis is to develop methodology to select experimental designs for the purpose of model discrimination within a Bayesian framework. This specific objectives are:

- To formulate a design selection criterion which reflects our objectives in model discrimination, and which is consistent with a Bayesian decision theoretic approach to model selection.
- To investigate methodology for efficient evaluation of experimental designs under this criterion, and to enable us to search for good designs under the proposed criterion.
- To apply the criterion to examples involving different sets of possible models, with differing levels of prior information. Specifically, we will apply the criterion to the selection of screening designs, follow-up runs and to a real example.
- To compare the use of the criterion to that of other criteria from the literature for a variety of examples.

In Chapter 2, we define the Penalised Model Discrepancy (PMD) criterion for design selection and describe methodology for evaluating our objective function and searching for good designs. In Chapter 3, we consider the problem of choosing designs for screening experiments for a variety of model spaces. We compare the use of the criterion to the MD, HD and F criteria from the literature. In Chapter 4, we apply the PMD

criterion to an example from the literature, where follow-up runs must be chosen to build on information from an initial experiment. We again compare our criterion to the MD, HD and F criteria. In Chapter 5, we apply similar ideas to the selection of follow-up runs for an experiment in tribology. Finally, in Chapter 6 we investigate approaches to the high computational burden that arises when large model spaces are used. Conclusions, a summary and ideas for further work are presented in Chapter 7.

Chapter 2

The Penalised Model Discrepancy Criterion

In this chapter we describe the approach taken to Bayesian model selection. This leads to the formulation of an objective function for comparing designs and a criterion for design selection, given in Section 2.2. The implementation of our criterion is discussed in Section 2.3.

2.1 A Decision Theoretic Approach to Model Selection

Suppose that data, $\mathbf{y}$, are obtained from an experiment that uses design $d_n \in \mathcal{D}(n)$, where $\mathcal{D}(n)$ is the set of all n-point exact designs. Then prior model probabilities $P(m_i)$ are updated to give posterior model probabilities $P(m_i|\mathbf{y}, d_n)$ via (1.7) and parameter distributions $f(\boldsymbol{\beta}_i, \sigma^2|\mathbf{y}, d_n, m_i)$ via (1.1).
We define a loss function $L(i, j)$ to be the loss incurred in selecting model m_i when model m_j is true in the sense described in Section 1.3. We choose a model m_i which minimises the expected loss

$$E(L(i,j)|\mathbf{y}, d_n) = \sum_j P(m_j|\mathbf{y}, d_n)L(i,j), \tag{2.1}$$

which is known as the Bayes risk; see for example, Berger (1985). The simplest form of

loss function is a loss of 1 for choosing an incorrect model, and a loss of zero for choosing the correct model, i.e.

$$L(i,j) = 0 \text{ if } i = j \text{ and } L(i,j) = 1 \text{ otherwise.} \tag{2.2}$$

This loss function leads us to always select the model with highest posterior probability. The loss function used to select a model in this thesis is constructed by assigning a loss of 1 to every model term from the true model that is not present in the selected model, and a loss of c for every extra term in the selected model that is not in the true model. This loss function is defined by

$$L(i,j) = \mid \mathcal{S}_j \setminus \mathcal{S}_i \mid + c \mid \mathcal{S}_i \setminus \mathcal{S}_j \mid, \tag{2.3}$$

where $\mathcal{S}_i$ is the set of terms in the selected model m_i and $|\mathcal{S}|$ denotes the size of the set $\mathcal{S}$. We call this the *Penalised Model Discrepancy* (PMD) loss function. An advantage of the PMD loss function is that it has simple interpretation when $c = 1$ as the *actual* number of terms by which the chosen model differs from the true model.

2.1.1 Formulation of the Expected Loss

As described in Section 2.1, a model is selected that minimises the expected loss. For the loss function (2.3), the expected loss from selecting model m_i, given the data $\mathbf{y}$ obtained using design d_n, is

$$E_j\left(L(i,j)|\mathbf{y}, d_n\right) = \sum_{j=1}^{M} p(m_j|\mathbf{y}) L(i,j) = \sum_{j=1}^{M} p(m_j|\mathbf{y}) \mid \mathcal{S}_j \setminus \mathcal{S}_i \mid + c \sum_{j=1}^{M} p(m_j|\mathbf{y}) \mid \mathcal{S}_i \setminus \mathcal{S}_j \mid . \tag{2.4}$$

We now reformulate (2.4) as a function of the probabilities of individual terms being present in the model. This reformulation will be used in later chapters to gain understanding of which particular terms are included in the chosen model.
The number of terms in m_j that are not in model m_i can be expressed as :

$$\mid \mathcal{S}_j \setminus \mathcal{S}_i \mid = \sum_{\ell \in \mathcal{S}} I(\ell \in \mathcal{S}_j) I(\ell \notin \mathcal{S}_i),$$

where $\mathcal{S} = \bigcup_{i=1}^{M} \mathcal{S}_i$ and $I(\ell \in \mathcal{S}_j) = 1$ if term ℓ is in model m_j, and 0 otherwise $(j = 1, \ldots, M)$. Then

$$\begin{aligned}\sum_{j=1}^{M} p(m_j|\mathbf{y}) \mid \mathcal{S}_j \setminus \mathcal{S}_i \mid &= \sum_{j=1}^{M} p(m_j|\mathbf{y}) \sum_{\ell \in \mathcal{S}} I(\ell \in \mathcal{S}_j) I(\ell \notin \mathcal{S}_i) \\ &= \sum_{\ell \in \mathcal{S}} I(\ell \notin \mathcal{S}_i) \sum_{j=1}^{M} p(m_j|\mathbf{y}) I(\ell \in \mathcal{S}_j) \\ &= \sum_{\ell \in \mathcal{S}} I(\ell \notin \mathcal{S}_i) p(\ell|\mathbf{y}).\end{aligned}$$

Here, $p(\ell|\mathbf{y}) = \sum_{j=1}^{M} p(m_j|\mathbf{y}) I(\ell \in \mathcal{S}_j)$ is the posterior probability that model term ℓ is present in the true model; see, for example, Box and Meyer (1986).

Applying a similar argument to the second summation in (2.4) gives

$$\sum_{j=1}^{M} p(m_j|\mathbf{y}) \mid \mathcal{S}_i \setminus \mathcal{S}_j \mid = \sum_{\ell \in \mathcal{S}} I(\ell \in \mathcal{S}_i) [1 - p(\ell|\mathbf{y})]. \tag{2.5}$$

Hence from (2.5) and (2.5), the expected loss is

$$E(L(i,j)|\mathbf{y}, d_n) = \sum_{j=1}^{M} p(m_j|\mathbf{y}) L(i,j) = \sum_{\ell \in \mathcal{S}} \{I(\ell \notin \mathcal{S}_i) p(\ell|\mathbf{y}) + c I(\ell \in \mathcal{S}_i) [1 - p(\ell|\mathbf{y})]\}. \tag{2.6}$$

For some choices of model spaces, this equation can provide insight into how the model selected relates to the posterior probabilities of the individual model terms. Define $\mathcal{T} \subseteq \mathcal{S}$ to be the set of terms such that

$$\ell \in \mathcal{T} \Leftrightarrow p(\ell|\mathbf{y}) > c [1 - p(\ell|\mathbf{y})],$$

for a chosen penalty c, that is, $p(\ell|\mathbf{y}) > \frac{c}{1+c}$. Then, provided that

$$\mathcal{T} = \mathcal{S}_i \text{ for some } 1 \leq i \leq M, \tag{2.7}$$

$\mathcal{S}_i$ will contain the terms of a model that minimises the expected loss (2.6). The value of this loss is then

$$\sum_{\ell \in \mathcal{S}} \min \{p(\ell|\mathbf{y}), c [1 - p(\ell|\mathbf{y})]\}. \tag{2.8}$$

The condition (2.7) applies for any prior distribution which gives non-zero prior probability to models composed of any subset of terms of $\mathcal{S}$, for example, relaxed weak heredity, see Chipman (1996), and Chapters 3 and 4 for definition and discussion.

2.2 Design Selection

The most general formulation of the criterion we use to select a design is to minimise the average expected loss of the model chosen, given the data, that is to minimise

$$E_{\mathbf{Y}|d_n}\left(\min_i E_j\left[L(i,j)|\mathbf{y},d_n\right]\right). \tag{2.9}$$

This expression is equivalent to equation (1.9) of Chapter 1 (see, also, Chaloner and Verdinelli, 1995) and can be applied to any choice of loss function. Throughout this thesis we restrict attention to the criterion with loss function (2.3), and this leads to our new criterion for obtaining designs.

Definition: The *Penalised Model Discrepancy* (PMD) optimal design over the set $\mathcal{D}_n$ of n-point designs is

$$d_n^* = \underset{\mathcal{D}_n}{\operatorname{argmin}}\ E_{\mathbf{Y}|d_n}\left(\min_i E_j\left[L(i,j)|\mathbf{y},d_n\right]\right), \tag{2.10}$$

where $L(i,j)$ is defined in equation (2.3).

An advantage of this criterion, compared with alternatives discussed in later chapters, is that it has straightforward interpretation: For the special case when $c = 1$, the expected loss in (2.3) may be is interpreted as the expected number of terms by which the chosen model differs from the true model. If $c \neq 1$, the objective function is a weighted sum of the expected number of additional terms and missing terms compared to the true model. The use of a value of $c < 1$ means that the inclusion of extra terms in the model is penalised less harshly than the omission of important terms, as a model with extra terms is still a 'correct' model for predictive purposes. This enables us to retain model terms if we do not have strong evidence that they are inactive, which may be necessary, particularly in the early stages of experimentation. Conversely, use of $c > 1$ would encourage the selection of a simpler model, possibly missing more active terms. This is further discussed and illustrated in Chapter 3.

2.3 Implementation

In this section we describe how the PMD criterion has been implemented in software, which is written in C. The PMD objective function cannot be calculated analytically, hence simulation is used to evaluate it. Further, as an exhaustive evaluation of all

possible designs cannot be made, a search algorithm is used to find good designs. For other computation required for this thesis, we have used R (R Development Core Team 2008).

2.3.1 Evaluation of the Objective Function

In each simulation, we generate a data set from the prior predictive distribution of $\mathbf{Y}$, calculate the posterior model probabilities, and select the model with the lowest expected loss. The losses incurred due to the models selected are averaged over a large number, s, of simulations to evaluate the objective function. The algorithm used to evaluate the objective function is:

1. Calculate the parts of the posterior model probability that do not depend on $\mathbf{y}$. We have, from Chapter 1, Section 1.1.3 that this probability involves $\mathbf{y}$ only through the term a_i^*, given by

$$a_i^* = a + \boldsymbol{\mu}_i'(\mathbf{V}_i^{-1} - \mathbf{V}_i^{-1}\mathbf{V}_i^*\mathbf{V}_i^{-1})\boldsymbol{\mu}_i + \mathbf{y}'(\mathbf{I}_n - \mathbf{X}_i\mathbf{V}_i^*\mathbf{X}_i')\mathbf{y}_i - 2\mathbf{y}_i'\mathbf{X}_i\mathbf{V}_i^*\mathbf{V}_i^{-1}\boldsymbol{\mu}_i.$$

   In this expression we may calculate $\mathbf{V}_i^*$, $2\mathbf{X}_i\mathbf{V}_i^*\mathbf{V}_i^{-1}\boldsymbol{\mu}_i$, $\mathbf{I}_n - \mathbf{X}_i\mathbf{V}_i^*\mathbf{X}_i'$ and

$$a + \boldsymbol{\mu}_i'(\mathbf{V}_i^{-1} - \mathbf{V}_i^{-1}\mathbf{V}_i^*\mathbf{V}_i^{-1})\boldsymbol{\mu}_i,$$

   for each of the models before we begin the simulation.

2. Generate $n \times 1$ vectors $\mathbf{Z}_k$ $(k = 1, \ldots, s)$ each containing independent $N(0, 1)$ deviates.

3. Draw a random sample of s models from $\mathcal{M}$ using probability sampling in conjunction with the prior model probabilities $P(m_i)$, for $i = 1, \ldots, M$. This sample provides a set of 'true' models from which data sets are generated. Without loss of generality, label these models, which may not be distinct, as $u_1, \ldots, u_s$.

4. Generate values for the variances σ_k^2 $(k = 1, \ldots, s)$ independently from an $IG(a/2, d/2)$ distribution.

5. Obtain values of the regression parameters $\boldsymbol{\beta}_k^{\text{sim}}$ for the kth simulation $(k = 1, \ldots, s)$ as random draws from the prior distribution $N(\boldsymbol{\mu}_{u_k}, \sigma_k^2\mathbf{V}_{u_k})$, where $\boldsymbol{\mu}_{u_k}$ and $\mathbf{V}_{u_k}$ are the prior mean and variance respectively of model u_k.

6. Create $n \times 1$ vectors of values of $\mathbf{Y}$ defined by

$$\mathbf{Y}_k = \boldsymbol{X}_{u_k} \boldsymbol{\beta}_k^{\text{sim}} + \sigma_k \mathbf{Z}_k,$$

for $k = 1, \ldots, s$, where $\boldsymbol{X}_{u_k}$ is the $n \times p_{u_k}$ model matrix for model u_k.

7. For $k = 1, \ldots, s$, start from the original prior distributions $P(m_i)$, $f(\boldsymbol{\beta}_i, \sigma | m_i)$, and calculate the posterior model probabilities $P(m_i|\mathbf{Y}_k)$ for $i = 1, \ldots, M$ using Bayes' theorem, see equation (1.7).

8. Find the expected loss of the model chosen by calculating $\sum_j L(i,j) P(m_j|\mathbf{y})$ for $i = 1, \ldots, M$ and taking the lowest value obtained.

9. Calculate the average expected loss over the s chosen models selected from the s simulation runs.

In step 8, computational savings may be made in one of two ways:
First,

i arrange the models in order of prior probability,

ii calculate the expected loss for the first model in this list in the usual way.

iii for each subsequent model, add up the terms in the expression (2.1) until

- either the expected loss becomes greater than that of the current best model, in which case we move on to the next model, or
- all terms of the sum have been included and the expected loss is still lower than that of the best model, in which case the model currently under consideration is now the best.

Alternatively, provided condition (2.7) holds, the expected loss of the model selected can be found from equation (2.8).

2.4 Searching for Designs

We use a modified Fedorov Exchange Algorithm (MFEA, Cook and Nachtsheim, 1980) to search for good n-point designs under the PMD criterion. The algorithm works as follows:

1. A set of N candidate points is created, from which the rows of the design may be chosen. For example, if we are dealing with f factors, each at two levels, we may allow all possible combinations of the factor levels, giving a candidate list of 2^f points.

2. Start with an n-point design consisting of points randomly selected, with replacement, from the candidate list. Evaluate the objective function for this design.

3. For the first design point, form a new design by exchanging this point with the first candidate point. Evaluate the objective function for the new design formed. If this value is lower than before the exchange was made, keep the new point in the design and record the objective function value as the best obtained so far. Otherwise, do not retain the exchange. Repeat the procedure for the lth candidate point ($l = 2, \ldots, N$).

4. Repeat step 3 for the ith design point ($i = 2, \ldots, n$).

5. If an improvement in the value of the objective function exceeding ϵ, where ϵ is some pre-determined small number, has been achieved through step 3, retain this design and repeat the process from step 3. If a sufficient improvement is not obtained, stop and return the current design.

In practice, several tries of the algorithm are made from random starting designs and the best design found is selected. This is to try to overcome the problem of the search becoming stuck near a local optimum.

2.4.1 Adaptive Simulation Size

We refine this algorithm by the use of an adaptive simulation size. In the early stages of the search, the differences between designs are expected to be larger than at later steps when the algorithm is approaching a local minimum. Hence, a small simulation size may be sufficient for the first few steps of the search, but a larger simulation is required later to give sufficient precision in our estimate of the objective function to discriminate between designs in the presence of Monte Carlo error. An adaptive sample size was also used in the simulated annealing algorithm of Muller, Sanso and De Iorio (2004).
Let R_k (for Bayes risk) be the expected loss (2.1) given the current design, at the kth step of our search. This is a random variable and, at the kth step, a random sample of values of R_k values is used to estimate $E(R_k)$, the objective function, by the sample mean, $\bar{R}_k$. Let s_k be the sample size used at the kth step of the search. At the $(k+1)$th step, the search moves to a new design if the estimated objective function for the new design is lower than that for the current design, that is, if $\bar{R}_{k+1} < \bar{r}_k$, where $\bar{r}_k$ is the observed sample mean. Suppose that δ is the decrease in value of the objective function observed on the last occasion on which a move was made to a new design. Then, at the kth step we would like to be able to detect a change of magnitude δ with reasonable probability in the presence of the Monte Carlo variation in R_{k+1}. Hence, if $E(R_{k+1}) = \bar{r}_k - \delta$, we require that

$$P(\bar{R}_{k+1} < \bar{r}_k) > 1 - \alpha. \tag{2.11}$$

The variance of R_{k+1} is unknown and is estimated by the sample variance, v_k, from the kth simulation. By the Central Limit Theorem, $\bar{R}_{k+1}$ is approximately distributed as $N(E(R_{k+1}), v_k/s_{k+1})$. Hence, (2.11) is satisfied provided that

$$\Phi\left(\frac{\delta}{\sqrt{v_k/s_{k+1}}}\right) > 1 - \alpha,$$

where Φ is the standard $N(0,1)$ cumulative distribution function. Thus, we require

$$s_{k+1} > \left(\frac{\Phi^{-1}(1-\alpha)}{\delta}\right)^2 v_k.$$

Therefore, at each step of the search, we set s_{k+1} to be the smallest integer to achieve this.

2.5 Summary

In this chapter we have described the Bayesian decision theoretic approach taken to model selection in this thesis. The new design criterion proposed (the PMD criterion) follows naturally from minimising the expectation of the expected loss of the model chosen and has the advantage of having a straightforward interpretation. It is not possible to calculate the objective function analytically, so a method for its estimation has been proposed using simulation. We have also given the steps of the Modified Fedorov Algorithm that is used in this thesis to find good designs by search, and described some methods of making computational savings which have been incorporated into the algorithm.

Chapter 3

Screening Experiments

3.1 Introduction

In this chapter we consider experiments where the aim is to identify active effects from a large number of possible effects. Typically these experiments take place early in an investigation when there is not much prior information available. Hence non-informative priors will be used in the analysis of the results. First, in Section 3.2, we investigate the designs of Sun *et al.* (2002) (known as 'main effects orthogonal designs') for a very simple model space in which each model has only main effects and a mean. In later sections, interactions are introduced. In Section 3.3 the model space is composed of models with all main effects and exactly one interaction. The sensitivity of the designs to the hyperparameters of the prior distributions is investigated in Section 3.3.4. In Section 3.4 models containing any subset of main effect and interaction terms are permitted. For this model space a comparison is made between the designs selected under the PMD criterion and the designs selected under criteria from the literature. Finally, in Section 3.4.6, an evaluation is made of the designs selected under the PMD criterion and three criteria from the literature defined in Section 1.3 in terms of several indicators of their capability of selecting a 'correct model'.

Throughout this chapter, unless otherwise stated, the hyperparameters for the prior distributions of $\boldsymbol{\beta}$ and σ^2 are $a = 200, d = 15, \boldsymbol{\mu} = \mathbf{0}$ and $\mathbf{V}_i = \mathbf{I}_{p_i}$ (The hyperparameters are introduced in Section 1.1.3 and the sensitivity of the PMD criterion to the values

used is investigated in Section 3.3.4). These values of a and d give a mean for the prior distribution of σ^2 of around 15.

3.2 Main Effects Only Models

We first investigate designs for a very simple model space composed of models that may contain any subset of main effect terms. To assign probabilities to each possible model, it is assumed that each main effect is included independently with prior probability p=0.5. A modified Fedorov exchange algorithm search (Section 2.4) was used to find designs for $f = 5, 6, 7, 8$ factors and run sizes from $f + 1$ to 16. Note that for $f + 1$ runs the designs are saturated for the largest model.

Figure 3.1 shows the value of the PMD objective function for the designs obtained plotted against the different numbers of runs investigated. For each number of factors, the objective function values for the best design found at each number of runs are connected by a line. For numbers of runs for which a main effect orthogonal design exists (8, 12 or 16 runs), these designs were the best found under the PMD criterion. This figure shows how the value of the PMD objective function decreases as the number of runs increases, and increases with the number of factors. For 5 and 6 factors, the objective function decreases most rapidly as the number of runs increases up to 8, the first available orthogonal design. For all numbers of factors shown, especially 7 and 8, there is a sharper decrease in the value of the PMD objective function between 11 and 12 runs than from 12 to 13 runs, showing the benefit of an orthogonal design, and the relative lack of improvement from using one more run than needed for an orthogonal design.

3.3 Models Containing All Main Effects and one Two-factor Interaction

Another simple model space was examined by Li (2005) who assumed that all main effect terms are known to be active, and there is possibly a single active two-factor interaction. The main objective of the experiment is to detect this interaction. The

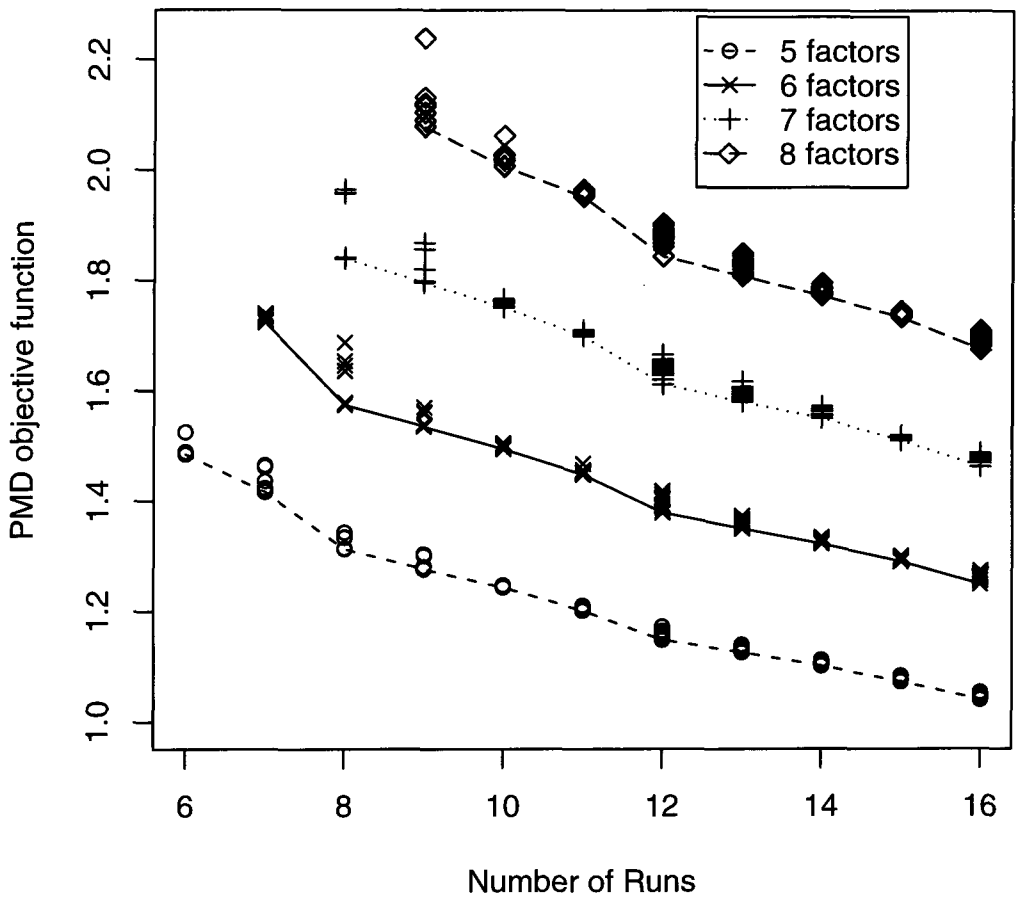

Figure 3.1: Value of the PMD objective function vs run size for 5, 6, 7 and 8 factor designs using models containing main effects only.

model space, $\mathcal{M}$, is composed of models

$$m_{ij} : E(\mathbf{Y}) = \beta_0 + \sum_{l=1}^{f} \beta_l x_l + \beta_{ij} x_i x_j, \tag{3.1}$$

for $i, j = 1 \ldots f; i \neq j$, i.e. the set of models with all f linear effects and one interaction. We assume that all models are equally likely to be 'correct' and assign $P(m_{ij}) = |\mathcal{M}|^{-1}$.

3.3.1 Ranking of Main Effects Orthogonal Designs Under Four Criteria

The 16-run main effects orthogonal designs of Sun *et al.* (2002) for 3 to 9 factors were ranked under each of the PMD, F, MD and HD criteria. The top ten designs under each criterion are given in Table 3.1, using the design labels of Sun *et al.*. The full factorial designs and regular fractional factorial designs are indicated in the table.
For 3 to 5 factors, 16 runs are sufficient to allow full factorial or resolution V regular fractional factorial designs that have all main effects and interactions clear of each other and this design is ranked highest by all the criteria. The best main effects orthogonal design for 3 factors is two replicates of the full factorial. For 4 factors, the best design is the full factorial and for 5 factors the best is the 2_V^{5-1} half fraction $ABCDE = I$.
For 6 to 9 factors, where the regular design has some interactions completely aliased with other interactions or main effects, the PMD criterion does not rank regular designs highly. In contrast,

- the MD criterion selects a regular design for all numbers of factors.
- the F criterion selects a regular design for 6 to 8 factors, with a regular design ranked second best for 9 factors.
- The HD criterion selects regular designs for 6 to 8 factors but the regular design is only ranked sixth for 9 factors.

The disagreement between the designs selected using the PMD criterion and the other criteria becomes more pronounced for larger numbers of factors. Figure 3.2 shows the rank correlation between the objective function of the PMD and those of each of the other criteria for the set of 16 run orthogonal main effects designs for 3 to 9 factors. This plot indicates that designs that are good under the other criteria are not necessarily

Table 3.1: Ranking of 16-run main effects orthogonal designs of Sun *et al.* (2002) for 3-9 factors under four criteria, using models containing all main effects and one 2-factor interaction. *Full factorial, †Regular design.

# factors	3				4				5				6			
Criterion	PMD	F	MD	HD	PMD	F	MD	HD	PMD	F	MD	HD	PMD	F	MD	HD
Ranking 1	2*	2*	2*	2*	3*	3*	3*	3*	4†	4†	4†	4†	13	5†	5†	5†
2	3	3	3	3	4	4	4	4	5	5	3	5	19	8	8	8
3	1	1	1	1	5	5	2	5	8	3	5	3	20	13	4†	13
4	-	-	-	-	1	2	5	2	7	8	7	7	24	4†	14	19
5	-	-	-	-	2	1	1	1	10	7	8	8	22	14	13	14
6	-	-	-	-	-	-	-	-	11	10	2	10	6	19	19	12
7	-	-	-	-	-	-	-	-	2	2	10	11	23	6	6	20
8	-	-	-	-	-	-	-	-	6	11	11	9	26	12	12	24
9	-	-	-	-	-	-	-	-	3	9	9	2	27	24	7	22
10	-	-	-	-	-	-	-	-	9	6	6	6	7	20	15	4†

# factors	7				8				9			
Criterion	PMD	F	MD	HD	PMD	F	MD	HD	PMD	F	MD	HD
Ranking 1	32	6†	6†	6†	68	6†	6†	6†	71	25	4†	25
2	49	12	12	12	67	18	18	18	36	4†	25	53
3	55	28	28	28	39	42	4	42	79	53	17	84
4	53	32	5	32	77	77	26	77	70	17	53	44
5	43	21	11	49	66	48	17	41	32	84	5	17
6	54	33	22	33	72	17	48	48	82	44	44	4†
7	27	11	33	21	42	4	77	67	77	5	3	71
8	45	5	32	31	36	26	42	68	69	65	84	83
9	30	22	21	55	73	41	41	17	68	22	12	74
10	50	49	49	36	74	40	12	76	84	10	10	55

good under the PMD criterion. For example, the highest ranked design under the F criterion for each of 6 to 9 factors are ranked 24, 53, 80 and 82 under the PMD criterion out of the 27, 55, 80 and 87 designs evaluated respectively.

The reason the difference between the F and PMD is that the two criteria have different objectives. The PMD criterion is aimed that discriminating between models, and so chooses designs with minimal aliasing of the terms that vary between models, i.e. the interaction terms. The F criterion selects designs which enable precise estimates of individual terms, particularly those with high prior probability, i.e. the main effects terms, even though good estimates of these terms are not directly useful in discriminating between the models.

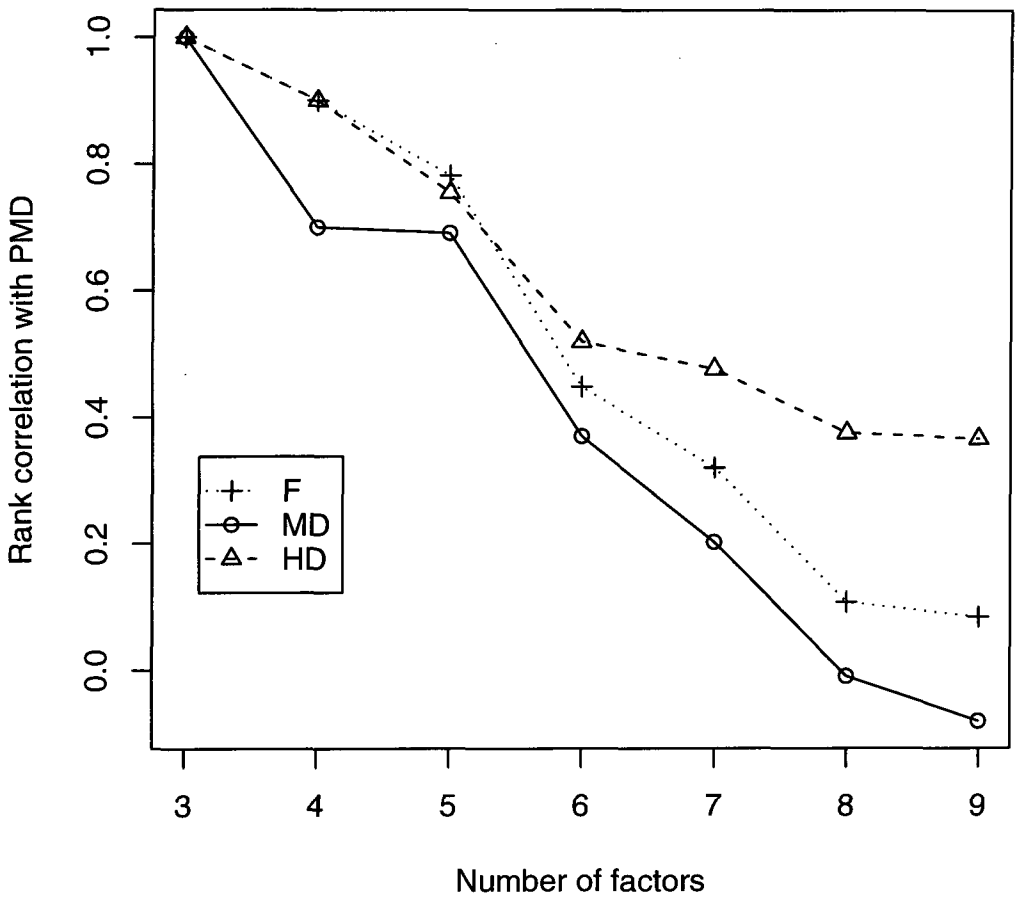

Figure 3.2: Correlation between the ranks of 16-run main effect orthognal designs under the PMD criterion and the ranks under each of the F, MD and HD criteria.

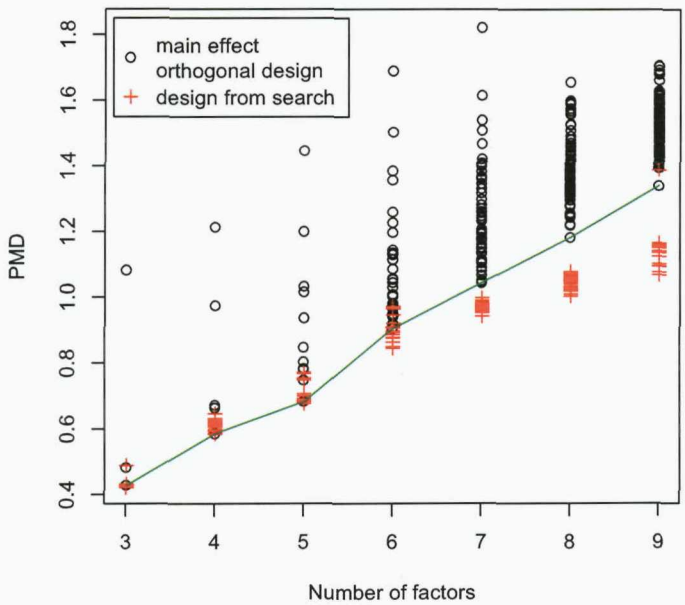

Figure 3.3: Values of the PMD objective function for all the 16-run main effect orthogonal designs together with designs found by algorithmic search.

3.3.2 Non-orthogonal designs

An MFEA search was run for designs of 16 runs in $3, \ldots, 9$ factors, and repeated 20 times for each number of runs. For 3 to 5 factors, it was not possible to improve on the best main effects orthogonal design. For $6, \ldots, 9$ factors, it is possible to make an improvement compared to the best main effects orthogonal design. Figure 3.3 shows the PMD values of all main effects orthogonal designs for $3, \ldots, 9$ factors, with a line connecting the best for each number of factors. Also plotted are the PMD values for the designs returned from 20 tries of the Modified Federov Exchange Algorithm at each number of factors.

3.3.3 Explanation of Results

For 3-5 factors, the best main effects orthogonal design has not only main effects orthogonal to each other, but also all effects (main effects and 2-factor interactions) are

orthogonal to each other. It is for these numbers of factors that no improvement on the best main effects orthogonal design was found with an unrestricted search. Such designs do not exist for greater numbers of factors in 16 runs.

For 7-9 factors, the best design found did not have main effects orthogonal, but did have all main effects orthogonal to all 2-factor interactions. For 6 factors, designs with this property were returned as local optimum designs, but were not the overall best found.

3.3.4 Sensitivity

Prior Model Probabilities

To assess the sensitivity of the design ranking to changes in the prior information, suppose that we have 7 factors. Suppose also that models m_{1j} $(j = 1 \dots 7)$, i.e. those containing an interaction involving the first factor, are considered to be more (or less) likely, a priori, than the other models. There are 15 models that do not include an interaction involving factor 1, and 6 models that do include an interaction involving factor 1. The whole model space consists of these 21 models and $\sum_{1 \le i < j \le 7} P(m_{ij}) = 1$. Hence we may incorporate such prior information by setting

$$p(m_{ij}) = \frac{1}{15 + 6\alpha} \quad i, j = 2, \dots, 7, \ i < j; \quad P(m_{1j}) = \frac{\alpha}{15 + 6\alpha}, \ j = 2, \dots, 7 \ (0 \le \alpha)$$

where m_{ij} is defined in (3.1). The prior probabilities of all models are equal when $\alpha = 1$. Figure 3.4 shows the changing performance of three designs as α varies from 0 to 4. We observe that the relative performance of the designs may change, especially if we have strong prior information on which models are more likely to be true (α near 0 or $\alpha \gg 1$). For example, the red line corresponds to a design for which at least one of the models m_{1j} $(j = 2 \dots 7)$ is indistinguishable from one or more models m_{ij} $(i, j = 2 \dots 7; \ i < j)$, when all models are assigned the same prior probability and the Normal-Inverse Gamma prior distribution, defined in Section 1.4 is used for the model parameters. Hence, although this design is poor when $\alpha = 1$, it is more useful when prior information indicates which of two sets of models is more likely.

The green line corresponds to a design where models that are indistinguishable under uniform prior information are only of the type m_{ij} $(i, j = 2 \dots 7; i < j)$; this design becomes more useful if the prior probability is concentrated on the first 6 models ($\alpha > 1$).

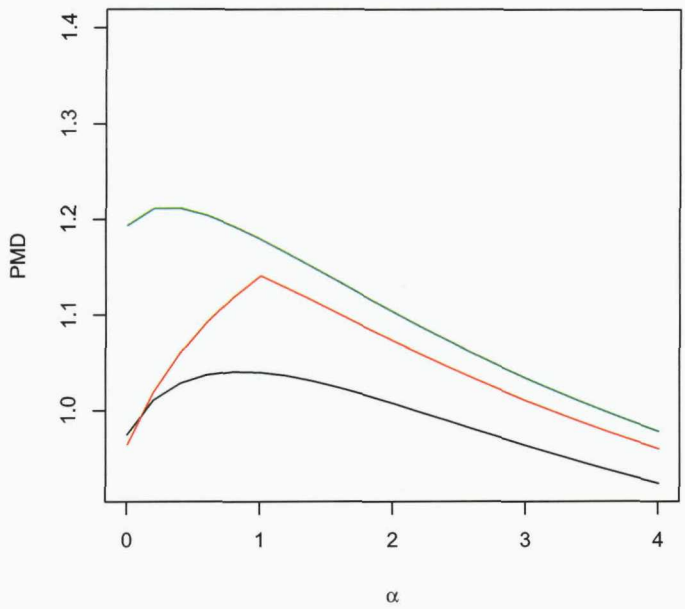

Figure 3.4: PMD values for three 16 run main effect orthogonal designs in 7 factors evaluated for $0 < \alpha < 4$

The results indicated in black correspond to a design which enables discrimination between all the models; it is a good choice for all α, although a more informative prior will improve its performance.

All 55 16-run main effects orthogonal designs in seven factors are displayed in the plot in Figure 3.5. In addition to the colour coding described above, designs shown in blue have pairs of indistinguishable models where the interactions of neither model involves the first factor and has indistinguishable pairs where exactly one of the models has an interaction involving the first factor. The objective function of these designs behaves similarly to a combination of the red and green lines.

The grey lines in the figure represent designs that also have pairs of indistinguishable models each of which contains an interaction term involving the first factor, in addition to the types of pairs that are not distinguishable for the designs in blue. These perform poorly for all values of α.

