Analysis of Optimization Experiments

V. ROSHAN JOSEPH[*]

Georgia Institute of Technology, Atlanta, GA 30332-0205

and

JAMES DILLON DELANEY[†]

Carnegie Mellon University, Pittsburgh, PA 15213-3890

Abstract

The typical practice for analyzing industrial experiments is to identify statistically significant effects with a 5% level of significance and then to optimize the model containing only those effects. In this article, we illustrate the danger in utilizing this approach. We propose methodology using the practical significance level, which is a quantity that a practitioner can easily specify. We also propose utilizing empirical Bayes estimation which gives shrinkage estimates of the effects. Interestingly, the mechanics of statistical testing can be viewed as an approximation to empirical Bayes estimation, but with a significance level in the range of 15-40%. We also establish the connections that our approach has with a less known but intriguing technique proposed by Taguchi, known as the beta coefficient method. A real example and simulations are used to demonstrate the advantages of the proposed methodology.

Key Words: Empirical Bayes method; Practical significance level; Shrinkage estimation; Variable selection.

[*] Dr. Joseph is an Assistant Professor in the H. Milton Stewart School of Industrial and Systems Engineering at Georgia Tech. He is a Member of ASQ. His email address is roshan@isye.gatech.edu.

[†] Dr. Delaney is a Visiting Assistant Professor in the Department of Statistics at Carnegie Mellon University. His email address is jdelaney@stat.cmu.edu.

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 perhaps 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 affecting 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 effects. 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 at the discretion of the experimenter. The usual recommendation is to choose levels that minimize the total cost. The foregoing might be adequate, but it is in the step of identifying the significant factors where something can go wrong.

The basic flaw in the methodology 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 significance 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 methodology 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, of which -1 is the existing level. 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 $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and σ^2 is known. Suppose that $\sigma = 10$. 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

$$\text{p-value} = 2\Phi\left(-\frac{|\widetilde{\beta}_1|}{\sigma/\sqrt{2}}\right) = .2888,$$

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

Now let us take a different view of this problem, one which stresses that optimization is the objective. It is easier to use a Bayesian framework to demonstrate what is happening. Let $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. Then, under the improper prior distribution ($p(\boldsymbol{\beta}) \propto 1$), the posterior distribution of β_1 given the data ($\boldsymbol{y}$) is $\mathcal{N}(\widetilde{\beta}_1, \sigma^2/2)$. Thus

$$Pr(\beta_1 > 0|\boldsymbol{y}) = \Phi\left(\frac{\widetilde{\beta}_1}{\sigma/\sqrt{2}}\right) = .8556.$$

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 design 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? We may not want to change the setting unless the improvement of .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 $|2\widehat{\beta}_1| > \Delta$, where $\widehat{\beta}_1$ is the posterior mean of β_1, which in this case is the same as $\tilde{\beta}_1$ (note that when x is changed from -1 to 1, the predicted mean response is changed by $2\widehat{\beta}_1$). 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". We should note that practical significance is not a new concept (see, e.g., Montgomery 2004), however, we have not seen it being used as part of any rigorous methodology.

An objection to this methodology as it is described so far, could be that it does not consider the randomness in the response (i.e., $\hat{\beta}_1$ does not depend on σ^2). One approach to overcome this problem is to compute the probability $Pr(|2\beta_1| > \Delta|\boldsymbol{y})$ and declare x_1 as practically significant when this probability is high enough. However, when more than one factor is involved in the experiment, the computation of such probabilities becomes difficult. Therefore, we propose a different approach. We show that empirical Bayes (EB) estimation, assuming a proper prior distribution for β_1, gives an estimate that shrinks as σ increases. Thus when σ is large enough, the expected improvement is less than Δ, suggesting no change should be made for that factor setting.

Many authors have advocated using a Bayesian approach for the analysis of experiments. Box and Meyer (1993) proposed a Bayesian approach to identify active effects from a screening experiment. See also the follow-up papers by Meyer et al. (1996) and Chipman et al. (1997). Even though model selection is made less ambiguous than what is encountered in the usual statistical testing approach, this work does not account for the optimization objective of an experiment. Peterson (2004) proposed a Bayesian approach when the objective of the experiment is optimization. His approach is to find the settings of the factors that maximize the probability that the predicted response is within some desirable range. This approach incorporates the uncertainty in the model parameters through the calculation of the posterior predictive density of the response (see also Miro-Quesada et al. 2004). Rajagopal and Del Castillo (2005) proposed an extension of this approach that also takes into account the uncertainty of the model form. This is achieved by obtaining the posterior predictive distribution averaged over a candidate class of models. One major difference between their work and what we propose here is that we try to balance the optimization objective with the cost of implementing particular factor settings through the introduction of practical significance levels. However, whereas Rajagopal and Del Castillo (2005) does, we do not incorporate the model uncertainty. Moreover, because we use an empirical Bayes approach the uncertainty associated with hyper-parameters is also lost. Nevertheless, our approach leads to a procedure that is very close to the statistical testing procedure, which should readily appeal to practitioners. In fact, we show that the model selection in a version of our approach can be approximated by a statistical testing procedure with a higher than typically used significance level in the range of 15-40%.

The details of the proposed methodology for optimization experiments are described in the following sections. As discussed before, it differs from the existing approach mainly in the use of two concepts: practical significance level and EB estimation. First, we present a real experiment to serve as motivation for adoption of this methodology.

An Example

Consider the experiment reported by Hellstrand (1989) with the objective of reducing the failure rate of deep groove ball bearings (see also Box et al. 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 osculation refers to the ratio between the ball diameter and the radius of the outer ring of the bearing. The response is failure rate (y), which is defined as the reciprocal of the average time to failure (in hours) $\times$ 100. The design and the data are given in Table 1.

The estimates of the seven effects are given in Table 2. Because this is an unreplicated

Table 1: Design Matrix and Data, the Bearing Experiment

	Factor			failure
Run	x_1	x_2	x_3	rate
1	-1	-1	-1	5.882
2	-1	-1	1	5.263
3	-1	1	-1	3.846
4	-1	1	1	6.250
5	1	-1	-1	4.000
6	1	-1	1	4.762
7	1	1	-1	1.176
8	1	1	1	0.781

Table 2: Parameter Estimates and Significance, the Bearing Experiment

			Approx. p-value	
Effect	$\widehat{\beta}_i$	$\lvert t_{PSE}\rvert$	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_1x_2	-0.719	0.918	0.31	> 0.40
x_1x_3	0.177	0.226	> 0.40	> 0.40
x_2x_3	-0.233	0.298	> 0.40	> 0.40
$x_1x_2x_3$	-0.523	0.667	> 0.40	> 0.40

experiment, the usual t-values cannot be computed in order to test the significance of each effect. A common remedy is to use a half-normal plot (Daniel 1959) and identify the large effects that appear to be outliers as the significant effects. The half normal plot of the effects is given in Figure 1. 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 the many other methods.) The t_{PSE} values from applying Lenth's method are given in Table 2. Two types of critical values may be used: the individual error rate (IER) and the experiment-wise error rate (EER). At the 5% significance level the critical value for IER is 2.30 (Wu and Hamada 2000). Because the t_{PSE} values are much lower than

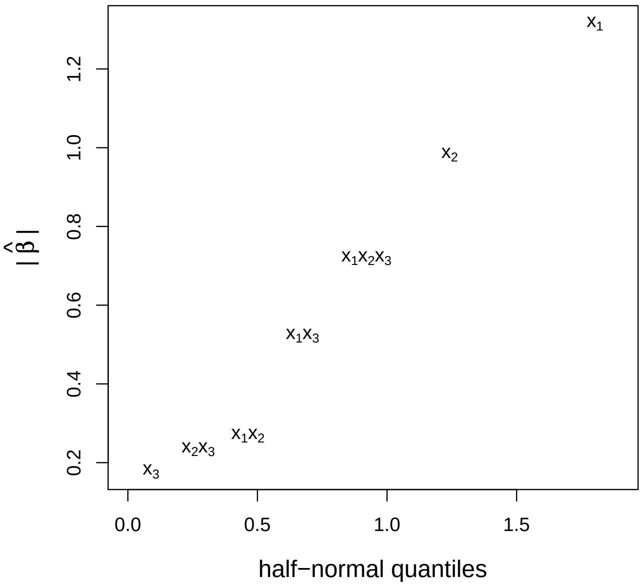

Figure 1: Half Normal Plot for the Bearing Experiment

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 2. We can see that the p-values are large enough to conclude that none of the effects are significant.

