Contributions to the Analysis of Experiments Using Empirical Bayes Techniques

A Thesis
Presented to
The Academic Faculty

by

James Dillon Delaney

In Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

H. Milton Stewart School of Industrial and Systems Engineering
Georgia Institute of Technology
August 2006

Contributions to the Analysis of Experiments Using Empirical Bayes Techniques

Approved by:

Dr. Roshan Joseph Vengazhiyil, advisor
H. Milton Stewart School of Industrial and
Systems Engineering
Georgia Institute of Technology

Dr. C. F. Jeff Wu
H. Milton Stewart School of Industrial and
Systems Engineering
Georgia Institute of Technology

Dr. Kwok-Leung Tsui
H. Milton Stewart School of Industrial and
Systems Engineering
Georgia Institute of Technology

Dr. Shreyes Melkote
George W. Woodruff School of Mechanical
Engineering
Georgia Institute of Technology

Dr. Ming Yuan
H. Milton Stewart School of Industrial and
Systems Engineering
Georgia Institute of Technology

Date Approved: 21 June 2006

To Beatrice.

ACKNOWLEDGEMENTS

The research presented here would not at all have been possible without the guidance, support, and patience of my advisor, Dr. Roshan Joseph Vengazhiyil. To Dr. Joseph, I am forever grateful. In addition I would like to thank all of the committee members for agreeing to serve on my committee. I am quite fortunate to have had the opportunity to present this work before and receive advice from such esteemed academicians as Professor C. F. Jeff Wu, Professor Kwok-Leung Tsui, Professor Shreyes Melkote, and Dr. Ming Yuan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

SUMMARY

This work is composed of two chapters. Both chapters contribute to the field of the analysis of physical experiments by addressing some practical limitations and offering alternatives to the existing methodology. The first chapter primarily addresses the issue of how to estimate the many factorial effects in highly fractionated designs. This is achieved through the application of *nearly objective* Bayes techniques. These techniques employ a functionally induced prior for the model parameters that have the highly desirable property of incorporating the concepts of effect hierarchy and effect heredity. The second chapter addresses a common "second step" in industrial settings, where often the entire purpose of the experiment is that of finding the optimal factor settings. Optimization experiments require the determination of settings for all of the factors so that a desired response can be achieved. With this as our primary objective, we make the case for an alternative to the standard practice: estimation followed by the use of statistical testing or the application of model selection algorithms, and finally the optimization of some reasonable parsimonious model. Instead, we propose the estimation techniques described in the first chapter in addition to a method of determining significance based on a criteria directly related to the problem at hand.

In the first chapter we focus on the estimation of a large number of effects from an experimental design with only a small number of runs. A full factorial experimental design over even a moderate number of multi-level factors may become infeasible to carry-out since the number of runs increases very rapidly with the number of factors. As a result, highly fractionated designs are employed in practice. However, while now the frequentist analysis may be carried out on this reduced run size, other problems

are introduced. For instance, we can only estimate a small subset of the factorial effects. The quantity of effects we can estimate is limited by the degrees of freedom available from this reduced run size. In addition, special techniques must be employed to resolve aliasing.

Bayes techniques have been suggested to address these issues. However, the common hierarchical model Bayesian approach to the design and analysis of experiments is typically encumbered by the daunting task of specifying a prior distribution for the large number of parameters in the linear model. Such a prior should also reflect a belief in the well known experimental design properties of effect hierarchy and effect heredity. Recently it has been proposed that we may specify a *functional prior* on the underlying transfer function. Through this functional prior, we are able to reduce the task of prior parameter specification to that of only a few hyper-parameters. When carefully selected, this functional prior may also incorporate the properties of effect hierarchy and effect heredity. Previously, this functionally induced prior was developed for two level experiments. Here we have extended these concepts for three and higher level designs. These designs play a very important role in industrial experiments.

The prior specification for multi-level factors requires that an interesting distinction be made between qualitative and quantitative factors. Such a distinction was not necessary in the case of 2-level factors. However, the Gaussian process functional prior assumption that we employ enables us to seamlessly integrate this aspect of multi-level factors in the modeling through the choice of an appropriate class of correlation functions. The application of the methodology is demonstrated with the analysis of two real world examples.

In the second chapter, we focus on what to do next, after estimation, in the case of an optimization experiment. Again, cost constraints may require that an experimental design's run size be kept small. In many such cases, not having enough data may be solely to blame for not being able to conclude an effect's significance

via a standard frequentist statistical test. This is particularly troublesome in an optimization experiment, where we wish to determine the optimal settings for all of the factors based on the experimental output. Another problem associated with frequentist hypothesis testing is that the choice of a significance level, α, tends to be completely arbitrary and has little connection to the real world problem.

A convenient property of the empirical Bayes estimates obtained in the first chapter is that they already incorporate information about uncertainty through the prior specification and the data. These estimators can be characterized as shrinkage estimates. In this chapter, some special known cases of the empirical Bayes estimator are discussed. For instance, connections are drawn to the so-called James-Stein estimator as well as the Beta Coefficient Method of Taguchi. Discussion of these special cases allow us to fully appreciate the functionally induced prior empirical Bayes estimator that is recommended here for the purpose of analyzing experiments.

After obtaining the empirical Bayes estimates, for an optimization experiment, it may not be desirable to perform additional statistical hypothesis testing or model selection. Instead, we may wish to use these estimates to determine factor settings which balance the goal of optimizing the response with the cost of changing factors from their current settings. Simulation results provide support for the conclusion that the recommended procedure is superior to frequentist estimation and hypothesis testing, with respect to a metric that should be of particular interest in optimization experiments. On average, the proposed techniques dictate factor settings that yield response values closer to our objective. Finally, we complete the analysis of a real world optimization experiment that is first visited in chapter one.

CHAPTER I

FUNCTIONALLY INDUCED PRIORS FOR THE DESIGN AND ANALYSIS OF EXPERIMENTS

1.1 INTRODUCTION

In the analysis of a typical experiment with any number of three and higher level factors, the amount and nature of the calculations in the traditional analysis do not facilitate the adoption of analysis strategies that can be easily automated. Historically, frequentist analysis strategies have had to rely upon tedious calculations to establish the aliasing relationships that comprise the design's degrees of freedom. Tools like half-normal plots and interaction plots would be used to perform variable selection and determine optimal factor settings, respectively, while computing was used to perform calculations necessary to estimate effects. While sometimes adequate, the traditional approach can be quite time consuming and does not lend itself well to utilization of the computing power that is now available.

Designs of three-level and four-level factors figure prominently in physical experiments. For example, all of the case studies reported in Taguchi, Chowdhury, and Taguchi (2000) use mixed two, three, and higher level designs. See Taguchi (1987) and Wu and Hamada (2000) for several other examples. Beginning with three-level factors, much more information about the shape of the response surface can be extracted from a good design. Unfortunately, the run size of full factorial 3^p and 4^p designs can be prohibitively large. Fractional factorial designs are used for reducing the run size, but they lead to aliasing of the effects. Several Bayesian approaches for estimating the effects from fractional designs have been suggested in the literature.

The technique used to incorporate prior information plays an important role in both optimal design choice and the subsequent estimation of effects and variable selection.

Some previous approaches to Bayesian methods to the design and analysis of experiments have focused on Bayesian hierarchical models that require eliciting or estimating many hyper-parameters in order to specify priors for a linear model's parameters. See for instance the review of the literature by Chaloner and Verdinelli (1995). Problems arise in both the proper specification of the numerous priors as well as potential complications with calculation. An interesting Bayesian approach, with model priors that facilitate the incorporation of principles like *effect heredity* is suggested by Chipman, Hamada, and Wu (1997). In that paper, estimation is through a Gibbs sampling procedure.

The specification of a prior for the model parameters is not a trivial matter. The sheer quantity of the parameters is a major problem, but there are several other issues. For example, consider a 3^2 design. Suppose u_1 and u_2 represent the two coded variables of the first factor and u_3 and u_4 those of the second factor. Then the linear model that we would like to fit is

$$Y = \beta_0 + \beta_1 u_1 + \beta_2 u_2 + \beta_3 u_3 + \beta_4 u_4 + \beta_5 u_1 u_3 + \beta_6 u_1 u_4 + \beta_7 u_2 u_3 + \beta_8 u_2 u_4 + \epsilon.$$

What should be the prior distribution for the β's? The usual approach is to take them as $\mathcal{N}(0, \tau_0^2)$, see for example Chipman et al. (1997). Although this choice looks reasonable, several questions remain unanswered. For example, by the effect hierarchy principle (see Hamada and Wu 1992), we know that a two-factor interaction (2fi) is less likely to be significant than a main effect. Therefore is it ideal to use the same distribution for a main effect and 2fi? Moreover, we can use different coding schemes to represent the two degrees of freedom for each factor. How should we change the prior specification depending on the coding scheme? Are the two effects of the same factor, say β_1 and β_2, equally important? It is known that if we use a linear-quadratic system, then the linear effect is more important than the quadratic effect. How do we

incorporate such a difference in the prior? Is it reasonable to take all the parameters to be independent? How should the prior be modified depending on the type of factor viz. qualitative or quantitative? It is imperative to develop a coherent and systematic approach to prior specification, so that we can answer all these questions.

In this chapter we propose the use of functionally induced priors for prior specification (Joseph 2006). Here a prior using a Gaussian process is postulated for the underlying transfer function and then the prior distribution for all the model parameters is induced from it. The work in Joseph (2006) focuses on two-level experiments. Mitchell, Morris, and Ylvisaker (1995) and Kerr (2001) have also studied the use of stochastic processes for the design of two-level experiments. Here we extend the approach for the case of three and higher level experiments. The extension is not trivial as there are many issues involved in higher level experiments that are not present in two-level experiments. For example, the type of factor, the type of correlation function, the type of coding scheme, the mixed-level nature of the experiments, etc. become important when dealing with higher level experiments, but are irrelevant for two-level experiments.

A very nice property of the functionally induced prior is that it agrees with many widely accepted principles in the design and analysis of experiments such as effect sparsity, effect hierarchy, and effect heredity (Wu and Hamada 2000). The introduction of these priors has provided for a very nice setting that enables the automation of many analytical tasks, that in previous approaches would have required a great deal of time consuming manual work.

The chapter is organized as follows. We begin by reviewing the general functionally induced prior Bayesian framework. We present a decomposition result that is extremely useful for studying three-level, four-level, $\cdots$, and mixed-level designs. The results are different for the case of qualitative and quantitative factors. In Section 1.3, we present the results for qualitative factors. The results of this section are very simple

and general, so that they can be used with any number of levels. In Section 1.4, the building blocks for applying the Bayesian methodology to three-level and four-level quantitative factors are presented. Here, there is a brief discussion of complications that can arise due to the choice of coding-scheme for the model matrix. We also demonstrate that a direct consequence of the functionally induced prior is a systematic methodology for ordering the effects. The utility of this Bayesian setting is illustrated through examples where the forward variable selection procedure is adapted to designs with three-level and four-level factors. This appears in Section 1.5. That this functionally induced prior has interesting implications for optimal design is demonstrated by an example in Section 1.6. Finally, concluding remarks and suggestions for future research are given in Section 1.7.

1.2   GENERAL METHODOLOGY

Suppose that there are p factors $\boldsymbol{x} = (x_1, x_2, \cdots, x_p)'$, where the factor x_i is experimented with at m_i levels. Assume the model

$$Y = f(\boldsymbol{x}) + e, \ \ e \sim \mathcal{N}(0, \sigma^2),$$

where e represents the random error in the response due to the uncontrollable variables in the system. The transfer function f could be nonlinear and highly complex, but we would like to approximate it by a linear model containing the main effects and interactions of the factors. The factor x_i can be represented by $m_i - 1$ coded variables and the interactions can be defined through the products of these coded variables. Thus we would like to approximate $f(\boldsymbol{x})$ by

$$f(\boldsymbol{x}) \approx \sum_{i=0}^{q-1} \beta_i u_i,$$

where $q = \prod_{i=1}^{p} m_i$. For example, in the 3^2 design discussed in Section 1, we let $u_5 = u_1 u_3$, $u_6 = u_1 u_4$, $u_7 = u_2 u_3$, and $u_8 = u_2 u_4$.

As the number of factors and/or the number of levels increase, the total number of parameters (q) can become very large. Therefore, postulating a prior distribution for $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_{q-1})'$ is a difficult task. Joseph (2006) used a simple idea to overcome this problem. The idea is to postulate a functional prior for the transfer function and use that to induce a prior for all of the parameters in the linear model. Therefore, let

$$f(\boldsymbol{x}) \sim GP(\mu_0, \sigma_0^2 \psi),$$

where μ_0 is the mean and $\sigma_0^2 \psi$ is the covariance function of the Gaussian process (GP). The covariance function is defined as $cov(Y(\boldsymbol{x}), Y(\boldsymbol{x} + \boldsymbol{h})) = \sigma_0^2 \psi(\boldsymbol{h})$. Because there are q parameters in the linear model, they can be chosen to exactly match the function values at q points. A simple choice for the q points is the full factorial design. Let $\boldsymbol{U}$ be the $q \times q$ model matrix for the parameter $\boldsymbol{\beta}$ and let $\boldsymbol{\Psi}$ be the corresponding correlation matrix. To simplify the results, consider instead $f(\boldsymbol{x}) = \mu_0 + \sum_{i=0}^{q-1} \beta_i u_i$ at the q points in the full factorial design. Then,

$$\boldsymbol{\beta} \sim \mathcal{N}\left(0, \sigma_0^2 \boldsymbol{U}^{-1} \boldsymbol{\Psi}(\boldsymbol{U}^{-1})'\right).$$

For obvious reasons, we call this a *functionally induced prior distribution*. For large q, the variance-covariance matrix is huge, which can be difficult to construct and handle. Therefore it is important to simplify the representation of the above matrix so that the results can be easily used in practice. We achieve this under some assumptions.

Assume that the correlation function ψ has a product correlation structure of the form:

$$\psi(\boldsymbol{h}) = \prod_{j=1}^{p} \psi_j(|h_j|). \tag{1}$$

Let $\boldsymbol{U}_j$ be the model matrix for factor x_j and let $\boldsymbol{\Psi}_j$ be the corresponding correlation matrix. For example, for a 3-level factor with possible levels 1, 2, and 3, the model

matrix using orthogonal polynomial coding (with common column lengths of $\sqrt{3}$) is

$$\boldsymbol{U}_j = \begin{pmatrix} 1 & -\sqrt{\frac{3}{2}} & \sqrt{\frac{1}{2}} \\ 1 & 0 & -\sqrt{2} \\ 1 & \sqrt{\frac{3}{2}} & \sqrt{\frac{1}{2}} \end{pmatrix} \tag{2}$$

and the correlation matrix is

$$\boldsymbol{\Psi}_j = \begin{pmatrix} 1 & \psi_j(1) & \psi_j(2) \\ \psi_j(1) & 1 & \psi_j(1) \\ \psi_j(2) & \psi_j(1) & 1 \end{pmatrix}. \tag{3}$$

Now we have the following result. All of the proofs are given in the Appendix A.

THEOREM 1 *Under the product correlation structure in (1):*

$$var(\boldsymbol{\beta}) = \sigma_0^2 \bigotimes_{j=1}^{p} \boldsymbol{U}_j^{-1} \boldsymbol{\Psi}_j (\boldsymbol{U}_j^{-1})'.$$

The impact of this theorem is that we can focus on each factor, one at a time, in choosing whatever coding scheme and correlation function suits our modeling needs. For example, to construct the variance-covariance matrix in a $2^{p_2} \times 3^{p_3} \times 4^{p_4}$ design, we only need to establish the structure of $\boldsymbol{U}_j^{-1} \boldsymbol{\Psi}_j (\boldsymbol{U}_j^{-1})'$ for a two-level, three-level, and four-level factor. The results can then be combined by taking Kronecker products to get the desired variance-covariance matrix for any values of p_2, p_3, and p_4.

In the following sections, we investigate the structure of the variance-covariance matrix, so that the result can be easily interpreted and applied in the design and analysis of experiments. The choice of correlation functions and coding schemes depend on the type of factors. Therefore we study the case of qualitative and quantitative factors separately.

1.3   QUALITATIVE FACTORS

By *qualitative factor* we mean a factor whose levels are nominal. That is, a qualitative factor might be the name of: the vendor for a part, a machine, a method, etc.

1.3.1 Correlation Function

For a qualitative factor, we should assign equal correlation between any two levels. This is because among our prior assumptions, there is no information as to how to either order these factor levels or to determine the relative distances between any of the levels. As mentioned previously, we assume that the prior Gaussian process is stationary. So for the jth factor, we need only be concerned with $h_j = |x_{ij} - x_{kj}|$, for two runs i and k. That is,

$$\psi_j(h_j) = \begin{cases} 1 & \text{if } h_j = 0 \\ \rho_j & \text{if } h_j \neq 0 \end{cases},$$

where $0 < \rho_j < 1$. Then the $m_j \times m_j$ correlation matrix has the *compound symmetry* form:

$$\Psi_j = \begin{pmatrix} 1 & \rho_j & \cdots & \rho_j \\ \rho_j & 1 & \cdots & \rho_j \\ \vdots & \cdots & \ddots & \vdots \\ \rho_j & \rho_j & \cdots & 1 \end{pmatrix}. \tag{4}$$

1.3.2 Prior Distribution

Suppose that for whatever coding schemes are selected for each of the single factor model matrices $\boldsymbol{U}_j$ for $j = 1, \ldots, p$, we impose only the restrictions that the first column of each $\boldsymbol{U}_j$ is $\boldsymbol{1}_{m_j}$ to correspond to the "y-intercept" effect, and that the remaining $m_j - 1$ columns of each $\boldsymbol{U}_j$ are a set of mutually orthogonal contrasts normalized to the length $\sqrt{m_j}$. Thus $\boldsymbol{U}'_j \boldsymbol{U}_j = m_j I_{m_j}$, where I_{m_j} is the identity matrix of dimension m_j. Then for Ψ_j as in (4),

$$\boldsymbol{U}'_j \Psi_j \boldsymbol{U}_j = m_j \begin{pmatrix} 1 + (m_j - 1)\rho_j & 0 & \cdots & 0 \\ 0 & 1 - \rho_j & \cdots & 0 \\ \vdots & \cdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 - \rho_j \end{pmatrix}. \tag{5}$$

7

Of course, since the columns of U_j are mutually orthogonal vectors, each with squared length m_j, we have:

$$\boldsymbol{U}_j^{-1}\boldsymbol{\Psi}_j(\boldsymbol{U}_j^{-1})' = \frac{1}{m_j^2}\boldsymbol{U}_j'\boldsymbol{\Psi}_j\boldsymbol{U}_j. \tag{6}$$

Now we can propose the following very general result for the joint prior distribution of the effects for a design incorporating some number of qualitative factors with any mixture of levels. This follows directly from Theorem 1, Equation (5) and Equation (6). Let

$$\tau_0^2 = \sigma_0^2 \prod_{j=1}^{p} \frac{1 + (m_j - 1)\rho_j}{m_j} \qquad \text{and} \qquad r_j = \frac{1 - \rho_j}{1 + (m_j - 1)\rho_j}.$$

Let $\delta_{ij} = 1$ if β_i includes the factor j and 0 otherwise.

PROPOSITION 1 *For factorial experiments on p qualitative factors, if we use an orthogonal coding for each factor and correlation matrix as in (4), then*

$$\beta_i \sim \mathcal{N}(0, \tau_0^2 \prod_{j=1}^{p} r_j^{\delta_{ij}}), \quad i = 0, 1, \cdots, \prod_{j=1}^{p} m_j - 1$$

and the effects are all mutually independent.

Note that because of the independence, the variance-covariance matrix is diagonal, which makes it very easy to construct. While the expression in the above proposition may seem a bit complicated, it is easy to summarize what is happening. The variance of any effect depends not on what the interpretation of the effect is, which contrasts are involved, but only on which factors are involved in that effect. For further clarity, consider the following example.

Example: Suppose we have two factors: A and B each experimented at three levels. Let a_1 and a_2 represent the two coded variables of factor A and b_1 and b_2 that of

factor B. With the correlation matrix for factor A (and similarly for factor B):

$$\boldsymbol{\Psi}_A = \begin{pmatrix} 1 & \rho_A & \rho_A \\ \rho_A & 1 & \rho_A \\ \rho_A & \rho_A & 1 \end{pmatrix},$$

we have $\tau_0^2 = \frac{\sigma_0^2}{9}(1+2\rho_A)(1+2\rho_B)$, $r_A = (1-\rho_A)/(1+2\rho_A)$, and $r_B = (1-\rho_B)/(1+2\rho_B)$. So that from Proposition 1: $\beta_0 \sim \mathcal{N}(0, \tau_0^2)$, β_{a_1} and $\beta_{a_2} \sim \mathcal{N}(0, \tau_0^2 r_A)$, β_{b_1} and $\beta_{b_2} \sim \mathcal{N}(0, \tau_0^2 r_B)$, and $\beta_{a_1 b_1}$, $\beta_{a_1 b_2}$, $\beta_{a_2 b_1}$, and $\beta_{a_2 b_2} \sim \mathcal{N}(0, \tau_0^2 r_A r_B)$.

Since each $0 < r_j < 1$ can be specified or estimated, both concepts of effect hierarchy and effect heredity are appropriately integrated into the prior. Generally, as the number of factors involved in an interaction increases, the *a priori* variance around the effect's mean, which is zero, decreases, justifying effect hierarchy. If a particular ρ_j is small, then the corresponding r_j is large, which would suggest a comparatively larger variance for effects that include that factor than those interactions of the same order that do not, justifying effect heredity.

There is a very simple case of Proposition 1 which arises when all of the correlation matrices for the factors are the same. When this occurs, the marginal prior of the effect depends on whether that effect is a "main effect" (me), "two-factor interaction" (2fi), ..., "p-factor interaction" (pfi):

COROLLARY 1 . *For $r_1 = r_2 = \ldots = r_p = r$,*

$$\begin{aligned} \beta_0 &\sim \mathcal{N}(0, \tau_0^2) \\ \beta_{me} &\sim \mathcal{N}(0, \tau_0^2 r) \\ \beta_{2fi} &\sim \mathcal{N}(0, \tau_0^2 r^2) \\ &\vdots \\ \beta_{pfi} &\sim \mathcal{N}(0, \tau_0^2 r^p), \end{aligned}$$

and the effects are all mutually independent.

1.3.3 Coding Schemes

The restrictions imposed on our model matrix to obtain the convenient result of Proposition 1 actually admit many reasonable choices for coding schemes. We would naturally find it desirable to estimate an overall mean effect, so the leading column of 1's is not really an imposing constraint. That the other columns need be orthogonal contrasts is also quite natural. We are still free to choose contrasts that have a sensible interpretation for the type of factor we are considering in this section, a qualitative factor. Below we discuss two such coding schemes that satisfy the assumptions of Proposition 1, but have been suggested in the frequentist design setting, indicating their value in interpretation.

For ease of implementation and interpretation, the orthogonal contrast coding scheme we recommend for a qualitative factor is Helmert coding (see Harville 1997). Other commonly used alternative coding schemes present problems. For instance, the effects from orthogonal polynomial coding do not have a natural interpretation for a qualitative factor. Although we do note that for two-level and three-level factors, Helmert coding and orthogonal polynomial coding are the same. Wu and Hamada (2000) offers some other alternatives. For example, for a three-level factor, the choice of using two of the following coding vectors: $\boldsymbol{D}_{01} = (-1, 1, 0)$, $\boldsymbol{D}_{02} = (-1, 0, 1)$, or $\boldsymbol{D}_{12} = (0, -1, 1)$ yield estimates for interpretable effects, however they are not mutually orthogonal. The problem more generally with treatment coding or zero sum coding is that the columns of $\boldsymbol{U}_j$ would not be mutually orthogonal. This would violate the assumptions that led to Proposition 1. Helmert coding, on the other hand, along with providing for the calculation of effects that may be interesting for the analysis of a qualitative factor, is quite easy to implement for any number of levels. In Helmert coding, the first effect is the difference between the second level and the first level. The second effect is the difference between the third level and the average of the first two, etc. Below is the model matrix that makes the interpretation

of effects more obvious:

$$\begin{pmatrix} 1 & -1 & -1 & -1 & \cdots & & -1 \\ 1 & 1 & -1 & -1 & \cdots & & -1 \\ 1 & 0 & 2 & -1 & \cdots & & -1 \\ 1 & 0 & 0 & 3 & \cdots & & -1 \\ \vdots & \vdots & \vdots & \vdots & \ddots & & \vdots \\ 1 & 0 & 0 & 0 & \cdots & (m_j - 1) \end{pmatrix}.$$

We need to "normalize" each column to have the same squared length, m_j. To accomplish this, the factors to multiply each column by are:

$$\left(1, \sqrt{\frac{m_j}{2}}, \sqrt{\frac{m_j}{6}}, \sqrt{\frac{m_j}{12}}, \ldots, \sqrt{\frac{1}{m_j - 1}} \right).$$

One should not feel restricted to using the above recommended coding scheme. Any set of mutually orthogonal contrasts will do. So if there is a set of such effects that is more interesting to the experimenter, they should be used. For a four-level qualitative factor, Wu and Hamada (2000) offers a convenient coding scheme. Their recommendation provides effects that can be interpreted as differences between pairs of levels. The model matrix is below:

$$\mathbf{U}_j = \begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & 1 & 1 & 1 \end{pmatrix}.$$

The above coding scheme can be extended to factors with 8, 12, 16, ... levels using Hadamard matrices. Such a coding scheme using only $\{-1, 1\}$ would not be naturally applicable to a three-level factor or a five-level factor. Whereas model matrices based on Helmert coding are easy to construct and provide interpretable effects for any number of factor levels.

11

1.4   QUANTITATIVE FACTORS

When a factor is continuous or discrete, but ordinal, where there exists some way to quantify the differences between the factor's level, we may treat it as a *quantitative* factor. For a quantitative factor, we may wish to run the experiment at evenly spaced levels, but this might not always be possible. Below we make recommendations for each situation.

