# A Bayesian Hierarchical Approach to Dual Response Surface Modelling

Younan Chen Keying Ye
Department of Statistics, Virginia Tech, Blacksburg, VA, 24061

## Abstract

In modern quality engineering, dual response surface methodology is a powerful tool to monitor an industrial process by using both the mean and the standard deviation of the measurements as the responses. The least squares method in regression is often used to estimate the coefficients in the mean and standard deviation models, and various decision criteria are proposed by researchers to find the optimal conditions. Based on the inherent hierarchical structure of the dual response problems, we propose a hierarchical Bayesian approach to model dual response surfaces. Such an approach is compared with two frequentist least squares methods by using two real data sets and simulated data.

**Key Words:** Bayesian hierarchical model; dual response surface; off-line quality control; genetic algorithm; optimization.

## 1 Introduction

Much of response surface methodology (RSM), particularly in the early years, was focused on finding operating conditions that resulted in an optimum of the mean response with the homogeneity assumption on the variances. During the last two decades, industrial statisticians and practitioners have become aware that they can no longer focus themselves only

on the expected value of the response of interest. Instead, the variability of the response also needs to be considered. A common problem in an industrial process is to find the operating condition that achieves the target value for the mean of a process characteristic and minimizes the process variability. The pioneering work has been credited to Taguchi ([16]), who developed a package of tools which were viewed unfavorably by many researchers and practitioners for lack of statistical foundation (see [14]).

The dual response surface approach, first introduced by Myers and Carter ([13]) and revitalized by Vining and Myers ([18]), suggests that the process characteristic and its process variability form a dual response system (DRS), and two separate models are established for the response and its variance. In statistics, this approach allows the use of all regression tools to approximate the two response surfaces. In practice, the two separate models give the analyst a more scientific understanding of the total process, and thus allow them to see what levels of the control factors can lead to satisfactory values of the response as well as the variance.

Like other optimization work in RSM, the dual response optimization problem also consists of the following three stages. The first stage is to build an optimal experiment so that the information among the responses and the control factors can be obtained efficiently. The second stage is to build two models based on the data from the experiment, one for the process characteristic and the other for the process variance. The last stage is to search for the optimal operating condition throughout the region of interest under certain optimization criterion based on the established models. The third stage results cannot be trusted if models built in the second stage do not reflect the dual response surfaces well. The second stage is the focus of this paper, and all work is done under the assumption that data have already been collected. Model building efficiency is usually evaluated by comparing the performance of a product or a process at the found operating conditions, therefore optimization criteria and algorithms are inevitably involved.

Following Vining and Myers' article ([18]), several optimization formulations and procedures have been proposed for the DRS problem, for example, in Del Castillo and Montgomery ([5]), Lin and Tu ([10]), Copeland and Nelson ([3]), Ames *et al.* ([1]), Kim and Lin ([9]) and Tang and Xu ([17]). The above optimization work is confined to the third stage and is carried out under the assumption that the established models approximate the true response

surfaces well. Therefore the accuracy of the optimization results largely relies on whether the two established models are good approximations of the true dual response surfaces or not. If the two estimated models fit the surfaces poorly, then the true response and its variance at the chosen operating condition are very likely to be far from the specified requirement.

In this paper, a Bayesian hierarchical regression modelling approach is proposed to dual response surface. The sample means are used as the response for the mean model and log-normal distributions are assumed for the variances. The estimates of the coefficients in the two models are based on the posterior inference. A hybrid of local optimization algorithms and the genetic algorithm is adopted to search for the optimal operating conditions under two common optimization criteria. In Section 2, the frequentist modelling approach initially proposed by Vining and Myers ([18]) is briefly reviewed and the basic idea of the Bayesian hierarchical model is sketched. Section 3 presents a brief introduction of the genetic algorithm, discusses its advantages and disadvantages. Furthermore, a hybrid optimization method is proposed. In Section 4, a Bayesian hierarchical model is developed and the associated computation issues are discussed. After the model development, the Bayesian approach is compared with the frequentist methods in [18] by using two real data sets and simulated data. The theoretical details of the Bayesian approach are placed in the Appendix.

# 2 Least Square Methods and Bayesian Hierarchical Modelling

## 2.1 Review of the frequentist least square methods

Let $\mathbf{x}$ represent a $k \times 1$ vector of independent factors under the experimenter's control and $\mathbf{X}$ an $N \times p$ matrix used in the model building, where $N$ is the number of different design locations and $p$ is the number of coefficients in the model. If the model is of complete second-order, the matrix $\mathbf{X}$ consists of the $k$ coordinates of the design points plus the intercept, the interaction terms and the quadratic terms. Suppose that the experimenter seeks to optimize the process for $\mathbf{x} \in R$ where $R$ is the region of interest (usually the experimental region).

Vining and Myers ([18]) built two full second order models for the sample mean and the

sample standard deviation respectively:

$$\overline{\mathbf{y}} = \mathbf{X}\underline{\beta} + \varepsilon, \quad \text{and} \quad \mathbf{s} = \mathbf{X}\underline{\gamma} + \eta, \tag{2.1}$$

where $\overline{\mathbf{y}}$ is the vector of the sample means, $\mathbf{s}$ the vector of the sample standard deviations at the design points, $\underline{\beta}$ and $\underline{\gamma}$ are the coefficient vectors to be estimated, and $\varepsilon$ and $\eta$ are the error terms in their respective model.

The least squares method is a natural choice to estimate the coefficients for the dual response models without any assumption on the distributions of the sample mean or the sample standard deviation. Vining and Myers ([18]) mentioned that the generalized least squares (GLS) method should be pursued to estimate $\underline{\beta}$ in order to take into account the heterogeneity of variances in dual response problems:

$$\widehat{\underline{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\overline{\mathbf{y}} \quad \text{and} \quad \widehat{\underline{\gamma}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{s},$$

where $\hat{\mathbf{V}}$ is the estimated covariance matrix of the mean responses obtained in the design. $\hat{\mathbf{V}}$ is diagonal with the assumption that the random errors are independent from design point to design point. Due to computational difficulty, only ordinary least squares estimates were calculated in their paper. The diagonal elements of $\hat{\mathbf{V}}$ can be obtained in two ways. One is to use the sample variance at each design point. The other way is to use predicted values obtained through the variance model, by which the information in the process variance model can be incorporated into the mean response model. When the number of replicates is small, the latter one is usually preferred.

One pitfall with the modelling approach in (2.1) is that it is possible to yield negative predicted values of the standard deviation, even if the true mean standard deviations are positive throughout the region of interest. If the response at a point with the negative predicted standard deviation happened to be the optimum under certain criterion, it would arise difficulty in explaining the process performance at the picked optimal point, and cause confusion among the practitioners. Hence, we are trying to seek a more natural modelling mechanism that would fit the dual response system well.

## 2.2 Bayesian hierarchical modelling

Baysian hierarchical models mainly deal with data involving multiple parameters that can be treated as related in certain way by the structure of the data. Suppose there are observed data $y_{ij}$'s, where $y_{ij}$ is the $j^{th}$ observation in the $i^{th}$ group. The vector $\underline{\theta}_i$ (could be a scalar) contains the parameter(s) to determine the distribution of $y_{ij}$ in the $i^{th}$ group, and they are believed to be connected with parameters $\underline{\phi}$. The parameters $\underline{\phi}$ are at a higher level and are called *hyperparameters*. Conditional on $\underline{\theta}_i$, the data $y_{ij}$'s are assumed to be independently distributed, and given the hyperparameters $\underline{\phi}$, the parameters $\underline{\theta}_i$'s have the common density $\pi(\underline{\theta}|\underline{\phi})$.

