A Bayesian Approach to Model Selection and Response Optimization on Data with a Multistratum Structure

by

Xiaohua Liu

A Thesis submitted to the Faculty of Graduate and Postdoctoral Studies of

The University of Manitoba

in partial fulfilment of the requirements of the degree of

Doctor of Philosophy

Department of Statistics

Faculty of Science

University of Manitoba

Winnipeg, Manitoba

Canada

© Xiaohua Liu, 2026

Abstract

Response optimization is a process of identifying the input variable settings that optimize the response. Since the true model is usually unknown, Bayesian model averaging is a popular approach to account for model uncertainty. Model selection is a process of choosing a best model from a set of candidates. Zellner's g-prior is a popular prior for use in Bayesian model selection. However, the methods cannot be applied to data obtained from multistratum designs, which arises naturally in industrial experiments due to the inconvenient and impractical complete randomization. Since the designs have complicated model structures, the marginal likelihood does not have a conjugate form, which poses a challenge for the research. This thesis investigates these issues and fills the gap in the existing literature on multistratum designs.

The first objective is to apply Bayesian model averaging and a predictive approach to investigate the optimization problem for data obtained from multistratum designs. With the posterior probabilities of models as weights, we consider the weighted average of the predictive densities of the response over all potential models. The goal of the optimization is to identify the values of the factors that result in a maximum probability of a response within a given range.

Two situations are considered.

The first situation is when the error terms follow multivariate normal distributions. Given the error variances, the marginal likelihood is derived, and the posterior predictive distribution of a future response is shown to follow a Student's t-distribution. When the distributions of the error variances are specified, the marginal likelihood is evaluated using bridge sampling. The cumulative posterior predictive probability is then computed using the multivariate Student's t-distribution and the Markov chain Monte Carlo (MCMC) samples obtained from bridge sampling.

The second situation considered is when the error terms follow Student's t-distributions. Stan is used to fit the Bayesian model, and the resulting posterior samples are used to compute the marginal likelihood and the cumulative posterior predictive probability. In particular, the marginal likelihood is estimated using bridge sampling. The model-robust optimization approaches under these two situations are illustrated through examples.

The second objective is to incorporate stratum-specific information into the prior formulation and propose a generalized g-prior for model selection within a Bayesian framework for data with a multistratum structure. Posterior distributions of parameters are derived and various choices of g-priors are considered. The proposed method is illustrated through examples using different types of multistratum designs. The comparison is made with existing model selection methods.

Acknowledgements

I wish to express my deepest gratitude to my thesis advisor, Dr. Po Yang, for her patient guidance, enthusiastic encouragement, and insightful critiques. She consistently provided invaluable direction whenever I faced a challenging crossroads in my research. This thesis would not have been possible without her unwavering support, her belief in my potential, and the intellectual freedom she granted me to explore my own ideas.

I also extend my sincere thanks to the other members of my committee, Dr. Saman Muthukumarana, Dr. Saumen Mandal, and Dr. Steven Zheng, for their valuable input. Their thoughtful feedback and challenging questions during my defense and throughout the writing process have undoubtedly strengthened the final quality of this work. I am particularly grateful for the time they dedicated to reviewing my manuscript with such careful attention to detail.

My sincere thanks go to the staff who provide the backbone for our department. I am deeply appreciative of Ms. Margaret Smith, administrative assistant. I am also grateful to the graduate assistants, Ms. Rosa Di Noto and Ms. Sunny Joo, for their constant availability and patience.

Finally, I am profoundly grateful to my parents. Their unwavering support and encouragement have been the bedrock of my entire academic journey, and especially during the writing of this thesis. They have been my pillars of strength, offering not only moral support but also the quiet sacrifices that made this pursuit possible. I dedicate this accomplishment to them, with all my love and endless gratitude for believing in me long before I ever believed in myself.

Contents

Abstract i

Acknowledgements iii

Contents v

List of Figures ix

1 Introduction 1

1.1 Design of experiments and multistratum designs 1

1.2 Model selection 3

1.3 Response optimization 5

1.4 Research objective and thesis organization 7

2 Preliminaries 10

2.1 Model selection and Bayesian approach 11

2.1.1 AIC and BIC 11

2.1.2 Posterior model probability 12

2.1.3 g-prior 14

2.2 Bayesian model averaging and response optimization 17

2.3 Bridge sampling 20

2.4 Multistratum designs 23

2.4.1 Split-plot design 24

2.4.2 Three-stratum design 25

3 Response Optimization on Data with a Multistratum Structure Under Normal Assumptions of Random Effects and Errors 30

3.1 Introduction 30

3.2 The model, Bayesian model averaging and optimization problem 31

3.3 Posterior predictive distribution 33

3.4 Posterior model probability 41

3.5 Optimization procedure 42

3.6 Example: Staggered-level design with simulated data 45

3.6.1 Selection of optimal factor levels by individual models . 49

3.6.2 Selection of optimal factor levels by Bayesian model averaging 49

3.6.3 Sensitivity analysis of parameters 57

4 Response Optimization on Data with a Multistratum Structure Under Non-normal Assumptions of Random Effects and Errors 60

4.1 Introduction 60

4.2 A motivating example and prior assumptions 62

4.3 Optimization procedure 67

4.4 Model selection 68

4.5 Optimal factor levels 69

4.6 Model-robustness analysis 73

4.7 Simulation study 74

4.7.1 Sensitivity analysis of π 78

5 Model selection using a generalized g-prior for data with a multistratum structure 79

5.1 Introduction 79

5.2 Generalized g-prior and posterior parameter estimates 80

5.3 Posterior model probabilities 85

5.4 Examples 88

5.4.1 Example 1 90

5.4.2 Example 2 97

5.4.3 Example 3 101

6 Conclusion and future work 109

Bibliography **116**

List of Figures

3.1 Contour plots of the model-averaged posterior probability z for each level combination (w,s). 53

3.2 Comparison of Bayesian model averaging, AIC, and BIC over 150 simulation replicates. 55

3.3 Boxplot comparison of Bayesian model averaging, AIC, and BIC over 150 simulation replicates. 56

4.1 QQ plots of whole plot and subplot errors. 65

4.2 Contour plots of the model-averaged posterior probability z. . . 71

4.3 Comparison of Bayesian model averaging, AIC, and BIC over 20 simulation replicates. 76

4.4 Boxplot comparison of Bayesian model averaging, AIC, and BIC over 20 simulation replicates. 77

5.1 Marginal posterior probabilities of the effects for the design in Table 5.1 96

5.2 Marginal posterior probabilities of the effects for the design in Table 5.4 101

5.3 (a) Marginal posterior probabilities of the effects for the design in Table 5.6 106

5.3 (b) Marginal posterior probabilities of the effects for the design in Table 5.6 107

Chapter 1

Introduction

1.1 Design of experiments and multistratum designs

Design of experiments are widely applied to social sciences, medicine, physics, engineering, as well as industrial settings. Experimental design enables experiments to optimize the data collection process, thereby increasing the likelihood of detecting effects and effectively testing hypotheses that address their research questions (Box et al., 2005). The most fundamental experimental design is the completely randomized design, where subjects or units are assigned at random to treatment groups. This means that every experimental unit has an equal probability of receiving any treatment, and the assignment of one unit does not affect the assignment of another. This process ensures fair and unbiased allocation, reducing the impact of confounding variables and allowing for precise estimation

of treatment effects (Montgomery, 2017).

However, in industrial experiments, incomplete randomization is common due to technical constraints, costs, or time limitations. Such experiments typically involve system-level factors or processing factors that are difficult or expensive to change, and therefore not reset independently for each run.

The split-plot design, originally developed by Fisher (1925) for use in agricultural experiments, is a popular method for the incomplete randomization experiments. It is commonly used in industry and has been extensively studied in the last two decades (Parker et al., 2007; Jones and Goos, 2012; Mylona et al., 2014; Xiao et al., 2025). A split-plot design involves two stages of randomization and is therefore also called a two-stratum design. Letsinger et al. (1996) extended response surface methodology to accommodate split-plot structures. Jones and Goos (2007) developed an algorithm for constructing split-plot designs. Literature reviews for the research on split-plot designs can be found in Jones and Nachtsheim (2009) and Cortes et al. (2018). More complicated designs include split-split-plot designs (Jones and Goos, 2009) and staggered-level designs (Arnouts and Goos, 2012) which have three stages of randomization and therefore have three-stratum structures. The resulting analysis must account for a complex error structure associated with each stratum. Arnouts and Goos (2015) considered the staggered-level designs for response surface models. Staggered-level designs with the inclusion of a block factor is considered in Goos et al. (2024).

