

# Model-Robust Process Optimization using Bayesian Model Averaging

Ramkumar Rajagopal

Enrique del Castillo

Dept. of Industrial & Manufacturing Engineering

The Pennsylvania State University

University Park, PA 16802

December 8, 2003

UNDER REVIEW. PLEASE DO NOT QUOTE WITHOUT AUTHORS' PERMISSION.

## Abstract

Traditional approaches for process optimization start by fitting a model and then optimizing the model to obtain optimal operating settings. These methods do not account for any uncertainty in the parameters of the model or in the form of the model. Bayesian approaches have been proposed recently to account for the uncertainty on the parameters of the model, assuming the model form is known [15]. This paper presents a Bayesian predictive approach to process optimization that accounts for the uncertainty in the model *form*, also accounting for the uncertainty of the parameters given each potential model. It is proposed to optimize the model-averaged posterior predictive density (MAP) of the response where the weighted average is taken using the model posterior probabilities as weights. The resulting model-robust optimization is illustrated with two experiments from the literature, one involving a mixture experiment and the other a small composite design.

Keywords: Response Surface Methodology, Bayesian Model Averaging, Predictive Density.

## 1 Introduction: Process Optimization

In the “end game” of Response Surface Methodology (RSM, see [2, 14]), optimization of a process traditionally consists of two steps. The first step is to design the experiment, collect data and fit a model, usually of second order or higher to allow for curvature. Once the model is fitted, the next step is to optimize the response based on the fitted model and obtain estimated optimal operating settings. The second step in this process strongly depends on the assumption that the fitted model is the correct representation of this process.

It is possible that a second different model, which arguably fits the data as well as the first model, provides considerably different optimal operating conditions (see our example section for cases when this occurs). A frequentist approach, common in RSM practice, is to assess the effect of the uncertainty of the parameter estimates on the optimum by computing a confidence region on the location of the optimum (e.g., see [17]). A more recent Bayesian approach implicitly considers the uncertainty of the parameters given the model form [15, 13]. The technique used by these authors involves obtaining the posterior predictive density of the response based on the assumed model, and maximizing the probability of obtaining the predicted response to lie within certain limits or specifications.

In this paper, the Bayesian predictive approach density is taken one step further by averaging over possible competing models. Here, no single model is assumed. Instead, as a first step, the Bayesian posterior probabilities for all possible models (belonging to a class (or classes) of models that are appropriate for the process) are calculated. Once the model posteriors are determined, the next step is to determine the posterior predictive density of the response for each of the competing models. The model-averaged posterior predictive density (MAP) is then computed by taking the weighted average of the densities over all competing models. The model posteriors computed earlier are used as the weights. The MAP is then used to maximize the probability of obtaining a response value within the given specification limits.

As the uncertainty in the model is more acute in the case where there are fewer runs, the examples provided in the later sections will focus on smaller designs. However, the main idea can be applied to any design where the form of the best model is in question.

In the next section, the technical details about the application of Bayesian model averaging to process optimization are discussed. The predictive approach we adopt focuses on making inferences on future values of the “observable”  $y$  [6]. For doing this, the posterior predictive density of the response under a particular choice of priors and the assumed likelihood needs to be derived. This is discussed in section 2 and the details are shown in

Appendix B. This is followed by two examples, one of which is a mixture experiment and the other a small composite design.

## 2 Bayesian Model Averaging

We consider a process with a single response variable  $y$  which is dependent on a  $(p \times 1)$  vector of regressors  $\mathbf{x}$  that are in turn functions of the  $k$  controllable factors. It is assumed that a suitable experiment with  $n$  runs has been designed and carried out and the data from the experiment is available. The vector of responses from the experiment is given by the  $(n \times 1)$  vector  $\mathbf{y}$ . Each observation of the model is assumed to be generated from a model linear in the parameters of the form

$$y = \mathbf{x}'\boldsymbol{\beta} + \epsilon \quad (1)$$

where,  $\epsilon$  is the error term, and  $\boldsymbol{\beta}$  is the vector of process parameters (i.e.,  $\mathbf{x}$  is in model form). The particular functional form of the  $\mathbf{x}'\boldsymbol{\beta}$  term (a function of the  $k$  controllable factors) is not known with certainty. In this paper, we focus on the case where  $\epsilon$  is normally distributed.

It is assumed that the goal of the optimization is to identify the values of the controllable factors that result in a response  $y$  such that  $L < y < U$  where  $L$  denotes a given lower bound (or specification) and  $U$  denotes a given upper bound. The approach adopted here maximizes the posterior predictive probability of obtaining a response  $y$  within these bounds. In this procedure, we first list the potential models that are under consideration based on a family or families of models. Next, the posterior probability of each of these  $i$  models given the experimental data,  $P(M_i|\mathbf{y})$ , is calculated. The posterior predictive density of the response is then calculated for each model  $M_i$  as a function of the controllable variables. This is denoted by  $P(y^*|M_i, \mathbf{x}^*, \mathbf{y})$ , where  $y^*$  is the predicted value of the response at a new set of observed regressors  $\mathbf{x}^*$ . In order to average the predictive density of the response over all competing models, we take the weighted average of  $P(y^*|M_i, \mathbf{x}^*, \mathbf{y})$ , using the model posteriors,  $P(M_i|\mathbf{y})$ , as the weights. The model-averaged posterior predictive density (MAP)

is thus of the form of a mixture of distributions over all competing models, namely:

$$MAP = P(y^*|\mathbf{x}^*, \mathbf{y}) = \sum_i P(y^*|\mathbf{x}^*, \mathbf{y}, M_i)P(M_i|\mathbf{y}). \quad (2)$$

The optimal control variables are then determined by maximizing the probability that the predicted response lies within the target bounds, i.e.,

$$\max_{x_1, \dots, x_k} P(L \leq Y^* \leq U) = \sum_i \left[ \int_L^U P(y^*|M_i, \mathbf{x}^*, \mathbf{y}) dy^* \right] P(M_i|\mathbf{y}) \quad (3)$$

where the maximization is over the  $k$  control factors  $(x_1, x_2, \dots, x_k)$  that  $\mathbf{x}^*$  depends on. It should be pointed out that this approach does not average the optimal levels of the controllable factors for each model. Instead, the optimal levels of the controllable factors are prescribed by averaging the predictive density of the response over all models. Constraints on the controllable factors  $x_i$  can be included in (3) if desired.

## 2.1 Calculating model posteriors

There is considerable literature on the calculation of model posterior probabilities (e.g., see [10], [19] and the references therein). The most common approach is to assume a candidate list of models based on a class (or classes) of models with  $f_i$  out of the total  $k$  factors present in model  $M_i$ . A useful method to determine model priors, proposed by Meyer et al. [10, 11, 12] is to choose the model priors based on the active factors (i.e., the factors present in each model). Denote the probability of factor  $j$  to be active as  $\pi_j$ ,  $j \in \{1, \dots, k\}$ . Assuming that the prior probabilities of active factors are independent, the model prior is given by

$$P(M_i) = \prod_{j \in M_i} (\pi_j) \prod_{j' \notin M_i} (1 - \pi_{j'}). \quad (4)$$

If  $\pi_j = \pi$ ,  $\forall j \in \{1, \dots, k\}$ , then  $P(M_i) = \pi^{f_i}(1 - \pi)^{k-f_i}$ . Let  $r_i$  be the number of terms in model  $M_i$  and  $t_i$  be the number of terms in model  $M_i$  excluding the constant term. Thus, if the model includes a constant term, we have that  $t_i = r_i - 1$ , otherwise  $t_i = r_i$ . Let  $\mathbf{X}_i$  be the  $(n \times r_i)$  design matrix corresponding to  $M_i$ . The posterior probability of  $M_i$  is given by

Bayes' theorem,

$$P(M_i|\mathbf{y}) = \frac{P(\mathbf{y}|M_i)P(M_i)}{\sum_i P(\mathbf{y}|M_i)P(M_i)}, \quad (5)$$

where  $P(\mathbf{y}|M_i)$  is the marginal likelihood of the model given the data. This marginal is defined as:

$$P(\mathbf{y}|M_i) = \int_{\sigma^2} \int_{\boldsymbol{\beta}_i} P(\mathbf{y}|M_i, \sigma^2, \boldsymbol{\beta}_i) P(\sigma^2, \boldsymbol{\beta}_i|M_i) d\boldsymbol{\beta}_i d\sigma^2, \quad (6)$$

where  $P(\mathbf{y}|M_i, \sigma^2, \boldsymbol{\beta}_i)$  is the likelihood function. Under the assumption of normality of the error terms, the likelihood is given by:

$$P(\mathbf{y}|M_i, \sigma^2, \boldsymbol{\beta}_i) \propto \sigma^{-n} \exp \left[ \frac{-1}{2\sigma^2} (\mathbf{y} - \mathbf{X}_i \boldsymbol{\beta}_i)' (\mathbf{y} - \mathbf{X}_i \boldsymbol{\beta}_i) \right]. \quad (7)$$

Let  $P(\sigma^2, \boldsymbol{\beta}_i|M_i)$  be the joint prior of the model parameters. The parameters  $\boldsymbol{\beta}_i$  and  $\sigma^2$  are assumed to be independent a priori. The priors on the parameters are chosen as

$$\boldsymbol{\beta}_i \sim N(\mathbf{0}, \boldsymbol{\Sigma}_i \sigma^2) \quad (8)$$

$$P(\sigma^2) \propto \frac{1}{\sigma^2} \quad (9)$$

$$P(\boldsymbol{\beta}_i, \sigma^2) = P(\sigma^2)P(\boldsymbol{\beta}_i) \quad (10)$$

Here, we choose  $\boldsymbol{\Sigma}^{-1} = (\mathbf{X}_i' \mathbf{X}_i) \mathbf{V}_i$ , where  $\mathbf{V}_i = \frac{1}{g} \begin{pmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{t_i} \end{pmatrix}$ , and  $g$  is a parameter whose value is to be chosen. Thus, the priors on all the  $\beta_i$ 's except for the constant term are assumed to be normally distributed using Zellner's g-prior [21]. The priors on the  $\beta_i$  for the constant term and on  $\log(\sigma^2)$  are assumed to be non-informative. A discussion on our choice of priors is included in section 4.