Nonhierarchical models are usually inappropriate for a data set with such a hierarchical structure. If all $y_{ij}$'s are assumed to come from the same population distribution, the difference among the groups is inevitably neglected. If the uniqueness of each group needs to be introduced into the model without using a hierarchical model, too many parameters have to be entered into the model and the data will be overfitted. As a result, the estimated model may fit the existing data well but will produce poor predictions for future observations.

What a Bayesian hierarchical model does is to express the relationship among the parameters $\underline{\theta}_i$'s with a prior distribution: $\underline{\theta}_i$'s are treated as variables and $\underline{\phi}$ as the parameters in the prior distribution of $\underline{\theta}_i$'s. In this way, the only unknown parameters in the model are $\underline{\phi}$. Denote by $\pi(\underline{\theta}, \underline{\phi})$ the joint prior distribution of $\underline{\theta}$ and $\underline{\phi}$, $\pi(\underline{\theta}|\underline{\phi})$ the conditional prior distribution of $\underline{\theta}$ given $\underline{\phi}$, and $\pi(\underline{\phi})$ the prior distribution of $\underline{\phi}$. Furthermore, let $f(\text{data}|\underline{\theta})$ represent the probability density function of the data given $\underline{\theta}$, and $\pi(\underline{\theta}, \underline{\phi}|\text{data})$ the posterior distribution of the parameters. The joint prior distribution of parameters can be expressed as:

$$\pi(\underline{\theta}, \underline{\phi}) = \pi(\underline{\theta}|\underline{\phi}) \cdot \pi(\underline{\phi}),$$

and the posterior distribution is:

$$\pi(\underline{\theta}, \underline{\phi}|\text{data}) \propto f(\text{data}|\underline{\theta}, \underline{\phi}) \cdot \pi(\underline{\theta}, \underline{\phi}) = f(\text{data}|\underline{\theta}) \cdot \pi(\underline{\theta}|\underline{\phi}) \cdot \pi(\underline{\phi}),$$

since $f(\text{data}|\underline{\theta}, \underline{\phi})$ depends only on $\underline{\theta}$, and the hyperparameters $\underline{\phi}$ affect data only through $\underline{\theta}$. The prior distribution facilitates the incorporation of subjective information into the analysis of the current data set. If little is known about $\underline{\phi}$, a noninformative prior distribution can be specified followed by a check on the propriety of the joint posterior distribution.

Bayesian analysis is based on the posterior inference. Estimates of parameters are usually summary statistics of the marginal posterior distributions, such as the posterior mean, median, mode, standard deviation, etc. In simple nonhierarchical Bayesian models, it is often easy to analytically derive the marginal posterior distributions and obtain the summary statistics. However, in more complicated models, especially if the parameters are of multiple dimensions, it is often hard or impossible to analytically express the marginal distribution of each parameter. A common strategy is to use simulation techniques, such as the MCMC Gibbs sampling ([15]), to generate observations for the marginal posterior distributions and calculate the summary statistics numerically.

Bayesian hierarchical modelling of dual response surfaces is a natural application as the hierarchical structure is a built-in property of dual response surface problems: the parameters in the mean response model, including the variance, determine the distribution of the process characteristic at each design point; and the distribution of that variance can be further determined by the coefficients in the process variance model.

# 3 The Optimization Algorithm

In this paper, we use the combination of the genetic algorithm (GA) and the generalized reduced gradient algorithm (GRG) to achieve dual response optimization. GRG is a widely used local optimization algorithm in DRS and a detailed description can be found in Castillo an Montgomery ([5]).

The genetic algorithm was pioneered by Holland ([7]), and applications of GA to problems of mathematical optimizations owe much to De Jong ([4]). Recently it has emerged as an increasingly popular family of methods for global optimization, mainly because it offers a simple way (for both continuous and discrete variables) to solve for near-global optimum even for poorly behaved functions. GA is done by iteratively applying principles of "Darwinian natural selection" to a population of computer representations of the solution domain. The algorithm attempts to mimic the natural evolution of a population by allowing solutions to reproduce, creating new solutions, and to compete for survival in the next iteration. Every generation will have members that may not be an improvement over the previous generation and far from the global optimum. After many generations (iterations), the best solution is

usually near the global optimum.

Instead of the point-to-point search in the traditional methods, GA proceeds from one population to another, and thus it sweeps through the parameter space in many directions simultaneously and thereby reduce the probability of convergence to false optima. The optimal locations resulting from GA are always close to each even for poorly behaved surfaces.

The main limitation is that, like other direct-search methods, convergence of the genetic algorithm does not necessarily occur at a point where the gradient is exactly 0. That is, the solution from GA does not guarantee the exact global optimum, but near global optimum. If a very precise result is required, GA may suffer from excessively slow convergence because of its fundamental requirement of neglecting the local information. However, GA will typically find a point that is close enough to the optimum so that a gradient-type algorithm will efficiently converge to the exact global optimum if the found location is used as the starting point.

On the other hand, classical "hill-climbing" methods, such as the Quasi-Newton method and GRG, are well known to exploit all local information in an efficient way, though they often result in different local optima if different starting points are given. To take the advantages of both types of the algorithms, we proceed the hybrid method to the optimization work: GA is first used to find a near-global optimal point, and then the chosen point is served as the starting point for GRG to work out the exact global optimum.

# 4 Bayesian Hierarchical Modelling in DRS

The Bayesian hierarchical model is suggested as a new approach to dual response surface. The coefficients in the dual models can be estimated with some location parameters of the posterior distributions. Consequently, the predicted mean and variance can be obtained at any point within the design region.

## 4.1 Notations and Assumptions

Let $\mathbf{x}$ represent a $k \times 1$ vector of independent control factors, $\mathbf{X}$ an $N \times p$ matrix used in the mean response model and $\mathbf{Z}$ an $N \times q$ matrix in the variance model, where $N$ is the number of distinct design points in the experiment, $p$ and $q$ are the numbers of parameters in the mean and variance models respectively; and $\underline{x}_i'$, $\underline{z}_i'$ are the respective row vectors associated with the matrices $\mathbf{X}$ and $\mathbf{Z}$ for the $i^{th}$ location in the experiment. Denote by $\mathbf{Y} = (y_{11}, y_{12}, \ldots, y_{1n_1}, y_{21}, \ldots, y_{Nn_N})'$ the vector of observations from the experiment, and $y_{ij}$ the $j^{th}$ response observed at the $i^{th}$ design point, where $i = 1, 2, \ldots, N$ and $j = 1, 2, \ldots, n_i$; and the sample variance and population variance at the $i^{th}$ design point are denoted by $s_i^2$ and $\sigma_i^2$ respectively. Furthermore, for all the parameters, $\underline{\beta}$ represents a $p \times 1$ vector of coefficients of the mean response model, $\underline{\gamma}$ a $q \times 1$ vector of coefficients of the variance response model, and $\delta^2$ the hyperparameter in the distribution of $\sigma_i^2$'s.

Note that $\mathbf{X}$ and $\mathbf{Z}$ are not necessarily the same, and quite often $q$ is less than $p$ as variance models are usually simpler than mean response models.