Finally, for the designs corresponding to the yellow lines, there is at least one pair of

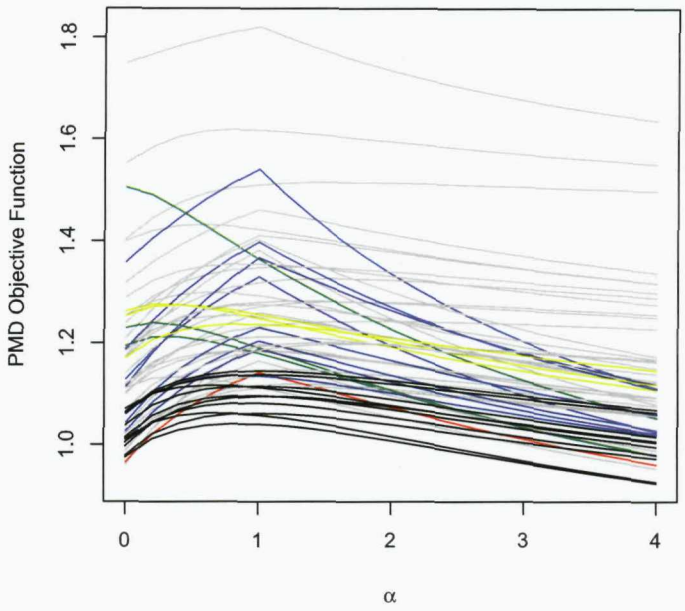

Figure 3.5: PMD values for the 55 main effects orthogonal designs for 7 factors in 16 runs evaluated for $0 \leq \alpha \leq 4$.

indistinguishable models which have interactions involving the first factor. Also, there is at least one pair of indistinguishable models with neither interaction involving the first factor. Finally, there are no pairs of indistinguishable models such that one model has an interaction involving factor 1 and the other does not. The objective function for these designs stays near the middle of the range of the range as α varies.

Prior Variance-Covariance Matrix (I)

We also assessed the sensitivity of the objective function to changes in the prior variance-covariance matrix. We evaluate the objective function for all 16-run orthogonal designs in 7 factors, using the prior distributions given in Section 3.1, except that

$$\mathbf{V} = \left(\begin{array}{c|c} \mathbf{I}_8 & 0 \\ \hline 0 & \lambda \end{array}\right),$$

that is, the prior variance of β_9, the coefficient of the interaction term, equals λ. We evaluated and ranked the designs using the PMD objective function for 10 different

values of λ, at intervals of 0.2 from 0.2 to 2. We found that the best design remained the same for each value of λ and that two designs were always ranked second or third, although their order varied with λ. However, the relative ranking of other designs varied a lot as λ was changed, for example, design 6 was ranked 30th for $\lambda = 1$, but 15th for $\lambda = 0.2$.

To investigate why the ranking of designs changes over the range $0.2 \leq \lambda \leq 1$, we have plotted in Figure 3.6 the number of pairs of models which are distinguishable (i.e. will not always have identical posterior probabilities) when given equal prior probabilities and the same prior parameter distribution, against

$$\sum_{1 \leq i \leq 8 < j \leq 29} S_{ij}^2 \tag{3.2}$$

on the vertical axis. Here $S_{ij}^2 = ((\mathbf{X}'\mathbf{X})_{ij})^2$ where $\mathbf{X}$ is the model matrix for a model including the intercept, the 7 main effects and all 21 2-factor interactions between them and $(\mathbf{X}'\mathbf{X})_{ij}$ denotes the (i,j)th element of $(\mathbf{X}'\mathbf{X})$. In expression (3.2), i denotes the intercept or a main effect term and j is an interaction, the expression measures the extent to which a design has aliasing between interactions (which vary between models) and other effects (which are included in all models). We observe that for designs whose ranking improves as λ increases, there is a lot of aliasing between interactions and other effects, relative to the other designs with a similar number of distinguishable models. This is because for small λ, the magnitude of the interaction effects generated is small, and hence will not greatly increase the probability of the true model when it is partially aliased with another effect in the same model. However, as λ increases, the magnitude of the interaction effect increases and so the probability of the model containing that effect will be increased even when aliasing is present.

In this case, the same design would be chosen for all λ in the range studied. However, these results show that, in general, an incorrect prior specification of the relative sizes of effects common to all models and those which vary between models might lead to the choice of an inappropriate design.

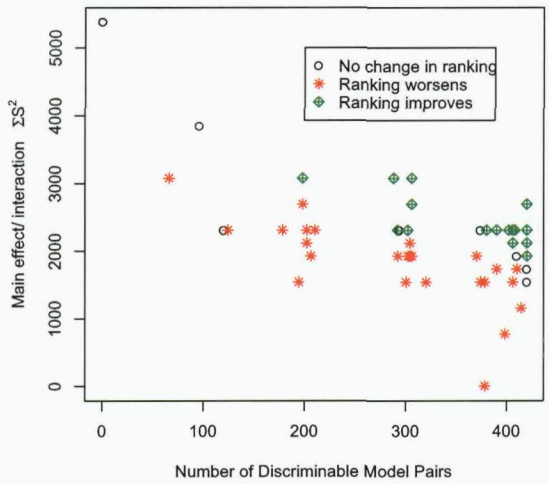

Figure 3.6: Features of designs leading to a change in ranking as λ varies.

Prior Variance-Covariance Matrix (II)

We now investigate a second form of prior variance-covariance matrix, namely $\mathbf{V} = \lambda\mathbf{I}$. As before, we use the values of λ at intervals of 0.2, from 0.2 to 2, and evaluate the PMD criterion for the 55 orthogonal 16-run designs in 7 factors. The value of λ changes the sizes of the model parameters relative to the random errors, as $\boldsymbol{\beta}|\sigma \sim N(0, \sigma^2\mathbf{V})$ and $\boldsymbol{\epsilon} \sim N(0, \sigma^2)$

For each value of λ, we ranked the designs from 1 (best, i.e. lowest PMD) to 55 (highest PMD). The results for the six designs with most improved ranks, the six that worsen in ranking the most, and the three best designs, are shown in Table 3.2.

An indication of why the ranking of some designs changes with λ is given by the $S_{ij}^2 = (\mathbf{X}'\mathbf{X}_{ij})^2$, $(i \neq j)$ values, which show the extent of aliasing between effects in the design, and which may equal 0, 64 or 256 for these designs. As an example, compare designs 10 and 45. Design 10 has $S_{ij}^2 = 0$ for a large number of pairs i, j, which enables discrimination between models containing these effects even when random error is large compared to the effect sizes. Design 45, on the other hand, has partial aliasing between many effects, so is less useful than design 10 for low values of λ. However, for large λ, models which have partially aliased effects may be better discriminated, but those with

Table 3.2: Change in ranking of designs under PMD with increasing effect size relative to error

Design no.	41	50	45	42	51	52
Ranking for $\lambda = 0.2$	50	15	22	37	27	35
Ranking for $\lambda = 2$	42	7	9	20	8	12
$\#S_{ij}^2 = 0$	656	680	668	662	644	644
$\#S_{ij}^2 = 64$	144	132	144	144	168	168
$\#S_{ij}^2 = 256, i \neq j$	12	0	0	6	0	0
Design no.	19	10	7	30	31	11
Ranking for $\lambda = 0.2$	12	9	20	5	19	18
Ranking for $\lambda = 2$	25	21	31	13	26	24
$\#S_{ij}^2 = 0$	710	728	746	710	686	728
$\#S_{ij}^2 = 64$	96	72	48	96	120	72
$\#S_{ij}^2 = 256, i \neq j$	6	12	18	6	6	12

Design no.	32	49	55
Ranking for $\lambda = 0.2$	1	2	3
Ranking for $\lambda = 2$	1	2	3
$\#S_{ij}^2 = 0$	692	680	686
$\#S_{ij}^2 = 64$	120	132	126
$\#S_{ij}^2 = 256, i \neq j$	0	0	0

full aliasing ($S_{ij}^2 = 256, i \neq j$) are still indistinguishable, so the ranking of designs such as 10 worsens. The three best designs were included in Table 3.2 for comparison. These designs have no full aliasing and less partial aliasing than most other designs, and are also good under the Expected Prediction Difference criteria of Li (2005).

Hyperparameters of the prior distribution

The value of the PMD criterion for a design is invariant to the scale parameter, a, of the prior distribution of σ, provided we are using zero means for the prior distributions of parameters. To show this, let us follow the steps of our computer code. First, we select models at random from the prior distribution, then, for a given model, we generate random values of $\sigma^2 = \frac{a}{\psi}$ where $\psi \sim \chi_d^2$. We then generate the other model parameters from the prior distribution $\boldsymbol{\beta} \sim N(\mathbf{m}, \sigma^2\mathbf{V})$ by $\boldsymbol{\beta} = \sigma\mathbf{C}\mathbf{z}_0 + \mathbf{m}$, where $\mathbf{z}_0$ is a p-dimensional vector of $N(0,1)$ random deviates and $\mathbf{C}$ is the Cholesky decomposition of V such that $\mathbf{C}'\mathbf{C} = \mathbf{V}$. We generate samples of $\mathbf{y}$ as $\mathbf{z}\sigma + \mathbf{X}_t\boldsymbol{\beta}$ where $\mathbf{z}$ is an n-dimensional vector of $N(0,1)$ random deviates and $\mathbf{X}_t$ is the model matrix of the true model. Then, for $\mathbf{m} = 0$, $\mathbf{y} = \sigma(\mathbf{z} + \mathbf{X}_t\mathbf{C}\mathbf{z}_0)$ and, if a_i^* is the value of a^* obtained for

model i,

$$\begin{aligned} a_i^* &= a + \mathbf{y}'(\mathbf{I}_n - \mathbf{X}_i\mathbf{V}^*\mathbf{X}'_i)\mathbf{y} && (3.3)\\ &= a + \sigma^2(\mathbf{z} + \mathbf{X}_t\mathbf{C}\mathbf{z}_0)'(\mathbf{I}_n - \mathbf{X}_i\mathbf{V}^*\mathbf{X}'_i)(\mathbf{z} + \mathbf{X}_t\mathbf{C}\mathbf{z}_0) && (3.4)\\ &= a + \frac{a}{\psi}(\mathbf{z} + \mathbf{X}_t\mathbf{C}\mathbf{z}_0)'(\mathbf{I}_n - \mathbf{X}_i\mathbf{V}^*\mathbf{X}'_i)(\mathbf{z} + \mathbf{X}_t\mathbf{C}\mathbf{z}_0). && (3.5) \end{aligned}$$

The likelihood given model i, is

$$f(\mathbf{y}|i) = \frac{|\mathbf{V}_i^*|^{1/2}a^{d/2}\Gamma(d^*/2)}{|\mathbf{V}_i|^{1/2}\pi^{n/2}\Gamma(d/2)}(a_i^*)^{-d^*/2},$$

where $d^* = d + n$. The part of the likelihood that depends on a is

$$a^{d/2}(a_i^*)^{-d^*/2} = a^{d/2}a^{-d^*/2}\left[1 + \psi^{-1}(\mathbf{z} + \mathbf{X}_t\mathbf{C}\mathbf{z}_0)'(\mathbf{I}_n - \mathbf{X}_i\mathbf{V}^*\mathbf{X}'_i)(\mathbf{z} + \mathbf{X}_t\mathbf{C}\mathbf{z}_0)\right]^{-d^*/2}.$$

Hence there is a common factor of $a^{(d-d^*)/2}$ in all model likelihoods, which cancels when we calculate the posterior model probabilities. Therefore the PMD value is independent of a when the prior means of the regression parameters are 0.

The effect of the shape hyperparameter, d, appears to be small enough to not really affect the value of the PMD criterion or the ranking of designs. If we use the same $\mathbf{V}$ for all models, then the part of the likelihood that differs between models is $|\mathbf{V}_i^*|^{\frac{1}{2}}(a_i^*)^{-\frac{d^*}{2}}$. We may divide all likelihoods by $(\frac{1}{a})^{\frac{d^*}{2}}$ without changing the posterior model probabilities, to get

$$|\mathbf{V}_i^*|^{\frac{1}{2}}(1 + \frac{r_i}{\Psi})^{-\frac{d^*}{2}}, \tag{3.6}$$

where

$$r_i = (\mathbf{z} + \mathbf{X}_t\mathbf{C}\mathbf{z}_0)'(\mathbf{I}_n - \mathbf{X}_i\mathbf{V}^*\mathbf{X}'_i)(\mathbf{z} + \mathbf{X}_t\mathbf{C}\mathbf{z}_0).$$

If we assume that Ψ, a random variable, is close to its modal value of $d + 2$, then (3.6) is approximated by

$$|\mathbf{V}^*|^{\frac{1}{2}}(1 + \frac{r_i}{d+2})^{-\frac{d^*}{2}}. \tag{3.7}$$

Using the fact that $(1 + \frac{x}{n})^n \to e^x$ as $n \to \infty$, (3.7) has a limit, as $d \to \infty$, of

$$|\mathbf{V}^*|^{\frac{1}{2}}e^{-r_i/2}.$$

For the 11 orthogonal designs for 5 factors in 16 runs, we evaluated the PMD criterion for $d = 5, 10, 15, 20, 25$. Also, we generated random normal residuals $\mathbf{z}, \mathbf{z}_0$ and random

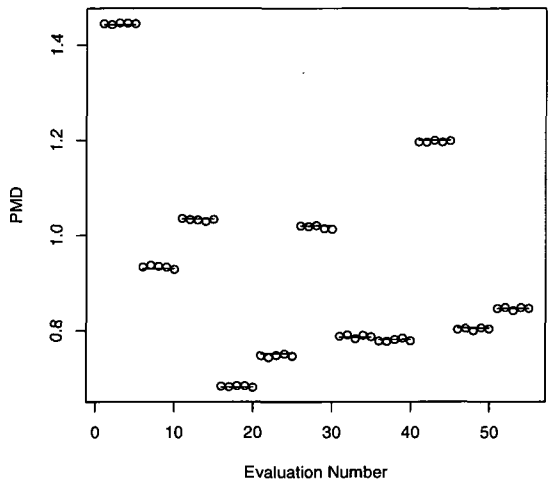

Figure 3.7: PMD value for 11 designs, repeated for 5 values of d.

selections of true models t. These were used to calculate r_i and $e^{-r_i/2}$ for all models. In this model space, $L(i,j) = 2$ for $i \neq j$, so our selected model is that with the highest posterior probability. Hence, as the prior model probabilities are equal, we were able to obtain an approximation to the PMD value for a design as $d \to \infty$ by calculating the mean of

$$2\left(1 - \frac{\max_i\{|\mathbf{V}_i^*|^{\frac{1}{2}}e^{-r_i/2}\}}{\sum_i |\mathbf{V}_i^*|^{\frac{1}{2}}e^{-r_i/2}}\right) \tag{3.8}$$

over 50 000 simulations.

The results are shown in Figure 3.7. Each group of five points gives the PMD values for a design at each of the five values of d that we used. The horizontal line through the group is the value obtained for the limit (3.8) for that design. We see that, even for small d, the PMD value is close to the asymptotic limit. Although the individual values of $|\mathbf{V}_i^*|^{\frac{1}{2}}e^{-r_i/2}$ for the models may not be close to the corresponding values of (3.6), the ratio used to obtain the model probabilities does appear to converge quickly as d increases.

Summary of Sensitivity Analysis

We have investigated the sensitivity of the PMD criterion to the prior model probabilities and hyperparameters of the prior distribution for a model space consisting of models involving seven factors that contain all main effect terms and one 2-factor interaction. In all cases, 16-run main effects orthogonal designs were evaluated and the best design remained the same despite the different prior distributions used. Firstly, the prior model probabilities were altered so that models containing an interaction involving the first factor were more or less likely than the others. For most designs, the PMD objective function was highest when all models had equal prior probability. However, the PMD objective function for designs with aliased pairs of interactions that did not involve the first factor increased as the prior probability for the indistinguishable models was increased. For all prior model probabilities studied, the best two designs remained the same. The sensitivity of the criterion to the prior variance-covariance matrix was investigated by changing the magnitude of the interaction term in comparison to the main effects. If the interaction was larger than the main effects, then designs that had fewer indistinguishable pairs of models were ranked more highly, even if they had partial aliasing between main effect and interaction terms. We also tried multiplying the prior variance-covariance matrix by a constant, to change the relative magnitude of the model coefficients and the error term. If the error was large, the designs with the least partial aliasing were more effective, whereas, with small error, minimising the number of totally aliased interactions was more important. Over the ranges studied, the best three designs remained the same. The scale hyperparameter, a, does not affect the value of the PMD objective function if the prior means of the regression coefficients are zero. The shape hyperparameter, d, also appears to have little effect on the PMD objective function.

3.4 Models containing any Subset of Main Effect and 2-factor Interaction Terms.

Bingham and Chipman (2007), in their work on the HD criterion, use a model space where any subset of the possible main effect and 2-factor interaction terms is permissible - marginality is not enforced. If p is the prior probability that a main effect term is

present, then 2-factor interaction terms are included with probability

$$p_0 = 0.01p,\ p_1 = 0.5p,\ \text{or}\ p_2 = p \tag{3.9}$$

given the presence of $0, 1$ or 2 respectively of the main effect terms of the factors involved. Bingham and Chipman give a formula to find the value of p that gives a specified expectation of the number of active effects. Bingham and Chipman applied their criterion to this model space for two examples: 5 factors in 12 runs and 6 factors in 16 runs. For 5 factors,Bingham and Chipman chose p to be 0.429, and their formula gave an expectation of about four active effects. For 6 factors, the value of p used was 0.410 (giving about five active effects expected). Because of the large size of the model space, Bingham and Chipman approximate their objective function by using a subset of the models with greatest prior probability. They used 40 models for searches and 400 models for the final evaluation of designs.

We will compare the F, HD, MD and PMD criteria on both examples. We will also use 400 models for the final evaluation, and investigate the use of different numbers of models in our searches. An investigation of the sensitivity of the objective function to the values of p and c is presented, as well as the results of simulation studies which indicate the performance of the designs for different loss functions. Finally, we investigate the set of 16-run main effects orthogonal designs for each of 3 to 9 factors, and evaluate each of these under all four criteria, using the 400 models with highest prior probability.

3.4.1 Obtaining the subset of models with highest prior probability

We use a model space where a model may have any combination of main effects and 2-factor interactions. For f factors, this space contains $2^{f+\binom{f}{2}}$ models. Even for moderate f, this number can be very large, for example, for 9 factors there are over 3.5×10^{13} models. This means that we cannot obtain the subset of models with highest prior probability by simply calculating the probability for every model and sorting the models accordingly. However, we may use the fact that many models have equal prior probability to obtain the top 400 (for example) models. Suppose a model contains m main effect terms and, respectively, i_0, i_1 and i_2 interaction terms for which $0, 1$ and 2 of the factors involved have the corresponding main effect term in the model. The prior

probability for such a model is

$$p^m(1-p)^{f-m}(0.01p)^{i_0}(1-0.01p)^{\binom{f-m}{2}-i_0}(0.5p)^{i_1}(1-0.5p)^{m(f-m)-i_1}p^{i_2}(1-p)^{\binom{m}{2}-i_2}, \tag{3.10}$$

provided $0 \leq m \leq f$, $0 \leq i_0 \leq \binom{f-m}{2}$, $0 \leq i_1 \leq m(f-m)$ and $0 \leq i_2 \leq \binom{m}{2}$. The number of models with the same values of m, i_0, i_1 and i_2, and hence the same prior probability, is $\binom{f}{m}\binom{f-m}{i_0}\binom{m(f-m)}{i_1}\binom{\binom{m}{2}}{i_2}$. We also use the probabilities (3.9) of Bingham and Chipman. We are therefore able to define classes of models with the same prior probability, and rank these classes according to the size of this probability. The number of classes is much smaller than the number of models. For example, there are 7582 classes for 9 factors. Starting from the class with highest probability, we use the minimum number of classes necessary for the inclusion of at least 400 models. For each class, it is possible to identify all the constituent models and their probabilities, and so obtain a list of the models of highest prior probability. This may be truncated at the desired number of models, and the probabilities then standardised to sum to 1.

3.4.2 A further look at the prior model probabilities

The classes of models that make up the 447 models with highest prior probability for 6 factors with $p = 0.410$ are given in Table 3.3. The relative probabilities of the models raises some questions about whether the prior model probabilities accurately represent our beliefs about which models are more likely. For example, models with $m = 1, i_0 = 0 = i_2, i_1 = 1$ have higher prior probability than models with $m = 2, i_0 = i_1 = i_2 = 0$, i.e. $P(I + A + AB) > P(I + A + B)$. This is in apparent disagreement with Bingham and Chipman's assumption of effect hierarchy, that lower-order effects are more likely to be important than higher-order effects. This occurs because, given the main effect terms of both A and B (but no other main effect term) are present in the model, there is a fairly high probability for each of the models $I + A + B$, $I + A + B + AB$, $I + A + B + AX$, $I + A + B + BX$ (where X represents any factor other than A or B). However, if A is the only main effect term present, only $I + A$ and models of the form $I + A + AX$ have any sizeable probability. To remove this problem, we suggest two possibilities (a) using a prior with strong heredity, that is, an

Table 3.3: Composition of 447 models with highest prior probability for 6 factors

m	i_0	i_1	i_2	Model Probability	Number of models
0	0	0	0	0.040	1
1	0	0	0	8.9×10^{-3}	6
1	0	1	0	2.3×10^{-3}	30
2	0	0	0	1.9×10^{-3}	15
2	0	0	1	1.3×10^{-3}	15
1	0	2	0	5.9×10^{-4}	60
2	0	1	0	4.8×10^{-4}	120
3	0	0	0	3.6×10^{-4}	20
2	0	1	1	3.3×10^{-4}	120
3	0	0	1	2.5×10^{-4}	60

Table 3.4: Composition of 447 models with highest prior probability for 6 factors, p_1=0.1.

m	i_0	i_1	i_2	Model Probability	Number of models
0	0	0	0	0.040	1
1	0	0	0	0.023	6
2	0	0	0	8.4×10^{-3}	15
2	0	0	1	5.8×10^{-3}	15
3	0	0	0	2.0×10^{-3}	20
3	0	0	1	1.4×10^{-3}	60
1	0	1	0	9.8×10^{-4}	30
3	0	0	2	9.5×10^{-4}	60
3	0	0	3	6.6×10^{-4}	20
2	0	1	0	3.6×10^{-4}	120
4	0	0	0	3.0×10^{-4}	15
2	0	1	1	2.5×10^{-4}	120

interaction term may only be included if both main effect terms in the factors involved are included, or (b) reducing p_1 from the current value of $0.5p$. For example, the top classes of models when $p_1 = 0.1p$ are given in Table 3.4. The ordering of the models by prior probability is now more intuitively reasonable.

3.4.3 5 factors in 12 runs

For 5 factors in 12 runs, the HD-optimal design is a Plackett-Burman design, which is orthogonal in the main effects. For the PMD, MD and F criteria, a different design is chosen, which is given in Table 3.5. This design is not balanced, however the columns representing the first four main effect terms are orthogonal to each other.

For 6 factors in 16 runs, the HD-optimal design is design 13 from the list of orthogonal

Table 3.5: The best 12 run design in 5 factors found under the PMD, MD and F criteria for Bingham and Chipman's model space

-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

main effects designs from Sun et al. This is also the best design found under the F criterion, and is the best design from this list under the PMD criterion. The strengths of this design are

1. As an orthogonal main effects design, all main effect terms may be estimated independently.

2. It has no pairs of effects (main effects or interactions) totally aliased, one of nine orthogonal main effects designs with this property.

3. Of these nine designs, this one has the joint most pairs of main effect and interaction terms orthogonal to each other (78 out of 90 possible pairs, the joint most with design 19). Out of all possible orthogonal main effects designs, only one has all these pairs orthogonal to each other, the 2^{6-2}_{IV} fraction (design 5).

4. Design 13 has more pairs of interactions orthogonal to each other (93 out of 105) than design 19 (87 out of 105).

Under the MD criterion, however, the best main effects orthogonal design is the resolution IV regular design. This is also the best design returned by a search under this criterion. The second and third best designs under this criterion also have total aliasing between pairs of 2-factor interactions.

In Figure 3.8, we have plotted the ranks of the 6 factor 16 run main effects orthogonal designs under each of the four criteria against each other. Designs denoted by a '+' do not have any pairs of terms totally aliased. The PMD criterion completely splits the designs, so that the nine designs with this property are ranked $1, \ldots, 9$ under the PMD criterion, out of the 27 designs. The regular fractional factorial designs in 16 runs and 6 factors are marked by a '*'. The regular designs are the 2^{6-2}_{III} design defined by I=ABC=ADEF, which is ranked second under the F criterion, and third under the MD, but only 23$^{\text{rd}}$ under the PMD criterion, and the 2^{6-2}_{IV} design defined by $I = ABCD = ABEF$, which is ranked 1$^{\text{st}}$ under the MD criterion.
The worst design under all four criteria, design 1, is two copies of a 2^{6-3}_{III} fraction.

3.4.4 Results of Design Searches

In addition to evaluating the main effects orthogonal designs under our criterion for this example, we also run searches, using different numbers of models to evaluate the objective function, performing 20 searches for each set of models. Figure 3.9 shows the PMD objective function for various designs, evaluated using 400 models. The upper part of Figure 3.9 shows that all the searches consistently find designs that are better under the majority of the main effects orthogonal designs. The lower part of the figure shows the results of the same searches, but focuses on the designs with lower values of the PMD objective function. In this figure, it can be seen that the best design found by a search using 40 models to evaluate the objective function is slightly worse than the best main effects orthogonal design. The searches that used 100 or 400 models to evaluate the objective function found designs that had slightly lower values of the PMD objective function than the best main effects orthogonal design.

3.4.5 Sensitivity

We observe the effect of using a different value of p to give different prior model probabilities. This is shown in Figure 3.10, where each line corresponds to one 16-run main effects orthogonal design in 6 factors. We see that, although the value of the PMD objective function increases with p, as with a greater p a greater number of models have

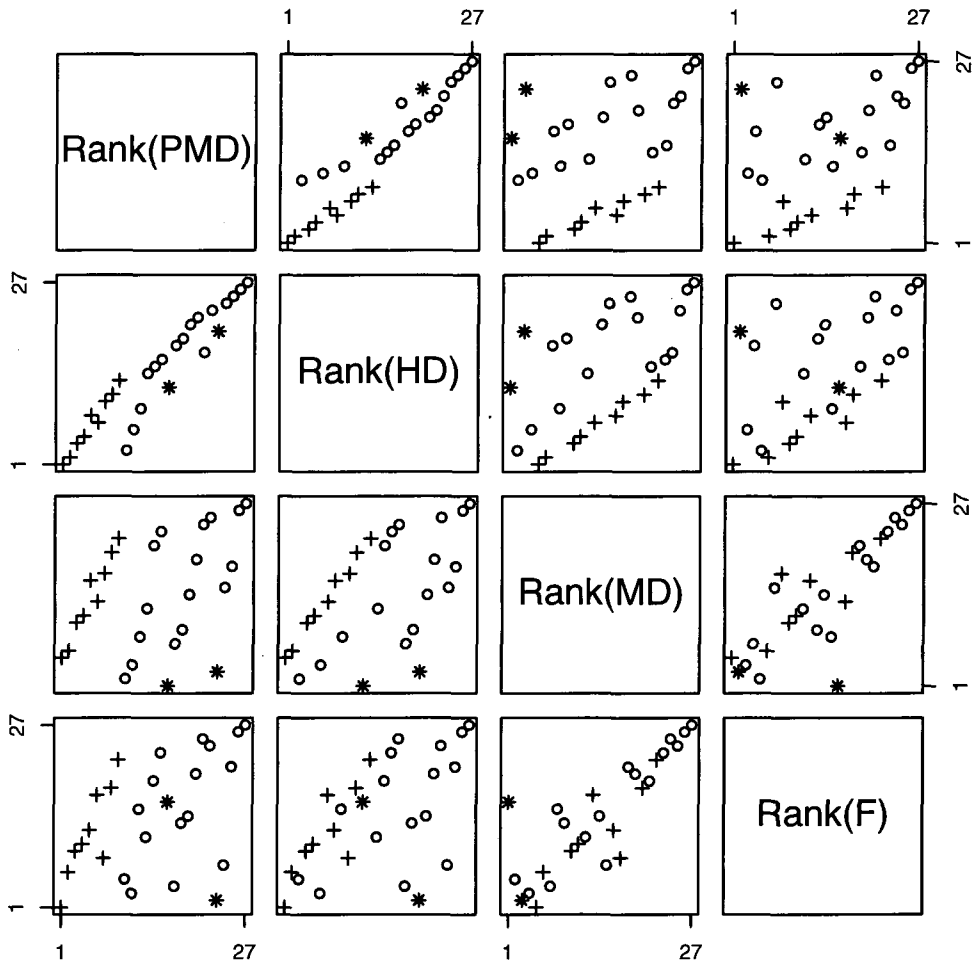

Figure 3.8: Ranks for each of the four criteria for the 16 run main effects orthogonal designs for 6 factors; + indicates designs with no pairs of terms completely aliased, * indicates regular fractions.

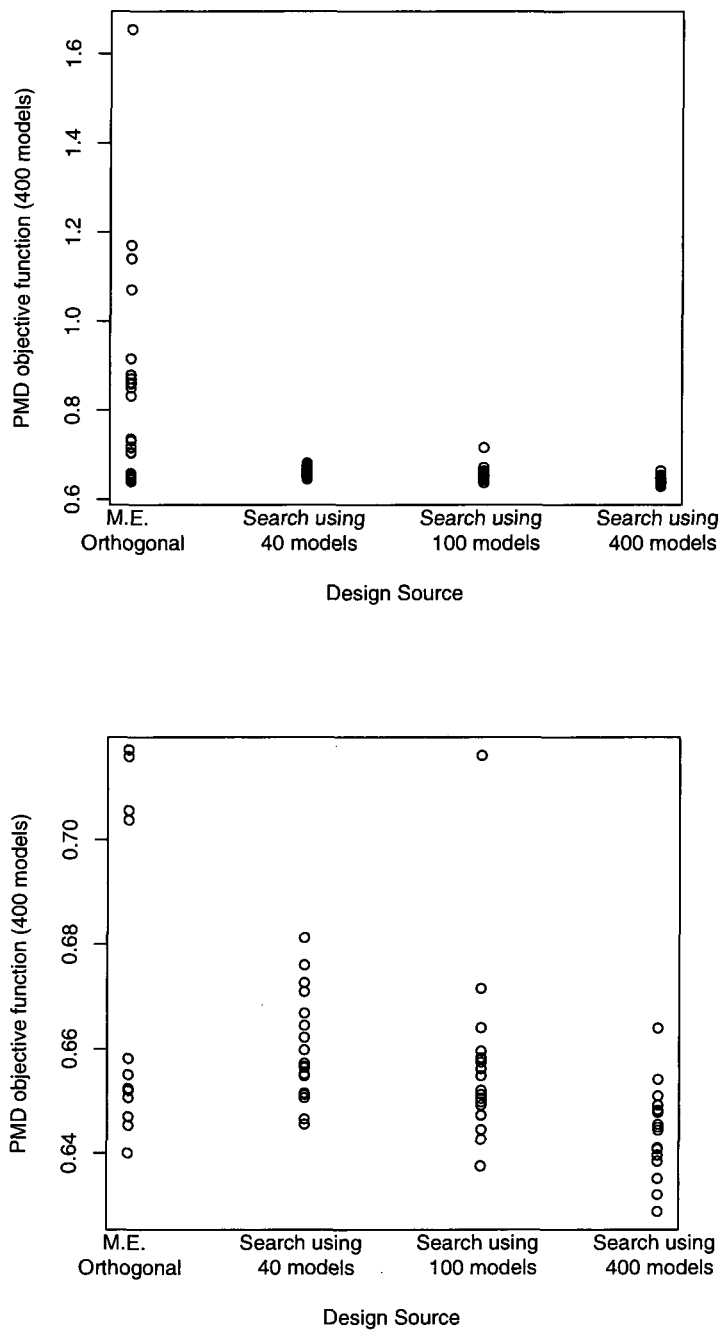

Figure 3.9: Comparison of PMD objective function values for 16 run 6 factor designs from different sources. In the second plot a magnified view of the lower section of the first plot is shown.

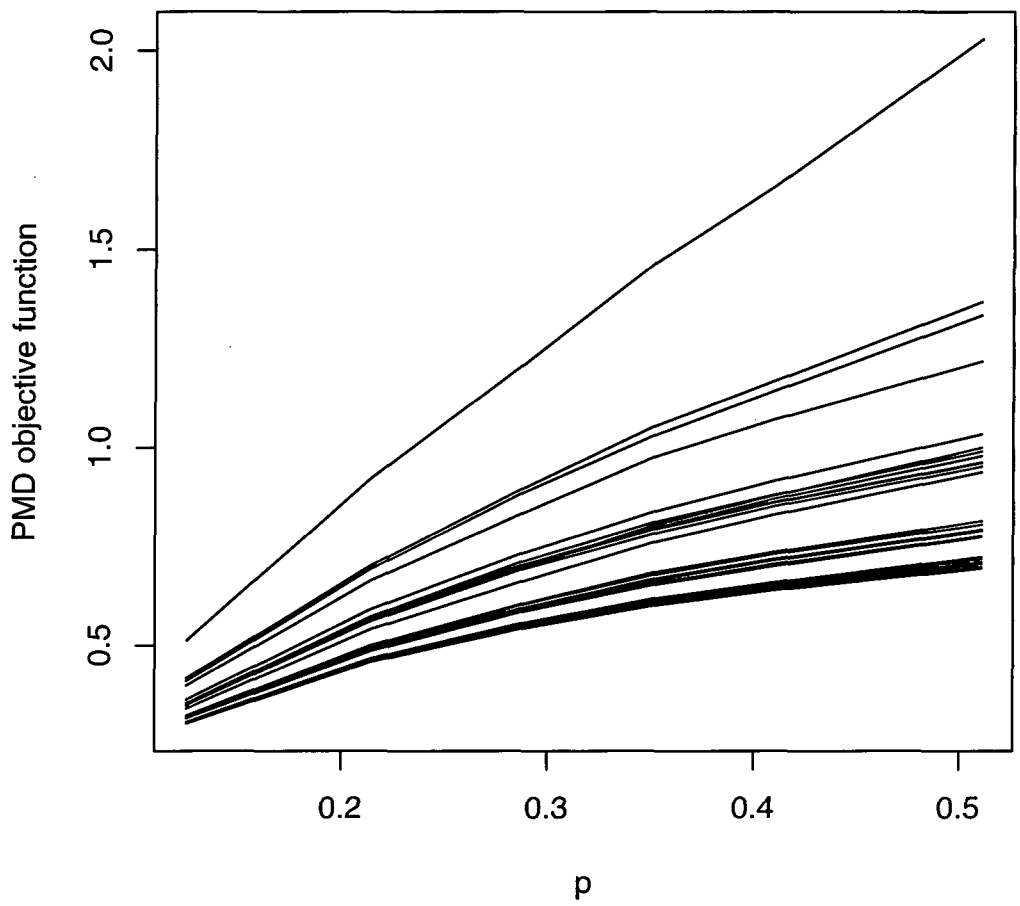

Figure 3.10: The effect of changing p on PMD objective function for 16 run 6 factor main effect orthogonal designs.

a reasonable prior probability, the ordering of the designs remains about the same. In particular, the best design remains constant over this range.

We also test the sensitivity of our criterion to changing the value of c. We may use a different value of c depending on whether we are interested in building a parsimonious model or not discarding factors that may have an important effect. Altering c between 0.25 and 4 gives no real change to the ranking of the main effects orthogonal designs - the top fifteen remain in exactly the same order. Designs returned from searches can

improve on the best main effects orthogonal design, with different designs having the lowest PMD objective function value when c is less than or greater than 1.
We can generalise our loss function to allow a different value of c to be used for main effect or interaction terms, so that our loss function may more accurately reflect the experimenter's interests. We let $c_1 = 1$ be the loss for wrongly including a main effect term and c_2 be the loss for wrongly including an interaction term. Twenty searches were performed at each of $c_2 = 0.25, 0.5, 2$ and 4. All the designs found, in addition to the main effect orthogonal designs, were evaluated using $c_2 = 0.25, 0.5, 1, 2$ and 4. Out of these designs, the main effects orthogonal design 13 was the best at all values of c_2 used; the searches did not manage to improve upon this design. The second-placed design changed as c_2 is altered. For $c_2 = 0.25$, the main effects orthogonal design 19 is second best, however, when $c_2 > 1$, a better design was one that was found in a search using $c_2 = 4$, which has less partial aliasing between the interaction terms.

3.4.6 Comparison via Simulation Studies

We would like to see how effective the designs chosen by the various criteria are for the task they are chosen to do - i.e. selecting a model. The analysis corresponding to use of a design selected by the PMD criterion is to choose the model that minimises the expected loss function (2.4) using the posterior model probabilities provided by our data. Although no analysis is explicitly given for the other criteria, one sensible analysis is to select the model with highest posterior probability.
To evaluate the performance of a design, we simulate 10000 models from the priors and generate responses from these models. Four methods are used to select a model based on the posterior model probabilities:

1. Select the model that minimises the PMD loss (2.1) using c=1.
2. Select the model with highest posterior probability, equivalent to using a 0-1 loss function.
3. As (1) but using c=0.5.
4. Select the model that minimises the expected value of the squared number of terms difference between the true and selected models (squared PMD loss).

For each of these methods of model selection, we find the percentage of time that the correct model (the one that was used to generate the data) is selected, the average R^2 value of the model selected, the average number of terms that are incorrectly included or excluded and average number of terms that the chosen model is wrong by compared to the true model.

For 12 runs in 5 factors, the HD criterion selects a Plackett-Burman design. A different design is chosen for the PMD, F and MD criteria. Table 3.6 shows that the Plackett-Burman design performs better in either method of model selection, selecting the correct model more often when using the model with highest posterior probability, and a model with lower PMD loss when using this method of model selection.