From the assumed priors and from equation (7), the integral in equation (6) can be computed and yields (see [12], [19]):

$$P(\mathbf{y}|M_i) \propto \gamma^{-t_i} |\boldsymbol{\Sigma}_i^{-1} + \mathbf{X}_i' \mathbf{X}_i|^{-\frac{1}{2}} S_i^{-\frac{(n-1)}{2}}, \quad (11)$$

where  $\gamma$  is such that

$$\frac{g}{\gamma^2} \mathbf{V}_i = \boldsymbol{\Sigma}^{-1}. \quad (12)$$

Then, by omitting the constant denominator in equation (5), we get

$$P(M_i|\mathbf{y}) \propto \pi^{f_i}(1 - \pi)^{k-f_i} \gamma^{-t_i} |\boldsymbol{\Sigma}_i^{-1} + \mathbf{X}_i' \mathbf{X}_i|^{-\frac{1}{2}} S_i^{-\frac{(n-1)}{2}}, \quad (13)$$

where

$$S_i = (\mathbf{y} - \mathbf{X}_i \hat{\boldsymbol{\beta}}_i)' (\mathbf{y} - \mathbf{X}_i \hat{\boldsymbol{\beta}}_i) + \hat{\boldsymbol{\beta}}_i' \boldsymbol{\Sigma}_i^{-1} \hat{\boldsymbol{\beta}}_i \quad (14)$$

$$= \mathbf{y}' \mathbf{y} - \mathbf{y}' \mathbf{X}_i (\boldsymbol{\Sigma}_i^{-1} + \mathbf{X}_i' \mathbf{X}_i)^{-1} \mathbf{X}_i' \mathbf{y}, \quad (15)$$

and

$$\hat{\boldsymbol{\beta}}_i = (\boldsymbol{\Sigma}_i^{-1} + \mathbf{X}_i' \mathbf{X}_i)^{-1} \mathbf{X}_i' \mathbf{y}. \quad (16)$$

Here,  $S_i$  is the Bayesian analog to the residual sum of squares and  $\hat{\boldsymbol{\beta}}_i$  gives the parameter estimates for model  $M_i$ . The probabilities for each of the models computed from the above equations are scaled by dividing each one of them by the sum of all the probabilities in order to obtain the model posterior probabilities for the models that sum to 1.

It should be noted that not all models in the original candidate list will have significant posterior probabilities. Hence, we choose a subset of  $m$  models from the original list based on the calculated posteriors. Methods for choosing this subset are discussed in appendix A.

## 2.2 Calculating the predictive density

The predictive density for the new response  $y^*$  at a new set of regressors  $\mathbf{x}^*$  for a given model  $M_i$  is given by:

$$P(y^*|M_i, \mathbf{x}^*, \mathbf{y}) = \int_{\sigma^2} \int_{\boldsymbol{\beta}_i} P(y^*|M_i, \mathbf{x}^*, \mathbf{y}, \sigma^2, \boldsymbol{\beta}_i) P(\boldsymbol{\beta}_i, \sigma^2|\mathbf{y}, M_i) d\boldsymbol{\beta}_i d\sigma^2, \quad (17)$$

where  $P(y^*|M_i, \mathbf{x}^*, \mathbf{y}, \sigma^2, \boldsymbol{\beta}_i)$  is the likelihood function, and  $P(\boldsymbol{\beta}_i, \sigma^2|\mathbf{y}, M_i)$  is the joint posterior of the model parameters [18]. Based on the observed data, the probability that the predicted response lies between the lower and upper bounds for a given model at a given set of regressors  $\mathbf{x}^*$  is obtained using the cumulative posterior predictive density, that is:

$$P(L \leq Y^* \leq U|M_i, \mathbf{x}^*, \mathbf{y}) = \int_L^U P(y^*|M_i, \mathbf{x}^*, \mathbf{y}) dy^*. \quad (18)$$

If the model parameters  $\log(\sigma^2)$  and  $\beta_i$  are assumed to have non-informative priors, then it has been shown by Press [19] that the predictive density obeys a  $t_{n-r_i}$  distribution. For the priors assumed here, the predictive density is shown in appendix B to follow a  $t_{n-1}$  distribution. The cumulative posterior predictive density can thus be obtained from the c.d.f. of a t-distribution that is very easy to compute using the incomplete beta function. This avoids the use of any numerical methods for the integration in equation (18). The cumulative predictive density is computed by:

$$P\left(\frac{y^* - \mathbf{x}^{*'}\hat{\beta}_i}{\hat{\sigma}_i\sqrt{1 + \mathbf{x}^{*'}(\Sigma_i^{-1} + \mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{x}^*}} < t | M_i, \mathbf{x}^*, \mathbf{Y}\right) = \frac{1}{2} \left[ 1 + I_{\frac{t^2}{\nu+t^2}}\left(\frac{1}{2}, \frac{\nu}{2}\right) \right], \quad (19)$$

where  $I_z(a, b)$  is the incomplete beta function,  $\nu = n - 1$ , and  $\hat{\sigma}_i^2 = S_i/(n - 1)$ . The objective function in equation (3) is thus the cumulative model averaged posterior predictive density, and can be calculated at a given observation  $\mathbf{x}^*$  using equations (13) and (19).

### 3 Examples

There are two hyper-parameters to be chosen in the priors, namely  $\pi$  and  $g$ . In both of the examples below, we choose a value of  $\pi = 0.5$ , which implies equal prior chances of a factor being active or inactive. The parameter  $g$  is chosen based on the value of  $\gamma$  that gives the lowest posterior probability for the null model (model with just the constant term). This choice is suggested by Meyer *at al.* [10], and we discuss it further in section 4 below. We also study the sensitivity of the solution to the choice of priors by solving the optimization problem for various values of these parameters.

The optimizations were carried out using MATLAB's *fmincon* routine. This function uses a sequential quadratic programming method. This is used to maximize the cumulative MAP between the upper and lower bounds, over all feasible values of  $x_1, \dots, x_k$ . As with most nonlinear programming algorithms, this method requires an initial starting point  $x_1, \dots, x_k$ . In order to avoid local optimums, we utilized different random starting values arranged in a latin hypercube (see [20]) to better cover the feasible region. In the two examples that

$x_1$	$x_2$	$x_3$	$y$
1.000	0.000	0.000	18.90
0.000	1.000	0.000	15.20
0.000	0.000	1.000	35.00
0.500	0.500	0.000	16.10
0.500	0.000	0.500	18.90
0.000	0.500	0.500	31.20
0.333	0.333	0.333	19.30
0.666	0.167	0.167	18.20
0.167	0.666	0.167	17.70
0.167	0.167	0.666	30.10
0.333	0.333	0.333	19.00

Table 1: Mixture data from [5] where the response is Glass Transition temperature

follow, convergence to the same point was always achieved, so the optimality of the solutions obtained seems to be well established.