1.4.1   Correlation Function

When the levels are equally spaced, the correlation matrix $\boldsymbol{\Psi}_j$ has the *symmetric Toeplitz* form given in (25), with $\psi_j(h_j) \to 0$ as $|h_j| \to \infty$. There are many parametric forms for $\psi_j(h_j)$. For example, generally we could make use of the two parameter exponential correlation function:

$$\psi_j(h_j) = \exp\left(-\theta_j |h_j|^{\alpha_j}\right) \qquad 0 < \alpha_j \leq 2 \qquad 0 < \theta_j < \infty. \tag{7}$$

This is the most popular correlation function used in computer experiments, but other correlation functions such as the Matérn correlation function, cubic correlation function, etc. could also be used (see Santner et al. 2003). A convenient special case of the exponential correlation function is when the parameter $\alpha_j = 2$. This case is useful for modeling functions that are infinitely differentiable. Let $\rho_j = \exp(-\theta_j)$. Then

$$\boldsymbol{\Psi}_j = \begin{pmatrix} 1 & \rho_j & \cdots & \rho_j^{(m_j-1)^2} \\ \rho_j & 1 & \cdots & \rho_j^{(m_j-2)^2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_j^{(m_j-1)^2} & \rho_j^{(m_j-2)^2} & \cdots & 1 \end{pmatrix}, \tag{8}$$

which will be used in most of the examples presented here.

We suggest that when the levels are not evenly spaced for factor j, that instead of using the values $x_j \in \{1, 2, \ldots, m_j\}$ for the arguments of the correlation function, that the end points: 1 and m_j be used for the first and last levels, and that the

other levels be represented by interpolating between these points. For example, if the unevenly spaced levels for a quantitative three-level factor are: $25, 30, 37$, then use the quantities $1, 11/6, 3$, instead of $1, 2, 3$ in the correlation function. One might anticipate that the factor levels would have to be grossly unevenly spaced for it to result in any noticeable changes in practice. However, this is one part of our suggested methodology where some caution should be exercised.

The most common coding scheme for quantitative factors is *orthogonal polynomial coding* (see Wu and Hamada 2000). Unfortunately, a general result like Proposition 1 does not exist for quantitative factors under this coding scheme. Therefore, we examine the most important cases of three-level and four-level designs in detail.

1.4.2 Prior Distribution for Three-Level Experiments

Let us consider an experiment with p quantitative, evenly spaced three-level factors. We assume a correlation matrix for each factor $j = 1, 2, \ldots, p$ of the form:

$$\boldsymbol{\Psi}_j = \begin{pmatrix} 1 & \psi_j(1) & \psi_j(2) \\ \psi_j(1) & 1 & \psi_j(1) \\ \psi_j(2) & \psi_j(1) & 1 \end{pmatrix}.$$

The model matrix using orthogonal polynomial coding (with common column lengths of $\sqrt{3}$) is:

$$\boldsymbol{U}_j = \begin{pmatrix} 1 & -\sqrt{\frac{3}{2}} & \sqrt{\frac{1}{2}} \\ 1 & 0 & -\sqrt{2} \\ 1 & \sqrt{\frac{3}{2}} & \sqrt{\frac{1}{2}} \end{pmatrix}.$$

So that by matrix multiplication, we have:

$$\boldsymbol{U}_j' \boldsymbol{\Psi}_j \boldsymbol{U}_j = \begin{pmatrix} 3 + 4\psi_j(1) + 2\psi_j(2) & 0 & -\sqrt{2}(\psi_j(1) - \psi_j(2)) \\ 0 & 3(1 - \psi_j(2)) & 0 \\ -\sqrt{2}(\psi_j(1) - \psi_j(2)) & 0 & 3 - 4\psi_j(1) + \psi_j(2) \end{pmatrix}. \quad (9)$$

Notice that the "quadratic" and "y-intercept" effects are going to be (negatively) correlated. This is an important difference from qualitative factors, where this matrix

13

was diagonal. We can now propose expressions for the model parameters' marginal prior distributions. For the following proposition, let:

$$\tau_0^2 = \frac{\sigma_0^2}{3^{2p}} \prod_{j=1}^{p} (3 + 4\psi_j(1) + 2\psi_j(2)),$$

$$r_{jl} = \frac{3 - 3\psi_j(2)}{3 + 4\psi_j(1) + 2\psi_j(2)}, \qquad r_{jq} = \frac{3 - 4\psi_j(1) + \psi_j(2)}{3 + 4\psi_j(1) + 2\psi_j(2)}.$$

Let $l_{ij} = 1$ if β_i includes the linear effect of factor j and 0 otherwise. Similarly, $q_{ij} = 1$ if β_i includes the quadratic effect of factor j and 0 otherwise. Then the following expressions follow directly from (9) and Theorem 1:

PROPOSITION 2 *For p quantitative three-level factors, with a model matrix coded according to orthogonal polynomial contrasts we have:*

$$\beta_i \sim \mathcal{N}\left(0, \tau_0^2 \prod_{j=1}^{p} r_{jl}^{l_{ij}} r_{jq}^{q_{ij}}\right), \quad i = 0, 1, \cdots, 3^p - 1.$$

Note that unlike the result in Proposition 1, the β_i's are not independent. We will consider some properties of a special case of this result. Suppose for each factor we assume a 3×3 correlation matrix in the form of (8), with $\rho_1 = \rho_2 = \ldots = \rho_p = \rho$. Let,

$$\tau_0^2 = \frac{\sigma_0^2}{3^{2p}} (3 + 4\rho + 2\rho^4)^p, \qquad r_l = \frac{3 - 3\rho^4}{3 + 4\rho + 2\rho^4}, \qquad r_q = \frac{3 - 4\rho + \rho^4}{3 + 4\rho + 2\rho^4}.$$

Then the following expressions illustrate a useful special case of Proposition 2:

COROLLARY 2 . *For p quantitative three-level factors, with a model matrix coded according to orthogonal polynomial contrasts, if we further assume $\rho_1 = \rho_2 = \cdots = \rho_p = \rho$ in (8), then:*

$$\beta_0 \sim \mathcal{N}\left(0, \tau_0^2\right)$$
$$\beta_l \sim \mathcal{N}\left(0, \tau_0^2 r_l\right)$$
$$\beta_q \sim \mathcal{N}\left(0, \tau_0^2 r_q\right)$$

14

$$\beta_{ll} \sim \mathcal{N}\left(0, \tau_0^2 r_l^2\right)$$

$$\beta_{lq} \sim \mathcal{N}\left(0, \tau_0^2 r_l r_q\right)$$

$$\vdots$$

$$\beta_{q\cdots q} \sim \mathcal{N}\left(0, \tau_0^2 r_q^p\right).$$

To clarify the above notation, the subscript lq, for example, indicates that β_{lq} is the interaction effect between the linear effect of any one of the p factors and the quadratic effect of any of the other $p - 1$ remaining factors.

We have for all $\rho \in (0, 1)$

$$0 < r_l^3 < r_q < r_l^2 < r_l < 1. \tag{10}$$

It is quite common to say that a quadratic effect is less important than a linear effect (notationally $q \prec l$). Because $r_q < r_l$, we now have a mathematical justification of the above statement. Similarly, $r_q < r_l^2$ shows that $q \prec ll$, which is an interesting result. The property (10) can be used to order many higher order effects. For example, $qq \prec llq \prec lq \prec ll$.

The ordering of effects is important for properly defining a design criterion similar to minimum aberration. Cheng and Ye (2005) proposes two rules:

$(a) \quad : \quad l \succ q \succ ll \succ lq \succ lll \succ qq \succ llq \succ llll \succ lqq \succ lllq \succ qqq \succ llqq \succ lqqq \succ qqqq,$

$(b) \quad : \quad l \succ q \succ ll \succ lq \succ qq \succ lll \succ llq \succ lqq \succ qqq \succ llll \succ lllq \succ llqq \succ lqqq \succ qqqq.$

Rule (a) is obtained by ordering effects first by the degree of the polynomial, and then within that by the number of factors involved in the interaction, whereas rule (b) is by ordering effects first by the number of terms in the interaction, and then by the degree of that polynomial. Alternatively, the rule implied by (10) is:

$$l \succ ll \succ q \succ lll \succ lq \succ llll \succ llq \succ qq \succ lllq \succ lqq \succ llqq \succ qqq \succ lqqq \succ qqqq.$$

As the number of factors increases, the ordering rule can get more complicated. For example, when there are five three-level factors, and $\alpha = 2$, we would need a result like (10) to include a comparison of the two terms r_l^5 and r_q^2. Numerically, it can be shown that: for $\rho \in (0, 0.357)$ or $(0.847, 1)$,

$$0 < r_l^5 < r_q^2 < r_l^3 < r_q < r_l^2 < r_l < 1,$$

whereas for $\rho \in (0.357, 0.847)$,

$$0 < r_q^2 < r_l^5 < r_l^3 < r_q < r_l^2 < r_l < 1.$$

This should be enough to order the orthogonal polynomial factorial effects for five three-level factors when the value of ρ is known. When $\alpha = 1$ the ordering of effects additionally depends on ρ with as few as four factors. The nice thing about the Bayesian approach is that we do not need to worry about these complicated ordering of effects, it will be automatically built-in in the design and analysis of experiments.

1.4.3 Prior Distribution for Four-Level Experiments

Let us now consider an experiment with p quantitative, evenly spaced four-level factors. We assume a correlation matrix for each evenly spaced factor $j = 1, 2, \ldots, p$ of the form:

$$\boldsymbol{\Psi}_j = \begin{pmatrix} 1 & \psi_j(1) & \psi_j(2) & \psi_j(3) \\ \psi_j(1) & 1 & \psi_j(1) & \psi_j(2) \\ \psi_j(2) & \psi_j(1) & 1 & \psi_j(1) \\ \psi_j(3) & \psi_j(2) & \psi_j(1) & 1 \end{pmatrix}. \tag{11}$$

As we did before we may attempt to use orthogonal polynomial coding, albeit with some reservations, in anticipation that some of the off diagonal terms in the prior parameter covariance matrix will be nonzero. The "normalized" model matrix for

the jth four-level factor using orthogonal polynomial coding is:

$$
U_j = \begin{pmatrix}
1 & -\frac{3}{\sqrt{5}} & 1 & -\frac{1}{\sqrt{5}} \\
1 & -\frac{1}{\sqrt{5}} & -1 & \frac{3}{\sqrt{5}} \\
1 & \frac{1}{\sqrt{5}} & -1 & -\frac{3}{\sqrt{5}} \\
1 & \frac{3}{\sqrt{5}} & 1 & \frac{1}{\sqrt{5}}
\end{pmatrix}.
$$

Now this model matrix would enable us to calculate "y-intercept", "linear", "quadratic", and "cubic" effects. However, as suggested by the matrix calculation below, the prior covariance matrix is in fact not diagonal. There are some nonzero covariances between the "y-intercept" and the "quadratic" effect as well as between the "linear" and "cubic" effects. For the following equation, let us denote: $\psi_j(1) = \psi_{j1}$, $\psi_j(2) = \psi_{j2}$, and $\psi_j(3) = \psi_{j3}$. Then we obtain:

$$
U_j' \Psi_j U_j =
$$

$$
\begin{pmatrix}
4 + 6\psi_{j1} + 4\psi_{j2} + 2\psi_{j3} & 0 & -2(\psi_{j1} - \psi_{j3}) & 0 \\
0 & 4 + 2\psi_{j1} - \frac{12}{5}\psi_{j2} - \frac{18}{5}\psi_{j3} & 0 & -2\psi_{j1} + \frac{16}{5}\psi_{j2} - \frac{6}{5}\psi_{j3} \\
-2(\psi_{j1} - \psi_{j3}) & 0 & 4 - 2\psi_{j1} - 4\psi_{j2} + 2\psi_{j3} & 0 \\
0 & -2\psi_{j1} + \frac{16}{5}\psi_{j2} - \frac{6}{5}\psi_{j3} & 0 & 4 - 6\psi_{j1} + \frac{12}{5}\psi_{j2} - \frac{2}{5}\psi_{j3}
\end{pmatrix}.
$$

$$(12)$$

Now using the above result and Theorem 1, we can obtain a result similar to Proposition 2. For notational simplicity, we will only provide a special case where the correlation matrix is as in (8) with $\rho_1 = \rho_2 = \ldots = \rho_p = \rho$. Let

$$
\tau_0^2 = \frac{\sigma_0^2}{4^{2p}}(4 + 6\rho + 4\rho^4 + 2\rho^9)^p, \qquad r_l = \frac{4 + 2\rho - \frac{12}{5}\rho^4 - \frac{18}{5}\rho^9}{4 + 6\rho + 4\rho^4 + 2\rho^9},
$$

$$
r_q = \frac{4 - 2\rho - 4\rho^4 + 2\rho^9}{4 + 6\rho + 4\rho^4 + 2\rho^9}, \qquad r_c = \frac{4 - 6\rho + \frac{12}{5}\rho^4 - \frac{2}{5}\rho^9}{4 + 6\rho + 4\rho^4 + 2\rho^9}.
$$

PROPOSITION 3 *For p quantitative four-level factors, with a model matrix coded according to orthogonal polynomial contrasts, if we further assume $\rho_1 = \rho_2 = \cdots = \rho_p = \rho$ in (8), then:*

$$
\beta_i \sim \mathcal{N}\left(0, \tau_0^2 r_l^{l_i} r_q^{q_i} r_c^{c_i}\right),
$$

17

where l_i, q_i, and c_i are the number of linear, quadratic, and cubic terms in β_i.

Thus, using the notations used in Corollary 2, $\beta_c \sim \mathcal{N}(0, \tau_0^2 r_c), \beta_{lc} \sim \mathcal{N}(0, \tau_0^2 r_l r_c)$, etc. Note that as in Corollary 2, these effects are not independent.

At this point, it should be abundantly clear that it is a trivial matter to construct Proposition 1, Proposition 2, and Proposition 3-like results for designs of any mixture of factors with possibly different numbers of levels, possibly different types: qualitative or quantitative, and different parametric forms for the correlation functions. The user need only construct a model matrix U_j and correlation matrix Ψ_j appropriate for each factor and then deduce the full factorial prior covariance results using Theorem 1.

1.4.4 Coding Schemes

In this section on quantitative factors, we presented results for three-level and four-level factors assuming that orthogonal polynomial coding was the most desirable coding scheme. This coding scheme does indeed have some nice properties. The model matrix is easy to generate for a factor of any number of levels. The effects generated from such a model matrix are also easy to interpret for a quantitative factor.

One drawback to the orthogonal polynomial coding scheme for a design containing a quantitative factor is that the resulting prior covariance matrix is not diagonal. In fact constructing the matrix $\boldsymbol{R} = \tau_0^{-2} \, var(\boldsymbol{\beta})$, which will be used in the estimation, is not a trivial matter. If the matrix $\boldsymbol{R}$ is not calculated directly, which in itself could be prohibitively computationally intensive, it is quite a difficult matter of accounting to calculate and position these off-diagonal elements correctly in the matrix. In addition, the matrix $\boldsymbol{R}$ represented in its full form, may be quite large, requiring sparse matrix techniques.

So suppose instead that our motivation was to find an orthogonal coding scheme for a quantitative factor's model matrix $\boldsymbol{U}_j$ that produces a diagonal $\boldsymbol{U}_j^{-1} \boldsymbol{\Psi}_j (\boldsymbol{U}_j^{-1})'$.

Let $\mathbf{\Lambda}_j = diag(\lambda_{j,1}, \lambda_{j,2}, \ldots, \lambda_{j,m_j})$, with each $\lambda_{j,k}$, $k = 1, \ldots, m_j$ being the eigenvalues of $\mathbf{\Psi}_j$ and $\boldsymbol{E}_j$ is a $m_j \times m_j$ matrix whose columns are orthonormal eigenvectors corresponding to those eigenvalues. Now by letting $\boldsymbol{U}_j = \sqrt{m_j}\boldsymbol{E}_j$, we obtain

$$\boldsymbol{U}_j^{-1}\mathbf{\Psi}_j(\boldsymbol{U}_j^{-1})' = \frac{1}{m_j}\mathbf{\Lambda}_j, \tag{13}$$

which is a diagonal matrix. Now by Theorem 1, the variance-covariance matrix is also diagonal. Therefore, the matrix R could be easily constructed. A related idea exists in Steinberg and Bursztyn (2004), which contains a procedure for data analysis that involves relating regression coefficients to those produced from the eigenvectors of the correlation matrix of the corresponding random field model. However, using a model matrix whose columns are proportional to the eigenvectors of $\mathbf{\Psi}_j$ presents its own problems in the context here. The coding scheme will vary with the correlation matrix. That is, the model matrix $\boldsymbol{U}_j$ will depend on $\psi_j(1), \psi_j(2), \ldots, \psi_j(m_j - 1)$. Also, that the leading column of this coding scheme will not precisely be a column of ones, failing to yield a true "y-intercept" effect, complicates construction of the full design model matrix with Kronecker products less predictable. So rather than suggesting the use of this "eigen-Coding" scheme here, we merely use the observation that orthogonal polynomial coding is very nearly the coding scheme obtained from the eigenvectors of $\mathbf{\Psi}_j$'s as evidence in support of the belief that dismissing the off-diagonal elements of the matrix $\boldsymbol{R}$ may be acceptable in practice.

In Figure 1 we see a comparison of the orthogonal polynomial coding scheme and the eigen-coding scheme. From the figure we see the curvature in the y-intercept and linear effects demonstrating their dependence on the quadratic and cubic effects respectively. In addition, the plots demonstrate the relative "closeness" of the orthogonal polynomial effects and each of their corresponding eigen-coding effect. For a single factor it is also easy to verify numerically that the correlation between a polynomial effect and its corresponding eigen-coding effect is very high. For reference, through numerical studies it can be shown that a single evenly-spaced four-level
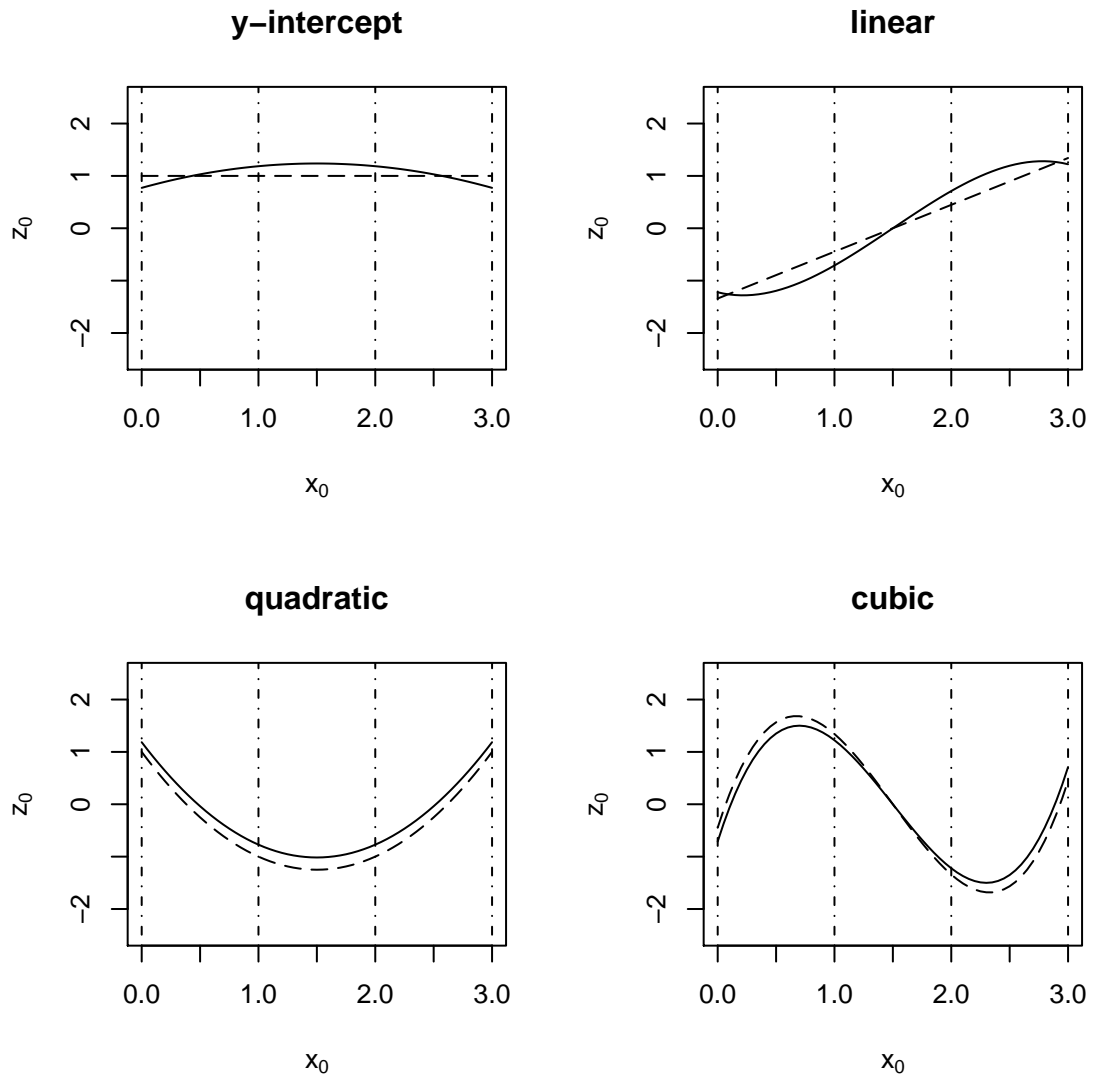
Figure 1: Comparison of Eigen-Coding (solid) and Orthogonal Polynomial Coding (dashed) for $\rho = 0.5$

factor with a correlation matrix like (8), the correlation between a polynomial effect and its corresponding eigen-coded effect is greater than 0.97 for all $\rho \in (0, 1)$.

1.5 EXAMPLES

We need the following notation. Let $\boldsymbol{D}$ be the design matrix, which has n rows and p columns corresponding to the p factors and $\boldsymbol{y} = (y_1, \cdots, y_n)'$ be the response values obtained from the experiment. Let $\boldsymbol{U}_D$ be the model matrix generated from D and $\boldsymbol{\Psi}_D$ the corresponding correlation matrix. Let $var(\boldsymbol{\beta}) = \tau_0^2 \boldsymbol{R}$, where the construction of the matrix $\boldsymbol{R}$ was discussed in detail in the previous sections. The examples presented in this section do not have replicates. Because we do not have any information about σ^2, we set $\sigma^2 = 0$. We obtain

$$\widehat{\boldsymbol{\beta}} = \mathbb{E}(\boldsymbol{\beta}|\boldsymbol{y}) = \frac{\tau_0^2}{\sigma_0^2} \boldsymbol{R} \boldsymbol{U}'_D \boldsymbol{\Psi}_D^{-1} (\boldsymbol{y} - \mu_0 \mathbf{1}_n)$$

and

$$var(\boldsymbol{\beta}|\boldsymbol{y}) = \tau_0^2 \left(\boldsymbol{R} - \frac{\tau_0^2}{\sigma_0^2} \boldsymbol{R} \boldsymbol{U}'_D \boldsymbol{\Psi}_D^{-1} \boldsymbol{U}_D \boldsymbol{R} \right).$$

A general expression for τ_0^2/σ_0^2 is given by

$$\frac{\tau_0^2}{\sigma_0^2} = \frac{\prod_{j=1}^p sum(\boldsymbol{\Psi}_j)}{q^2}, \tag{14}$$

where $q = \prod_{i=1}^p m_i$ and $sum(\boldsymbol{\Psi}_j)$ denotes the sum of all the elements of the matrix $\boldsymbol{\Psi}_j$. We can calculate the ratios

$$t_i = \frac{|\widehat{\beta}_i|}{\widehat{\sigma}_{\beta_i}},$$

to identify the important effects, where $\widehat{\sigma}_{\beta_i}^2$ is the diagonal element in $var(\boldsymbol{\beta}|\boldsymbol{y})$ corresponding to β_i. The most important effect is the one with the largest t_i. The other important effects can be similarly identified one-by-one using a forward selection strategy as explained in Joseph (2006).

The hyper-parameters can be estimated using empirical Bayes methods. Let $\boldsymbol{\rho} = (\rho_1, \cdots, \rho_p)'$. Then

$$\widehat{\boldsymbol{\rho}} = \arg \min_{\boldsymbol{\rho}} n \log \widehat{\sigma}_0^2 + \log \det(\boldsymbol{\Psi}_D),$$

21

$$\widehat{\mu}_0 = (\mathbf{1}_n' \boldsymbol{\Psi}_D^{-1} \mathbf{1}_n)^{-1} \mathbf{1}_n' \boldsymbol{\Psi}_D^{-1} \boldsymbol{y},$$

and

$$\widehat{\sigma}_0^2 = \frac{1}{n} (\boldsymbol{y} - \widehat{\mu}_0 \mathbf{1}_n)' \boldsymbol{\Psi}_D^{-1} (\boldsymbol{y} - \widehat{\mu}_0 \mathbf{1}_n).$$

For numerical stability, we must put some mild constraints on the feasible region of $\boldsymbol{\rho}$ in the above optimization, such as $\rho_i \in [0, 0.99]$. We could have instead implemented the penalized likelihood recommendations from Li and Sudjianto (2005). There are some additional considerations in the empirical Bayes step for estimating $\boldsymbol{\rho}$. It is important to obtain the constrained *global* optimum. Most software will converge on some local optima. We employ a naive approach to global optimization and implement a sequence of local optimizations over randomly generated initial values, choosing the best local optimum as the global optimum. We caution that it is possible to begin the algorithm with a value for $\boldsymbol{\rho}$ that is not a true global optimum due to either the precautions taken to prevent inverting an ill-conditioned $\boldsymbol{\Psi}$ matrix, or by not being able to pragmatically do an exhaustive search of the feasible region for all of the local optima. The $\boldsymbol{\rho}$ is estimated only at step 0 of the forward selection procedure. We use this estimate for each subsequent step. From this estimate of $\boldsymbol{\rho}$, we are able to calculate the factor τ_0^2/σ_0^2, as well as the matrices $\boldsymbol{R}$ and $\boldsymbol{\Psi}_D$ used in the calculations at all later steps. The first example illustrates a situation where the matrix $\boldsymbol{R}$ has nonzero off-diagonal elements. This matrix can be constructed through the explicit matrix calculations suggested by Theorem 1. However, we found that the diagonal approximation to $\boldsymbol{R}$ in this example is adequate for discovering the first few important effects. In the second example, $\boldsymbol{R}$ is a diagonal matrix. So for this example constructing $\boldsymbol{R}$ is a simple matter, where the diagonal elements of $\boldsymbol{R}$ correspond to the appropriate factor calculation preceding each of the propositions and entered into the matrix $\boldsymbol{R}$ in the order the effects appear as columns of $\boldsymbol{U}_D$.

Here we emphasize the ease with which the methodology of Joseph (2006) is extended beyond two-level experiments. In addition, we stress that very often the

procedure is entirely automatic, yielding no ambiguity in situations where the traditional frequentist approach would require deeper attention. Finally, we note that there may exist situations where the iterative procedure is unnecessary. That is, a quick proxy to the forward selection may be to use a half-normal plot to select effects in Step 0. This technique seems to produce results equivalent to the forward selection when the significant effects after k steps form a projection of the factor space onto a lower dimensional, but orthogonal factor space.