The best 16 run main effects orthogonal designs for 6 factors differed for the four criteria. Design 13 was best for the PMD, MD and HD criteria, whereas design 5 would be selected under the F criterion. The second placed designs were 19 under PMD and HD, 4 under F and 8 under MD. The results of a simulation on these five designs are displayed in Table 3.7. The searches under the PMD criterion described in Section 3.4.4 found some 16-run, 6-factor designs with a lower PMD objective function value than the best main effects orthogonal design. The results of a simulation on the best three of these designs are shown in Table 3.8. The simulation shows that these designs perform slightly better at model selection than the best main effects orthogonal designs. We also evaluate all the other 16 run 6 factor main effect orthogonal designs in the set provided by Sun et al. (2002), using this simulation. These designs are then ranked on their performance in the simulation, according to the proportion of time that the correct model was selected, when selecting the model with highest posterior probability and also the average PMD loss of the model selected when selecting a model to minimise this loss. Table 3.9 shows the rank correlation between simulation performance and the four design criteria, where the simulation is based on either selecting the model with highest posterior probability or the model that minimises the PMD loss function. The PMD and HD criteria correlate more strongly with the performance in the simulations than the MD and F criteria.

Table 3.6: Performance of 5 factor 12 run designs in simulation study. % correct denotes the % of times that the model selected by each of the four methods was the same as the model used to generate the data.

Performance Measure	Design		
	F-optimal	Plackett-Burman 1	Plackett-Burman 2
% correct, PMD loss (c=1)	47.35	47.09	47.21
% correct, 0-1 loss	48.2	47.9	48.0
% correct, PMD loss (c=0.5)	44.41	44.47	44.53
% correct, squared PMD loss (c=1)	45.46	45.04	45.14
Mean R^2, PMD loss (c=1)	0.71	0.71	0.71
Mean R^2, 0-1 loss	0.7	0.71	0.71
Mean R^2, PMD loss (c=0.5)	0.73	0.73	0.73
Mean R^2,squared PMD loss (c=1)	0.71	0.71	0.71
Average modal posterior model probability	0.48	0.48	0.48
Mean terms wrong by, PMD loss (c=1)	0.8	0.82	0.81
Mean terms wrong by, 0-1 loss	0.81	0.83	0.83
Mean terms wrong by, PMD loss (c=0.5)	0.84	0.86	0.86
Mean terms wrong by, squared PMD loss (c=1)	0.81	0.83	0.82
Mean terms incorrectly included, PMD loss (c=1)	0.14	0.14	0.14
Mean terms incorrectly included, 0-1 loss	0.14	0.16	0.16
Mean terms incorrectly included, PMD loss (c=0.5)	0.28	0.29	0.29
Mean terms incorrectly included, squared PMD loss (c=1)	0.15	0.15	0.15
Mean terms incorrectly omitted, PMD loss (c=1)	0.66	0.68	0.67
Mean terms incorrectly omitted, 0-1 loss	0.67	0.68	0.67
Mean terms incorrectly omitted, PMD loss (c=0.5)	0.56	0.58	0.57
Mean terms incorrectly omitted, squared PMD loss (c=1)	0.65	0.68	0.67

Table 3.7: Performance of designs 4,5,8,13,19 (16 run, 6 factor main effects orthogonal designs) under each criterion from simulation.

	Design Number				
Performance Measure, method of model selection	4	5	8	13	19
% correct, PMD loss (c=1)	46.86	49.38	53.76	55.23	55.37
% correct, 0-1 loss	48.56	51.02	54.72	55.77	55.99
% correct, PMD loss (c=0.5)	45.51	49.18	52.46	53.44	53.41
% correct, squared PMD loss (c=1)	43.58	45.38	51.29	53.46	53.54
Mean R^2, PMD loss (c=1)	0.68	0.67	0.68	0.69	0.69
Mean R^2, 0-1 loss	0.69	0.69	0.69	0.69	0.69
Mean R^2, PMD loss (c=0.5)	0.7	0.7	0.71	0.7	0.71
Mean R^2,squared PMD loss (c=1)	0.65	0.63	0.67	0.69	0.69
Average modal posterior model probability	0.49	0.51	0.54	0.56	0.56
Mean terms wrong by, PMD loss (c=1)	0.91	0.82	0.69	0.63	0.63
Mean terms wrong by, 0-1 loss	0.97	0.85	0.7	0.64	0.64
Mean terms wrong by, PMD loss (c=0.5)	0.95	0.86	0.72	0.66	0.66
Mean terms wrong by, squared PMD loss (c=1)	0.92	0.84	0.7	0.64	0.64
Mean terms incorrectly included, PMD loss (c=1)	0.21	0.15	0.12	0.1	0.1
Mean terms incorrectly included, 0-1 loss	0.26	0.2	0.14	0.11	0.12
Mean terms incorrectly included, PMD loss (c=0.5)	0.35	0.28	0.22	0.2	0.2
Mean terms incorrectly included, squared PMD loss (c=1)	0.18	0.11	0.1	0.1	0.11
Mean terms incorrectly omitted, PMD loss (c=1)	0.71	0.67	0.57	0.53	0.53
Mean terms incorrectly omitted, 0-1 loss	0.71	0.64	0.56	0.53	0.53
Mean terms incorrectly omitted, PMD loss (c=0.5)	0.61	0.58	0.5	0.46	0.46
Mean terms incorrectly omitted, squared PMD loss (c=1)	0.74	0.73	0.6	0.54	0.54

Table 3.8: Performance of best three 16-run designs found for 6 factors, from simulation.

	Design		
Performance Measure, method of model selection	Found #1	Found #2	Found#3
% correct, PMD loss (c=1)	55.31	55.11	55.05
% correct, 0-1 loss	55.8	55.69	55.63
% correct, PMD loss (c=0.5)	53.38	53.18	53.25
% correct, squared PMD loss (c=1)	53.61	53.32	53.34
Mean R^2, PMD loss (c=1)	0.69	0.69	0.69
Mean R^2, 0-1 loss	0.69	0.69	0.69
Mean R^2, PMD loss (c=0.5)	0.7	0.7	0.7
Mean R^2,squared PMD loss (c=1)	0.69	0.69	0.68
Average modal posterior model probability	0.56	0.56	0.56
Mean terms wrong by, PMD loss (c=1)	0.63	0.63	0.64
Mean terms wrong by, 0-1 loss	0.64	0.64	0.65
Mean terms wrong by, PMD loss (c=0.5)	0.66	0.66	0.67
Mean terms wrong by, squared PMD loss (c=1)	0.64	0.64	0.65
Mean terms incorrectly included, PMD loss (c=1)	0.1	0.1	0.1
Mean terms incorrectly included, 0-1 loss	0.11	0.11	0.12
Mean terms incorrectly included, PMD loss (c=0.5)	0.2	0.2	0.2
Mean terms incorrectly included, squared PMD loss (c=1)	0.11	0.11	0.11
Mean terms incorrectly omitted, PMD loss (c=1)	0.53	0.53	0.54
Mean terms incorrectly omitted, 0-1 loss	0.53	0.53	0.53
Mean terms incorrectly omitted, PMD loss (c=0.5)	0.46	0.46	0.47
Mean terms incorrectly omitted, squared PMD loss (c=1)	0.53	0.54	0.54

Table 3.9: Spearman's rank correlations to compare the rankings under each criterion with those from the simulation study, for the 16 run 6 factor main effect orthogonal designs of Sun et al. (2002).

	Method of model selection	
Criterion	Highest Posterior Probability	PMD loss
F	0.44	0.47
PMD	0.98	0.98
MD	0.33	0.36
HD	0.90	0.91

3.4.7 Choice of c

Our simulation studies may also give us some insight into the effects of changing c, and some guidance on values of c that could be used. Figure 3.11 shows the results of a simulation done on main effects orthogonal design 13 using the top 400 models for $p = 0.410$. Although the full set of models for this value of p gives the expected number of effects as 5, in this reduced set the expectation is about 2.21.

It can be seen that the losses due to the two types of error are not evenly balanced for $c = 1$ - many more terms are incorrectly omitted than are spuriously included. This is due to the high prior probability of models with small numbers of terms and a lack of information on which terms are active provided by the design. Using a value of $c \approx 0.334$ (calculated from the simulation) gives, on average, an equal number of terms (0.377) incorrectly missed or included. The total number of terms wrong is therefore on average 0.754, compared to 0.625 for $c = 1$.

Because the average number of model terms is 2.21, and there are 21 possible model terms (not including the intercept), the average number of terms wrong is

$$\begin{aligned} &2.21P(\text{ Term not included } | \text{ Term is in true model }) + \\ &(21 - 2.21)P(\text{ Term included } | \text{ Term not in true model }). \end{aligned} \tag{3.11}$$

Using $c = 0.334$, the two terms in this sum are equal. If, however, we would like the two conditional probabilities to be equal, we should use $c \approx 0.082$ (from simulation), which

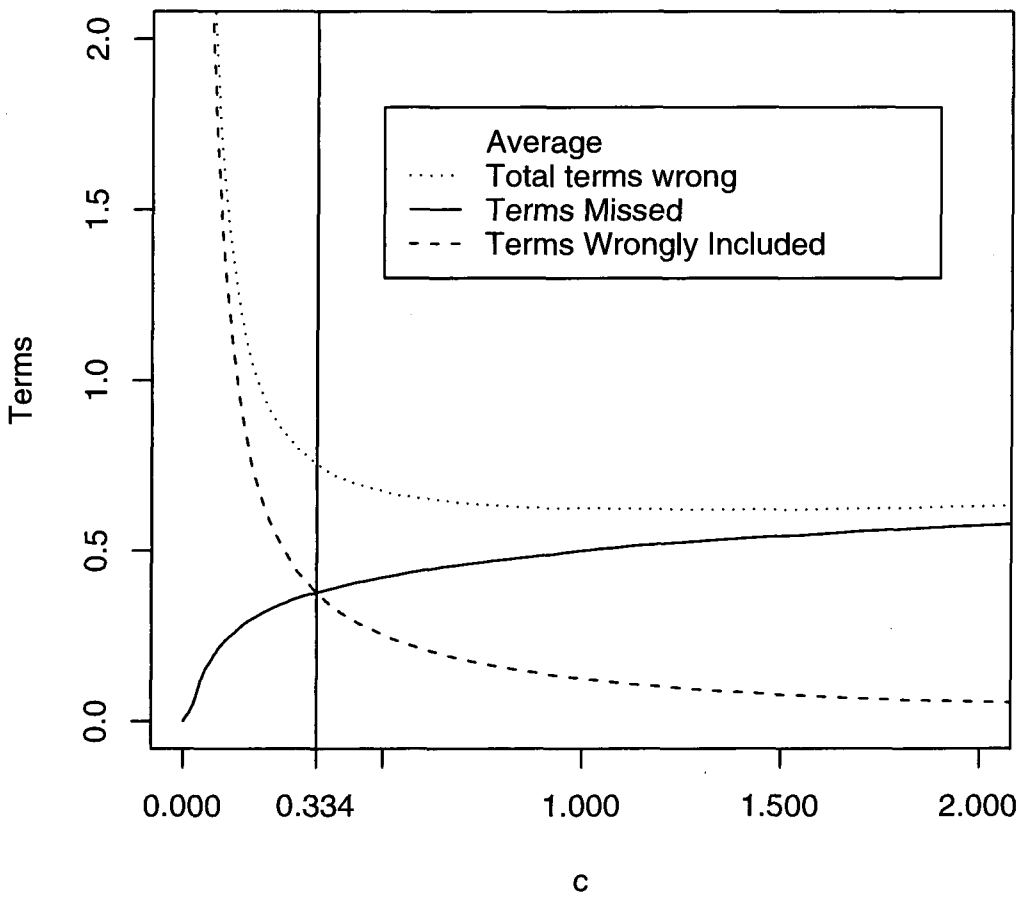

Figure 3.11: The number of terms incorrectly included or excluded as c changes, from simulation, for main effect orthogonal design 13.

gives

$$\begin{aligned} P(\text{ Term not included } \mid \text{ Term is in true model }) &= \\ P(\text{ Term included } \mid \text{ Term not in true model }) &= 0.092. \end{aligned} \tag{3.12}$$

This is the intersection point in Figure 3.12, which shows how the conditional probabilities of incorrectly including or not including a term vary with c. If are 'equally interested in avoiding both types of error', the intuitive choice of $c = 1$ may in fact be too high, depending on what we mean by this statement. The values of c given here are specific to this design and set of models. However, they may be determined using the posterior model probabilities obtained when simulating to evaluate the PMD objective function, so do not require much additional computational expense.

Alternatively, if there are actual financial losses associated with missing or incorrectly including terms, these can be used to set c. For example, retaining spurious terms in our model means that future experiments will require more runs to enable the extra parameters to be estimated. However, omitting a term may mean that possible process improvements may be missed, which has associated financial cost. We should then set c to be the cost, per term, of the extra runs required in future experiments divided by the expected cost, in terms of not being able to optimise the response, of missing a term from the model.

3.4.8 Main Effects Orthogonal Designs in 3 to 9 factors.

For each of $f = 3, \ldots, 9$ factors, we obtain the 400 models with highest prior probability, using p such that the expected number of active effects is $f - 1$. Using this model space, with the corresponding probabilities re-standardised to sum to 1, we evaluate the 16-run main effects orthogonal designs in the catalogue provided by Sun *et al.*, under all four criteria. The ranking of the designs under each criterion is given in Table 3.10.

When three to five factors are used, a regular design is best under all four criteria, as in the previous example (see Table 3.1). For six to nine factors, the best designs under the PMD and HD criteria are similar, and the designs ranked highly by the F criterion are also ranked highly under the MD. However, designs that are good under the PMD and HD criteria do not usually rank highly under the F and MD criteria.

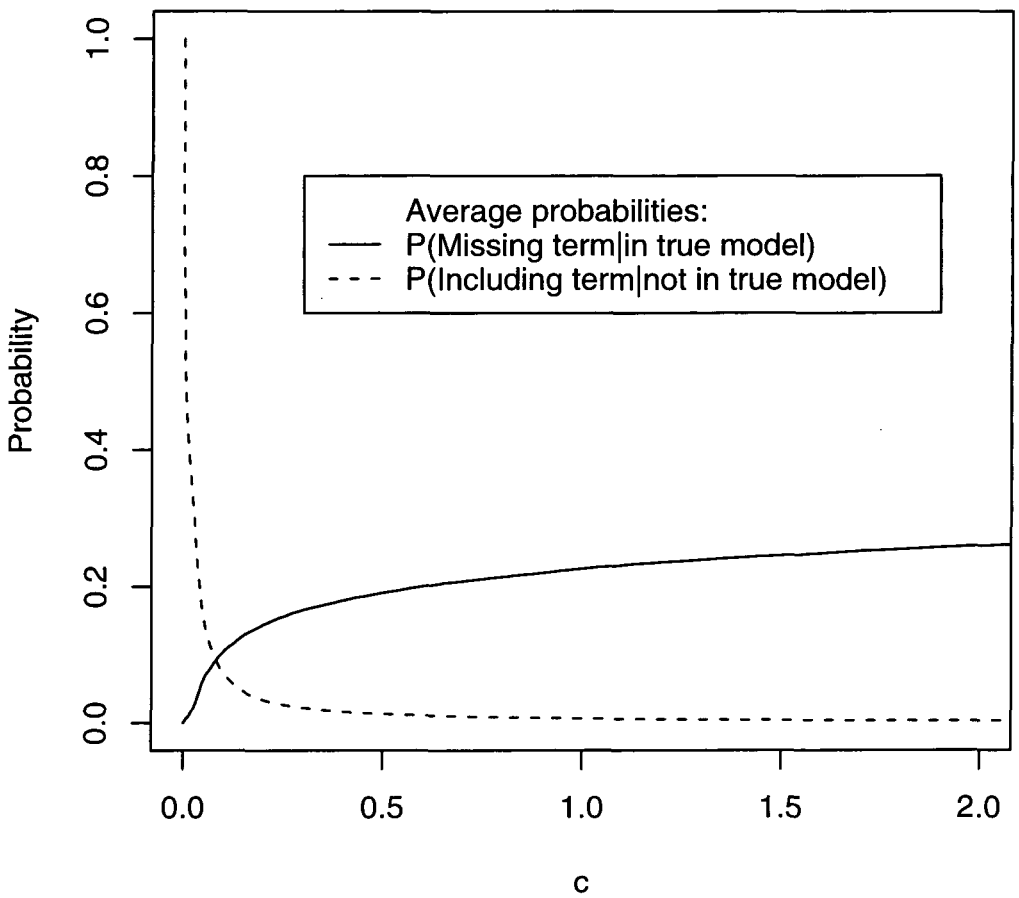

Figure 3.12: The probability of terms being incorrectly included or excluded as c changes, from simulation, for main effect orthogonal design 13.

Table 3.10: Ranking of 16-run main effects orthogonal designs in 3 to 9 factors under all four criteria, using Bingham and Chipman's model space *Full factorial, †Regular design.

Number of	3				4				5				6			
factors	PMD	F	MD	HD	PMD	F	MD	HD	PMD	F	MD	HD	PMD	F	MD	HD
Ranking 1	2*	2*	2*	2*	3*	3*	3*	3*	4†	4†	4†	4†	13	13	5†	13
2	3	3	3	3	4	4	4	4	5	5	3	5	19	4†	8	19
3	1	1	1	1	5	5	2	5	7	8	5	7	24	14	4†	8
4					2	2	5	2	8	7	7	8	20	8	14	24
5					1	1	1	1	10	10	8	3	22	6	13	20
6									11	3	2	10	23	19	19	14
7									3	2	10	11	18	24	6	22
8									9	11	11	9	26	5	12	23
9									2	9	9	2	27	20	7	12
10									6	6	6	6	8	15	24	18

Number of	7				8				9			
factors	PMD	F	MD	HD	PMD	F	MD	HD	PMD	F	MD	HD
Ranking 1	32	12	6†	32	67	6†	6†	77	76	4†	4†	84
2	49	6†	12	49	68	18	18	42	84	25	25	83
3	55	28	28	28	72	4	4	67	83	53	53	65
4	53	32	5	55	77	77	17	68	74	17	17	76
5	43	21	11	53	76	42	77	76	75	84	84	74
6	54	11	33	43	42	17	48	61	65	44	10	75
7	39	33	32	33	50	48	26	79	72	22	5	86
8	50	5	22	21	71	26	42	50	73	32	32	72
9	45	22	21	54	70	12	12	41	86	5	22	66
10	52	49	49	31	69	58	58	71	66	10	44	87

3.5 Summary

In this chapter, we have investigated the use of the PMD criterion for selecting designs for screening experiments. Three model spaces that could be used in a screening situation have been tried, allowing either main effects only, all main effects plus one 2-factor interaction, or any subset of main effects and 2-factor interactions. For the model space consisting of models containing any combination of the main effect terms, we have run Modified Federov Algorithm searches for various numbers of factors and runs. For numbers of runs where main effects orthogonal designs are available, these are the best designs found under the PMD criterion for this model space.

The model space consisting of models that contain all main effect terms plus exactly one 2-factor interaction was investigated, by ranking the set of 16-run main effects orthogonal

designs in 3 to 9 factors and comparing this to the MD, HD and F criteria. Searches for good 16-run designs were also run, and for 6 or more factors, it was possible to find non-orthogonal designs that performed better than the best orthogonal design. For this model space, we have also shown how changing the parameters of the prior distribution affects the objective function and the ranking of main effect orthogonal designs. Finally, we use a much larger space of possible models, allowing any combination of main effect and interaction terms, as used by Bingham and Chipman (2007). This model space is approximated by using a subset of models with high prior probability. The 16-run main effect orthogonal designs in 3 to 9 factors are evaluated and ranked, and the results compared to those for the MD, HD and F criteria. We notice that agreement between the rankings of designs under the four criteria lessens as the number of factors is increased. Simulation studies have been used to evaluate the performance of designs chosen under the four criteria against several indicators of suitability for model selection. The studies show that the PMD criterion agrees closely with the performance for the design in model selection, as does the HD. For the examples used in the simulation study, the four criteria select quite similar designs. There were only small differences in the performance of the designs in the simulation study. If an example was found where very different designs were selected by each criterion, a simulation study could provide useful information about the strengths and weaknesses of designs selected using each criterion. The simulations also gave an indication of what values of c should be used, and suggest that using $c = 1$, though intuitive, may be too large if we are equally interested in identifying active effects and omitting inactive effects from our model.

Chapter 4

Follow-up Experiments

In this chapter we investigate the use of the PMD criterion in the practical situation where prior information is available from an initial experiment of n runs and it is desired to select a further n^* follow-up runs that provide as much further information as possible to enable better discrimination between the models. Following Meyer *et al.* (1996), an injection moulding example from the literature is used to make a comparison of the PMD criterion with the MD, F and HD criteria, which were defined in Section 1.3.1. The performances of various different designs under each of the four criteria are compared, and underlying reasons for the choice of particular designs are discussed. A second example on a chemical reactor, where data from a full factorial is available, is used to investigate the sensitivity of our methodology to the choice of design for the initial experiment. In addition, the results from the analysis are used to compare the conclusions from a PMD design with those from an MD design (given by Meyer *et al.* 1996). Throughout this chapter we have made use of the R package BsMD (Barrios 2004) for producing graphs of factor probabilities, evaluating the MD objective function and searching for good designs under the MD criterion.

4.1 First Example

Box, Hunter and Hunter (1978) described an experiment for an injection moulding process, where 8 factors were varied - see Table 4.1. The aim of the experiment was to

Table 4.1: Design and Results for the Injection Moulding Example from Box, Hunter and Hunter (1978, p398)

A	B	C	D	E	F	G	H	y
-1	-1	-1	1	1	1	-1	1	14.0
1	-1	-1	-1	-1	1	1	1	16.8
-1	1	-1	-1	1	-1	1	1	15.0
1	1	-1	1	-1	-1	-1	1	15.4
-1	-1	1	1	-1	-1	1	1	27.6
1	-1	1	-1	1	-1	-1	1	24.0
-1	1	1	-1	-1	1	-1	1	27.4
1	1	1	1	1	1	1	1	22.6
1	1	1	-1	-1	-1	1	-1	22.3
-1	1	1	1	1	-1	-1	-1	17.1
1	-1	1	1	-1	1	-1	-1	21.5
-1	-1	1	-1	1	1	1	-1	17.5
1	1	-1	-1	1	1	-1	-1	15.9
-1	1	-1	1	-1	1	1	-1	21.9
1	-1	-1	1	1	-1	1	-1	16.7
-1	-1	-1	-1	-1	-1	-1	-1	20.3

Factors

A Mould Temperature
B Moisture Content
C Holding Pressure
D Cavity Thickness
E Booster Pressure
F Cycle time
G Gate Size
H Screw Speed

Response

y Shrinkage (%).

identify those factors that had an important effect on the percentage shrinkage in an injection moulding process. The design used for the experiment was a 2^{8-4}_{IV} fractional factorial design, with generators $I = ABDH = ACEH = BCFH = ABCG$.
Meyer *et al.* (1996) analysed the results of this experiment using a Bayesian approach, discussed below, and subsequently found a follow-up design using the MD criterion of Box and Hill (1967), defined in Section 1.3.1. The set of possible models they considered was all linear models which contain

i any subset of factor main effects and also

ii all possible 2- and 3-factor interactions between those factors whose main effects are included in the model.

The automatic inclusion of interactions in a model in (ii) is known as 'effect forcing'. Meyer *et al.* used prior probabilities which differ slightly from those in Section 1.1.3. These prior distributions will be used in this chapter. The prior model probabilities are calculated under the assumption that each factor has the same probability p of being

Table 4.2: Posterior model probabilities for the initial experiment

Model Number	Factors in Model	Posterior Probability
m_1	C	0.0002
m_2	C E	0.0004
m_3	A C E	0.2356
m_4	A C H	0.2356
m_5	A E H	0.2356
m_6	C E H	0.2356
m_7	A C E H	0.0566

active, independently of each other, so that

$$P(m_i) = p^{f_i}(1-p)^{k-f_i}, \tag{4.1}$$

where f_i is the number of factors in model m_i and $k = 8$ is the total number of factors. Meyer *et al.* used $p = 0.25$. An improper prior for the overall mean β_0 and standard deviation σ is used, so that $f(\beta_0, \sigma) \propto \sigma^{-1}$. This is equivalent to using values of $a = 0$ and $d = -1$ for the hyperparameters in the inverse Gamma distribution, and allowing the prior variance of β_0 to tend to ∞. The rest of the coefficients in $\boldsymbol{\beta}_i$ are given Normal prior distributions with mean 0 and standard deviation $\gamma\sigma$, where γ is a scaling parameter estimated by Meyer *et al.* to be 2 from the data. As in Chapter 1, the posterior distribution for model m_i is still a Normal Inverse Gamma distribution, with

$$\begin{aligned}
\mathbf{V}_i^* &= \left(\frac{1}{\gamma^2}\begin{pmatrix} 0 & 0 \\ 0 & \mathbf{I}_{p_i} \end{pmatrix} + \mathbf{X}_i'\mathbf{X}_i\right)^{-1} && (4.2)\\
\boldsymbol{\mu}_i^* &= \mathbf{V}_i^*\mathbf{X}_i'\mathbf{y} && (4.3)\\
a_i^* &= \mathbf{y}'\mathbf{y} - (\boldsymbol{\mu}_i^*)'(\mathbf{V}_i^*)^{-1}\boldsymbol{\mu}_i^* \\
d^* &= n - 1.
\end{aligned}$$

From the results of the initial experiment, seven models were found to account for 99.98% of the posterior probability, as shown in Table 4.2. All these models contain only factors A, C, E and H. The aliasing scheme of this design (in particular the relation

Table 4.3: Posterior means for the model parameters in 3-factor models after the initial experiment (reproduced from Meyer *et al.* (1996))

Parameter	Posterior Mean
I	19.75
A = CEH	-0.17
C = AEH	1.364
E = ACH	-0.94
H = ACE	0.297
AC = EH	0.223
AE = CH	1.14
AH = CE	-0.149

$I = ACEH$) and the effect forcing prior distribution result in models $m_3, \ldots, m_6$ being indistinguishable in that they have the same posterior model probabilities. This is because models $m_3, \ldots, m_6$ have the same model matrix $\mathbf{X}$ but with columns assigned to different explanatory variables. Hence the parameters have the same posterior distributions but are attributed to different effects in each model. The posterior means for the regression parameters in the three-factor models are also identical, but assigned to different effects. These are summarised in Table 4.3.

4.1.1 Searching for Follow-up Runs

As models containing only factors A, C, E and H account for virtually all the posterior probability, following Meyer *et al.* (1996), we searched for a set of $n^* = 4$ follow-up runs, from a list of candidate points composed of all combinations of these factors at levels ± 1. The points in the candidate list are given in Table 4.4 together with numerical labels. Note that points 1-8 are combinations of levels of the factors A, C, E, H which have already been run in the initial 8 factor experiment. It is assumed that the remaining four factors are inactive, and can be set to some convenient level in the follow-up runs. As an alternative to the Bayesian criteria discussed here, there are also frequentist methods for selecting follow-up runs. For example, Box *et al.* (1978) construct a follow-up design for this experiment using the methods of Daniel (1962) to algebraically remove the aliasing between pairs of effects. The semi-foldover, or $\frac{3}{4}$ designs of John (1971) are another way to augment a half fractional factorial design without using all the runs required to

Table 4.4: Candidate points for the injection moulding experiment

Point Number	A	C	E	H
1	-1	-1	-1	-1
2	-1	-1	1	1
3	-1	1	-1	1
4	-1	1	1	-1
5	1	-1	-1	1
6	1	-1	1	-1
7	1	1	-1	-1
8	1	1	1	1
9	-1	-1	-1	1
10	-1	-1	1	-1
11	-1	1	-1	-1
12	-1	1	1	1
13	1	-1	-1	-1
14	1	-1	1	1
15	1	1	-1	1
16	1	1	1	-1

complete a full factorial. The three best sets of follow-up runs under the MD criterion were found by Meyer *et al.* to be $(9, 9, 12, 15)$, $(9, 12, 14, 15)$ and $(9, 11, 12, 15)$.

4.2 Comparison between the PMD and MD Criteria

The PMD criterion, defined in Section 2.2 with $c = 1$, and the search algorithm outlined in Section 2.4 were used to find sets of four follow-up runs to compare with those of Meyer *et al.* (1996). The prior distributions for the models and parameters were taken to be the posterior distributions from the analysis of the initial experiment of Meyer *et al.* Table 4.5 gives the best sets of four follow-up runs we found. Also shown are the values of the PMD and MD objective functions for the best follow-up designs under each of the criteria.

The table shows that the best follow-up designs differ. However, the best design under PMD is the third best under the MD criterion. The best three designs under the PMD criterion consist of four distinct design points. This contrasts with the best design under the MD criterion, which has a point repeated. For both criteria, the best follow-up designs contain only points which were not included in the initial design, which has

Table 4.5: Best sets of four follow-up runs for the objective function values under the MD and PMD criteria; for design point labels, see Table 4.4

Found under MD criterion		
Runs	MD	PMD
9 9 12 15	85.7	0.080
9 12 14 15	84.4	0.054
9 11 12 15	83.6	0.041
Found under PMD criterion		
Runs	MD	PMD
9 11 12 15	83.6	0.041
11 12 15 16	47.2	0.042
10 11 12 15	50.4	0.044

$I = ACEH$ as a defining contrast. Thus they have the main effect of A completely aliased with the interaction $-CEH$, which is the 'reverse' of the aliasing in the design used in the initial experiment. When we analyse the follow-up experiment, using the posterior from the initial experiment as a prior, we are effectively analysing a 20-run experiment. Therefore adding four points from a complement of the initial design breaks the aliasing in the initial design. This means that the follow-up runs provide information on which effects in a pair of effects that were aliased after the initial experiment is in the true model. For example, analysis of the data from the initial experiment gives a comparatively large positive posterior mean (Table 4.3) for the parameter corresponding to $C = AEH$ in the equivalent 3-factor models labelled $m_3, \ldots, m_6$ in Table 4.2. For all the designs in Table 4.5, the points at which $C = 1$ will have $AEH = -1$ (where AEH denotes multiplication of columns in Table 4.4) and vice-versa. Suppose also that higher values of the response, $\mathbf{Y}^*$, are obtained at the runs of the follow-up experiment at which $C = 1$. Then the posterior mean from the follow-up runs will remain positive for C in those models which contain C, and will move towards zero for AEH in the models containing this term. Hence, the posterior probability will be increased for models that contain C and will be decreased for those containing AEH.

Some insights can be gained into how the results under the PMD criterion with $c = 1$ relate to those obtained from the 0,1 loss function (Equation 2.2). From the results of the initial experiment, the four most probable models all differ from each other by

exactly eight terms. This can be seen from Table 4.2 when the impact of effect forcing is taken into account. Therefore, the expected loss for one of these models after the follow-up experiment will be approximately 8 times the sum of the posterior probabilities of the rest of models $m_3, \ldots, m_6$ in Table 4.2. Hence, if the total of the posterior probabilities of the other models is negligible, then the minimum value of (2.1) is

$$\begin{aligned}
\min_i E(L(m_i, m_j)|\mathbf{Y}) &= \min_i \sum_{m \in \mathcal{M}} P(m_i|\mathbf{Y}) L(m_i, m_j) \\
&\approx \min_{3 \le i \le 6} \sum_{j=3}^{6} P(m_i|\mathbf{Y}) L(m_i, m_j) \\
&= \min_{3 \le i \le 6} \left[8 \sum_{j \neq i} P(m_j|\mathbf{y}) \right] \\
&= 8 \left\{ 1 - \max_i \left[P(m_i|\mathbf{y}) \right] \right\}.
\end{aligned}$$

This indicates that our approach will, for this example, give similar results to choosing the model with highest posterior probability, that is, the model found using the 0,1 loss function.

In order to provide a more general comparison of the PMD criterion with that of Meyer et al. (1996), we randomly generated four-point follow-up designs from the list of candidate points (Table 4.4), and evaluated and ranked them under both criteria. The results are plotted in Figure 4.1. This figure also includes the ranks of all designs composed of four distinct points that were not included in the original design. The colours identify the number of new points used (that is, not in the initial design), and different plotting characters show the number of distinct points in the design.

The plot indicates a general trend that designs which are good under one criterion tend to be good under the other (rank correlation = 0.92). We observe from the evaluated designs that, designs which perform well under the PMD criterion are those which have four distinct points. This is not necessarily the case for the MD criterion. For both criteria, the better performing designs contain no points that are replicates of the runs in the initial design.

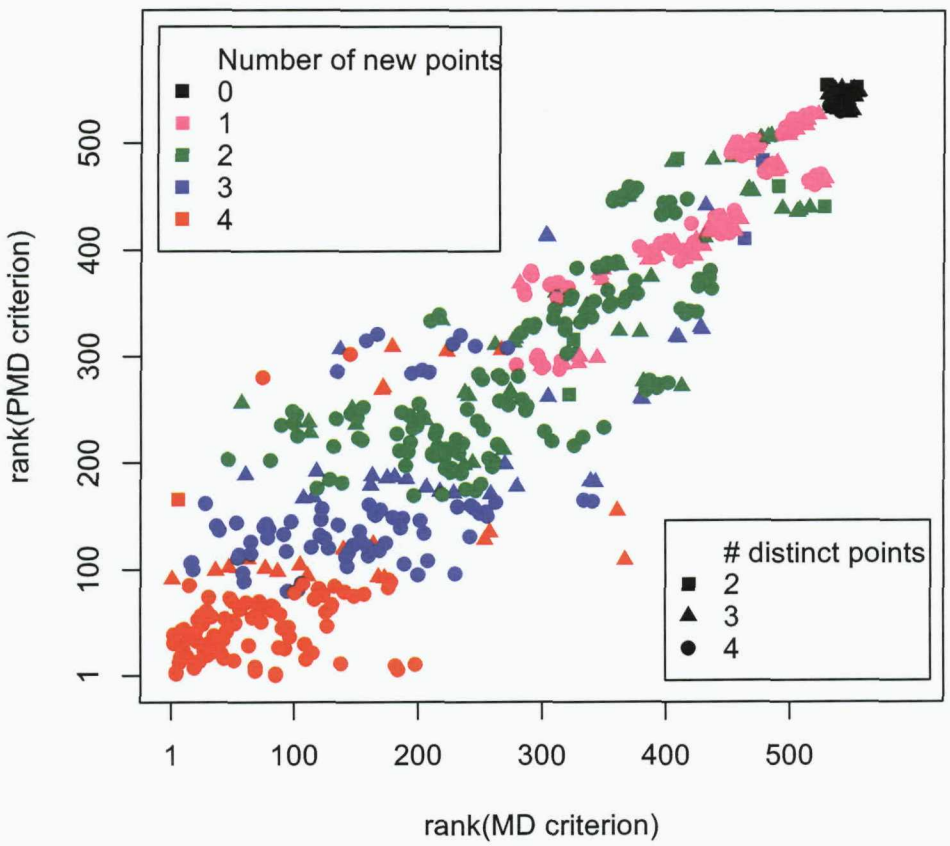

Figure 4.1: Ranking of MD vs PMD objective function values for random sample of four-point follow-up designs formed from candidate points in Table 4.4, plus all designs formed of four distinct new points.

4.3 Comparison between the PMD and F Criteria.

In the discussion of the Meyer *et al.* (1996) paper, Jones and DuMouchel proposed an alternative to the MD criterion, called the F criterion (described in Section 1.3.1). In Figure 4.2 the rankings are plotted for the same set of designs as in Figure 4.1 under the PMD and F criteria. The plot indicates a slightly stronger relationship between these criteria (rank correlation = 0.94) than between the PMD and MD. It is clear that the best designs under both criteria are those formed entirely of distinct points which are not present in the first stage design.

Figure 4.2 shows that some designs perform poorly under the PMD criterion compared to other designs with the same number of new and distinct points, and a similar F value. These tend to be designs that have either A assigned level 1, or H assigned level -1 for each of the new design points. For example, the design ranked 66th under the F criterion and 280th under the PMD criterion consists of candidate points 13,14,15 and 16 from Table 4.4, and has factor A set to level 1 at all four points. These sets of follow-up runs are plotted in red in Figure 4.3, and can be identified as being above and to the left of the trend line in Figure 4.2. These are obviously undesirable in terms of estimating the associated main effects. Also such a follow-up design may lead to main effect terms cancelling with 2-factor interactions and making their partially aliased joint effect close to 0, so that it is hard to tell if either effect is important. Designs with either A set to 1 for all points or H assigned level -1 for all points perform badly even when compared to other designs with one factor kept at a constant level. If a follow-up design has, for example, A equal to 1 at all points, then the columns of the model matrix representing the terms C and AC will be identical, as will be the columns for E and AE, and the columns for H and AH. The confounding between the terms for E and AE, and between H and AH give rise to cancelling. The prior means for the coefficients of E and AE in the three-factor models are -0.94 and 1.14 respectively (refer to Table 4.3 for the prior means of the effects at the second stage). Therefore, if A is set to level 1, the average difference between the expected response at the high and low levels of E will be $2 \times (-0.94 + 1.14) = 0.40$, much less than the $2 \times -0.94 = 1.88$ it would be without the cancelling with the AE interaction.

Table 4.6 shows the difference between the expected response at the high and low levels

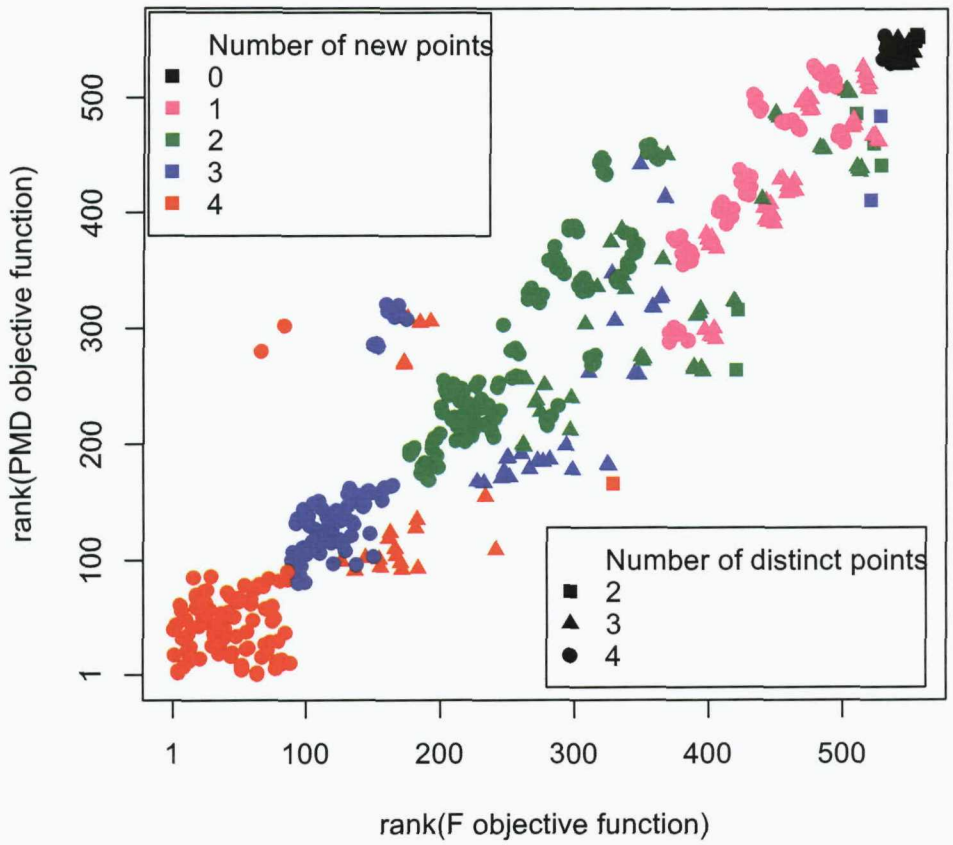

Figure 4.2: Ranks under F and PMD criteria for a random sample of designs formed from candidate points in Table 4.4, plus all designs formed of four distinct new points.