### 3.1 Example 1: Mixture Experiment

This example, taken from Frisbee et al. [5], shows a mixture experiment where the response is glass transition temperature of films cast from poly(DL-lactide) (PLA), and the controllable variables are amounts of non-ionic surfactants, namely, Polaxamer 188 NF ( Pluronic<sup>®</sup> F68), Ployoxyethylene 40 monostearate (Myrj<sup>®</sup> 52-S) and Polyoxyethylene sorbitan fatty acid ester NF (Tween<sup>®</sup> 60). The authors are interested in finding the composition of the controllable factors that minimize the glass transition temperature. The data from [5] is given in Table 1. The authors fit a regression equation that is given by:

$$y = 18.50x_1 + 13.88x_2 + 36.06x_3 - 35.21x_1x_3 + 19.55x_2x_3. \quad (20)$$

Based on the fitted equation, Frisbee *et al.* [5] use contour plots to determine the minimal plateau region for glass transition temperature. However, as the experiment consisted of only 11 runs, the accuracy of the model used is suspect. There are a some other regression models that provide a reasonable fit to the data, and each of these would result in a different

optimal solution. Peterson [16] suggested that a different class of models, namely a Becker model [1], also represent adequately this process. The Becker model is of the form

$$y = b_1x_1 + b_2x_2 + b_3x_3 + b_4 \min(x_1, x_2) + b_5 \min(x_1, x_3) + b_6 \min(x_2, x_3). \quad (21)$$

The ordinary least square (OLS) regression statistics for the models in equation (20) and (21), as well as for all other models belonging to these two classes of models are shown in Table 2. Higher order terms in each model were considered only if the corresponding lower order terms were present. In the table, each row represents a competing model and under the columns containing the model terms (effects), a ‘1’ indicates that the term is present in the model and a ‘0’ indicates otherwise. The OLS statistics shown in the table are based on the sum of squares of the residuals ( $SSE$ ), the total sum of squares ( $SST$ ), and the standard error ( $S.E.$ ). We note that since the mixture models are fitted without the constant term, the  $SSE/SST$  ratio is greater than 1 for some models. This means that the model  $y = \bar{y}$ , where  $\bar{y}$  is the mean of the observed responses, fits the data better than the models for which the  $SSE/SST$  ratio is greater than 1. Even based on the OLS statistics, there are many possible models that can be used to represent the process.

<i>Model No.</i>	const	A	B	C	AB	AC	BC	$\min(A, B)$	$\min(A, C)$	$\min(B, C)$	$\frac{SSE}{SST}$	$\frac{SSE/n-r_i}{SST/n-1}$	<i>S.E.</i>	$P(M_i data)$
1	0	1	1	1	0	0	0	0	1	1	0.0223	0.0372	1.32	0.3557
2	0	1	1	1	0	0	0	0	1	0	0.0696	0.0994	2.1572	0.143
3	0	1	1	1	0	1	1	0	0	0	0.0478	0.0797	1.9323	0.138
4	0	1	1	1	0	0	0	1	1	1	0.0182	0.0365	1.3071	0.091
5	0	1	1	1	0	1	0	0	0	0	0.091	0.13	2.4675	0.0803
6	0	1	1	1	1	1	1	0	0	0	0.0384	0.0769	1.8972	0.0759
7	0	1	1	1	1	1	0	0	0	0	0.083	0.1383	2.5451	0.0392
8	0	1	1	1	0	0	0	1	1	0	0.0692	0.1153	2.3232	0.0279
9	0	1	1	1	0	0	0	0	0	0	0.2252	0.2815	3.6304	0.0146
10	0	1	1	1	0	0	1	0	0	0	0.1875	0.2678	3.5413	0.0119
11	0	1	1	1	1	0	0	0	0	0	0.2146	0.3066	3.7886	0.0068
12	0	1	1	1	1	0	1	0	0	0	0.1753	0.2922	3.6991	0.0058
13	0	1	1	1	0	0	0	0	0	1	0.212	0.3028	3.7654	0.0037
14	0	1	1	1	0	0	0	1	0	0	0.2119	0.3027	3.7646	0.0034
15	1	0	0	0	0	0	0	0	0	0	1	1	6.8426	0.0015
16	0	1	1	1	0	0	0	1	0	1	0.1901	0.3169	3.8519	0.0011
17	0	0	1	1	0	0	0	0	0	0	1.0888	1.2098	7.5262	0.0001
18	0	1	0	1	0	0	0	0	0	0	1.2697	1.4107	8.1273	0.0001
19	0	0	1	1	0	0	1	0	0	0	1.0671	1.3338	7.9026	0.0001
20	0	1	0	1	0	1	0	0	0	0	1.0958	1.3697	8.0083	0
21	0	0	1	1	0	0	0	0	0	1	1.044	1.305	7.8169	0
22	0	0	0	1	0	0	0	0	0	0	2.7513	2.7513	11.35	0
23	0	1	0	1	0	0	0	0	1	0	1.1881	1.4852	8.339	0
24	0	0	1	0	0	0	0	0	0	0	7.2363	7.2363	18.407	0
25	0	1	0	0	0	0	0	0	0	0	7.5341	7.5341	18.7818	0
26	0	1	1	0	0	0	0	0	0	0	4.9586	5.5096	16.0613	0
27	0	1	1	0	1	0	0	0	0	0	4.9138	6.1422	16.9584	0
28	0	1	1	0	0	0	0	1	0	0	4.9459	6.1824	17.0138	0

Table 2: Least square regression statistics and posterior probabilities for competing models for example 1

In order to perform Bayesian model-robust optimization, the first step was to define the prior parameters required for model averaging. Using the method proposed by Meyer et al. [10], a value of  $\gamma = 10$  was chosen. The parameter  $\pi$  was then set at 0.5 for all the factors. The sensitivity of the optimal solution with respect to the chosen parameters is discussed later. Model posteriors were calculated for all models discussed earlier. The resulting posterior probabilities are shown in the last column of Table 2 for each model.

Since there are no constant terms in the mixture models considered, we use Zellner's  $g$ -prior for all the  $\beta_i$ , i.e.,  $\Sigma_i^{-1} = (1/g)(\mathbf{X}_i' \mathbf{X}_i)$  in equation (8). In this case, the cumulative posterior predictive density is given by (see appendix B),

$$P\left(\frac{y^* - \mathbf{x}^{*'} \hat{\beta}_i}{\hat{\sigma}_i \sqrt{1 + \mathbf{x}^{*'} (\Sigma_i^{-1} + \mathbf{X}_i' \mathbf{X}_i)^{-1} \mathbf{x}^*}} < t | M_i, \mathbf{x}^*, \mathbf{Y}\right) = \frac{1}{2} \left[ 1 + I_{\frac{t^2}{\nu + t^2}} \left( \frac{1}{2}, \frac{\nu}{2} \right) \right], \quad (22)$$

where  $I_z(a, b)$  is the incomplete beta function,  $\nu = n$ , and  $\hat{\sigma}_i^2 = S_i/n$ .

Based on the model posteriors, only models with  $P(M_i | data) > 0.0279$  were considered for averaging as they accounted for 95% of the probability. Table 3 shows these models. The model numbers correspond to the respective models in Table 2. As the objective is to minimize the response, the lower bound  $L$  was set at  $\infty$  and the upper bound  $U$  was set at 18 for illustration purposes. The optimization of the MAP resulted in point(0.133, 0.867, 0) as the optimum levels of the controllable factors, where the probability of obtaining  $Y^* \in (-\infty, 18)$  was 0.9388. Table 3 also shows the optimal values of the controllable factors at which the posterior predictive densities of each individual model is maximized for  $Y^* \in (-\infty, 18)$ . The maximum value of the posterior predictive density is given in the column labeled ' $z^*$ '. It can be seen that the optimal solution can vary drastically based on the model chosen.

Figure 1 shows the cumulative MAP plotted on a 2-D simplex as well as a 3-D plot. The 2-D plot shows the points at which  $P(-\infty < Y^* < 18)$  was evaluated, with the squares representing points where  $P(-\infty < Y^* < 18) > 0.7$ . The 3-D plot shows the same points with cumulative MAP plotted on the vertical axis. Figure 2 shows  $P(-\infty < Y^* < 18 | M_i)$ , plotted at the same points for the eight competing models with the squares representing

<i>ModelNo.</i>	$P(M_i data)$	$x_1^*$	$x_2^*$	$x_3^*$	$z^*$
1	0.3557	0	1	0	0.9998
2	0.1430	0	1	0	0.9969
3	0.1380	0	1	0	0.8933
4	0.0910	0.2756	0.7244	0	0.9998
5	0.0803	0	1	0	0.7358
6	0.0759	0.2249	0.7751	0	0.8993
7	0.0392	0.3243	0.6757	0	0.7695
8	0.0279	0.2469	0.7531	0	0.9960

Table 3: Optimum for individual models for example 1

points where the individual predictive density is greater than 0.7. In order to better understand the importance of maximizing the MAP, Table 4 shows the probabilities of conformance,  $P(-\infty < Y^* < 18)$  for various cases of the true model and the assumed model. The table shows the value of  $P(L \leq Y^* \leq U|M_i, \mathbf{y}, x_1, \dots, x_k)$  where  $M_i$  is the true model and control factors  $x_1, \dots, x_k$  are set at their optimal values obtained from solving from maximizing this probability using the *assumed* model. Thus, for example, if the assumed model is model 1, then the probability of conformance is maximized at the point  $(0, 1, 0)$ , as shown in table 3, yielding a probability of 0.9998. However, this is actually the probability of conformance only if the true model is also model 1. If, for example, it so happens that the true model is model 7, then the probability of having  $Y^* \in (-\infty, 18)$  is actually 0.6584 when using the solution point  $(0, 1, 0)$ , obtained with the wrong model. Similarly, the last column on the table shows  $P(L \leq Y^* \leq U|M_i, \mathbf{y}, x_1, \dots, x_k)$  for the true model, evaluated at the solution  $x_1, \dots, x_k$  obtained from maximizing the MAP. Based on the column statistics, it can be seen that operating at the point which maximizes the MAP has highest average probability of conformance (and among lowest std. deviation of this probabilities) compared to probabilities provided by solutions obtained by assuming single one of the competing models. The MAP also has higher minimum probability of conformance, thus it improves the worst-case scenario (worst true model). Therefore, it is seen that regardless the true process model (within the assumed family of models), the solution obtained using the model-average