By examining the data in Table 1, we can see that run numbers 7 and 8 produce failure rates that are much lower than those of the other runs. It appears that keeping osculation and temperature both at their high values is beneficial. Hellstrand (1989) identified this setting using simple interaction plots. He confirmed this choice of factor settings through observing vastly improved bearing performance in its subsequent use. These settings yielded a substantial improvement in failure rate that would have been missed if we were to rely upon only the statistical test of significance.

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

$$TC = E\{L(Y)\} + C(\boldsymbol{x}), \tag{1}$$

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 methodology that can be used without requiring knowledge of the actual form of the cost function. To achieve this, we will identify the factors that have a practically significant effect on $E\{L(Y)\}$ and use only those factors in order to minimize $E\{L(Y)\}$. The settings of the other factors may be selected so as to minimize the cost. This is similar to the existing approach, 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 $E\{L(Y)\}$ that contains only the practically significant factors and then optimize it. A factor will be identified as practically significant if it can make a change in the response more than a prescribed practical significance level Δ. Note that in this approach each factor is tested for practical significance; not each effect (main effects, interaction effects, etc.). These concepts will be made clear with some examples.

Let us look at the bearing experiment again. First consider a model with only the main effects. The model is given by

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

Failure 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 $E(Y)$ which is estimated by $\widehat{y}$. Suppose that the existing level of failure rate is 5 and a 5% decrease is considered to be a significant improvement, then we can take $\Delta = .05 \times 5 = .25$. 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 .25 and so all of the factors are identified as practically significant, a very different conclusion from that arrived at from the application of the statistical significance level.

We note that the expected loss function in the foregoing formulation should be replaced with other measures of interest, if they are more meaningful to the problem at hand. For example, if the quality characteristic is of the nominal-the-best variety, we may perform the analysis on the mean and variance separately; if the quality characteristic is of the larger-the-better variety, we may perform the analysis on the mean; and so on. It is important to choose a measure that is readily understandable by the experimenter, so that Δ can be easily chosen.

Now consider the full linear model with interactions,

$$\widehat{y} = 3.995 - 1.315x_1 - 0.982x_2 + 0.269x_3 - 0.719x_1x_2 - 0.177x_1x_3 + 0.233x_2x_3 - 0.523x_1x_2x_3. \quad (2)$$

In statistical hypothesis testing, the magnitudes of all seven effects are tested for their significance. In that setting, there is a penalty associated with including each effect in the model. In the proposed methodology only the significance of a factor is considered. There is no penalty for including an interaction effect when their parent effects are already in the model, because we only consider the cost associated with each factor, not with each effect. This is more reasonable when optimization is the objective.

To apply the practical significance level to each factor, we would need to know the effect of each factor. But then since interactions are present, the effect of a factor changes with the settings of the other factors. When there are factors present that have more than two levels, we might consider their quadratic, cubic, etc. 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 introduce the concept of the *impact* of a factor with respect to the optimal setting of $\boldsymbol{x}$.

Let $E\{L(Y)\} = g(\boldsymbol{x})$ 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

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

where $\boldsymbol{x}_{(i)}$ denotes all of the factors except x_i. The impact is the maximum change in $E\{L(Y)\}$, with respect to x_i. 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 $\mathit{impact}(x_i) = |2\beta_i|$, which would coincide exactly with the usual definition for a factorial effect.

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

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

The two factors x_i and x_j would be identified as practically insignificant if $\mathit{impact}(x_i, x_j) < 2\Delta$, in addition to each of $\mathit{impact}(x_i) < \Delta$ and $\mathit{impact}(x_j) < \Delta$. By extending these definitions to more than two factors, we can see that the set of practically insignificant factors is the largest set of factors such that every subset among these factors has an impact less than the practical significance level times the number of elements in that subset. This largest set of insignificant factors can be found through an exhaustive search. Much more efficient algorithms can be developed, which will be examined in a subsequent article. We note that in the case of a main effects model, the set of insignificant factors is easily obtained by calculating the individual impacts.

By optimizing the full linear model in (2), we obtain $x_1^* = 1$, $x_2^* = 1$, and $x_3^* = 1$. Now we need to compute the impacts. First consider the impact of x_1. By fixing $x_2^* = x_3^* = 1$ in (2), we obtain $g(x_1, \boldsymbol{x}^*_{(1)}) = 3.515 - 1.315x_1 - 0.719x_1 - 0.177x_1 - .523x_1$. Thus, the impact of x_1 is

$$\mathit{impact}(x_1) \;=\; 2 \times |-1.315 - 0.719 - 0.177 - 0.523| = 5.469.$$

Similarly, the impact of the other two factors can be computed as

$$\mathit{impact}(x_2) \;=\; 2 \times |-0.523 + 0.233 - 0.719 - 0.982| = 3.981,$$

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

Because all of these impacts are more than .25, they are all identified as practically significant. Thus all three factors should be changed to their higher levels so as to minimize the failure rate. This is a much different conclusion than what we obtain using the statistical significance tests.

This result is consistent with the conclusion of Hellstrand (1989), except therein the factor cage design (x_3) was not considered significant. Can the observed effect of x_3 be due merely 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.

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 $u_7 = x_1 x_2 x_3$. Assume that σ^2 is known. If unknown, it can be estimated from replicates. Let $\boldsymbol{u} = (1, u_1, u_2, \cdots)'$ and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \cdots)'$. Then $Y = \boldsymbol{u}'\boldsymbol{\beta} + \epsilon$. In a Bayesian framework, we would need to specify a prior distribution for $\boldsymbol{\beta}$. For notational simplicity, rewrite the model as $Y = \mu + \boldsymbol{u}'\boldsymbol{\beta} + \epsilon$, where μ denotes the prior mean for β_0. We assume the multivariate normal prior: $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$.

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

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

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

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\Sigma}\boldsymbol{U}'(\boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{U}' + \sigma^2 \boldsymbol{I})^{-1}(\boldsymbol{y} - \mu \boldsymbol{1}). \tag{3}$$

The unknown hyper-parameters (i.e., parameters in the prior distribution) can be estimated using empirical Bayes (EB) methods (see e.g. Carlin and Louis 2000). One common approach to implementing the EB method is to use maximum likelihood estimation of the hyper-parameters using the marginal distribution of the response. The marginal distribution of $\boldsymbol{y}$ can be obtained by integrating out $\boldsymbol{\beta}$. We obtain $\boldsymbol{y} \sim \mathcal{N}(\mu\mathbf{1}, \boldsymbol{U\Sigma U}' + \sigma^2 I)$. Thus, the log-likelihood of the marginal distribution of $\boldsymbol{y}$ is given by

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

The log-likelihood can be maximized with respect to μ, and the parameters in $\boldsymbol{\Sigma}$ to obtain their estimates.