Trinca and Gilmour (2001) developed methods for constructing general response surface multistratum designs, including split-plot and staggered-level

designs. Trinca and Gilmour (2015) introduced a stratum-by-stratum algorithm to construct multistratum designs. The construction of sequential and follow-up multistratum designs based on Bayesian approach can be found in Lin and Yang (2019, 2022). Gilmour et al. (2026) proposed a pure error estimation method for the variance components of multistratum designs.

1.2 Model selection

Model selection is the process of identifying the most suitable statistical model from a set of candidate models. The goal is to choose a model that generalizes well to new data while avoiding overfitting. Bayesian model selection has a well-established background. The direct approach is to use the indicators in the model, like Gibbs variable selection, which was suggested by Dellaportas et al. (2002). Stochastic search variable selection (George and McCulloch, 1993) is a more computationally intensive method using the indicators. It was extended to multivariate case by Brown et al. (1998). Popular methods for model selection include the stepwise selection (forward, backward, or hybrid), which is a sequential approach that adds or removes predictors based on p-values or criteria-based approaches, such as Akaike's Information Criterion (AIC) (Akaike, 1974) and the Bayesian Information Criterion (BIC) (Schwarz, 1978), which penalize model complexity, as well as Bayesian methods such as Bayes factors and posterior model probabilities. Madigan and Raftery (1994) introduced the Occam's window approach for model selection. According to the principles of Occam's window, if the posterior probability of a model is significantly lower

than that of the best model within its class, the model will be eliminated. The method provides an overall strategy for model selection and addresses model uncertainty in a highly efficient manner. It enables effective searching through the vast space of models.

The computation of marginal likelihoods and, then, posterior model probabilities is a challenge in Bayesian model comparison when certain parameters cannot be integrated out. Bridge sampling has emerged as a particularly attractive approach due to its statistical efficiency and relative ease of implementation. Originally introduced by Meng and Wong (1996), bridge sampling estimates the ratio of normalizing constants by drawing samples from two distributions and leveraging an optimal bridge function to minimize asymptotic variance. The method has since been extensively developed and applied across diverse statistical contexts. Gronau et al. (2017) provided a comprehensive tutorial on bridge sampling, demonstrating its application to a wide range of models and introducing the bridgesampling R package, which has become a standard tool for computing marginal likelihoods from posterior samples.

Zellner's g-prior (Zellner, 1986) for regression models play a critical role in model comparisons. Its key appeal lies in the fact that the posterior mean becomes a weighted average of the ordinary least squares estimate and the prior mean, and the Bayes factor for model comparison collapses to a simple function of the residual sum of squares, enabling efficient analytical expressions without MCMC. The g-prior method has been applied in generalized linear model (Bové and Held, 2011) and logistic model (Hanson et al., 2014). The mixtures of g-priors have also been extensively studied (Liang et al., 2008; Li and Clyde,

2018; Kang and Jeong, 2024).

1.3 Response optimization

Response surface methodology provides a set of techniques for designing experiments, analyzing data, and interpreting results in empirical investigations of processes to achieve an optimal response. Response surface optimization is a mathematical and statistical technique used to optimize the response of a system or process by analyzing and manipulating the input variables. It is widely used in engineering, manufacturing and process optimization (Wu and Zhang, 2007; Khuri and Mukhopadhyay, 2010; Myers et al., 2016).

Traditional response optimization approaches often rely on selecting a single model and treating the chosen model as if it were the true generator of the data. Under the independent, identically normal distributed error and some vague prior assumptions, the posterior predictive density for a given model follows a Student's t-distribution, Press (1982) described the methodology in details. However, this method disregards the inherent uncertainty in the process of model selection. Consequently, it can lead to over confidence in the resulting inferences and decisions, making them riskier than perceived. To address this issue, Bayesian model averaging offers a statistical framework that addresses model uncertainty by combining predictions from multiple models. The idea of combining models to account for this uncertainty dates back to Barnard (1963), with Roberts (1965) later proposing the simultaneous consideration of two competing models. Madigan and Raftery (1994) demonstrated that averaging

predictions across multiple models yields better predictive performance than relying on a single model. Hoeting et al. (1999) offered a tutorial introduction to Bayesian model averaging.

Rajagopal and del Castillo (2005) conducted the process optimization by incorporating the g-prior using Bayesian model averaging for a single response. Peterson (2004) introduced a posterior predictive approach which accounts for both the correlations between response variables and the model parameter uncertainty for multiple responses. Miro-Quesada et al. (2004) extended the Bayesian predictive approach for multiple response optimization to robust parameter design. Ng (2010) proposes a Bayesian model averaging framework for multiple response optimization that accounts for response model uncertainty, parameter uncertainty, and response correlations, and further addresses the design of follow-up experiments to improve quality.

For multistratum designs, since the number of levels in each stratum is small, it is difficult to assess the distribution of the random stratum effects. Moreover, in practice, the random errors may not follow a normal distribution. Rajagopal et al. (2005) studied such case with the inclusion of noise factors. They proposed a methodology that yields a solution robust to non-normal error terms. Robinson et al. (2009) demonstrated the effectiveness of the Bayesian approach in analyzing split-plot experiments, particularly its flexibility in handling various functionals of the response distribution. By employing an empirical approximation of the posterior distribution for the objective function of interest, Robinson et al. (2012) proposed a Bayesian method for process optimization in a broad class of robust parameter design models, encompassing both normal

and non-normal responses, within the split-plot design framework. Peterson (2023) studied a practical experiment and assumed that the random stratum effects follow Student's t-distributions to provide protection against possible outliers in the random effects. The posterior predictive distribution for a single model was computed using Stan (Carpenter et al., 2017; Stan Development Team, 2020). For a literature review on response optimization using Bayesian predictive approach, see del Castillo and Reis (2020).

1.4 Research objective and thesis organization

Despite the growing body of work on multistratum designs and Bayesian response optimization, some important issues remain unaddressed in the literature. First, the response optimization for data obtained from multistratum designs has not been investigated in the literature. This includes two cases: (1) random errors are normally distributed, and (2) random errors are not normally distributed, which often appears in practice. Second, while Zellner's g-prior is widely used for variable selection in linear models, its extension to multistratum settings has received little attention. To fill these gaps, this thesis develops a Bayesian model averaging approach for response optimization that accounts for model uncertainty under both normal and non-normal error distributions, using bridge sampling and Stan to handle complex hierarchical structures. Furthermore, a generalized g-prior is proposed that incorporates stratum-specific information, enabling model selection in multistratum designs with various choices of the g hyperparameter. The remainder of the thesis is organized as follows.

Chapter 2 introduces preliminary knowledge that are related to the research in this thesis, including g-prior, posterior model probability, Bayesian model averaging, bridge sampling, and multistratum designs.

In Chapter 3, we present a Bayesian approach to response optimization for data with a multistratum structure under the assumption of normally distributed error terms. First, we introduce the general model for multistratum designs and prior assumptions. Then, we derive a closed-form expression showing that the posterior predictive density follows a Student's t-distribution. Next, we derive the posterior model probabilities and outline the optimization procedure. We illustrate the methodology using a 28-run staggered-level design with simulated data. Bridge sampling is used to estimate the marginal likelihood. Finally, we provide a simulation study comparing the performance of Bayesian model averaging with AIC and BIC.

In Chapter 4, we investigate the situations where error terms do not follow normal distributions. A practical example with data from a split-plot design is utilized to explain our methodology. We assume that error terms follow Student's t-distributions to accommodate potential outliers. First, we describe the model and prior assumptions. Then, we describe the optimization procedure. We fit the model in Stan and use the bridge sampling to estimate the marginal likelihood. Next, we perform model selection followed by searching for the optimal factor levels. A model-robustness analysis evaluates the performance of Bayesian model averaging relative to individual models. A simulation study is conducted for comparing Bayesian model averaging with AIC and BIC. Finally, we carry out a sensitivity analysis for the prior model probability.