We further assume independence among observations and parameters as follows:

- Independence among $y_{ij}$'s conditional on $\underline{\beta}$ and $\sigma_i^2$'s:

$$corr(y_{ij}, y_{i'j'}|\underline{\beta}, \sigma_i^2, \sigma_{i'}^2) = 0, \quad \text{for either} \quad i \neq i' \quad \text{or} \quad j \neq j' \quad \text{or both.}$$

- Independence among $\sigma_i^2$'s conditional on $\underline{\gamma}$ and $\delta^2$:

$$corr(\sigma_i^2, \sigma_{i'}^2|\underline{\gamma}, \delta^2) = 0, \quad \text{for} \quad i \neq i'.$$

- Independence among $\underline{\beta}$, $\underline{\gamma}$, and $\delta^2$ in the prior specification:

$$\pi(\underline{\beta}, \underline{\gamma}, \delta^2) = \pi(\underline{\beta}) \cdot \pi(\underline{\gamma}) \cdot \pi(\delta^2).$$

The independence among parameters assumption is made for simplicity, though it may need to be modified in the presence of subjective information.

## 4.2 Model building

The Bayesian hierarchical model is built in the following three stages:

- **Stage 1** Since variances are assumed to be different at various locations in dual response surface problems, a natural choice for the process characteristic is a heteroscedastic normal distribution conditional on $\underline{\beta}$ and the variance $\sigma_i^2$ at the $i^{th}$ location:

$$y_{ij}|\underline{\beta}, \sigma_i^2 \sim N(\underline{x}_i'\underline{\beta}, \sigma_i^2), \quad -\infty < y_{ij} < \infty, \quad \text{for} \quad i = 1, 2, \ldots, N; \quad j = 1, 2, \ldots, n_i.$$

  Hence the distribution of the sample mean at the $i^{th}$ location is $N(\underline{x}_i'\underline{\beta}, \sigma_i^2/n_i)$.

- **Stage 2** The logarithm transformation is a common technique used in variance modelling. Bartlett and Kendall ([2]) used the logarithm of the sample variance as the response for a linear regression model, and Engel and Huele ([6]) used the logarithm of the residual squares. Similarly, in the Bayesian model, the variance $\sigma_i^2$ is assumed to have a log-normal distribution conditional on $\underline{\gamma}$ and $\delta^2$:

$$\ln \sigma_i^2|\underline{\gamma}, \delta^2 \sim N(\underline{z}_i'\underline{\gamma}, \delta^2) \quad \text{for} \quad i = 1, 2, \ldots, N.$$

- **Stage 3** Set the prior distributions for parameters, $\underline{\beta}$, $\underline{\gamma}$, and $\delta^2$. Informative priors are always preferred if prior information with respect to parameters is available from previous studies or subjective knowledge. In choosing the form of the informative priors, those from the conjugate family are of the most interest. The practical advantage of using a conjugate prior distribution is that we know the posterior distribution follows the same parametric form as of the prior distribution. However, due to the complexity of the model we consider here, it is hard to find the real conjugate priors. Therefore the commonly used prior distributions for such kind of parameters are assumed as follows:

$$\pi(\underline{\beta}) \sim \text{MVN}_p(\underline{\mu}_\beta, \Sigma_\beta), \quad \pi(\underline{\gamma}) \sim \text{MVN}_q(\underline{\mu}_\gamma, \Sigma_\gamma), \quad \pi(\delta^2) \sim \text{Inverse-Gamma}(\alpha, \tau),$$

  where $\underline{\mu}_\beta$, $\Sigma_\beta$, and $\underline{\mu}_\gamma$, $\Sigma_\gamma$ are the respective mean vectors and covariance matrices of $\underline{\beta}$ and $\underline{\gamma}$ in the prior distributions, and $\alpha$, $\tau$ are the shape and scale parameters of the Inverse-Gamma distribution, respectively.

  If information available is in the form apparently not belonging to the conjugate family, we need to elicit more realistic prior distributions to incorporate such information. However, if there is no subjective information, noninformative priors could be used.

As discusses in Section 2.2, if the marginal posterior distributions of the parameters of interest cannot be analytically expressed, simulation techniques such as the MCMC Gibbs sampling method may be used. To use the Gibbs sampling technique, the joint posterior distribution and full conditional posterior distributions for using the above conventional prior distributions and noninformative distributions are derived and proved in Appendix A.

For a dual response surface problem without any subjective information on parameters, the following independent prior distributions can be used:

$$\pi(\underline{\beta}) \propto 1 \quad \pi(\underline{\gamma}) \propto 1 \quad \pi(\delta^2) \propto \frac{1}{\delta^2} \exp(-\frac{\lambda}{\delta^2}),$$

where $\lambda$ is a small positive number. The joint and conditional posterior distributions by using the above noninformative priors are derived in Appendix B. The propriety of the posterior distribution by using these improper priors are discussed in Appendix C.

## 4.3 Computation

The Gibbs sampler is used in the posterior simulation. Based on Theorem A.2 and Corollary B.2, the full posterior conditional distributions of $\underline{\beta}$, $\underline{\gamma}$ and $\delta^2$ can be generated directly but the distributions of $\sigma_i^2$'s are uncommon, so the rejection method (e.g. see [15]) is used. To simulate observations for the desired density $f(x)$ with the rejection method, we have two requirements: first, a well-known density $g(x)$ that we can easily generate observations from; and second, the ratio between $f(x)$ and $g(x)$ must have an upper bound $M$. The basic rejection method is to generate an observation from $g(x)$ first, and then compare the value $\frac{f(x)}{Mg(x)}$ with a uniform variate $\nu$, where $\nu \sim U(0,1)$, to decide whether the observation is accepted or not. The desired density $f(x)$ is often called the *target density*, and $g(x)$ is called the *instrumental density*.

The target density to be simulated is $\pi(\sigma_i^2|\text{others})$ in Corollary B.2 and the instrumental density proposed is:

$$g(\sigma_i^2) \sim \text{Inverse-Gamma}\left(\frac{n_i}{2}, \frac{(n_i-1)s_i^2 + n_i(\bar{y}_i - \underline{x}_i'\underline{\beta})^2}{2}\right).$$

The ratio between the target density and the instrumental density is

$$R_i = \frac{\pi(\sigma_i^2|\text{others})}{g(\sigma_i^2)} \propto \exp\left[-\frac{1}{2\delta^2}(\ln\sigma_i^2 - \underline{z}_i'\underline{\gamma})^2\right] \quad \text{for} \quad i = 1, 2, \ldots, N.$$

When $\sigma_i^2 = \exp\{\underline{z}_i'\underline{\gamma}\}$ , the ratio $R_i$ achieves its maximum, which is the ratio between the normalizing constants of the target density and the instrumental density.

It is expected that adding a small value of $\lambda$ should exert no influence on the conditional posterior distributions of $\underline{\beta}$, $\underline{\gamma}$, $\delta^2$ and $\sigma_i^2$'s. Usually $\frac{1}{2}(\underline{d} - Z\underline{\gamma})'(\underline{d} - Z\underline{\gamma})$ is very large relative to $\lambda$ if $\lambda$ is chosen to be tiny, so the conditional posterior distribution of $\delta^2$ is little affected, and the whole simulation procedure remains almost unchanged. The printing ink data and the catapult data in Section 5 are used for sensitivity analysis to study the effect of changing $\lambda$ on the optimal locations found. $\lambda = 0.1, 0.001$, and $10^{-6}$ are tried, and it is found that the predicted performance at the chosen locations is fairly robust to the choice of $\lambda$ (refer to Tables 1 and 2 in Sections 5.1 and 5.2, respectively).