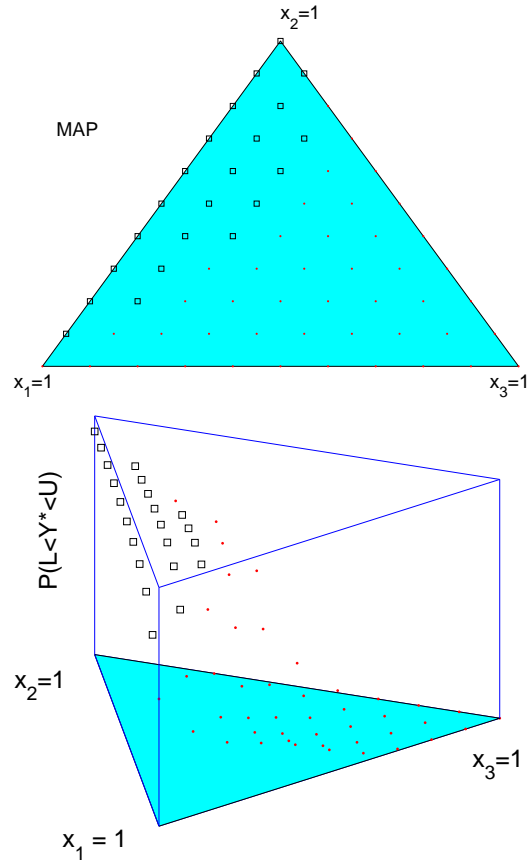


Figure 1: Simplex and 3-D plot of cumulative model-averaged posterior predictive density for example 1

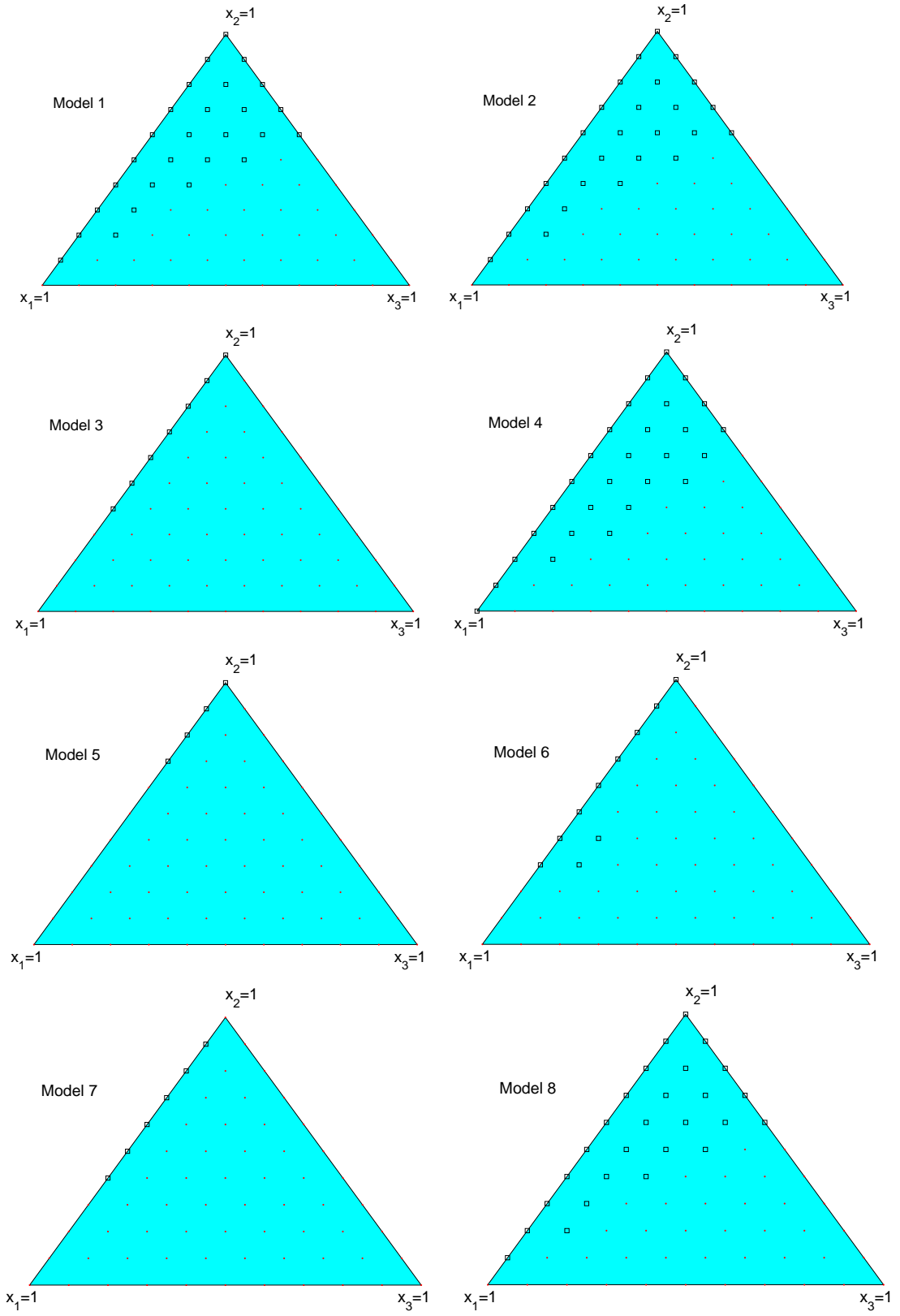


Figure 2: Simplex plot of cumulative posterior predictive densities for individual models for example 1

approach provides an operating point that gives relative high probabilities of conformance. It is in this sense that we can say the solutions obtained are *robust* to the uncertainty in the form of the true model.

<i>Assumed Model</i> $\rightarrow$	1	2	3	4	5	6	7	8	MAP
<i>True Model</i> $\downarrow$									
1	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9997	0.9998	0.9998
2	0.9969	0.9969	0.9969	0.9962	0.9969	0.9965	0.9958	0.9964	0.9967
3	0.8933	0.8933	0.8933	0.8517	0.8933	0.8623	0.8399	0.8579	0.8778
4	0.9997	0.9997	0.9997	0.9998	0.9997	0.9998	0.9998	0.9998	0.9998
5	0.7358	0.7358	0.7358	0.7057	0.7358	0.7129	0.6980	0.7099	0.7240
6	0.8434	0.8434	0.8434	0.8976	0.8434	0.8993	0.8933	0.8990	0.8920
7	0.6584	0.6584	0.6584	0.7680	0.6584	0.7623	0.7695	0.7653	0.7373
8	0.9949	0.9949	0.9949	0.9960	0.9949	0.9960	0.9958	0.9960	0.9957
Min	0.6584	0.6584	0.6584	0.7057	0.6584	0.7129	0.6980	0.7099	0.7240
Max	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
Mean	0.8903	0.8903	0.8903	0.9019	0.8903	0.9036	0.8990	0.9030	0.9029
Std. Dev.	0.1342	0.1342	0.1342	0.1170	0.1342	0.1157	0.1194	0.1161	0.1173

Table 4: Model-Robustness analysis for example 1. Table gives  $P(L \leq Y^* \leq U | M_i, \mathbf{y}, x_1, \dots, x_k)$  where  $M_i$  is the true model, evaluated at the settings  $x_1, \dots, x_k$  obtained from maximizing the probability of conformance using the assumed model.

Table 5 shows the sensitivity of the solution with respect to the chosen parameters  $\gamma$  and  $\pi$ . It is seen that the sensitivity of the solution to  $\pi$  is dependent on  $\gamma$ . At the value of  $\gamma$  chosen, the optimal controllable variables as well as the optimal predictive density are insensitive to the choice of  $\pi$ .

### 3.2 Example 2: Small-composite Design

The second example uses data from Czitrom and Spagon [4] for a chemical vapor deposition (CVD) process. The goal of the experiment was to investigate the Uniformity and Stress responses. This example illustrates the model-averaging approach on the first response. The central composite inscribed (CCI) design that was used and the experimental data are shown in Table 6. There are two controllable factors: Pressure and ratio of the gaseous reactants  $H_2$  and  $WF_6$  (denoted by  $H_2/WF_6$ ). The goal was to minimize the response, as a smaller value of “Uniformity” indicates a more uniform layer being deposited on a wafer. The models considered included combinations of main effects, two-way interactions and quadratic effects. In all the models higher order effects were included only if the corresponding main effect(s) is(are) present in the model. Table 7 lists these models along with their least square regression statistics and posterior probabilities. The prior on the factors,  $\pi$ , was set at 0.5 and a value of  $\gamma = 2$  was chosen using the method described in section 4.