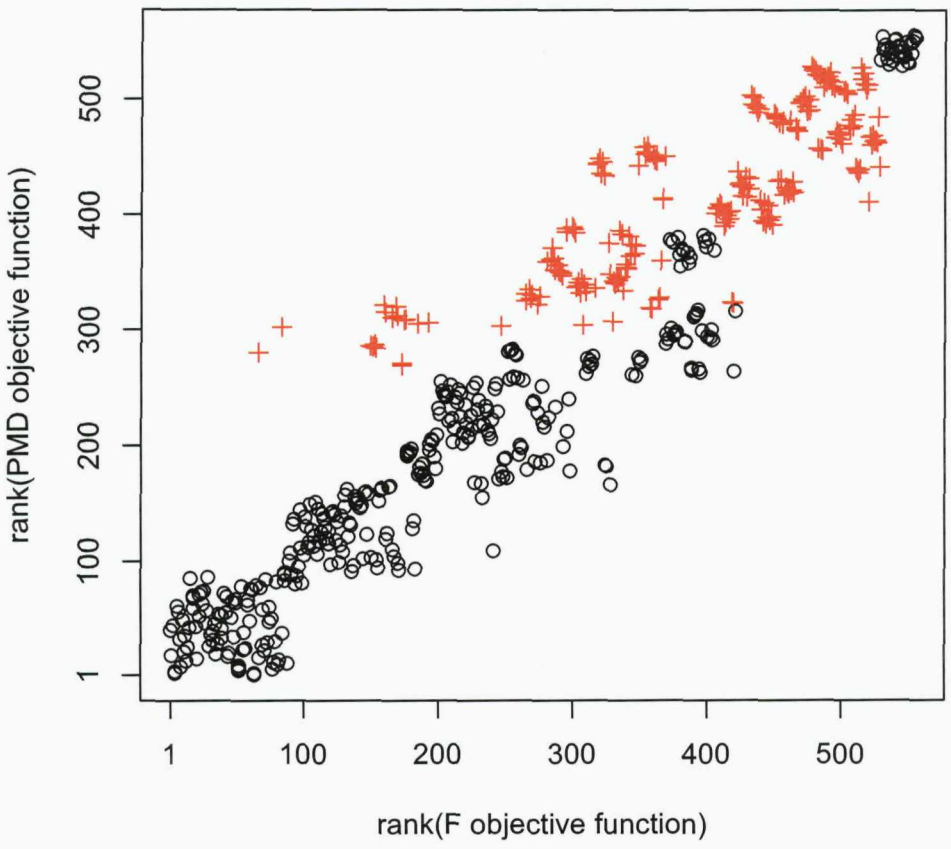

Figure 4.3: Ranking under F and PMD Criteria for the Set of Designs used in Figure 4.2. Designs in Red have either A set to 1 or H to -1 for all new points.

Table 4.6: Apparent sizes of effects for follow-up designs with different terms aliased.

Aliasing	Difference in $\mathbf{X}\beta$ between high/low values of factor.				
	A	C	E	H	Average Absolute Value
A=1	-0.34	3.16	0.40	0.32	1.06
A=-1	-0.34	2.28	-4.16	0.88	1.92
C=1	0.10	2.72	-2.16	2.88	1.96
C=-1	-0.78	2.72	-1.60	-1.68	1.70
E=1	1.94	2.44	-1.88	1.04	1.82
E=-1	-2.62	3.00	-1.88	0.16	1.92
H=1	-0.62	5.00	-1.44	0.60	1.92
H=-1	-0.06	0.44	-2.32	0.60	0.86

of each factor, due to the main effect and aliased interaction terms, for having each possible aliasing of a main effect with the mean. Designs for which A is equal to 1, or H to -1 lead to these differences being small, on average, so detecting the presence of main effect terms is more difficult, leading to greater uncertainty and a higher PMD value. This is reflected in the ranks of these designs under the PMD criterion, but not as much under the F criterion. Example of such designs are the two designs with four distinct new points that are separated from the rest in Figure 4.2, ranked 66 and 82 under the F criterion but 280 and 302 respectively under the PMD.

4.4 Comparison to the HD criterion.

We use the same set of randomly chosen designs as before to compare the HD with the PMD criterion, see Figure 4.4. The rank correlation is 0.97 for these two criteria, the strongest correlation with PMD of all the criteria. The HD criterion selects the same best design as the PMD criterion, and, in contrast to the MD and F criteria, designs with A set to 1 or H set to -1 at all new points are given a low ranking, as under the PMD criterion.

4.5 Second Example

A second example of selecting follow-up runs to build on the information from a first-stage experiment was also investigated by Meyer *et al.* The example uses data from a 2^5 full factorial experiment involving a chemical reactor, described by Box *et al.*

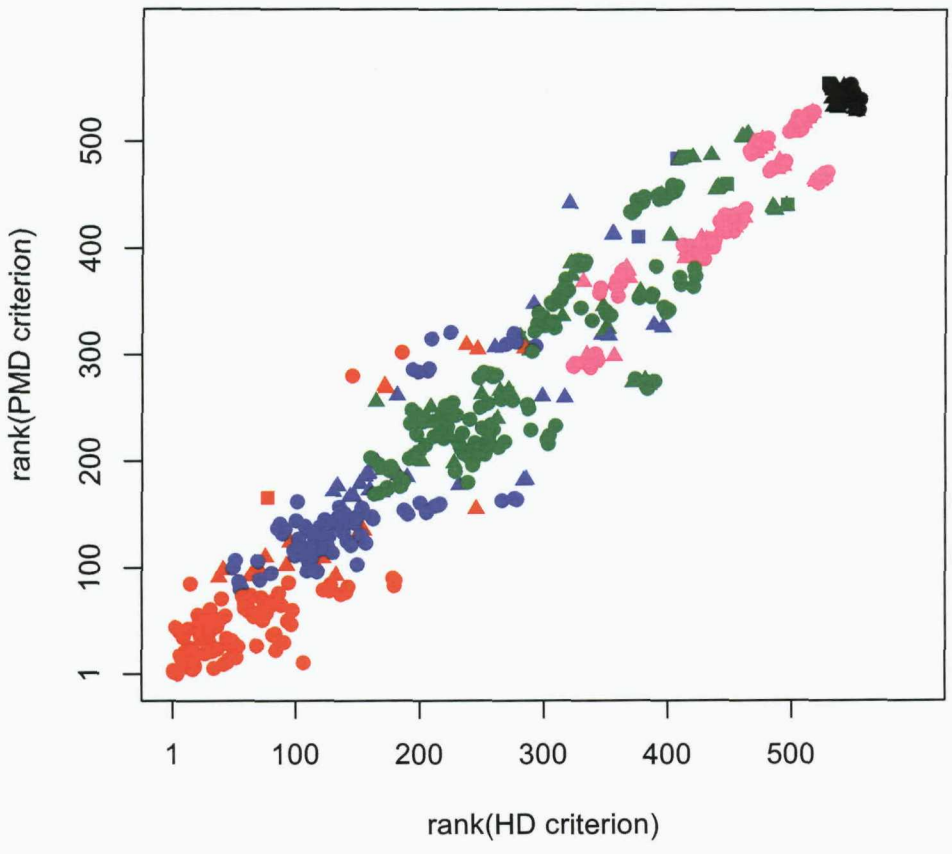

Figure 4.4: Ranking of HD vs PMD objective function values for random sample of four-point follow-up designs formed from candidate points in Table 4.4, plus all designs formed of four distinct new points.

Table 4.7: Top 10 models after first stage for second example, as given in Meyer *et al.* (1996).

Probability	Model
0.231	I
0.134	I B
0.075	I D
0.070	I A
0.055	I A D AD
0.055	I A B AB
0.055	I B D BD
0.052	I E
0.051	I C
0.032	I B C BC

(2005). The eight runs are a 2^{5-2}_{III} design with defining relation $I = ABD = ACE$. Meyer *et al.* used the MD criterion to select a four-run follow-up design with points chosen from the full factorial. The responses obtained at these points in the actual experiment are known. Hence this example has the advantage of allowing an indication of how the design would have performed in practice to be gained. The results from the initial 8-run experiment and four follow-up runs may also be compared to those from using all 32 runs.

4.5.1 Initial Experiment

The prior distributions of models and parameters are set up as in Section 4.1 except that in order to make a direct comparison with the results of Meyer *et al.* (1996) the value $\gamma = 0.4$ was used. In the analysis of the first eight runs, weak evidence was found for factors B, D and E being active, but it was inconclusive. The top 10 models are given in Table 4.7 and the effect probabilities in Table 4.8.

Unlike our first example, there is insufficient evidence to justify dropping any of the factors. We made 20 tries of the search algorithm described in Section 2.4 for a 4-run follow-up design under the PMD criterion. The five best follow-up designs, together with the best five under the MD criterion (from Meyer *et al.*, Table 8) were then evaluated under each criterion. Table 4.9 lists the follow-up points together with the values of the two objective functions. The labels of the $2^5 = 32$ possible design points are given in

Table 4.8: Effect probabilities from the runs of the 2^{5-2}_{III} initial experiment in the second example

Effect	Probability
I	1
A	0.27
B	0.38
C	0.17
D	0.29
E	0.17
AB	0.10
AC	0.04
AD	0.10
AE	0.04
BC	0.07
BD	0.10
BE	0.06
CD	0.05
CE	0.04
DE	0.05
ABC	0.02
ABD	0.03
ABE	0.02
ACD	0.02
ACE	0.01
ADE	0.02
BCD	0.02
BCE	0.01
BDE	0.02
CDE	0.02

Appendix C.

It can be seen from the table that there are no designs that are in the top 5 under both criteria. However, designs that are good under one of the criteria are generally good under the other. Figure 4.5 shows the PMD and MD objective functions for 500 designs randomly selected from the 32 candidate points. There is a strong correlation (-0.8) between the two objective functions; all the designs listed in Table 4.9 are in the top 2% of all possible 4-run designs under each of the criteria.

Table 4.9: Best follow-up runs for the second example under the PMD and MD criteria

Search Criterion	Follow-up runs	PMD	MD
PMD	2 4 10 12	1.951	0.549
PMD	25 26 27 28	1.955	0.529
PMD	4 10 12 18	1.957	0.545
PMD	18 20 26 28	1.961	0.504
PMD	9 10 12 27	1.966	0.560
MD	4 10 11 26	1.972	0.615
MD	4 10 11 28	1.975	0.610
MD	4 10 26 27	1.967	0.608
MD	4 10 12 27	1.971	0.606
MD	4 11 12 26	1.975	0.603

4.5.2 Performance of the PMD and MD-optimal follow-up designs

In this subsection, we follow Meyer *et al.* (1996) in evaluating designs for this example by comparing the conclusions from the Bayesian analysis of the follow-up runs with those found from the PMD criterion and also from the analysis of the full data set for the 2^5 experiment. Analysis of the full 2^5 experiment gives posterior probabilities close to 1 for the factors B, D and E and near zero for the others. The MD-optimal design gives each of these factors a probability greater than 0.6 of being active, whilst A and C each have probabilities less than 0.2, giving broad agreement with the results from the full data set. The posterior probabilities of the factors using the MD-optimal design, and the two best designs under the PMD criterion are shown in Figure 4.6, together with the probabilities obtained by using the first eight runs only (that is, the initial design). The two best designs under the PMD criterion both give higher posterior probabilities for factors B, D and E than for A and C, although the results are less conclusive than for the MD-optimal design.

The ten models with highest posterior probability using the MD-optimal and each of the two designs found using the PMD criterion are given in Tables 4.10-4.12.

A more detailed summary of the distribution of models and parameters after the first stage may be found in Table 4.13. In addition to the posterior probabilities of all model terms, we have also sampled from the model-averaged posterior distribution for the regression parameters, using methodology described in more detail in Section 5.6. From this we have obtained the mean of each component of $\boldsymbol{\beta}$, given that it is included in the

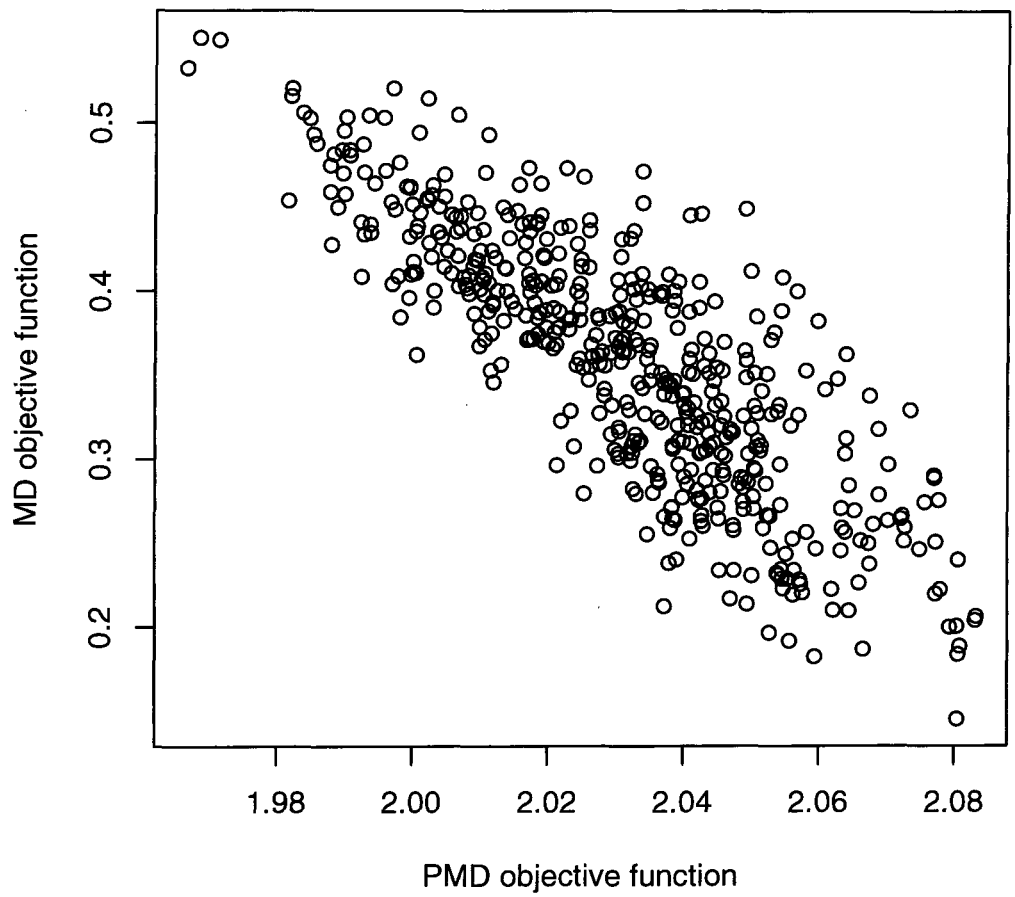

Figure 4.5: PMD and MD objective function values for 500 randomly selected designs for the second example

model, and a 95% credible interval for its value, formed from the 2.5 and 97.5 percentiles of the sampled values. This is an interval, in this case centred on a parameter's marginal posterior median, for which the probability that the parameter lies within it is 0.95. The credible intervals for all terms (except the intercept) contain zero. However, for several of the parameters the credible interval are not centred close to zero, and contains values of much higher magnitude on one side of zero than the other. For example, the evidence suggests that coefficient of B is positive.

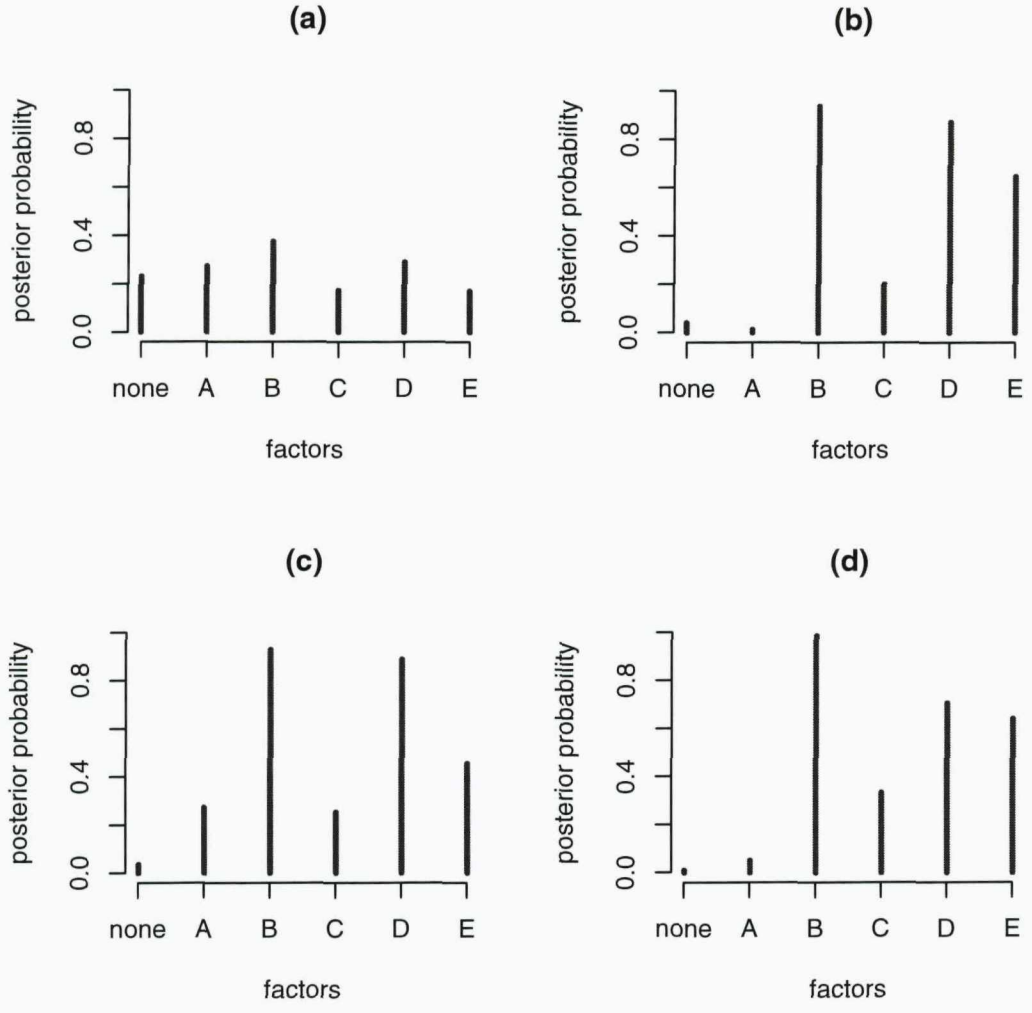

Figure 4.6: Posterior factor probabilities using different 4-run follow-up experiments (second example) from (a) 8 runs of 2^{5-2}_{III}, and with additional runs (b) 4,10,11,26, (c) 2,4,10,12, (d) 25,26,27,28.

The parameter distributions summarised in Table 4.13 suggest that the $A = BD$, $B = AD$ and $D = AB$ effects are likely to be large and positive, hence the higher probability of these factors compared to the other two. The need to estimate the parameters associated with these factors is reflected in the designs chosen by the two criteria. The MD-optimal design consists of points for which the column representing ABD is set to -1, reversing the first stage aliasing in a similar way to our first example.

Table 4.10: Factors in the 10 most probable models after using MD-optimal follow-up (4 10 11 26).

Probability	Factors in Model
0.462	B D E
0.209	B D
0.172	B C D E
0.064	B
0.041	none
0.016	B C D
0.006	D
0.005	E
0.004	B C
0.004	C

Table 4.11: Factors in the 10 most probable models after using PMD-optimal follow-up (2 4 10 12).

Probability	Factors in Model
0.283	B D
0.232	B D E
0.109	A B C D
0.095	A B D E
0.089	B C D E
0.049	B
0.036	none
0.015	B C D
0.013	D

The best design found under the PMD criterion maintains orthogonality for factors B and D, but sets A to +1 for all points and C and E to -1 for all points. The posterior distribution of the BC interaction after the first stage suggests that in it is usually negative in those models that contain it. Therefore, setting C to -1 has the advantage that the main effect of B will not be cancelled with the BC interaction - setting C to -1 gives a better chance of detecting the main effect of B, if it is active. The top five follow-up designs under both criteria have C set to -1 for all points.

The new parameter distributions obtained after using the MD-optimal follow-up runs is given in Table 4.14. Terms with a posterior probability of being active that is > 0.5 are shown in bold. Resolving the first stage aliasing and gaining extra information means that the effects with high probability are those involving factors B, D and E, not A. The 95% probability intervals for terms such as BD and DE no longer include zero, giving us

Table 4.12: Factors in the 10 most probable models after using 2nd best follow-up under the PMD criterion (25 26 27 28).

Probability	Factors in Model
0.366	B D E
0.164	B D
0.144	B C E
0.095	B C D E
0.080	B
0.065	B C D
0.018	B C
0.017	A B E
0.011	A B
0.010	none

a more definite indication of the direction of the associated effects. Summaries of the posterior parameter distributions after using the top two follow-up experiments under the PMD criterion are given in Tables 4.15 and 4.16. The best design under the PMD criterion seems to have been less successful than the MD-optimal design, as factor E has a posterior probability of being active that is < 0.5, yet an analysis of the full factorial experiment, mentioned by Meyer *et al.* (1996), shows E to be almost definitely active. The second best design under the PMD criterion does indicate that E is active.

4.5.3 Alternative initial designs

Although Meyer *et al.* use the $\frac{1}{4}$ fraction defined by $ABD = ACE$ as the starting design, we could equally have chosen any of the 15 possible regular 2^{5-2}_{III} fractional factorial designs for the initial experiment. We repeated the above investigation using each of these initial designs in turn. We observed how the initial design affects the posterior factor probabilities after the initial experiment and the best selection of follow-up runs (Table 4.17). For each initial design, twenty MFEA searches under the PMD criterion were performed and the best four-point follow-up design obtained. These are given in Table 4.18. For each design, an indication of the settings of each factor is given. Factors for which balance is retained are indicated by 'b'. Factors that take only one level in the follow-up design are denoted by -1/+1 as appropriate. Several general trends may be noticed by comparing the design to the factor probabilities after the first stage from Table 4.17. Firstly, factors with high probability after the initial fraction are often given

Table 4.13: Summary of posterior distribution of components of β after the first stage.

Term	P(In model)	Mean|in model	95% credible interval | in model	
I	1	64.595	27.415	101.741
A	0.271	3.283	-5.198	11.823
B	0.375	5.497	-2.821	13.83
C	0.172	-0.2	-9.005	8.643
D	0.291	3.575	-4.858	12.269
E	0.17	-1.016	-10.255	8.141
A B	0.103	3.282	-4.547	11.015
A C	0.038	-0.872	-8.988	7.051
A D	0.104	5.029	-2.881	12.581
A E	0.038	-0.207	-8.337	7.689
B C	0.066	-3.395	-11.06	3.911
B D	0.104	3.035	-4.591	10.392
B E	0.057	1.67	-6.039	9.371
C D	0.047	1.672	-6.048	9.06
C E	0.038	3.018	-5.222	11.258
D E	0.052	-3.244	-11.128	4.739
A B C	0.015	1.367	-5.533	8.445
A B D	0.03	0.163	-9.784	10.475
A B E	0.015	-2.929	-9.7	4.373
A C D	0.016	-2.974	-9.937	3.646
A C E	0.007	0.202	-11.005	11.899
A D E	0.015	1.564	-5.106	8.406
B C D	0.016	-0.777	-7.407	6.752
B C E	0.014	3.094	-3.842	9.755
B D E	0.016	-0.166	-6.9	6.884
C D E	0.015	4.646	-2.574	11.781

balanced follow-up runs, whereas factors with very low probability may be set to a constant level. The initial fractions defined by $ABC = BDE = 1$ and $ACE = BDE = 1$ lead to high (> 0.3) posterior probabilities for the three factors B, D and E. The follow-up runs chosen are the same in these two cases, and consist of a run with all factors set to the high level (point 32) and three runs with exactly 1 out of factors B, D and E set to -1 and everything else set to +1.

4.5.4 Single stage design

An alternative to using an eight-run initial design plus a four-run follow-up experiment would be to choose a twelve run design before beginning experimentation. For

Table 4.14: Summary of posterior distribution of components of β after MD-optimal follow-up (4 10 11 26).

Term	P(In model)	Mean|in model	95% probability interval | in model	
I	1	65.045	59.405	70.291
A	0.011	-0.587	-9.623	7.208
B	**0.938**	**9.166**	**2.26**	**14.313**
C	0.197	0.454	-4.157	5.362
D	**0.873**	**5.253**	**0.143**	**10.018**
E	**0.646**	**-1.55**	**-4.921**	**2.794**
A B	0.008	1.917	-4.173	7.783
A C	0.002	1.626	-2.386	5.432
A D	0.006	1.474	-4.661	10.635
A E	0.004	0.299	-3.303	4.539
B C	0.193	-3.384	-8.397	1.195
B D	**0.865**	**5.678**	**0.964**	**10.24**
B E	**0.639**	**2.011**	**-1.716**	**5.457**
C D	0.189	1.353	-3.173	6.149
C E	0.172	2.307	-1.864	6.633
D E	**0.637**	**-4.76**	**-8.014**	**-0.077**
A B C	0.002	1.908	-1.901	5.147
A B D	0.006	-1.759	-15.168	8.408
A B E	0.003	0.013	-3.84	2.745
A C D	0.002	-1.197	-5.618	4.633
A C E	0.001	0.177	-4.182	3.72
A D E	0.003	1.213	-5.046	4.832
B C D	0.188	-1.683	-6.074	2.564
B C E	0.172	3.329	-1.228	8.255
B D E	**0.636**	**-0.552**	**-4.102**	**2.554**
C D E	0.172	4.974	0.316	9.426

comparison, 12 runs corresponding to a Plackett-Burman design were selected from the data. The posterior marginal factor probabilities are shown in Figure 4.7, for two values of γ. The analysis for the initial eight runs, from which the probabilities shown in figure 4.6 (a) are calculated, used $\gamma = 0.4$. Following Meyer *et al.* (1996), the analysis for the initial eight runs, combined with four follow-up runs was performed using $\gamma = 1.2$, from which the probabilities shown in Figure 4.6 (b)-(d) are calculated. From comparing the posterior factor probabilities in Figures 4.6 (b)-(d) and 4.7, the evidence on which factors are active is much more conclusive when a single stage 12-run experiment is used. This may be partly because of the blocking factor which was used for the follow-up runs, but does not have to be used in a single stage design. The 12 run Plackett-Burman

Table 4.15: Summary of posterior distribution of components of β after PMD optimal follow-up (2 4 10 12).

Term	P(In model)	Mean|in model	95% probability interval | in model	
I	1	64.298	56.519	71.341
A	0.355	2.734	-2.57	8.988
B	**0.934**	**7.381**	**0.889**	**13.216**
C	0.32	-0.116	-3.954	4.077
D	**0.899**	**4.869**	**-0.753**	**10.538**
E	0.487	-1.366	-4.711	2.877
A B	0.337	3.162	-1.753	8.218
A C	0.188	-0.513	-5.227	4.185
A D	0.337	4.805	-0.275	10.183
A E	0.171	0.234	-4.531	4.983
B C	0.316	-2.976	-7.626	1.825
B D	**0.875**	**4.574**	**-0.742**	**9.613**
B E	0.483	1.614	-2.721	5.514
C D	0.314	1.357	-3.21	5.916
C E	0.159	2.337	-2.36	7.042
D E	0.482	-3.895	-7.884	1.305
A B C	0.187	1.26	-3.414	5.892
A B D	0.325	0.846	-5.25	6.931
A B E	0.17	-2.733	-7.476	2.144
A C D	0.187	-2.717	-7.435	2.149
A C E	0.045	0.008	-5.801	5.887
A D E	0.17	1.283	-3.416	5.981
B C D	0.314	-1.303	-5.847	3.253
B C E	0.159	2.944	-1.711	7.554
B D E	0.481	-0.557	-4.701	3.46
C D E	0.159	4.442	-0.399	9.171

design has the advantage of having all main effects orthogonal to each other, which the combined eight-run initial experiment and four follow-up runs do not have. Therefore, in this example, if we plan to run 12 runs it is better to design a 12-run experiment than use an initial eight-run experiment and select four follow-up runs based on the results of the initial experiment. However, using an initial experiment gives us the option of stopping after fewer runs if the results are already conclusive, or dropping factors with low probability, as in the injection moulding example.

Table 4.16: Summary of posterior distribution of components of β after using 2nd best follow-up under the PMD criterion (25 26 27 28).

Term	P(In model)	Mean|in model	95% probability interval | in model	
I	1	63.903	56.994	69.466
A	0.063	3.092	-3.461	8.186
B	**0.982**	**9.245**	**3.543**	**14.504**
C	0.388	-0.48	-4.716	3.571
D	**0.641**	**5.342**	**0.456**	**8.995**
E	**0.713**	**-1.552**	**-4.767**	**2.543**
A B	0.062	4.334	-1.758	9.034
A C	0.016	0.167	-4.663	5.483
A D	0.018	2.692	-3.424	8.611
A E	0.039	-1.663	-6.323	2.802
B C	0.387	-5.504	-9.344	-0.4
B D	**0.639**	**5.286**	**0.865**	**9.26**
B E	**0.712**	**2.947**	**-0.56**	**6.277**
C D	0.167	1.533	-3.855	6.185
C E	0.305	4.257	-1.573	8.212
D E	0.493	-5.115	-8.642	0.827
A B C	0.016	2.681	-3.394	7.45
A B D	0.018	-0.586	-8.029	6.425
A B E	0.039	-4.672	-9.009	1.412
A C D	0.004	-0.676	-6.636	4.952
A C E	0.005	1.518	-5.553	8.802
A D E	0.014	-0.491	-5.668	4.698
B C D	0.167	-1.53	-6.35	2.893
B C E	0.305	4.286	-1.43	8.083
B D E	0.493	-0.136	-3.324	4.031
C D E	0.112	4.052	-1.009	9.052

4.6 Summary

In this chapter, we have shown how the PMD criterion can be applied to the problem of selecting follow-up runs after an initial experiment, using two examples from the literature and have compared its performance with those of the MD, HD and F criteria. In the first example we showed that the PMD criterion has the advantage over the MD of selecting follow-up runs with distinct points which can 'reverse' the aliasing scheme relative to the initial experiment. It was also noted that follow-up runs with certain patterns of aliasing, which induce effect cancelling, perform badly under the PMD criterion. We have shown that there is generally a strong positive correlation between

Table 4.17: Factor probabilities after all possible regular 8-run first stage experiments

Defining	P(Factor included in model)					
Relation	none	A	B	C	D	E
ABC=ADE=1	0.23	0.24	0.39	0.22	0.22	0.18
*ABD=ACE=1	0.23	0.27	0.37	0.17	0.29	0.17
ABE=ACD=1	0.26	0.18	0.37	0.19	0.26	0.2
ABC=BDE=1	0.19	0.14	0.37	0.14	0.41	0.36
ABD=BCE=1	0.22	0.3	0.37	0.18	0.27	0.18
ABE=BCD=1	0.2	0.17	0.41	0.32	0.29	0.17
ABC=CDE=1	0.23	0.22	0.41	0.23	0.23	0.17
ACD=BCE=1	0.25	0.19	0.41	0.18	0.24	0.2
ACE=BCD=1	0.2	0.16	0.45	0.28	0.29	0.15
ABD=CDE=1	0.23	0.28	0.31	0.23	0.28	0.18
ACD=BDE=1	0.2	0.17	0.3	0.17	0.41	0.33
ADE=BCD=1	0.22	0.21	0.36	0.27	0.26	0.18
ABE=CDE=1	0.25	0.19	0.35	0.2	0.28	0.19
ACE=BDE=1	0.18	0.13	0.34	0.14	0.47	0.32
ADE=BCE=1	0.24	0.2	0.38	0.2	0.24	0.2

*Design used by Meyer *et al.* (1996).

the ordering of the sets of follow-up runs using the PMD objective function and the orderings under each of the MD, HD and F criteria. We noted, however, that there are some important exceptions including sets of four follow-up runs where only three runs are distinct which performed well under the MD criterion and much less well under the PMD criterion.

In the second example we have investigated the effect of the initial design on the follow-up runs chosen. Initial designs with different aliasing schemes give slightly different posterior probabilities for the factors, so follow-up runs that concentrate on different factors are chosen. Comparison with a 12-run Plackett-Burman design shows that, in this example, if 12 experimental runs are available, it is better to choose a 12-run single stage design than to select four follow-up runs to resolve aliasing in an initial eight-run experiment.

Throughout this chapter we have used the set of models and prior probabilities from Meyer *et al.* to allow comparison with their results. Similarly, we have kept the assumption of effect forcing. This assumption, particularly for 3 factor interactions, seems unrealistic. The use of a resolution *IV* initial design, as in the first example, where main effects are confounded with three-factor interactions, suggests that

Table 4.18: Summary of designs chosen under the PMD criterion for all possible regular starting fractions. Key: b=factor is balanced in follow-up runs, -1= factor set to -1 level for all follow-up runs, +1 factor is set to +1 for all follow-up runs.

Starting fraction	Follow-up runs	Setting of factors				
Defining relation	chosen by PMD	A	B	C	D	E
ABC=ADE=1	9 10 11 28	b	b	-1	+1	
ABD=ACE=1	2 4 10 12	+1	b	-1	b	-1
ABE=ACD=1	9 11 25 27	-1	b	-1	+1	b
ABC=BDE=1	16 24 30 32	+1		+1		
ABD=BCE=1	25 26 27 28	b	b	-1	+1	+1
ABE=BCD=1	9 13 27 31	-1	b	b	+1	b
ABC=CDE=1	9 11 13 15	-1	b	b	+1	-1
ACD=BCE=1	9 11 25 27	-1	b	-1	+1	b
ACE=BCD=1	11 13 27 19	-1	b	b	+1	b
ABD=CDE=1	29 30 31 32	b	b	+1	+1	+1
ACD=BDE=1	3 15 23 31	-1	+1		b	b
ADE=BCD=1	26 28 30 32	+1	b	b	+1	+1
ABE=CDE=1	9 11 13 15	-1	b	b	+1	-1
ACE=BDE=1	16 24 30 32	+1		+1		
ADE=BCE=1	9 11 25 27	-1	b	-1	+1	b

three-factor interactions are viewed as unlikely *a priori*. It would be interesting to explore the use of a less restrictive model space in relation to these and other examples.

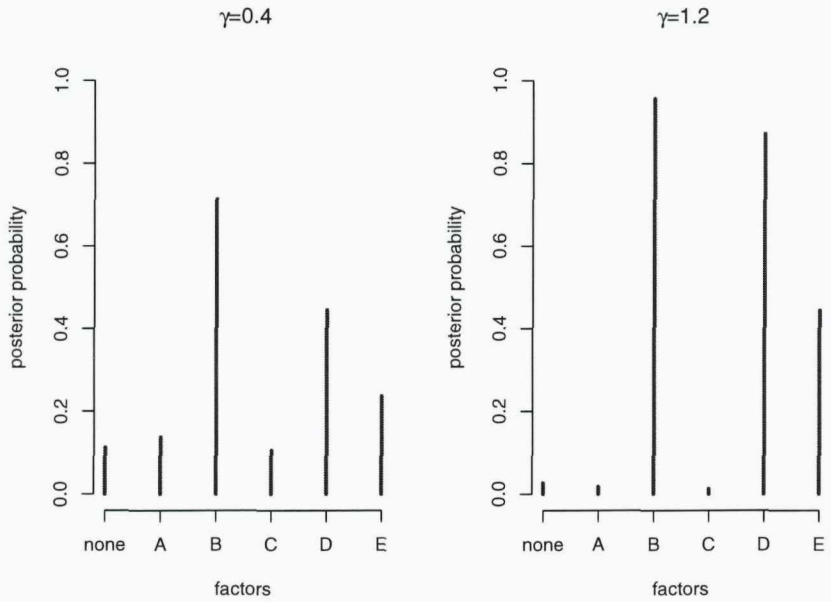

Figure 4.7: Posterior factor probabilities from a 12-run Plackett-Burman design for the second example, with two values of γ. Runs used are 1,3,6,12,14,15,18,23,24,25,28,29.

Chapter 5

Application to a Tribology Experiment

5.1 Introduction

In this chapter, a small programme of experimentation is described, which was carried out by the Tribology group in the School of Engineering Sciences at the University of Southampton. This example demonstrates how the PMD criterion may be used to select a small number of follow-up runs after an initial experiment, and also shows the type of analysis that may be carried out on the data obtained from the experiment. The researchers were interested in simulating the effect of contaminated oil on the wear in a pin and disc assembly and finding out which contaminants had an important effect on wear. The aim of the particular experiment described here was to screen for important process effects.

The experimental setup consisted of a pin moving relative to a disc, with oil between them; see Figure 5.1. There were six factors that could be set in the experiment: Disc material (steel or silicon), Pin material (steel or silicon), Soot (% by weight), Oxidation (hours), Concentration of H_2SO_4 (mM) and Moisture (%ml). Four response variables were measured: Charge (pC), Coefficient of Friction, Temperature (°C) and Wear Scar Radius (mm). The observations were time-consuming to obtain and hence only about 20 runs could be made in the experiment. A linear model with main effects and possibly

Table 5.1: Coding scheme for explanatory variables.