Chapter 5 focuses on model selection for multistratum data using a generalized g-prior. We first introduce the generalized g-prior and derive the posterior distributions of parameters. Then, we develop the posterior model probabilities Next, we present four examples to illustrate our method. For some examples, three effect size scenarios (large, moderate, small) are considered, and the performance of different g-prior choices for model selection is evaluated. The comparisons with AIC and BIC are also provided. Finally, we further assess the different choices of g-prior by examining the marginal posterior probabilities of individual effects.

Chapter 6 concludes the thesis by summarizing the main contributions and discussing directions for future research.

Chapter 2

Preliminaries

This chapter reviews essential concepts and tools for the thesis. It introduces classical model selection methods, Bayesian model averaging, modern computational methods, and multistratum designs. These preliminaries support the methodologies developed in later chapters. Section 2.1 introduces key concepts for selecting among competing statistical models, including the information criteria AIC and BIC, posterior model probabilities from a Bayesian perspective, and Zellner's g-prior for regression coefficients. Section 2.2 introduces Bayesian model averaging as a coherent framework for accounting for model uncertainty in response optimization, where predictions from multiple candidate models are combined using posterior model probabilities. Section 2.3 introduces bridge sampling, a Monte Carlo technique for estimating marginal likelihoods by linking the posterior distribution to a proposal distribution through a bridge function. Section 2.4 describes multistratum designs, including split-plot, split-split-plot, and staggered-level designs, which arise in industrial experiments due to practical

constraints on complete randomization.

2.1 Model selection and Bayesian approach

Model selection is the process of comparing a set of candidate statistical models and choosing the one that is most appropriate according to a specified criterion. In practice, we rarely know the true data-generating process, so we must rely on data to guide us toward a model that best captures the underlying structure while remaining simple. This process balances goodness-of-fit with model complexity to achieve optimal predictive performance and interpretability.

2.1.1 AIC and BIC

AIC and BIC are popular model selection methods. Both aim to balance goodness-of-fit with model complexity, preventing overfitting.

AIC estimates the relative amount of information lost by a given model. It is defined as:

$$\text{AIC} = -2\ln(\hat{L}) + 2k,$$

where: $\hat{L}$ is the maximized value of the likelihood function for the model and k is the number of estimated parameters in the model. Given a set of candidate models, the one with the smallest AIC is preferred. AIC rewards goodness-of-fit (via $-2\ln\hat{L}$) but penalizes the number of parameters (via $2k$). It is derived from information theory and is asymptotically efficient since as $n \to \infty$, it tends to select the model that minimizes the mean squared error of prediction.

BIC is based on Bayesian principles and approximates the marginal likelihood of the model. It is defined as:

$$\text{BIC} = -2\ln(\hat{L}) + k\ln(n).$$

As with AIC, smaller BIC values indicate a better model. The penalty term $k\ln(n)$ grows more quickly with n than AIC's $2k$ when $n > 7$. Consequently, BIC tends to select simpler models than AIC for moderate to large sample sizes. BIC is consistent, that is, given a true finite-dimensional model among the candidates, BIC selects it with probability approaching 1 as $n \to \infty$.

2.1.2 Posterior model probability

The Bayesian approach to statistics and data analysis offers a coherent framework for learning from data by combining prior information about unknown parameters with information from the observed data. For a given model M, Bayes' theorem updates the prior distribution of the parameters to obtain the posterior distribution.

Assume that $P(\theta|M)$ is the prior distribution of the parameter vector $\boldsymbol{\theta}$. It represents the uncertainty about $\boldsymbol{\theta}$ before observing the data. The posterior distribution of $\boldsymbol{\theta}$ given the data $\mathbf{Y}$ can be obtained using Bayes' theorem:

$$P(\boldsymbol{\theta} \mid M, \mathbf{Y}) = \frac{L(\mathbf{Y} \mid M, \boldsymbol{\theta})\, P(\boldsymbol{\theta} \mid M)}{L(\mathbf{Y} \mid M)}$$

where $L(\mathbf{Y} \mid M, \boldsymbol{\theta})$ is the likelihood of the data and

$$L(\mathbf{Y} \mid M) = \int L(\mathbf{Y} \mid M, \boldsymbol{\theta}) P(\boldsymbol{\theta} \mid M) d\boldsymbol{\theta}$$

is the marginal likelihood, a normalizing constant. $P(\boldsymbol{\theta} \mid M, \mathbf{Y})$ updates the belief about theta after incorporating the data.

Box and Meyer (1993) suggested to choose the model priors based on the active factors, that is, the factors that appear in each model. Let M_i be a model in the model space and π_j, $j = 1, 2, \ldots, m$, be the probability that a factor j is active. Assume that the prior probabilities of active factors are independent, the prior for the model M_i (Box and (Box and Meyer, 1993)) is defined as

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

If $\pi_j = \pi$ for $j = 1, 2, \cdots, m$, then

$$P(M_i) = \pi^{f_i} (1 - \pi)^{m - f_i}, \tag{2.1}$$

where f_i is the number of factors in the model M_i.

The posterior probability of M_i given the data can be calculated by Bayes' theorem,

$$P(M_i|\mathbf{Y}) = \frac{L(\mathbf{Y}|M_i) P(M_i)}{\sum_i L(\mathbf{Y}|M_i) P(M_i)}, \tag{2.2}$$

The posterior model probability quantifies how likely each model is, given the observed data and our prior beliefs. It can be used to select the most plausible

model or select a set of potential models that are used for Bayesian model averaging in order to account for model uncertainty.

2.1.3 g-prior

In Bayesian analysis, various types of priors can be employed, depending on the nature of the problem and the available information. Zellner's g-prior (Zellner, 1986) is a Bayesian prior distribution used in regression analysis to facilitate model selection. It was introduced to address the limitations of traditional prior distributions and is a popular conjugate prior for regression coefficients. Due to its analytical convenience and its connection to the predictor structure, Zellner's g-prior is widely used in Bayesian model comparison.

Consider the normal linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.3}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of responses, $\mathbf{X}$ is an $n \times (p+1)$ design matrix, p is the number of terms in the model excluding the constant. $\boldsymbol{\beta}$ is a $(p+1) \times 1$ vector of coefficients, $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, and σ^2 is the variance of random errors.

Conditional on σ^2, the prior distribution of $\boldsymbol{\beta}$ is assumed to follow the multivariate normal:

$$\boldsymbol{\beta} \mid \sigma^2 \sim N\big(\tilde{\boldsymbol{\beta}},\, g\,\sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\big), \tag{2.4}$$

where $\tilde{\boldsymbol{\beta}}$ is a prior mean and $g > 0$ is a scalar hyperparameter. σ^2 is usu-

ally assigned an (improper) non-informative prior $p(\sigma^2) \propto 1/\sigma^2$ or a proper inverse-gamma prior.

The name "g-prior" comes from the factor g that scales the covariance matrix. The prior covariance is proportional to $(\mathbf{X}^T\mathbf{X})^{-1}$, the same structure as the frequentist sampling variance of the ordinary least square (OLS) estimator

$$\hat{\boldsymbol{\beta}}_{\text{OLS}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}.$$

It reflects the scale and correlation structure of the predictors.

Assuming a non-informative prior for σ^2, that is, $p(\sigma^2) \sim 1/\sigma^2$, then the posterior distribution of $\boldsymbol{\beta}$ given σ^2 and $\mathbf{Y}$ is

$$\boldsymbol{\beta} \mid \sigma^2, \mathbf{Y} \sim N\Big(\frac{1}{1+g}\tilde{\boldsymbol{\beta}} + \frac{g}{1+g}\hat{\boldsymbol{\beta}}_{\text{OLS}},\ \sigma^2\frac{g}{1+g}(\mathbf{X}^T\mathbf{X})^{-1}\Big),$$

where the posterior mean is a weighted average of the prior mean and the OLS estimate.

The marginal posterior of $\boldsymbol{\beta}$ given Y is a multivariate t-distribution with n degrees of freedom:

$$\boldsymbol{\beta} \mid \mathbf{Y} \sim t_n\Big(\frac{1}{1+g}\tilde{\boldsymbol{\beta}} + \frac{g}{1+g}\hat{\boldsymbol{\beta}}_{\text{OLS}},\ s^2/n\frac{g}{1+g}(\mathbf{X}^T\mathbf{X})^{-1}\Big). \tag{2.5}$$

where $s^2 = (\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{OLS}})^\top(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}_{\text{OLS}}) + \frac{1}{g+1}(\hat{\boldsymbol{\beta}}_{\text{OLS}} - \tilde{\boldsymbol{\beta}})^\top\mathbf{X}^\top\mathbf{X}(\hat{\boldsymbol{\beta}}_{\text{OLS}} - \tilde{\boldsymbol{\beta}})$.