The approach is very general. It can be applied to any type of designs: regular, nonregular, orthogonal, nonorthogonal, and mixed-level designs. The only condition required for the existence of EB estimates is that the matrix $\boldsymbol{U\Sigma U}' + \sigma^2 I$ be nonsingular.

When $\boldsymbol{U}$ is orthogonal and $\boldsymbol{\Sigma}$ is diagonal some simplifications can be observed. Below we consider three special structures for $\boldsymbol{\Sigma}$. They are presented in the order of increasing complexity. The last covariance structure is the most preferred, however the discussion of the first two is provided because it reveals interesting insights into the overall approach.

Before proceeding further, we should mention a disadvantage of using EB methods. In the EB method the uncertainty due to the hyper-parameter estimation is not accounted for (see Berger 1985). We can overcome this problem by using a hierarchical Bayes approach, where a second stage prior is postulated for the hyper-parameters. However, this comes at the expense of increased computation. In this article we focus on EB methods and leave further comparisons with the hierarchical Bayes methods for future research.

Identical Variances Prior

Consider the bearing experiment again. Assume that $\boldsymbol{\Sigma} = \tau^2\boldsymbol{I}$. Because Table 1 reflects a full factorial design, the columns of $\boldsymbol{U}$ are orthogonal. Thus $\boldsymbol{U\Sigma U}' = 8\tau^2\boldsymbol{I}$. From (3), we obtain

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

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

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

Thus

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

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

The marginal log-likelihood simplifies to

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

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

$$\widehat{\boldsymbol{\beta}} = \left(1 - \frac{\sigma^2}{s^2}\right)_+ \widetilde{\boldsymbol{\beta}}. \tag{4}$$

Thus the EB 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 EB estimation is well-known in the statistical literature. However, we have not seen it advanced as an alternative to statistical testing in the analysis of experiments. Note that if replicates are available, then σ^2 can be estimated based on the

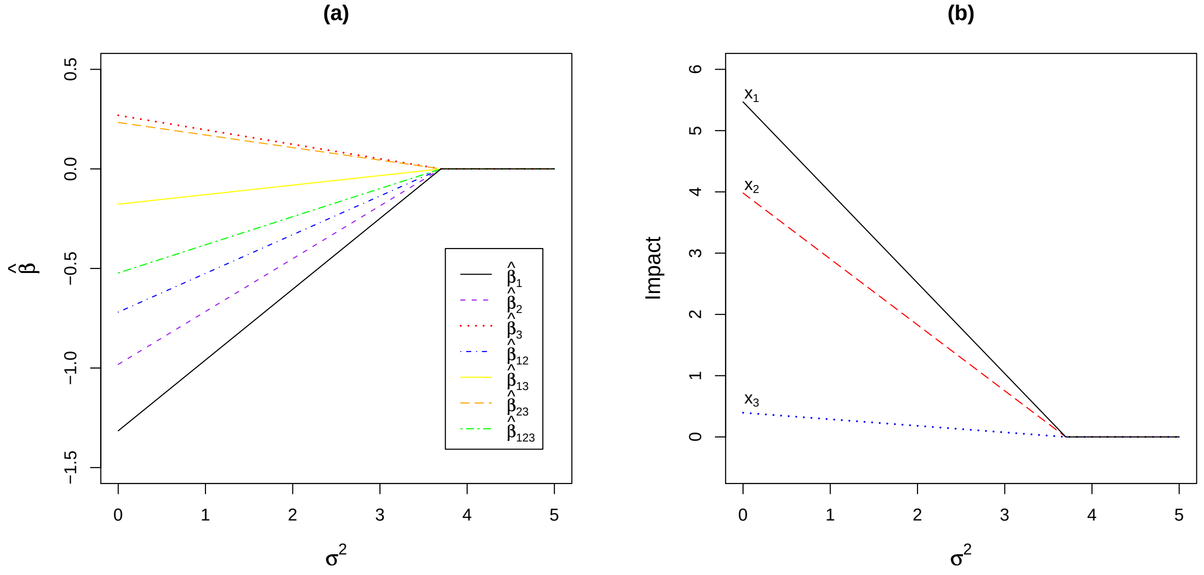

Figure 2: Bearing Experiment With Equal Prior Variances: (a) Coefficients (b) Impacts

sample variance of the replicates. If replicates are not available, then a reasonable guess value should be used.

The coefficients $\hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_7$ are plotted in Figure 2(a) against σ^2 (note that $\hat{\beta}_0 = 0$). We can see that as σ^2 increases, the $\hat{\beta}$'s decrease to 0. The impacts of the three factors can be calculated as before and are plotted in Figure 2(b). We can see that the impact of x_3 is practically insignificant at the 5% level ($\Delta = .25$) when $\sigma^2 > 1.4$. Therefore, if σ^2 is as large as 1.4, then x_3 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 failure rate substantially. This is a conclusion completely different from that obtained using the statistical tests of significance.

Unequal Variances Prior

Now consider a more general form for $\mathbf{\Sigma}$. As before, let the β_i's be independent but with possibly different prior variances: τ_i^2. Then $\mathbf{\Sigma} = diag(\tau_0^2, \tau_1^2, \cdots, \tau_7^2)$. From (3), 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

$$l = constant - \frac{1}{2}\sum_{i=0}^{7}\log(8\tau_i^2 + \sigma^2) - \frac{1}{2}\sum_{i=0}^{7}\frac{8\widetilde{\beta}_i^{\,2}}{8\tau_i^2 + \sigma^2}. \tag{5}$$

Minimizing 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 $H_0 : \beta_i = 0$ where $\sigma_{\widetilde{\beta}_i}$ denotes the standard error of $\widetilde{\beta}_i$. Because $\sigma_{\widetilde{\beta}_i} = \sigma/\sqrt{8}$, we obtain

$$\widehat{\beta}_i = \left(1 - \frac{1}{z_i^2}\right)_+ \widetilde{\beta}_i. \tag{6}$$

This shows that $\widehat{\beta}_i$ shrinks completely to 0 if $|z_i| \leq 1$. This threshold is equivalent to using $\alpha = 31.73\%$ in statistical testing. That is, when $|z_i| > 1$, the i^{th} coefficient is identified as statistically significant at the 31.73% level and $\widetilde{\beta}_i$ is used in the model. Whereas with EB estimation, a value smaller than $\widetilde{\beta}_i$ is used and as $|z_i|$ decreases, $\widehat{\beta}_i$ decreases continuously from $\widetilde{\beta}_i$ to zero.

The estimates of the coefficients are plotted in Figure 3(a). In addition, the impacts for the three factors at their optimal settings are plotted in Figure 3(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 ($\Delta = .25$) 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. Note that in this example, we do not have an estimate of σ^2. An engineer working in the deep groove ball bearing process can possibly give a reasonable estimate. But we can argue that σ^2 should be much less than 6.7, because the sample variance of the observed failure rates is only 3.6. Thus our analysis strongly supports the conclusion that at least x_1 and x_2 are practically significant.

Although we used the 2^3 design to derive (6), the result is much more general. It can be applied to fractional factorial designs (regular and nonregular) 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 below and is proved in the Appendix.