Variable	Label	-1 level	0 level	+1 level
Disc Material	A	Steel	-	Silicon
Pin Material	B	Steel	-	Silicon
Soot (% wt)	C	0	5	10
Oxidation (hours)	D	0	5	10
H_2SO_4 (mM)	E	0	1.25	2.5
Moisture (% ml)	F	0	1.25	2.5

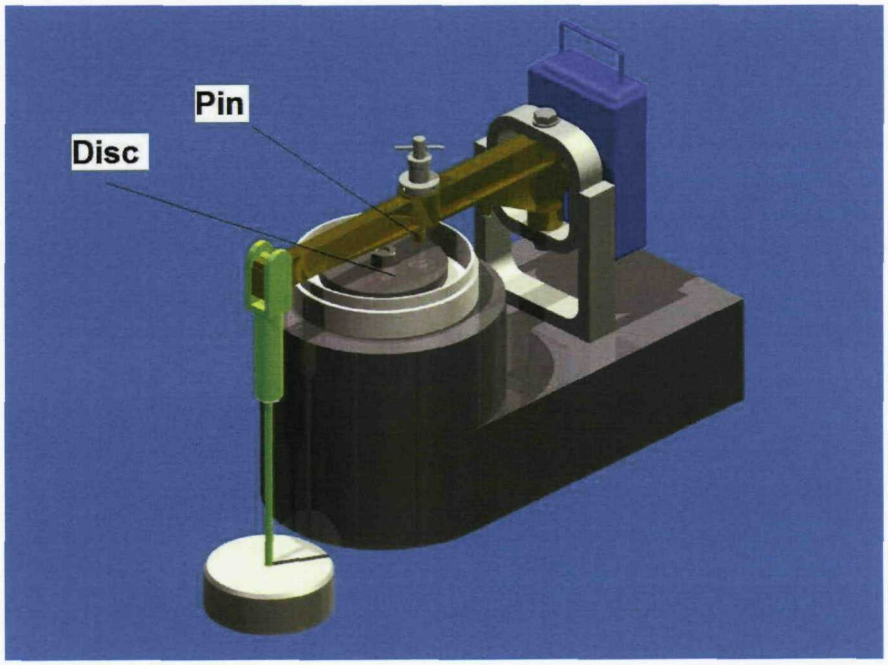

Figure 5.1: Experimental rig for tribology example

some two-factor interactions was believed likely to produce a reasonable approximation to the response.

5.2 Initial Experiment

The design chosen for the initial experiment was a 2^{6-2}_{IV} fractional factorial with defining relation $ABDE = ACEF = BCDF = I$ combined with four 'centre points' with the quantitative factors (C, D, E, F) set to the 0 level for the four possible combinations of the two qualitative factors (A, B). These four points were to be used for model checking. The factors were labelled and coded as shown in Table 5.1. Each of the four response

variables was measured at each run of the experiment.

5.3 Models and Prior Distributions

For each response variable, the model space adopted included all linear models formed from main effect and 2-factor interaction terms in the six factors, subject to strong heredity (see Chipman 1996). That is, an interaction term may not be present unless both main effects of the factors involved in the interaction are also in the model. We make this restriction so that the model that we select is easily interpreted. Nelder (1998) showed that an analysis using models which obey strong heredity has the advantage of not being affected by re-coding of the factors.

The model probabilities are built up from the individual effect probabilities, as described in Section 4.1. Each main effect term is included in the model independently with the same probability p_m. Each interaction term is included in the model with probability p_i, provided both corresponding main effects are present; and with probability 0 otherwise. The prior distribution on σ^2 is an improper non-informative prior $f(\sigma^2) \propto \sigma^{-2}$, equivalent to an inverse-gamma distribution with $a = d = 0$. Conditional on σ^2, the regression parameters $\boldsymbol{\beta}$ are distributed as $N(0, \mathbf{I}\sigma^2)$.

5.4 Analysis of the Initial Experiment

In this section, the analysis of the initial experiment is described, starting with a preliminary investigation based on simple frequentist methods. A full Bayesian analysis is then given. The design and data from the experiment are given in Table 5.2.

Run 15 was an outlier with respect to the response variable charge, with value 45.5 pC. A re-run was performed at this combination of factor levels under the same laboratory conditions and the new response value used instead and is recorded in Table 5.2.

5.4.1 Preliminary Analysis

A half-normal effects plot was used to identify unusually large effects or combinations of effects (Daniel 1959). A brief description of findings for each variable is now given. The

Table 5.2: Initial runs of tribology experiment

Run	A	B	C	D	E	F	Charge (pC)	Coefficient of Friction	Temp. (°C)	Wear Scar (mm)	Temp. Difference (°C)
1	-1	1	-1	1	-1	-1	0.03	0.11	40.32	0.11	22.32
2	-1	1	1	1	-1	1	0.02	0.09	54.96	0.26	33.96
3	-1	-1	0	0	0	0	0.02	0.17	66.34	0.48	41.34
4	-1	-1	-1	1	1	1	0.02	0.08	38.27	0.14	18.27
5	-1	-1	1	-1	-1	1	0.02	0.08	64.48	0.42	41.98
6	-1	-1	-1	-1	-1	-1	0.02	0.11	36.70	0.14	18.10
7	-1	1	-1	-1	1	1	0.02	0.09	41.07	0.11	20.07
8	-1	1	0	0	0	0	0.04	0.11	59.47	0.24	39.37
9	-1	-1	1	1	1	-1	0.02	0.10	50.06	0.45	30.56
10	-1	1	1	-1	1	-1	0.03	0.08	58.95	0.25	36.85
11	1	1	-1	1	1	-1	5.39	0.07	54.00	0.12	28.00
12	1	-1	-1	-1	1	-1	14.41	0.07	55.88	0.26	32.88
13	1	-1	0	0	0	0	26.75	0.08	56.00	0.26	32.00
14	1	-1	1	-1	1	1	10.68	0.07	63.93	0.32	33.93
15	1	1	1	-1	-1	-1	13.16	0.10	60.86	0.31	29.86
16	1	1	1	1	1	1	3.57	0.09	65.63	0.23	40.63
17	1	1	-1	-1	-1	1	6.75	0.10	56.20	0.18	32.20
18	1	1	0	0	0	0	28.6	0.10	58.89	0.22	32.89
19	1	-1	1	1	-1	-1	6.05	0.05	57.02	0.40	32.02
20	1	-1	-1	1	-1	1	6.82	0.06	51.36	0.21	28.36

half-normal plots for the charge and coefficient of friction responses are given in Appendix B.

Wear Scar Radius: The main effects of B and C (pin material and soot) stand out for this response. The $AC = EF$ interaction term is the next largest, but is close to the line of effects assumed to be non-active (see Figure 5.2).

Temperature: The largest effects on the temperature response appear to be those of A and C (disc material and soot) and the two-factor interaction between them (see Figure 5.3).

Charge: The preliminary analysis indicated that a log transformation would produce a response variable that would better fit the modelling assumptions (normal distribution with constant error variance). The normal effects plots of the transformed charge variable is shown in Figure B.1. This plot shows that the main effect of A (disc material) is much larger than all other effects. Figure B.2 shows a normal effects plot with A removed to give a clearer assessment of the next two largest effects, D (oxidation) and

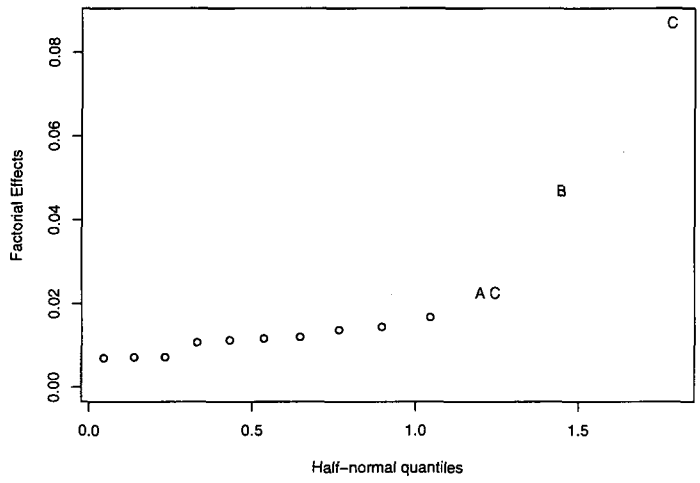

Figure 5.2: Half Normal Plot for Wear Scar.

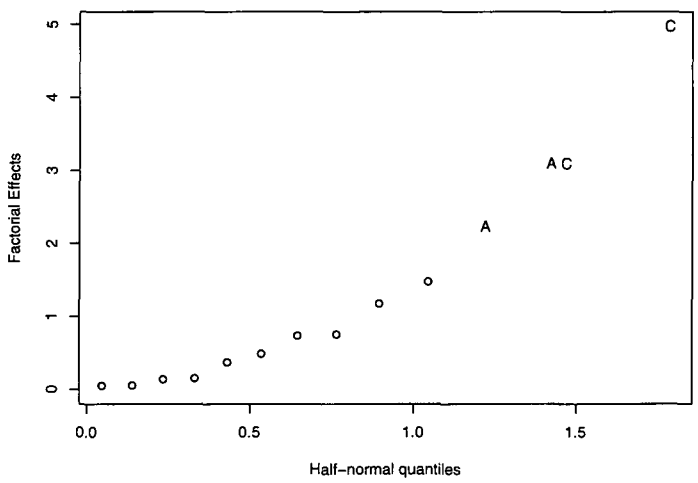

Figure 5.3: Half Normal Plot for Temperature.

the aliased interactions $AD = BE$ relative to the sizes of the other factorial effects.

Coefficient of Friction: Run 3, one of the 'centre points', is an outlier with respect to coefficient of friction. It was decided to set aside the centre points for model checking, and perform an analysis based on the results from the fractional factorial. The half-normal plot for the coefficient of friction response does not closely follow a straight line. However, residual diagnostics and the Box-Cox stabilisation (Box and Cox 1964) do not suggest that transformation is required. A few effects stand out as being larger than the rest, which are the main effects of A and B (disc and pin materials), the interaction between them ($AB = DE$), and the $AD = BE$ interaction.

Based on this preliminary analysis, the log transform was retained for charge. The preliminary analysis provided a useful understanding of the data for the more detailed modelling.

A useful preliminary way to explore any relationships that may exist between the four responses given in Table 5.2 is to plot them against each other, as shown in Figure 5.4. There weak is evidence of a correlation between the wear scar and temperature, and, to a lesser extent, between log(charge) and temperature. A possible analysis for data with multiple response variables is multivariate analysis of variance (MANOVA). However, this technique is most effective with a reasonable degree of correlation between the response variables, which is not evident for all pairs of variables in this data and so is thought unnecessary.

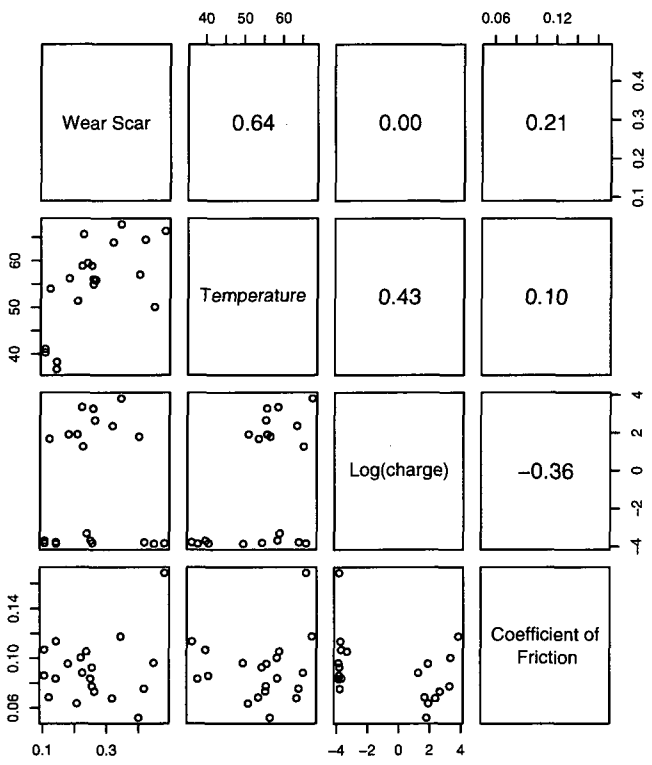

Figure 5.4: Matrix plot and correlations of the four responses given in Table 5.2

5.4.2 Bayesian Analysis

The data from the initial experiment that were analysed consist of a set of 19 runs (with the outlying response obtained from a centre point excluded) for the coefficient of friction, and the full set of 20 observations for each of the three remaining responses. The prior distributions outlined in Section 5.3 are used, with $p_m = p_i = 0.5$. For each response the posterior probabilities of all models in the set were calculated were calculated using R (R Development Core Team, 2008). The results are summarised for each response in the following tables in two ways. First, the most probable models are listed. Second, the marginal probabilities for each factorial effect are given, formed by summing the probabilities of all models containing that effect. These have been used also by Meyer and Wilkinson (1998), who call them model-averaged effect probabilities. An effect probability formed by summing over the top 10 models by posterior probability is also given (re-normalised so that a term that is present in all the top 10 models is given marginal probability 1). This is given following the reasoning (although not the method) of Madigan and Raftery (1994), who state that models with a low posterior probability have effectively been 'discredited' and that averaging over a smaller set of models more accurately reflects our model uncertainty. The results for the wear scar response are given in Tables 5.4 and 5.3, for the results under the other three responses, see Appendix B.

Table 5.3: Top 10 most probable models for Wear Scar at First Stage

Probability	Model Terms			
0.349	I	B	C	
0.158	I	C		
0.114	I	B	C	BC
0.026	I	B	C	F
0.024	I	B	C	E
0.022	I	B	C	D
0.022	I	C	F	
0.021	I	C	E	
0.02	I	C	D	
0.019	I	A	B	C

Table 5.4: Model-averaged Effect Probabilities for Wear Scar for the initial experiment.

Factorial Term	Probability (All models)	Probability (In top 10 models)
I	1.00	1.000
A	0.09	0.025
B	0.71	0.715
C	0.97	1.000
D	0.10	0.054
E	0.10	0.058
F	0.11	0.062
AB	0.01	0.000
AC	0.03	0.000
AD	0.00	0.000
AE	0.00	0.000
AF	0.00	0.000
BC	0.17	0.147
BD	0.01	0.000
BE	0.01	0.000
BF	0.01	0.000
CD	0.02	0.000
CE	0.02	0.000
CF	0.03	0.000
DE	0.00	0.000
DF	0.00	0.000
EF	0.00	0.000

5.4.3 Summary of Results after the initial experiment

The analysis of half-normal plots from the initial experiment provides evidence that several of the factors are active and have an effect on some of the responses. Charge is much higher when the disc material (A) is silicon. There is also some evidence to suggest that oxidation (D) and the $AD = BE$ interaction have some effect. When coefficient of friction is the response, the main effects A and B and the $AD = BE$ and $AB = DE$ interactions stand out in the half-normal plot (figure B.3). For temperature, the largest effects are A, C and the $AC = EF$ interaction. For wear scar, the effects of B, C and the $BC = DF$ interaction are the largest.

A Bayesian analysis shows that, for the log(charge) response, disc material is almost definitely having an effect. The next most likely model term is the main effect of Oxidation (D). For coefficient of friction, the model with highest posterior probability contains the intercept term only, however, some other model terms retain a reasonable posterior probability, such as B and A, with probabilities 0.3 and 0.22 respectively. The main effect terms of B and C have high posterior probability for the wear scar response. There is also some evidence to suggest that the BC interaction is active, as it appears in the third most likely model and has the highest posterior probability of any interaction term under any of the responses in this analysis. For temperature, C has a high posterior probability (0.48). The second most likely model term is the main effect of A. The marginal probabilities for the interaction terms are very small - with the exception of the BC interaction under the wear scar response, all are less than 0.03. However, the strong heredity prior that was used will have had some effect on these results. Some interactions that were seen to be large in the half-normal plots do not have high posterior probability because the parent main effects are small hence the models that contain them have low posterior probability. A follow-up experiment was planned to provide further information about the importance of effects for which the evidence is so far inconclusive. Such effects include the BC interaction for the wear scar response and the main effects of C and A for temperature. For charge, the main effect of A is almost certainly active, but a follow-up experiment will help determine whether any other factors have an effect. For the coefficient of friction response the follow-up runs will help confirm or negate the weak evidence for the main effect of B from the initial experiment.

5.5 Selection of Follow-up Runs

It was possible to perform a further six experimental runs in order to provide more information on which main effects and interactions affect each response, and to help distinguish between effects aliased in the initial runs. For each of the four responses, the posterior distribution obtained in the analysis of Section 5.4.2 was used as a prior for evaluating follow-up designs under the PMD criterion. Twenty searches were performed for each response, using the Modified Fedorov Exchange Algorithm described in Chapter 2. In addition, 80 designs were generated randomly, for comparison. All 160 designs were evaluated under the PMD criterion for all four responses. It was decided to base selection of a design on its performance with respect to the wear scar and temperature response only, as these produced the greatest variation in PMD values. Hence there was less scope to choose a bad design for the other two responses. Because of the difference in the ranges of values of the objective function for the different responses, a ranking of the designs under each response was employed rather than values of the objective function. This ranking was used to identify a minimax design, i.e. a design that has the minimum value of

$$\max\{\text{Rank for wear scar, Rank for temperature}\}$$

over the set of designs considered.

The ranks of the designs with respect to the two responses are plotted in Figure 5.5. The colours (black, blue, red, green) correspond to the responses used in the search - log(charge), coefficient of friction, wear scar and temperature respectively. Designs plotted in magenta were generated randomly. This plot shows the search algorithm to be working fairly effectively - the designs found for a particular response are generally ranked higher under that response than those found for different responses or obtained at random, apart from a few instances of designs where the algorithm became trapped in local optima.

The minimax design has ranks of 18 and 19 for the responses wear scar and temperature respectively. This is the design plotted in red on the lower left-hand corner of Figure 5.5, and is the design that we would ideally use. However, the researchers had only three

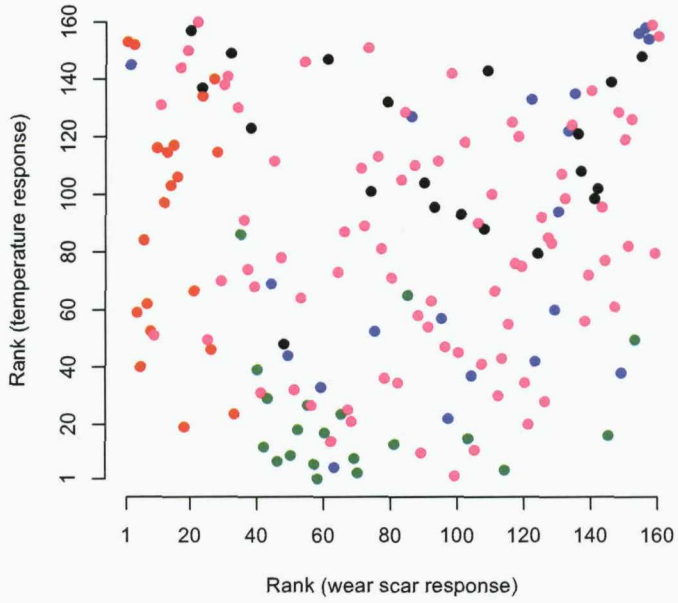

Figure 5.5: Ranks of Follow-up Designs for Wear Scar and Temperature Responses. Found for log (charge)• coefficient of friction• wear scar• temperature• random•.

steel discs left for experimentation, and this design requires the use of steel discs for four runs. In Figure 5.6 we have plotted the rankings of the designs again, but designs using three or fewer steel discs are indicated by squares. The minimax design (of those that are permissible) is ranked 41 and 31 with respect to wear scar radius and temperature respectively, and is one of the designs generated randomly. There is another possible design, with rankings 42 and 12 that was found using temperature as a response. Although not strictly the minimax design, we use this design because of its better performance with temperature as a response. The experiment was run using this design, and data collected for the four responses. The design and the responses obtained from it are given in Table 5.5. It may be noted that this design maintains balance and orthogonality for factors A and C, the two most important factors at the first stage when using temperature as a response. The factor B is not balanced, but is orthogonal to C, which is of use when wear scar is the response and models containing B, C and BC are among those of high posterior probability after the first stage.

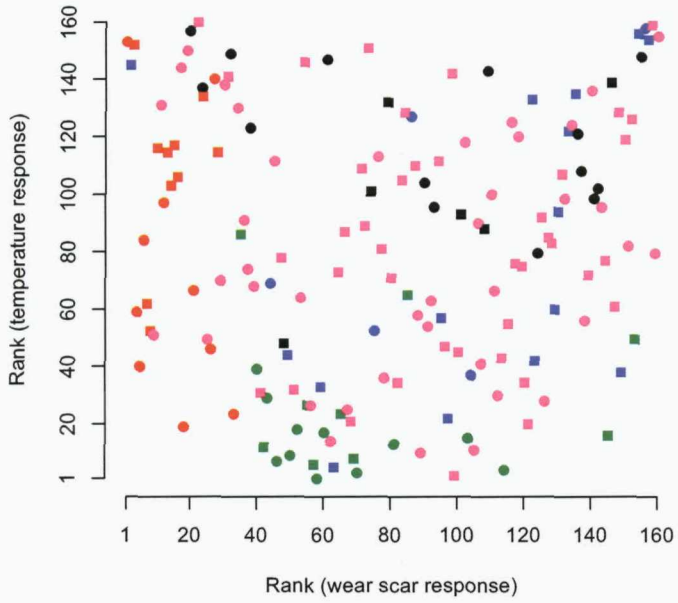

Figure 5.6: Ranks of Follow-up Designs for Wear Scar and Temperature Responses. Found for log (charge)• coefficient of friction• wear scar• temperature• random•. Squares denote designs requiring 3 or fewer steel discs.

Data on temperature difference (temperature of apparatus - ambient temperature) became available for all experimental runs after the second stage, so this was used in preference to the temperature variable previously used.

Table 5.5: Follow-up runs for Tribology Experiment

Run	A	B	C	D	E	F	Charge (pC)	Coefficient of Friction	Temp. (°C)	Wear Scar (mm)	Temp. Difference (°C)
1	1	1	-1	-1	1	-1	6.65	0.08	37.33	0.12	17.33
2	-1	-1	1	1	1	-1	0.02	0.09	52.93	0.42	34.93
3	-1	1	1	-1	-1	1	0.02	0.10	55.21	0.21	37.21
4	1	-1	-1	-1	-1	1	5.15	0.06	38.17	0.26	20.17
5	1	-1	1	1	1	-1	13.94	0.07	47.70	0.34	29.70
6	-1	-1	-1	1	-1	1	0.02	0.16	46.53	0.23	28.53

5.6 Analysis for Second Stage experiment

We perform a Bayesian analysis at the second stage, again calculating the model averaged probabilities of each term, and finding the top 10 most probable models. For each model term we also calculate the maximum value of c which would see it included in the model selected using the PMD loss function (2.3). This value of c is simply $\frac{p}{1-p}$, where p is the posterior model-averaged probability of the term. In this example, the models selected using $c = 1$ turn out to be the models with highest posterior probability. Another analysis performed is to show the likely size and direction of effects using histograms of the model-averaged posterior distribution of the parameters, similar to those produced by Meyer and Wilkinson (1998). To create these histograms, we sample 10 000 models, with replacement, from the posterior model distribution. For each model, we calculate $a^*, d^*, \boldsymbol{\mu}^*$ and $\mathbf{V}^*$. From these, we take one sample of σ^2 from a $a^*\chi^{-2}_{d^*}$ distribution and $\boldsymbol{\beta}$ from a $N(\boldsymbol{\mu}^*, \sigma^2\mathbf{V}^*)$ distribution. For each regression term, we can plot a histogram of the values that the associated elements of $\boldsymbol{\beta}$ take when we sample from models containing that parameter. These histograms must be interpreted with reference to the model-averaged effect probabilities in Tables 5.6, 5.8, 5.10 and 5.12 as for each term there are models that do not contain it, where its value is effectively zero. These models have not been included on the histograms because the relative height of the 'spike' at zero is dependent on the width of the bins used in constructing the histogram. However, the existence of these models, and their combined probability, should be borne in mind when considering the posterior parameter distributions.

Table 5.6: Model-Averaged Effect Probabilities for Wear Scar at Second Stage.

Factorial Term	Probability (All models)	Probability (In top 10 models)	In PMD model chosen when $c <$
I	1.00	1.000	NA
A	0.10	0.058	0.115
B	0.95	0.967	19.211
C	1.00	1.000	497.127
D	0.08	0.031	0.082
E	0.08	0.046	0.083
F	0.09	0.051	0.102
AB	0.02	0.000	0.020
AC	0.04	0.023	0.043
AD	0.00	0.000	0.002
AE	0.00	0.000	0.002
AF	0.00	0.000	0.001
BC	0.24	0.213	0.322
BD	0.01	0.000	0.014
BE	0.01	0.000	0.013
BF	0.02	0.000	0.016
CD	0.02	0.000	0.017
CE	0.01	0.000	0.014
CF	0.03	0.013	0.026
DE	0.00	0.000	0.001
DF	0.00	0.000	0.001
EF	0.00	0.000	0.003

Table 5.7: Top 10 most probable models for Wear Scar at Second Stage

Probability	Model Terms				
0.487	I	B	C		
0.168	I	B	C	BC	
0.032	I	B	C	F	
0.029	I	A	B	C	
0.028	I	C			
0.027	I	B	C	E	
0.026	I	B	C	D	
0.02	I	A	B	C	AC
0.011	I	B	C	F	CF
0.011	I	B	C	E	BC

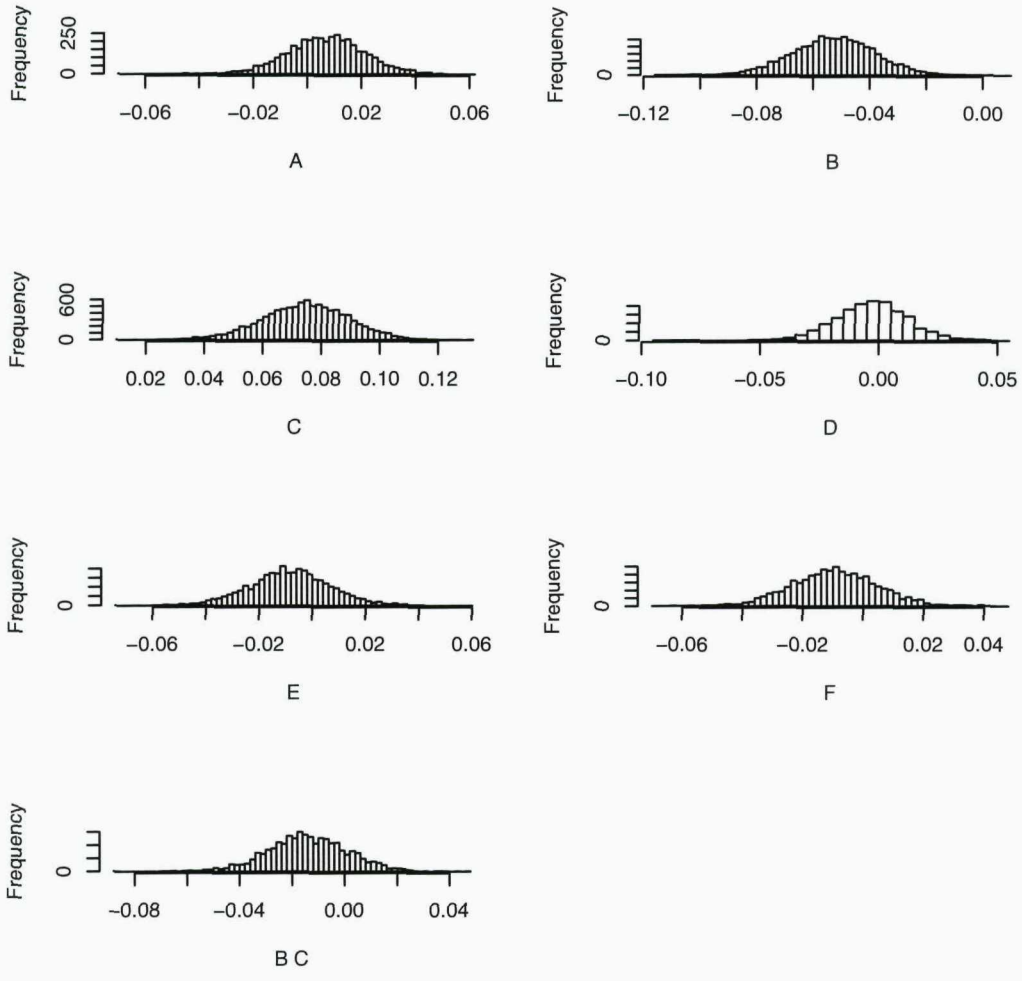

Figure 5.7: Model-Averaged Posterior Parameter Distributions for Wear Scar at Second Stage

Table 5.8: Model-averaged Effect Probabilities for Temperature Difference at Second Stage.

Factorial Term	Probability (All models)	Probability (In top 10 models)	In PMD model chosen when $c <$
I	1.00	1.000	NA
A	0.12	0.088	0.136
B	0.10	0.059	0.113
C	0.89	0.944	7.846
D	0.10	0.072	0.114
E	0.11	0.061	0.120
F	0.14	0.107	0.167
AB	0.00	0.000	0.001
AC	0.03	0.027	0.033
AD	0.00	0.000	0.002
AE	0.00	0.000	0.002
AF	0.00	0.000	0.002
BC	0.02	0.000	0.016
BD	0.00	0.000	0.001
BE	0.00	0.000	0.001
BF	0.00	0.000	0.002
CD	0.02	0.014	0.017
CE	0.02	0.000	0.016
CF	0.02	0.021	0.026
DE	0.00	0.000	0.001
DF	0.00	0.000	0.002
EF	0.00	0.000	0.005

Table 5.9: Top 10 most probable models for Temperature Difference at Second Stage

Probability	Model Terms			
0.469	I	C		
0.072	I	C	F	
0.052	I	A	C	
0.051	I	C	E	
0.05	I	B	C	
0.049	I	C	D	
0.047	I			
0.022	I	A	C	AC
0.018	I	C	F	CF
0.012	I	C	D	CD

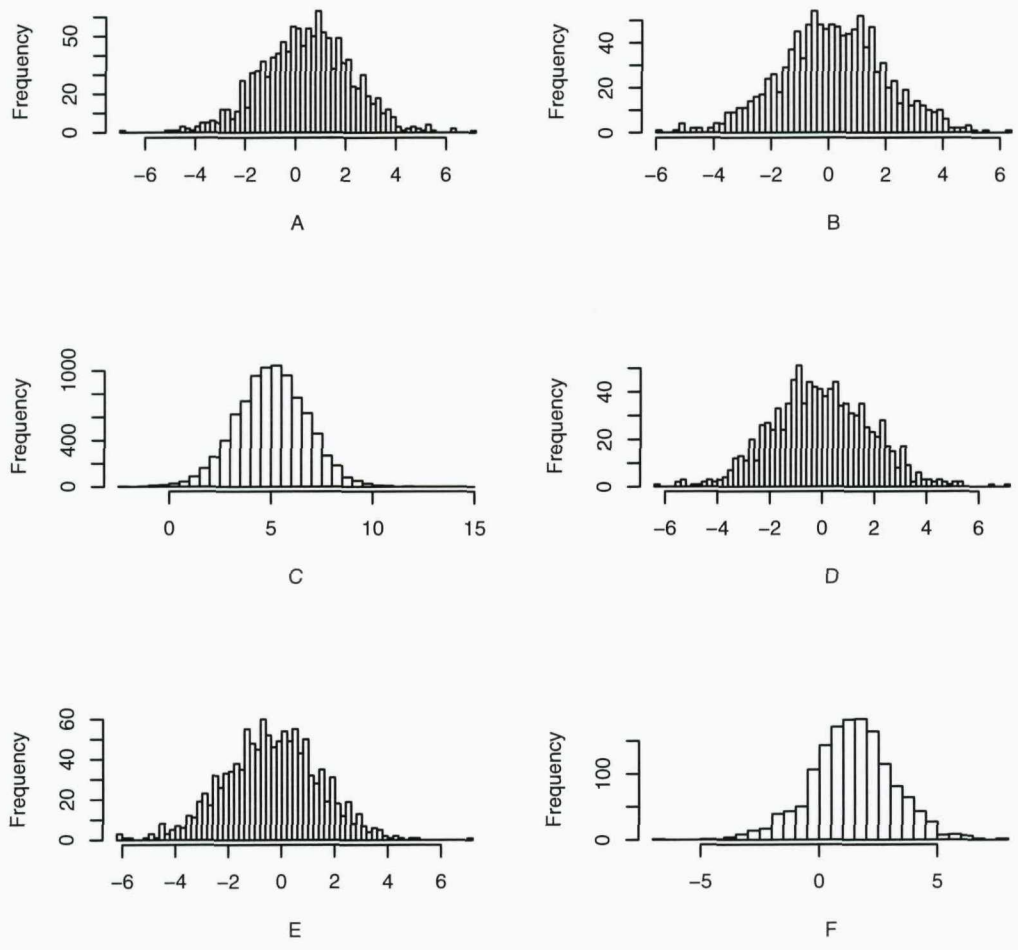

Figure 5.8: Model-averaged Posterior Parameter Distributions for Temperature Difference at Second Stage

Table 5.10: Effect Probabilities for log(charge) at Second Stage.

Factorial Term	Probability (All models)	Probability (In top 10 models)	In PMD model chosen when $c <$
I	1.00	1.000	NA
A	1.00	1.000	6×10^{12}
B	0.12	0.091	0.135
C	0.11	0.082	0.119
D	0.10	0.078	0.114
E	0.10	0.059	0.105
F	0.16	0.123	0.184
AB	0.03	0.021	0.029
AC	0.02	0.016	0.021
AD	0.02	0.014	0.018
AE	0.02	0.000	0.017
AF	0.03	0.027	0.035
BC	0.00	0.000	0.002
BD	0.00	0.000	0.002
BE	0.00	0.000	0.002
BF	0.00	0.000	0.002
CD	0.00	0.000	0.001
CE	0.00	0.000	0.001
CF	0.00	0.000	0.002
DE	0.00	0.000	0.001
DF	0.00	0.000	0.002
EF	0.00	0.000	0.002

Table 5.11: Top 10 most probable models for log(charge) at Second Stage

Probability	Model Terms			
0.512	I	A		
0.087	I	A	F	
0.063	I	A	B	
0.06	I	A	C	
0.058	I	A	D	
0.054	I	A	E	
0.024	I	A	F	AF
0.019	I	A	B	AB
0.015	I	A	C	AC
0.012	I	A	D	AD

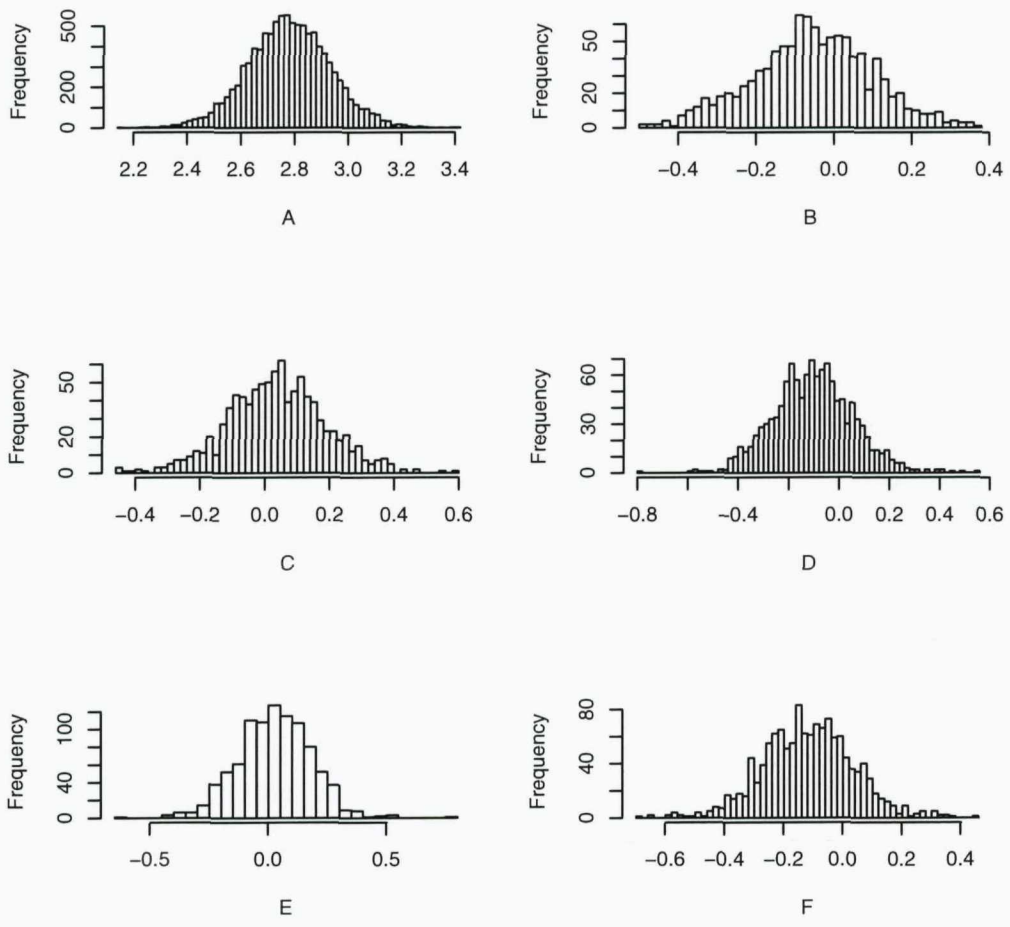

Figure 5.9: Marginal Posterior Parameter Distributions for log(charge) at Second Stage

Table 5.12: Model-Averaged Effect Probabilities for Coefficient of Friction at Second Stage.