If a proper inverse-gamma prior is used for σ^2, the conditional posteriors remain normal-inverse-gamma, and the marginal for $\boldsymbol{\beta}$ is again a multivariate t. The prior is invariant under scale transformations of the predictors (when

an intercept is included appropriately) and under linear reparameterizations. This is important for default prior choices. The posterior mean is a shrinkage estimator that shrinks the OLS estimate toward the prior mean. The amount of shrinkage is controlled by g. For a large g, the posterior mean approaches the OLS estimate; for a small g, it is dominated by the prior. The g-prior is widely used in Bayesian model averaging and variable selection because the Bayes factor between any two nested models can be computed analytically once g is chosen. This allows efficient exploration of large model spaces. More common is to treat g as a hyperparameter to be estimated or to use a mixture of g-priors (e.g., the hyper-g prior) to achieve robustness. The g-prior is favored for its computational convenience, automatic scaling and invariance properties.

The posterior probability for a model M is

$$L(\mathbf{Y} \mid M) \propto (1+g)^{-\frac{p+1}{2}} \left(\mathbf{Y}^T\mathbf{Y} - \frac{g}{1+g} \mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \right)^{-\frac{n}{2}}$$

and by Bayes' theorem, model selection can be conducted using the posterior model probability:

$$P(M \mid \mathbf{Y}) \propto (1+g)^{-\frac{p+1}{2}} \left(\mathbf{Y}^T\mathbf{Y} - \frac{g}{1+g} \mathbf{Y}^T\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y} \right)^{-\frac{n}{2}} \times \prod_{j \in M} (\pi_j) \times \prod_{j' \notin M} (1-\pi_{j'})$$

2.2 Bayesian model averaging and response optimization

Bayesian model averaging provides a coherent framework for accounting for model uncertainty inherent in statistical inference. Traditional approaches typically select a single best model and proceed as if it had generated the data, an approach that ignores uncertainty about the model form itself and often leads to overconfident inferences and predictions. In contrast, Bayesian model averaging averages over a set of candidate models, weighting each by its posterior model probability, thereby accommodating model uncertainty. To find the posterior model probabilities using Equation (5.9), we need to specify model priors. Box and Meyer (1993) suggested to choose the model priors based on the active factors, that is, the factors that appear in each model. Let M_i be a model in the model space and π_j, $j = 1, 2, \ldots, m$, be the probability that a factor j is active, where m is the number of factors in the experiment. Assume that the prior probabilities of active factors are independent, the prior for the model M_i (Box and Meyer, 1993) is defined as

For a given model M_i and a new set of regressors $\mathbf{x}^*$, the posterior predictive density of a future response Y^* is defined as

$$P(Y^* \mid M_i, \mathbf{x}^*, \mathbf{Y}) = \int_\theta P(Y^* \mid M_i, \mathbf{x}^*, \mathbf{Y}, \theta)\, P(\theta \mid \mathbf{Y}, M_i)\, d\theta,$$

where $P(Y^* \mid M_i, \mathbf{x}^*, \mathbf{Y}, \theta)$ is the likelihood and $P(\theta \mid \mathbf{Y}, M_i)$ the joint posterior of the model parameters (Press, 1982).

Under the model (2.3) and some conjugate priors for $\boldsymbol{\beta}$ and σ^2, Y^* follows a Student's t-distribution. In particular, with the diffuse prior

$$p(\boldsymbol{\beta}, \sigma^2) \propto \frac{1}{\sigma^2},$$

it can be shown that

$$\mathbf{Y}^* \mid \mathbf{Y} \sim t_{n-(p+1)}\Big(\mathbf{x}^{*\top}\hat{\boldsymbol{\beta}},\ s^2\big(1+\mathbf{x}^{*\top}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{x}^*\big)\Big), \tag{2.6}$$

where

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top\mathbf{Y}, \qquad s^2 = \frac{(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})^\top(\mathbf{Y}-\mathbf{X}\hat{\boldsymbol{\beta}})}{n-(p+1)}.$$

The Bayesian model-averaged posterior predictive density is defined as

$$P(Y^* \mid \mathbf{x}^*, \mathbf{Y}) = \sum_i P(Y^* \mid M_i, \mathbf{x}^*, \mathbf{Y})\, P(M_i \mid \mathbf{Y}). \tag{2.7}$$

Therefore the model-averaged interval probability is a weighted sum of the interval probabilities from each model M_i, weighted by the posterior model probabilities $P(M_i \mid \mathbf{Y})$, that is,

$$P(L < Y^* < U \mid \mathbf{x}^*, \mathbf{Y}) = \sum_i P(L < Y^* < U \mid M_i, \mathbf{x}^*, \mathbf{Y})\, P(M_i \mid \mathbf{Y}).$$

For each model M_i, assume that the posterior predictive distribution of Y^* is a Student's t-distribution with $\nu_i = n - p_i - 1$ degrees of freedom, where p_i is the number of terms in M_i excluding the constant. To find the interval probability, L

and U can be standardized to t-scores:

$$t_{L,i} = \frac{L - \mathbf{x}^{*\top}\hat{\boldsymbol{\beta}}_i}{s_i\sqrt{1+\mathbf{x}^{*\top}(\mathbf{X}_i^\top\mathbf{X}_i)^{-1}\mathbf{x}^*}}, \quad t_{U,i} = \frac{U - \mathbf{x}^{*\top}\hat{\boldsymbol{\beta}}_i}{s_i\sqrt{1+\mathbf{x}^{*\top}(\mathbf{X}_i^\top\mathbf{X}_i)^{-1}\mathbf{x}^*}}.$$

Let $F_{t_{\nu_i}}(\cdot)$ denote the cumulative distribution function of the Student's t-distribution with ν_i degrees of freedom. The interval probability given model M_i is:

$$P(L < Y^* < U \mid M_i, \mathbf{x}^*, \mathbf{Y}) = F_{t_{\nu_i}}\big(t_{U,i}\big) - F_{t_{\nu_i}}\big(t_{L,i}\big), \tag{2.8}$$

where the the cumulative distribution function can be calculated using the regularized incomplete beta function. Johnson et al. (1995), that is,

$$F_{t_{\nu_i}}\big(t\big) = \frac{1}{2} + \frac{1}{2} I_{\frac{t^2}{\nu_i+t^2}}\left(\frac{1}{2}, \frac{\nu_i}{2}\right).$$

Hence, the interval probability given model M_i becomes

$$P(L < Y^* < U \mid M_i, \mathbf{x}^*, \mathbf{Y}) = \frac{1}{2}\left[I_{\frac{t_{U,i}^2}{\nu_i+t_{U,i}^2}}\left(\frac{1}{2}, \frac{\nu_i}{2}\right) - I_{\frac{t_{L,i}^2}{\nu_i+t_{L,i}^2}}\left(\frac{1}{2}, \frac{\nu_i}{2}\right)\right].$$

Then, the model-averaged interval probability is

$$\begin{aligned}
&P(L < Y^* < U \mid \mathbf{x}^*, \mathbf{Y}) \\
&= \sum_i \Big[F_{t_{\nu_i}}\big(t_{U,i}\big) - F_{t_{\nu_i}}\big(t_{L,i}\big)\Big] P(M_i \mid \mathbf{Y}), \\
&= \sum_i \frac{1}{2}\left[I_{\frac{t_{U,i}^2}{\nu_i+t_{U,i}^2}}\left(\frac{1}{2}, \frac{\nu_i}{2}\right) - I_{\frac{t_{L,i}^2}{\nu_i+t_{L,i}^2}}\left(\frac{1}{2}, \frac{\nu_i}{2}\right)\right] P(M_i \mid \mathbf{Y}).
\end{aligned}$$

2.3 Bridge sampling

Bridge sampling is a Monte Carlo technique in Bayesian statistics for estimating marginal likelihoods, or more generally, ratios of normalizing constants (Gronau et al., 2017). By linking the posterior distribution to a suitable proposal distribution through a bridge function, it enables stable estimation of the ratio between their normalizing constants. This method is commonly used to compute marginal likelihoods for model comparison.