1.5.1 Blood Glucose Experiment

Hamada and Wu (1992) analyzed an experiment designed to study blood glucose reading levels from a testing device. In this experiment, there was one two-level factor and seven three-level factors (Table 1). The three-level factors were all considered to be quantitative factors. These factors did not all have evenly spaced levels, but they were approximately evenly spaced. The design was a nonregular fraction of a $2^1 \times 3^7$ design, the 18-run design popularized by Taguchi (1987). The design and the data are given in Table 2.

Table 1: Factors and Levels, the Blood Glucose Experiment

	Factor	Level 1	Level 2	Level 3
A.	wash	no	yes	
B.	microvial volume (ml)	2.0	2.5	3.0
C.	caras H_2O level (ml)	20	28	35
D.	centrifuge RPM	2100	2300	2500
E.	centrifuge time (min)	1.75	3	4.5
F.	(sensitivity, absorption)	(0.10,2.5)	(0.25,2)	(0.50,1.5)
G.	temperature (^{0}C)	25	30	37
H.	dilution ratio	1:51	1:101	1:151

In the frequentist analysis, it is computationally cumbersome to entertain all of the 4,374 possible factorial effects. Therefore, we consider only the main effects and two-factor interactions. This analysis identifies the effects $B_l H_q$, $B_q H_q$, $E_l G_l$, AH_q,

Table 2: Design Matrix and Data, the Blood Glucose Experiment

Run	A	G	B	C	D	E	F	H	Mean Reading
1	1	1	1	1	1	1	1	1	97.94
2	1	1	2	2	2	2	2	2	83.40
3	1	1	3	3	3	3	3	3	95.88
4	1	2	1	1	2	2	3	3	88.86
5	1	2	2	2	3	3	1	1	106.58
6	1	2	3	3	1	1	2	2	89.57
7	1	3	1	2	1	3	2	3	91.98
8	1	3	2	3	2	1	3	1	98.41
9	1	3	3	1	3	2	1	2	87.56
10	2	1	1	3	3	2	2	1	88.11
11	2	1	2	1	1	3	3	2	83.81
12	2	1	3	2	2	1	1	3	98.27
13	2	2	1	2	3	1	3	2	115.52
14	2	2	2	3	1	2	1	3	94.89
15	2	2	3	1	2	3	2	1	94.70
16	2	3	1	3	2	3	1	2	121.62
17	2	3	2	1	3	1	2	3	93.86
18	2	3	3	2	1	2	3	1	96.10

$\cdots$, as having high explanatory power, which is shown in Figure 2(b). Thus, the frequentist approach does not lead to a model satisfying effect hierarchy or effect heredity.

By contrast, the proposed methodology does respect effect hierarchy and effect heredity, and is able to entertain all of the factorial effects. In step 0 of the Bayesian forward selection, the empirical Bayes estimate of the correlation matrix parameters is given by the vector,

$$\hat{\boldsymbol{\rho}} = (0.93, 0.00, 0.99, 0.99, 0.98, 0.98, 0.99, 0.00)'.$$

Figure 3(a) shows the half-normal plot of the t_i ratios at this step. We can see that $B_l H_q$ is the most significant effect. After selecting this effect and continuing with the forward selection, we identify the effects $B_q H_q$, B_l, B_q, $\cdots$ as having high explanatory power. This is shown in the R^2-plot in Figure 3(b).

Of course, the principles of effect hierarchy and effect heredity can be enforced in the frequentist forward selection through some modifications, such as the strategy presented in Hamada and Wu (1992). However, we believe that the Bayesian strategy is more elegant and efficient. For example, if a three-factor interaction effect is significant, the frequentist analysis will miss it, but the Bayesian analysis will identify it with high probability. Indeed, the Bayesian analysis seems to be more powerful than the frequentist analysis, as can be seen by comparing the half-normal plots of both of the analyses at step 0; that is, all of the significant effects can be identified even at step 0 of the Bayesian analysis. In the next section, we provide an example where the frequentist analysis fails, but the Bayesian analysis succeeds.

Implementation of the Bayesian methodology of Chipman et al. (1997) was also illustrated through this example. One of the most significant differences in the Bayesian methodology presented here versus that of Chipman et al. (1997) is how the prior belief in effect heredity is incorporated into the model. In the procedure presented above, specification of effect heredity is through the parameter space as a consequence of our functional prior assumption. In Chipman et al. (1997), effect heredity is reflected through prior specification in the model space. In the methodology presented here, effect heredity was a direct consequence of the functionally induced prior on $\boldsymbol{\beta}$, whereas in Chipman et al. (1997) hierarchical priors on all subset models had to be specified in order to incorporate prior beliefs about heredity. The technique of Chipman et al. (1997) does offer the advantage of great flexibility in enabling the incorporation of other possible *a priori* beliefs about relationships between effects through adding on to the hierarchical prior structure. However, the procedure described in this paper is a fairly automatic methodology that quite naturally imposes effect hierarchy and effect heredity. Moreover, the extension of the prior specification to include three and higher order interactions, cubic, fourth order terms, etc. is more difficult to implement with the hierarchical priors compared with the functionally

(a) Frequentist Forward Selection (Step 0)
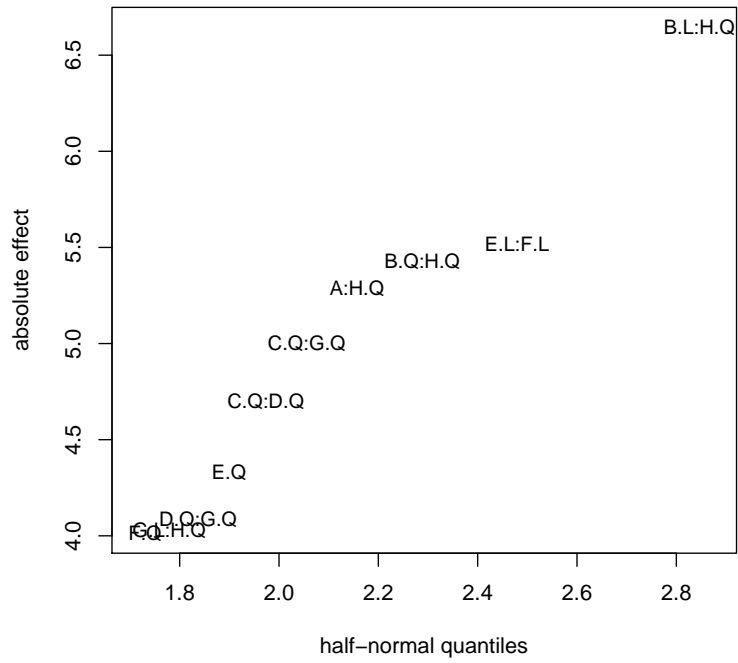
(b) Frequentist Forward Selection
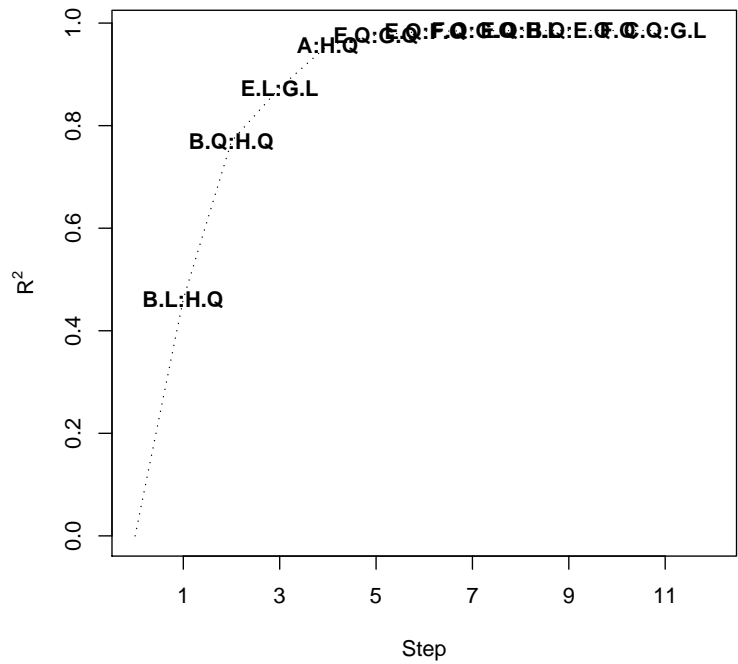
Figure 2: Frequentist Analysis of the Blood Glucose Experiment

(a) Bayesian Forward Selection (Step 0)

(b) Bayesian Forward Selection
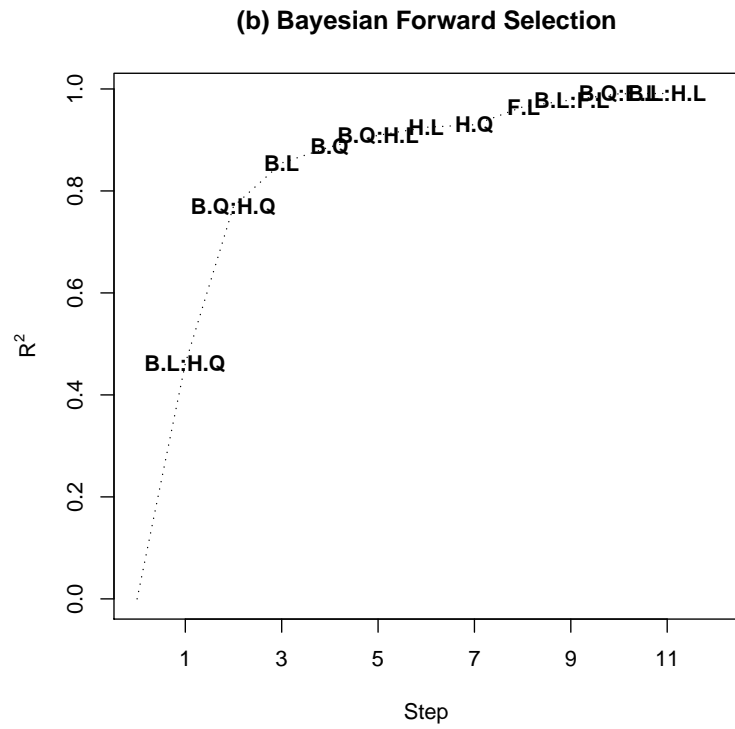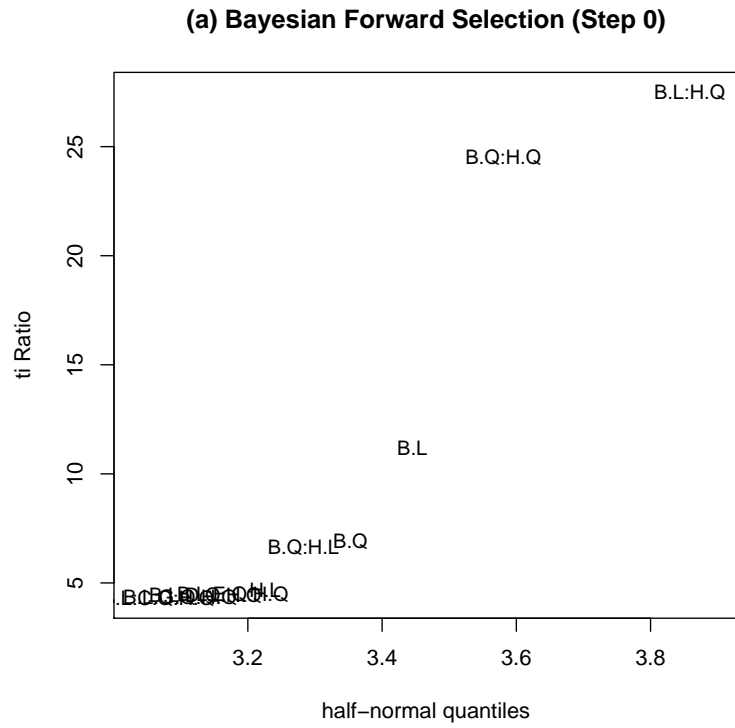
Figure 3: Bayesian Analysis of the Blood Glucose Experiment

induced priors.

1.5.2 Router Bit Experiment

Phadke (1989) reported on an experiment designed to help improve the lifetime of a router bit used to cut printed circuit boards. This experiment was also analyzed by Wu and Hamada (2000). The experiment is an unreplicated fraction of a $2^7 \times 4^2$ design. The factors and levels are shown in Table 3 and the design and data are given in Table 4. There are only 32 runs and 2,048 possible effects to consider. The two four-level factors: (D) "bit type" and (E) "spindle position" are treated as qualitative factors. So in analyzing this experiment, we have two types of factors: seven two-level factors and two qualitative four-level factors.

Table 3: Factors and Levels, the Router Bit Experiment

Factor		Level			
A.	suction (in of Hg)	1	2		
B.	x-y feed (in/min)	60	80		
C.	in-feed (in/min)	10	50		
D.	bit type	1	2	3	4
E.	spindle position	1	2	3	4
F.	suction foot	SR	BB		
G.	stacking height (in)	3/16	1/4		
H.	Slot depth (mils)	60	100		
J.	speed (rpm)	30000	40000		

The coding scheme that we used for the four-level factors is the Wu-Hamada recommendation highlighted in a previous section. Here, those main effects are labeled D_1, D_2, D_3 and E_1, E_2, E_3. Figure 4(a) shows the half-normal plot from a traditional analysis. The effects D_2, G, J, GJ and AF appear to be significant. Note that each of them represents a set of aliased effects. Assuming three and higher order interactions are negligible, one can show that

$$AF = -D_2H = -CE_2 = BD_3 = D_1E_3 = E_1G \qquad (15)$$

Table 4: Design Matrix and Data, the Router Bit Experiment

Run	A	B	C	D	E	F	G	H	J	Lifetime
1	−	−	−	1	1	−	−	−	−	3.5
2	−	−	−	2	2	+	+	−	−	0.5
3	−	−	−	3	4	−	+	+	−	0.5
4	−	−	−	4	3	+	−	+	−	17.0
5	−	+	+	3	1	+	+	−	−	0.5
6	−	+	+	4	2	−	−	−	−	2.5
7	−	+	+	1	4	+	−	+	−	0.5
8	−	+	+	2	3	−	+	+	−	0.5
9	+	−	+	4	1	−	+	+	−	17.0
10	+	−	+	3	2	+	−	+	−	2.5
11	+	−	+	2	4	−	−	−	−	0.5
12	+	−	+	1	3	+	+	−	−	3.5
13	+	+	−	2	1	+	−	+	−	0.5
14	+	+	−	1	2	−	+	+	−	2.5
15	+	+	−	4	4	+	+	−	−	0.5
16	+	+	−	3	3	−	−	−	−	3.5
17	−	−	−	1	1	−	−	−	+	17.0
18	−	−	−	2	2	+	+	−	+	0.5
19	−	−	−	3	4	−	+	+	+	0.5
20	−	−	−	4	3	+	−	+	+	17.0
21	−	+	+	3	1	+	+	−	+	0.5
22	−	+	+	4	2	−	−	−	+	17.0
23	−	+	+	1	4	+	−	+	+	14.5
24	−	+	+	2	3	−	+	+	+	0.5
25	+	−	+	4	1	−	+	+	+	17.0
26	+	−	+	3	2	+	−	+	+	3.5
27	+	−	+	2	4	−	−	−	+	17.0
28	+	−	+	1	3	+	+	−	+	3.5
29	+	+	−	2	1	+	−	+	+	0.5
30	+	+	−	1	2	−	+	+	+	3.5
31	+	+	−	4	4	+	+	−	+	0.5
32	+	+	−	3	3	−	−	−	+	17.0

and

$$D_2 = AG = BE_3 = E_1F,$$

whereas the effects G, J, and GJ are clear (not aliased with any of the main effects or two-factor interactions). Follow-up experiments can be used to de-alias the above effects (see, e.g, Meyer, Steinberg, and Box 1996). An alternative to running a follow-up experiment is the approach in Wu and Hamada (2000) which appeals to the widely accepted principles of effect hierarchy and effect heredity. In that analysis, effect hierarchy was manually applied to the aliasing relationships to select the main effect D_2 as opposed to one of the two-factor interactions with which it is aliased. Similarly, effect heredity was used to justify selecting either the interaction D_2H or E_1G as opposed to the other four two-factor interaction effects. However neither of these two principles enable breaking the tie between D_2H and E_1G. Wu and Hamada (2000) argued that because the four spindles are synchronized, the effect of G should not vary substantially with the spindle position; thus ruling out the E_1G interaction, so that D_2H was the effect identified as significant.

In Step 0 of the proposed method, we obtain the empirical Bayes estimates of $\boldsymbol{\rho}$,

$$\hat{\boldsymbol{\rho}} = (0.99, 0.99, 0.99, 0.71, 0.99, 0.99, 0.60, 0.09, 0.56)'.$$

Figure 4(b) shows that the effects J, GJ, D_2, HJ, D_2H, G, and GHJ seem to be significant, which are the same as the first seven effects identified by the Bayesian forward selection strategy. Note that in the Bayesian analysis no confusion is created by the aliasing relationships. For example, at step 0, the t_i ratios for the effects in (15) are: $t_{AF} = 0.14$, $t_{D_2H} = 42.33$, $t_{CE_2} = 0.10$, $t_{BD_3} = 0.61$, $t_{D_1E_3} = 0.43$, and $t_{E_1G} = 0.70$. Thus D_2H stands out very clearly from the others as the significant effect. This could not be achieved using the frequentist analysis. Wu and Hamada (2000) were able to choose D_2H but only after applying expert knowledge of the process. Whereas the Bayesian approach was able to identify this effect through

mere data analysis. We also note that a reasonable, potentially significant three-factor interaction is identified by the Bayesian analysis, which was not even possible in the frequentist analysis.

By proposing the aforementioned Bayesian analysis, we are not trying to discourage the use of follow-up experiments. If a decision has to be made based on a one-shot experiment, the Bayesian analysis will be able to provide a unique answer. On the other hand, if resources do exist to perform follow-up experiments, then even in this situation, this type of Bayesian analysis can yield very useful information. For example, based on the t_i ratios, we can order the effects in (15): $D_2H \succ E_1G \succ BD_3 \succ D_1E_3 \succ AF \succ CE_2$. This ordering is immensely helpful for the optimal choice of follow-up runs. In frequentist analysis all six of these effects would be viewed as equally important and thus some of these additional resources will be spent on de-aliasing unimportant effects.

1.6 AN EXPERIMENTAL DESIGN EXAMPLE

A functionally induced prior is extremely useful for finding an optimal experimental design. In this section we will demonstrate its usefulness through an example. The detailed development of design criteria for three and higher level designs and their construction are left for future work.

Consider an $OA(16, 2^4 4^1)$. It can be constructed from an $OA(16, 2^{15})$ using the method of replacement as follows. Denote the 15 columns of $OA(16, 2^{15})$ by $\mathbf{1}, \mathbf{2}, \mathbf{3}, \mathbf{4}, \mathbf{12}, \mathbf{13}, \cdots, \mathbf{1234}$, where the two levels in each column are coded as -1 and 1. Let A be the four-level factor and B, C, D, and E be the four two-level factors. The columns $\mathbf{1}, \mathbf{2}$, and $\mathbf{12}$ can be combined to form the four-level factor. Now, how should the four two-level factors be assigned to the remaining 12 columns? Consider the following two choices given in Wu and Hamada (2000):

$$d_1 : \ A, \mathbf{3}, \mathbf{4}, \mathbf{23}, \mathbf{134},$$

(a) Traditional Analysis

(b) Bayesian Forward Selection (Step 0)
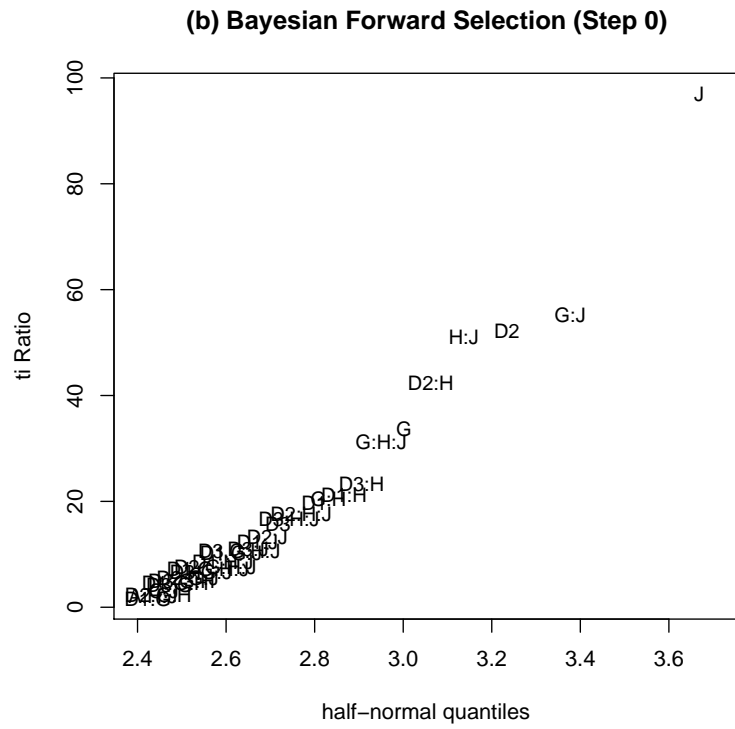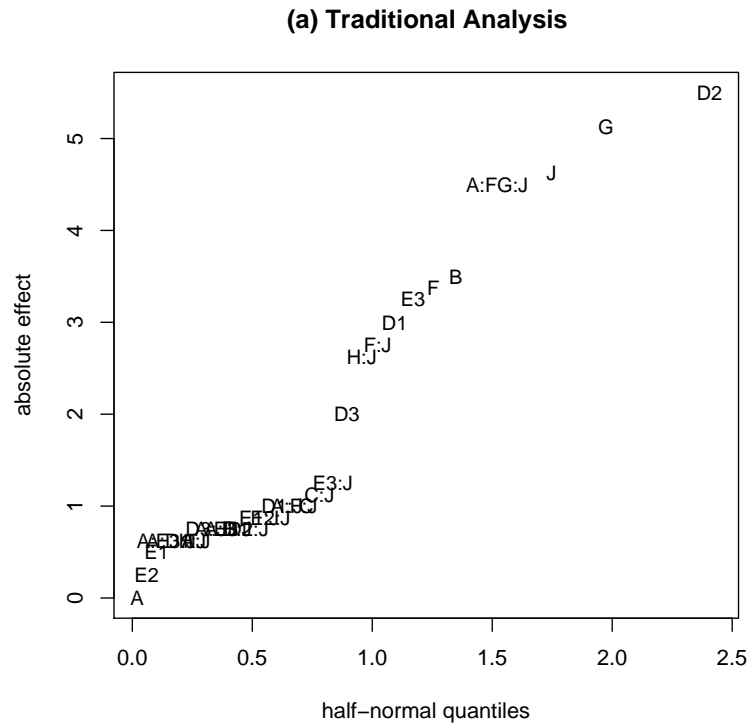
Figure 4: Analysis of the Router Bit Experiment

32

and

$$d_2 : \quad A, \mathbf{3}, \mathbf{4}, \mathbf{34}, \mathbf{124}.$$

The factor A can be represented by three dummy variables a_1, a_2, and a_3. Using Wu-Hamada coding $a_1 = \mathbf{1}$, $a_2 = \mathbf{2}$, and $a_3 = \mathbf{12}$. Then the defining contrast subgroup of the two designs are given by

$$d_1 : \quad \boldsymbol{I} = a_1 BCE = a_2 BD = a_3 CDE,$$

and

$$d_2 : \quad \boldsymbol{I} = BCD = a_3 CE = a_3 BDE.$$

In the case of two-level factors Joseph (2006) has shown that the posterior variance of β_0 can be minimized by minimizing $\sum_{i \in J_0(d)} R_{ii}$, where $J_0(d)$ denotes the indices of the effects in the defining contrast subgroup of design d. Denote this objective function by $W_0(d)$. If the four-level factor is a qualitative factor, then by Proposition 1, the β_i's are independent and we can use the result in Joseph (2006). Let the prior variances of the four-level factor be $\tau_0^2 r_4$ and that of the two-level factors $\tau_0^2 r_2$. Then,

$$W_0(d_1) = 1 + r_4 r_2^2 + 2 r_4 r_2^3,$$

and

$$W_0(d_2) = 1 + r_4 r_2^2 + (1 + r_4) r_2^3.$$

Since $r_4 < 1$, we can see that $W_0(d_1) < W_0(d_2)$. Therefore d_1 is a better design than d_2. This agrees with the minimum aberration criterion proposed by Wu and Zhang (1993). This example shows that there may be some interesting connections between the Bayesian criterion and the minimum aberration criterion. Note that the minimum aberration criterion in Wu and Zhang (1993) considers only qualitative factors. Cheng and Ye (2005) proposed design criteria for quantitative factors using indicator functions. This extension can be easily made using the Bayesian approach presented here, because we only need to change the correlation function. We leave the details for future research.

1.7 CONCLUSIONS

Typically, frequentist methods in the analysis of three and higher level experiments require significant work to resolve ambiguities. First the analyst, must identify the aliasing relationships. In the case of a regular fraction, this will enable the analyst to make variable selection decisions based on the well known principles of effect sparsity, effect hierarchy, and effect heredity. After analyzing the data, there may still be the need to run a follow-up experiment to resolve issues that arise from aliasing. In nonregular designs, such as the 18-run designs, the traditional approach would only consider estimating main effects, because of complex aliasing. Modern techniques such as those presented in Wu and Hamada (2000), while adequate in extracting more information from these designs, do not lend themselves well to being an automatic procedure. Here we have extended the use of functionally induced priors to designs that involve three-level and four-level factors. From this exposition, the procedure for extending the ideas for fractions of factorials not directly addressed here should be obvious. These tools provide a major step toward a reasonable fully automatic procedure for analyzing experimental data. Not only are the procedures well grounded in theory that facilitate the above mentioned principles of analysis of experiments, but the procedures are easy to implement and yield credible empirical results.

In the general framework, a Gaussian process over the design space induces a joint prior distribution for the linear model's parameters. From this, some additional assumptions about experimental design can be validated. Yet two effect ordering principles for three-level designs from Cheng and Ye (2005) could be challenged as a consequence of the theory here. We could be more specific about when the ordering assumptions are valid and explain why. Moreover new rules can be obtained when the assumptions are not valid.

We make a distinction between *qualitative* factors and *quantitative* factors. This becomes important with three-level and higher designs. We also provide a consistent

and logical way of addressing this distinction through the specification of the correlation function that partially characterizes the underlying Gaussian Process. This approach fits into the Gaussian process functional prior framework seamlessly.