Proposition 1. Let

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

and

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

$\boldsymbol{U}$ is an $n \times s$ matrix such that $\boldsymbol{U}'\boldsymbol{U} = n\boldsymbol{I}_s$. Then the 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. $H_1 : \beta_i \neq 0$ with σ^2 known, that is $z_i = \frac{\widetilde{\beta}_i}{\sigma/\sqrt{n}}$.

The methodology can easily be extended to the case of an unknown σ. If an estimate of σ can be obtained from replicates, then

$$t_i = \widetilde{\beta}_i / \widehat{\sigma}_{\widetilde{\beta}_i} \tag{7}$$

has a t distribution with the appropriate degrees of freedom. Thus, the EB estimate in (6) becomes

$$\widehat{\beta}_i = \left(1 - \frac{1}{t_i^2}\right)_+ \widetilde{\beta}_i. \tag{8}$$

Similarly, the estimator in (4) is approximately $\widehat{\boldsymbol{\beta}} = (1 - 1/F)_+ \widetilde{\boldsymbol{\beta}}$, where F is the usual F-statistic for the overall test of significance (i.e., testing none of the effects are significant against at least one of them is significant). Thus, when the identical variances prior is

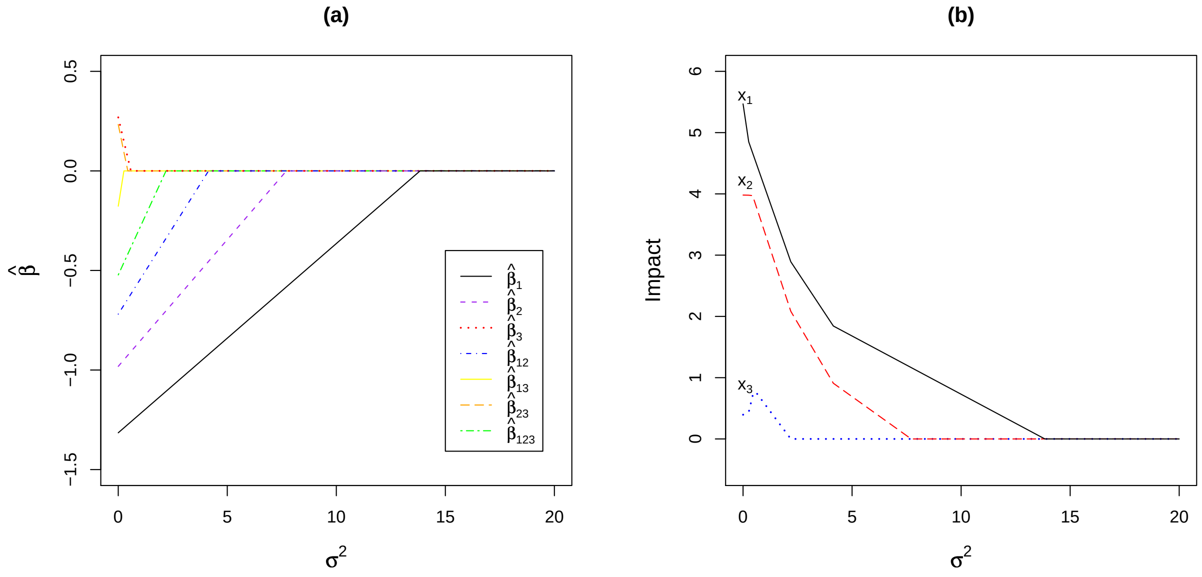

Figure 3: Bearing Experiment With Unequal Prior Variances: (a) Coefficients (b) Impacts

used, the shrinkage factor depends on the overall F-test statistic, whereas when the unequal variances prior is used, there are many shrinkage factors, each depending on the individual t-test statistics.

Heredity Prior

As it is described so far, the methodology does not incorporate the principles of effect hierarchy and effect heredity (Hamada and Wu 1992). Because the main effects, two-factor interactions, and the three-factor interaction are all treated the same way, effect hierarchy is not reflected in the methodology. Because an interaction term can appear in the model without any of its parent factors, effect heredity is also not reflected in the methodology. We can remedy this. Joseph (2006) and Joseph and Delaney (2007) 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. Then, from (3),

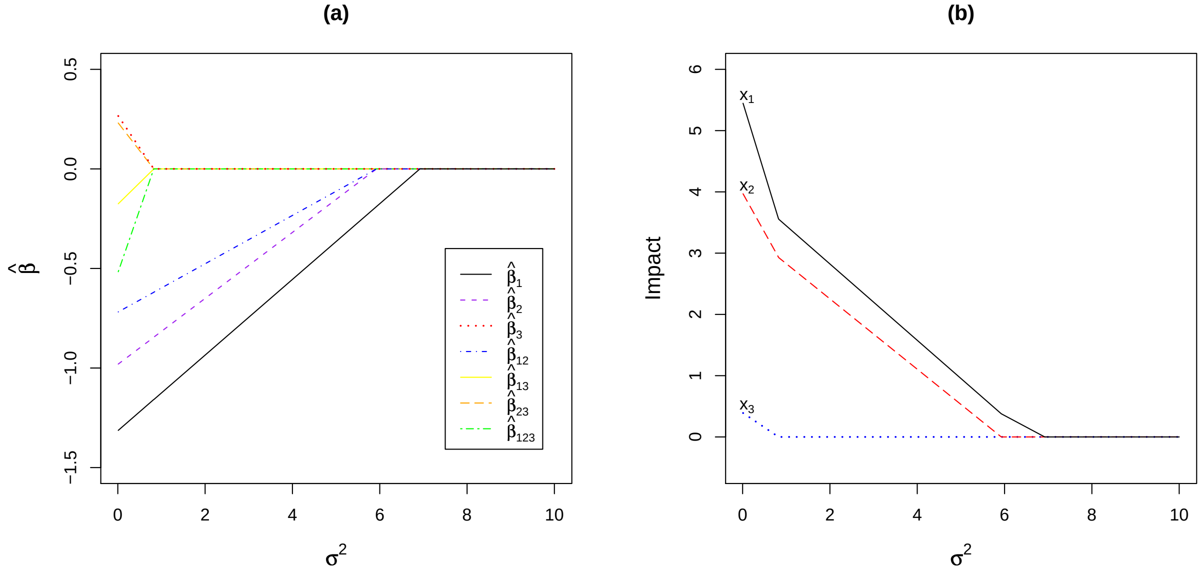

Figure 4: Bearing Experiment With Heredity Prior: (a) Coefficients (b) Impacts

we obtain

$$\widehat{\beta}_i = \frac{8\tau^2 \boldsymbol{R}_{ii}}{8\tau^2 \boldsymbol{R}_{ii} + \sigma^2} \widetilde{\beta}_i,$$

where $\boldsymbol{R}_{ii}$ represents the $(i+1)^{th}$ diagonal element of $\boldsymbol{R}$. Furthermore, the marginal log-likelihood becomes

$$l = constant - \frac{1}{2} \sum_{i=0}^{7} \log\left(8\tau^2 \boldsymbol{R}_{ii} + \sigma^2\right) - \frac{1}{2} \sum_{i=0}^{7} \frac{8\widetilde{\beta}_i^{\,2}}{8\tau^2 \boldsymbol{R}_{ii} + \sigma^2}. \tag{9}$$

We may numerically maximize this log-likelihood in order to find the EB estimates for the hyper-parameters μ, r_1, r_2, r_3, and τ^2.

The consequence of assuming the heredity prior can readily be discerned from the plot of the coefficients that is given in Figure 4(a). Coefficients are zeroed in groups as σ^2 increases. For instance, both $\widehat{\beta}_2$ and the interaction $\widehat{\beta}_{1,2}$ are zero for all $\sigma^2 > 6$. The separation between the significant factors and insignificant factors 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 4(b) are once again consistent with the conclusion of Hellstrand (1989).