# 5 Methods Comparison

In this section, the proposed Bayesian hierarchical model (henceforth referred as BAYES) is applied to two data sets. The coefficients in the two models, $\underline{\beta}$ and $\underline{\gamma}$, are estimated with their respective posterior medians. Posterior medians are preferred over the posterior means for their robustness to outliers. The optimization results from the Bayesian model are compared with the two frequentist least squares methods in [18]: one is to model the mean and standard deviation with linear regressions and estimate both sets of parameters with ordinary least squares (OOLS); and the other is similar to OOLS except that the mean response models coefficients are estimated with generalized least squares (GOLS). Furthermore, simulation is used to assess the modelling efficiency of each method by comparing the performance at the chosen optimal operating conditions.

To compare the optimization result from each model, optimal operating conditions are searched under two criteria in the literature. The first one, proposed by Vining and Myers ([18]), strictly sets the mean response at the target value, and then minimizes the variability subject to this constraint (henceforth referred to as "target is the best" criterion). The second one, proposed by Lin and Tu ([10]), minimizes the sum of deviance around the target value and the variability. This measure is defined as $(\text{mean} - \text{target})^2 + \text{variance}$, which is very similar to the definition of the mean squared error ($MSE$ criterion).

## 5.1 Printing Ink Data

Box and Draper (1987) outlined an experiment involving printing ink, which has been widely adopted in dual response surface analysis for illustration (e.g. see [1], [3] and [9]). The purpose of the experiment was to study the effects of speed ($x_1$), pressure ($x_2$), and distance ($x_3$), upon a printing machine's ability to apply coloring inks upon package labels. The original experiment used a $3^3$ complete factorial design with three runs at each design point for a total of 81 runs. All previous analyses considered the situation that we wish to minimize the process variability around a target value of 500 for the response. The same situation is considered in this paper. Since the experimental region is cuboidal with each factor taking values at $-1, 0, 1$, optimal conditions are searched throughout the cuboidal region.

Vining and Myers ([18]) assumed full second order models for both the mean and variance models so that the dual response optimization algorithm in Myers and Carter ([13]) could be carried out. All subsequent analyses on this data set has been done based on this assumption, though some authors have pointed out that the assumption of two full second order models does not necessarily hold (e.g. see [10]). In this paper, we will still stick to this assumption, i.e., the model matrices $\mathbf{X}$ and $\mathbf{Z}$ are identical, both containing columns for the intercept, linear terms, quadratic terms and cross product terms in order.

Optimization results from the three models are displayed in Table 1. In Table 1, $x_a$ denotes the optimal location found under the "target is the best" criterion, and $\hat{\mu}_a$, $\hat{\sigma}_a^2$, and $\widehat{MSE}_a$ are respectively the estimated mean, variance and $MSE$ at the location $x_a$. Similarly, notations $\hat{\mu}_b$, $\hat{\sigma}_b^2$ and $\widehat{MSE}_b$ are for the results found under the $MSE$ criterion. These subscripts will also be used for in Table 2.

Though the only difference between the OOLS and GOLS models is in the estimation of $\beta$, the optimal operating conditions picked using the two methods are not close to each other and the predicted means, variances, and $MSE$s also differ greatly. This may cause confusion among the practitioners on which point should be viewed as "optimal". On the other hand, the optimal locations picked out using the Bayesian hierarchical methods for different small values of $\lambda$ are very close to each other, and their respective predicted performances (such as predicted means, variances and $MSE$s) are quite similar.

Table 1: Optimization results comparison for the printing ink data

| Models | OOLS | GOLS | BAYES | | |
|---|---|---|---|---|---|
| | | | $\lambda = 10^{-1}$ | $\lambda = 10^{-3}$ | $\lambda = 10^{-6}$ |
| $\hat{x}_a$ | (1.0000, 0.1163, -0.2584) | (1.0000, 0.7189 -0.4484) | (0.4788, 0.1695, 0.1002) | (0.4769, 0.1729, 0.0984) | (0.4757, 0.1759, 0.0975) |
| $\hat{\mu}_a$ | 500 | 500 | 500 | 500 | 500 |
| $\hat{\sigma}_a^2 = \widehat{MSE}_a$ | 2034.4 | 2494.7 | 27.824 | 28.472 | 28.536 |
| $\hat{x}_b$ | (1.000, 0.0715, -0.2503) | (1.0000, 0.5537, -0.3864) | (0.4783, 0.1695, 0.0998) | (0.4765, 0.1726, 0.0981) | (0.4753, 0.1757, 0.0972) |
| $\hat{\mu}_b$ | 494.672 | 493.036 | 499.845 | 499.833 | 499.840 |
| $\hat{\sigma}_b^2$ | 1977.6 | 2421.6 | 27.783 | 28.416 | 28.483 |
| $\widehat{MSE}_b$ | 2005.9 | 2470.0 | 27.808 | 28.444 | 28.508 |

## 5.2 Catapult Data

The catapult data is from a Roman-style catapult experiment, first adopted by Luner ([11]) to illustrate for teaching purpose the use of RSM in quality improvement. The requirement for the catapult performance is that it must throw projectiles a distance of 80 inches with a high degree of precision. Based on the knowledge of the catapult production process, three factors, arm length ($x_1$), full stop angle ($x_2$), and pivot height($x_3$), are included as potential significant factors to catapult performance. A complete randomized central composite design (CCD) is carried out: 8 factorial points and 6 axial points with 3 replicates as each point, and 18 replicates at the central point. The response $y$ is defined as the distance, in inches, from the base of the catapult to the place where the projectile reaches the ground.

As in the previous analysis, full second order models are assumed for both the mean and the variance in each method in the analysis of the catapult data. Optimal points, $\mathbf{x}_0$'s, are searched under the two optimization criteria used for the printing ink data over the spherical region, $\mathbf{x}_0'\mathbf{x}_0 \leq 3$, since the design region is almost spherical. Table 2 summarizes the optimization results from the three methods. For this data, the optimal points found using the OOLS and GOLS methods are not as different as those in the printing ink data

set, and their predicted performances are much closer. The results from the BAYES method by using three different $\lambda$ values are still very similar to each other.

Table 2: Optimization results comparison for the catapult data

| Models | OOLS | GOLS | BAYES | | |
|---|---|---|---|---|---|
| | | | $\lambda = 10^{-1}$ | $\lambda = 10^{-3}$ | $\lambda = 10^{-6}$ |
| $\hat{x}_a$ | (0.0420, -0.2742, -0.2394) | (0.0435, -0.2829 -0.2227) | (-0.0616, -0.2285, -0.2302) | (-0.0573, -0.2177, -0.2345) | (-0.0618, -0.2427, -0.2263) |
| $\hat{\mu}_a$ | 80 | 80 | 80 | 80 | 80 |
| $\hat{\sigma}_a^2 = \hat{MSE}_a$ | 5.308 | 5.449 | 13.867 | 14.572 | 13.744 |
| $\hat{x}_b$ | (0.0595, -0.2741, -0.2633) | (0.0611, -0.2815, -0.2473) | (-0.0451, -0.1820, -0.2007) | (-0.0443, -0.1669, -0.2067) | (-0.0430, -0.1947, -0.1974) |
| $\hat{\mu}_b$ | 79.760 | 79.756 | 80.822 | 79.831 | 79.833 |
| $\hat{\sigma}_b^2$ | 5.070 | 5.210 | 14.167 | 14.167 | 14.100 |
| $\hat{MSE}_b$ | 5.127 | 5.270 | 14.842 | 14.195 | 14.128 |

# 6 Simulation Studies

Note that all the values of means, variances and $MSE$s in Tables 1 and 2 are predicted values given by the estimated models, and the locations suggested by each model are only optimal to the extent that the fitted model is correct. Therefore those values give us no reason to prefer one method to the others, as the true nature of the data sets is unknown to us. To assess the performance of each modelling approach, simulated data sets should be analyzed.