Models with  $P(M_i|data) > 0.0254$  were used for model averaging as they accounted for 95% of the probability. Based on these models and within the region  $\{-1 \leq x_1 \leq 1, -1 \leq x_2 \leq 1\}$ , the MAP was maximized for  $Y^* \in (-\infty, 5)$  at the point  $(1.0000, -0.9198)$  yielding a maximum probability of conformance of 0.8851. The optimum values of the controllable factors obtained by maximizing the individual predictive densities, and the maximum value of the predictive density for the individual models for  $Y \in (-\infty, 5)$  are given in table 8. It can be seen that for all the models the optimum value of  $x_1$  is 1, but the optimum setting for  $x_2$  can vary anywhere from -1 to 1. Figures ?? shows the surface plot of the cumulative posterior predictive density of the response in the region  $(-\infty, 5)$  for different possible values

$\gamma$	$\pi$	$x_1^*$	$x_2^*$	$x_3^*$	$z^*$
0.5	0.25	0.6841	0.3159	0.0000	0.7213
0.5	0.50	0.6703	0.3297	0.0000	0.6993
0.5	0.75	0.7433	0.2567	0.0000	0.6891
2	0.25	0.6167	0.3833	0.0000	0.6906
2	0.50	0.5043	0.4957	0.0000	0.6759
2	0.75	0.5043	0.4957	0.0000	0.6759
5	0.25	0.3233	0.6767	0.0000	0.8252
5	0.50	0.3232	0.6768	0.0000	0.8252
5	0.75	0.3233	0.6767	0.0000	0.8252
10	0.25	0.1330	0.8670	0.0000	0.9388
10	0.50	0.1330	0.8670	0.0000	0.9388
10	0.75	0.1330	0.8670	0.0000	0.9388
30	0.25	0.0213	0.9787	0.0000	0.9856
30	0.50	0.0213	0.9787	0.0000	0.9856
30	0.75	0.0213	0.9787	0.0000	0.9856
100	0.25	0.0000	1.0000	0.0000	0.9767
100	0.50	0.0000	1.0000	0.0000	0.9825
100	0.75	0.0000	1.0000	0.0000	0.9825

Table 5: Sensitivity of solution with respect to the parameters  $\gamma$  and  $\pi$  for example 1

<i>Coded Pressure</i>	<i>Coded <math>H_2/WF_6</math></i>	<i>Uniformity</i>
1	0	4.6
0	0	6.2
0.71	-0.71	3.4
-0.71	0.71	6.9
-1	0	7.3
0	0	6.4
-0.71	-0.71	8.6
0	-1	6.3
0.71	0.71	5.1
0	1	5.4
0	0	5

Table 6: Design and experimental data for CVD process [4]

<i>Model no.</i>	<i>constant</i>	<i>A</i>	<i>B</i>	<i>AB</i>	<i>A<sup>2</sup></i>	<i>B<sup>2</sup></i>	<i>R<sup>2</sup></i>	<i>R<sup>2</sup><sub>Adj</sub></i>	<i>S.E.</i>	<i>P(M<sub>i</sub> data)</i>
1	1	1	1	1	0	0	0.8703	0.8148	0.6145	0.2827
2	1	1	0	0	0	0	0.7186	0.6874	0.7982	0.2396
3	1	1	1	1	1	0	0.8715	0.7858	0.6607	0.108
4	1	1	1	1	0	1	0.8703	0.7839	0.6637	0.1053
5	1	1	0	0	1	0	0.7198	0.6498	0.8449	0.0907
6	1	1	1	0	0	0	0.7285	0.6607	0.8316	0.0671
7	1	1	1	1	1	1	0.8716	0.7431	0.7235	0.0416
8	1	1	1	0	1	0	0.7297	0.6139	0.8871	0.0254
9	1	1	1	0	0	1	0.7285	0.6122	0.8891	0.025
10	1	1	1	0	1	1	0.7298	0.5496	0.9581	0.0098
11	1	0	0	0	0	0	0	0	1.4276	0.0035
12	1	0	1	0	0	0	0.0099	-0.1001	1.4974	0.0009
13	1	0	1	0	0	1	0.0099	-0.2376	1.5882	0.0003

Table 7: Least square regression statistics and posterior probabilities for competing models for example 2

of the control factors.

Model-robustness analysis for the competing models is given in Table 9. (Note that models 2 and 5 are independent of the second factor,  $x_2$  ( $H_2/WF_6$ ). In the table, for the columns associated with these two models, the probabilities of conformance were evaluated at the point (1,0)). Similarly as in the previous example, it can be seen that the solution obtained by maximizing the MAP is robust to the uncertainty in the true model of the

<i>Model no.</i>	<i>P(M<sub>i</sub> data)</i>	<i>x<sub>1</sub><sup>*</sup></i>	<i>x<sub>2</sub><sup>*</sup></i>	<i>z<sup>*</sup></i>
1	0.2827	1	-1	0.9665
2	0.2396	1	N/A	0.8132
3	0.1080	1	-1	0.9569
4	0.1053	1	-0.9017	0.9618
5	0.0907	1	N/A	0.7776
6	0.0671	1	1	0.8477
7	0.0416	1	-0.9018	0.9464
8	0.0254	1	1	0.8178

Table 8: Optimum for individual models for example 2

process.

<i>Assumed Model</i> $\rightarrow$	1	2	3	4	5	6	7	8	MAP
<i>True Model</i> $\downarrow$									
1.00	0.9665	0.8680	0.9665	0.9648	0.8680	0.3935	0.9648	0.3935	0.9651
2.00	0.8132	0.8132	0.8132	0.8132	0.8132	0.8132	0.8132	0.8132	0.8132
3.00	0.9569	0.8323	0.9569	0.9540	0.8323	0.3733	0.9540	0.3733	0.9546
4.00	0.9612	0.8614	0.9612	0.9618	0.8614	0.3991	0.9618	0.3991	0.9618
5.00	0.7776	0.7776	0.7776	0.7776	0.7776	0.7776	0.7776	0.7776	0.7776
6.00	0.7379	0.8165	0.7379	0.7472	0.8165	0.8477	0.7472	0.8477	0.7455
7.00	0.9456	0.8318	0.9456	0.9464	0.8318	0.3749	0.9464	0.3749	0.9463
8.00	0.7043	0.7809	0.7043	0.7130	0.7809	0.8178	0.7130	0.8178	0.7114
Min	0.7043	0.7776	0.7043	0.7130	0.7776	0.3733	0.7130	0.3733	0.7114
Max	0.9665	0.8680	0.9665	0.9648	0.8680	0.8477	0.9648	0.8477	0.9651
Mean	0.8579	0.8227	0.8579	0.8598	0.8227	0.5996	0.8598	0.5996	0.8594
Std. Dev.	0.1111	0.0330	0.1111	0.1075	0.0330	0.2302	0.1075	0.2302	0.1082

Table 9: Model-robustness analysis for example 2. Table gives  $P(L \leq Y^* \leq U|M_i, \mathbf{y}, x_1, \dots, x_k)$  where  $M_i$  is the true model, evaluated at the settings  $x_1, \dots, x_k$  obtained from maximizing the probability of conformance using the assumed model.

Table 10 shows results of sensitivity analysis to the solution with respect to the parameters  $\pi$  and  $\gamma$ . The sensitivity of the solution to  $\pi$  is dependent on the value of the  $\gamma$  chosen. Here, also, it is seen that at a given value of  $\gamma$ , the solution is insensitive to the selection of the  $\pi$  parameter.

### 3.2.1 Pre-Posterior Analysis

In the above example the maximum model averaged posterior probability of conformance to the specifications  $(-\infty, 5)$  was 0.8851. In practice, a process engineer may feel that such probability of conformance is too small. There are two possible reasons for a relative low probability of conformance. The first reason is that the data is limited, and so given the available data, this is the highest probability of conformance that can be obtained. In this case, running more experiments and using the additional data could give a higher value of posterior probability of conformance, especially when the repeatability of the observed measures is high. The second reason is that the specification limits set by the process engineer are unrealistic. In such case there is no point in running more experiments as the additional data will not increase the probability of conformance. These two situations can be discerned by using a pre-posterior approach, as suggested by Peterson [15]. Table 11 shows the posterior probability of conformance,  $z^*$ , as well as the optimal levels of the control factors,  $(x_1^*, x_2^*)$ , and the mean and standard deviation estimates of the posterior response at  $(x_1^*, x_2^*)$  for two cases. Both cases use the same values for the hyper-parameters as before with  $\pi = 0.5$ , and  $\gamma = 2$ . The first case (labeled “data” in table 11) uses the original data that is shown in table 6, and the second case (labeled “data+replicate” in table 11) uses the original data along with a replicate of the original data appended to the data. The data used in the second case would be valid if the experimental observations are completely repeatable. The way to mimic more data is simply based on replicating the  $X_i$  matrices and changing the corresponding degrees of freedom in the MAP computations [15]. For each of these two cases, the results are shown for various values of specification limits. It can be seen from the table that for the specification limits used earlier  $(-\infty < Y^* < 5)$ , the posterior probability