In this section, we introduce bridge sampling for estimating the marginal likelihoods for a given model M.

Consider the identity:

$$1 = \frac{\int L(\mathbf{Y} \mid M, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid M) h(\boldsymbol{\theta} \mid M) g(\boldsymbol{\theta} \mid M) d\boldsymbol{\theta}}{\int L(\mathbf{Y} \mid M, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid M) h(\boldsymbol{\theta} \mid M) g(\boldsymbol{\theta} \mid M) d\boldsymbol{\theta}}, \tag{2.9}$$

where $g(\boldsymbol{\theta} \mid M)$ is referred to as the proposal distribution and $h(\boldsymbol{\theta} \mid M)$ the bridge function. Multiplying both sides of (2.9) by $L(\mathbf{Y} \mid M)$, we obtain

$$\begin{aligned} L(\mathbf{Y} \mid M) &= \frac{\int L(\mathbf{Y} \mid M, \boldsymbol{\theta})\, p(\theta \mid M)\, h(\theta \mid M)\, g(\theta \mid M)\, d\theta}{\int \frac{L(\mathbf{Y}|M,\boldsymbol{\theta})\, p(\theta|M)}{L(\mathbf{Y}|M)}\, h(\theta \mid M)\, g(\theta \mid M)\, d\theta} \\ &= \frac{\int L(\mathbf{Y} \mid M, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid M) h(\boldsymbol{\theta} \mid M) g(\boldsymbol{\theta} \mid M) d\boldsymbol{\theta}}{\int h(\boldsymbol{\theta} \mid M) g(\boldsymbol{\theta} \mid M) p(\boldsymbol{\theta} \mid M, \mathbf{Y}) d\boldsymbol{\theta}} \\ &= \frac{\mathbb{E}_{L(\boldsymbol{\theta})}[p(\mathbf{Y} \mid M, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid M) h(\boldsymbol{\theta} \mid M)]}{\mathbb{E}_{\text{post}}[h(\boldsymbol{\theta} \mid M) g(\boldsymbol{\theta} \mid M)]}. \end{aligned} \tag{2.10}$$

Then the marginal likelihood can be approximated by

$$\hat{L}(\mathbf{Y} \mid M) = \frac{\frac{1}{N_2} \sum_{i=1}^{N_2} L(\mathbf{Y} \mid M, \tilde{\boldsymbol{\theta}}_i) p(\tilde{\boldsymbol{\theta}}_i \mid M) h(\tilde{\boldsymbol{\theta}}_i \mid M)}{\frac{1}{N_1} \sum_{j=1}^{N_1} h(\boldsymbol{\theta}_j^* \mid M) g(\boldsymbol{\theta}_j^* \mid M)}, \tag{2.11}$$

where $\tilde{\boldsymbol{\theta}}_i$ and $\boldsymbol{\theta}_j^*$ are samples taken from $g(\boldsymbol{\theta} \mid M)$ and $p(\boldsymbol{\theta} \mid M, \mathbf{Y})$, respectively. It means that we need to take samples from both the proposal distribution and the posterior distribution of θ to estimate the marginal likelihood.

There are different choices of the proposal distribution and the bridge function. The proposal distribution $g(\boldsymbol{\theta})$ should closely approximate the posterior distribution and provide sufficient overlap with it. A convenient choice is a multivariate normal distribution whose first two moments are chosen to matched to those of the posterior distribution.

A bridge function chosen in Meng and Wong (1996) has the form

$$h(\boldsymbol{\theta} \mid M) = C \cdot \frac{1}{s_1 p(\mathbf{Y} \mid M, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid M) + s_2 p(\mathbf{Y} \mid M) g(\boldsymbol{\theta} \mid M)}, \tag{2.12}$$

where $s_1 = \frac{N_1}{N_2+N_1}$, $s_2 = \frac{N_2}{N_2+N_1}$, and C is a constant and is canceled out in (2.11) since both the numerator and the denominator of (2.11) include $h(\boldsymbol{\theta} \mid M)$. Thus, the value of C is not required. Meng and Wong (1996) showed that the bridge function minimizes the relative mean-squared error. Thus, this function is referred to as the "optimal bridge function". Since this optimal bridge function depends on the unknown $p(\mathbf{Y} \mid M)$, an iterative scheme is required to estimate the marginal likelihood. The formula on iteration $t+1$ can be obtained by applying (2.12) to (2.11) as follows:

$$
\begin{aligned}
\hat{p}(\mathbf{Y} \mid M)^{(t+1)} &= \frac{\frac{1}{N_2}\sum_{i=1}^{N_2} \frac{p(\mathbf{Y} \mid \tilde{\boldsymbol{\theta}}_i, M)\, p(\tilde{\boldsymbol{\theta}}_i \mid M)}{s_1\, p(\mathbf{Y} \mid \tilde{\boldsymbol{\theta}}_i, M)\, p(\tilde{\boldsymbol{\theta}}_i \mid M) + s_2\, \hat{p}(\mathbf{Y} \mid M)^{(t)}\, g(\tilde{\boldsymbol{\theta}}_i \mid M)}}{\frac{1}{N_1}\sum_{j=1}^{N_1} \frac{g(\boldsymbol{\theta}_j^* \mid M)}{s_1\, p(\mathbf{Y} \mid \boldsymbol{\theta}_j^*, M)\, p(\boldsymbol{\theta}_j^* \mid M) + s_2\, \hat{p}(\mathbf{Y} \mid M)^{(t)}\, g(\boldsymbol{\theta}_j^* \mid M)}} \\
&= \frac{\frac{1}{N_2}\sum_{i=1}^{N_2} \frac{p(y|\tilde{\boldsymbol{\theta}}_i)p(\tilde{\boldsymbol{\theta}}_i)}{s_1 p(y|\tilde{\boldsymbol{\theta}}_i)p(\tilde{\boldsymbol{\theta}}_i) + s_2 \hat{p}(\mathbf{Y}|M)^t g(\tilde{\boldsymbol{\theta}}_i)} \frac{1/g(\tilde{\boldsymbol{\theta}}_i)}{1/g(\tilde{\boldsymbol{\theta}}_i)}}{\frac{1}{N_1}\sum_{j=1}^{N_1} \frac{g(\boldsymbol{\theta}_j^*)}{s_1 p(y|\boldsymbol{\theta}_j^*)p(\boldsymbol{\theta}_j^*) + s_2 \hat{p}(\mathbf{Y}|M)^t g(\boldsymbol{\theta}_j^*)} \frac{1/g(\boldsymbol{\theta}_j^*)}{1/g(\boldsymbol{\theta}_j^*)}} \\
&= \frac{\frac{1}{N_2}\sum_{i=1}^{N_2} \frac{l_{2,i}}{s_1 l_{2,i} + s_2 \hat{p}(\mathbf{Y}|M)^{(t)}}}{\frac{1}{N_1}\sum_{j=1}^{N_1} \frac{1}{s_1 l_{1,j} + s_2 \hat{p}(\mathbf{Y}|M)^{(t)}}},
\end{aligned}
\tag{2.13}
$$

and $\tilde{\boldsymbol{\theta}}_i$ are samples taken from the proposal distribution and the posterior distribution, respectively, and $l_{1,j} = \frac{p(\mathbf{Y}|\boldsymbol{\theta}_j^*)p(\boldsymbol{\theta}_j^*)}{g(\boldsymbol{\theta}_j^*)}$ and $l_{2,i} = \frac{p(\mathbf{Y}|\tilde{\boldsymbol{\theta}}_i)p(\tilde{\boldsymbol{\theta}}_i)}{g(\tilde{\boldsymbol{\theta}}_i)}$.

The practical implementation consists of the following steps:

1. **Obtain posterior samples.** Draw $2N_1$ samples from the posterior distribution $p(\boldsymbol{\theta} \mid M, \mathbf{Y})$. Split them into two equal parts: the first N_1 samples are used to fit the proposal distribution and the remaining N_1 samples are used for the bridge estimator.

2. **Choose a proposal distribution.** Often a multivariate normal distribution is used because it is easy to sample from, has a closed-form density, and can often approximate a unimodal posterior distribution reasonably well through its mean vector and covariance matrix.