Shrinkage Estimation Methods

Taguchi (1987, chapter 19) criticized the use of statistical testing in experiments and proposed an intriguing method which he named the *beta coefficient method.* He notes on page 566 of his book that “when results obtained by experiment are actually applied, it is rare that an effect greater than the experimental results is obtained, and that in most cases less than the expected effect is obtained.” Therefore, he suggested that the effects obtained from an experiment should be shrunk towards 0 before making the prediction. Taguchi 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 introduce different notation, λ.

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

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

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

$$\lambda_i = 1 - \frac{\widehat{\sigma}^2/n}{\widetilde{\beta}_i^2} = 1 - \frac{1}{t_i^2}. \tag{10}$$

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 EB estimate in (8).

Although the EB approach (with unequal variances prior) leads to the same approach suggested by Taguchi, we note that the EB 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.

The value of using shrinkage estimation for improving efficiency is evident from its presence in the statistical literature (Gruber 1998). Recently, there has been a surge of interest in developing methods that are a combination of shrinkage and subset selection. These methods include the nonnegative (nn) garrote by Breiman (1995), the least absolute shrinkage and selection operator (lasso) by Tibshirani (1996), and least angle regression (LARS) by Efron et al. (2004). EB estimation has connections to the nn-garrote, which can be seen by considering orthogonal designs. Breiman (1995) showed that the estimate of β_i in the nn-garrote scheme is given by

$$\hat{\beta}_i = \left(1 - \frac{c}{\tilde{\beta}_i^2}\right)_+ \tilde{\beta}_i,$$

where $\tilde{\beta}_i$ is the least squares estimate and c is estimated from the data using cross validation methods. If we replace c with $\hat{\sigma}^2/n$, then the nn-garrote estimator is exactly the same as the one we obtained in (8). However, the EB estimate is more general than the nn-garrote estimate. This is because, in fractional factorial designs the number of effects to estimate is more than the number of runs, in which case least squares estimates do not exist. Thus, the nn-garrote estimates also do not exist, whereas the EB estimates can still be found (see Joseph 2006 and Joseph and Delaney 2007).

Statistical Testing as an Approximation

For the EB estimate in (6), the value of β_i in the estimated model can be expressed: $\lambda_i \widetilde{\beta}_i$, where $\lambda_i = (1 - 1/z_i^2)_+$. In statistical hypothesis testing, $\lambda_i = 0$ when $|z_i| \leq z_{\alpha/2}$ and $\lambda_i = 1$ otherwise. As discussed in the introduction, it is difficult to find a meaningful statistical significance level, α, for a given problem. However, the similarity of this testing procedure with the EB estimation reveals that statistical testing can be used as an approximation. A simple approximation is to take $z_{\alpha/2} = 1$, which yields $\alpha = 31.73\%$. But because the EB estimates shrink towards 0 when $z_i > 1$, we may prefer to search for an even closer approximate statistical test.

A plot of λ as a function of z is provided in Figure 5(a). The objective is to find the $z_{\alpha/2}$ that minimizes the absolute difference between the EB estimate and the estimate implicit in 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 $\alpha = 15.73\%$. At this level, the EB estimate of β_i is one half of the least squares estimate.

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 continue to be used for many more years to come. Moreover, the procedure using statistical testing is easier to implement than EB estimation. So if an investigator prefers to apply statistical testing, we recommend using $\alpha = 15\%$. Taguchi (1987) also noted that statistical testing is acceptable as a procedure as long as it is viewed as an approximation to his beta coefficient method. However, he did not suggest any optimal value for α as we did here.

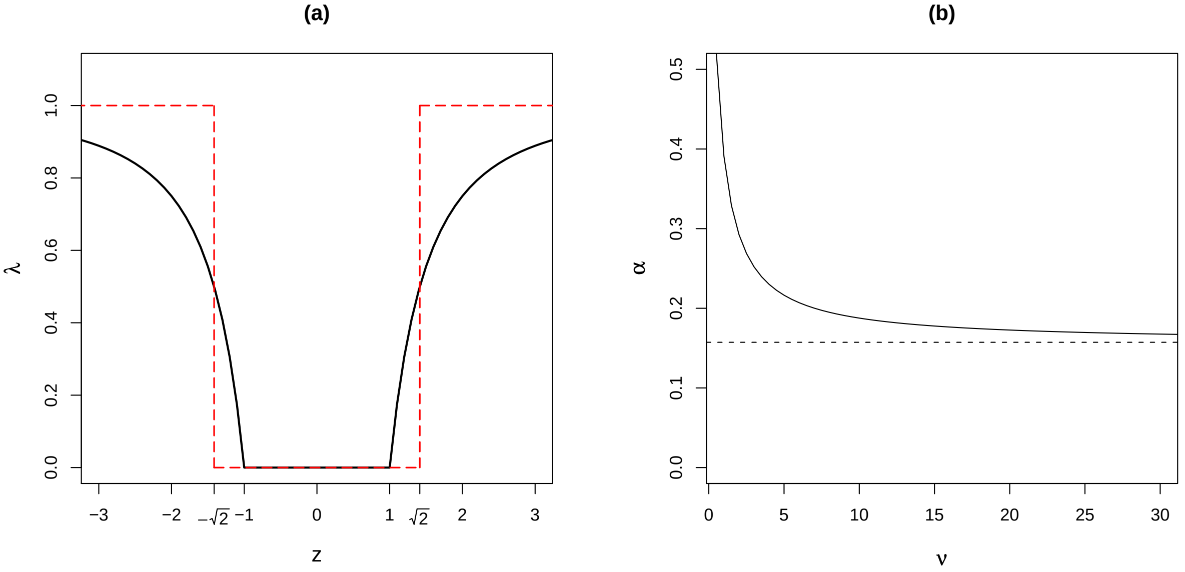

Figure 5: Testing as an Approximation: (a) Shrinkage as a Function of Critical Values: Hypothesis Testing (dashed), EB Estimation (solid), (b) Optimal significance levels for t-tests

Note that the significance level α is the probability of mistakenly declaring a given effect as nonzero. When many effects are tested simultaneously, the probability of mistakenly declaring at least one effect as nonzero will be larger than α. To overcome this problem, multiple testing methods such as the Bonferronni correction method and the studentized maximum modulus method are recommended in the literature (see, e.g. Wu and Hamada 2000). However, our derivation demonstrates that $\alpha = .15$ should be used irrespective of the number of effects being examined. Therefore, in optimization experiments, we recommend against using multiple testing procedures.

If σ can be estimated, then a t-statistic could be used for testing H_0: $\beta_i = 0$. Note that the optimal critical value is still $\sqrt{2}$, 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 $\alpha = .3918$. This approaches .1573 as $\nu \rightarrow \infty$ (see Figure 5(b)).

It is a common practice to use a higher significance level when stepwise regression methods are applied for variable selection. In this setting, Kennedy and Bancroft (1971) offer simulation results to suggest using a statistical significance level in the range of .10 to .25. Our theoretical results provide further justification for this choice. Moreover, a different level of α would be associated with different degrees of freedom. In fact, it is better to fix the t-critical value at $\sqrt{2}$ instead of specifying any particular α-level. Most statistical software use an F-critical value for entering or removing a variable from the model. This critical value should be set at $(\sqrt{2})^2 = 2$.