$\gamma$	$\pi$	$x_1^*$	$x_2^*$	$z^*$
0.5	0.25	1	-1	0.4878
0.5	0.50	1	-1	0.5191
0.5	0.75	1	-1	0.5218
1	0.25	1	-0.9151	0.7149
1	0.50	1	-0.9154	0.7487
1	0.75	1	-0.9154	0.7621
2	0.25	1	-0.9684	0.8428
2	0.50	1	-0.9198	0.8851
2	0.75	1	-0.9198	0.9094
5	0.25	1	-0.7760	0.8438
5	0.50	1	-0.7928	0.8992
5	0.75	1	-0.7930	0.9305
10	0.25	1	-1	0.8090
10	0.50	1	-0.5696	0.8557
10	0.75	1	-0.5995	0.8967
100	0.25	1	-0.9072	0.5563
100	0.50	1	0.5146	0.7038
100	0.75	1	0.7584	0.7944

Table 10: Sensitivity of solution with respect to parameters  $\pi$  and  $\gamma$  for example 2

	data					data+replicate				
	$z^*$	$x_1^*$	$x_2^*$	mean	std. dev.	$z^*$	$x_1^*$	$x_2^*$	mean	std. dev.
$Y^* < 5$	0.8851	1	-0.9198	3.5792	0.9670	0.9955	1	-1	2.8083	0.7599
$Y^* < 4$	0.6494	1	-1	3.5296	0.9896	0.9338	1	-1	2.8083	0.7599
$Y^* < 3$	0.3329	1	-1	3.5296	0.9896	0.5986	1	-1	2.8083	0.7599
$Y^* < 2$	0.1057	1	-1	3.5296	0.9896	0.1499	1	-1	2.8083	0.7599

Table 11: Pre-posterior analysis for example 2

of conformance increases from 0.8851 to 0.9955 when one more replicate is used. Therefore, this is evidence that in this case it is worth considering running additional experiments in order to obtain a higher posterior probability of conformance given the data. However, when the specification limits are set as  $(-\infty < Y^* < 2)$ , the posterior probability of conformance increases from 0.1057 to only about 0.1499. Thus, even when the repeatability of the process is high, the highest possible posterior probability of conformance is still very low. In this case, this is evidence that there is a need for re-designing the specification limits on the response.

## 4 Choice of Priors and Hyper-parameters

The previous sections were based on the assumption of an non-informative prior for  $\log(\sigma^2)$ , a non-informative prior on the  $\beta$  for the constant term, and a normally distributed  $g$ -prior for the remaining  $\beta$ 's. Other choices of prior that are typically considered in the literature are the use of a non-informative prior for all the  $\beta$ 's, or the use of a normally distributed  $g$ -prior for all the  $\beta$ 's.

A non-informative prior for all the  $\beta$ 's is the same as the prior we use earlier when  $g \rightarrow \infty$ . The non-informative prior is convenient for the calculation of the posterior predictive density since the resulting distribution is a  $t$ -distribution [19]. However, in the calculation of the model posteriors, this prior tends to favor the null model (i.e., a model with just the constant term). This can be explained based on Bayes' factors, since the model posteriors can be used

for model selection in a Bayesian hypothesis testing for the true model [8]. Fernandez *et al.* [3] use Bayes factors to recommend using a non-informative prior just for the constant term  $\beta$  rather than using the  $g$ -prior for all the  $\beta$ 's. They make the recommendation based on the ease of choosing the hyper-parameters when computing the Bayes' factors for the model posteriors.

For the priors chosen, there are only two hyper-parameters to be chosen, namely  $\pi$  and  $\gamma$ . In all cases here, we have assumed  $\pi = 0.5$ , so that the model posteriors are proportional to the marginal likelihood of the data. To choose  $\gamma$ , Meyer *et al.* [10] recommend using the value that minimizes the posterior probability of the null model. They use an empirical Bayesian approach to show that this value of gamma also maximizes the posterior density,  $p(\gamma|\mathbf{y})$ . As was done in the examples of section 3, a sensitivity analysis of the solutions with respect to variations on these two parameters should be conducted. Further justification for (essentially equivalent) priors as used here can be found in Meyer et al. [10].

## 5 Conclusion

A Bayesian methodology for process optimization is proposed that prescribes operating points that are robust to uncertainty in the response model. Analytical results have been derived to obtain closed form expressions for the cumulative model-averaged posterior predictive density. The results have been applied to two examples that demonstrate the advantages of model-averaging using the Bayesian predictive approach. For cases where the model-averaged posterior probability of conformance to the specifications is small, a pre-posterior analysis is recommended. As shown in example 2, this analysis could be used to determine if additional experiments could result in a higher probability of conformance or if the specifications were too demanding to start with.

**Acknowledgement.-** We thank Dr. John J. Peterson (Glaxo SmithKline Beecham) for his comments on an earlier draft of this paper.

## References

- [1] Becker, N.G., “Models for the response of a mixture”, *Journal of the Royal Statistical Society, Series B (Methodological)*, 30, (1968).
- [2] Box, G.E.P., and Draper, N.R., *Empirical model-building and response surfaces*, New York: Wiley (1987).
- [3] Fernandez, C., Ley, E., and Steel, M.F.J., “Benchmark Priors for Bayesian Model Averaging”, *Journal of Econometrics*, 100(2), 381-427, (2001).
- [4] Czitrom, V., and Spagon, P.D., *Statistical Case Studies for Industrial Process Improvement*, American Statistical Association and the Society for Industrial and Applied Mathematics, (1997).
- [5] Frisbee, S. E., and McGinity, J.W., “Influence of Nonionic Surfactants on the Physical and Chemical Properties of a Biodegradable Pseudolatex,” *Eur. J. Pharm Biopharm.*, 40(6), 355-363, (1994).
- [6] Geisser, S., *Predictive inference: an introduction*. New York: Chapman & Hall, (1993).
- [7] Johnson, N.L., Kotz, S., and Balakrishnan, N., *Continuous Univariate Distributions, Volume 2*, 2nd ed., John Wiley & Sons, New York, (1995).
- [8] Kass, R.E., and Raftery, A.E., “Bayes Factors”, *Journal of the American Statistical Association*, 90(430), 773-795, (1995).
- [9] Madigan, D., and Raftery, A.E., “Model Selection and Accounting for Model Uncertainty in Graphical Models using Occam’s Window”, *Journal of the American Statistical Association*, 89 (428), (1994).
- [10] Meyer, R.D., and Box, G., “Finding the Active Factors in Fractionated Screening Experiments”, *CQPI Report No. 80*, (1992).

- [11] Meyer, R.D., and Box, G., “Finding the Active Factors in Fractionated Screening Experiments”, *Journal of Quality Technology*, 25(2), 94-105, (1993).
- [12] Meyer, R.D., Steinberg, D.M., and Box, G., “Follow-up Designs to Resolve Confounding in Multifactor Experiments”, *Technometrics*, 38(4), 303-313, (1996).
- [13] Miro-Quesada, G., Del Castillo, E., and Peterson, J. P., “A Bayesian Approach for Multiple Response Surface Optimization in the Presence of Noise Variable”, to appear in *Journal of Applied Statistics*, (2004).
- [14] Myers, R.H., and Montgomery, D.C., *Response Surface Methodology: process and product optimization using designed experiments*, 2nd ed., New York: Wiley (2002).
- [15] Peterson, J.J., “A Bayesian Reliability Approach to Multiple Response Surface Optimization”, to appear in *Journal of Quality Technology*, (2004).
- [16] Peterson, J.J., personal communication, (2003).
- [17] Peterson, J.J., Cahya, S., and Del Castillo, E., “A General Approach to Confidence Regions for Optimal Factor Levels of Response Surfaces”, *Biometrics*, 58, pp. 422-431, (2002).
- [18] Press, S.J., *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference*, 2nd ed., Robert E. Krieger Publishing, Malabar, FL, (1982)
- [19] Press, S.J., *Subjective and Objective Bayesian Statistics*, 2nd ed., John Wiley & Sons, New York, NY, (2003).
- [20] Press, W.H., Teukolsky, S.A., Vetterling, W.T., and Flannery, B.P., *Numerical Recipes in C*, New York: Cambridge University Press, (1992).
- [21] Zellner, A., “On Assessing Prior Distributions and Bayesian Regression Analysis with g-prior Distributions”, *Bayesian Inference and Decision Techniques: Essays in Honour of Bruno de Finetti*, eds. Goel, P.K., and Zellner, A., Amsterdam, North-Holland, 1986.

## A Choosing Model Subset for Averaging

It is usually not necessary to average over all  $m$  models considered originally. There are different methods proposed in literature to choose a subset of these, either based on the marginal likelihood  $P(\mathbf{y}|M_i)$ , or on the model posterior probabilities  $P(M_i|\mathbf{y})$ . Madigan *et al.* [9] propose an algorithm based on Occam's window and Occam's razor to choose the subset of models. A simple criteria based on Occam's razor is to choose the subset of models,  $M_j$ , such that

$$\frac{\max_i P(M_i|\mathbf{y})}{P(M_j|\mathbf{y})} \leq c' \quad (\text{A.1})$$

Here,  $c'$  is a constant whose value remains to be chosen. The above criteria are useful when the number of candidate models is very large. In the examples we discuss, there are fewer candidate models. Also, as the focus of this paper is on process optimization, we use a simpler criteria to choose the subset of models from the original candidate list. Here, we order the model posteriors in descending order and include only the top  $m$  models, the sum of whose posteriors account for at least 95% of the total probability.