In our discussion of the examples, we note that some designs might be more likely than others to produce ambiguities in variable selection. When these arise, they could either be resolved manually or resolved through a simple automated procedure that respects the principles of effect sparsity, effect hierarchy and effect heredity. When different components of the $\boldsymbol{\rho}$ vector are used for each factor, this issue becomes increasingly less likely to be a concern.

This chapter also presents an example that illustrates how functionally induced priors can be used in optimal design. The example shows that the Bayesian criterion and the minimum aberration criterion may be related for the case of qualitative factors. The Bayesian criterion is more general, because there is no restriction on the type of design or the number of runs. Moreover, the Bayesian criterion can be easily extended to deal with the case of quantitative factors. We believe that some very useful optimal design results can be obtained by using the Bayesian methodology proposed in this chapter.

1.8 CONTRIBUTIONS

The research described in this chapter contributes to the body of knowledge in the field of the Design and Analysis of Experiments in the following ways:

1. A prior distribution which seamlessly incorporates the properties of effect hierarchy and effect heredity for the parameters in the linear model is developed for three and higher level experiments.

2. This prior is also designed to allow for the specialization for qualitative and quantitative factors through a simple correlation structure specification.

3. A new analysis methodology is provided which is computationally simpler than the existing Bayesian analysis methodologies.

4. A coherent methodology is described for obtaining an *a priori* ordering of the importance of factorial effects.

5. A coherent and general methodology is further advanced toward the goal of a completely automatic expert system for the design and analysis of experiments.

CHAPTER II

ANALYSIS OF OPTIMIZATION EXPERIMENTS

2.1 INTRODUCTION

Experiments are used for many purposes such as for optimizing a process, for developing a prediction model, for identifying important factors, and for validating a scientific theory. Among these, optimization is arguably the most important objective in industrial experiments (Taguchi 1987, Wu and Hamada 2000, Myers and Montgomery 2002, Montgomery 2004). However, the same type of data analysis is used irrespective of the underlying objective. Here we argue that the analysis of optimization experiments should be done in a different way.

The existing approach to data analysis is to first identify the statistically significant effects that influence the response. Analysis of variance, t-tests, half-normal plots, step-wise regression, and other variable selection techniques are used for this purpose. Once the significant effects are identified, a model is built involving only those factors. The model is then optimized to find the best settings of the factors. The factors that are not statistically significant are allowed to take any values in the experimental range. Their settings are left to the discretion of the experimenter. The usual recommendation is to choose levels that minimize the cost. The foregoing procedure is very intuitive and might be adequate, but it is in the identification of the significant factors where something can go wrong.

The basic flaw in the procedure is that the objective of optimization cannot be easily translated into meaningful quantities used in a significance test. An α level

of 5% is usually used for identifying the significant effects. But what is this signifi-
cance level's connection to the optimization of a machining process in order to reduce
dimensional variation or the optimization of a chemical process in order to improve
yield? Using a quantity in a procedure that has no direct connection to the objective
of the experiment can be misleading.

For example, consider an experiment with the objective of increasing the lifetime
of a product. A factor x is varied at two levels -1 and 1 in the experiment. Suppose
that the lifetimes observed at these two settings are 50 and 65 hours, respectively.
Consider the model $y = \beta_0 + \beta_1 x + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ with $\sigma = 10$. We obtain
the least squares estimate $\widetilde{\beta}_1 = 7.5$. Now to test the hypothesis $H_0 : \beta_1 = 0$ against
$H_1 : \beta_1 \neq 0$, we obtain

$$\text{p-value} = 2\Phi\left(-\frac{\widetilde{\beta}_1}{\sigma/\sqrt{2}}\right) = 0.2888,$$

where Φ is the standard normal distribution function. This level is much higher than
$\alpha = .05$, hence we would fail to reject H_0 and would conclude that the factor is not
significant.

Now let us take a different view of this problem, that with optimization as the
objective. It is easier to use a Bayesian framework to demonstrate what is happening.
Under the improper prior distribution, $p(\boldsymbol{\beta}) \propto 1$, $\boldsymbol{\beta} \in \mathbb{R}^2$, the posterior distribution
of β_1 given the data $(\boldsymbol{y})$ is $\mathcal{N}(\widetilde{\beta}_1, \sigma^2/2)$. Thus

$$Pr(\beta_1 > 0|\boldsymbol{y}) = \Phi\left(\frac{\widetilde{\beta}_1}{\sigma/\sqrt{2}}\right) = 0.8556.$$

In other words, if we set $x = 1$, then there is an 86% chance that the lifetime will be
higher than when $x = -1$. No matter what, we need to set x to some value. Thus
we should choose 1, a conclusion quite different from that obtained when using the
statistical test of significance.

What makes an optimization experiment different? When we optimize a product

or process, we need to select a level for the factor irrespective of whether it is statistically significant or not. A factor can be easily thrown out of a model, but cannot be thrown out of a product or process. Thus the application of a test of significance makes sense in the case of experiments where the objective is prediction or screening, but not when the objective is optimization. When developing a model for prediction and screening, one can focus on balancing model fit and size, but when developing a model for optimization, a balance should be made between the improvement that can be achieved and the cost associated with changing the level of factors.

In the example, suppose instead that the lifetime at $x = 1$ is 50.1 hours. Because this is greater than the lifetime at $x = -1$, there is still more than a 50% chance of achieving an improvement by changing the setting to $x = 1$. However, the improvement is very small. So, should we make this change? To answer this question, we may look into the cost associated with such a change. Suppose changing x from -1 to 1 reduces the cost, then the best decision after the experiment seems to be to choose $x = 1$. But if it increases the cost, then the decision is not easy. We may not want to change the setting unless the improvement of 0.1 hours is worth more to us than the increase in cost of producing the product with $x = 1$. Thus if the improvement is practically insignificant, then we may decide not to make any change. Let Δ denote the practical significance level. Then a change will be made if $|2\widetilde{\beta}_1| > \Delta$. For example, Δ could be taken to be 5% of the existing lifetime. Thus we will make a change if the improvement is more than 2.5 hours. Note that here, the use of the 5% level is much more meaningful than the 5% level used in the test of significance. It is much easier to say "make a change if it can result in at least a 5% improvement" than to say "make a change if the factor is statistically significant at the 5% level".

One immediate objection to this approach might be that it does not consider the randomness in the response. We will overcome this problem by modifying the estimation method of β_1. We will show that empirical Bayes estimation assuming a

proper prior distribution for β_1 will give an estimate that shrinks as σ increases. Thus when σ is large enough, the expected improvement becomes less than Δ, indicating no change for the factor level.

In the recent literature, Bayesian analysis techniques have been successfully implemented to address the problem of process optimization. In particular, Rajagopal and Castillo (2005) presents an intriguing fully Bayesian approach that obtains a posterior predictive distribution by averaging over candidate classes of models. From this, factor settings can be ranked based on their probability of yielding a response in some desirable range. Rather than requiring the specification of priors in the model space, our approach concentrates on the familiar, full factorial linear model and utilizes the approach of Chapter 1 for parameter prior elicitation and hyper-parameter estimation.

The details of the proposed analysis method are described in the following sections. It differs from the usual, frequentist analysis in two aspects: the statistical significance level is replaced with a practical significance level and the least squares estimation is replaced with empirical Bayes estimation. First, we present a real experiment to motivate the problem solution.

2.2   AN EXAMPLE

Consider the experiment reported by Hellstrand (1989) with the objective of reducing the wear rate of deep groove bearings (see also Box, Hunter, and Hunter 2005, pp. 209-211). A two-level full factorial design over three factors: osculation (x_1), heat treatment (x_2), and cage design (x_3), was used for the experiment. The design and the data are given in Table 5.

The estimates of the seven effects are given in Table 6. Because this is an unreplicated experiment, t-values cannot be computed in order to test the significance

Table 5: Design Matrix and Data, the Bearing Experiment

| | Factor | | | lifetime | wear |
Run	x_1	x_2	x_3	(hours)	rate
1	-1	-1	-1	17	5.882
2	-1	-1	1	19	5.263
3	-1	1	-1	26	3.846
4	-1	1	1	16	6.250
5	1	-1	-1	25	4.000
6	1	-1	1	21	4.762
7	1	1	-1	85	1.176
8	1	1	1	128	0.781

Table 6: Parameter Estimates and Significance, the Bearing Experiment

| | | | Approx. p-value | |
Effect	$\widetilde{\beta}_i$	$|t_{PSE}|$	IER	EER
x_1	-1.315	1.678	0.10	> 0.40
x_2	-0.982	1.253	0.19	> 0.40
x_3	0.269	0.343	> 0.40	> 0.40
$x_1 x_2$	-0.719	0.918	0.31	> 0.40
$x_1 x_3$	0.177	0.226	> 0.40	> 0.40
$x_2 x_3$	-0.233	0.298	> 0.40	> 0.40
$x_1 x_2 x_3$	-0.523	0.667	> 0.40	> 0.40

of each effect. Hamada and Balakrishnan (1998) provides an excellent and comprehensive review of the very many techniques that have been suggested for identifying active effects under this complication of unreplication. A common approach is to use a half-normal plot (Daniel 1959) and declare the large effects that appear to be outliers as the significant effects. The half normal plot of the effects is given in Figure 5. We can see that none of the effects seem to be significant.

A more formal approach for identifying significant effects in unreplicated experiments is to use the method proposed by Lenth (1989). (See Hamada and Balakrishnan (1998) for an excellent review of many other methods as well as Variyath, Abraham,

and Chen (2005) for a new approach using the jackknife that also examines this bearing example.) The t_{PSE} values from applying Lenth's method are given in Table 6. Two types of critical values may be used: the individual error rate (IER) and the experiment-wise error rate (EER). IER and EER critical values for Lenth's test are tabulated in Wu and Hamada (2000). At the 5% significance level the critical value for IER is 2.30. Because the t_{PSE} values are much lower than this value, none of the effects are found to be significant. The EER critical value is 4.87, which is much larger than the IER critical value, and thus the same conclusion would be obtained. We can also compute the p-values for each effect based on IER and EER. They are also shown in Table 6. We can see that the p-values are large enough to conclude that none of the effects are significant.
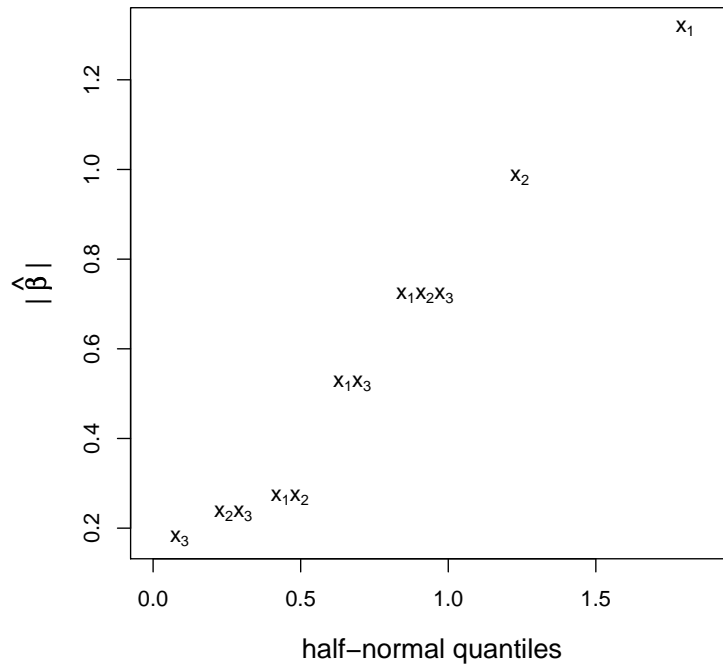
Figure 5: Half Normal Plot for the Bearing Experiment

By examining the data in Table 5, we can see that run numbers 7 and 8 produce

wear rates much lower than those of the other runs. It appears that keeping osculation and temperature simultaneously at their high values is beneficial. Hellstrand (1989) confirmed this choice of factor settings through observing vastly improved bearing performance in a particular application. These settings are said to yield a substantial improvement in wear rate that would have been missed if we were to rely upon only the statistical test of significance.

2.3 PRACTICAL SIGNIFICANCE LEVEL

Let Y denote the response and $\boldsymbol{x} = (x_1, x_2, \cdots, x_p)'$ the experimental factors. Let $L(Y)$ be an appropriate quality loss function that converts the units of the response measurements into dollars. Let $C(\boldsymbol{x})$ be the cost function that reflects the cost of running the process or producing the product at each of the particular settings for the factors. Then, our objective is to find the optimal settings for the factors that minimize the total cost

$$TC = \mathbb{E}\{L(Y)\} + C(\boldsymbol{x}), \tag{16}$$

where the expectation is taken with respect to the distribution of the response.

The form of the cost function $C(\boldsymbol{x})$ is problem-specific and can be difficult to obtain. Therefore, we propose a general strategy that can be used without requiring the knowledge of the actual form of the cost function. To achieve this, we will identify the factors that have a practically significant effect on $\mathbb{E}\{L(Y)\}$ and use only those factors in order to minimize $\mathbb{E}\{L(Y)\}$. The settings of the other factors may be selected so as to minimize the cost. This is similar to the existing strategy, except that practical significance is used instead of statistical significance and factor significance is used instead of effect significance.

To be more specific, we select a model for $\mathbb{E}\{L(Y)\}$ optimize it, and adopt that setting for a factor, only if it is a practically significant factor. A factor will be identified as practically significant if its effect on the response is more than a prescribed

practical significance level Δ.

For example, consider the bearing experiment again. It is easier to illustrate the concept if only the main effects are present in the model. The main effects model is given by

$$\widehat{y} = 3.995 - 1.315x_1 - 0.982x_2 + 0.269x_3.$$

The wear rate is a smaller-the-better (STB) characteristic and thus $L(Y) = KY$ is a reasonable loss function to use (see Joseph 2004). So we need to minimize the mean $\mathbb{E}(Y)$ which is estimated by $\widehat{y}$. Suppose that the existing level of wear rate is 5 and a 5% decrease is considered to be a significant improvement, then we can take $\Delta = .05 \times 5 = 0.25$. Each of the factors can independently make a change of two times its coefficient estimate (because they vary from -1 to 1). All of these are more than 0.25 and so all of the factors are identified as practically significant. Under the main effects model, an estimate of σ can be obtained. Using the ubiquitous independent t-test procedure we can find that the factor x_1 is statistically significant at the 5% level (p-value= 0.048) and the other two effects are not significant (p-values are 0.104 and 0.596), a very different conclusion from that arrived at from the application of the practical significance level.

Now consider the full linear model with interactions. It is given by

$$\widehat{y} = 3.995 - 1.315x_1 - 0.982x_2 + 0.269x_3 - 0.719x_1x_2 - 0.177x_1x_3 + 0.233x_2x_3 - 0.523x_1x_2x_3.$$

To apply the practical significance level to each factor, we need to know the effect of each factor. But, because interactions are present the effect of a factor changes with the levels of the other factors. When there are factors present having more than two levels, then we might consider their quadratic, cubic, etc. effects. Therefore, we need a more general concept than "effects". Therefore, we need a more general concept than "effects". To address this issue and alleviate any confusion with the definition of factorial effects, we define the *impact* of a factor with respect to the optimal setting

of $\boldsymbol{x}$.

Let $\mathbb{E}\{L(Y)\} = g(\boldsymbol{x})$ and let $\boldsymbol{x}^*$ minimize $g(\boldsymbol{x})$. The minimization is performed while constraining $\boldsymbol{x}$ within the experimental region. Define the *impact* of factor x_i as

$$imp(x_i) = \max_{x_i} g(x_i, \boldsymbol{x}^*_{(i)}) - \min_{x_i} g(x_i, \boldsymbol{x}^*_{(i)}),$$

where $x_{(i)}$ denotes all of the factors except x_i. The impact is the maximum change in $\mathbb{E}\{L(Y)\}$, when the factor x_i is changed from its best to worst settings. If this change is less than Δ, then we will identify the factor as practically insignificant. It is easy to see that if $g(\boldsymbol{x}) = \beta_0 + \sum_{i=1}^{p} \beta_i x_i$ and if the two levels are encoded by -1 and 1, then $imp(x_i) = |2\beta_i|$, which would coincide exactly with the usual definition of that factorial effect.

To identify two factors as practically insignificant, we should also consider their combined impact:

$$imp(x_i, x_j) = \max_{x_i, x_j} g(x_i, x_j, \boldsymbol{x}^*_{(i,j)}) - \min_{x_i, x_j} g(x_i, x_j, \boldsymbol{x}^*_{(i,j)}).$$

The two factors x_i and x_j would be identified as practically insignificant if $imp(x_i, x_j) < 2\Delta$, in addition to each of $imp(x_i) < \Delta$ and $imp(x_j) < \Delta$. In this manner, we can extend these definitions to any number of factors. In fact, we may define the set of practically insignificant factors as:

$$\mathcal{S}^* := \{\mathcal{S} \ni \forall s \subseteq \mathcal{S} : imp(s) < card\,(s)\,\Delta\}, \tag{17}$$

where $card(s)$ represents the number of elements in the set s. In words, the set of practically insignificant factors is *the largest set of factors such that every subset has an impact less than the practical significance level times the number of elements in that subset.* So we are not merely interested in obtaining the set of factors such that each factor's impact is less than Δ. We must also consider all possible combined impacts. The search for this largest set of insignificant factors can be performed through

45

an exhaustive search. However, we propose Algorithm 1 for identifying this set of insignificant factors. In this algorithm, $\mathcal{X}$ represents the set of p factors: $\{x_1, \ldots, x_p\}$. At each step we increase the cardinality of the set $\mathcal{S}^*$ by including the factor x^* that yields the smallest combined impact, so long as for this new $\mathcal{S}^*$, the marginal increase in combined impact from this step is still less than Δ.

Algorithm 1 Identify the Full Set of Insignificant Factors: $\mathcal{S}^*$

$\mathcal{S}^* \Leftarrow \varnothing$
$p \Leftarrow card\,(\mathcal{X})$
$k \Leftarrow 1$
while $k \leq p$ **do**
   $x^* \Leftarrow argmin_{x \in \mathcal{X} \setminus \mathcal{S}^*}\, imp(x \cup \mathcal{S}^*)$
   if $imp(x^* \cup \mathcal{S}^*) - imp(\mathcal{S}^*) \geq \Delta$ **then**
     return $\mathcal{S}^*$
   else
     $\mathcal{S}^* \Leftarrow x^* \cup \mathcal{S}^*$
     $k \Leftarrow k + 1$
   end if
end while
return $\mathcal{S}^*$

By optimizing the full linear model, we obtain $x_1^* = 1$, $x_2^* = 1$, and $x_3^* = 1$. Now the impact of the three factors can be computed as

$$imp(x_1) = 2 \times |-0.523 - 0.177 - 0.719 - 1.315| = 5.469,$$

$$imp(x_2) = 2 \times |-0.523 + 0.233 - 0.719 - 0.982| = 3.981,$$

$$imp(x_3) = 2 \times |-0.523 + 0.233 + 0.269 - 0.177| = 0.395.$$

Because all of these impacts are more than 0.25, they are all identified as practically significant. There are no insignificant factors. Thus all three factors should be changed to their higher levels to minimize the wear rate. This is a much different conclusion than what we obtain using the statistical significance tests.

This result agrees with the conclusion obtained by Hellstrand (1989), except therein the factor cage design (x_3) was not considered significant. Can the observed

effect of x_3 be entirely due to random error? Are we unnecessarily incurring a potential cost by forcing the cage design to its higher level? We will answer these questions in the next section.

2.4 EMPIRICAL BAYES ESTIMATION

Suppose that the response is related to the factors through the model $Y = \beta_0 + \sum_i \beta_i u_i + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and u_i's are functions of the factors. For example, in a 2^3 design we can take $u_1 = x_1$, $u_2 = x_2$, $u_3 = x_3$, $u_4 = x_1 x_2$, $u_5 = x_1 x_3$, $u_6 = x_2 x_3$, and $u_7 = x_1 x_2 x_3$. Let $\boldsymbol{u} = (1, u_1, u_2, \cdots)'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \cdots)'$. Then $Y = \boldsymbol{u}'\boldsymbol{\beta} + \epsilon$. To use Bayesian methods, we need to specify a prior distribution for $\boldsymbol{\beta}$. An excellent general reference for Bayesian analysis is provided by Gelman, Carlin, Stern, and Rubin (2004). For notational simplicity, rewrite the model as $Y = \mu + \boldsymbol{u}'\boldsymbol{\beta} + \epsilon$, where μ denotes the prior mean for β_0. We use the following multivariate normal prior: $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$.

Let $\boldsymbol{D}$ be the design matrix with n runs and $\boldsymbol{U}_D$ be the model matrix. Let $\boldsymbol{y}$ denote the data obtained from the experiment. Assuming the ϵ's are independent, we have the Bayesian model

$$\boldsymbol{y}|\boldsymbol{\beta} \sim \mathcal{N}(\mu\boldsymbol{1}_n + \boldsymbol{U}_D\boldsymbol{\beta}, \sigma^2\boldsymbol{I}_n) \text{ and } \boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}),$$

where $\boldsymbol{1}_n$ is a vector of 1's having length n and $\boldsymbol{I}_n$ is the n-dimensional identity matrix. Then the posterior mean of $\boldsymbol{\beta}$ given the data is

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\Sigma}\boldsymbol{U}'_D(\boldsymbol{U}_D\boldsymbol{\Sigma}\boldsymbol{U}'_D + \sigma^2\boldsymbol{I}_n)^{-1}(\boldsymbol{y} - \mu\boldsymbol{1}_n). \tag{18}$$

The unknown hyper-parameters in the model can be estimated using empirical Bayes methods. The log-likelihood of the marginal distribution of $\boldsymbol{y}$ is given by

$$l = constant - \frac{1}{2}\log det(\boldsymbol{U}_D\boldsymbol{\Sigma}\boldsymbol{U}'_D + \sigma^2\boldsymbol{I}_n) - \frac{1}{2}(\boldsymbol{y} - \mu\boldsymbol{1}_n)'(\boldsymbol{U}_D\boldsymbol{\Sigma}\boldsymbol{U}'_D + \sigma^2\boldsymbol{I}_n)^{-1}(\boldsymbol{y} - \mu\boldsymbol{1}_n).$$

The log-likelihood can be maximized with respect to μ, and the parameters in $\boldsymbol{\Sigma}$ to get their estimates. We consider three special structures for $\boldsymbol{\Sigma}$. They are presented

in the order of increasing complexity. The last covariance structure is the one we recommend, however the discussion of the first two is provided because it reveals additional insights into the overall procedure.

2.4.1 Identical Variances Prior

Consider the bearing experiment again. For simplicity, assume that $\boldsymbol{\Sigma} = \tau^2 \boldsymbol{I}_8$. Because $\boldsymbol{D}$ is a full factorial design, the columns of $\boldsymbol{U}_D$ are orthogonal. Thus $\boldsymbol{U}_D \boldsymbol{\Sigma} \boldsymbol{U}_D' = 8\tau^2 \boldsymbol{I}_8$. From (18), we obtain

$$\widehat{\boldsymbol{\beta}} = \frac{\boldsymbol{U}_D'(\boldsymbol{y} - \mu \mathbf{1}_8)}{8 + \sigma^2/\tau^2}.$$

The least squares estimate of $\boldsymbol{\beta}$ is given by

$$\widetilde{\boldsymbol{\beta}} = (\boldsymbol{U}_D' \boldsymbol{U}_D)^{-1} \boldsymbol{U}_D'(\boldsymbol{y} - \mu \mathbf{1}_8) = \frac{1}{8} \boldsymbol{U}_D'(\boldsymbol{y} - \mu \mathbf{1}_8).$$

Thus

$$\widehat{\boldsymbol{\beta}} = \frac{8}{8 + \sigma^2/\tau^2} \widetilde{\boldsymbol{\beta}},$$

which illustrates that the Bayes estimate shrinks the least squares estimate by the factor $8/(8 + \sigma^2/\tau^2)$.

The marginal log-likelihood simplifies to

$$l = constant - \frac{8}{2} \log(8\tau^2 + \sigma^2) - \frac{(\boldsymbol{y} - \mu \mathbf{1}_8)'(\boldsymbol{y} - \mu \mathbf{1}_8)}{2(8\tau^2 + \sigma^2)}.$$

Differentiating with respect to μ and τ^2 and equating to 0, we obtain the familiar solutions

$$\widehat{\mu} = \bar{y}$$

and

$$8\tau^2 + \sigma^2 = \frac{1}{8} \sum_{i=1}^{8} (y_i - \bar{y})^2.$$

Denote the right side of this equation, the sample variance of Y, by s^2. Then, because τ^2 cannot be negative, we obtain

$$\widehat{\tau}^2 = \frac{1}{8} \left(s^2 - \sigma^2 \right)_+,$$

where $(x)_+ = x$ if $x > 0$ and 0 otherwise. Thus

$$\widehat{\boldsymbol{\beta}} = \left(1 - \frac{\sigma^2}{s^2}\right)_+ \widetilde{\boldsymbol{\beta}}.$$

Thus the estimate of $\boldsymbol{\beta}$ decreases as σ^2 increases and becomes 0 when σ^2 exceeds the observed variance of Y. The above estimator may be recognized as the well-known positive-part James-Stein estimator (see Lehmann and Casella 1998, pg. 275). The connection between James-Stein estimation and empirical Bayes estimation is well-known in the statistical literature. However, we have not seen it advanced as an alternative to statistical testing for the analysis of experiments.

The coefficients $\beta_1, \beta_2, \cdots, \beta_7$ are plotted in Figure 6(a) against σ^2 (note that $\beta_0 = 0$). We can see that as σ^2 increases, the β's decrease to 0. The impacts of the three factors can be calculated as before and are plotted in Figure 6(b). We can see that the impact of x_3 is practically insignificant at the 5% level when $\sigma^2 > 1.4$. Therefore, it can be set to minimize the cost. This is exactly the same result obtained by Hellstrand (1989) with his subsequent experiments. The analysis shows that even in the presence of large random error, the two factors, osculation and heat treatment, have significant effects and can be adjusted to improve the wear rate substantially. This is a conclusion completely different from that obtained using the statistical tests of significance.