Factorial Term	Probability (All models)	Probability (In top 10 models)	In PMD model chosen when $c <$
I	1.00	1.000	NA
A	0.61	0.560	1.565
B	0.14	0.091	0.163
C	0.15	0.111	0.177
D	0.12	0.043	0.141
E	0.16	0.121	0.195
F	0.11	0.043	0.127
AB	0.03	0.000	0.033
AC	0.02	0.000	0.020
AD	0.02	0.000	0.021
AE	0.02	0.000	0.023
AF	0.01	0.000	0.010
BC	0.00	0.000	0.004
BD	0.00	0.000	0.003
BE	0.00	0.000	0.003
BF	0.00	0.000	0.002
CD	0.00	0.000	0.002
CE	0.00	0.000	0.005
CF	0.00	0.000	0.002
DE	0.00	0.000	0.003
DF	0.00	0.000	0.003
EF	0.00	0.000	0.003

Table 5.13: Top 10 most probable models for Coefficient of Friction at Second Stage

Probability	Model Terms		
0.295	I	A	
0.147	I		
0.047	I	E	
0.046	I	A	C
0.043	I	A	E
0.037	I	C	
0.034	I	A	B
0.033	I	B	
0.032	I	F	
0.032	I	D	

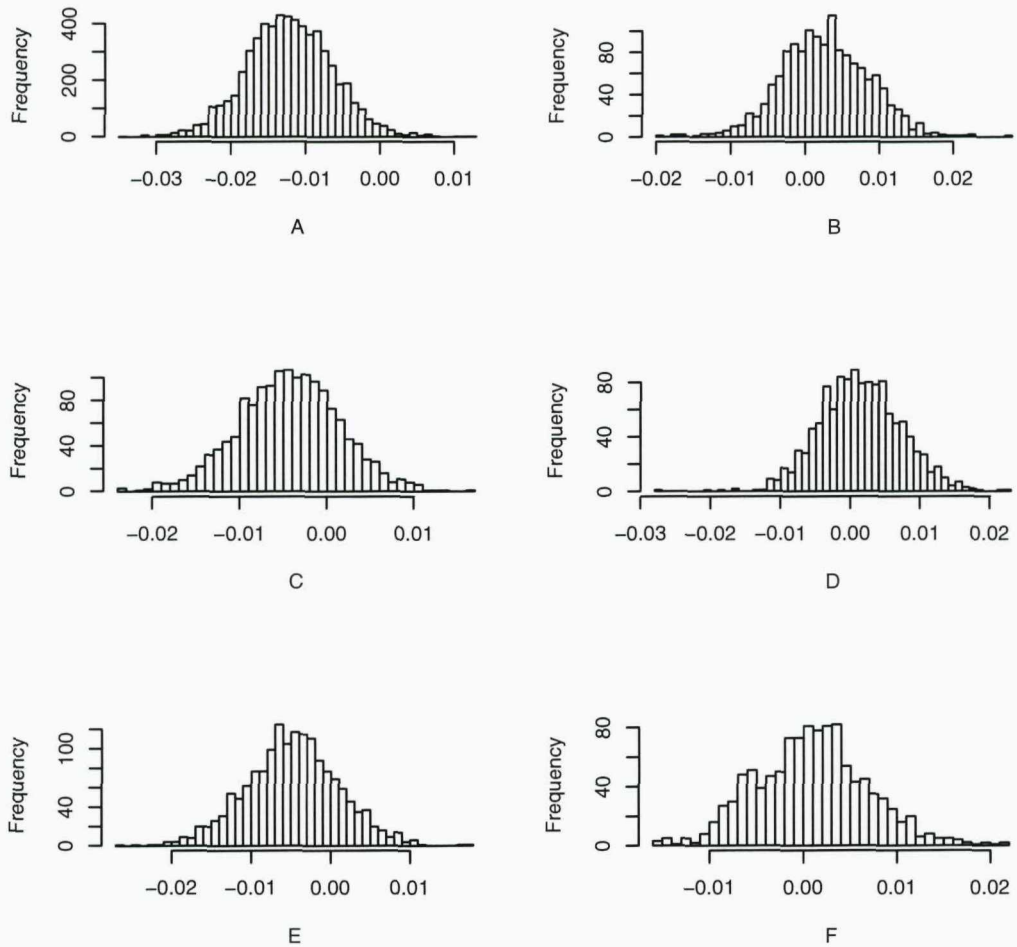

Figure 5.10: Model-Averaged Posterior Parameter Distributions for Coefficient of Friction at Second Stage

5.6.1 Summary of Results

For the wear scar response, the main effects of pin material (B) and soot (C) are very likely to be active. The BC interaction is also possibly active, with posterior probability 0.24, and is present in the second most likely model. The posterior means, for B, C and BC, averaged over the models containing them, are -0.052, 0.075 and -0.014 respectively. Both B and C were viewed a likely to be active after the initial experiment (with probabilities 0.71 and 0.97 respectively). The follow-up runs have provided more

evidence to confirm this. With temperature difference as the response, the main effect of C is active with high probability, and has an estimated mean of 5.04. No other model terms have a posterior probability greater than 0.14. After the initial experiment, for the temperature response, A was viewed as a potentially important factor, with a probability of 0.22, and large estimated effects for A and AC (see Figure 5.3). However after the follow-up experiment, which maintained the orthogonality of A and C, the probability that A is active has dropped to 0.12. With log(charge) as the response, the main effect of disc material (A) dominates, with posterior probability effectively equal to 1. This was also the case after the initial experiment. The posterior model-averaged mean for the effect of A is about 2.8, equating to charge being about 270 times greater for silicon discs. For coefficient of friction, factor A has a posterior probability of around 0.61 of being active. The expected effect of disc material, given that it is active, is around -0.01. After the initial experiment, the factor with highest marginal probability was B, with a probability of 0.30; the probability of A was 0.22. The extra data provided by the follow-up runs has reduced the probability of B to 0.14.

5.7 Conclusions

In this chapter, the PMD criterion was applied to the selection of follow-up runs for a real experiment, which was then performed. The data from the experiment was analysed by Bayesian methods consistent with the objectives of the experiment of model selection and effect screening. The multiple responses were handled by fitting separate models and performing separate design searches for each response, then selecting a minimax design using two of the responses. For the data from this experiment, it was felt that modelling a multivariate response would not be appropriate, however, the use of multivariate techniques would be a possible extension to this work.

The Bayesian final analysis of this data does not give very high posterior probability to many interaction terms. This may simply be because none were active, or that a larger experiment is required to give sufficient power to detect them. However, the prior distributions used may have had some effect. Changing the prior distributions, for example by not enforcing strong heredity, changing the value of p_i, or the prior variance

of the associated model parameters might lead to more models containing interactions having high posterior probability.

Chapter 6

Computational Approaches to Large Model Spaces

6.1 Introduction

As we have seen in Chapter 3, some situations, particularly in screening, result in the need to consider a large number of possible models. Evaluation of the PMD criterion requires posterior probabilities of all these models to be calculated at each run of simulation. If the model space is sufficiently large, evaluating the probabilities for all models is computationally infeasible. The specific model space that we will consider here is one with f possible factors, where the main effect of each factor is included independently with probability p_{main}, and, conditional on the presence of main effects i and j, the ij interaction is included with probability p_{int}. We will generally use $p_{\text{main}} = p_{\text{int}} = 0.5$ to represent a lack of prior information.

Therefore, there are

$$\sum_{f_m=0}^{f} \binom{f}{f_m} \times 2^{\binom{f_m}{2}}$$

possible models, where each term in the summation is the number of models that involve f_m factors. This rises rapidly with f, as demonstrated in Table 6.1. In this chapter we will discuss three methods of approximating the objective function for large model spaces. These are:

Table 6.1: Size of strong heredity model space for 3-9 factors.

No. of factors	No. of models
3	18
4	113
5	1450
6	40069
7	$\approx 2.35 \times 10^6$
8	$\approx 2.86 \times 10^8$
9	$\approx 7.12 \times 10^{10}$

(i) using a subset of the models with highest prior probability

(ii) a technique called Occam's window, and

(iii) using an MCMC scheme to explore the model space.

6.2 Using only the Models with Highest Prior Probability

One simple approach is, for the purposes of design, to use an approximation to the prior and set the prior probabilities of some models to 0. This is done by replacing the model space with the $M_0 < M$ models with highest prior probability, with the prior model probabilities re-normalised to sum to 1. This approach was used by Chipman (1996) to evaluate the HD objective function, and was used in Chapters 3 and 5 of this thesis. There are some drawbacks to this method. If the combined prior probability of the M_0 models used is not very close to 1, the approximation to the objective function may be poor. Figures 6.1 and 6.2 (using a semi-log scale) show that, for six factors, a relatively small set of models do account for most of the prior probability. In our model space, when $p_{\text{main}} = p_{\text{int}} = 0.5$, all models containing the same number of main effect terms will have the same prior probability. Therefore, there is not a unique set of M_0 models with highest prior probability unless M_0 corresponds to a cut-off point between different sizes of models. These cut-off points correspond to the 'kinks' in Figure 6.2. Using a value of M_0 that is not a cut-off between models of different sizes will mean that the main effect and interaction terms involving some of the factors will be given higher prior probability than others. This may lead to designs being selected that are good at estimating and

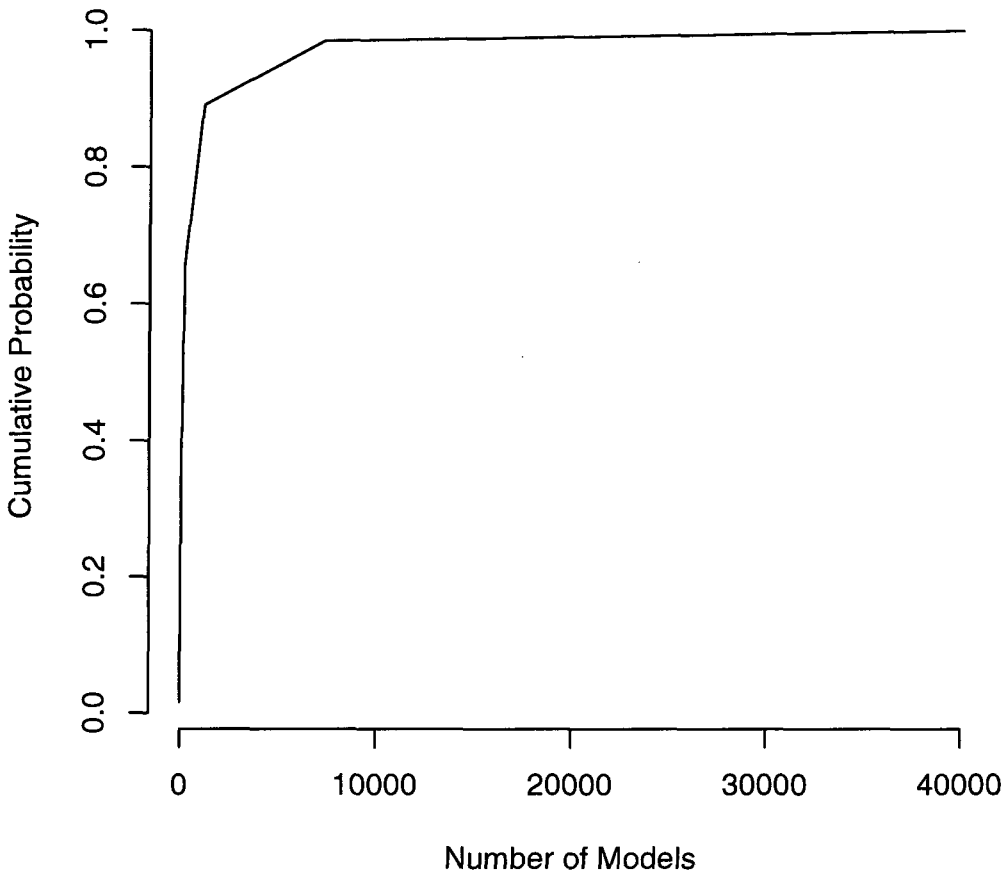

Figure 6.1: Cumulative prior probability for six-factor model space

discriminating between these effects at the expense of others, even though all factors are being viewed as equally important *a priori*.

Using a subset of the models may be unrepresentative of the whole space. For example, with 6 factors, we might use the 197 models with highest prior probability, which is a natural cut-off corresponding to all models involving three or less factors, with a combined prior probability of 0.66. However, under the prior given, we are just as likely to have six active factors as any given three; the prior probabilities for six-factor models are smaller purely because there are more of them, as there are more possible

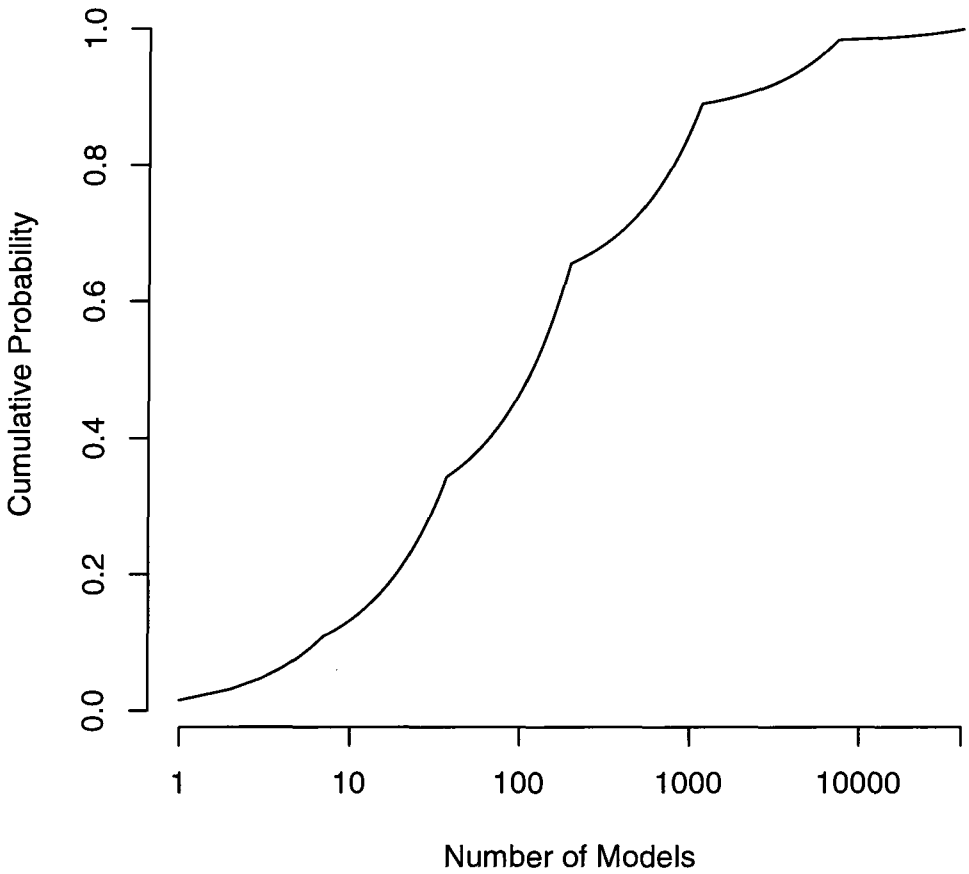

Figure 6.2: Cumulative prior probability for six-factor model space (semi-log scale)

combinations of interaction terms. Using just the top 197 models ignores a large set of possible models with a fairly high combined prior probability (0.34).

6.3 Occam's Window

Madigan and Raftery (1994) use the term 'Occam's window' to describe a method of accounting for model uncertainty in which only models m_i with a posterior probability

$$P(m_i|\mathbf{y}) > \frac{\max_l \{P(m_l|\mathbf{y})\}}{r} \tag{6.1}$$

are used in the calculation of posterior expectations, where $r > 1$ is a tuning parameter to be chosen. In our situation the expected posterior loss is approximated by

$$E_j(L(i,j)|d_n) = \sum_{j \in \mathcal{A}} P(m_j|\mathbf{y}, d_n)L(i,j).$$

where

$$\mathcal{A} = \left\{ m_k : P(m_k|\mathbf{y}) > \frac{\max_l \{P(m_l|\mathbf{y})\}}{r} \right\}.$$

Madigan and Raftery (1994) provide an algorithm for finding $\mathcal{A}$, for graphical models, which we have adapted for linear models. The algorithm builds up models from the null model by adding terms and comparing posterior probabilities. They also provide an algorithm for constructing $\mathcal{A}$ by removing terms from the saturated model. However, due to the structure of our model space, this is less useful, as a main effect term may not be removed unless all interactions in which it is involved have already been removed, so it is required to evaluate the posterior probability for more models in order to find $\mathcal{A}$ under this algorithm. The algorithm that we use to find $\mathcal{A}$ is a slightly altered version of the Up algorithm of Madigan and Raftery. We start with $\mathcal{A} = \emptyset$, and the set of models under consideration $\mathcal{C} = \{\text{constant model}\}$. At any stage of the algorithm, the model, m_i, that we are currently considering, was reached through a number of steps starting at the model with just a constant term, and adding one term at a time. For model m_i, let this set of models (including the constant model and m_i) be $\mathcal{V}(m_i)$. We also define two constants, O_L and O_R. Now let m_i^+ be a supermodel of m_i and consider the log posterior odds, $\log \frac{P(m_i|\mathbf{y})}{P(m_i^+|\mathbf{y})}$. If the log posterior odds is large and positive ($> O_R$), there is evidence in favour of the smaller model, if the log posterior odds are large and negative ($< O_L$), the evidence points towards the larger model. Intermediate values of the log odds mean that both models should be considered. Madigan and Raftery (1994) also rule that, if a smaller model is rejected in favour of a larger one, so are all its submodels, and if a larger model is rejected in favour of a smaller one, all its supermodels are also rejected. Combining these rules means that a model is not permissible if it has probability less than e^{O_L} times the probability of any of its supermodels, or if its probability is less than e^{-O_R} times the probability of one of its submodels.
The algorithm is as follows:

1. Select a model m_i from $\mathcal{C}$

2. $\mathcal{C} \leftarrow \mathcal{C} \setminus \{m_i\}, \mathcal{A} \leftarrow \mathcal{A} \cup \{m_i\}$

3. Select a supermodel m_i^+ of m_i by adding a permissible effect to i.

4. Evaluate $f(\mathbf{y} \mid \mathbf{X}, m_i^+)$.

5. Compute $B = \log\left(\frac{\sup_{m_j \in \mathcal{V}(m_i)}\{f(\mathbf{y}|\mathbf{X},m_j)P(m_j)\}}{f(\mathbf{y}|\mathbf{X},m_i^+)P(m_i^+)}\right)$

6. If $B < O_L, \mathcal{A} \leftarrow \mathcal{A} \setminus \{m_i\}$, if $m_i^+ \notin \mathcal{C}$ then $\mathcal{C} \leftarrow \mathcal{C} \cup \{m_i^+\}$

7. If $O_L \leq B \leq O_R$ then if $m_i^+ \notin \mathcal{C}$ then $\mathcal{C} \leftarrow \mathcal{C} \cup \{m_i^+\}$

8. If there are more supermodels of m_i, go to 3.

9. If $\mathcal{C} \neq \emptyset$ go to 1.

10. Finally, remove from $\mathcal{A}$ any models that do not satisfy the condition (6.1)

In the original formulation of Occam's window, if a model has a lower probability than one of its submodels, then that model always is rejected and not included in the final set $\mathcal{A}$. We do not wish to do this, as the number of parameters has already had an effect on the posterior probabilities, through both the prior and the likelihood. Hence we use $-O_L = O_R = \log(r)$ and only remove models from the final set on the basis of inequality (6.1).

6.3.1 Use of Occam's window

We compare the use of Occam's window with different values of O_L for 5-factor 16-run main effect orthogonal designs. We use $O_L = -O_R = -\log(r)$ for 10 different values of r. These are plotted against the PMD values calculated by using all the models (Figure 6.3). Also plotted are the PMD values obtained by using a reduced model set, of either 106 or 426 models. The sizes of these sets are chosen to be natural cut-off points that do not omit models of the same prior probability as models that are included, as discussed in Section 6.2.

We see that, as r increases, Occam's window becomes a better approximation to evaluation over all models, but even for $r = 2$, the ordering of designs is close to being correct.

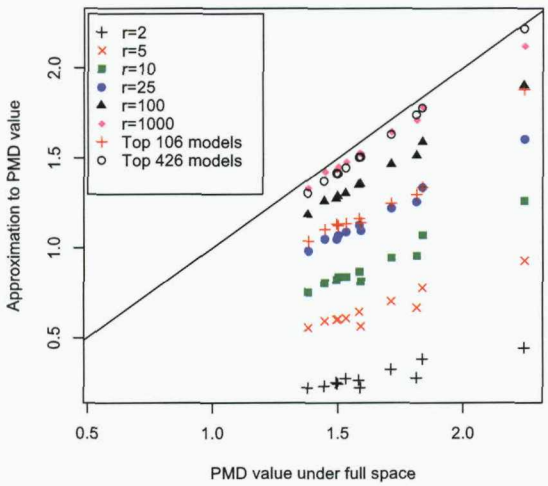

Figure 6.3: Comparison of PMD values for 5 factor 16 run main effects orthogonal designs using Occam's window or a reduced model space, and complete evaluation for the full model space.

The time taken to evaluate a design depends on the design and the value of r. Generally, worse designs will leave more models with substantial posterior probability so will take longer to evaluate. For a simulation size of 1000, we find the PMD value for each of the 5 factor 16 run main effects orthogonal designs, and measure the time taken (in seconds) using 10 values of r. Figure 6.4 shows how the time taken changes with r over this range, with each line corresponding to a design. This is related to the number of models contained in set $\mathcal{A}$, over which the PMD value is calculated. The average number of models in $\mathcal{A}$ for each 5 factor 16 run main effects orthogonal design is shown for different values of r in Figure 6.5.

Figure 6.6 shows how the relative time taken by the different designs remains similar as we change r, and there is a considerable time saving compared to calculating posterior probabilities for all models, for $r = 30$, for example, the time taken is a saving of at least 50% compared to using all models, for all 11 designs.

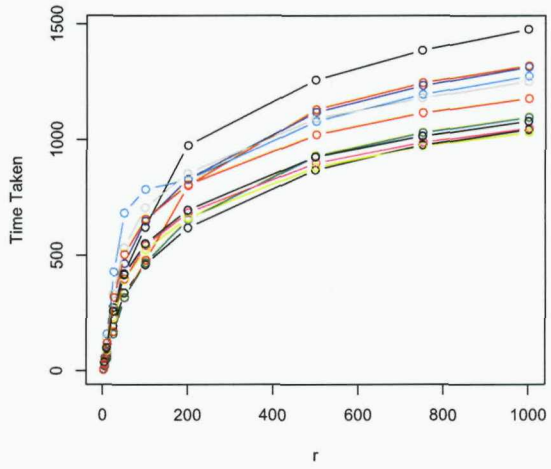

Figure 6.4: Time taken to find PMD vs r for 5 factor 16 run main effects orthogonal designs.

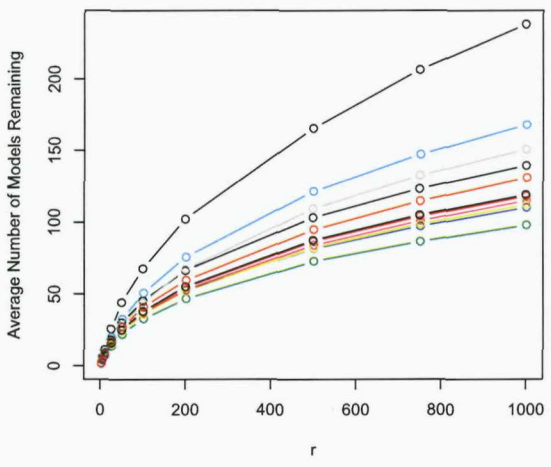

Figure 6.5: Average number of models remaining in $\mathcal{A}$ for different values of r for 5 factor 16 run main effects orthogonal designs.

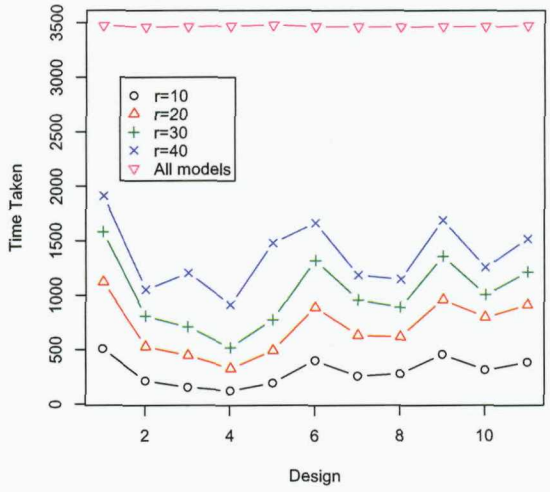

Figure 6.6: Time taken to find PMD for several values of r and by using all models for 5 factor 16 run main effects orthogonal designs.

Table 6.2: Top 10 main effects orthogonal designs for 6 factors in 16 runs evaluated using Occam's window and top 400 models by prior probability.

Rank	Occam's Window		Top 400 models	
	Design	PMD	Design	PMD
1	13	1.20	13	0.63
2	19	1.23	19	0.63
3	18	1.24	24	0.64
4	20	1.24	20	0.64
5	24	1.24	22	0.64
6	23	1.25	23	0.64
7	22	1.26	18	0.64
8	8	1.27	26	0.65
9	26	1.28	27	0.65
10	14	1.28	8	0.69

6.3.2 Results using Occam's Window

We use Occam's window with $r = 20$ to evaluate all 16-run main effects orthogonal designs in 6 factors. This is an example for which calculating posterior probabilities for all models would be prohibitively time-consuming. The top ten designs, are given in Table 6.2.

The order of top 10 designs are not dissimilar to the order for Bingham and Chipman's model space, given in Table 3.10 and repeated in Table 6.2 for comparison. In Figure 6.7, the PMD value of a design, calculated using Occam's window, is related to its aliasing structure. It can be seen that designs with fewer pairs of terms completely aliased have lower PMD objective function values. Pairs of terms i, j without any partial aliasing will have zero as the corresponding entry in $\mathbf{S}_{i,j}$, where $\mathbf{S} = \mathbf{X}'\mathbf{X}$ for the model containing all terms. Given the number of pairs of completely aliased terms, having more zero entries in $\mathbf{S}$ decreases the objective function.

6.4 MCMC Methods

An alternative approach to the computation of the PMD objective function for large model spaces is to construct a Markov Chain Monte Carlo (MCMC) scheme to move around the model space, with equilibrium distribution equal to the posterior model distribution, $P(m|\mathbf{y})$. We use the proportion of iterations that each term is included in the model visited by the chain to approximate the posterior model-averaged probability for that term. An approximation to the expected loss of the model chosen may then be calculated as

$$\min_i E_j \left[L(i,j)|\mathbf{y}, d_n\right] = \sum_{\ell \in \mathcal{S}} \min \left\{p(\ell|\mathbf{y}), c\left[1 - p(\ell|\mathbf{y})\right]\right\} \tag{6.2}$$

as described in Section 2.1.1. The expected losses of the models chosen are then averaged over many runs of simulation, as before. We will use a Metropolis-Hastings Algorithm to approximate $P(m|\mathbf{y})$, the steps of which are given below.

1. Select a model to start at. In this work we experiment with two methods. Firstly, we can randomly select a model from the prior model distribution. Alternatively we can start at the 'true' model from which the response $\mathbf{y}$ was generated.

2. At each iteration of the algorithm, propose a model j to move to from the current model i, using some proposal distribution $q(i,j)$.

3. Accept the proposed model with probability

$$\min\left\{\frac{P(m_j|\mathbf{y})q(m_j,m_i)}{P(m_i|\mathbf{y})q(m_i,m_j)}, 1\right\},$$

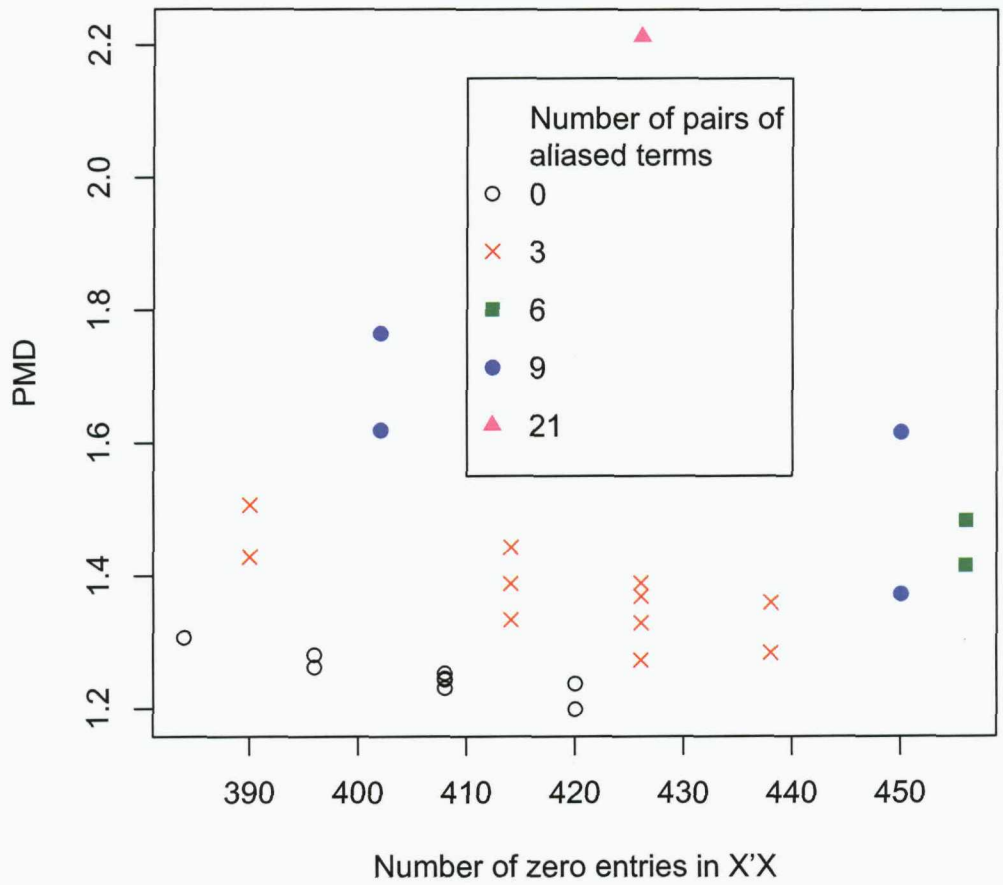

Figure 6.7: The effect of aliasing structure on the PMD objective function value for 16 run main effect orthogonal designs in 6 factors.

using the result that

$$\frac{P(m_j|\mathbf{y})}{P(m_i|\mathbf{y})} = \frac{f(\mathbf{y}|m_j)P(m_j)}{f(\mathbf{y}|m_i)P(m_i)} \tag{6.3}$$

to calculate this probability.

4. If model m_j is not accepted, remain at m_i.

5. Go to step 2. Repeat for a fixed number of iterations.

There are two types of proposal distribution that we will try. Firstly, we can propose m_j which differs from m_i by one term (we will call this a 1-step proposal). To do this, we randomly select a term ℓ from $\mathcal{S}$, with equal weight given to all terms. If $\ell \notin \mathcal{S}_i$, then m_j is model m_i with term ℓ added; if $\ell \in \mathcal{S}_i$, m_j is model m_i with term ℓ removed. If m_j is now a model that does not obey strong heredity, we go back to m_i and try again until a permissible model is reached. This is still only counted as one proposal. Hence $q(m_i, m_j)$ is equal to the reciprocal of the number of permissible moves from m_i. Alternatively, we can use an independence sampler based on the prior as the proposal distribution, where $q(m_i, m_j) = P(m_j)\ \forall i$. We also try using a chain where, at each step, we randomly select either of the proposal distributions, with equal probability (50:50 proposal). The idea of this proposal is to use the independence sampler to make large moves around the model space and the 1-step proposal to explore locally.

6.4.1 Use of MCMC Scheme

We apply the MCMC methods to the example from Section 6.3.1. Before we do this, we check that the use of Equation (2.8) is applicable. Recall that the expected loss given by this equation is that of the model that includes all terms with posterior probability $> \frac{c}{1+c}$. If this model does not obey strong heredity, we would not be permitted to select it, so the expected loss calculated would be incorrect. However, we can show that this is not the case. Suppose, for example, that the AB interaction is in the model that minimised the expected loss, that is, $P(AB|\mathbf{y}) > \frac{c}{1+c}$, where $P(AB|\mathbf{y})$ is the sum of the posterior probabilities of all models containing the term AB. Then

$$P(AB|\mathbf{y}) = \sum_j P(j|\mathbf{y})I(AB \in \mathcal{S}_j) < \sum_j P(j|\mathbf{y})I(A \in \mathcal{S}_j)I(B \in \mathcal{S}_j) = P(A \cap B|\mathbf{y}), \quad (6.4)$$

because $AB \in \mathcal{S}_j \Rightarrow A \in \mathcal{S}_j$ and $B \in \mathcal{S}_j$. Hence

$$P(AB|\mathbf{y}) > \frac{c}{1+c} \Rightarrow P(A \cap B|\mathbf{y}) > \frac{c}{1+c}. \quad (6.5)$$

So, if AB is included in the model, so will the main effect terms of A and B. Therefore, the model which minimises the expected loss is always permissible, so we can use Equation (2.8) to calculate this loss.

Table 6.3: Correlations between PMD values for 5 factor 16 run main effects orthogonal designs using the fulls model space and various MCMC approximations with 100 iterations.

Method	Comparison to values from full space:	
	Correlation	Rank Correlation
Random start, 1-step proposal	0.910	0.855
Start at True, 1-step proposal	0.983	0.900
Start at True, 50:50 Proposal	0.990	0.964
Start at True, Independence Sampler	0.993	0.991

To compare the different proposal distributions and methods of choosing a model to start from, we have evaluated the eleven 5 factor, 16 run main effect orthogonal designs using a chain of length 100 to approximate the full space of 1450 models. In Figure 6.8, the PMD objective function values obtained are plotted against the PMD values evaluated over the full model space. The correlations between each approximation and the evaluation over the full space are given in Table 6.3.

We observe that, when using the 1-step proposal distribution, starting from the true model gives an improved approximation to the PMD objective function, compared to choosing a starting model at random. This is because the true model will generally be in a region of high posterior probability, so the chain does not waste iterations moving away from a starting point that has low probability under the posterior. The other plots in Figure 6.8 and the correlations given in Table 6.3 show that the independence sampler is the most effective proposal distribution.

We also investigate the length of chain required to accurately approximate the objective function. Figure 6.9 shows how the PMD objective function values obtained from independence sampler MCMC approximations, starting at the true model, change as the chain length increases from 2 to 1000. The last point in each line is the objective function obtained using the full model space. The objective function values start off close together, and spread out towards the values from the full model space as the chain length increases. If we are only interested in the objective function as a means of selecting a design, then the ranking of the designs is more important than the actual values of the objective function. Using very short chains of 10 or even 5 iterations gives the correct top two designs. The rank correlation of the objective function values obtained from MCMC methods for various lengths of chain and by using the full model space are given

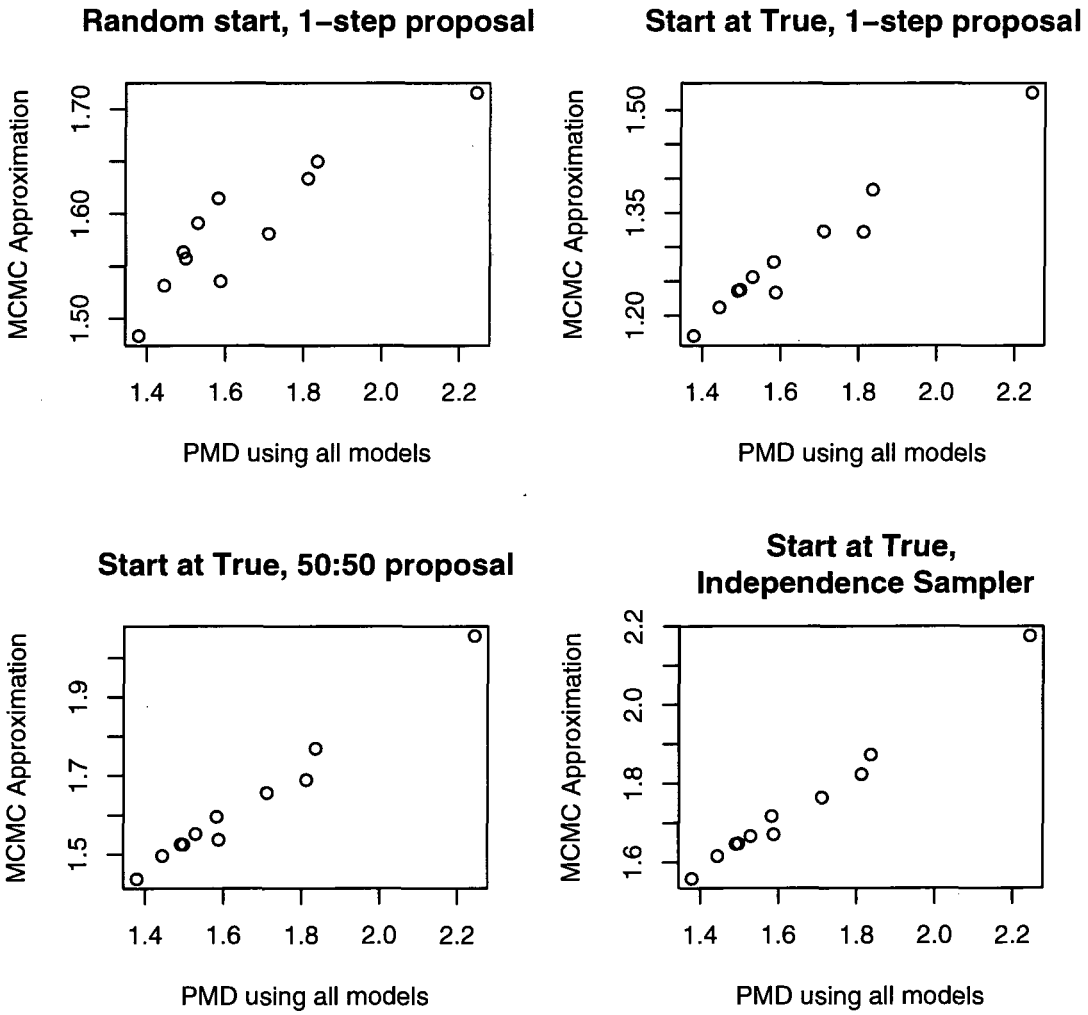

Figure 6.8: Comparison of PMD values for 5 factor 16 run main effects orthogonal designs using various MCMC approximations with 100 iterations.

in Figure 6.10. This shows that the ranking of designs using an MCMC approximation is very similar to that obtained using all the models. We can also compare the use of MCMC methods to using a reduced model space, of 106 or 426 models, as we did in Figure 6.3. In Figure 6.11 we have plotted the PMD objective function values obtained using the full model space against the MCMC approximations with various length chains and values obtained using a subset of the models. The manner in which the objective function approaches its limiting value as the chain length is increased is different to that

of the objective functions obtained via Occam's window as r is increased. When using Occam's window, the PMD values start off low for all designs, and increase with r. This is because increasing r increases the number of models that will be in $\mathcal{A}$, and decreases, on average, the posterior probability of the model chosen, when the probabilities of the models in $\mathcal{A}$ are re-normalised to sum to 1. Hence the expected loss of the model chosen increases with r. When using the MCMC algorithm, the objective function may be either over- or under-estimated with short chain lengths, but move towards the correct value as the chain length increases. With short chain lengths, the algorithm tends to overestimate the objective function for good designs and underestimate it for bad ones.