## B Calculation of Posterior Predictive Density

**Theorem:** For a single response process with  $k$  controllable factors, under the process model of the form given in equation (1) with normally distributed error terms, and under the priors on the factors and the parameters given by equations (4), (8), (9) and (10) with  $\Sigma^{-1} = (\mathbf{X}'_i \mathbf{X}_i) \mathbf{V}_i$ , where  $\mathbf{V}_i = \frac{1}{g} \begin{pmatrix} 0 & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{t_i} \end{pmatrix}$ , the cumulative posterior predictive density for the new response  $y^*$  at a new set of regressors  $\mathbf{x}^*$  for a given model  $M_i$ , is given by:

$$P \left( \frac{y^* - \mathbf{x}^{*'} \hat{\boldsymbol{\beta}}_i}{\hat{\sigma}_i \sqrt{1 + \mathbf{x}^{*'} (\Sigma_i^{-1} + \mathbf{X}'_i \mathbf{X}_i)^{-1} \mathbf{x}^*}} < t | M_i, \mathbf{x}^*, \mathbf{Y} \right) = \frac{1}{2} \left[ 1 + I_{\frac{t^2}{\nu + t^2}} \left( \frac{1}{2}, \frac{\nu}{2} \right) \right], \quad (\text{A.2})$$

where  $I_z(a, b)$  is the incomplete beta function,  $\nu = n - 1$ , and  $\hat{\sigma}_i^2 = S_i / (n - 1)$ .

**Proof:** The posterior predictive density is given by,

$$P(y^* | M_i, \mathbf{x}^*, \mathbf{y}) = \int_{\sigma^2} \int_{\boldsymbol{\beta}_i} P(y^* | M_i, \mathbf{x}^*, \mathbf{y}, \sigma^2, \boldsymbol{\beta}_i) P(\boldsymbol{\beta}_i, \sigma^2 | \mathbf{y}, M_i) d\boldsymbol{\beta}_i d\sigma^2, \quad (\text{A.3})$$

where  $P(y^*|M_i, \mathbf{x}^*, \mathbf{y}, \sigma^2, \boldsymbol{\beta}_i)$  is the likelihood function, and  $P(\boldsymbol{\beta}_i, \sigma^2|\mathbf{y}, M_i)$  is the joint posterior of the model parameters. Assuming normally distributed errors, the likelihood is:

$$P(y^*|M_i, \mathbf{x}^*, \mathbf{y}, \sigma^2, \boldsymbol{\beta}_i) \propto \sigma^{-1} \exp \left[ \frac{-1}{2\sigma^2} (y^* - \mathbf{x}^{*'} \boldsymbol{\beta}_i)' (y^* - \mathbf{x}^{*'} \boldsymbol{\beta}_i) \right], \quad (\text{A.4})$$

and

$$P(\boldsymbol{\beta}_i, \sigma^2|\mathbf{y}, M_i) \propto P(\mathbf{y}|M_i, \boldsymbol{\beta}_i, \sigma^2) P(\boldsymbol{\beta}_i|M_i, \sigma^2) P(\sigma^2|M_i), \quad (\text{A.5})$$

where  $P(\mathbf{y}|M_i, \boldsymbol{\beta}_i, \sigma^2)$  is the likelihood function given by equation (7).  $P(\boldsymbol{\beta}_i|M_i, \sigma^2)$  and  $P(\sigma^2|M_i)$  are the priors on the model parameters assumed to be of the form:

$$p(\sigma^2|M_i) \propto \frac{1}{\sigma^2}, \quad (\text{A.6})$$

and,

$$P(\boldsymbol{\beta}_i|M_i, \sigma^2) \propto \gamma^{-t_i} \sigma^{-t_i} \exp \left[ \frac{-1}{2\sigma^2} \boldsymbol{\beta}_i' \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}_i \right], \quad (\text{A.7})$$

where,

$$\frac{g}{\gamma^2} \mathbf{V}_i = \boldsymbol{\Sigma}^{-1}. \quad (\text{A.8})$$

Let  $k_{1,i} = \gamma^{-t_i}$ . Then,

$$P(\boldsymbol{\beta}_i, \sigma^2|\mathbf{y}, M_i) \propto k_{1,i} (\sigma^2)^{-\frac{n+t_i+2}{2}} \exp \left[ \frac{-1}{2\sigma^2} \{ (\mathbf{y} - \mathbf{X}_i \boldsymbol{\beta}_i)' (\mathbf{y} - \mathbf{X}_i \boldsymbol{\beta}_i) + \boldsymbol{\beta}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\beta}_i \} \right]. \quad (\text{A.9})$$

Let  $\Omega_i = (\mathbf{y}^* - \mathbf{x}^{*'} \boldsymbol{\beta}_i)' (\mathbf{y}^* - \mathbf{x}^{*'} \boldsymbol{\beta}_i) + (\mathbf{y} - \mathbf{X}_i \boldsymbol{\beta}_i)' (\mathbf{y} - \mathbf{X}_i \boldsymbol{\beta}_i) + \boldsymbol{\beta}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\beta}_i$ . This gives,

$$P(y^*|M_i, \mathbf{x}^*, \mathbf{y}) \propto k_{1,i} \int_{\sigma^2} \int_{\boldsymbol{\beta}_i} (\sigma^2)^{-\frac{n+t_i+3}{2}} \exp \left[ -\frac{\Omega_i}{2\sigma^2} \right] d\boldsymbol{\beta}_i d\sigma^2. \quad (\text{A.10})$$

By making a substitution  $u = \frac{\Omega_i}{2\sigma^2}$ , the above equation can be rewritten as,

$$P(y^*|M_i, \mathbf{x}^*, \mathbf{y}) \propto k_{1,i} \int_{\boldsymbol{\beta}_i} \left( \frac{\Omega_i}{2} \right)^{\frac{n+t_i+1}{2}} \left[ \int_0^\infty u^{\frac{n+t_i-1}{2}} \exp(-u) du \right] d\boldsymbol{\beta}_i. \quad (\text{A.11})$$

The inner integral inside the square brackets is a constant given by the gamma function,  $\Gamma \left( \frac{n+t_i+1}{2} \right)$ . Let  $k_{2,i} = k_{1,i} \Gamma \left( \frac{n+t_i+1}{2} \right)$ . Then,

$$P(y^*|M_i, \mathbf{x}^*, \mathbf{y}) \propto k_{2,i} \int_{\boldsymbol{\beta}_i} \left( \frac{\Omega_i}{2} \right)^{\frac{n+t_i+1}{2}} d\boldsymbol{\beta}_i. \quad (\text{A.12})$$

It can be shown that,

$$(\mathbf{y} - \mathbf{X}_i \boldsymbol{\beta}_i)'(\mathbf{y} - \mathbf{X}_i \boldsymbol{\beta}_i) + \boldsymbol{\beta}_i' \boldsymbol{\Sigma}_i^{-1} \boldsymbol{\beta}_i = S_i + (\boldsymbol{\beta}_i - \hat{\boldsymbol{\beta}}_i)'(\boldsymbol{\Sigma}_i^{-1} \mathbf{X}_i' \mathbf{X}_i)(\boldsymbol{\beta}_i - \hat{\boldsymbol{\beta}}_i), \quad (\text{A.13})$$

where  $S_i$  is defined in equation (15), and  $\hat{\boldsymbol{\beta}}_i$  is defined in equation (16). From the above equation, we can rewrite  $\Omega_i$  as,

$$\Omega_i = (y^* - \mathbf{x}^{*'} \boldsymbol{\beta}_i)'(y^* - \mathbf{x}^{*'} \boldsymbol{\beta}_i) + S_i + (\boldsymbol{\beta}_i - \hat{\boldsymbol{\beta}}_i)'(\boldsymbol{\Sigma}_i^{-1} + \mathbf{X}_i' \mathbf{X}_i)(\boldsymbol{\beta}_i - \hat{\boldsymbol{\beta}}_i). \quad (\text{A.14})$$

Define the  $(r_i \times 1)$  vector  $\mathbf{Q}_i$  and the scalar  $w_i$  as,

$$\mathbf{Q}_i = (\boldsymbol{\Sigma}_i^{-1} + \mathbf{X}_i' \mathbf{X}_i + \mathbf{x}^* \mathbf{x}^{*'})^{-1} (\hat{\boldsymbol{\beta}}_i' (\boldsymbol{\Sigma}_i^{-1} + \mathbf{X}_i' \mathbf{X}_i) + y^* \mathbf{x}^{*'}), \quad (\text{A.15})$$

and

$$\begin{aligned} w_i &= y^{*'} y^* + \hat{\boldsymbol{\beta}}_i' (\boldsymbol{\Sigma}_i^{-1} + \mathbf{X}_i' \mathbf{X}_i + \mathbf{x}^* \mathbf{x}^{*'}) \hat{\boldsymbol{\beta}}_i + S_i \\ &\quad - (\hat{\boldsymbol{\beta}}_i' (\boldsymbol{\Sigma}_i^{-1} + \mathbf{X}_i' \mathbf{X}_i) + y^* \mathbf{x}^{*'}) (\boldsymbol{\Sigma}_i^{-1} + \mathbf{X}_i' \mathbf{X}_i + \mathbf{x}^* \mathbf{x}^{*'})^{-1} (\hat{\boldsymbol{\beta}}_i' (\boldsymbol{\Sigma}_i^{-1} + \mathbf{X}_i' \mathbf{X}_i) + y^* \mathbf{x}^{*'}). \end{aligned} \quad (\text{A.16})$$