We use the following three different scenarios to do simulation:

- **Constant variance**:

$$y_{ij} \sim N(\underline{x}_i'\underline{\beta}, 1), \quad \text{for} \quad i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, n_j;$$

- Nonconstant variance where the standard deviation follows a normal distribution (referred to as **Nonconstant Stdev**):

$$y_{ij} \sim N(\underline{x}_i'\underline{\beta}, \sigma_i^2), \quad \sigma_i \sim N(\underline{z}_i'\underline{\gamma}, \eta^2), \quad \text{for} \quad i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, n_j.$$

  In generating the standard deviation, the minimum mean standard deviation over the experimental region is set to be much larger than $\eta$, so that the simulated standard deviation has little likelihood to be negative;

- Nonconstant variance where the variance follows a log-normal distribution (referred to as **Nonconstant Logvar**):

$$y_{ij} \sim N(\underline{x}_i'\underline{\beta}, \sigma_i^2), \quad \sigma_i^2 \sim \text{Log-Normal}(\underline{z}_i'\underline{\gamma}, \eta^2), \quad \text{for} \quad i = 1, 2, \ldots, N, \quad j = 1, 2, \ldots, n_j.$$

Since full second order models are assumed for both the mean and variance models in previous analyses, $\underline{x}_i'$ and $\underline{z}_i'$ will be set as identical in the simulation, both containing the intercept, linear terms, quadratic terms and interactions in order. For each scenario, 2500 data sets are generated from a $3^3$ ($N = 27$) factorial design with 3 replicates ($n_i = 3, \quad \text{for} \quad i = 1, 2, \ldots, 27$) at each design point. The simulation size is determined such that the largest standard error of the mean efficiency is less than 0.005. Each method builds two full second order models for the mean and variance surfaces and optimal points are searched over the cuboidal region ($-1 < x_1, x_2, x_3 < 1$), under the $MSE$ criterion. For the BAYES model, $\lambda = 0.001$ is assumed in the prior distribution of $\delta^2$ for the simulation studies as it has been shown that the effect of $\lambda$ is almost negligible in the analysis for the two real data sets.

The relative efficiency based on the $MSE$ criterion is introduced as a single-number measure of performance to see how well a model does in choosing "good" points overall. It is defined as the ratio between the true model's minimum $MSE$ and the true model's $MSE$ at the point proposed by that modelling approach:

$$\text{relative efficiency} = \frac{\min MSE_{\text{true}}}{MSE_{\text{point proposed}}}.$$

If the operating condition picked by a model performs closely to the true optimum, i.e., the relative efficiency is very near to 1, then the model is viewed as good. If the relative efficiency of a model is very small, that means the model picks a point far from optimum.

Table 3: Comparison of the relative efficiencies between the three methods using simulated Data. Simulation size is 2500 for each scenario. It is determined so that the largest standard error of the mean relative efficiency is less than 0.005.

| **Models** | **Constant Variance** | | | **Nonconstant Stdev** | | | **Nonconstant Logvar** | | |
|---|---|---|---|---|---|---|---|---|---|
| | Mean | Median | Stdev | Mean | Median | Stdev | Mean | Median | Stdev |
| OOLS | 0.9201 | 0.9561 | 0.0959 | 0.7714 | 0.8181 | 0.1693 | 0.5423 | 0.5500 | 0.2412 |
| GOLS | 0.9182 | 0.9552 | 0.0986 | 0.7695 | 0.8156 | 0.1707 | 0.5425 | 0.5463 | 0.2403 |
| BAYES | 0.9084 | 0.9484 | 0.1079 | 0.7693 | 0.8025 | 0.1642 | 0.7893 | 0.8337 | 0.1834 |

Table 3 gives numerical summaries of the relative efficiencies. When the variance is constant over the region of interest, all three methods do a good job, with the medians and means of the relative efficiency above 0.90. As the variance structure becomes complicated, their performances degrade differently. In the Nonconstant Stdev case, the results using three methods are not as good as those in the constant variance case but their performances are still fairly comparable. When the true variance has the logarithm structure, the BAYES model stands out among the three with the largest mean, median and the smallest variance of the relative efficiency. Case-wise speaking, the performance of the BAYES model is much more stable than the two least squares methods as the its performance does not decrease as dramatically as the OOLS and GOLS models when the variance structure becomes more complicated. Since in dual response surface problems the variance is assumed to be non-constant, the performance under the latter two scenarios should be given more emphasis. Based on the simulation studies on the two nonconstant variance cases, the BAYES model is recommended.

# 7 Summary and Conclusion

In the literature of dual response surface, much work has been devoted into the improvement of optimization criterion, the optimization algorithms and implementations. However, if the models established do not fit the true response surfaces well, then the optimization results based on the modelling are not trustworthy.

This paper is to develop a Bayesian hierarchical model for dual response system and assess its performance by comparing it to the two existing frequentist methods. From the studies on the literature data sets and the simulated data, it is shown that the results from the two frequentist methods are not stable: in the real data analysis, the two methods pick two optimal locations far away from each other under the same optimization criterion; in simulation studies, when the true variances have the logarithm structure, the relative efficiencies are only about 0.5. Contrary to the instability of the two least squares methods, the BAYES model keeps performing well in both the real data analysis and the simulation studies. The consistent performance of the BAYES model excludes ambiguity in parameter estimation. The analysts can simply choose the posterior medians as the estimates for the parameters.

In practice, the BAYES model retains the advantages when the dual response approach was initiated: they have two separate estimated models which allow them to monitor both the mean characteristic and the process variability. They can make any necessary adjustment and compromise as they have done under the frequentist models. Another advantage that the Bayesian approach possesses over the frequentist approach is in incorporating prior information from the previous studies into the current data analysis, which accords with the sequentiality nature of most response surface problems (e.g. see Myers and Montegomery [12]). Moreover, as the Bayesian methodology works with the probability distributions, it can be given the predictive distribution of the response at any point within the region of interest. Thus the analysts will be given a more comprehensive view of the response of interest rather than merely the mean and the variance. This would be of future research interest.

## APPENDIX

# A Posterior distributions (conventional informative priors)

**Theorem A.1.** Suppose $y_{ij}|\underline{\beta}, \sigma_i^2 \sim N(\underline{x}_i'\underline{\beta}, \sigma_i^2)$ and $\ln \sigma_i^2|\underline{\gamma}, \delta^2 \sim N(\underline{z}_i'\underline{\gamma}, \delta^2)$. The following prior distributions are assumed for $\underline{\beta}$, $\underline{\gamma}$, and $\delta^2$:

$$\pi(\underline{\beta}) \sim \text{MVN}_p(\underline{\mu}_\beta, \Sigma_\beta), \quad \pi(\underline{\gamma}) \sim \text{MVN}_q(\underline{\mu}_\gamma, \Sigma_\gamma), \quad \pi(\delta^2) \sim \text{Inverse-Gamma}(\alpha, \tau),$$

where $\underline{\mu}_\beta$, $\Sigma_\beta$, $\underline{\mu}_\gamma$, $\Sigma_\gamma$, $\alpha$, and $\tau$ are the respective mean vectors and covariance matrices of $\underline{\beta}$ and $\underline{\gamma}$ in the prior distributions, and $\alpha$, $\tau$ are the shape and scale parameters of the Inverse-Gamma distribution, respectively. Then the joint posterior distribution of $(\underline{\beta}, \underline{\gamma}, \sigma_1^2, \ldots, \sigma_N^2, \delta^2)$ is:

$$\begin{aligned}
&\pi(\underline{\beta}, \underline{\gamma}, \sigma_1^2, \ldots, \sigma_N^2, \delta^2 | Y) \\
\propto &\frac{1}{(\delta^2)^{N/2+\alpha+1} |\Sigma_\beta|^{1/2} |\Sigma_\gamma|^{1/2} \prod_{i=1}^N (\sigma_i^2)^{n_i/2+1}} \exp\left[ -\sum_{i=1}^N \frac{(n_i - 1)s_i^2 + n_i(\bar{y}_i - \underline{x}_i'\underline{\beta})^2}{2\sigma_i^2} \right] \\
\times &\exp\left[ -\frac{1}{2\delta^2}(\underline{d} - Z\underline{\gamma})'(\underline{d} - Z\underline{\gamma}) - \frac{1}{2}(\underline{\beta} - \underline{\mu}_\beta)'\Sigma_\beta^{-1}(\underline{\beta} - \underline{\mu}_\beta) - \frac{1}{2}(\underline{\gamma} - \underline{\mu}_\gamma)'\Sigma_\gamma^{-1}(\underline{\gamma} - \underline{\mu}_\gamma) - \frac{\tau}{\delta^2} \right],
\end{aligned}$$

where $\underline{d} = (d_1, d_2, \ldots, d_N)' = (\ln \sigma_1^2, \ln \sigma_2^2, \ldots, \ln \sigma_N^2)'$.

*Proof.* The joint posterior distribution is proportional to the product of the probability density of the data $Y$ and all the prior distributions since independence among observations and parameters is assumed in Section 4.1.

The probability density of $Y$ conditional on $\underline{\beta}$ and $(\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2)$ is:

$$f(Y | \underline{\beta}, \sigma_1^2, \ldots, \sigma_N^2) = \prod_{i=1}^N \frac{1}{(2\pi\sigma_i^2)^{n_i/2}} \exp\left[ -\frac{\sum_{j=1}^{n_i} (y_{ij} - \underline{x}_i'\underline{\beta})^2}{2\sigma_i^2} \right].$$

The probability density function of $\sigma_i^2$ given $\underline{\gamma}$ and $\delta^2$ is:

$$f(\sigma_i^2 | \underline{\gamma}, \delta^2) = \frac{1}{\sigma_i^2 \sqrt{2\pi\delta^2}} \exp\left[ -\frac{1}{2}(\ln \sigma_i^2 - \underline{z}_i'\underline{\gamma})^2 \right].$$

The product of the above conditional distributions and the assumed prior distributions on $\underline{\beta}$, $\underline{\gamma}$, and $\delta^2$ yield the joint posterior distribution of $(\underline{\beta}, \underline{\gamma}, \sigma_1^2, \ldots, \sigma_N^2, \delta^2)$. □

**Theorem A.2.** Suppose we have the same assumptions as in Theorem A.1. Let $\bar{Y}$ be the vector of sample means at design points, $\bar{Y} = (\bar{y}_1, \bar{y}_2, \ldots, \bar{y}_N)'$, $V$ the variance covariance

matrix for $\bar{Y}$, and $\underline{t} = (t_1, t_2, \ldots, t_N)'$. The full conditional posterior distributions are:

$$\begin{aligned}
\pi(\underline{\beta}|\text{others}) &\sim \text{MVN}_p\left((X'V^{-1}X + \Sigma_\beta^{-1})^{-1}(X'V^{-1}\bar{Y} + \Sigma_\beta^{-1}\underline{\mu}_\beta), (X'V^{-1}X + \Sigma_\beta^{-1})^{-1}\right), \\
\pi(\underline{\gamma}|\text{others}) &\sim \text{MVN}_q\left((Z'Z/\delta^2 + \Sigma_\gamma^{-1})^{-1}(Z'\underline{d}/\delta^2 + \Sigma_\gamma^{-1}\underline{\mu}_\gamma), (Z'Z/\delta^2 + \Sigma_\gamma^{-1})^{-1}\right), \\
\pi(\delta^2|\text{others}) &\sim \text{Inverse-Gamma}\left(N/2 + \alpha, \frac{(\underline{d} - Z\underline{\gamma})'(\underline{d} - Z\underline{\gamma}) + 2\tau}{2}\right), \\
\pi(\sigma_i^2|\text{others}) &\propto \frac{1}{(\sigma_i^2)^{n_i/2+1}} \exp\left\{-\frac{1}{2\sigma_i^2}[(n_i - 1)s_i^2 + n_i(\bar{y}_i - \underline{x}_i'\underline{\beta})^2] - \frac{1}{2\delta^2}(\ln\sigma_i^2 - \underline{z}_i'\underline{\gamma})^2\right\}, \\
&\quad \text{for} \quad i = 1, 2, \ldots, N.
\end{aligned}$$

*Proof.* To derive the conditional posterior distribution for certain variable(s), we simply keep terms involving the variable(s) of interest, and remove the other terms from the joint posterior distributions. □

# B Posterior distributions (noninformative priors)

**Corollary B.1.** Suppose $y_{ij}|\underline{\beta}, \sigma_i^2 \sim N(\underline{x}_i'\underline{\beta}, \sigma_i^2)$ and $\ln\sigma_i^2|\underline{\gamma}, \delta^2 \sim N(\underline{z}_i'\underline{\gamma}, \delta^2)$. In the absence of historical information, the following prior distributions are assumed for $\underline{\beta}$, $\underline{\gamma}$, and $\delta^2$:

$$\pi(\underline{\beta}) \propto 1 \quad \pi(\underline{\gamma}) \propto 1 \quad \pi(\delta^2) \propto \frac{1}{\delta^2}\exp(-\frac{\lambda}{\delta^2}),$$

where $\lambda$ is a small positive number. Then the joint posterior distribution of $(\underline{\beta}, \underline{\gamma}, \sigma_1^2, \ldots, \sigma_N^2, \delta^2)$ is:

$$\begin{aligned}
&\pi(\underline{\beta}, \underline{\gamma}, \sigma_1^2, \ldots, \sigma_N^2, \delta^2|Y) \\
\propto &\frac{1}{(\delta^2)^{N/2+1}\prod_{i=1}^N(\sigma_i^2)^{n_i/2+1}} \exp\left[-\sum_{i=1}^N \frac{(n_i - 1)s_i^2 + n_i(\bar{y}_i - \underline{x}_i'\underline{\beta})^2}{2\sigma_i^2} - \frac{(\underline{d} - Z\underline{\gamma})'(\underline{d} - Z\underline{\gamma}) + 2\lambda}{2\delta^2}\right].
\end{aligned}$$

**Corollary B.2.** Suppose we have the same assumptions as in Corollary B.1. The full

conditional posterior distributions are:

$$\begin{aligned}
\pi(\underline{\beta}|\text{others}) &\sim \text{MVN}_p((X'V^{-1}X)^{-1}X'V^{-1}\bar{Y}, (X'V^{-1}X)^{-1}),\\
\pi(\underline{\gamma}|\text{others}) &\sim \text{MVN}_q((Z'Z)^{-1}Z'\underline{d}, (Z'Z)^{-1}\delta^2),\\
\pi(\delta^2|\text{others}) &\sim \text{Inverse-Gamma}\left(\frac{N}{2}, \frac{(\underline{d}-Z\underline{\gamma})'(\underline{d}-Z\underline{\gamma})+2\lambda}{2}\right),\\
\pi(\sigma_i^2|\text{others}) &\propto \frac{1}{(\sigma_i^2)^{n_i/2+1}}\exp\left\{-\frac{1}{2\sigma_i^2}[(n_i-1)s_i^2+n_i(\bar{y}_i-\underline{x}_i'\underline{\beta})^2]-\frac{1}{2\delta^2}(\ln\sigma_i^2-\underline{z}_i'\underline{\gamma})^2\right\},\\
&\quad \text{for} \quad i=1,2,\ldots,N.
\end{aligned}$$

Both Corollaries can be proven easily by using Theorems A.1 and A.2.