3. **Transform the posterior samples.** It may be necessary to transform the posterior samples. For example, if a multivariate normal distribution is used as the proposal distribution and if the parameter space is bounded (e.g. $[0.1]$), then transform the posterior samples to the real line so that the range of the posterior distribution is the same as that of the proposal distribution.

4. **Fit the proposal distribution to the first batch of N_1 transformed posterior samples.**

5. **Generate N_2 samples from the proposal distribution.**

6. **Compute the quantities $l_{2,i}$ and $l_{1,j}$.** For each proposal sample $\tilde{\boldsymbol{\theta}}_i$ evaluate $l_{2,i}$. For each sample $\boldsymbol{\theta}_j^*$ from the second batch of N_1 transformed posterior samples evaluate $l_{1,j}$.

7. **Iterate the update rule.** Choose an initial guess $\hat{p}(\mathbf{y})^{(0)}$. Repeatedly apply (2.13) until convergence, e.g. until $\left|\hat{p}(\mathbf{y})^{(t+1)} - \hat{p}(\mathbf{y})^{(t)}\right| / \hat{p}(\mathbf{y})^{(t+1)} < 10^{-10}$.

2.4 Multistratum designs

A multistratum design is an experimental design where experimental units are organized into groups (strata) and cannot be randomized completely. In this section, we introduce some of the popular multistratum designs.

2.4.1 Split-plot design

The split-plot design, originally developed by Fisher (1925) for use in agricultural experiments has appeared in most industry experiments. Such design contains two types of factors, hard-to-change (or whole-plot) factors and easy-to-change (or subplot) factors, and thus divides the experimental runs into two strata: the top stratum contains the whole plots, while the lower stratum contains the individual subplot runs. This structure creates two levels of randomization and two corresponding error terms in the analysis.

Table 2.1 of Box et al. (2005) illustrates a split-plot experiment designed to study the effects of temperature and coating on the corrosion resistance of steel bars. Since altering temperature requires significant time and energy due to thermal inertia, it is assigned to whole plots to avoid frequent and impractical changes, thus, saves time and costs. Therefore, temperature (360°C, 370°C, 380°C) is considered as a hard-to-change factor or a whole-plot factor. Each whole plot (numbered 1 to 6) corresponds to a batch or run at a specific temperature. The six whole plots are randomly assigned to the three temperature levels, with each temperature appearing twice (replicates). For instance, whole plots 1 and 6 are both at 360°C, 2 and 5 at 370°C, and 3 and 4 at 380°C. While coating type (C_1, C_2, C_3, C_4) can be changed based on the randomized order in process, which is treated as the subplot factor, since it is easy to change coating types. Within each whole plot, four steel bars receive one of the four coatings. The order of coatings within each whole plot is randomized independently. For example, whole plot 1 has coatings in the order C_2, C_3, C_1, C_4, while whole plot 2 has C_1, C_3, C_4, C_2. After each experiment, the corrosion resistance (the

numerical values in Table 2.1) is measured for each combination.

Table 2.1: Split-plot design with data for studying the corrosion resistance of steel bars (Box et al., 2005)

Whole-plot	Temperature (°C)	Coating (randomized order)			
		C_2	C_3	C_1	C_4
1	360°	73	83	67	89
		C_1	C_3	C_4	C_2
2	370°	65	87	86	91
		C_3	C_1	C_2	C_4
3	380°	147	155	127	212
		C_4	C_3	C_2	C_1
4	380°	153	90	100	108
		C_4	C_1	C_3	C_2
5	370°	150	140	121	142
		C_1	C_4	C_2	C_3
6	360°	33	54	8	46

2.4.2 Three-stratum design

In practice, some designs may involve very-hard-to-change, hard-to-change, and easy-to-change factors, such as split-split-plot designs and staggered-level designs. These designs involve three levels of randomizations and consist of whole plots, subplots, and individual runs nested within subplots. The top stratum is the whole-plot stratum. During the experiment, the levels of the whole-plot factors, or very hard-to-change factors, are changed the least frequently. The levels of subplot factors or hard-to-change factors are changed more often. The sub-subplot factors, or easy-to-change factors, should be reset between runs, regardless of whether their levels are changed.

Table 2.2: 36-run I-optimal split-split-plot design with a very hard-to-change factors w, a hard-to-change factor s and three easy-to-change factors, t_1, t_2 and t_3.

Run	wp	sp	w	s	t_1	t_2	t_3
1	1	1	0	0	0	0	0
2	1	1	0	0	0	1	−1
3	1	1	0	0	1	0	0
4	1	2	0	−1	0	0	−1
5	1	2	0	−1	−1	1	1
6	1	2	0	−1	1	−1	1
7	2	3	0	0	−1	0	0
8	2	3	0	0	0	0	−1
9	2	3	0	0	0	−1	0
10	2	4	0	1	0	0	−1
11	2	4	0	1	1	1	1
12	2	4	0	1	−1	−1	1
13	3	5	−1	1	−1	1	−1
14	3	5	−1	1	0	0	1
15	3	5	−1	1	1	−1	−1
16	3	6	−1	−1	−1	−1	−1
17	3	6	−1	−1	1	1	−1
18	3	6	−1	−1	0	0	1
19	4	7	1	1	1	1	−1
20	4	7	1	1	−1	1	1
21	4	7	1	1	0	−1	0
22	4	8	1	0	1	−1	−1
23	4	8	1	0	−1	0	0
24	4	8	1	0	0	0	1
25	5	9	1	−1	1	1	0
26	5	9	1	−1	−1	−1	1
27	5	9	1	−1	−1	1	−1
28	5	10	1	1	1	−1	1
29	5	10	1	1	−1	−1	−1
30	5	10	1	1	0	1	0
31	6	11	0	−1	1	1	1
32	6	11	0	−1	1	−1	−1
33	6	11	0	−1	−1	0	0
34	6	12	0	0	1	0	0
35	6	12	0	0	0	1	0
36	6	12	0	0	−1	−1	1

Table 2.2 shows a 36-run split-split-plot design in Arnouts and Goos (2015) Table 14. It involves one very-hard-to-change factor w, one hard-to-change factor s, and two easy-to-change factors t_1 and t_2. In the design, the level of s must be reset whenever the level of w is changed in order to maintain the nested structure of the designs.

A staggered level design is a modern experimental design used when there are two or more hard-to-change factors. It was introduced as a cost-efficient and statistically powerful alternative to traditional designs like split-split-plot.

While split-split-plot designs require resetting factors at the same time, that creating a nested structure, staggered level designs allow each hard-to-change factor to be reset independently and at different points in the experiment, thus, do not have a nested structure. This reduces the number of physical resets compared to split-split-plot designs. The unique staggered pattern of adjustments is what gives the design its name.

Table 2.3: 20-run I-optimal staggered-level design and split-split-plot design with one very hard-to-change factor w, one hard-to-change factor s, and two easy-to-change factors t_1 and t_2.

	Staggered-level design				Split-split-plot design			
Run	w	s	t_1	t_2	w	s	t_1	t_2
1	0	1	0	1	0	−1	0	1
2	0	1	1	0	0	−1	1	−1
3	0	0	1	−1	0	0	0	0
4	0	0	0	0	0	0	−1	0
5	1	0	0	0	−1	−1	−1	−1
6	1	0	1	1	−1	−1	1	1
7	1	−1	−1	−1	−1	1	1	−1
8	1	−1	−1	−1	−1	1	−1	1
9	−1	−1	−1	1	1	1	−1	1
10	−1	−1	1	−1	−1	1	1	−1
11	1	−1	−1	−1	1	0	0	−1
12	−1	1	1	1	1	0	1	0
13	1	1	−1	0	1	−1	−1	0
14	1	1	0	−1	1	−1	1	1
15	1	0	0	0	1	0	0	0
16	1	0	−1	−1	1	0	−1	−1
17	0	0	0	1	0	1	1	1
18	0	0	−1	1	0	1	−1	−1
19	0	−1	1	1	0	0	−1	1
20	0	−1	−1	−1	0	0	0	0

Table 2.3 presents the I-optimal 20-run staggered-level design and split-split-plot design for response surface experiments in Arnouts and Goos (2015). Each design involves a very hard-to-change factor w, a hard-to-change factor s, and two easy-to-change factors t_1 and t_2. Different from the split-split-plot design, the level of s in the staggered-level design is reset at different points in time than the level of w. Because the levels of w and s are reset independently and at different intervals, they are not nested. Arnouts and Goos (2015) pointed out that the split-split-plot design requires 5 settings of w and 10 settings of s, while the

staggered-level design requires only 5 settings of w and 6 settings of s, making it the more cost-efficient.