Then by completing the squares, we get

$$\Omega_i = w_i + (\boldsymbol{\beta}_i - \mathbf{Q}_i)'(\boldsymbol{\Sigma}_i^{-1} + \mathbf{X}_i' \mathbf{X}_i + \mathbf{x}^* \mathbf{x}^{*'}) (\boldsymbol{\beta}_i - \mathbf{Q}_i). \quad (\text{A.17})$$

Thus, we have that the posterior predictive density is of the form:

$$P(y^* | M_i, \mathbf{x}^*, \mathbf{y}) \propto k_{2,i} \int_{\boldsymbol{\beta}_i} \frac{d\boldsymbol{\beta}_i}{[w_i + (\boldsymbol{\beta}_i - \mathbf{Q}_i)'(\boldsymbol{\Sigma}_i^{-1} + \mathbf{X}_i' \mathbf{X}_i + \mathbf{x}^* \mathbf{x}^{*'}) (\boldsymbol{\beta}_i - \mathbf{Q}_i)]^{\frac{n+t_i+1}{2}}}. \quad (\text{A.18})$$

The integral in the above equation is a matrix  $T$ -distribution (see, e.g., [18]), and thus integrates to a constant which is a function of  $y^*$ ,  $\mathbf{x}^*$  and  $M_i$ . We include all the constant terms that are independent of  $y^*$  in constant  $k_{3,i}$ , and rewrite the above equation by including only the constant term that includes  $y^*$  as,

$$P(y^* | M_i, \mathbf{x}^*, \mathbf{y}) \propto \frac{k_{3,i}}{w_i^{n/2}}. \quad (\text{A.19})$$

Using a well-known matrix identity (see [18]), we rewrite  $w_i$  as

$$w_i = S_i + \frac{(y^* - \mathbf{x}^{*'} \hat{\boldsymbol{\beta}}_i)^2}{1 + \mathbf{x}^{*'} (\boldsymbol{\Sigma}_i^{-1} + \mathbf{X}_i' \mathbf{X}_i)^{-1} \mathbf{x}^*}. \quad (\text{A.20})$$

We can integrate the joint posterior,  $P(\boldsymbol{\beta}_i, \sigma^2 | \mathbf{y}, M_i)$ , over  $\sigma^2$ , to obtain the marginal posterior distribution of  $\boldsymbol{\beta}_i$ . This gives

$$\boldsymbol{\beta}_i | \sigma^2, \mathbf{y}, M_i \sim N \left( \hat{\boldsymbol{\beta}}_i, \sigma^2 (\boldsymbol{\Sigma}_i^{-1} + \mathbf{X}_i' \mathbf{X}_i)^{-1} \right). \quad (\text{A.21})$$

Thus, we have that

$$Z_i = \frac{y^* - \mathbf{x}^* \hat{\boldsymbol{\beta}}_i}{\sigma \sqrt{1 + \mathbf{x}^{*'} (\boldsymbol{\Sigma}_i^{-1} + \mathbf{X}_i' \mathbf{X}_i)^{-1} \mathbf{x}^*}} \sim N(0, 1). \quad (\text{A.22})$$

Similarly, by integrating the joint posterior  $P(\boldsymbol{\beta}_i, \sigma^2 | \mathbf{y}, M_i)$  over  $\boldsymbol{\beta}_i$ , it can be shown that the marginal posterior distribution of  $\sigma^2$  is given by

$$\frac{\sigma^2}{S_i} \sim \text{inv-}\chi_{n-1}^2, \quad (\text{A.23})$$

(an inverse chi-square distribution) or in other words,

$$\frac{S_i}{\sigma^2} \sim \chi_{n-1}^2. \quad (\text{A.24})$$

If  $\hat{\sigma}_i^2 = S_i / (n - 1)$ ,

$$W_i = \frac{(n - 1) \hat{\sigma}_i^2}{\sigma^2} \sim \chi_{n-1}^2. \quad (\text{A.25})$$

Thus, if  $\mu_y^* = \mathbf{x}^* \hat{\boldsymbol{\beta}}_i$ , and  $\sigma_y^{*2} = \hat{\sigma}_i^2 [1 + \mathbf{x}^{*'} (\boldsymbol{\Sigma}_i^{-1} + \mathbf{X}_i' \mathbf{X}_i)^{-1} \mathbf{x}^*]$ , then equation (A.20) can be written as,

$$w_i = S_i \left[ 1 + \frac{1}{n - 1} \frac{(y^* - \mu_y^*)^2}{\sigma_y^{*2}} \right]. \quad (\text{A.26})$$

Thus, from equation (A.19),

$$P(y^* | M_i, \mathbf{x}^*, \mathbf{y}) \propto \left[ 1 + \frac{1}{n - 1} \frac{(y^* - \mu_y^*)^2}{\sigma_y^{*2}} \right]^{-n/2} \quad (\text{A.27})$$

The density above is a Student t with mean  $\mu_y^*$ , and variance  $\sigma_y^{*2}$ , with  $(n - 1)$  degrees of freedom. That is, the posterior predictive density is

$$y^* | M_i, \mathbf{x}^*, \mathbf{y} \propto t_{n-1} \left( \mathbf{x}^* \hat{\boldsymbol{\beta}}_i, \hat{\sigma}_i^2 [1 + \mathbf{x}^{*'} (\boldsymbol{\Sigma}_i^{-1} + \mathbf{X}_i' \mathbf{X}_i)^{-1} \mathbf{x}^*] \right). \quad (\text{A.28})$$

The cumulative posterior predictive density of the response, given the model and the data, at a given level of control factors can be computed using the c.d.f. of a  $t_\nu$  distribution with  $\nu = n - 1$ . This can easily be computed using the following identity (see [7]):

$$P\left(\frac{y^* - \mu_y^*}{\sigma_y^*} < t | M_i, \mathbf{x}^*, \mathbf{y}\right) = \frac{1}{2} \left[ 1 + I_{\frac{t^2}{\nu+t^2}} \left( \frac{1}{2}, \frac{\nu}{2} \right) \right]. \quad (\text{A.29})$$

**Corollary:** For mixture models (as in example 1), or models without intercept, the models are fitted without a constant term. In this case, we assume a Zellner's  $g$ -prior [21] on all the  $\beta_i$  in the models. Thus, for a single response process with  $k$  controllable factors, under the process model of the form given in equation (1) with normally distributed error terms, and under the priors on the factors and the parameters given by equations (4), (8), (9) and (10) with  $\Sigma^{-1} = (\mathbf{X}_i' \mathbf{X}_i) \mathbf{V}_i$ , where  $\mathbf{V}_i = \frac{1}{g} \mathbf{I}_{t_i}$ :

1. The posterior predictive density is

$$y^* | M_i, \mathbf{x}^*, \mathbf{y} \propto t_n \left( \mathbf{x}^* \hat{\beta}_i, \hat{\sigma}_i^2 \left[ 1 + \mathbf{x}^{*'} (\Sigma_i^{-1} + \mathbf{X}_i' \mathbf{X}_i)^{-1} \mathbf{x}^* \right] \right), \quad (\text{A.30})$$

where

$$\hat{\sigma}_i^2 = \frac{S_i}{n}, \quad (\text{A.31})$$

and

$$\frac{n \hat{\sigma}_i^2}{\sigma^2} = \frac{S_i}{\sigma^2} \sim \chi_n^2. \quad (\text{A.32})$$

2. The cumulative posterior predictive density for the new response  $y^*$  at a new observation  $\mathbf{x}^*$  for a given model,  $M_i$ , is

$$P\left(\frac{y^* - \mathbf{x}^{*'} \hat{\beta}_i}{\hat{\sigma}_i \sqrt{1 + \mathbf{x}^{*'} (\Sigma_i^{-1} + \mathbf{X}_i' \mathbf{X}_i)^{-1} \mathbf{x}^*}} < t | M_i, \mathbf{x}^*, \mathbf{y}\right) = \frac{1}{2} \left[ 1 + I_{\frac{t^2}{\nu+t^2}} \left( \frac{1}{2}, \frac{\nu}{2} \right) \right], \quad (\text{A.33})$$

where  $I_z(a, b)$  is the incomplete beta function,  $\nu = n$ , and  $\hat{\sigma}_i^2 = S_i/n$ .

Here the only difference in the computation of the cumulative posterior predictive density is in the degrees of freedom of the  $t$  distribution.