# C Propriety of the posterior distributions (noninformative priors)

**Theorem C.1.** Let $\Omega_{m\times m}$ be a real symmetric positive definite matrix having eigenvalues $\lambda_1 \le \lambda_2 \le \cdots \le \lambda_m$ with associated orthonormal $(m\times 1)$ eigenvectors $v_1, v_2, \cdots, v_m$. Then

$$\min\{\det(B'\Omega B) : B_{m\times n} \quad \text{real}, B'B = I_n\} = \lambda_1\lambda_2\cdots\lambda_n,$$

where $B = [v_1, v_2, \cdots, v_n]$.

Theorem C.1 and its proof can be found in Horn and Johnson [8]. It is used in the following to show that the posterior distribution by using the noninformative priors in Corollary B.1 is proper when the numbers of replicates at design points, $n_i$'s are all greater than 1.

**Theorem C.2.** Suppose we have the same assumptions as in Corollary B.1. Then the posterior distribution of $(\underline{\beta}, \underline{\gamma}, \sigma_1^2, \ldots, \sigma_N^2, \delta^2)$ is proper when $n_i > 1 \quad \text{for} \quad i = 1, 2, \ldots, N$.

*Proof.* Suppose the mean model matrix $X$ is of full row rank $p$, then $X'X$ is an $p\times p$ symmetric matrix with corresponding eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_p$ and normalized eigenvectors $c_1, c_2, \cdots, c_p$. With spectral decomposition, $X'X$ can be expressed as

$$X'X = C\Lambda C' \tag{C.2}$$

where $\Lambda = \text{diag}\{\lambda_i\}$, a diagonal matrix with $\lambda_i$ as the $i^{th}$ element, and $C$ is the orthogonal matrix $C = (c_1, c_2, \cdots, c_p)$. Since $C'C = I_p$, (C.2) can be written as

$$C'X'XC = \Lambda,$$

and then

$$\Lambda^{-\frac{1}{2}}C'X'XC\Lambda^{-\frac{1}{2}} = B'B = I_p, \tag{C.3}$$

where $B = XC\Lambda^{-\frac{1}{2}}$ and $\Lambda^{-\frac{1}{2}} = \text{diag}\{\lambda_i^{-\frac{1}{2}}\}$.

It is known from Corollary B.1 that the joint posterior distribution of $(\underline{\beta}, \underline{\gamma}, \sigma_1^2, \ldots, \sigma_N^2, \delta^2)$ is:

$$\begin{aligned}
&\pi(\underline{\beta}, \underline{\gamma}, \sigma_1^2, \ldots, \sigma_N^2, \delta^2|Y) \\
\propto &\frac{1}{(\delta^2)^{N/2+1}\prod_{i=1}^N(\sigma_i^2)^{n_i/2+1}} \exp\left[-\sum_{i=1}^N \frac{(n_i-1)s_i^2 + n_i(\bar{y}_i - \underline{x}_i'\underline{\beta})^2}{2\sigma_i^2} - \frac{(\underline{d} - Z\underline{\gamma})'(\underline{d} - Z\underline{\gamma}) + 2\lambda}{2\delta^2}\right].
\end{aligned}$$

After integrating over $\underline{\gamma}$, $\delta^2$ and $\underline{\beta}$, we have

$$\begin{aligned}
\pi(\sigma_1^2, \ldots, \sigma_N^2|Y) = &\frac{\mathcal{C}_1}{|X'D^{-1}X|^{1/2}[\underline{d}'(I - M_z)\underline{d} + 2\lambda]^{\frac{N-q}{2}}\prod_{i=1}^N(\sigma_i^2)^{n_i/2+1}} \\
&\times \exp\left[-\sum_{i=1}^N \frac{(n_i-1)s_i^2}{2\sigma_i^2} - \frac{1}{2}\overline{\underline{Y}}'(D^{-1} - M_{x,D})\overline{\underline{Y}}\right],
\end{aligned}$$

where $M_z = Z(Z'Z)^{-1}Z'$, $M_{x,D} = D^{-1}X(X'D^{-1}X)^{-1}X'D^{-1}$ and $\mathcal{C}_1$ is the normalizing constant in the marginal posterior distribution of $(\sigma_1^2, \ldots, \sigma_N^2)$.

Since both $(I - M_z)$ and $(D^{-1} - M_{x,D})$ are positive semidefinite, $\underline{d}'(I - M_z)\underline{d}$ and $\overline{\underline{Y}}'(D^{-1} - M_{x,D})\overline{\underline{Y}}$ are nonnegative, therefore

$$\pi(\sigma_1^2, \ldots, \sigma_N^2|Y) \leq \frac{\mathcal{C}_1\mathcal{C}_2}{|X'D^{-1}X|^{1/2}\prod_{i=1}^N(\sigma_i^2)^{n_i/2+1}} \exp\left[-\sum_{i=1}^N \frac{(n_i-1)s_i^2}{2\sigma_i^2}\right], \tag{C.4}$$

where $\mathcal{C}_2 = (2\lambda)^{\frac{N-q}{2}}$. Let

$$\frac{1}{\eta_i} = \frac{(n_i-1)s_i^2}{\sigma_i^2}, \quad \text{for} \quad i = 1, 2, \cdots, N.$$

Order $\eta$'s such that

$$\eta_{(1)} \geq \eta_{(2)} \geq \cdots \geq \eta_{(N)} \quad \text{and} \quad \frac{1}{\eta_{(1)}} \leq \frac{1}{\eta_{(2)}} \leq \cdots \leq \frac{1}{\eta_{(N)}}.$$

Then

$$\begin{aligned} X'D^{-1}X &= X'\text{diag}\{\frac{n_i}{\sigma_i^2}\}X = X'\text{diag}\left\{\frac{n_i}{(n_i-1)s_i^2\eta_i}\right\}X \\ &= \left(\text{diag}\left\{\sqrt{\frac{n_i}{(n_i-1)s_i^2}}\right\}X\right)'\text{diag}\left\{\frac{1}{\eta_i}\right\}\left(\text{diag}\left\{\sqrt{\frac{n_i}{(n_i-1)s_i^2}}\right\}X\right). \end{aligned}$$