Chapter 3

Response Optimization on Data with a Multistratum Structure Under Normal Assumptions of Random Effects and Errors

3.1 Introduction

Response optimization is a statistical method used to find the combination of input factor settings that produces the most desirable value of a response variable. By accounting for model uncertainty, Bayesian model averaging is a popular approach for improving average predictive performance, compared to relying solely on a single model. However, the response optimization methods available in the literature cannot be directly applied to data collected from multistratum

designs. This chapter considers we consider this problem for multistratum data using the Bayesian model averaging approach. Random effects and errors are assumed to follow normal distributions. We derive the marginal likelihood and posterior model probability. We show that the posterior predictive distribution of a future response is a Student's t-distribution. Bridge sampling method is used to estimate the marginal likelihood. The objective is to identify optimal values of factors that result in the maximum probability of the predicted response lying within target bounds.

The remainder of this chapter is structured as follows. Section 3.2 introduces the general multistratum model and the optimization problem. Section 3.3 develops the posterior predictive distribution for multistratum designs and derives a closed-form expression showing that the posterior predictive density follows a Student's t-distribution. The posterior model probabilities are derived in Section 3.4, and the optimization procedure is outlined in Section 3.5. Section 3.6 illustrates the methodology using a 28-run I-optimal staggered-level design with simulated data, including sensitivity analyses for the prior parameters π, τ, and α. A simulation study comparing the performance of Bayesian model averaging with AIC and BIC is also performed.

3.2 The model, Bayesian model averaging and optimization problem

Consider an n-run, s-stratum design with m factors $x_1, x_2, ..., x_m$. Assume that there are b_l units and m_l factors in the lth stratum, where $l = 1, \cdots, s-1, s$. Thus

$\sum_{l=1}^{s} m_l = m$ and $b_s = n$. Then, the model can be written as:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \sum_{l=1}^{s-1} \mathbf{U}_l \boldsymbol{\gamma}_l + \boldsymbol{\epsilon}, \tag{3.1}$$

where $\mathbf{Y}$ is an $n \times 1$ vector of the responses, $\mathbf{X}$ is the model matrix, $\boldsymbol{\beta}$ is the $(p+1) \times 1$ vector of effects of the model terms, p is the number of terms in the model excluding the constant term, $\mathbf{U}_l$ is the $n \times b_l$ indicator matrix for $\boldsymbol{\gamma}_l$, $\boldsymbol{\gamma}_l$ is the $b_l \times 1$ vector of the random effects for the lth stratum, and $\boldsymbol{\epsilon}$ is the $n \times 1$ vector of random errors. Note that $\mathbf{U}_s = \mathbf{I}_n$, where $\mathbf{I}_n$ is the $n \times n$ identity matrix and $\boldsymbol{\gamma}_l = \boldsymbol{\varepsilon}$. When $s = 2$, then it is a split-plot or blocked design.

Assume that $\boldsymbol{\gamma}_l \sim N(\mathbf{0}_{b_l}, \sigma_l^2 \mathbf{I}_{b_l})$, where $\mathbf{0}_{b_l}$ is the $b_l \times 1$ vector of zeros, $\boldsymbol{\epsilon} \sim N(\mathbf{0}_n, \sigma^2 \mathbf{I}_n)$, and $\boldsymbol{\gamma}_k$ and $\boldsymbol{\gamma}_h$ are independent for $k \neq h$. Define $\sigma_w^2 = \sum_{l=1}^{s-1} \sigma_l^2 + \sigma^2$ and $\rho_l = \sigma_l^2 / \sigma_w^2$ for $l = 1, \ldots, s-1$, and $\rho_s = \sigma^2 / \sigma_w^2$. Let $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_s)$. Then $\sum_{l=1}^{s} \rho_l = 1$.

The variance-covariance matrix of $\mathbf{Y}$ is

$$\begin{aligned}
\mathrm{Var}(\mathbf{Y}) &= \sum_{l=1}^{s-1} \mathbf{U}_l \,\mathrm{Var}(\boldsymbol{\gamma}_l) \mathbf{U}_l^{\mathsf{T}} + \mathrm{Var}(\boldsymbol{\epsilon}) \\
&= \sum_{l=1}^{s-1} \sigma_l^2 \mathbf{U}_l \mathbf{U}_l^{\mathsf{T}} + \sigma^2 \mathbf{I}_n. \\
&= \sigma_w^2 \left(\sum_{l=1}^{s-1} \frac{\sigma_l^2}{\sigma_w^2} \mathbf{U}_l \mathbf{U}_l^{\mathsf{T}} + \frac{\sigma^2}{\sigma_w^2} \mathbf{I}_n \right) \\
&= \sigma_w^2 \left(\sum_{l=1}^{s-1} \rho_l \mathbf{U}_l \mathbf{U}_l^{\mathsf{T}} + \rho_s I_n \right) \\
&= \sigma_w^2 \mathbf{V},
\end{aligned} \tag{3.2}$$

where

$$\mathbf{V} = \sum_{l=1}^{s-1} \rho_l \mathbf{U}_l \mathbf{U}_l^\top + \rho_s \mathbf{I}_n.$$

The goal of the optimization is to identify the factor levels that result in the maximum probability of the predicted response Y^* falling within a given region. Thus, the optimal factor levels can be determined by maximizing the model-averaged posterior probability z, which is defined as

$$\Pr(L \leq Y^* \leq U | \mathbf{x}^*, \mathbf{Y}) = \sum_i \left[\int_L^U p(Y^* | M_i, \mathbf{x}^*, \mathbf{Y}) dY^* \right] P(M_i | \mathbf{Y}), \tag{3.3}$$

where L and U are given lower bound and upper bound, respectively. The maximum model-averaged posterior probability is denoted by z^*.

3.3 Posterior predictive distribution

Assume that the prior for the total variance for a given model M and $\boldsymbol{\rho}$ follows the inverse-gamma distribution with parameters α and δ, that is,

$$\sigma_w^2 | M, \boldsymbol{\rho} \sim IG(\alpha, \delta). \tag{3.4}$$

For the coefficient $\boldsymbol{\beta}$, we will assume a non-informative prior for the intercept β_0 and independent normally distributed priors for the rest of the entries in $\boldsymbol{\beta}$, that is,

$$\boldsymbol{\beta} | M, \sigma_w^2, \boldsymbol{\rho} \sim N(\mathbf{0}_{p+1}, \sigma_w^2 \boldsymbol{K}_\tau{}^{-1}), \tag{3.5}$$

where

$$\boldsymbol{K}_\tau = \frac{1}{\tau^2}\begin{pmatrix} 0 & \mathbf{0}_p' \\ \mathbf{0}_p & \mathbf{I}_p \end{pmatrix}$$

and τ captures the magnitude of the effect relative to the experimental noise.

The marginal likelihood of M given $\boldsymbol{\rho}$ can be found by

$$L(\mathbf{Y}|M,\boldsymbol{\rho}) = \int_{\sigma_w^2}\int_{\boldsymbol{\beta}} L(\mathbf{Y}|M,\sigma_w^2,\boldsymbol{\beta},\boldsymbol{\rho})p(\boldsymbol{\beta}|M,\sigma_w^2,\boldsymbol{\rho})p(\sigma_w^2|M,\boldsymbol{\rho})d\boldsymbol{\beta}d\sigma_w^2.$$

Since $\mathbf{Y}$ is normally distributed with mean $\mathbf{X}\boldsymbol{\beta}$ and the variance-covariance matrix of $\mathbf{Y}$ is $\sigma_w^2\mathbf{V}$, we have

$$L(\mathbf{Y}|M,\sigma_w^2,\boldsymbol{\beta},\boldsymbol{\rho}) = (2\pi)^{-\frac{n}{2}}\sigma_w^{-n}\ |\mathbf{V}|^{-\frac{1}{2}}\ \exp[-\frac{1}{2\sigma_w^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})]. \tag{3.6}$$