6.4.2 Timings for MCMC Method

We now investigate how the time taken to evaluate the PMD objective function using the MCMC scheme is affected by the number of factors. An MCMC chain of fixed length will always perform the same number of marginal likelihood evaluations for any number of factors or size of model space. However, to find the marginal likelihood, it is necessary to calculate $\mathbf{V}^*$ (see Equation 1.6), which requires the inversion of a $p \times p$ matrix where p is equal to the number of terms in model i. The number of operations required to invert such a matrix is $O(p^3)$ (see Press, Teukolsky, Vettering and Flannery, 2002, for details). The times taken to estimate the PMD objective function for 16-run designs in $3, \ldots, 9$ factors, using an MCMC approximation with chain length 5000 and simulation size 1000, are plotted in Figure 6.12.

6.4.3 Diagnostics for MCMC

We can use graphical methods to asses the convergence of the Markov chain approximations of the model term probabilities and expected loss. Figure 6.13 shows, for 16-run, 9 factor design 1 (from the catalogue of Sun *et al.* (2002)), for one run of simulation, how the estimated probability of each term converges, for an MCMC approximation starting at the true model and using an independence sampler. Each line on the plot represents one of the 46 model terms, its probability being estimated by the proportion of steps of the chain so far that the model has included that term. We also

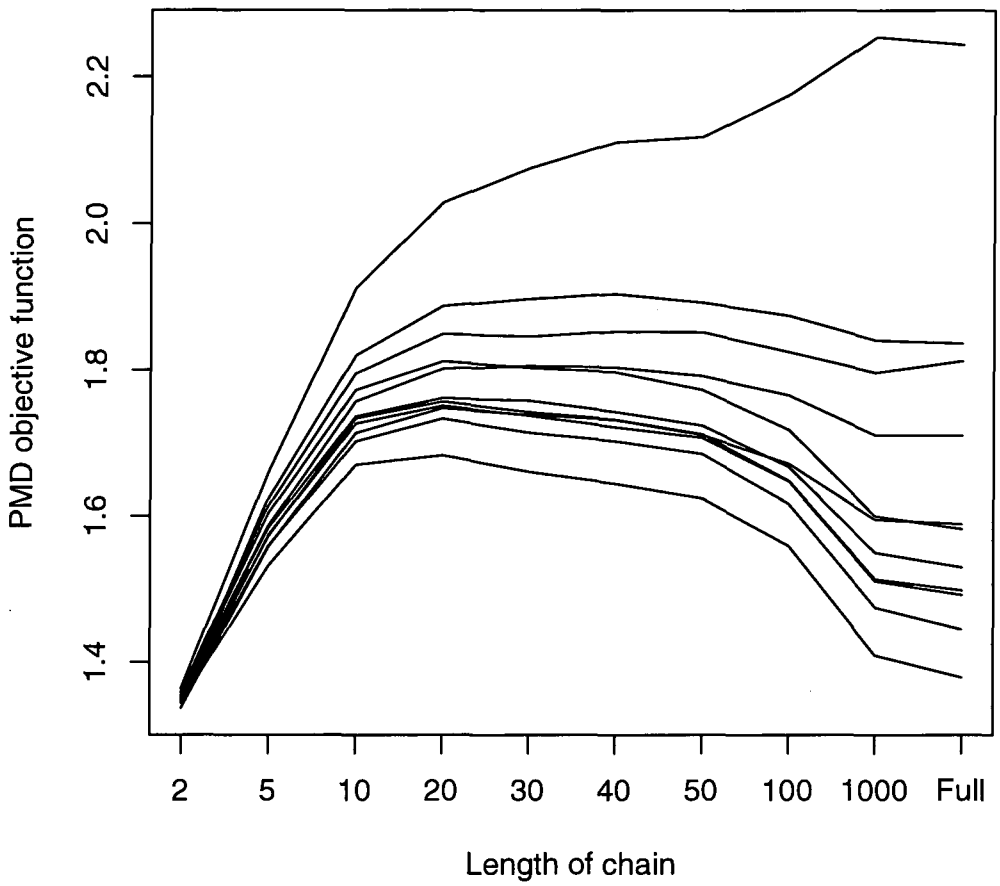

Figure 6.9: PMD values for 5 factor 16 run main effects orthogonal designs using the independence sampler with different numbers of iterations.

plot (in Figure 6.14) the progress of the estimate of expected loss as the chain moves around the model space. This is calculated using Equation (2.8).

After some initial movement, the estimates of the term probabilities settle down after the first 5000 steps of the chain, and do not alter much in the next 5000 steps. Similar behaviour can be seen in the estimation of the expected loss. Another example, for which the estimated expected loss is much lower, is shown in Figures 6.15 and 6.16. This example is for 16-run design 84 in nine factors (from the catalogue of Sun *et al.* (2002)).

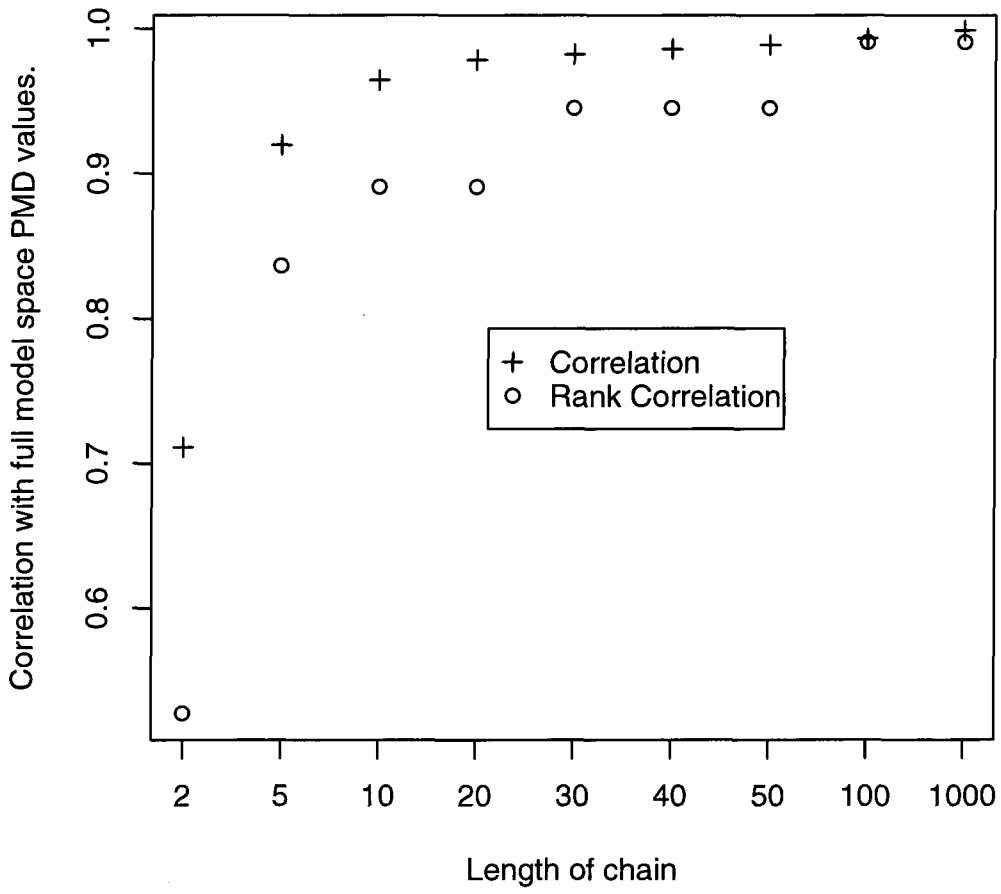

Figure 6.10: Correlations of PMD values from independence samplers with those from full model space for 5 factor 16 run main effect orthogonal designs.

We observe that most of the terms have a probability near 0 or 1, resulting in a low expected loss; there is only one term for which there is any real uncertainty about whether it should be included. Again, after about 5000 steps of the chain, the term probabilities and expected loss have just about converged.

In our first example, the proposed model was accepted about 8% of the time. In the second example, the posterior less closely resembled the prior and hence the proposal distribution, and probability was concentrated on a smaller set of models. For this

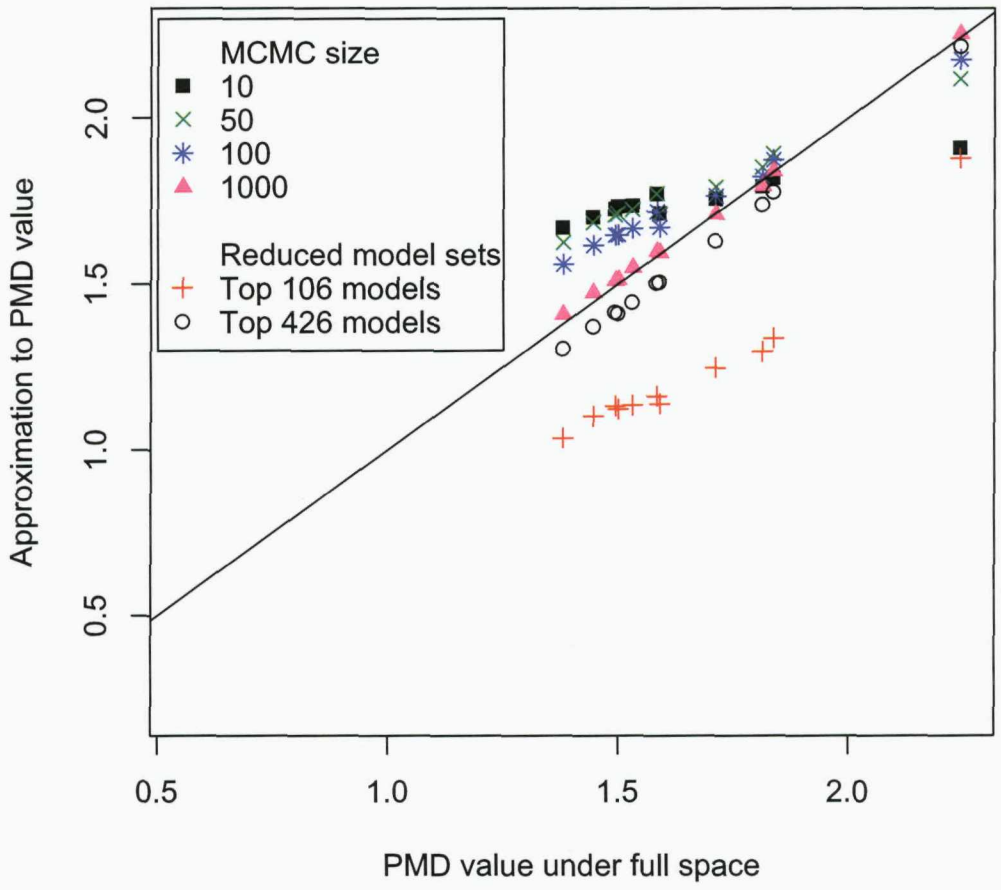

Figure 6.11: Comparison of PMD values for 5 factor 16 run main effects orthogonal designs using the independence sampler with different numbers of iterations.

example, the acceptance rate falls to about 4%.

6.4.4 Results from MCMC Method

We use the MCMC scheme to obtain approximations to the PMD objective function for the 16 run main effects orthogonal designs in 6 to 9 factors. The model space of interest allows models that contain any subset of the main effect and 2-factor interaction terms, subject to heredity. The sizes of the model spaces involved here (given in Table 6.1)

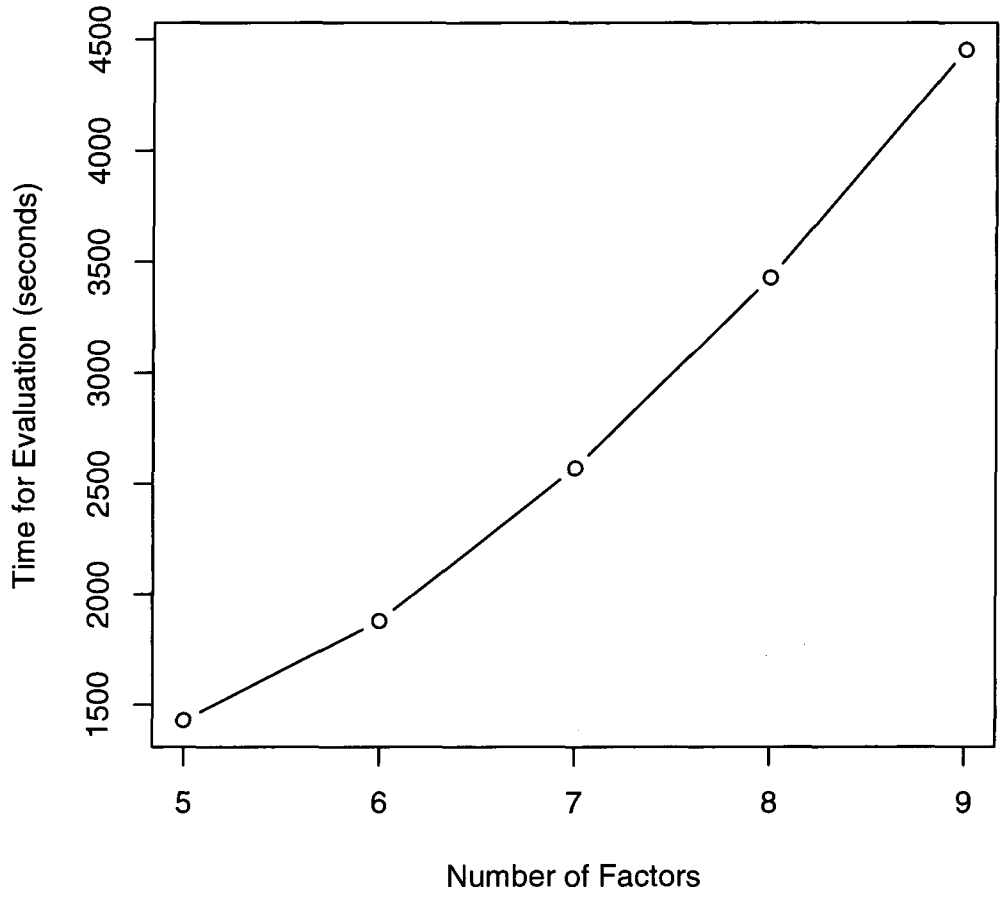

Figure 6.12: Time taken to perform 1000 evaluations of PMD objective function using a MCMC with chain length 5000 for 5 to 9 factors.

mean that an evaluation over the full model space would be impossible, and there would be difficulties in even using Occam's window as an approximation, as for any reasonable r, the size of $\mathcal{A}$ may become very large. For completeness, we present a list of the top ten best 16-run main effects orthogonal designs at each of 3 to 9 factors, although when there were < 6 factors, the evaluation was made using the full model space as use of the MCMC method was not required. These results are given in Table 6.4.

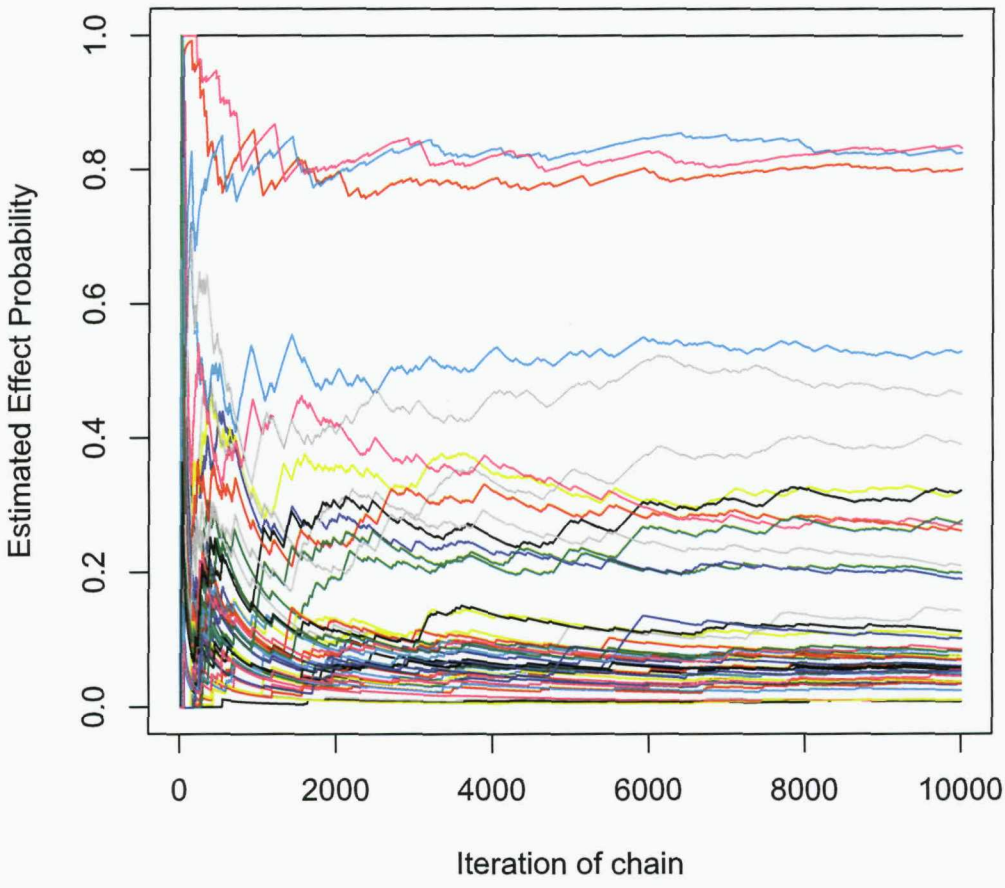

Figure 6.13: Example of convergence of term probabilities.

6.4.5 Comparison of the three methods

For the set of 16 run main effects orthogonal designs in 6 factors, we have used three different approximations to the objective function. These are: Occam's window with $r = 20$, evaluation using only the 197 models with highest prior probability, and an MCMC approximation using chains of length 5000. The objective function values obtained using each of these three methods for these 27 designs are plotted these against each other in Figure 6.17. The same designs have low PMD objective function values in

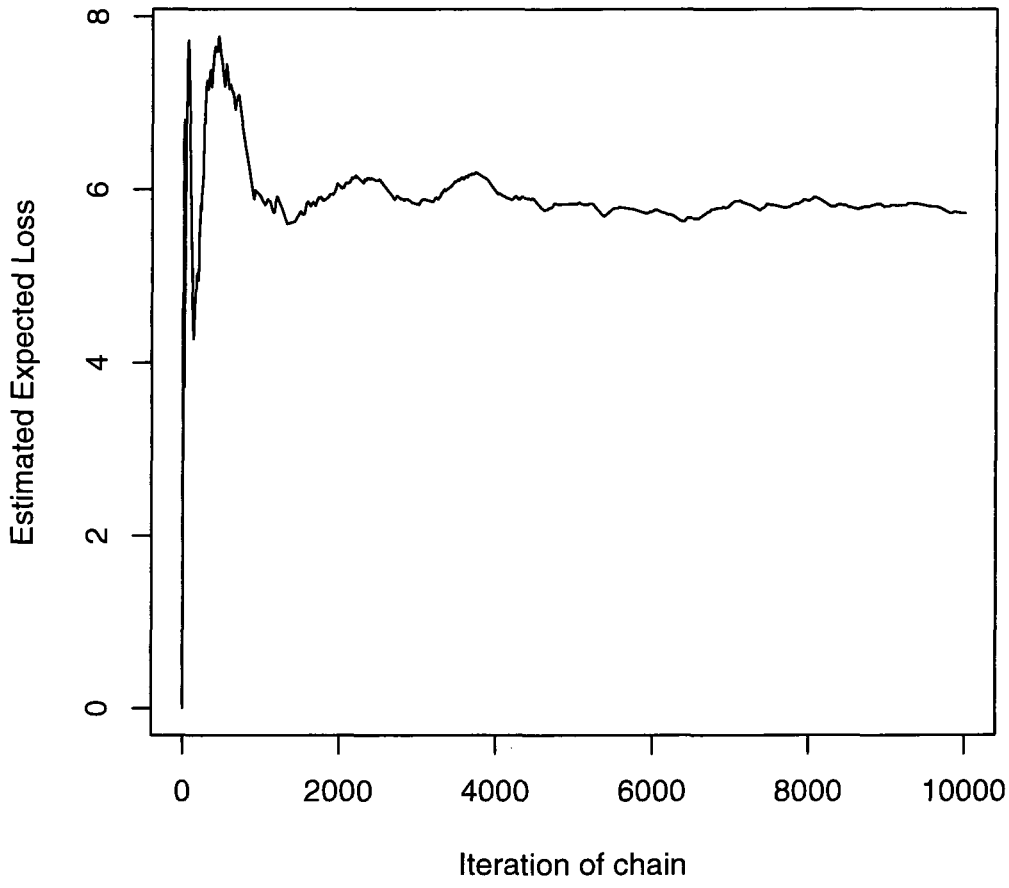

Figure 6.14: Example of convergence of estimated expected loss

each approximation, suggesting that these are the designs that would perform well if evaluations were done using the full model space.

6.5 Monte Carlo Approximations to other Objective Functions

We move on to evaluating the objective functions associated with the HD and MD criteria, for large model spaces. These criteria differ from the PMD in that it is not necessary to calculate the posterior model probabilities in order to evaluate the objective

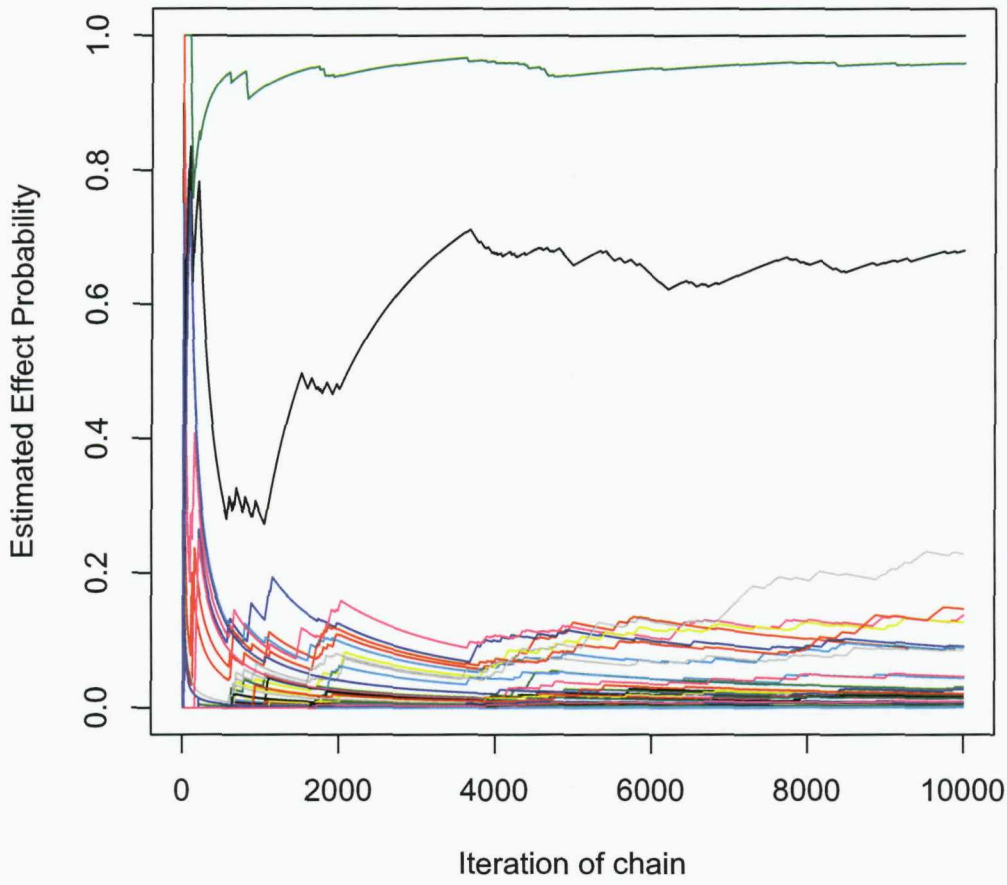

Figure 6.15: Example of convergence of term probabilities.

function. Rather, the *prior* model probabilities are used to find the weighted average distance (Kullback-Leibler or Hellinger) between the predictive posterior distributions of pairs of models. As posterior model probabilities are not required, we do not need to use MCMC methods. However, we can use a Monte Carlo approach of sampling pairs of models from the prior distribution, finding the distance between their predictive distributions, and averaging this distance over many sampled pairs of models. It is possible that this method will give a better approximation to the objective function evaluated over all models than is obtained by using a subset of the models with highest

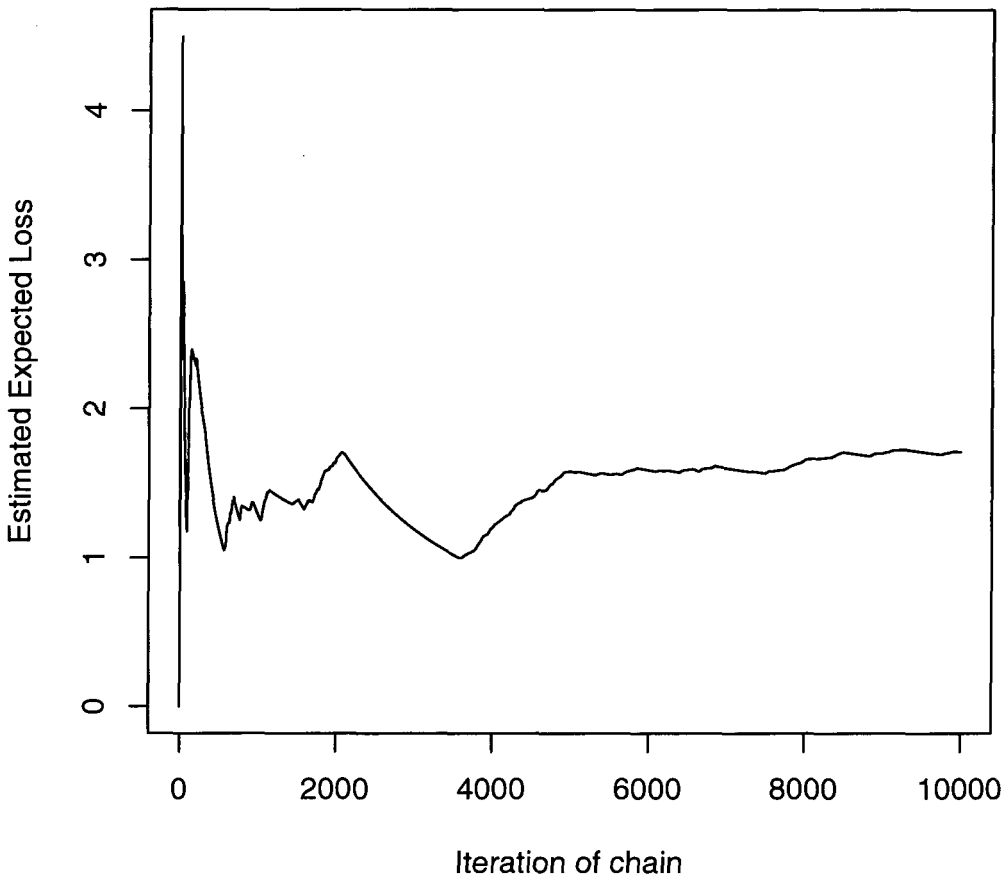

Figure 6.16: Example of convergence of estimated expected loss

prior probability. In comparison, Chipman (1996) obtain a subset of models with high prior probability by simulating models from the prior and discarding all but the most probable. The HD objective function is then evaluated over this set of models only. To test the method for these objective functions, we use Bingham and Chipman's model space for 4 factors, which contains 1024 models, using p calculated to give $E(\#\text{ effects}) = 3$. For this model space, we can obtain the exact values of the MD and HD objective functions for the five 16-run main effects orthogonal designs, evaluated over all $\binom{1024}{2} = 523776$ pairs of models, within a reasonable amount of computing time.

Table 6.4: Top 10 16 run main effects orthogonal designs under PMD criterion for 3 to 9 factors

	Number of Factors						
	3	4	5	6	7	8	9
Ranking 1	2	3	4	13	32	68	71
2	3	4	5	19	49	42	79
3	1	5	7	8	53	77	32
4		2	8	20	28	76	36
5		1	10	24	55	67	84
6			11	14	33	66	70
7			3	18	43	36	78
8			2	23	54	41	77
9			9	6	39	40	82
10			6	22	21	39	81

We will compare these exact values to those obtained from Monte Carlo evaluations as the simulation length is increased, and also to approximations from using a subset of the models with highest prior probability.

6.5.1 Methodology

To calculate the Monte Carlo Approximation to either objective function, we follow these steps:

1. Generate a model m_i from the prior model distribution. Each main effect, is independently included with probability p. Conditional on these, interaction terms are independently included with probability $0.01p, 0.5p$ or p if $0, 1$ or 2 of the associated main effect terms are included.

2. Independently generate another model m_j in exactly the same way.

3. For the HD criterion, calculate $\frac{1}{2}H(f_i, f_j)$ (see Section 1.3.1 or Bingham and Chipman (2007) for definition of $H(,))$. The $\frac{1}{2}$ is necessary because $HD = \sum_{i<j} P(m_i)P(m_j)H(f_i, f_j) = \frac{1}{2}\sum_{i,j} P(m_i)P(m_j)H(f_i, f_j) = \frac{1}{2}E_{i,j}(H(f_i, f_j))$ as $H(,)$ is symmetric and $H(i, i) = 0$. For the MD criterion, the distance function $I(m_i, m_j) = \int f(\mathbf{Y}|m_i) \log\left(\frac{f(\mathbf{Y}|m_i)}{f(\mathbf{Y}|m_j)}\right) d\mathbf{Y}$ is not symmetric, so we calculate $\frac{1}{2}(I(m_i, m_j) + I(m_j, m_i))$.

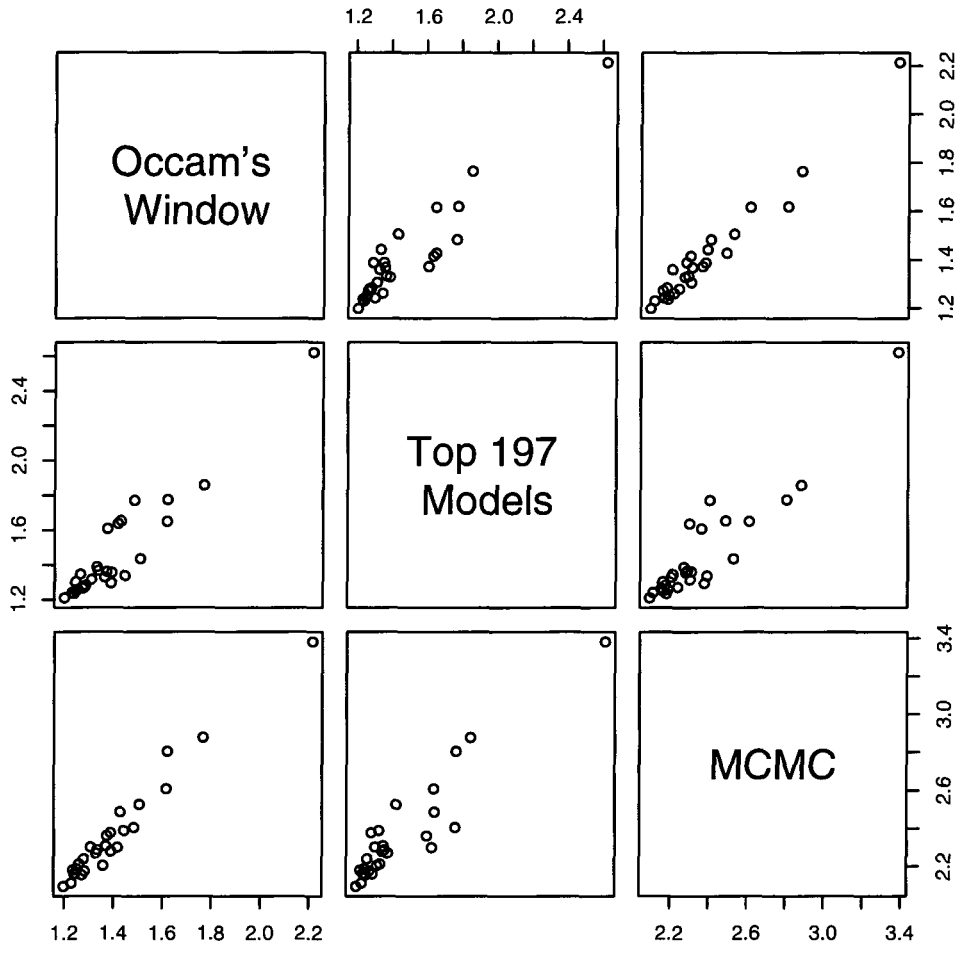

Figure 6.17: Comparison of approximations to the PMD objective function for 16 run 6 factor main effects orthogonal designs

4. Average the distances obtained over many samples of pairs of models from the prior.

To examine the performance of the method of using a subset of models with the highest prior probability, we first calculate the distances between all possible pairs of models. To find the approximation to the objective function that we would obtain by using, for example, the 100 models with highest prior probability, we sum the distances between the pairs formed using these 100 models, and divide by the square of the total

probability of this subset of models, to account for the fact that we would have re-normalised the probabilities to sum to 1. To enable fair comparison between this method and the Monte Carlo approach, we look at the number of pairs of models for which distances must be calculated. If we use the M_0 models with highest prior probability, then $M_0(M_0 - 1)/2$ pairs of models must be used.

6.5.2 Results for MD Objective Function

For each of the 16-run main effects orthogonal designs in four factors, we calculate the approximation to the MD objective function using the two methods. The progress of these approximations as the number of pairs of models used increases is shown in Figure 6.18. Each colour in the plot corresponds to one design, with solid and dashed lines indicating the Monte Carlo and subset of models approximations respectively.
We observe that the Monte Carlo approximation moves around initially, but quickly homes in on the correct value. The values of the objective function for the method of using a subset of models steadily increase to approach the true values, but always underestimate the objective function, even when a large number of models are used.
Another comparison that we can make is to plot the mean squared error of the objective function, compared to the true value, across the five designs, for both the methods. The results of this are shown on a semi-log scale in Figure 6.19, and again demonstrate that the Monte Carlo approximation quickly gets close to the true value for all the designs.

6.5.3 Results for HD Objective Function

The results for the HD objective function are similar to those for the MD. These are shown in Figures 6.20 and 6.21.

6.5.4 Best designs under MD and HD criteria

We use the Monte Carlo evaluation methods to rank the 16-run main effects orthogonal designs in 5-9 factors. The top ten designs presented here should be compared to those in Table 3.10, where evaluation was made using the 400 models with highest prior probability. The designs that are ranked highly, and their ordering, are similar under

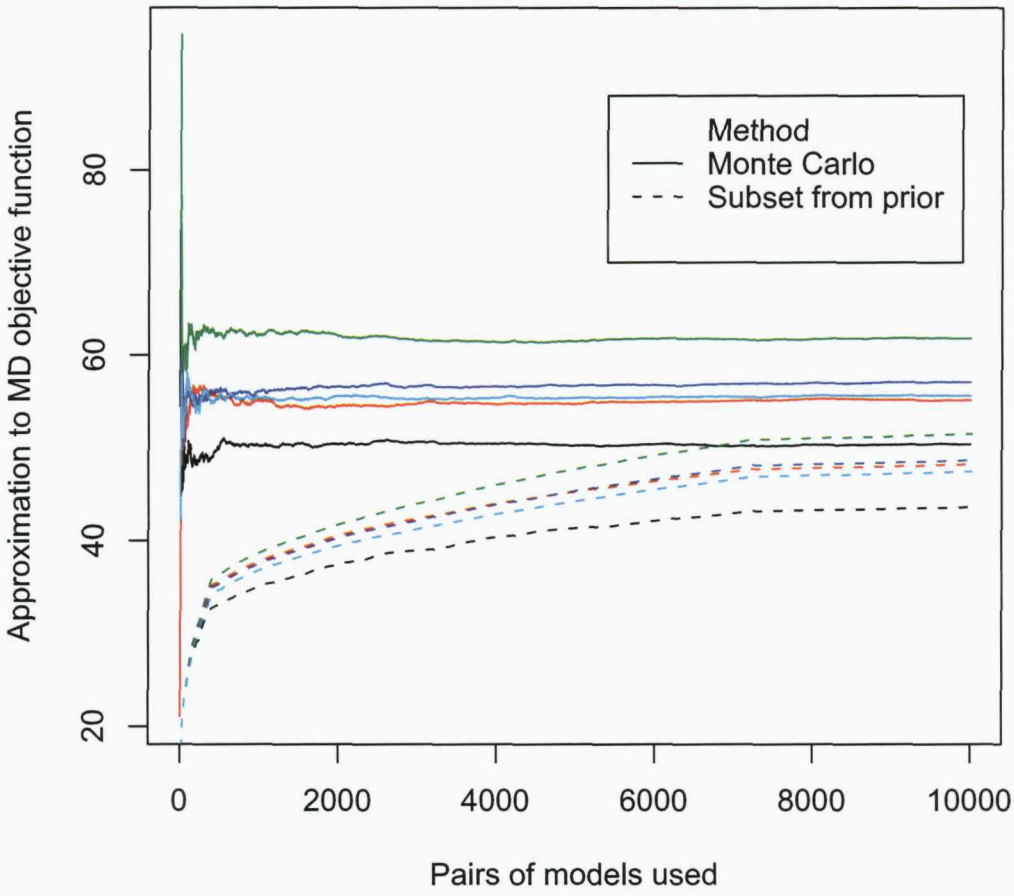

Figure 6.18: Comparison of Approximation Methods for MD Objective Function

both methods. However, the Monte Carlo method will be able to obtain a more accurate approximation to the objective function with less computational expense.