Suppose the diagonal matrix $\Lambda = \text{diag}\{\lambda_i\}$ consists of the eigenvalues of the symmetric matrix $\left(\text{diag}\left\{\sqrt{\frac{n_i}{(n_i-1)s_i^2}}\right\}X\right)'\left(\text{diag}\left\{\sqrt{\frac{n_i}{(n_i-1)s_i^2}}\right\}X\right)$, and the orthogonal matrix $P = (p_1, p_2, \cdots, p_N)$ contains the corresponding normalized eigenvectors. Take the determinant of $X'D^{-1}X$:

$$\begin{aligned} |X'D^{-1}X| &= \left|\left(\text{diag}\left\{\sqrt{\frac{n_i}{(n_i-1)s_i^2}}\right\}X\right)'\text{diag}\left\{\frac{1}{\eta_i}\right\}\left(\text{diag}\left\{\sqrt{\frac{n_i}{(n_i-1)s_i^2}}\right\}X\right)\right| \\ &= |\Lambda|\cdot|\Lambda^{-1/2}|\cdot|P'|\cdot\left|X_*'\text{diag}\left\{\frac{1}{\eta_i}\right\}X_*\right|\cdot|P|\cdot|\Lambda^{-1/2}| \\ &= |\Lambda|\cdot\left|\Lambda^{-1/2}P'X_*'\text{diag}\left\{\frac{1}{\eta_i}\right\}X_*P\Lambda^{-1/2}\right| = |\Lambda|\cdot\left|U'\text{diag}\left\{\frac{1}{\eta_i}\right\}U\right|, \end{aligned}$$

where $X_* = \text{diag}\left\{\sqrt{\frac{n_i}{(n_i-1)s_i^2}}\right\}X$, and $U = X_*P\Lambda^{-1/2}$. With (C.3), it is known that $U'U = \Lambda^{-1/2}P'X_*'X_*P\Lambda^{-1/2} = I$. Use Theorem C.1:

$$|X'D^{-1}X|^{1/2} = |\Lambda|^{1/2}\cdot\left|U'\text{diag}\left\{\frac{1}{\eta_i}\right\}U\right| \geq |\Lambda|^{1/2}\cdot\prod_{i=1}^{p}\frac{1}{\eta_{(i)}}.$$

Let $f^*(\sigma_1^2, \cdots, \sigma_N^2)$ denote the right side of (C.4), then

$$\begin{aligned} f^*(\sigma_1^2, \cdots, \sigma_N^2) &\leq \frac{\mathcal{C}_1\mathcal{C}_2\mathcal{C}_3}{\left(\prod_{i=1}^{p}\frac{1}{\eta_{(i)}}\right)^{1/2}\left(\prod_{i=1}^{N}\eta_i^{n_i/2+1}\right)}\exp\left(-\sum_{i=1}^{N}\frac{1}{2\eta_i}\right) \\ &= \frac{\mathcal{C}_1\mathcal{C}_2\mathcal{C}_3}{\left(\prod_{i=p+1}^{N}\eta_{(i)}\right)^{1/2}\left(\prod_{i=1}^{N}\eta_i^{(n_i+1)/2}\right)}\exp\left(-\sum_{i=1}^{N}\frac{1}{2\eta_i}\right), \end{aligned} \tag{C.5}$$

where $\mathcal{C}_3 = \prod_{i=1}^{N}[(n_i-1)s_i^2]^{n_i/2+1}/|\Lambda|^{1/2}$.

Break the exponential form $\exp\left(-\sum_{i=1}^{N}\frac{1}{2\eta_i}\right)$ into two parts and write one of them in

terms of order statistics $\eta_{(i)}$, then the right hand side of (C.5) becomes:

$$\begin{aligned}
& \mathcal{C}_1\mathcal{C}_2\mathcal{C}_3 \cdot \frac{1}{\prod_{i=p+1}^{N} \eta_{(i)}^{1/2}} \exp\left(-\sum_{i=1}^{N} \frac{1}{4\eta_{(i)}}\right) \cdot \frac{1}{\prod_{i=1}^{N} \eta_i^{(n_i+1)/2}} \exp\left(-\sum_{i=1}^{N} \frac{1}{4\eta_i}\right) \\
\leq \; & \mathcal{C}_1\mathcal{C}_2\mathcal{C}_3 \cdot \frac{1}{\prod_{i=p+1}^{N} \eta_{(i)}^{1/2}} \exp\left(-\sum_{i=p+1}^{N} \frac{1}{4\eta_{(i)}}\right) \cdot \frac{1}{\prod_{i=1}^{N} \eta_i^{(n_i+1)/2}} \exp\left(-\sum_{i=1}^{N} \frac{1}{4\eta_i}\right) \\
\leq \; & (2e^{-1/2})^{N-p}\mathcal{C}_1\mathcal{C}_2\mathcal{C}_3 \cdot \frac{1}{\prod_{i=1}^{N} \eta_i^{(n_i+1)/2}} \exp\left(-\sum_{i=1}^{N} \frac{1}{4\eta_i}\right), \qquad \text{(C.6)}
\end{aligned}$$

since $\max\left(\frac{1}{\eta_{(i)}} \exp(-\frac{1}{4\eta_{(i)}})\right) = 4e^{-1}$ for any $i$.

The right hand side of (C.6) is integrable when all $n_i$'s are greater than 1, therefore the posterior distribution is also integrable and thus proper. □

# References


[1] Ames, A. E., Mattucci, N., Szonyi, G., and Hawkins, D. M. Quality loss function for optimization across mulitple response surfaces. *Journal of Quality Technology 29* (1997), 339–346.

[2] Bartlett, M. S., and Kendall, D. G. The statistical analysis of variance heterogeneity and the logarithmic transformation. *Journal of the Royal Statistical Society B8* (1946), 128–150.

[3] Copeland, K. A. F., and Nelson, P. R. Dual response optimization via direct function minimization. *Journal of Quality Technology 28* (1996), 331–336.

[4] De Jong, K. A. *An analysis of the behavior of a class of genetic adaptive systems.* PhD thesis, 1975.

[5] Del Castillo, E., and D.C., M. A nonlinear programming solution to the dual response problem. *Journal of Quality Technology 25* (1993), 199–204.

[6] Engel, J., and Huele, A. F. A generalized linear modelling approach to robust design. *Technometrics 38* (1996), 365–373.

[7] Holland, J. H. *Adaptation in natural and artificial systems.* Ann Arbor, MI: University of Michigan Press, 1975.

[8] Horn, R. A., and Johnson, C. R. *Matrix Analysis.* Cambridge University Press, Cambridge, 1985.

[9] Kim, K. J., and Lin, D. K. J. Dual response surface optimization: A fuzzy modelling approach. *Journal of Quality Technology 30* (1995), 1–10.

[10] Lin, D. k. J., and Tu, W. Dual response surface optimization. *Journal of Quality Technology 27* (1995), 34–39.

[11] Luner, J. J. Achieving continuous improvement with the dual response approach: A demonstration of the roman catapult. *Quality Engineering 6* (1994), 691–705.

[12] Myers, R. H., and C, M. D. *Response Surface Methodology: Process and Product Optimizaiton Using Design Experiments.* John Wileys & Sons, Inc, 1995.

[13] Myers, R. H., and Carter, W. H. Response surface techniques for dual response systems. *Technometrics 15* (1973), 301–317.

[14] Nair, V. N., Abraham, B., MacKay, J., Box, G., Kacker, R., Lorenzen, T., Lucas, J., Myers, R., Vining, G., Nelder, J., Phadke, M., Sacks, J., Welch, W., Shoemaker, A., Tsui, K., Taguchi, S., and Wu, C. F. J. Taguchi's parameter design: A panel discussion. *Technometrics 34* (1992), 127–161.

[15] Robert, C. P., and Casella, G. *Monte Carlo Statistical Methods.* Springer-Verlag Inc, 1985.

[16] Taguchi, G. Introduction to quality engineering: designing quality into products and processes. Tech. rep., Asian Productivity Organization, Tokyo, 1985.

[17] Tang, L. C., and Xu, K. A. A unified approach for dual response optimization. *Journal of Quality Technology 34* (2002), 437–447.

[18] Vining, G. G., and Myers, R. H. Combining taguchi and response surface philosophies: A dual response surface approach. *Journal of Quality Technology 22* (1990), 38–45.