2.4.2 Unequal Variances Prior

Now consider a more general form for $\boldsymbol{\Sigma}$. As before, let the β_i's be independent but with possibly different prior variances: τ_i^2. Then $\boldsymbol{\Sigma} = diag(\tau_0^2, \tau_1^2, \cdots, \tau_7^2)$. We obtain,

$$\widehat{\beta}_i = \frac{8\tau_i^2}{8\tau_i^2 + \sigma^2} \widetilde{\beta}_i,$$

and the marginal log-likelihood becomes

$$l = constant - \frac{1}{2}\sum_{i=0}^{7} \log(8\tau_i^2 + \sigma^2) - \frac{1}{2}\sum_{i=0}^{7} \frac{8\widetilde{\beta}_i^2}{8\tau_i^2 + \sigma^2}. \tag{19}$$

(a)

(b)
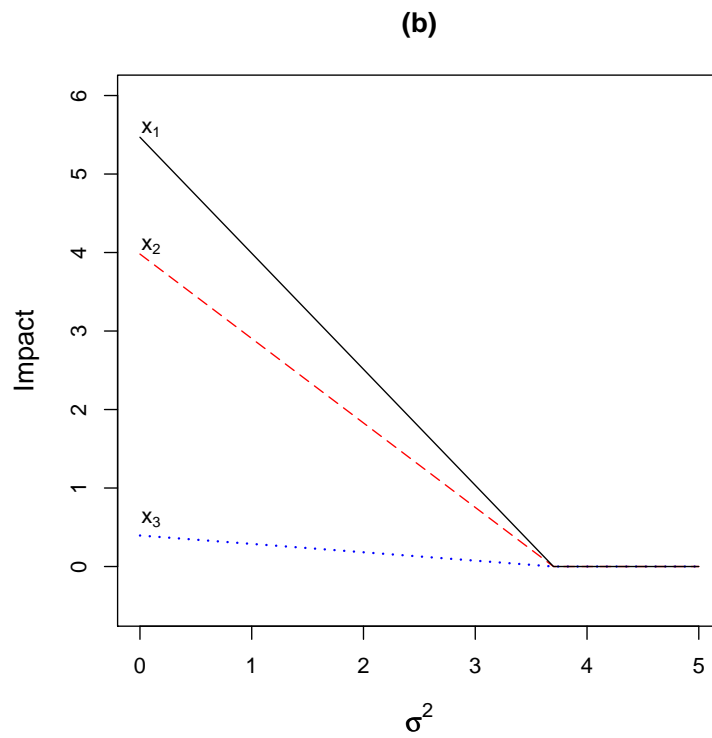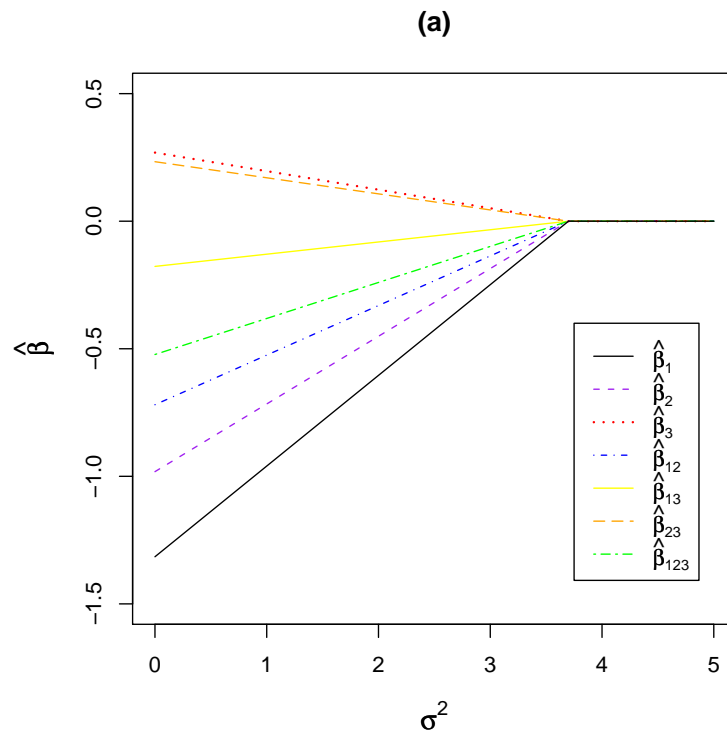
Figure 6: Bearing Experiment With Equal Prior Variances: (a) Coefficients (b) Impacts

50

Maximizing l, we obtain $\widehat{\tau}_i^2 = (\widetilde{\beta}_i^2 - \sigma^2/8)_+$. Let

$$z_i = \frac{\widetilde{\beta}_i}{\sigma_{\widetilde{\beta}_i}},$$

which is the usual test statistic for testing $H_0 : \beta_i = 0$ when $\sigma_{\widetilde{\beta}_i}$ denotes the standard error of $\widetilde{\beta}_i$. Because $\sigma_{\widetilde{\beta}_i} = \sigma/\sqrt{8}$, we obtain

$$\widehat{\beta}_i = \left(1 - \frac{1}{z_i^2}\right)_+ \widetilde{\beta}_i. \tag{20}$$

This shows that $\widehat{\beta}_i$ shrinks completely to 0 if $|z_i| \leq 1$. This threshold is equivalent to using an α level of about 32% in statistical testing. That is, when $|z_i| > 1$, the ith coefficient is identified as statistically significant at the 32% level and $\widetilde{\beta}_i$ is used in the model. Whereas in the EB shrinkage procedure, a value smaller than $\widetilde{\beta}_i$ is used and as $|z_i|$ increases, $\widehat{\beta}_i$ increases continuously to $\widetilde{\beta}_i$. A more detailed comparison of the *hard thresholding rule* of statistical testing and the *soft* or *continuous thresholding rule* of this shrinkage estimator are provided in Section 2.5.

Here we address the interesting connection between shrinkage estimators and subset regression techniques. If we rewrite (20) as:

$$\widehat{\beta}_i = \left(1 - \frac{\sigma_{\widetilde{\beta}_i}^2}{\widetilde{\beta}_i^2}\right)_+ \widetilde{\beta}_i$$

then we can recognize that this is very similar to the nonnegative (nn-) garrote coefficients of Breiman (1995). However, in the nn-garrote, the values $\sigma_{\widetilde{\beta}_i}^2$ are replaced by a single parameter that is estimated through minimizing squared errors in a cross-validation scheme. Note however, that the EB estimate recommended here is more general than any of the techniques mentioned above to which we are drawing comparisons. We may use the EB estimate for all factorial effects from a fractional factorial design. Its representation does not depend on the existence of a corresponding least-squares estimate.

The estimates of the coefficients are plotted in Figure 7(a). In addition, the impacts for the three factors at their optimal settings are plotted in Figure 7(b). We can

see that the coefficients shrink to 0 at a slower rate. The impact of x_3 is practically insignificant at the 5% level when $\sigma^2 > 1.7$. The impacts of x_1 and x_2 are practically significant, provided $\sigma^2 < 12.5$ and $\sigma^2 < 6.7$, respectively.

Although we used the 2^3 design to derive the result in (20), the result is much more general. It can be applied to fractional factorial designs and to designs with factors having more than two levels. The only restriction is that the model matrix corresponding to the effects that we are trying to estimate should be orthogonal. The proposition is formally stated and proved in Appendix B.

The approach can easily be extended to the case of an unknown σ. If an estimate of σ can be obtained, then $t_i = \widetilde{\beta}_i / \widehat{\sigma}_{\widetilde{\beta}_i}$ has a t distribution. Thus, we obtain

$$\widehat{\beta}_i = \left(1 - \frac{1}{t_i^2}\right)_+ \widetilde{\beta}_i. \tag{21}$$

2.4.3   Heredity Prior

The foregoing analysis does not incorporate the principles of effect hierarchy and effect heredity (Hamada and Wu 1992). The effect hierarchy principle is not incorporated because the main effects, two-factor interactions, and the three-factor interaction are all treated the same way. The effect heredity principle is not incorporated because an interaction term can appear in the model without any of its parent factors. Joseph (2006) and Chapter 1 of the present work show that these principles can easily be incorporated into the analysis through the prior specification. Let $\boldsymbol{\Sigma} = \tau^2 \boldsymbol{R}$, where $\boldsymbol{R} = diag(1, r_1, r_2, r_3, r_1 r_2, r_1 r_3, r_2 r_3, r_1 r_2 r_3)$, and $r_i \in [0, 1]$ for all i.

For convenience, let us introduce the indicator variables:

$$\gamma_{i,j} = \begin{cases} 1 & \text{if effect } i \text{ includes factor } j \\ 0 & \text{otherwise.} \end{cases}$$

So that, for example, $\gamma_{0,1} = \gamma_{0,2} = \gamma_{0,3} = 0$, and $\gamma_{1,1} = 1$, while $\gamma_{1,2} = \gamma_{1,3} = 0$. Then

(a)

(b)
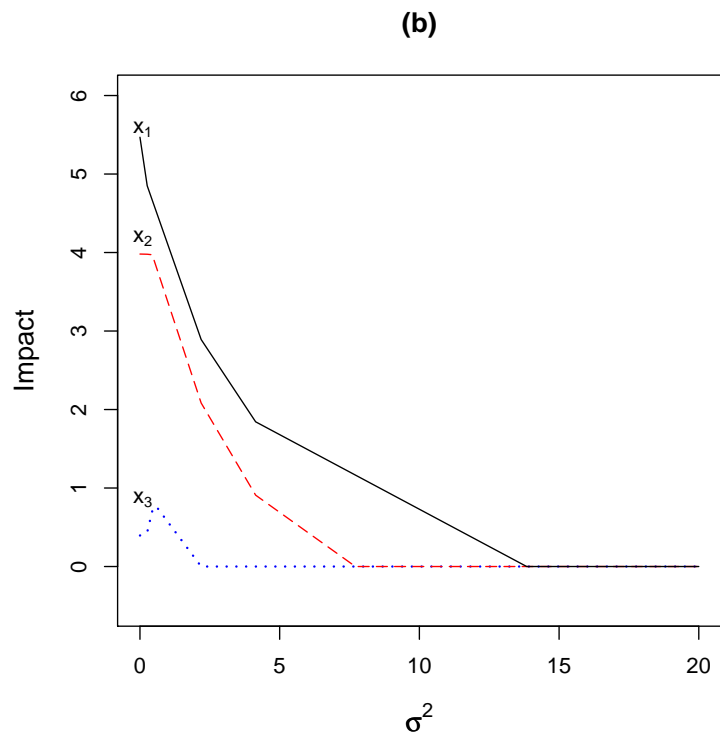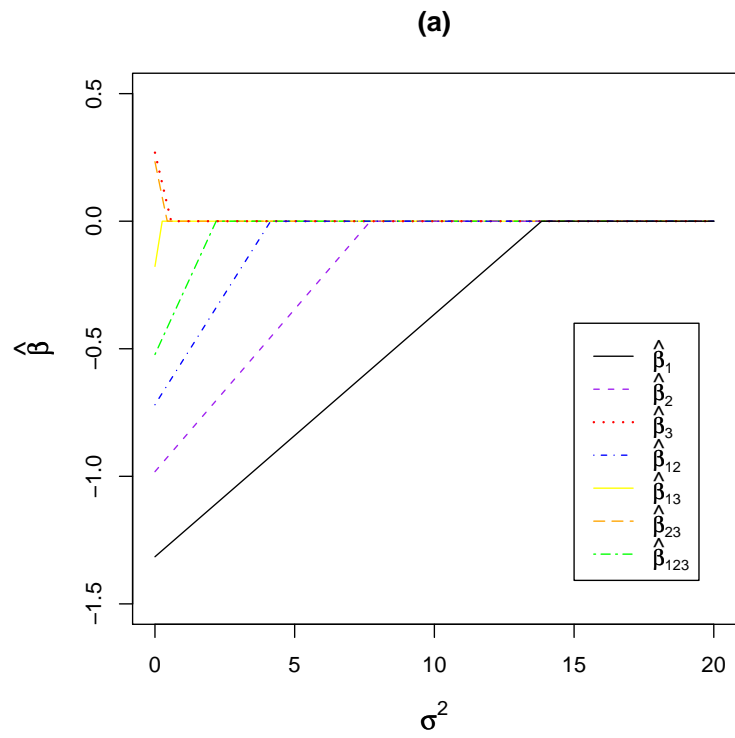
Figure 7: Bearing Experiment With Unequal Prior Variances: (a) Coefficients (b) Impacts

we have

$$\widehat{\beta_i} = \frac{8\tau^2 r_1^{\gamma_{i,1}} r_2^{\gamma_{i,2}} r_3^{\gamma_{i,3}}}{8\tau^2 r_1^{\gamma_{i,1}} r_2^{\gamma_{i,2}} r_3^{\gamma_{i,3}} + \sigma^2} \widetilde{\beta}_i,$$

and the marginal log-likelihood becomes

$$l = constant - \frac{1}{2} \sum_{i=0}^{7} \left(\log \left(8\tau^2 \prod_{j=1}^{3} r_j^{\gamma_{i,j}} + \sigma^2 \right) + \frac{8\widetilde{\beta}_i^2}{8\tau^2 \prod_{j=1}^{3} r_j^{\gamma_{i,j}} + \sigma^2} \right). \qquad (22)$$

We may numerically maximize this log likelihood in order to find empirical Bayes' estimates for the hyper-parameters $\mu, r_1, r_2, r_3,$ and τ^2.

The consequence of assuming the heredity model can be readily discerned from the plot of the coefficients given in Figure 8(a). Coefficients approach zero in groups as σ^2 increases. For instance, both $\widehat{\beta}_2$ and the interaction $\widehat{\beta}_{1,2}$ are zero for $\sigma^2 > 6$. Overall the coefficients shrink at a rate that is much more rapid than with just the unequal prior variances assumption. The separation between the significant effects and insignificant effects is quite discernable. For $\sigma^2 < 6.3$, x_1 is practically significant, for $\sigma^2 < 5.5$, x_2 is practically significant, and for $\sigma^2 < 0.3$, x_3 is practically significant. That is, the factor x_3, cage design, is practically insignificant under virtually all assumptions for the error variance. The impacts in Figure 8(b) are once again consistent with the conclusion of Hellstrand (1989).

2.5 STATISTICAL TESTING AS AN APPROXIMATION

For the empirical Bayes estimate in (20), the value of β_i in the estimated model can be written: $\lambda_i \widetilde{\beta}_i$, where $\lambda_i = (1 - 1/z_i^2)_+$. If statistical testing is used, then $\lambda_i = 0$ when $|z_i| \leq z_{\alpha/2}$ and 1 otherwise. As discussed in the introduction, it is difficult to find a meaningful value of α for a given problem. However, the similarity of this testing procedure with the empirical Bayes procedure reveals that statistical testing can be used as an approximation. A simple approximation is to take $z_{\alpha/2} = 1$, which gives an α level of 31.73%. But because the empirical Bayes estimates shrink towards 0 when $z_i > 1$, we may prefer to search for an even closer approximate statistical test.
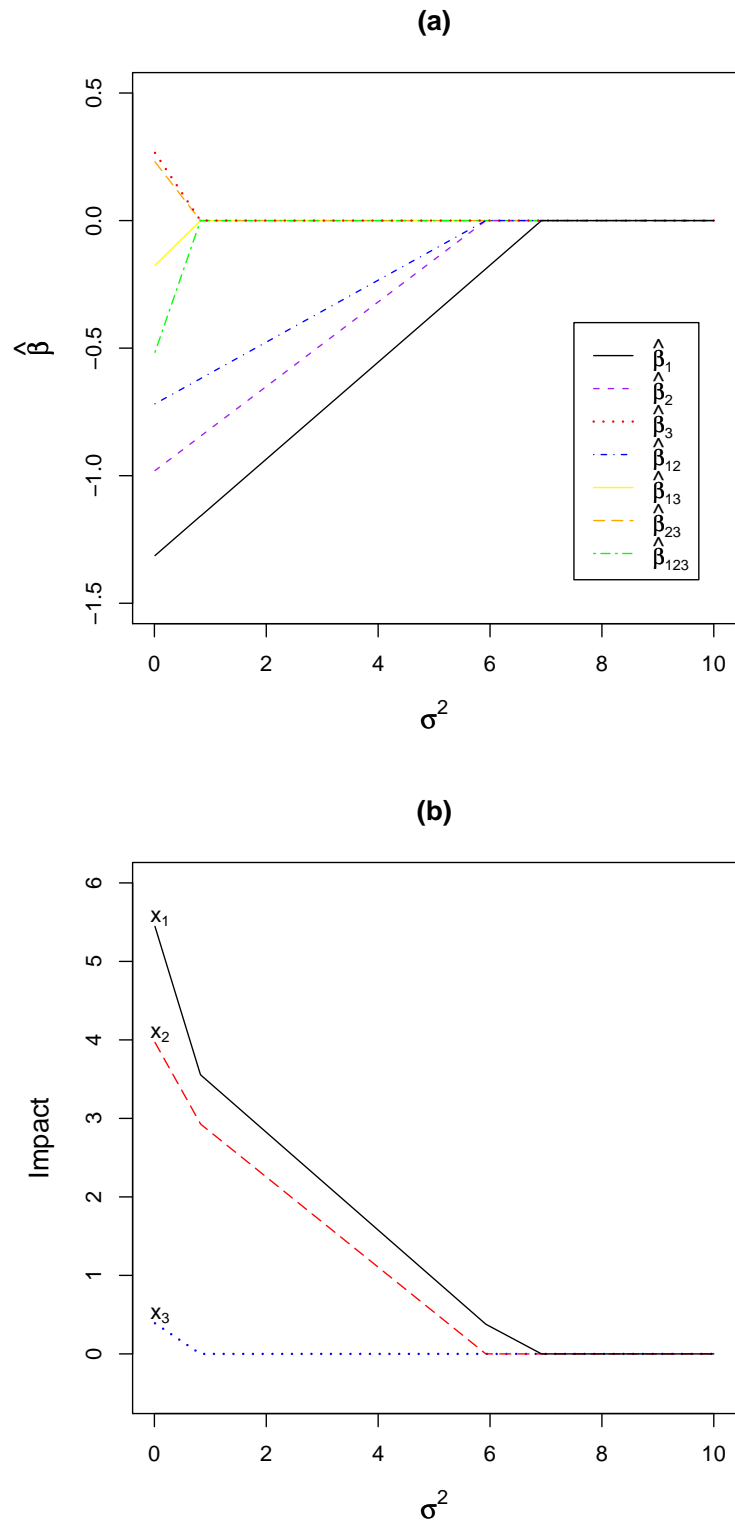
Figure 8: Bearing Experiment With Heredity Prior: (a) Coefficients (b) Impacts

A plot of λ as a function of z is provided in Figure 9(a). The objective is to find the $z_{\alpha/2}$ that minimizes the absolute difference between the empirical Bayes estimate and the estimate after using statistical testing. Under the null hypothesis, $z \sim \mathcal{N}(0,1)$. Thus, we minimize

$$\int_1^{z_{\alpha/2}} \{(1 - \frac{1}{z^2}) - 0\}\phi(z)\, dz + \int_{z_{\alpha/2}}^\infty \{1 - (1 - \frac{1}{z^2})\}\phi(z)\, dz,$$

where $\phi(z)$ is the standard normal density function. By differentiating with respect to $z_{\alpha/2}$ and equating to 0, we obtain

$$(1 - \frac{1}{z_{\alpha/2}^2})\phi(z_{\alpha/2}) - \frac{1}{z_{\alpha/2}^2}\phi(z_{\alpha/2}) = 0.$$

Solving, we obtain $z_{\alpha/2} = \sqrt{2}$. This corresponds to an α level of 15.73%. At this level, the empirical Bayes estimate of β_i is one half of the least squares estimate.

If σ can be estimated, then a t-statistic would be used for testing H_0: $\beta_i = 0$. Note that the optimal critical value remains the same as $\sqrt{2}$ irrespective of the distribution of the test statistic. Therefore, the optimal significance level in a t-test can be obtained by solving for α in $t_{\alpha/2,\nu} = \sqrt{2}$, where ν represents the degrees of freedom for the error. For $\nu = 1$, we obtain $\alpha = 0.3918$. This approaches 0.1573 as $\nu \to \infty$ (see Figure 9(b)).

Because of the popularity of statistical testing and its primacy in the analysis techniques described in many textbooks on the design and analysis of experiments, we envision that it will be continued to be used for many more years to come. Moreover, the procedure using statistical testing is easier to implement than the empirical Bayes procedure. So if an investigator prefers to apply statistical testing, we do recommend using the α level of 15%. Our derivation clearly demonstrates that this liberal level should be used irrespective of the number of effects being examined. Therefore, for optimization experiments, we additionally recommend against incorporating in procedures intended for multiple testing such as the Bonferronni correction method, studentized maximum modulus method, etc. There does exist previous work
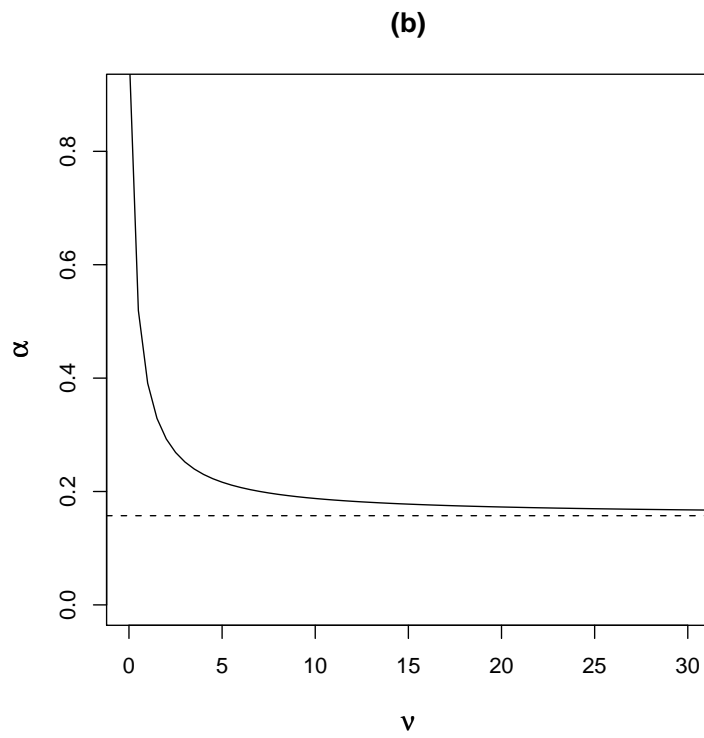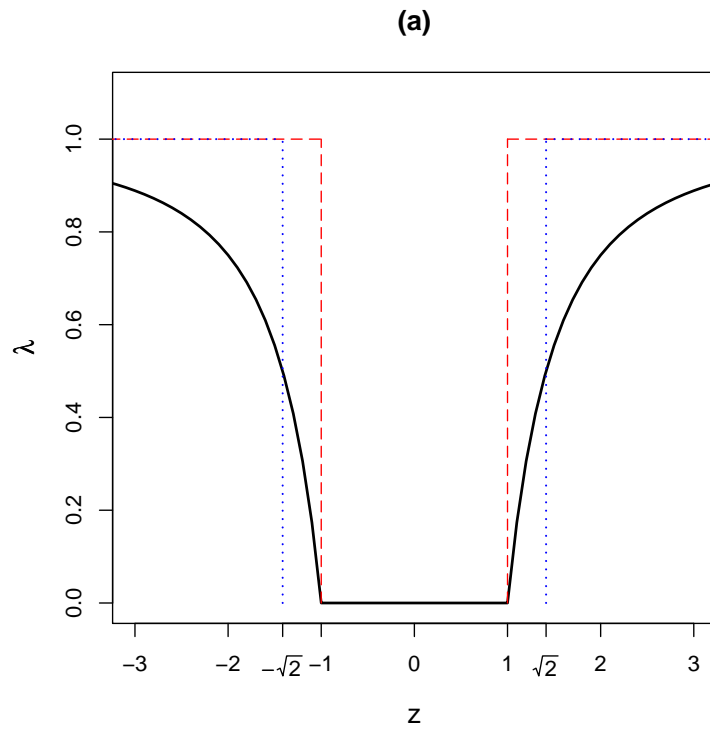
Figure 9: Testing as an Approximation: (a) Shrinkage Coefficient as a Function of Critical Values, (b) Optimal t-tests

in attempting to provide guidance in the choice of α-level for frequentist point null hypothesis testing. For example, Kennedy and Bancroft (1971) also make the case for the use of a more liberal, $0.10 \leq \alpha \leq 0.25$, significance level, in the context of the sequential tests involved in forward selection. However, Berger and Sellke (1987) demonstrates that in univariate point-null hypothesis testing, even fairly conservative choices for α, can lead to an unacceptably high posterior probability for the null in an objective Bayes setting.

2.6   TAGUCHI'S BETA COEFFICIENT METHOD

Taguchi (1987, chapter 19) criticized the use of statistical testing in experiments and proposed an intriguing method which he named the *beta coefficient method*. From his experience he found that the predicted value from the experiment is always an over estimate of the true value. Therefore, he suggested that the effects obtained from the experiment should be shrunk towards 0 before making the prediction. He denoted the shrinkage factor by the parameter β and so he named the method the beta coefficient method. But because the variable β is more commonly used for denoting the linear model parameters, we use different notation.

Taguchi developed his method using an analysis of variance model and sum of squares calculations, but for the consistency of exposition, we explain his method using the regression model set up used throughout this chapter. Let λ_i denote the shrinkage applied to the least squares estimate $\widetilde{\beta}_i$. The objective is to find the λ_i that minimizes the mean squared error $\mathbb{E}\{(\lambda_i\widetilde{\beta}_i - \beta_i)^2\}$. Because $\mathbb{E}(\widetilde{\beta}_i) = \beta_i$, we obtain

$$\mathbb{E}\{(\lambda_i\widetilde{\beta}_i - \beta_i)^2\} = \lambda_i^2 \, var(\widetilde{\beta}_i) + (1 - \lambda_i)^2\beta_i^2.$$

Differentiating with respect to λ_i and equating to 0, we obtain

$$\lambda_i = \frac{\beta_i^2}{\beta_i^2 + var(\widetilde{\beta}_i)} = 1 - \frac{var(\widetilde{\beta}_i)}{\beta_i^2 + var(\widetilde{\beta}_i)}.$$

If the columns in the model matrix are orthogonal, then $var(\widetilde{\beta}_i) = \sigma^2/n$. An unbiased estimate of $\beta_i^2 + \sigma^2/n$ is $\widetilde{\beta}_i^2$. Thus λ_i can be estimated by

$$\lambda_i = 1 - \frac{\widehat{\sigma^2}/n}{\widetilde{\beta}_i^2} = 1 - \frac{1}{t_i^2}. \tag{23}$$

Because λ_i must be nonnegative, modifying the estimate to $\lambda_i = (1 - 1/t_i^2)_+$ is required. This produces the shrinkage coefficient suggested by Taguchi. This is exactly the same as the empirical Bayes shrinkage coefficient in (21). Taguchi used sum of squares to derive the result and thus the shrinkage coefficient is obtained as $(1 - 1/F_i)_+$, where F_i is the F-ratio from the analysis of variance table. It is easy to show that $F_i = t_i^2$. Consequently, the two shrinkage coefficients are equivalent.