Simulation

We use simulation to investigate the properties of the proposed methodology for optimization experiments. Below, we consider the estimation of the main effects from a design that is a 12-run orthogonal array over 11 factors, with the model matrix (see, e.g. Wu and Hamada 2000):

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

Models were simulated with the following mechanism:

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

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

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

Similar models are widely used in the analysis of experiments (see, e.g., Chipman et al. 1997). η_i is an indicator variable that controls whether effect i is active or inactive. This happens with probability γ and $1 - \gamma$ respectively. When the effect is active ($\eta_i = 0$), the coefficient is drawn from $\mathcal{N}(0, 1)$ and when it is inactive ($\eta_i = 1$), the coefficient is drawn from $\mathcal{N}(0, \tau^2)$, where $\tau^2 << 1$. By making τ^2 small, we ensure that the inactive effects are small.

Without loss of generality, we assume $\mu = 0$. First consider the case when σ^2 is known. Then, for each of these models we carry out estimation and variable selection in the traditional frequentist manner, using statistical hypothesis testing. The significance levels of $\alpha = .0045$, $\alpha = .0500$, and $\alpha = .1573$ respectively correspond to: the Bonferroni adjustment to $\alpha = .05$ to properly account for simultaneous testing, the α-level required for declaring significance in many publications, and the level we recommend as an approximation to the EB procedure as discussed above, respectively. The hypothesis test corresponding to these levels will be applied to the standard least squares parameter estimates. For EB estimation, we use (6) or (8) depending on whether σ^2 is known or unknown. In addition, results are presented for a variety of levels of the practical significance level (Δ) applied to the EB estimates, so as to identify up to 11 factor impacts. N=10,000 random models were generated for many different values for σ^2, τ^2, and γ.

We assume that Y is a larger-the-better quality characteristic. For each simulation j, we have a true model for the response $y_j(\boldsymbol{x})$. Let $\boldsymbol{x}^\dagger$ represent the true optimal factor settings. Then $y_j(\boldsymbol{x}^\dagger) = \sum_{i=1}^{11} |\beta_i^{(j)}|$, where $\beta_i^{(j)}$ is the coefficient of x_i in the jth model. For simplicity, assume that the existing setting is 0 for all of the factors. Thus $y_j(\boldsymbol{0}) = 0$. Therefore $y_j(\boldsymbol{x}^\dagger)$ can be viewed as the improvement obtained in an experiment with the j^{th} true model. Thus the average improvement is $1/N \sum_{i=1}^{N} y_j(\boldsymbol{x}^\dagger)$. Let $\boldsymbol{x}^*$ denotes the factor settings we would choose when employing an estimation and thresholding methodology. The average improvement with that methodology is $1/N \sum_{i=1}^{N} y_j(\boldsymbol{x}^*)$. Thus the percent improvement can be calculated by

$$\% \text{ Improvement} = \frac{\sum_{j=1}^{N} y_j(\boldsymbol{x}^*)}{\sum_{j=1}^{N} y_j(\boldsymbol{x}^\dagger)} \times 100.$$

Because our objective in an optimization experiment is to balance quality with cost, we need a metric to asses the cost. In a main effects model, the cost might be considered to be proportional to the number of active effects. Therefore, we use the average number of active effects as a second metric to evaluate the performance of different techniques. Therefore, a good methodology should yield high values for the % Improvement and low values for the Number of Active Effects.

Table 3: Simulation results ($\gamma = 0.5$, $\tau^2 = 0.001$, σ^2 known)

% Improvement									
	Statistical Significance (α)			Practical Significance (Δ)					
σ^2	0.0045	0.0500	0.1573	0.0	0.1	0.2	0.3	0.4	0.5
0.0	100	100	100	100	98	96	96	95	94
0.5	81	88	91	93	93	92	91	90	89
1.0	68	80	86	90	89	88	87	87	85
2.0	51	68	78	84	83	82	81	80	79
5.0	25	47	61	70	70	69	68	67	66
10.0	11	30	46	57	57	56	55	54	53

Number of Active Effects									
	Statistical Significance (α)			Practical Significance (Δ)					
σ^2	0.0045	0.0500	0.1573	0.0	0.1	0.2	0.3	0.4	0.5
0.0	11.00	11.00	11.00	11.00	5.88	5.05	4.82	4.60	4.39
0.5	3.16	4.09	5.14	6.36	5.93	5.51	5.10	4.71	4.37
1.0	2.39	3.50	4.68	6.02	5.69	5.36	5.03	4.71	4.41
2.0	1.60	2.78	4.12	5.60	5.35	5.09	4.83	4.57	4.32
5.0	0.71	1.87	3.29	4.94	4.77	4.59	4.40	4.22	4.04
10.0	0.33	1.31	2.74	4.47	4.34	4.20	4.06	3.92	3.79

Table 3 displays the two metrics for comparing the proposed methodology with existing hypothesis testing techniques for scenarios that could easily characterize some real experiments. Here we use $\gamma = .5$ and $\tau^2 = .001$. From this table, it is quite clear that the settings selected when using EB estimation, in particular when $\Delta = 0$, yield superior results with respect to the optimization experiment objective of improving the response that would indeed be realized. For example when $\sigma^2 = 1$, the % Improvement with our methodology is 90, whereas if an $\alpha = .05$ statistical testing procedure is applied, then the % Improvement is only 80. The situation is far less favorable for statistical testing with the Bonferroni adjustment, which results in an improvement of only 68%. The approximate statistical test using $\alpha = .1573$ is actually quite good with an average % Improvement of 86. This is still uniformly worse than EB estimation. As σ^2 increases, the improvement obtained by the EB procedure is much higher than the other procedures.

The second panel of Table 3 reveals that the average Number of Active Effects is larger for our methodology. With $\gamma = .5$, the average Number of Active Effects should be close to 5.5. We can see that this number is only achieved in our methodology when Δ is positive. As Δ increases, the Number of Active Effects is reduced, however, the % Improvement is also reduced. Figure 6 shows the plot of these two metrics over Δ when $\sigma^2 = 0$. This figure suggests that using a small value for Δ, a large reduction in the average Number of Active Effects can be achieved while sacrificing only a small amount in % Improvement.

That the same patterns so far revealed in this section are reproducible for different combinations of γ and τ^2 is illustrated in Figure 7. In this figure, the two metrics are plotted as a function of γ. Each line represents a different value for τ^2, for either the usual statistical test of significance procedure or EB estimation . In these plots, we fix $\sigma^2 = 1$. Also, the statistical significance level of $\alpha = .05$ and practical significance level of $\Delta = 0$ are used. Notice that for virtually any values of γ and τ^2, the % Improvement using the settings suggested by EB estimation is superior to that when using the statistical z-test.

We also checked to see how the EB response estimates, after applying a practical significance rule, compare with the least squares response estimates after applying statistical testing. Table 4 provides mean squared error calculations from these simulations. We can see that for these small values of Δ, the mean squared errors are generally much smaller after using EB and these practical significance levels than with the models based on least squares and statistical testing.