6.6 Conclusions

In this chapter we have shown that, if we wish to consider more than a few factors and allow the possibility of 2-factor interactions, then the number of possible models can grow very rapidly if we do not use a restrictive prior such as effect forcing. This can

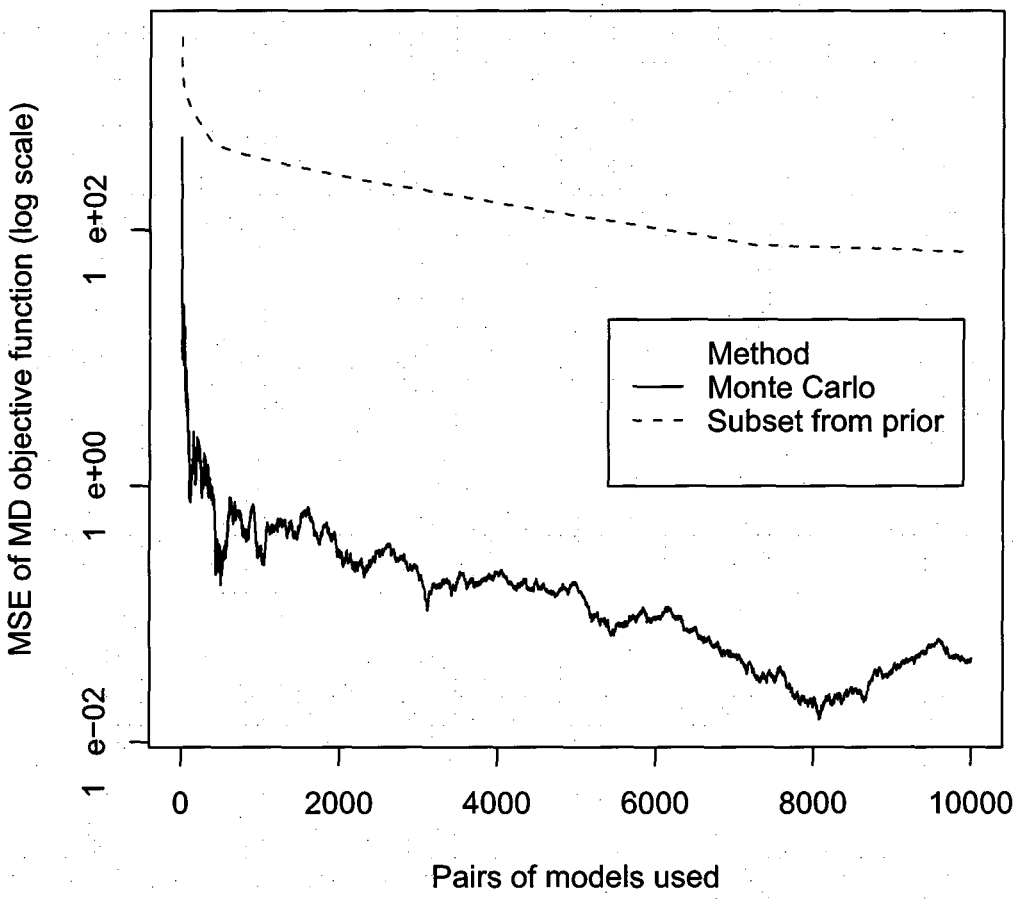

Figure 6.19: Mean squared errors of approximation methods for MD objective function

increase the amount of computation required to obtain the objective functions for the PMD, HD and MD criteria. Three approaches were considered for the PMD: only considering a subset of the models with relatively high prior probability, evaluating the objective function over a subset of models with high posterior probability using Occam's window, and using an MCMC scheme to estimate marginal term probabilities from which we can calculate the objective function.

The use of a subset of models with high prior probability, equivalent to changing the prior to give a prior probability of zero for some models, is a fairly straightforward way

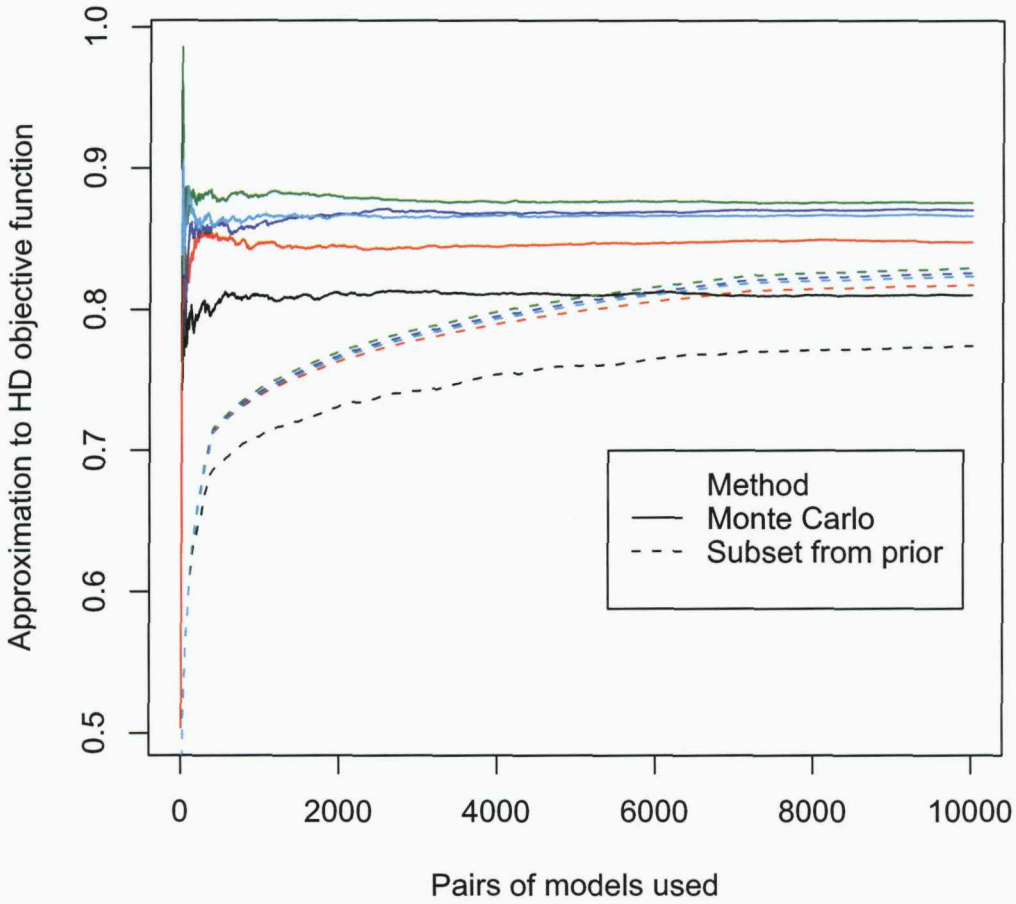

Figure 6.20: Comparison of approximation methods for HD objective function

to reduce the computation required. For the model spaces discussed in this chapter, this method can provide a reasonable approximation to the objective function. We have shown that some care is needed in selecting which models to include in the set used to evaluate the objective function if all factors are equally important and we require our prior to reflect this. Occam's window allows us to evaluate the objective function over a set of models with high posterior probability. Madigan and Raftery (1994) argue that this is not merely a computational convenience, but rather, a natural Bayesian extension of the principle of Occam's razor, which states that we should accept the simplest

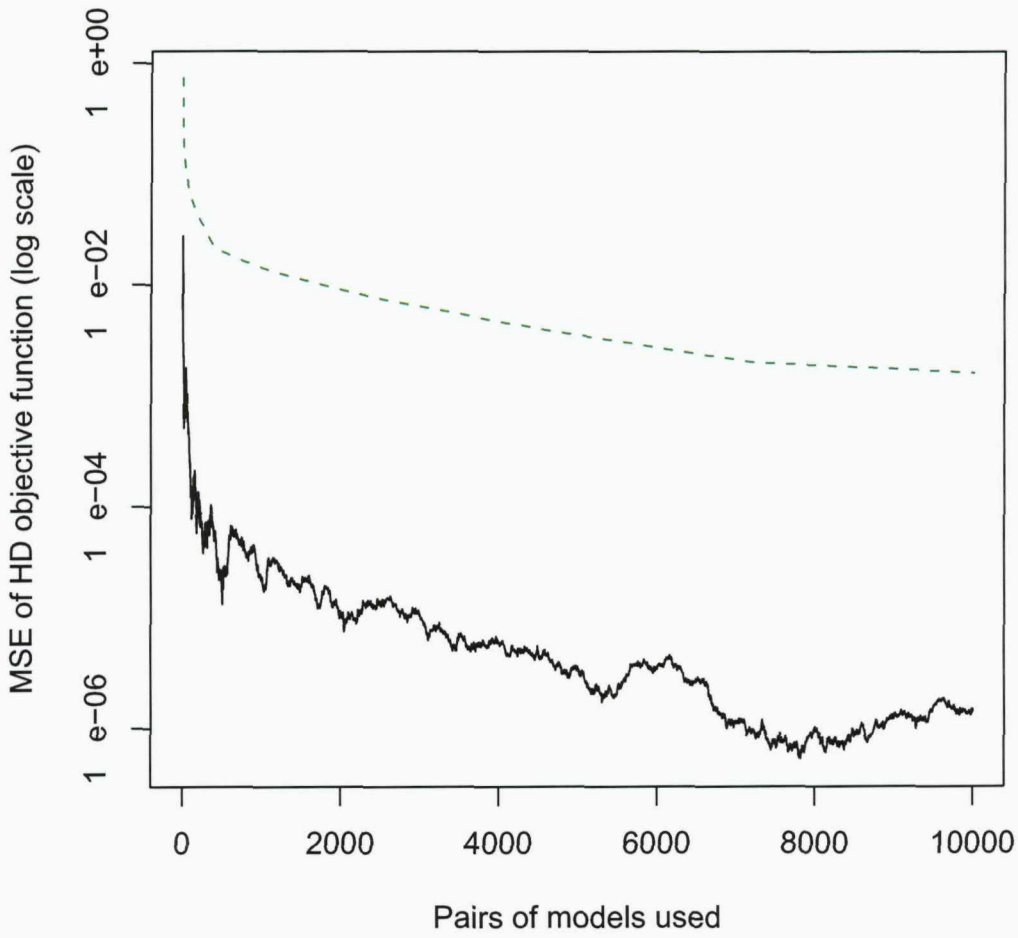

Figure 6.21: Mean squared errors of approximation methods for HD objective function

hypothesis which explains the data. In Occam's window, the simplest set of models which can explain the data are selected. Computationally, use of Occam's window can allow the use of larger model spaces than could be investigated by evaluation of posterior probabilities for all models. However, the set of accepted models can use up a lot of computer memory, and the algorithm can be slow for larger numbers of factors. For example, we were not able to evaluate designs for 7 or more factors in the model space described in this chapter for any reasonable value of r.

The MCMC scheme does not require as much to be held in the computer memory as

Table 6.5: Top 10 16-run main effects orthogonal designs under HD criterion for 3 to 9 factors

	Number of Factors						
	3	4	5	6	7	8	9
Ranking 1	2	3	4	19	55	77	84
2	3	4	5	13	32	76	83
3	1	5	8	20	49	42	86
4		2	7	24	33	67	53
5		1	10	22	28	68	65
6			3	8	43	79	74
7			11	23	54	50	75
8			9	14	53	71	71
9			2	18	36	48	44
10			6	26	39	41	81

Table 6.6: Top 10 16-run main effects orthogonal designs under MD criterion for 3 to 9 factors

	Number of Factors						
Ranking	3	4	5	6	7	8	9
1	2	3	4	5	6	6	4
2	3	4	3	4	5	4	5
3	1	5	5	8	12	5	3
4		2	7	14	11	3	2
5		1	8	6	22	9	25
6			2	13	28	12	10
7			10	19	7	26	12
8			11	3	4	18	7
9			9	7	33	17	9
10			6	12	21	20	13

Occam's window, only details of the current and proposed models. This enables us to look at even larger model spaces, such as those for 7, 8 or 9 factors in this chapter. As with all MCMC methods, it is important to check that the chain converged. The acceptance rate of proposed moves is quite low, so the value of the objective function may move slowly without having converged.

We have also introduced Monte Carlo methods for evaluating the MD and HD objective functions. These appear to produce an accurate approximation to the objective function with less computation than would be required by using a subset of models with high prior probability.

Chapter 7

Summary and Further Work

7.1 Summary

In this thesis we have examined the problem of choosing an experimental design for model selection in a Bayesian framework. The Penalised Model Discrepancy criterion was introduced in Chapter 2, and minimises the expectation of a weighted sum of the expected number of terms wrongly omitted and incorrectly included in the chosen model. Software has been written to evaluate the PMD objective function and to search for good designs under the criterion.

In Chapter 3 the PMD criterion was applied to the selection of designs for screening experiments, for several model spaces suitable for screening situations. If only main effects are considered, the PMD criterion selects a main effects orthogonal design if one is available. For a model space where all main effects are always included, and exactly one two-factor interaction, a non-orthogonal design may be preferred over an orthogonal design if it reduces aliasing involving 2-factor interactions. A study was also made into the sensitivity of the PMD criterion to the hyperparameters of the prior distribution for this model space. For a larger model space, where any subset of main effect and 2-factor interaction terms is permitted subject to strong heredity, attention was focussed on 16-run main effect orthogonal designs. A comparison was made to other Bayesian criteria for choosing designs for model discrimination. Of the criteria studied, the HD criterion of Bingham and Chipman (2007) agreed most closely with the PMD in its

ranking of designs. The MD and F criteria selected similar designs to each other, and tended to rank regular designs more highly than the PMD and HD criteria. A simulation study was also made on the model discrimination performance of designs selected by the four criteria. There was not a large difference in the performance of the designs in the simulations, although designs chosen by the PMD criterion were usually slightly better. These simulation studies was also used to show the dependence of the average number of model terms incorrectly omitted or included on c, the loss incurred by incorrectly including a term. The simulation studies showed that, for the particular situation studied, using $c = 1$ can lead to far more terms being missed than wrongly included, suggesting that a smaller value of c may be more appropriate.

Two examples of the use of the PMD criterion for the selection of follow-up runs were presented in Chapter 4. In the first, injection moulding, example the PMD criterion selected distinct points that were not repeats of runs from the initial experiment. The F and HD criteria performed similarly; however, the MD criterion selected a design with a repeated run. The PMD criterion was also seen to penalise designs more heavily than the F criterion if they had aliasing which gave rise to cancelling of effects. In the second, chemical reactor, example, an investigation was made into the effect of the initial design on the follow-up runs chosen. Different initial designs result in different factors having high marginal posterior probability after the initial experiment; factors with high probability are often chosen to be balanced in the follow-up runs.

The tribology example in Chapter 5 demonstrated the use of the PMD criterion for a real experiment. This experiment shows one way that the PMD criterion may be used to select follow-up runs when there are multiple responses to an experiment. This example also showed the type of Bayesian analysis which may be used for experiments designed using the PMD criterion.

In Chapter 6 we addressed the computational issues that arise when the space of potential models is very large, and produced algorithms that were implemented in our programs for the evaluation of the PMD objective function. These algorithms are effective in reducing the time required to produce a good estimate of the objective function. Code in R for evaluation of the MD and HD objective functions when the model space is large, was also produced. This made use of a Monte Carlo method which

was shown to produce a more accurate approximation to the true objective function value, for less computational expense, than is obtained by using a subset of models with high prior probability.

7.2 Further Work

One possible extension to this work is to look at generalised linear models instead of the linear models that we have used here. In most generalised linear models, such as those where the response follows a binomial or Poisson distribution, there is not a closed form expression for the posterior parameter distributions or model likelihoods. This potentially adds another layer of computation required, as MCMC methods are needed to produce these. However, if the size of the model space is very large, we need to use Markov Chain methods to approximate the posterior model probabilities anyway, as described in Chapter 6. It would be possible to use methods such as Reversible Jump Markov Chain Monte Carlo (Green 1995) to expand this idea to include generalised linear models. Alternatively, we can use a Laplace approximation to the posterior marginal likelihood of each model, as given in Kass and Raftery (1995).

One important use of any statistical model is that it may be used to predict, or give a predictive distribution for, future observations from the response being modelled. Therefore design criteria that are based on the ability of the design to select a model and estimate its parameters so that it will produce good predictions would be useful. The PMD criterion treats the omission of any term from the model chosen, compared to the 'true' model as equally important. This would not be the case with prediction based criteria, for which the omission of terms that make a large difference to the expected response would incur a higher loss.

In the examples presented in Chapters 3 and 4, the HD and PMD criteria generally select similar designs, and rank designs in a similar order. Evaluation of the HD objective function does not require simulation, so requires less time to compute than the PMD. Therefore, use could be made of the HD criterion to quickly find designs that perform well under the PMD criterion. For example, a shortlist of designs found by a search under the HD criterion could be evaluated under the PMD and the best design

selected. Alternatively, the HD criterion could be used within the Modified Fedorov Exchange Algorithm to order the candidate points at step 3 of the algorithm given in 2.4 by the improvement in the HD objective function each point would make if swapped in to the design. This step would reduce (on average) the number of designs which are evaluated at each step under the PMD criterion before one is found that reduces the objective function, and should reduce the computation time required.

In this thesis we have only considered designs with factors at two levels. If the model space under consideration contains models with quadratic or higher order terms then more than 2 levels are necessary to allow estimation of these terms. A modified Fedorov exchange algorithm might still be used to search for designs if we are satisfied with restricting the factors to a limited number of levels. Alternatively, methods such as simulated annealing (see, for example, Brooks and Morgan, 1995) could be used to allow variables to be set at any level within a permitted range.

Appendices

Appendix A

Derivation of HD objective function for non-zero prior means.

The prior predictive distribution under model i is $\mathbf{Y} \sim N(\mathbf{X}_i\boldsymbol{\mu}_i, a_i(\mathbf{I} + \mathbf{X}_i\mathbf{V}_i\mathbf{X}_i'))$ (see O'Hagan and Forster). Let $\hat{\mathbf{Y}}_i = \mathbf{X}_i\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i = \mathbf{I} + \mathbf{X}_i\mathbf{V}_i\mathbf{X}_i'$. Then the posterior predictive density of model i is

$$f_i = \frac{1}{(2\pi)^{n/2}|a_i\boldsymbol{\Sigma}_i|^{1/2}} \exp\left(-\frac{1}{2}\left(\mathbf{Y} - \hat{\mathbf{Y}}_i\right)' \frac{\boldsymbol{\Sigma}_i^{-1}}{a_i}\left(\mathbf{Y} - \hat{\mathbf{Y}}_i\right)\right)$$

and

$$\begin{aligned}
f_i^{\frac{1}{2}} f_j^{\frac{1}{2}} &= \frac{1}{(2\pi)^{n/2}|a_i\boldsymbol{\Sigma}_i|^{1/4}|a_j\boldsymbol{\Sigma}_j|^{1/4}} \exp\left(-\frac{1}{4}\left[\left(\mathbf{Y} - \hat{\mathbf{Y}}_i\right)' \frac{\boldsymbol{\Sigma}_i^{-1}}{a_i}\left(\mathbf{Y} - \hat{\mathbf{Y}}_i\right) + \left(\mathbf{Y} - \hat{\mathbf{Y}}_j\right)' \frac{\boldsymbol{\Sigma}_j^{-1}}{a_j}\left(\mathbf{Y} - \hat{\mathbf{Y}}_j\right)\right]\right) \\
&= \frac{1}{(2\pi)^{n/2}|a_i\boldsymbol{\Sigma}_i|^{1/4}|a_j\boldsymbol{\Sigma}_j|^{1/4}} \exp\left(-\frac{1}{2}\left[\mathbf{Y}'\left(\frac{\boldsymbol{\Sigma}_i^{-1}}{2a_i} + \frac{\boldsymbol{\Sigma}_j^{-1}}{2a_j}\right)\mathbf{Y} - \mathbf{Y}'\left(\frac{\boldsymbol{\Sigma}_i^{-1}\hat{\mathbf{Y}}_i}{2a_i} + \frac{\boldsymbol{\Sigma}_j^{-1}\hat{\mathbf{Y}}_j}{2a_j}\right)\right.\right. \\
&- \left.\left.\left(\frac{\boldsymbol{\Sigma}_i^{-1}\hat{\mathbf{Y}}_i}{2a_i} + \frac{\boldsymbol{\Sigma}_j^{-1}\hat{\mathbf{Y}}_j}{2a_j}\right)'\mathbf{Y} + \left(\frac{\hat{\mathbf{Y}}_i'\boldsymbol{\Sigma}_i^{-1}\hat{\mathbf{Y}}_i}{2a_i} + \frac{\hat{\mathbf{Y}}_j'\boldsymbol{\Sigma}_j^{-1}\hat{\mathbf{Y}}_j}{2a_j}\right)\right]\right) \\
&= \frac{\exp\left(-\frac{1}{2}\left(\frac{\hat{\mathbf{Y}}_i'\boldsymbol{\Sigma}_i^{-1}\hat{\mathbf{Y}}_i}{2a_i} + \frac{\hat{\mathbf{Y}}_j'\boldsymbol{\Sigma}_j^{-1}\hat{\mathbf{Y}}_j}{2a_j}\right) + \frac{1}{2}\left(\frac{\boldsymbol{\Sigma}_i^{-1}\hat{\mathbf{Y}}_i}{2a_i} + \frac{\boldsymbol{\Sigma}_j^{-1}\hat{\mathbf{Y}}_j}{2a_j}\right)'\mathbf{B}^{-1}\left(\frac{\boldsymbol{\Sigma}_i^{-1}\hat{\mathbf{Y}}_i}{2a_i} + \frac{\boldsymbol{\Sigma}_j^{-1}\hat{\mathbf{Y}}_j}{2a_j}\right)\right)}{(2\pi)^{n/2}|a_i\boldsymbol{\Sigma}_i|^{1/4}|a_j\boldsymbol{\Sigma}_j|^{1/4}} \\
&\times \exp\left(\frac{-1}{2}\left(\mathbf{Y} - \mathbf{B}^{-1}\left(\frac{\boldsymbol{\Sigma}_i^{-1}\hat{\mathbf{Y}}_i}{2a_i} + \frac{\boldsymbol{\Sigma}_j^{-1}\hat{\mathbf{Y}}_j}{2a_j}\right)\right)'\mathbf{B}\left(\mathbf{Y} - \mathbf{B}^{-1}\left(\frac{\boldsymbol{\Sigma}_i^{-1}\hat{\mathbf{Y}}_i}{2a_i} + \frac{\boldsymbol{\Sigma}_j^{-1}\hat{\mathbf{Y}}_j}{2a_j}\right)\right)\right)
\end{aligned}$$

where $\mathbf{B} = \left(\frac{\boldsymbol{\Sigma}_i^{-1}}{2a_i} + \frac{\boldsymbol{\Sigma}_j^{-1}}{2a_j}\right)$.

We know that

$$\int \frac{\exp\left(-\frac{1}{2}\left[\left(\mathbf{Y} - \mathbf{B}^{-1}\left(\frac{\boldsymbol{\Sigma}_i^{-1}\hat{\mathbf{Y}}_i}{2a_i} + \frac{\boldsymbol{\Sigma}_j^{-1}\hat{\mathbf{Y}}_j}{2a_j}\right)\right)'\mathbf{B}\left(\mathbf{Y} - \mathbf{B}^{-1}\left(\frac{\boldsymbol{\Sigma}_i^{-1}\hat{\mathbf{Y}}_i}{2a_i} + \frac{\boldsymbol{\Sigma}_j^{-1}\hat{\mathbf{Y}}_j}{2a_j}\right)\right)\right]\right)}{(2\pi)^{n/2}|\mathbf{B}^{-1}|^{1/2}} d\mathbf{Y} = 1$$

Hence

$$\int f_i^{1/2} f_j^{1/2} d\mathbf{Y} = \frac{\exp\left(\left(\frac{\hat{\mathbf{Y}}_i' \boldsymbol{\Sigma}_i^{-1} \hat{\mathbf{Y}}_i}{2a_i} + \frac{\hat{\mathbf{Y}}_j' \boldsymbol{\Sigma}_j^{-1} \hat{\mathbf{Y}}_j}{2a_j}\right) - \left(\frac{\boldsymbol{\Sigma}_i^{-1} \hat{\mathbf{Y}}_i}{2a_i} + \frac{\boldsymbol{\Sigma}_j^{-1} \hat{\mathbf{Y}}_j}{2a_j}\right)' \left(\frac{\boldsymbol{\Sigma}_i^{-1}}{2a_i} + \frac{\boldsymbol{\Sigma}_j^{-1}}{2a_j}\right)^{-1} \left(\frac{\boldsymbol{\Sigma}_i^{-1} \hat{\mathbf{Y}}_i}{2a_i} + \frac{\boldsymbol{\Sigma}_j^{-1} \hat{\mathbf{Y}}_j}{2a_j}\right)\right)}{\left|\left(\frac{\boldsymbol{\Sigma}_i^{-1}}{2a_i} + \frac{\boldsymbol{\Sigma}_j^{-1}}{2a_j}\right)\right|^{1/2} |a_i \boldsymbol{\Sigma}_i|^{1/4} |a_j \boldsymbol{\Sigma}_j|^{1/4}}$$

When all terms have a prior mean of zero, the numerator is equal to 1 and this expression reduces to the one found in appendix A of Bingham and Chipman.

Appendix B

Analysis of first stage results for the tribology example

B.1 Half-normal Plots

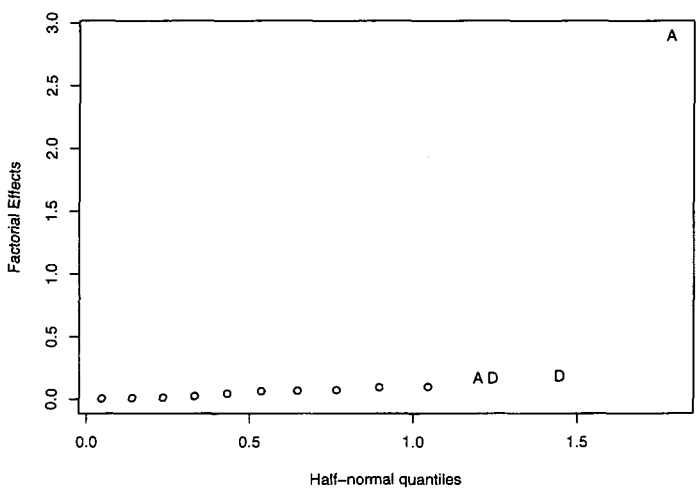

Figure B.1: Half Normal Plot for log(charge)

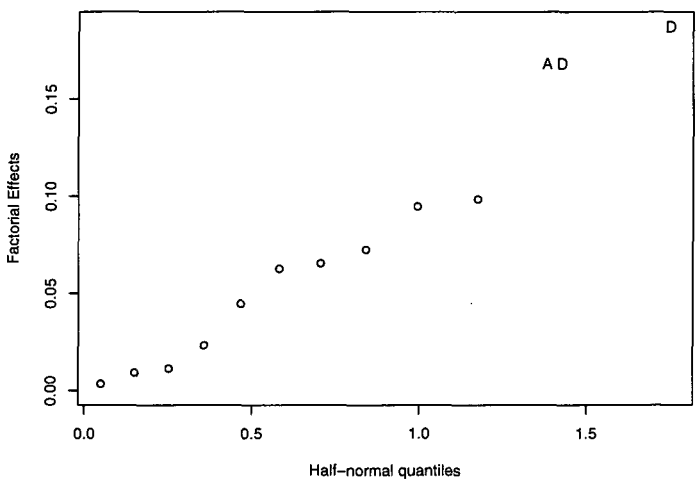

Figure B.2: Half Normal Plot for log(charge), without effect A.

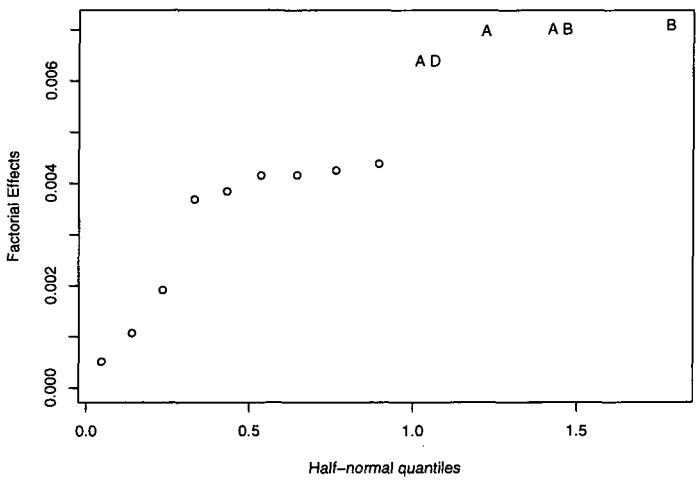

Figure B.3: Half Normal Plot for Coefficient of Friction.

B.2 Bayesian Analysis

Table B.1: Marginal or Model-averaged Effect Probabilities for Temperature for the initial experiment.

Factorial Term	Probability (All models)	Probability (In top 10 models)
I	1.00	1.000
A	0.22	0.162
B	0.12	0.052
C	0.48	0.444
D	0.15	0.107
E	0.13	0.056
F	0.14	0.095
AB	0.00	0.000
AC	0.02	0.000
AD	0.00	0.000
AE	0.00	0.000
AF	0.00	0.000
BC	0.01	0.000
BD	0.00	0.000
BE	0.00	0.000
BF	0.00	0.000
CD	0.01	0.000
CE	0.01	0.000
CF	0.01	0.000
DE	0.00	0.000
DF	0.00	0.000
EF	0.00	0.000

Table B.2: Top 10 most probable models for Temperature based on 20 observations

Probability	Model Terms
0.224	I C
0.175	I
0.071	I A
0.051	I A C
0.049	I D
0.044	I F
0.043	I E
0.039	I B
0.032	I C D
0.028	I C F

Table B.3: Marginal or Model-averaged Effect Probabilities for log(Charge) for the initial experiment.

Factorial Term	Probability (All models)	Probability (In top 10 models)
I	1.00	1.000
A	1.00	1.000
B	0.10	0.060
C	0.11	0.083
D	0.17	0.132
E	0.11	0.083
F	0.12	0.092
AB	0.02	0.000
AC	0.02	0.016
AD	0.04	0.033
AE	0.02	0.016
AF	0.02	0.019
BC	0.00	0.000
BD	0.00	0.000
BE	0.00	0.000
BF	0.00	0.000
CD	0.00	0.000
CE	0.00	0.000
CF	0.00	0.000
DE	0.00	0.000
DF	0.00	0.000
EF	0.00	0.000

Table B.4: Top 10 most probable models for log(Charge) based on 20 observations.

Probability	Model Terms
0.49	I A
0.088	I A D
0.066	I A F
0.06	I A E
0.059	I A C
0.054	I A B
0.03	I A D AD
0.017	I A F AF
0.015	I A E AE
0.014	I A C AC

Table B.5: Marginal or Model-averaged Effect Probabilities for Coefficient of Friction for the initial experiment.

Factorial term	Probability (All models)	Probability (In top 10 models)
I	1.00	1.000
A	0.22	0.150
B	0.30	0.265
C	0.14	0.072
D	0.17	0.110
E	0.18	0.116
F	0.16	0.083
AB	0.02	0.000
AC	0.01	0.000
AD	0.01	0.000
AE	0.01	0.000
AF	0.01	0.000
BC	0.01	0.000
BD	0.01	0.000
BE	0.01	0.000
BF	0.01	0.000
CD	0.00	0.000
CE	0.00	0.000
CF	0.00	0.000
DE	0.01	0.000
DF	0.00	0.000
EF	0.01	0.000

Table B.6: Top 10 most probable models for Coefficient of Friction based on 19 observations.

Probability	Model Terms
0.218	I
0.132	I B
0.086	I A
0.066	I E
0.062	I D
0.062	I F
0.054	I C
0.026	I A B
0.02	I B E
0.019	I B D

Appendix C

Candidate Points for the Reactor Experiment

Table C.1: Candidate points for the reactor experiment

Run	A	B	C	D	E
1	-1	-1	-1	-1	-1
2	1	-1	-1	-1	-1
3	-1	1	-1	-1	-1
4	1	1	-1	-1	-1
5	-1	-1	1	-1	-1
6	1	-1	1	-1	-1
7	-1	1	1	-1	-1
8	1	1	1	-1	-1
9	-1	-1	-1	1	-1
10	1	-1	-1	1	-1
11	-1	1	-1	1	-1
12	1	1	-1	1	-1
13	-1	-1	1	1	-1
14	1	-1	1	1	-1
15	-1	1	1	1	-1
16	1	1	1	1	-1
17	-1	-1	-1	-1	1
18	1	-1	-1	-1	1
19	-1	1	-1	-1	1
20	1	1	-1	-1	1
21	-1	-1	1	-1	1
22	1	-1	1	-1	1
23	-1	1	1	-1	1
24	1	1	1	-1	1
25	-1	-1	-1	1	1
26	1	-1	-1	1	1
27	-1	1	-1	1	1
28	1	1	-1	1	1
29	-1	-1	1	1	1
30	1	-1	1	1	1
31	-1	1	1	1	1
32	1	1	1	1	1

Bibliography

Agboto, V., Li, W. and Nachtsheim, C. (2006) Screening designs for model discrimination. *Submitted to Technometrics.*

Atkinson, A. and Federov, V. (1975a) The design of experiments for discriminating between two rival models. *Biometrika*, **62**, 57–70.

— (1975b) Optimal design: Experiments for discriminating between several models. *Biometrika*, 289–303.

Barrios, E. (2004) Bayesian screening and model selection. BsMD. URL `http://cran.r-project.org/src/contrib/Descriptions/BsMD.html`.

Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis.* Springer.

Bingham, D. and Chipman, H. (2007) Incorporating prior information in optimal designs for model selection. *Technometrics*, **49**, 155–163.

Box, G. and Cox, D. (1964) An analysis of transformations. *Journal of the Royal Statistical Society Series B*, 211–252.

Box, G. and Hill, W. (1967) Discrimination among mechanistic models. *Technometrics*, **9**, 57–68.

Box, G., Hunter, J. and Hunter, W. (1978) *Statistics for Experimenters, 1st Edition.* Wiley.

— (2005) *Statistics for Experimenters, 2nd Edition.* Wiley.

Box, G. and Meyer, R. (1986) Dispersion effects from fractional designs. *Technometrics*, **28**, 19–27.

Brooks, S. and Morgan, B. (1995) Optimization using simulated annealing. *The Statistician*, **44**, 241–257.

Chaloner, K. and Verdinelli, I. (1995) Bayesian experimental design: A review. *Technometrics*, **10**, 273–304.

Chipman, H. (1996) Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, **24**, 17–36.

Cook, R. and Nachtsheim, C. (1980) A comparison of algorithms for constructing exact D-optimal designs. *Technometrics*, **22**, 315–324.

Daniel, C. (1959) Use of half-normal plots in interpreting factorial two-level experiments. *Technometrics*, **1**, 311–341.

— (1962) Sequences of fractional replicates in the 2^{p-q} series. *Journal of the American Statistical Association*, 403–429.

DuMouchel, W. and Jones, B. (1994) A simple Bayesian modification of D-optimal designs to reduce dependence on an assumed model. *Technometrics*, **36**, 37–47.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Jeffreys, H. (1939) *Theory of Probability.* Oxford University Press.

John, P. W. M. (1971) *Statistical Design and Analysis of Experiments.* SIAM Classics in Applied Mathematics, Philadelphia.

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.

Jones, B. A., Li, W., Nachtsheim, C. J. and Ye, K. Q. (2007) Model discrimination-another perspective on model-robust designs. *Journal of Statistical Planning and Inference*, **137**, 1576–1583.

Kass, R. E. and Raftery, A. E. (1995) Bayes factors. *Journal of the American Statistical Association*, **90**, 773–795.

Li, W. (2005) Screening designs for model selection. In *Screening Methods for Experimentation in Industry, Drug Discovery and Genetics.* (eds. A. Dean and S. Lewis), 207–234. Springer Verlag, New York.

Lindley, D. (1972) *Bayesian Statistics - A Review.* SIAM, Philadelphia.

Madigan, D. and Raftery, A. (1994) Model selection and accounting for model uncertainty in graphical models using Occam's window. *Journal of the American Statistical Association*, **89**, 1535–1546.

Meyer, R., Steinberg, D. and Box, G. (1996) Follow-up designs to resolve confounding in multi-factor experiments. *Technometrics*, **38**, 303–313.

Meyer, R. and Wilkinson, R. (1998) Bayesian variable assessment. *Communications in Statistics - Theory and Methods*, **27**, 2675–2705.

Muller, P., Sanso, B. and De Iorio, M. (2004) Optimal Bayesian design by inhomogeneous Markov chain simulation. *Journal of the American Statistical Association*, **99**, 788–798.

Nelder, J. A. (1998) The selection of terms in response-surface models - how strong is the weak heredity principle ? *The American Statistician*, **52**, 315–318.

O'Hagan, A. and Forster, J. (2004) *Kendall's Advanced Theory of Statistics, volume 2B: Bayesian Inference, 2nd edition.* Arnold, London.

Plackett, R. and Burman, J. (1946) The design of optimum multifactor experiments. *Biometrika*, **33**, 305–325.

Press, W. H., Teukolsky, S., Vettering, W. T. and Flannery, B. (2002) *Numerical Recipes in C++.* Cambridge University Press, Cambridge.

R Development Core Team (2008) *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-project.org`. ISBN 3-900051-00-3.

Srivastava, J. (1975) *Designs for searching non-negligible effects. In: A Survey of Statistical Design and Linear Models.* North-Holland, New York.

Sun, D., Li, W. and Ye, Q. (2002) An algorithm for sequentially constructing non-isomorphic orthogonal designs and its applications. *Tech. Rep. SUNYSB-02-13*, State University of New York at Stony Brook.