We note that replacing the numerator and denominator by unbiased estimators in the expression for λ_i would not generally produce an unbiased estimate of λ_i. Indeed, several other estimates are possible. For example, the maximum likelihood estimate of λ_i is

$$\lambda_i = 1 - \frac{\widehat{\sigma^2}/n}{\widetilde{\beta}_i^2 + \widehat{\sigma^2}/n} = 1 - \frac{1}{1 + t_i^2},$$

is a very different estimate. Taguchi does not sufficiently detail his line of reasoning for how he arrived at the estimate in (23), but he did seem to have the right intuition to obtain the correct estimate.

Taguchi (1987) predicted that his method would completely replace the statistical testing methods used in the analysis of experiments. However, it did not happen. We believe that the justification given through the empirical Bayes method will make this method more popular in the future. We also note that this empirical Bayes perspective admits an even more general procedure that can be used with any type of design (it need not be orthogonal) and that easily incorporates effect hierarchy and heredity. This should lead to better models and better decision making.

2.7   SIMULATION

We use simulation to investigate the properties of the proposed procedure for optimization experiments. In particular, we are interested in providing evidence to support our expectation that this method performs better than the usually applied statistical testing techniques. This is most easily revealed through simulation of main effects models. We are especially encouraged by how well the proposed methodology performs compared to the frequentist technique when σ^2 is not known.

2.7.1   Main Effects Modeling

Below, we consider the estimation of the main effects from a design that is a 12-run orthogonal array over 11 factors, with model matrix:

$$
\boldsymbol{U} = \begin{pmatrix}
1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 \\
1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 \\
1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 \\
1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 \\
1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 & -1 \\
1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 & 1 \\
1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 & 1 \\
1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 & 1 \\
1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 & -1 \\
1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 & 1 \\
1 & 1 & -1 & 1 & 1 & 1 & -1 & -1 & -1 & 1 & -1 & 1 \\
1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1
\end{pmatrix}.
$$

Models were simulated with the following mechanism:

$$
f(\beta_i | \eta_i) = \eta_i \mathcal{N}(0, \tau^2) + (1 - \eta_i)\mathcal{N}(0, 1) \qquad i = 1, \ldots, 11
$$

$$
\eta_i = \begin{cases} 1 & \text{with probability } 1 - \gamma \\ 0 & \text{with probability } \gamma. \end{cases}
$$

$$Y = \mu + \boldsymbol{U\beta} + \varepsilon \qquad \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

The use of a normal mixture model of this sort to represent a linear model with some combination of active and inactive effects is not uncommon in Bayes hierarchical modeling. For instance, Chipman, Hamada, and Wu (1997) uses a similar mixture model assumption for their model selection technique. Without loss of generality, we assume $\mu = 0$ and σ^2 to be known. Then, for each of these models we carry out estimation and variable selection in the traditional frequentist method, using statistical hypothesis testing. The significance levels of $\alpha = 0.0045$, $\alpha = 0.0500$, and $\alpha = 0.1573$ correspond to the Bonferroni adjustment to $\alpha = 0.05$ to properly account for simultaneous testing, the α-level required for declaring significance in many publications, and the level we would recommend as an approximation to the empirical Bayes procedure presented in this paper, respectively. In addition, results are presented for a variety of levels of the thresholding parameter, Δ. N=10,000 random models were generated for many different settings of σ^2, τ^2, and γ.

We assume that Y is a larger the better quality characteristic. Table 7 and Table 8 display some metrics for comparing the proposed procedure with the existing frequentist techniques for a couple of scenarios that could easily characterize some real experiments. For the moment we assume that the value of σ^2 is known. With 11 main effects and a parameter value of $\gamma = 0.2$, around two effects are expected to be active, characterizing factor sparsity. When an effect is active, its coefficient is drawn from a $\mathcal{N}(0, 1)$ distribution and should be much larger than the coefficient from an inactive effect which is drawn from a distribution that is very tightly concentrated around zero, $\mathcal{N}(0, \tau^2 = 0.001)$. We also consider models when several more factors should be active, $\gamma = 0.5$.

From these simulations, we intend to compare the performance of using the EB shrinkage estimator combined with the practical thresholding level Δ, with some typical frequentist hypothesis testing approaches. The different techniques are used

for estimation, parameter thresholding and selecting optimal factor settings.

The metrics by which we compare the procedures are: *% Improvement, Estimation Error*, and *Number of Active Effects*. By % Improvement we hope to quantify the quality of the guidance provided by each procedure for determining factor settings. For each run j, we have a true model for the response $y_j(\boldsymbol{x})$. Let $\boldsymbol{x}^\dagger$ represent the true optimal factor settings. We would choose these settings if we knew the true response function. Whereas $\boldsymbol{x}^*$ denote the factor settings we would choose based on the estimated model and thresholding technique we use. Then

$$\% \text{ Improvement} = \frac{\frac{1}{N} \sum_{j=1}^{N} y_j(\boldsymbol{x}^*) - y_j(\boldsymbol{0})}{\frac{1}{N} \sum_{j=1}^{N} y_j(\boldsymbol{x}^\dagger) - y_j(\boldsymbol{0})}.$$

So this ratio reflects the proportion of the maximum possible improvement in the response that is actually obtained on average with that technique. Estimation error is averaged over all of the model i coefficients in the following way:

$$\text{Estimation Error} = \sqrt{\frac{1}{N(n-1)} \sum_{j=1}^{N} \sum_{i=1}^{n-1} \left(\beta_{i,j} - \widehat{\beta}_{i,j} \right)^2}.$$

And finally, the number of active effects is the average over all of the simulation runs of the estimated effects that are nonzero. When an effect estimate is 0, either because of failure to reject H_0 in the statistical test, or by shrinkage, or practical significance thresholding, the setting selected is 0. For the purposes of these metrics, this is equivalent to determining the setting for that factor to be either $+1$ or -1 based on a "coin flip" or by arbitrarily setting it to its midpoint, if such a setting exists.

From these tables, it is quite clear that the settings selected when using the EB estimators, in particular when $\Delta = 0.0$, yield superior results with respect to the optimization experiment objective of improving the response that would be realized. This pattern is most pronounced when σ^2 is large. Here there is also a distinct pattern of slightly better performance for the proposed method with regard to criteria

Table 7: ME Simulation ($\gamma = 0.2$, $\tau^2 = 0.001$, σ^2 known)

% Improvement								
	Frequentist (α)			Practical Thresholding (Δ)					
σ^2	0.0045	0.0500	0.1573	0.0	0.2	0.4	0.6	0.8	1.0
0.0	1.00	1.00	1.00	1.00	0.88	0.87	0.85	0.82	0.78
0.5	0.74	0.81	0.84	0.86	0.85	0.83	0.81	0.78	0.74
1.0	0.62	0.74	0.79	0.83	0.81	0.79	0.77	0.74	0.71
2.0	0.46	0.63	0.72	0.77	0.76	0.74	0.71	0.69	0.66
5.0	0.23	0.43	0.55	0.64	0.63	0.61	0.59	0.57	0.54
10.0	0.10	0.27	0.41	0.52	0.51	0.49	0.47	0.46	0.44

Estimation Error								
	Frequentist (α)			Practical Thresholding (Δ)					
σ^2	0.0045	0.0500	0.1573	0.0	0.2	0.4	0.6	0.8	1.0
0.0	0.00	0.00	0.00	0.00	0.03	0.03	0.05	0.06	0.08
0.5	0.14	0.15	0.17	0.14	0.14	0.14	0.13	0.13	0.14
1.0	0.20	0.21	0.24	0.20	0.20	0.20	0.20	0.19	0.19
2.0	0.28	0.30	0.34	0.28	0.28	0.28	0.28	0.28	0.27
5.0	0.40	0.45	0.54	0.43	0.43	0.43	0.43	0.43	0.43
10.0	0.47	0.60	0.74	0.59	0.59	0.59	0.59	0.59	0.59

Number of Active Effects								
	Frequentist (α)			Practical Thresholding (Δ)					
σ^2	0.0045	0.0500	0.1573	0.0	0.2	0.4	0.6	0.8	1.0
0.0	11.00	11.00	11.00	11.00	2.03	1.83	1.66	1.50	1.34
0.5	1.28	1.97	3.11	4.67	3.56	2.60	1.95	1.54	1.30
1.0	0.98	1.73	2.94	4.51	3.73	2.97	2.31	1.81	1.44
2.0	0.67	1.48	2.73	4.37	3.77	3.20	2.68	2.20	1.79
5.0	0.31	1.08	2.35	4.08	3.68	3.28	2.91	2.55	2.21
10.0	0.16	0.84	2.10	3.88	3.59	3.31	3.02	2.75	2.48

Table 8: ME Simulation ($\gamma = 0.5$, $\tau^2 = 0.001$, σ^2 known)

	% Improvement								
	Frequentist (α)			Practical Thresholding (Δ)					
σ^2	0.0045	0.0500	0.1573	0.0	0.2	0.4	0.6	0.8	1.0
0.0	1.00	1.00	1.00	1.00	0.97	0.95	0.93	0.90	0.86
0.5	0.81	0.88	0.92	0.93	0.92	0.90	0.88	0.85	0.81
1.0	0.69	0.81	0.87	0.90	0.88	0.87	0.84	0.81	0.78
2.0	0.50	0.68	0.78	0.84	0.82	0.80	0.77	0.75	0.71
5.0	0.25	0.46	0.61	0.70	0.69	0.67	0.64	0.62	0.60
10.0	0.11	0.30	0.45	0.57	0.55	0.54	0.52	0.50	0.48

	Estimation Error								
	Frequentist (α)			Practical Thresholding (Δ)					
σ^2	0.0045	0.0500	0.1573	0.0	0.2	0.4	0.6	0.8	1.0
0.0	0.00	0.00	0.00	0.00	0.02	0.04	0.06	0.09	0.13
0.5	0.21	0.19	0.19	0.17	0.17	0.18	0.18	0.19	0.20
1.0	0.30	0.27	0.27	0.24	0.25	0.25	0.25	0.26	0.26
2.0	0.43	0.38	0.38	0.34	0.34	0.34	0.35	0.35	0.35
5.0	0.61	0.58	0.60	0.51	0.51	0.51	0.52	0.52	0.52
10.0	0.70	0.75	0.82	0.68	0.68	0.68	0.68	0.68	0.68

	Number of Active Effects								
	Frequentist (α)			Practical Thresholding (Δ)					
σ^2	0.0045	0.0500	0.1573	0.0	0.2	0.4	0.6	0.8	1.0
0.0	11.00	11.00	11.00	11.00	5.10	4.65	4.21	3.80	3.42
0.5	3.17	4.11	5.17	6.41	5.54	4.74	4.09	3.60	3.20
1.0	2.41	3.51	4.72	6.08	5.40	4.75	4.13	3.61	3.16
2.0	1.57	2.80	4.14	5.63	5.11	4.59	4.08	3.59	3.15
5.0	0.72	1.87	3.31	4.98	4.62	4.24	3.86	3.50	3.16
10.0	0.33	1.31	2.74	4.49	4.22	3.94	3.66	3.38	3.12

of Estimation Error. However, the third panel demonstrates that if the goal is a parsimonious model for prediction, then the proposed method with $\Delta = 0$ is clearly not the best choice. Figure 10 illustrates a pattern that we expect to observe in real experiments. With even a very small choice for Δ, say 0.2, the model is of the appropriate size and not much is lost in % Improvement. In fact, in Section 2.8 we observe that in the router bit experiment for even small values of Δ, the number of practically significant factors is small, so the complexity of the model that is to be optimized is dramatically reduced. This consequently increases the number of practically insignificant factors which can be set to levels that reduce costs.

For the simple case when $\sigma^2 = 0$, many of the values in the tables and all of Figure 10 could have been calculated analytically, or at least via a very simple numerical integration, rather than simulation. This is a consequence of the fact that the underlying probability distribution is a mixture of half-normal distributions. Since $y(\mathbf{0}) = 0$ and we are only considering main effects models, we could instead calculate:

$$\% \text{ Improvement} = \frac{\gamma e^{-\frac{\Delta^2}{8}} + (1 - \gamma)\tau e^{-\frac{\Delta^2}{8\tau^2}}}{\gamma + \tau (1 - \gamma)}.$$

In the denominator, we use the well known expression for the mean of a half-normal random variable. See, for instance Johnson, Kotz, and Balakrishnan (1994). That is, more generally, when $X \sim \mathcal{N}(\mu, \sigma^2)$, then

$$\mathbb{E}(|X|) = \mu + \sqrt{\frac{2}{\pi}}\sigma.$$

When $\sigma^2 = 0$, we also note that:

$$\text{Estimation Error} = \sqrt{\int_0^{\frac{\Delta}{2}} 2x^2 \left(\gamma\phi\left(x\right) + (1 - \gamma)\frac{1}{\tau}\phi\left(\frac{x}{\tau}\right)\right) dx},$$

and

$$\text{Number of Effects} = 22 \left(1 - \gamma\Phi\left(\frac{\Delta}{2}\right) - (1 - \gamma)\Phi\left(\frac{\Delta}{2\tau}\right)\right),$$

where $\phi(\cdot)$ represents the standard normal density function and $\Phi(\cdot)$ represents the standard normal cumulative distribution function. When $\sigma^2 > 0$, the tabulated values

could also be expressed as integrals that could be evaluated numerically. However, the expressions are more complex. And when σ^2 is unknown, the expressions and numerical integration would be even more complicated. So in every case, we approximate these values through the simulation explicitly described above. The high number of simulation runs, N=10,000, does not take much time on today's standard desktop computer.

That the same patterns so far revealed in this section are reproducible for different combinations of γ and τ^2 is illustrated in the following "interaction" plots. The same three metrics are now plotted as a function of γ. Each line represents a different value of τ^2, for either of the usual frequentist statistical test of significance procedure or the empirical Bayes shrinkage estimation procedure. In these plots, $\sigma^2 = 1$ and is assumed to be known. The statistical significance level of $\alpha = 0.05$ and practical thresholding rule of $\Delta = 0$ are used in these three plots. Notice that for virtually any values for γ and τ^2, the % Improvement using the settings suggested by the EB parameter estimates is superior to that when using the statistical z-test.

2.7.2 Main Effects With Unknown σ^2

In this section we examine the case when σ^2 is not known but may be estimated from the data. The important distinction we make here with the simulation in the previous section is that since we are estimating σ^2, the frequentist tests of statistical significance involve *t-statistics* rather than *z-statistics*. For simplicity, we assume that m center points are incorporated into each experimental design. The use of center points in this manner is illustrated in an exercise from Wu and Hamada (2000, page. 146).

In Table 9 and Table 10 we provide the simulation results for when $m = 3$ that correspond to Table 7 and Table 8, respectively. The advantage of the proposed factor optimization procedure over the procedure that involves a frequentist t-test with small

(a)

(b)
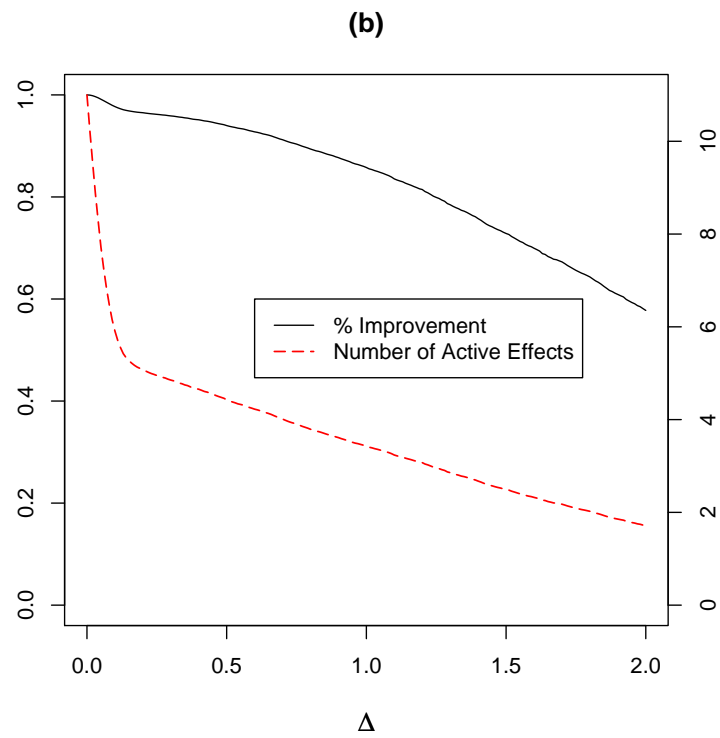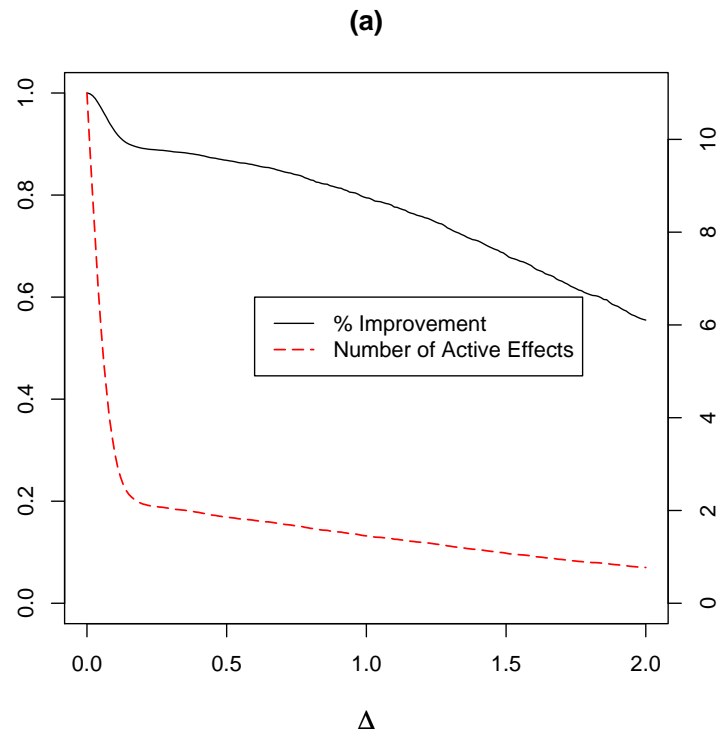
Figure 10: ME Simulation, Choosing Δ ($\sigma^2 = 0$, $\tau^2 = 0.001$) (a) $\gamma = 0.2$ (b) $\gamma = 0.5$
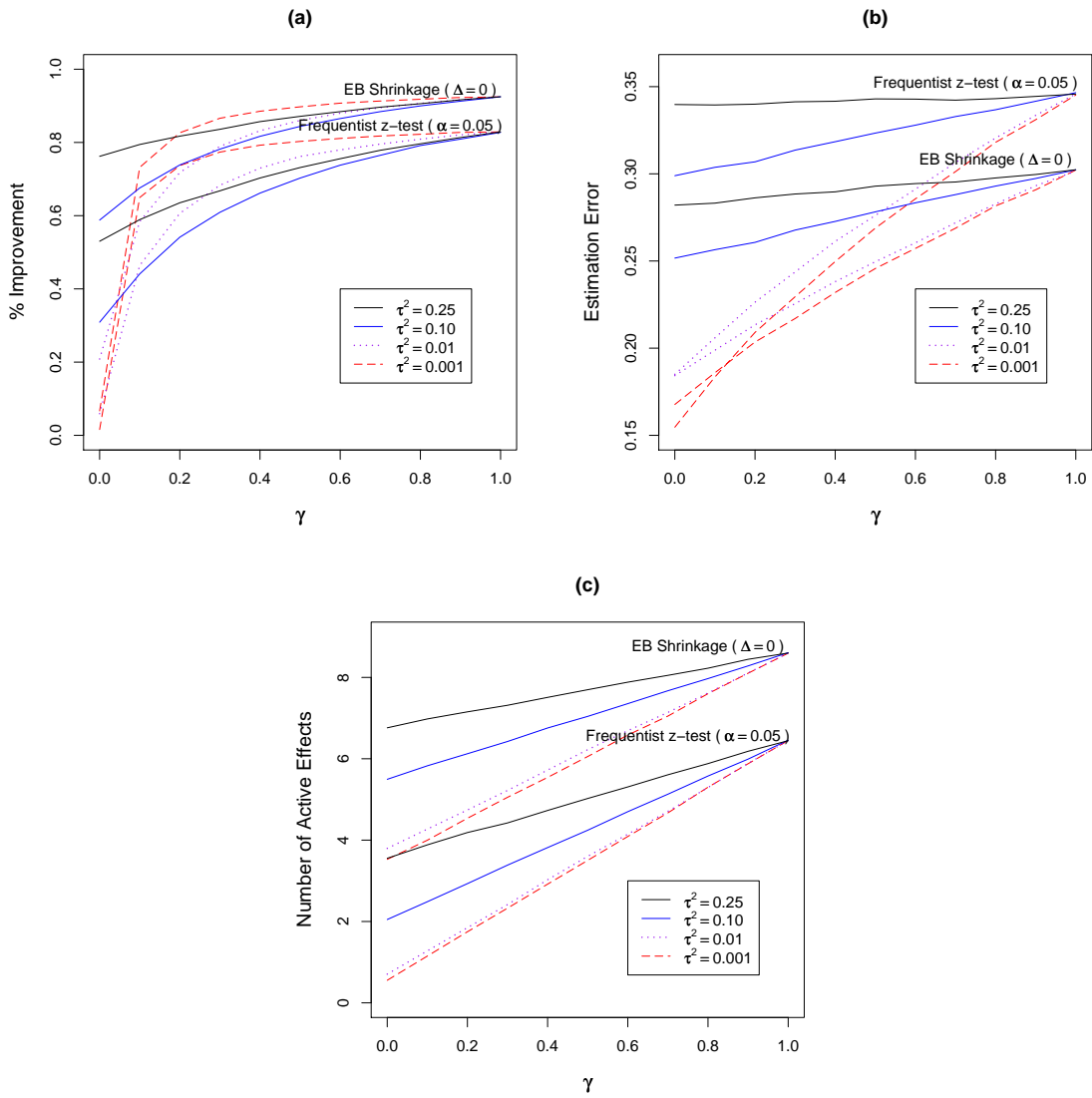
Figure 11: ME Simulation, Varying γ and τ^2 ($\sigma^2 = 1$) (a) Ratio, (b) Estimation Error, (c) Model Size

ν is profound. For example, when $m = 3$, $\gamma = 0.50$, $\tau^2 = 0.001$, and $\sigma^2 = 1.0$, on average, we would expect to obtain 90% of the true optimum, by choosing the settings that are dictated by our EB estimates. Whereas, in the frequentist setting, the usual t-test would only yield 54% of the true optimal response, on average.

However, as $m \to \infty$ the advantage of the proposed procedure over the utilization of a frequentist test quickly begins to resemble that smaller, but distinct, advantage demonstrated in the previous section. This is illustrated in the plots of % Improvement provided in Figure 12. Additional simulation results appear in Appendix C.

2.8   ROUTER BIT EXPERIMENT, REVISITED

Once again, recall the Router Bit Experiment of Phadke (1989). This experiment is used to illustrate the application of the induced, heredity prior in Section 1.5, above, and is also analyzed in Wu and Hamada (2000). The experiment is a 32 run regular fraction of a $2^7 \times 4^2$ factorial design, where the objective is to maximize the lifetime of router bits. Therefore, we consider this problem to be an optimization experiment. We achieve the objective of determining optimal factor settings by applying the methodology presented in this chapter. That is, we find the optimal settings that minimize the total cost (16). Here, we primarily focus on the objective of maximizing the expected bit lifetime.

To find the optimal factor settings, we calculate the EB estimates of the 2,048 full factorial model parameters, optimize the response with respect to all of the factors, and apply a practical significance level thresholding rule Δ to the factor impacts. For the factors that were deemed practically insignificant, the setting adopted is the current factor setting.