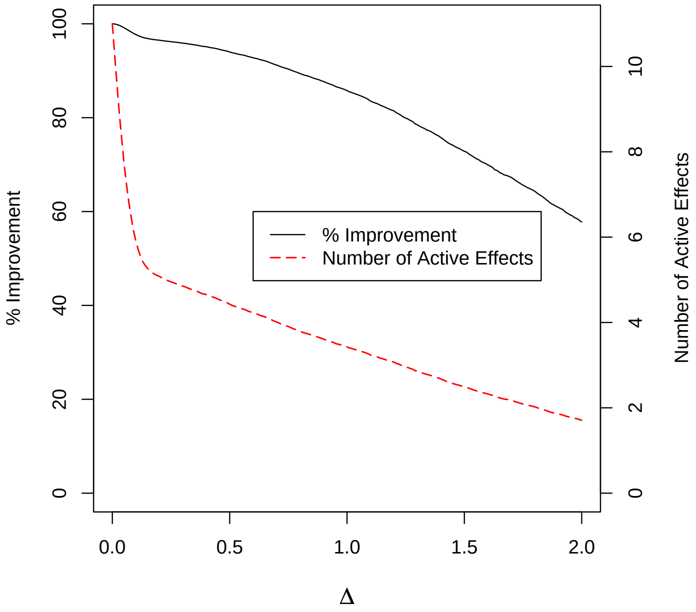

Figure 6: Simulation, Choosing Δ ($\gamma = 0.5$, $\tau^2 = 0.001$, and $\sigma^2 = 0$)

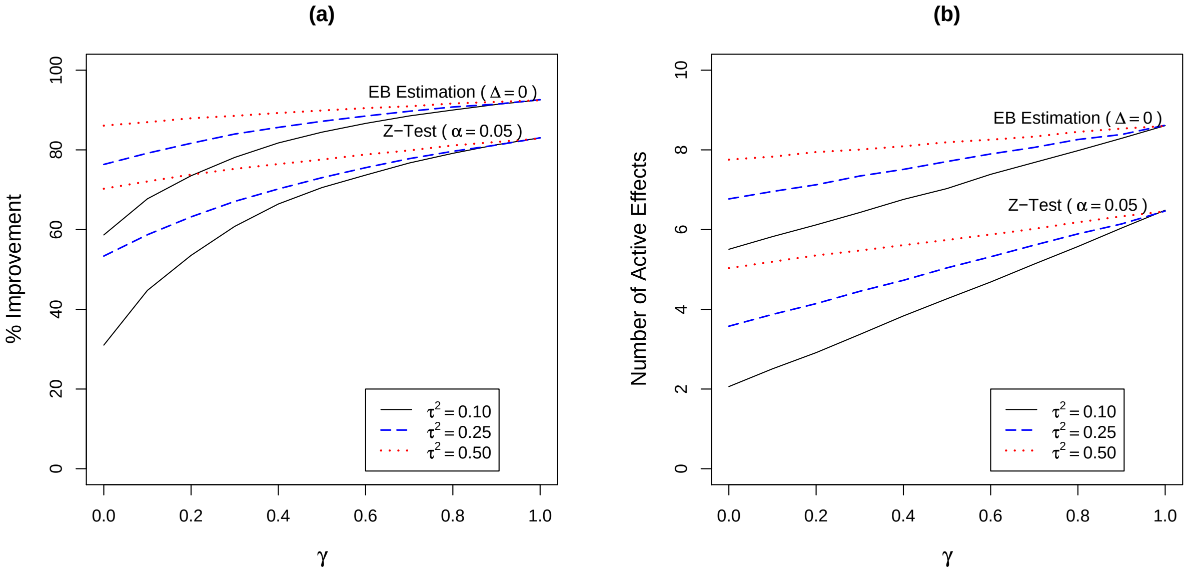

Figure 7: Simulation by varying γ and τ^2 ($\sigma^2 = 1$ and known) (a) % Improvement and (b) Number of active effects.

Table 4: Simulation results ($\gamma = 0.5$, $\tau^2 = 0.001$, σ^2 known)

Mean Squared Error of Response Estimates									
	Statistical Significance (α)			Practical Significance (Δ)					
σ^2	0.0045	0.0500	0.1573	0.0	0.1	0.2	0.3	0.4	0.5
0.0	0.00	0.00	0.00	0.00	0.01	0.03	0.04	0.06	0.10
0.5	1.00	0.42	0.22	0.14	0.16	0.19	0.23	0.28	0.34
1.0	2.63	1.08	0.55	0.33	0.37	0.42	0.48	0.55	0.64
2.0	6.08	2.74	1.41	0.81	0.88	0.97	1.07	1.18	1.30
5.0	13.70	7.24	4.14	2.53	2.65	2.79	2.94	3.10	3.25
10.0	19.08	12.20	7.64	5.04	5.20	5.34	5.52	5.69	5.87

Now consider the case when σ^2 is not known but may be estimated from the data. Because we are estimating σ^2, the 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, which are then used for estimating σ^2 (see, e.g. Wu and Hamada (2000, pp. 146)). We are using center points, because the addition of them into the design will not alter the estimates of β_i's.

When m is small, the advantage of the proposed methodology over that which involves a t-test is profound. For example, when $m = 3$, $\gamma = .50$, $\tau^2 = .001$, and $\sigma^2 = 1.0$, on average, we would expect to obtain 90% Improvement by choosing the settings that are dictated by EB estimation, whereas, in the hypothesis testing setting, the usual t-test would only yield 54% Improvement. However, as $m \to \infty$ the advantage of the proposed methodology over the utilization of the hypothesis test quickly begins to resemble the smaller, but distinct, advantage demonstrated earlier. This is illustrated in the plots of % Improvement provided in Figure 8. The results from employing the approximate statistical test ($|t| < \sqrt{2}$) is also shown. The approximate test is much better than the usual t-test with $\alpha = .05$, however, not as good as the EB estimation methodology.

In the simulation, we did not need to incorporate the heredity prior model, because each true model did not contain any interactions. The superiority of the heredity prior analysis over main effects analysis is demonstrated in Joseph (2006) and Joseph and Delaney (2007).

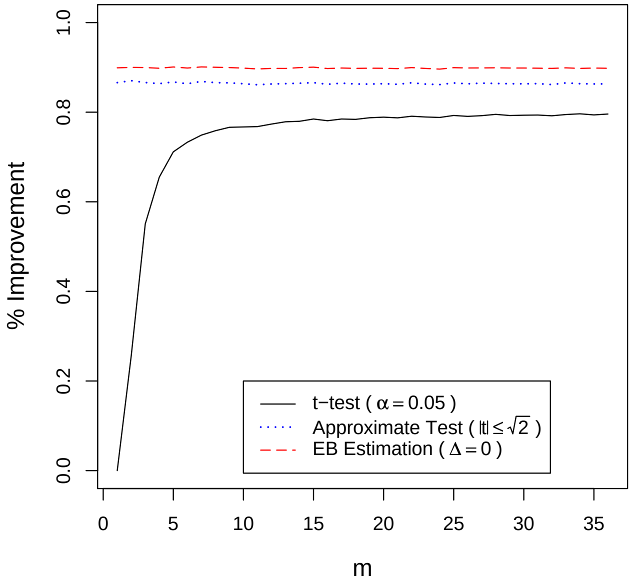

Figure 8: % Improvement as $m \to \infty$ ($\sigma^2 = 1.0$ and unknown, $\gamma = 0.50$, and $\tau^2 = 0.001$)

Therefore, we would also expect better performance for our methodology for analyzing the natural extensions of the simulations performed here to models involving interactions. In fact, the benefit from using the practical significance level will be much more pronounced when interactions are involved. To see this, suppose all of the interactions (two-factor interactions up to 11-factor interactions) are considered in our simulations. Then there are a total of $2^{11} - 1 = 2047$ effects as opposed to only 11 in the main effects model. Thus a huge reduction in the Number of Active Effects can be achieved with a small level of Δ. A real example illustrating the use of Δ with heredity prior and additional simulation results are given in Delaney (2006).