Unfortunately, when the model matrix is not orthogonal, there is not a general, convenient result like Proposition 4 for expressing the EB parameter estimates in

Table 9: ME Simulation ($\gamma = 0.2$, $\tau^2 = 0.001$, $m = 3$ centerpoints)

% Improvement								
	Frequentist (α)			Practical Thresholding (Δ)					
σ^2	0.0045	0.0500	0.1573	0.0	0.2	0.4	0.6	0.8	1.0
0.0	1.00	1.00	1.00	1.00	0.88	0.87	0.85	0.82	0.78
0.5	0.16	0.64	0.80	0.87	0.85	0.83	0.81	0.78	0.75
1.0	0.09	0.50	0.72	0.83	0.82	0.80	0.78	0.75	0.72
2.0	0.04	0.35	0.61	0.77	0.76	0.74	0.72	0.69	0.66
5.0	0.02	0.19	0.42	0.65	0.64	0.62	0.60	0.58	0.55
10.0	0.01	0.12	0.31	0.54	0.53	0.51	0.50	0.48	0.47

Estimation Error								
	Frequentist (α)			Practical Thresholding (Δ)					
σ^2	0.0045	0.0500	0.1573	0.0	0.2	0.4	0.6	0.8	1.0
0.0	0.00	0.00	0.00	0.00	0.03	0.04	0.05	0.06	0.08
0.5	0.39	0.20	0.17	0.16	0.16	0.15	0.14	0.14	0.14
1.0	0.42	0.28	0.23	0.22	0.22	0.22	0.21	0.20	0.20
2.0	0.44	0.35	0.33	0.31	0.31	0.31	0.31	0.30	0.29
5.0	0.44	0.45	0.50	0.48	0.48	0.48	0.48	0.48	0.47
10.0	0.46	0.54	0.67	0.67	0.67	0.67	0.67	0.67	0.67

Number of Active Effects								
	Frequentist (α)			Practical Thresholding (Δ)					
σ^2	0.0045	0.0500	0.1573	0.0	0.2	0.4	0.6	0.8	1.0
0.0	11.00	11.00	11.00	11.00	2.04	1.85	1.68	1.52	1.36
0.5	0.26	1.50	2.93	5.62	4.28	3.03	2.15	1.63	1.35
1.0	0.16	1.20	2.70	5.48	4.52	3.54	2.68	2.03	1.58
2.0	0.10	0.95	2.41	5.30	4.62	3.90	3.20	2.57	2.05
5.0	0.07	0.72	2.09	5.08	4.63	4.16	3.68	3.21	2.76
10.0	0.06	0.66	2.00	4.98	4.66	4.33	3.98	3.63	3.28

Table 10: ME Simulation ($\gamma = 0.5$, $\tau^2 = 0.001$, $m = 3$ centerpoints)

% Improvement								
	Frequentist (α)			Practical Thresholding (Δ)					
σ^2	0.0045	0.0500	0.1573	0.0	0.2	0.4	0.6	0.8	1.0
0.0	1.00	1.00	1.00	1.00	0.97	0.95	0.93	0.90	0.86
0.5	0.17	0.69	0.87	0.93	0.92	0.91	0.88	0.85	0.81
1.0	0.10	0.54	0.78	0.90	0.88	0.87	0.84	0.81	0.78
2.0	0.05	0.39	0.67	0.84	0.83	0.81	0.78	0.76	0.72
5.0	0.02	0.22	0.47	0.71	0.70	0.68	0.66	0.64	0.61
10.0	0.01	0.14	0.34	0.58	0.57	0.56	0.54	0.53	0.51

Estimation Error								
	Frequentist (α)			Practical Thresholding (Δ)					
σ^2	0.0045	0.0500	0.1573	0.0	0.2	0.4	0.6	0.8	1.0
0.0	0.00	0.00	0.00	0.00	0.02	0.04	0.06	0.09	0.13
0.5	0.61	0.30	0.20	0.18	0.18	0.18	0.19	0.19	0.21
1.0	0.66	0.42	0.30	0.26	0.26	0.26	0.26	0.26	0.27
2.0	0.68	0.53	0.41	0.36	0.36	0.36	0.36	0.37	0.37
5.0	0.70	0.65	0.61	0.55	0.55	0.55	0.55	0.55	0.55
10.0	0.71	0.73	0.79	0.74	0.74	0.74	0.74	0.74	0.74

| Number of Active Effects								
	Frequentist (α)			Practical Thresholding (Δ)					
σ^2	0.0045	0.0500	0.1573	0.0	0.2	0.4	0.6	0.8	1.0
0.0	11.00	11.00	11.00	11.00	5.10	4.66	4.24	3.82	3.43
0.5	0.56	2.92	4.74	7.07	6.04	5.05	4.24	3.66	3.23
1.0	0.32	2.20	4.15	6.76	5.98	5.18	4.43	3.78	3.25
2.0	0.19	1.64	3.58	6.42	5.82	5.19	4.56	3.98	3.45
5.0	0.10	1.06	2.80	5.89	5.48	5.05	4.60	4.15	3.71
10.0	0.08	0.85	2.41	5.49	5.19	4.87	4.54	4.20	3.87
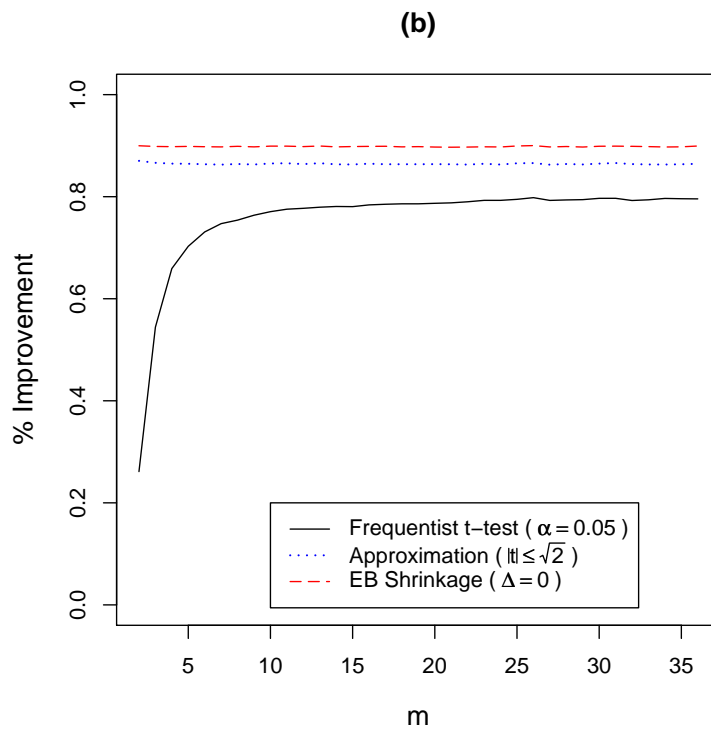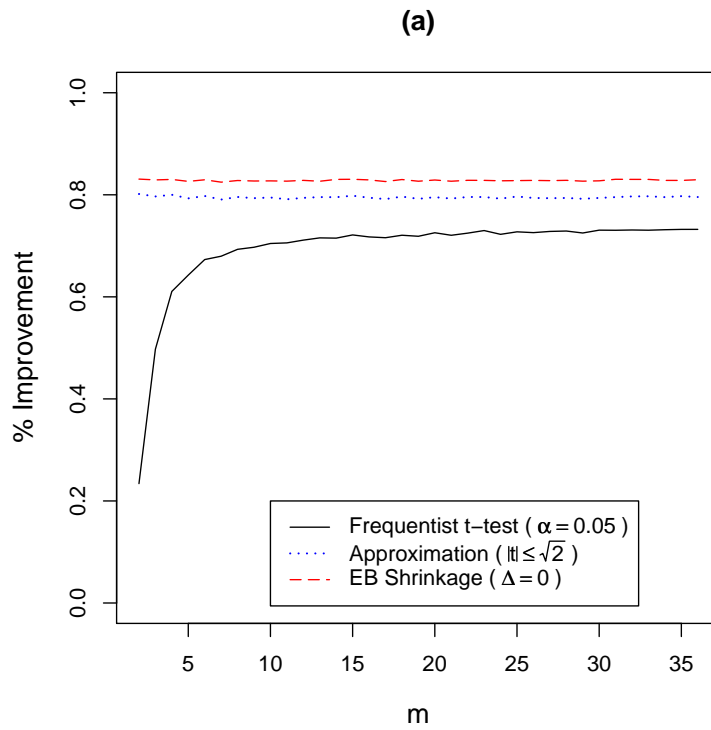
Figure 12: % Improvement as $m \to \infty$ for (a) $\sigma^2 = 1.0$, $\gamma = 0.25$, and $\tau^2 = 0.001$ (b) $\sigma^2 = 1.0$, $\gamma = 0.50$, and $\tau^2 = 0.001$

terms of shrinkage factors multiplied by the least squares estimates.

There do exist some additional scenarios when an expression of the form $\widehat{\beta}_i = \lambda_i \widetilde{\beta}_i$ can be found. For instance, see Joseph (2006), expression (18), for estimators of this form for each of the parameters when the experiment is a positive, regular fraction of a 2^p design.

Here the model matrix is $32 \times 2{,}048$. However, as described at the beginning of Section 2.4, we may still use the EB estimator obtained from the posterior mean (18), after plugging in the covariance matrix parameters obtained from maximizing the integrated likelihood.

The sampling error, σ^2 is unknown, but we assume it is practically limited by the precision of the observations. That is, the exact times the router bits failed during the experiment are not known. The bits are only inspected at regular intervals, after each 100 inches of cutting (in the X-Y plane). The recorded lifetimes, y_i, are the midpoint of that interval during which the bit is determined to have failed, in units of 100s of inches. So suppose that with high probability the recorded lifetime is within 50 inches of the true lifetime for that particular bit. That is,

$$P\left(\mu_{y_i} - 0.5 < Y_i < \mu_{y_i} + 0.5\right) > 0.9973$$

This statement implies that there may be other, very minor, sources of measurement error, besides the interval censoring, and that $\sigma < \frac{1}{6}$. We perform the analysis twice: first with $\sigma^2 = 0$ and then with $\sigma^2 = \frac{1}{36}$.

From Figure 13 it is made apparent how even a small practical significance level, Δ, can dramatically simplify the model that needs to be considered in the $\mathbb{E}\{L(Y)\}$ portion of (16).

Recall that in Section 1.5, the analysis is performed with $\sigma^2 = 0$. We obtain EB estimates of: $\widehat{\mu} = 5.8125$, $\widehat{\sigma_0^2} = 74.9462$, and

$$\widehat{\boldsymbol{\rho}} = (0.99, 0.99, 0.99, 0.71, 0.99, 0.99, 0.60, 0.09, 0.56)'.$$

73

We then obtain the EB estimates for the 2,048 full factorial model parameters. Note that since the two four level factors are qualitative, optimizing $\widehat{y}(\boldsymbol{x})$ is a trivial exercise of evaluating $\widehat{y}(\boldsymbol{x})$, at all $\boldsymbol{x} \in \mathcal{X}$, the 2,048 discrete design points, and comparing.
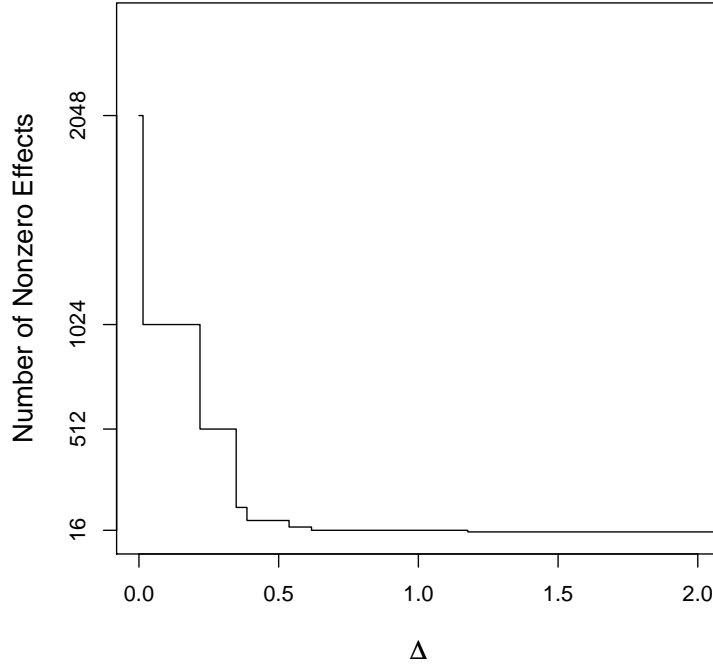
Figure 13: Router Bit Experiment: Model Size as Δ Increases

Factor E represents a variable "spindle position" which does not require us to determine its setting. The optimal settings for the remaining factors, under the full factorial parameter model, are:

$$A = -, B = -, C = +, D = 4, F = -, G = -, H = +, J = + \tag{24}$$

Now suppose that we are interested in making a factor setting change, as long as it offers at least a 10% improvement in the expected lifetime (at the current settings). This implies

$$\Delta = 0.10\widehat{y} = 0.10(3.5) = 0.35.$$

Recall that this value for Δ corresponds to a marginal lifetime of 35 inches of cutting. In order to apply the practical significance level, Δ, we wish to identify the largest set of insignificant effects that exists, such that all of its subsets are also insignificant, as defined in (17). Below, we systematically identify the set $\mathcal{S}^*$ for the Router Bit Experiment. First, after examining the individual factor impacts,

$$imp(\boldsymbol{x}) = (0.39, 0.22, 0.54, 15.54, 0.35, 0.01, 0.75, 0.62, 1.18),$$

we discover that the factor, D, has an enormous impact on router bit lifetime at the optimum, whereas the factors J, G, H, C, A, and E are merely practically significant. The subset $\{B, E, F\}$ has the potential to be practically insignificant. Recall that a k-factor combined impact has to be greater than $k\Delta$ for us to identify that something in that subset of factors as practically significant. So we could evaluate the combined impacts of all subsets of $\{B, E, F\}$ to determine the set $\mathcal{S}^*$, but instead, here we will illustrate the application of Algorithm 1.

$$\mathcal{S}^* = \varnothing$$

Step 1

$$imp(\boldsymbol{x}) = (0.39, 0.22, 0.54, 15.54, 0.35, 0.01, 0.75, 0.62, 1.18)$$

$$imp(F) - imp(\varnothing) = 0.01 < 0.35 \text{ (continue)}$$

$$\mathcal{S}^* = \{F\}$$

$$\mathcal{X} \setminus \mathcal{S}^* = (A, B, C, D, E, G, H, J)$$

Step 2

$$imp(F, \boldsymbol{x}) = (0.41, 0.23, 0.55, 15.58, 0.37, 0.82, 0.94, 1.18)$$

$$imp(F, B) - imp(F) = 0.22 < 0.35 \text{ (continue)}$$

$$\mathcal{S}^* = \{B, F\}$$

$$\mathcal{X} \setminus \mathcal{S}^* = (A, C, D, E, G, H, J)$$

Step 3

$$imp(F, B, \boldsymbol{x}) \quad = \quad (0.63, 0.77, 15.79, 0.58, 1.23, 0.98, 1.71)$$

$$imp(F, B, E) - imp(F, B) \quad = \quad 0.353 > 0.35 \text{ (STOP)}$$

$$\mathcal{S}^* \quad = \quad \{B, F\}$$

The results of the algorithm indicate that A, B, C, D, E, G, H, and J are practically significant. As a result, we suggest that these factors be set to the levels determined in (24). The remaining factors should have their settings determined by minimizing $C(\boldsymbol{x})$ in (16). So for minimizing $\mathbb{E}\{L(Y)\}$ in (16), the settings

$$A = -, B = -, C = +, D = 4, G = -, H = +, J = +$$

are recommended.

The settings for D, G, H, and J are consistent with the recommended settings in the analysis of Wu and Hamada (2000). However, only the effects D_2, G, J, and the interactions GJ and D_2H appear in their model for predicted lifetime. As a result there is ambiguity in the optimal setting for the four level factor D. This factor could be set to its level 1 or its level 4, to attain the same predicted lifetime. In order to resolve this uncertainty in the setting for factor D, Wu and Hamada (2000, page 270) employ the $D \times H$ interaction plot. From this plot, it is quite clear that $D = 4$ and $H = +$ is the optimal setting.

With the procedure discussed here, we were able to identify these optimal setting for factors D and H automatically. In addition, here we have applied a criteria with a practical connection to the optimization problem at hand, for when to consider changing a factor from its most inexpensive operating conditions. In addition, we anticipate that in selecting optimal settings for two additional factors, A and B, we may realize additional improvement in the mean lifetime.

When we change to $\sigma^2 = \frac{1}{36}$, the numerical results are virtually the same as those observed with $\sigma^2 = 0$. The recommended factor settings are identical to those reported here for $\sigma^2 = 0$.

2.9   CONCLUSIONS

The deep groove bearing design optimization example of Hellstrand (1989) illustrates a common and profound challenge encountered by quality technology practitioners. There are some undesirable practical consequences associated with the rigorous application of frequentist statistical hypothesis testing procedures that can prevent obtaining sufficient guidance in the design process. As is often the case, cost constraints keep the run size of an experiment quite small. In this particular example, a small run size may be to blame for not being able to conclude from a standard statistical test that two of the three factors are indeed significant. Unfortunately, the usual recommendation to just "collect more data" is usually not a practical solution. The engineer may have to make decisions with just the data that is presently available.

Another difficulty with statistical testing is that there seems to be a blind devotion to the use of an $\alpha = 0.05$ significance level, without much reflection on what this actually means and whether it has any practical connection with the problem at hand. In fact, if we are to rigorously adhere to the meaning of a test of significance at the $\alpha = 0.05$ level, then we would have to apply the correct simultaneous testing procedure when we examine the size of multiple factorial effects; thereby magnifying the probability we will be unable to identify any significant effects.

In an optimization experiment, the sole objective is determining the particular factor settings that will yield a desired response. In such a situation, we should be able to identify an amount of improvement in the response that is not large enough to be of practical significance. Thus, *practical significance* provides a much more meaningful criteria for determining whether changing a factor's setting is "worth it" than does an α-level. Further, when we focus on the objective of determining optimal factor settings, we might be able to ignore other metrics for evaluating our estimation and model selection procedure.

The procedure we recommend for the analysis of optimization experiments centers

around an overall objective function which balances quality and cost. We suggest the empircal Bayes estimator presented in Joseph (2006) that has many desirable properties. It shrinks the coefficients and incorporates the effect heredity principle. Based on these estimates, we may find the optimal settings for the factors. Further, we may calculate the impact that a factor level change can have near this optimal and determine whether this is large enough to be of practical interest.

There are special cases to the empirical Bayes estimator discussed herein, that have received some previous attention. In particular, connections are drawn to the so-called James-Stein estimator as well as the Beta Coefficient Method of Taguchi. And for those that are bound to using a frequentist point-null test, we suggest an α-level that serves as an approximation to using the recommended procedure.

The simulation results provide support for the conclusion that the recommended procedure is superior to frequentist testing for identifying factor settings that, on average, yield response values closer to our objective without unduly increasing the cost. This is the goal of optimization experiments. Finally, we come full-circle by revisiting the router bit experiment that is analyzed in Chapter 1, illustrating the application of the techniques discussed in this chapter to this real world example.

2.10 CONTRIBUTIONS

The research described in this chapter contributes to the body of knowledge in the Design and Analysis of Experiments in the following ways:

1. Through real examples and simulation, it is shown that the widely used statistical tests of hypothesis are not appropriate for optimization experiments.

2. An alternative analysis, using empirical Bayes methods, is proposed. Its connections to James-Stein estimation and Taguchi's Beta Coefficient Method are established.

3. Two concepts, *practical significance level* and *factor impacts*, are introduced as tools for obtaining optimal factor settings.

APPENDIX A

PROOFS

Proof of Theorem 1

We first need to establish a general result concerning the construction of our correlation matrices over all p factors. Observe that for the following result to hold, that we can define the $(m_1 m_2 \cdots m_p) \times (m_1 m_2 \cdots m_p)$ full factorial model matrix over all p factors, $\boldsymbol{U}$, using whatever coding scheme we desire. Suppose we construct our full model matrix via a Kronecker product of the individual factor model matrices, taken in increasing order of the frequency the levels change. The run order for the full factorial design corresponds to one where the first factor's levels are changing the slowest and the pth factor's levels are changing the quickest:

$$\boldsymbol{U} = \boldsymbol{U}_1 \otimes \boldsymbol{U}_2 \otimes \cdots \otimes \boldsymbol{U}_p = \bigotimes_{j=1}^{p} \boldsymbol{U}_j.$$

Now, let $\boldsymbol{\Psi}$ denote the correlation matrix corresponding to the full factorial design over all p factors. The $m_j \times m_j$ correlation matrix corresponding to factor j denoted by $\boldsymbol{\Psi}_j$ will have the general structure of a symmetric Toeplitz matrix due to the stationarity assumption imposed on the Gaussian process in each factor:

$$\boldsymbol{\Psi}_j = \begin{pmatrix} 1 & \psi_j(1) & \dots & \psi_j(m_j - 1) \\ \psi_j(1) & 1 & \ddots & \psi_j(m_j - 2) \\ \vdots & \ddots & \ddots & \vdots \\ \psi_j(m_j - 1) & \psi_j(m_j - 2) & \dots & 1 \end{pmatrix}. \qquad (25)$$

Then $\boldsymbol{\Psi}$ has a convenient block symmetric structure. Let the matrix $\boldsymbol{\Psi}_{(i)}$ represent the correlation matrix for the full factorial design over the last $p - i$ factors. Then,

since the first factor's levels are changing the slowest, and we have assumed a product correlation function structure, the correlation matrix has the following block form:

$$
\boldsymbol{\Psi} = \begin{pmatrix}
\boldsymbol{\Psi}_{(1)} & \psi_1(1)\boldsymbol{\Psi}_{(1)} & \cdots & \psi_p(m_1-1)\boldsymbol{\Psi}_{(1)} \\
\psi_1(1)\boldsymbol{\Psi}_{(1)} & \boldsymbol{\Psi}_{(1)} & \ddots & \psi_1(m_1-2)\boldsymbol{\Psi}_{(1)} \\
\vdots & \ddots & \ddots & \vdots \\
\psi_1(m_1-1)\boldsymbol{\Psi}_{(1)} & \psi_1(m_1-2)\boldsymbol{\Psi}_{(1)} & \cdots & \boldsymbol{\Psi}_{(1)}
\end{pmatrix}, \quad (26)
$$

where each of the blocks are $(m_2 m_3 \cdots m_p) \times (m_2 m_3 \cdots m_p)$. This matrix follows from the fact that the first $(m_2 m_3 \cdots m_p)$ runs in the full factorial design only differ among the last $p-1$ factors, in the same way the full factorial design with $p-1$ factors varies. Each run in the second block of $(m_2 m_3 \cdots m_p)$ runs differs from the first run in the full factorial design by one level in the first factor, and then in the same way as the full factorial design differs among the last $p-1$ factors, etc. Hence $\boldsymbol{\Psi} = \boldsymbol{\Psi}_1 \otimes \boldsymbol{\Psi}_{(1)}$. Noting that $\boldsymbol{\Psi}_{(p-1)} = \boldsymbol{\Psi}_p$, we obtain

$$
\boldsymbol{\Psi} = \boldsymbol{\Psi}_1 \otimes \boldsymbol{\Psi}_{(1)} = \boldsymbol{\Psi}_1 \otimes \boldsymbol{\Psi}_2 \otimes \boldsymbol{\Psi}_{(2)} = \cdots = \boldsymbol{\Psi}_1 \otimes \cdots \otimes \boldsymbol{\Psi}_p = \bigotimes_{j=1}^{p} \boldsymbol{\Psi}_j.
$$

Now through the properties of the Kronecker product operator, we can prove Theorem 1:

$$
\begin{aligned}
var(\boldsymbol{\beta}) &= \sigma_0^2 \boldsymbol{U}^{-1} \boldsymbol{\Psi} (\boldsymbol{U}^{-1})' \\
&= \sigma_0^2 (\bigotimes_{j=1}^{p} \boldsymbol{U}_j)^{-1} \bigotimes_{j=1}^{p} \boldsymbol{\Psi}_j \left((\bigotimes_{j=1}^{p} \boldsymbol{U}_j)^{-1} \right)' \\
&= \sigma_0^2 \bigotimes_{j=1}^{p} \boldsymbol{U}_j^{-1} \bigotimes_{j=1}^{p} \boldsymbol{\Psi}_j \left(\bigotimes_{j=1}^{p} \boldsymbol{U}_j^{-1} \right)' \\
&= \sigma_0^2 \bigotimes_{j=1}^{p} \boldsymbol{U}_j^{-1} \bigotimes_{j=1}^{p} \boldsymbol{\Psi}_j \bigotimes_{j=1}^{p} (\boldsymbol{U}_j^{-1})' \\
&= \sigma_f^2 \bigotimes_{j=1}^{p} \boldsymbol{U}_j^{-1} \boldsymbol{\Psi}_j (\boldsymbol{U}_j^{-1})'.
\end{aligned}
$$

$\diamond$

Proof of Equation (5)

We have $\boldsymbol{\Psi}_j = (1 - \rho_j)\boldsymbol{I}_{m_j} + \rho_j \boldsymbol{J}_{m_j}$, where $\boldsymbol{J}_{m_j}$ is a $m_j \times m_j$ square matrix of 1's.

Then:

$$
\begin{aligned}
\boldsymbol{U}'_j \boldsymbol{\Psi}_j \boldsymbol{U}_j &= (1 - \rho_j)\boldsymbol{U}'_j \boldsymbol{U}_j + \rho_j \boldsymbol{U}'_j \boldsymbol{J}_{m_j} \boldsymbol{U}_j \\[2mm]
&= m_j(1 - \rho_j)I_{m_j} + \rho_j \boldsymbol{U}'_j
\begin{pmatrix}
m_j & 0 & \cdots & 0 \\
\vdots & \vdots & \cdots & 0 \\
m_j & 0 & \cdots & 0
\end{pmatrix} \\[2mm]
&= m_j(1 - \rho_j)I_{m_j} + \rho_j
\begin{pmatrix}
m_j^2 & 0 & \cdots & 0 \\
0 & 0 & \cdots & 0 \\
\vdots & \vdots & \cdots & 0 \\
0 & 0 & \cdots & 0
\end{pmatrix} \\[2mm]
&= m_j
\begin{pmatrix}
1 - \rho_j + m_j \rho_j & 0 & \cdots & 0 \\
0 & 1 - \rho_j & \cdots & 0 \\
\vdots & & \ddots & \vdots \\
0 & 0 & 0 & 1 - \rho_j
\end{pmatrix}.
\end{aligned}
$$

$\diamondsuit$

Proof of Equation (14)

Let $\boldsymbol{\Sigma} = \boldsymbol{U}^{-1}\boldsymbol{\Psi}(\boldsymbol{U}^{-1})'$. Then $var(\beta_0) = \tau_0^2 = \sigma_0^2 \boldsymbol{\Sigma}_{1,1}$. But since $\boldsymbol{U}$ is the orthogonal full factorial model matrix with the leading column of $\boldsymbol{U}$ assumed to be $\mathbf{1}_q$, we have that $\boldsymbol{U}^{-1}$ has as its leading row: $\frac{1}{q}\mathbf{1}_q'$. So that:

$$
\frac{\tau_0^2}{\sigma_0^2} = \frac{\mathbf{1}_q' \boldsymbol{\Psi} \mathbf{1}_q}{q^2} = \frac{sum(\boldsymbol{\Psi})}{q^2} = \frac{sum(\bigotimes_{j=1}^{p} \boldsymbol{\Psi}_j)}{q^2} = \frac{\prod_{j=1}^{p} sum(\boldsymbol{\Psi}_j)}{q^2}.
$$

$\diamondsuit$

APPENDIX B

A GENERALIZATION OF THE EMPIRICAL BAYES ESTIMATE

PROPOSITION 4 *Let*

$$\boldsymbol{y} = \mu_0 \boldsymbol{1}_n + \boldsymbol{U}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n)$$

and

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}), \qquad where \ \boldsymbol{\Sigma} = diag(\tau_0^2, \ldots, \tau_{s-1}^2).$$

$\boldsymbol{U}$ *is an $n \times s$ matrix such that $\boldsymbol{U}'\boldsymbol{U} = n\boldsymbol{I}_s$. Then the empirical Bayes (EB) estimate is*

$$\widehat{\beta}_i = (1 - \frac{1}{z_i^2})_+ \widetilde{\beta}_i,$$

where $\widetilde{\boldsymbol{\beta}} = (\boldsymbol{U}'\boldsymbol{U})^{-1}\boldsymbol{U}'\boldsymbol{y}$, the ordinary least squares estimate of $\boldsymbol{\beta}$ and z_i is the test statistic for testing $H_0 : \beta_i = 0$ vs. $H_1 : \beta_i \neq 0$, that is $z_i = \frac{\widetilde{\beta}_i}{\sigma/\sqrt{n}}$.

Proof of Proposition 4

By EB estimate, we mean the estimate obtained for $\widehat{\beta}_i$ after plugging the estimates of τ_i^2, $i = 0, \ldots, s - 1$ that maximize the integrated likelihood, or equivalently:

$$\widehat{\boldsymbol{\tau}}^2 = \underset{\boldsymbol{\tau}^2}{argmin} \left[log \left| \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}' + \sigma^2 \boldsymbol{I}_n \right| + (\boldsymbol{y} - \mu_0 \boldsymbol{1}_n)'(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}' + \sigma^2 \boldsymbol{I}_n)^{-1}(\boldsymbol{y} - \mu_0 \boldsymbol{1}_n) \right]$$

into the expression for the posterior mean of $\boldsymbol{\beta}$, conditional on $\boldsymbol{\tau}^2$.

In order for the $n \times s$ matrix $\boldsymbol{U}$ to yield $\boldsymbol{U}'\boldsymbol{U} = n\boldsymbol{I}_s$, it must be that $s \leq n$. When $s < n$, there exist orthogonal columns that can be appended to $\boldsymbol{U}$, say $\boldsymbol{V}$,

such that: $\boldsymbol{W} = [\boldsymbol{U}, \boldsymbol{V}]$, where $\boldsymbol{W}$ is $n \times n$ and $\boldsymbol{W}'\boldsymbol{W} = n\boldsymbol{I}_n$. Since we are not particularly interested in the "effects" represented by the columns of $\boldsymbol{V}$ and as we demonstrate below, the optimization problem is separable, we can extend the matrix $\boldsymbol{\Sigma}$ in an arbitrary way. Let $\boldsymbol{S} = diag(\tau_0^2, \ldots, \tau_{s-1}^2, \tau_s^2, \ldots \tau_{n-1}^2)$ represent the $n \times n$ prior covariance matrix for these n orthogonal effects. In terms of these matrices, the -2 log likelihood is

$$l = log|\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}' + \sigma^2\boldsymbol{I}_n| + (\boldsymbol{y} - \mu_0\boldsymbol{1}_n)'(\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}' + \sigma^2\boldsymbol{I}_n)^{-1}(\boldsymbol{y} - \mu_0\boldsymbol{1}_n).$$

Note that

$$\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}' + \sigma^2\boldsymbol{I}_n = \boldsymbol{W}\boldsymbol{S}\boldsymbol{W}' + \sigma^2\boldsymbol{W}\boldsymbol{W}^{-1}(\boldsymbol{W}')^{-1}\boldsymbol{W}'.$$

Now, since $\boldsymbol{W}$ is orthogonal,

$$\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}' + \sigma^2\boldsymbol{I}_n = \boldsymbol{W}\boldsymbol{S}\boldsymbol{W}' + \frac{\sigma^2}{n}\boldsymbol{W}\boldsymbol{W}'.$$

Or simply,

$$\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}' + \sigma^2\boldsymbol{I}_n = \boldsymbol{W}\, diag\left(\tau_0^2 + \frac{\sigma^2}{n}, \tau_1^2 + \frac{\sigma^2}{n}, \ldots, \tau_{n-1}^2 + \frac{\sigma^2}{n}\right)\boldsymbol{W}'. \qquad (27)$$

Consider the determinant,

$$
\begin{aligned}
|\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}' + \sigma^2\boldsymbol{I}_n| &= \left|\boldsymbol{W}\, diag\left(\tau_0^2 + \frac{\sigma^2}{n}, \tau_1^2 + \frac{\sigma^2}{n}, \ldots, \tau_{n-1}^2 + \frac{\sigma^2}{n}\right)\boldsymbol{W}'\right| \\
&= \left|diag\left(\tau_0^2 + \frac{\sigma^2}{n}, \tau_1^2 + \frac{\sigma^2}{n}, \ldots, \tau_{n-1}^2 + \frac{\sigma^2}{n}\right)\boldsymbol{W}'\boldsymbol{W}\right| \\
&= \left|diag\left(n\tau_0^2 + \sigma^2, n\tau_1^2 + \sigma^2, \ldots, n\tau_{n-1}^2 + \sigma^2\right)\right| \\
&= \prod_{i=0}^{n-1}\left(n\tau_i^2 + \sigma^2\right).
\end{aligned}
$$

So that its log is:

$$log|\boldsymbol{W}\boldsymbol{S}\boldsymbol{U}' + \sigma^2\boldsymbol{I}_n| = \sum_{i=0}^{n-1} log(n\tau_i^2 + \sigma^2). \qquad (28)$$

84

Now consider the inverse of (27),

$$
\begin{aligned}
(\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}' + \sigma^2 \boldsymbol{I}_n)^{-1} &= \left(\boldsymbol{W} \, diag(\tau_0^2 + \frac{\sigma^2}{n}, \tau_1^2 + \frac{\sigma^2}{n}, \ldots, \tau_{n-1}^2 + \frac{\sigma^2}{n}) \boldsymbol{W}' \right)^{-1} \\
&= (\boldsymbol{W}')^{-1} \, diag\left(\frac{n}{n\tau_0^2 + \sigma^2}, \frac{n}{n\tau_1^2 + \sigma^2}, \ldots, \frac{n}{n\tau_{n-1}^2 + \sigma^2} \right) \boldsymbol{W}^{-1} \\
&= \frac{1}{n^2} \boldsymbol{W} \, diag\left(\frac{n}{n\tau_0^2 + \sigma^2}, \frac{n}{n\tau_1^2 + \sigma^2}, \ldots, \frac{n}{n\tau_{n-1}^2 + \sigma^2} \right) \boldsymbol{W}'.
\end{aligned}
$$

Let

$$
\boldsymbol{D} = diag\left(\frac{n}{n\tau_0^2 + \sigma^2}, \frac{n}{n\tau_1^2 + \sigma^2}, \ldots, \frac{n}{n\tau_{n-1}^2 + \sigma^2} \right).
$$

So that,

$$
\begin{aligned}
(\boldsymbol{y} - \mu_0 \boldsymbol{1}_n)'(\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}' + \sigma^2 \boldsymbol{I}_n)^{-1}(\boldsymbol{y} - \mu_0 \boldsymbol{1}_n) &= \frac{1}{n^2}(\boldsymbol{y} - \mu_0 \boldsymbol{1}_n)'\boldsymbol{W}\boldsymbol{D}\boldsymbol{W}'(\boldsymbol{y} - \mu_0 \boldsymbol{1}_n) \\
&= \widetilde{\boldsymbol{\beta}}' \boldsymbol{D} \widetilde{\boldsymbol{\beta}} \\
&= \sum_{i=0}^{n-1} \frac{n}{n\tau_i^2 + \sigma^2} \widetilde{\beta}_i^{\,2}. \qquad (29)
\end{aligned}
$$

Thus, from (28) and (29), we see that the finding of $\boldsymbol{\tau}^2$ that maximizes the integrated likelihood is equivalent to solving the convenient *separable* optimization problem:

$$
\widehat{\boldsymbol{\tau}^2} = \underset{\boldsymbol{\tau}^2 \geq \boldsymbol{0}}{argmin} \sum_{i=0}^{n-1} \left[log(n\tau_i^2 + \sigma^2) + \frac{n}{n\tau_i^2 + \sigma^2} \widetilde{\beta}_i^{\,2} \right].
$$

Differentiating with respect to τ_i^2, we obtain the partial derivatives:

$$
\frac{\partial l}{\partial \tau_i^2} = \frac{n}{n\tau_i^2 + \sigma^2} - \frac{n^2}{(n\tau_i^2 + \sigma^2)^2} \widetilde{\beta}_i^{\,2}, \quad \forall i = 1, \ldots, n-1.
$$

Setting the partial derivatives to zero and solving for τ_i^2, yields

$$
n(n\widehat{\tau_i^2} + \sigma^2) = n^2 \widetilde{\beta}_i^{\,2}.
$$

So that

$$
\widehat{\tau_i^2} = \left(\widetilde{\beta}_i^{\,2} - \frac{\sigma^2}{n} \right)_+
$$

is feasible. We can verify that $\widehat{\boldsymbol{\tau}^2} = (\widehat{\tau_0^2}, \widehat{\tau_1^2}, \ldots, \widehat{\tau_{n-1}^2})'$ is a constrained *global minimizer* by observing that:

$$
\forall \tau_i^2 < \widetilde{\beta}_i^{\,2} - \frac{\sigma^2}{n} : \frac{\partial l}{\partial \tau_i^2} = \frac{n}{n\tau_i^2 + \sigma^2} - \frac{n^2}{(n\tau_i^2 + \sigma^2)^2} \widetilde{\beta}_i^{\,2} < 0
$$

85

and

$$\forall \tau_i^2 > \widetilde{\beta_i}^2 - \frac{\sigma^2}{n} : \frac{\partial l}{\partial \tau_i^2} = \frac{n}{n\tau_i^2 + \sigma^2} - \frac{n^2}{(n\tau_i^2 + \sigma^2)^2}\widetilde{\beta_i}^2 > 0.$$

That is, l is monotonically decreasing for $\tau_i^2 < \widetilde{\beta_i}^2 - \frac{\sigma^2}{n}$ and monotonically increasing for $\tau_i^2 > \widetilde{\beta_i}^2 - \frac{\sigma^2}{n}$.

From Joseph (2006), we have an expression for the posterior mean of $\boldsymbol{\beta}$. Plugging in the EB estimators $\widehat{\tau_i^2}$, $\forall i = 0, \dots n-1$, into (18) yields:

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}_{EB}} &= \widehat{\boldsymbol{S}}\boldsymbol{W}' \left(\boldsymbol{W}\widehat{\boldsymbol{S}}\boldsymbol{W}' + \sigma^2\boldsymbol{I}_n \right)^{-1} (\boldsymbol{y} - \mu_0\boldsymbol{1}_n) \\
&= \frac{1}{n^2}\widehat{\boldsymbol{S}}\boldsymbol{W}'\boldsymbol{W}\widehat{\boldsymbol{D}}\boldsymbol{W}'(\boldsymbol{y} - \mu_0\boldsymbol{1}_n) \\
&= \widehat{\boldsymbol{S}}\widehat{\boldsymbol{D}}\widetilde{\boldsymbol{\beta}} \\
&= diag\left(\frac{n\widehat{\tau_0^2}}{n\widehat{\tau_0^2} + \sigma^2}, \frac{n\widehat{\tau_1^2}}{n\widehat{\tau_1^2} + \sigma^2}, \dots, \frac{n\widehat{\tau_{n-1}^2}}{n\widehat{\tau_{n-1}^2} + \sigma^2} \right)\widetilde{\boldsymbol{\beta}}
\end{aligned}
$$

Note that when $n\widetilde{\beta_i}^2 > \sigma^2$,

$$\frac{n\widehat{\tau_i^2}}{n\widehat{\tau_i^2} + \sigma^2} = \frac{n\widetilde{\beta_i}^2 - \sigma^2}{n\widetilde{\beta_i}^2} = 1 - \frac{1}{z_i^2}.$$

And when $n\widetilde{\beta_i}^2 < \sigma^2$,

$$\frac{n\widehat{\tau_i^2}}{n\widehat{\tau_i^2} + \sigma^2} = \frac{0}{\sigma^2} = 0.$$

Therefore,

$$\widehat{\beta_i} = (1 - \frac{1}{z_i^2})_+\widetilde{\beta_i} \quad \forall i = 0, \dots, n-1.$$

And if the effects $\widehat{\beta_s}, \dots, \widehat{\beta_{n-1}}$ are not of interest, then they can simply be ignored. $\diamondsuit$

APPENDIX C

ADDITIONAL SIMULATION RESULTS

The following tables characterize the performance of the proposed optimization experiement analysis methodology versus the traditional, frequentist methodology employing a statistical test of significance. Random models were generated with the same mechanism as described in Section 2.7.1. However, we additionally assume that σ^2 is unknown, but we guess that it is $\sigma^2_{guess} = 1$, without using centerpoints. The procedure is compared to that using the independent t-tests with $\alpha = 0.05$, with different assumptions on the number of degrees of freedom (ν) that are "left over" for estimating this variance. The summary tables that follow further demonstrate the strong performance of the proposed methodology. In addition, the tables provide an adequate baseline (when $\nu = 1$) for putting this performance into perspective.

Table 11: ME Simulation ($\gamma = 0.25$, $\tau^2 = 0.001$, $\sigma^2_{guess} = 1$)

% Improvement									
	t-test $\alpha = 0.05$ (ν)			Practical Thresholding (Δ)					
σ^2	1	2	10	0.0	0.2	0.4	0.6	0.8	1.0
0.0	0.00	0.43	0.74	0.88	0.86	0.84	0.81	0.78	0.75
0.5	0.00	0.43	0.74	0.87	0.85	0.83	0.81	0.78	0.74
1.0	0.00	0.43	0.73	0.85	0.84	0.82	0.79	0.77	0.73
2.0	0.00	0.43	0.71	0.82	0.81	0.79	0.77	0.74	0.71
5.0	0.01	0.45	0.67	0.75	0.74	0.73	0.72	0.70	0.68
10.0	0.02	0.45	0.61	0.66	0.66	0.65	0.64	0.62	0.61

Estimation Error									
	t-test $\alpha = 0.05$ (ν)			Practical Thresholding (Δ)					
σ^2	1	2	10	0.0	0.2	0.4	0.6	0.8	1.0
0.0	0.50	0.29	0.13	0.07	0.08	0.08	0.10	0.11	0.13
0.5	0.50	0.29	0.16	0.14	0.14	0.14	0.14	0.15	0.16
1.0	0.50	0.30	0.21	0.21	0.21	0.21	0.21	0.20	0.20
2.0	0.50	0.32	0.34	0.33	0.33	0.33	0.33	0.32	0.31
5.0	0.50	0.48	0.60	0.57	0.58	0.57	0.57	0.57	0.56
10.0	0.50	0.77	0.89	0.86	0.86	0.86	0.86	0.85	0.85

Average Number of Active Effects									
	t-test $\alpha = 0.05$ (ν)			Practical Thresholding (Δ)					
σ^2	1	2	10	0.0	0.2	0.4	0.6	0.8	1.0
0.0	0.00	0.59	1.42	2.13	2.01	1.88	1.74	1.59	1.44
0.5	0.00	0.62	1.48	3.47	2.83	2.32	1.95	1.69	1.50
1.0	0.00	0.65	1.70	4.80	4.02	3.25	2.62	2.12	1.74
2.0	0.00	0.70	2.48	6.13	5.38	4.61	3.85	3.15	2.55
5.0	0.01	1.27	4.25	7.61	7.03	6.38	5.69	5.01	4.33
10.0	0.02	2.43	5.73	8.50	8.04	7.53	6.96	6.39	5.81

Table 12: ME Simulation ($\gamma = 0.50$, $\tau^2 = 0.001$, $\sigma^2_{guess} = 1$)

% Improvement									
	t-test $\alpha = 0.05$ (ν)			Practical Thresholding (Δ)					
σ^2	1	2	10	0.0	0.2	0.4	0.6	0.8	1.0
0.0	0.00	0.45	0.79	0.93	0.91	0.89	0.87	0.83	0.79
0.5	0.00	0.45	0.78	0.92	0.90	0.88	0.85	0.82	0.79
1.0	0.00	0.45	0.77	0.90	0.88	0.86	0.84	0.81	0.77
2.0	0.00	0.46	0.75	0.87	0.85	0.84	0.82	0.79	0.76
5.0	0.01	0.47	0.70	0.79	0.78	0.77	0.75	0.73	0.71
10.0	0.02	0.47	0.64	0.70	0.70	0.69	0.67	0.66	0.64

Estimation Error									
	t-test $\alpha = 0.05$ (ν)			Practical Thresholding (Δ)					
σ^2	1	2	10	0.0	0.2	0.4	0.6	0.8	1.0
0.0	0.70	0.40	0.18	0.10	0.10	0.11	0.13	0.15	0.18
0.5	0.70	0.41	0.22	0.18	0.18	0.18	0.19	0.20	0.22
1.0	0.71	0.42	0.27	0.24	0.25	0.25	0.25	0.26	0.27
2.0	0.71	0.45	0.38	0.36	0.36	0.36	0.36	0.36	0.36
5.0	0.70	0.57	0.62	0.59	0.59	0.59	0.59	0.59	0.59
10.0	0.70	0.82	0.89	0.86	0.86	0.86	0.86	0.86	0.86

Average Number of Active Effects									
	t-test $\alpha = 0.05$ (ν)			Practical Thresholding (Δ)					
σ^2	1	2	10	0.0	0.2	0.4	0.6	0.8	1.0
0.0	0.00	1.17	2.85	4.25	4.02	3.77	3.49	3.19	2.89
0.5	0.00	1.23	2.93	5.18	4.60	4.09	3.65	3.29	2.97
1.0	0.00	1.27	3.08	6.08	5.39	4.73	4.12	3.59	3.14
2.0	0.00	1.39	3.66	6.96	6.33	5.65	4.98	4.34	3.74
5.0	0.01	1.93	4.98	8.05	7.53	6.95	6.33	5.69	5.06
10.0	0.04	2.93	6.13	8.69	8.28	7.81	7.29	6.75	6.20

Table 13: ME Simulation ($\gamma = 0.25$, $\tau^2 = 0.25$, $\sigma^2_{guess} = 1$)

% Improvement									
	t-test $\alpha = 0.05$ (ν)			Practical Thresholding (Δ)					
σ^2	1	2	10	0.0	0.2	0.4	0.6	0.8	1.0
0.0	0.00	0.21	0.59	0.89	0.85	0.80	0.74	0.67	0.59
0.5	0.00	0.23	0.60	0.86	0.83	0.78	0.73	0.67	0.61
1.0	0.00	0.24	0.60	0.83	0.80	0.76	0.72	0.66	0.61
2.0	0.00	0.26	0.59	0.78	0.75	0.72	0.69	0.65	0.60
5.0	0.00	0.31	0.56	0.67	0.66	0.64	0.62	0.60	0.57
10.0	0.01	0.33	0.50	0.56	0.56	0.55	0.54	0.52	0.50

Estimation Error									
	t-test $\alpha = 0.05$ (ν)			Practical Thresholding (Δ)					
σ^2	1	2	10	0.0	0.2	0.4	0.6	0.8	1.0
0.0	0.66	0.50	0.29	0.16	0.17	0.18	0.21	0.25	0.29
0.5	0.66	0.50	0.32	0.23	0.23	0.24	0.26	0.28	0.31
1.0	0.66	0.50	0.36	0.29	0.29	0.30	0.31	0.32	0.34
2.0	0.66	0.52	0.43	0.38	0.38	0.39	0.39	0.40	0.41
5.0	0.66	0.62	0.64	0.60	0.60	0.60	0.61	0.61	0.61
10.0	0.66	0.85	0.90	0.87	0.87	0.87	0.87	0.87	0.87

Average Number of Active Effects									
	t-test $\alpha = 0.05$ (ν)			Practical Thresholding (Δ)					
σ^2	1	2	10	0.0	0.2	0.4	0.6	0.8	1.0
0.0	0.00	0.69	3.05	6.77	6.07	5.33	4.57	3.81	3.13
0.5	0.00	0.79	3.40	7.03	6.36	5.65	4.90	4.17	3.49
1.0	0.00	0.90	3.65	7.23	6.59	5.89	5.16	4.44	3.74
2.0	0.00	1.14	4.12	7.54	6.94	6.28	5.59	4.89	4.21
5.0	0.00	1.88	5.19	8.19	7.69	7.13	6.53	5.90	5.27
10.0	0.02	2.91	6.17	8.73	8.32	7.85	7.35	6.80	6.25

Table 14: ME Simulation ($\gamma = 0.50$, $\tau^2 = 0.25$, $\sigma^2_{guess} = 1$)

% Improvement									
	t-test $\alpha = 0.05$ (ν)			Practical Thresholding (Δ)					
σ^2	1	2	10	0.0	0.2	0.4	0.6	0.8	1.0
0.0	0.00	0.33	0.69	0.92	0.89	0.86	0.81	0.76	0.70
0.5	0.00	0.33	0.69	0.90	0.87	0.84	0.80	0.75	0.70
1.0	0.00	0.34	0.68	0.87	0.85	0.82	0.78	0.74	0.69
2.0	0.00	0.36	0.68	0.83	0.81	0.79	0.76	0.72	0.68
5.0	0.00	0.39	0.63	0.74	0.73	0.71	0.69	0.67	0.64
10.0	0.01	0.40	0.57	0.64	0.63	0.62	0.61	0.59	0.57

Estimation Error									
	t-test $\alpha = 0.05$ (ν)			Practical Thresholding (Δ)					
σ^2	1	2	10	0.0	0.2	0.4	0.6	0.8	1.0
0.0	0.79	0.53	0.27	0.15	0.16	0.18	0.20	0.24	0.27
0.5	0.79	0.53	0.32	0.23	0.23	0.24	0.26	0.28	0.31
1.0	0.79	0.54	0.36	0.29	0.29	0.30	0.31	0.33	0.35
2.0	0.79	0.56	0.44	0.39	0.39	0.40	0.40	0.41	0.43
5.0	0.79	0.67	0.65	0.61	0.61	0.61	0.62	0.62	0.62
10.0	0.79	0.87	0.91	0.87	0.87	0.87	0.87	0.87	0.87

Average Number of Active Effects									
	t-test $\alpha = 0.05$ (ν)			Practical Thresholding (Δ)					
σ^2	1	2	10	0.0	0.2	0.4	0.6	0.8	1.0
0.0	0.00	1.26	3.97	7.37	6.75	6.09	5.40	4.71	4.06
0.5	0.00	1.36	4.21	7.56	6.96	6.31	5.63	4.94	4.29
1.0	0.00	1.45	4.41	7.66	7.09	6.48	5.82	5.16	4.49
2.0	0.01	1.69	4.81	7.95	7.40	6.81	6.18	5.54	4.90
5.0	0.01	2.34	5.63	8.45	7.98	7.46	6.90	6.31	5.71
10.0	0.04	3.24	6.41	8.86	8.47	8.02	7.51	7.01	6.48

Bibliography

James O. Berger and Thomas Sellke. Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American Statistical Association*, 82(397):112–122, 1987.

George E. P. Box, J. Stuart Hunter, and William G. Hunter. *Statistics for Experimenters*. Wiley, New York, NY, 2nd edition, 2005.

Leo Breiman. Better subset regression using the nonnegative garrote. *Technometrics*, 37:373–384, 1995.

Kathryn Chaloner and Isabella Verdinelli. Bayesian experimental design: A review. *Statistical Science*, 10(3):273–304, 1995.

Shao-Wei Cheng and Kenny Ye. Geometric isomorphism and minimum aberration. *Annals of Statistics*, 32(5):2168–2185, 2005.

Hugh Chipman, Michael Hamada, and C. F. Jeff Wu. A bayesian variable-selection approach for analyzing designed experiments with complex aliasing. *Technometrics*, 39:372–381, 1997.

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

Andrew Gelman, John B. Carlin, Hal S. Stern, and Donald B. Rubin. *Bayesian Data Analysis*. Chapman and Hall/CRC, Boca Raton, FL, 2nd edition, 2004.

Michael Hamada and N. Balakrishnan. Analyzing unreplicated factorial experiments: A review with some new proposals. *Statistica Sinica*, 8(11):1–41, 1998.

Michael Hamada and C. F. Jeff Wu. Analysis of designed experiments with complex aliasing. *Journal of Quality Technology*, 24:130–137, 1992.

David A. Harville. *Matrix Algebra from a Statistician's Perspective*. Springer, New York, NY, 1997.

C. Hellstrand. The necessity of modern quality improvement and some experience with its implementation in the manufacture of rolling bearings [and discussion]. *Philosophical Transactions of the Royal Society of London, Series A, Mathematical and Physical Sciences*, 327(1596):529–537, 1989.

Norman L. Johnson, Samuel Kotz, and N. Balakrishnan. *Continuous Univariate Distributions*. Wiley, New York, NY, 2nd edition, 1994.

V. Roshan Joseph. Quality loss functions for nonnegative variables and their applications. *Journal of Quality Technology*, 32(1596):129–138, 2004.

V. Roshan Joseph. A bayesian approach to the design and analysis of fractionated experiments. *Technometrics*, 48:219–229, 2006.

William J. Kennedy and Theodore A. Bancroft. Model building for prediction in regression based upon repeated significance tests. *The Annals of Mathematical Statistics*, 42(4):1273–1284, 1971.

M. Kathleen Kerr. Bayesian optimal fractional factorials. *Statistica Sinica*, 11:605–630, 2001.

Erich L. Lehmann and George Casella. *Theory of Point Estimation*. Springer, New York, NY, 2nd edition, 1998.

Russell V. Lenth. Quick and easy analysis of unreplicated factorials. *Technometrics*, 31:469–473, 1989.

Runze Li and Agus Sudjianto. Analysis of computer experiments using penalized likelihood in gaussian kriging models. *Journal of the American Statistical Association*, 47(2):111–120, 2005.

R. Daniel Meyer, David M. Steinberg, and George E. P. Box. Follow-up designs to resolve confounding in multifactor experiments, (with discussion). *Technometrics*, 38:303–332, 1996.

Toby J. Mitchell, Max D. Morris, and Donald Ylvisaker. Two level fractional factorials and bayesian prediction. *Statistica Sinica*, 5:559–573, 1995.

Douglas C. Montgomery. *Design and Analysis of Experiments*. Wiley, New York, NY, 2004.

Raymond H. Myers and Douglas C. Montgomery. *Response Surface Methodology*. New York: Wiley, 2nd edition, 2002.

Madhav Shridhar Phadke. *Quality Engineering Using Robust Design*. Prentice Hall, Englewood Cliffs, NJ, 1989.

Ramkumar Rajagopal and Enrique Del Castillo. Model-robust process optimization using bayesian model averaging. *Technometrics*, 47:152–163, 2005.

Thomas J. Santner, Brian J. Williams, and William I. Notz. *The Design and Analysis of Computer Experiments*. Springer, New York, NY, 2003.

David M. Steinberg and Dizza Bursztyn. Data analytic tools for understanding random field regression models. *Technometrics*, 46(4):411–420, 2004.

Genichi Taguchi. *System of Experimental Design, Vol. 1 & Vol. 2*. Unipub/Kraus International, White Plains, NY, 1987.

Genichi Taguchi, Subi Chowdhury, and Shin Taguchi. *Robust Engineering*. McGraw-Hill, New York, NY, 2000.

Asokan Mulayath Variyath, Bovas Abraham, and Jiahua Chen. Analysis of performance measures in experimental design using jackknife. *Journal of Quality Technology*, 37:91–100, 2005.

C. F. Jeff Wu and Michael Hamada. *Experiments: Planning, Analysis, and Parameter Design Optimization*. Wiley, New York, NY, 2000.

C. F. Jeff Wu and Runchu Zhang. Minimum aberration designs with two-level and four-level factors. *Biometrika*, 80:203–209, 1993.