Conclusions

The deep groove bearing design optimization example of Hellstrand (1989) illustrates a common and profound challenge encountered by practitioners. There are some undesirable practical consequences associated with the rigorous application of statistical hypothesis testing procedures, in that it can prevent the gathering of sufficient guidance for 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 the use of an $\alpha = .05$ 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 $\alpha = .05$ 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 effects as significant.

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 a statistical significance 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 methodology we recommend for the analysis of optimization experiments centers around an overall objective function which balances quality and cost. We suggest the EB estimator presented in Joseph (2006) and Joseph and Delaney (2007) because of its many desirable properties. It shrinks the coefficients and incorporates the effect hierarchy and heredity principles. 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 setting and determine whether this is large enough to be of practical interest.

We have demonstrated EB estimation using three different prior specifications. This choice of prior has an effect on the final results. Our recommendation is to use the heredity prior. However, the amount of computation involved in obtaining the estimates is much more than for the other priors. When only a main effects model is considered, the use of either the heredity or unequal variances prior leads to the same result. But when interaction and polynomial terms are present in the model, the heredity prior should offer a more reasonable and interpretable model. We have shown that when the unequal variances prior is used, the EB estimator is the same as the beta coefficient method of Taguchi.

We found that statistical testing can be viewed as an approximation to the proposed EB procedure. This is a very interesting and useful result because statistical testing in general is much easier to implement than the EB procedure. We showed that using a z-critical or t-critical value of $\sqrt{2}$ in statistical testing approximates the EB procedure (with the unequal variances prior). This leads to a statistical significance level that ranges from 15 to 40% depending on the error degrees of freedom. That this statistical significance level is more liberal than the one that is universally used, and much more liberal than the typical implementation for the simultaneous testing of multiple parameters, should come as no surprise to practitioners experienced with analyzing engineering process optimization experiments.

The simulation results provide support for the conclusion that the recommended methodology is superior to statistical hypothesis 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.

Acknowledgments

The authors thank the Editor and a referee for their valuable comments and suggestions. This research was supported by the U.S. National Science Foundation grant CMMI-0448774.

APPENDIX: Proof of Proposition 1

Proof: 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 marginal log-likelihood is

$$l = constant - \frac{1}{2}\, log\, det\left(\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}' + \sigma^2\boldsymbol{I}_n\right) + (\boldsymbol{y} - \mu_0\mathbf{1}_n)'(\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}' + \sigma^2\boldsymbol{I}_n)^{-1}(\boldsymbol{y} - \mu_0\mathbf{1}_n).$$

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

$$det\left(\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}' + \sigma^2\boldsymbol{I}_n\right) = \prod_{i=0}^{n-1}\left(n\tau_i^2 + \sigma^2\right), \tag{11}$$

and,

$$(\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}' + \sigma^2\boldsymbol{I}_n)^{-1} = \frac{1}{n^2}\boldsymbol{W}\boldsymbol{\Lambda}\boldsymbol{W}',$$

where

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

Let $\widetilde{\boldsymbol{\beta}} = (\boldsymbol{W}'\boldsymbol{W})^{-1}\boldsymbol{W}'\boldsymbol{y}$. So that,

$$(\boldsymbol{y} - \mu\mathbf{1}_n)'(\boldsymbol{W}\boldsymbol{S}\boldsymbol{W}' + \sigma^2\boldsymbol{I}_n)^{-1}(\boldsymbol{y} - \mu\mathbf{1}_n) = \sum_{i=0}^{n-1}\frac{n}{n\tau_i^2 + \sigma^2}\widetilde{\beta}_i^2. \tag{12}$$

Thus, from (11) and (12), we see that the finding of $\boldsymbol{\tau}^2$ that maximizes the integrated likelihood is equivalent to solving the convenient *separable* optimization problem:

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

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

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

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.

Plugging in the EB estimators $\widehat{\tau_i^2}$, for all $i = 0, \ldots n-1$, into (3) yields:

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

Note that when $n\widetilde{\beta_i}^2 > \sigma^2$, we have $n\widehat{\tau_i^2}/(n\widehat{\tau_i^2} + \sigma^2) = 1 - 1/z_i^2$, and when $n\widetilde{\beta_i}^2 < \sigma^2$, we have $n\widehat{\tau_i^2}/(n\widehat{\tau_i^2} + \sigma^2) = 0$. Therefore,

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

And if the effects $\widehat{\beta}_s, \ldots, \widehat{\beta}_{n-1}$ are not of interest, then they can simply be ignored. ♢

References

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

Box, G. E. P.; Hunter, J. S.; and Hunter, W. G. (2005). *Statistics for Experimenters.* Wiley, New York, NY, 2nd edition.

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

Breiman, L. (1995). "Better Subset Regression Using the Nonnegative Garrote". *Technometrics*, 37, pp. 373–384.

Carlin, B. P. and Louis, T. A. (2000). *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman and Hall/CRC, Boca Raton, FL.

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

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

Delaney, J. D. (2006). *Contributions to the Analysis of Experiments Using Empirical Bayes Techniques.* PhD thesis, Georgia Institute of Technology, Atlanta, GA.

Efron, B.; Johnstone, I.; Hastie, T.; and Tibshirani, R. (2004). "Least Angle Regression". *Annals of Statistics*, 32, pp. 407–499.

Gruber, M. (1998). *Improving Efficiency By Shrinkage.* Marcel Dekker, New York, NY.

Hamada, M. and Balakrishnan, N. (1998). "Analyzing Unreplicated Factorial Experiments: A Review With Some New Proposals". *Statistica Sinica*, 8(11), pp. 1–41.

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

Hellstrand, C. (1989). "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), pp. 529–537.

Joseph, V. R. (2004). "Quality Loss Functions for Nonnegative Variables and Their Applications". *Journal of Quality Technology*, 32, pp. 129–138.

Joseph, V. R. (2006). "A Bayesian Approach to the Design and Analysis of Fractionated Experiments". *Technometrics*, 48, pp. 219–229.

Joseph, V. R. and Delaney, J. D. (2007). "Functionally Induced Priors for the Analysis of Experiments". *Technometrics*, 49, pp. 1–11.

Kennedy, W. J. and Bancroft, T. A. (1971). "Model Building For Prediction in Regression Based Upon Repeated Significance Tests". *Annals of Mathematical Statistics*, 42, pp. 1273–1284.

Lehmann, E. and Casella, G. (1998). *Theory of Point Estimation.* Springer, New York, NY, 2nd edition.

Lenth, R. V. (1989). "Quick and Easy Analysis of Unreplicated Factorials". *Technometrics*, 31, pp. 469–473.

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

Miro-Quesada, G.; Del Castillo, E.; and Peterson, J. J. (2004). "A Bayesian Approach for Multiple Response Surface Optimization in the Presence of Noise Variables". *Journal of Applied Statistics*, 31(3), pp. 251–270.

Montgomery, D. C. (2004). *Design and Analysis of Experiments.* Wiley, New York, NY.

Myers, R. H. and Montgomery, D. C. (2002). *Response Surface Methodology.* New York: Wiley, 2nd edition.

Peterson, J. J. (2004). "A Posterior Predictive Approach to Multiple Response Surface Optimization". *Journal of Quality Technology*, 36(2), pp. 139–153.

Rajagopal, R. and Del Castillo, E. (2005). "Model-Robust Process Optimization Using Bayesian Model Averaging". *Technometrics*, 47, pp. 152–163.

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

Tibshirani, R. (1996). "Regression Shrinkage and Selection via the LASSO". *Journal of the Royal Statistical Society, Sec. B*, 58, pp. 267–288.

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