Note that $\mathbf{V}$ is related to $\boldsymbol{\rho}$. In addition, (5.1) and (5.2) indicates that

$$p(\sigma_w^2|M,\boldsymbol{\rho}) = \frac{\delta^\alpha}{\Gamma(\alpha)}\sigma_w^{-2\alpha-2}\exp[-\frac{\delta}{\sigma_w^2}], \tag{3.7}$$

and

$$\begin{aligned} p(\boldsymbol{\beta}|M,\sigma_w^2,\boldsymbol{\rho}) &= (2\pi)^{-\frac{p+1}{2}}|\sigma_w^2\boldsymbol{K}_\tau{}^{-1}|^{-\frac{1}{2}}\exp[-\frac{1}{2\sigma_w^2}\boldsymbol{\beta}^T\boldsymbol{K}_\tau\boldsymbol{\beta}] \\ &= (2\pi)^{-\frac{p+1}{2}}\tau^{-p}\sigma_w^{-p}\ \exp[-\frac{1}{2\sigma_w^2}\boldsymbol{\beta}^T\boldsymbol{K}_\tau\boldsymbol{\beta}], \end{aligned}$$

respectively. It follows that

$$\begin{aligned} L(\mathbf{Y}|M,\boldsymbol{\rho}) = \int_{\sigma_w^2}\int_{\boldsymbol{\beta}}&(2\pi)^{-\frac{n+p+1}{2}}\frac{\delta^\alpha}{\Gamma(\alpha)}\sigma_w^{-(n+p+2\alpha+2)}\tau^{-p}\ |\mathbf{V}|^{-\frac{1}{2}} \\ &\exp\{-\frac{1}{2\sigma_w^2}[(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\ +\boldsymbol{\beta}^T\boldsymbol{K}_\tau\boldsymbol{\beta}+2\delta]\}d\boldsymbol{\beta}d\sigma_w^2. \end{aligned}$$

Note that

$$(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\ +\boldsymbol{\beta}^T\boldsymbol{K}_\tau\boldsymbol{\beta}+2\delta = S+(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}+\boldsymbol{K}_\tau)(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}), \tag{3.8}$$

where

$$S=\mathbf{Y}^T\mathbf{V}^{-1}\mathbf{Y}-\ \mathbf{Y}^T\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}+\boldsymbol{K}_\tau)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y}+2\delta$$

and

$$\hat{\boldsymbol{\beta}}=(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}+\boldsymbol{K}_\tau)^{-1}\mathbf{X}^T\mathbf{V}^{-1}\ \mathbf{Y}$$

is the Bayesian estimator for $\boldsymbol{\beta}$ (Lin, 2018). Then,

$$\begin{aligned}
&L(\mathbf{Y}|M,\boldsymbol{\rho})\\
&=\int_{\sigma_w^2}\int_{\boldsymbol{\beta}}(2\pi)^{-\frac{n+p+1}{2}}\frac{\delta^\alpha}{\Gamma(\alpha)}\tau^{-p}|\mathbf{V}|^{-\frac{1}{2}}\ \sigma_w^{-(n+2\alpha+1)}\exp[-\frac{1}{2\sigma_w^2}S]\\
&\sigma_w^{-(p+1)}\exp\{-\frac{1}{2\sigma_w^2}[(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}+\boldsymbol{K}_\tau)(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})]\}d\boldsymbol{\beta}d\sigma_w^2\\
&=\frac{(2\delta)^\alpha\Gamma(\frac{n+2\alpha-1}{2})}{\sqrt{2}\Gamma(\alpha)}\pi^{-\frac{n}{2}}\tau^{-p}|\mathbf{V}|^{-\frac{1}{2}}|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}+\boldsymbol{K}_\tau|^{-\frac{1}{2}}S^{-\frac{n+2\alpha-1}{2}}\\
&\propto\tau^{-p}|\mathbf{V}|^{-\frac{1}{2}}|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}+\boldsymbol{K}_\tau|^{-\frac{1}{2}}S^{-\frac{n+2\alpha-1}{2}}.
\end{aligned}$$

The following theorem presents the posterior predictive distribution of the new response for a given model at a new set of regressors $\mathbf{x}^*$.

Theorem 3.3.1. *For an n-run, s-stratum and m-factor design, the new response Y^* at a new $\mathbf{x}^*$ for a given model M and the data follows a Student's t-distribution,*

that is,

$$Y^*|M,\mathbf{x}^*,\mathbf{Y},\boldsymbol{\rho} \sim t_{n+2\alpha-1}\left(\mathbf{x}^{*T}\hat{\boldsymbol{\beta}},\ \frac{S}{n+2\alpha-1}\left[1+\mathbf{x}^{*T}(\boldsymbol{K}_\tau+\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{x}^*\right]\right).$$

Proof: The posterior predictive density given the data and $\boldsymbol{\rho}$ can be found by

$$p(Y^*|M,\mathbf{x}^*,\mathbf{Y},\boldsymbol{\rho}) = \int_{\sigma_w^2}\int_{\boldsymbol{\beta}} L(Y^*|M,\mathbf{x}^*,\mathbf{Y},\sigma_w^2,\boldsymbol{\beta},\boldsymbol{\rho})p(\sigma_w^2,\boldsymbol{\beta}|M,\mathbf{Y},\boldsymbol{\rho})d\boldsymbol{\beta}d\sigma_w^2,$$

where $L(Y^*|M,\mathbf{x}^*,\mathbf{Y},\sigma_w^2,\boldsymbol{\beta},\boldsymbol{\rho})$ is the likelihood function and $p(\sigma_w^2,\beta|M,\mathbf{Y},\boldsymbol{\rho})$ is the joint posterior of σ_w^2 and β. Assume that Y^* is normally distributed with mean $\mathbf{x}^*\boldsymbol{\beta}$ and variance σ_w^2, we get

$$L(Y^*|M,\mathbf{x}^*,\mathbf{Y},\sigma_w^2,\boldsymbol{\beta},\boldsymbol{\rho}) = (2\pi)^{-\frac{1}{2}}\sigma_w^{-1}\ \exp[-\frac{1}{2\sigma_w^2}(Y^*-\mathbf{x}^{*T}\boldsymbol{\beta})^T(Y^*-\mathbf{x}^{*T}\boldsymbol{\beta})].$$

By (3.6), (3.7), and (3.8),

$$\begin{aligned}
&p(\sigma_w^2,\boldsymbol{\beta}|M,\mathbf{Y},\boldsymbol{\rho})\\
&= L(\mathbf{Y}|M,\sigma_w^2,\boldsymbol{\beta},\boldsymbol{\rho})p(\boldsymbol{\beta}|M,\sigma_w^2,\boldsymbol{\rho})p(\sigma_w^2|M,\boldsymbol{\rho})\\
&= (2\pi)^{-\frac{n+p+1}{2}}\frac{\delta^\alpha}{\Gamma(\alpha)}\sigma_w^{-(n+p+2\alpha+2)}\tau^{-p}\mid\mathbf{V}|^{-\frac{1}{2}}\\
&\times\exp\{-\frac{1}{2\sigma_w^2}[(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\ +\boldsymbol{\beta}^T\boldsymbol{K}_\tau\boldsymbol{\beta}+2\delta]\}.
\end{aligned}$$

Define the scalar Ω as

$$\Omega = \ (Y^*-\mathbf{x}^{*T}\boldsymbol{\beta})^T(Y^*-\mathbf{x}^{*T}\boldsymbol{\beta})\ +(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\ +\boldsymbol{\beta}^T\boldsymbol{K}_\tau\boldsymbol{\beta}+2\delta.$$

By (3.8), Ω can be written as

$$\Omega = (Y^* - \mathbf{x}^{*T}\boldsymbol{\beta})^T(Y^* - \mathbf{x}^{*T}\boldsymbol{\beta}) + S + (\boldsymbol{\beta} - \hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X} + \boldsymbol{K}_\tau)(\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}).$$

It follows that

$$\begin{aligned}
&p(Y^*|M, \mathbf{x}^*, \mathbf{Y}, \boldsymbol{\rho}) \\
&= \int_{\sigma_w^2}\int_{\boldsymbol{\beta}} L(Y^*|M, \mathbf{x}^*, \mathbf{Y}, \sigma_w^2, \boldsymbol{\beta}, \boldsymbol{\rho}) p(\sigma_w^2, \boldsymbol{\beta}|M, \mathbf{Y}, \boldsymbol{\rho}) d\boldsymbol{\beta} d\sigma_w^2 \\
&= \int_{\sigma_w^2}\int_{\boldsymbol{\beta}} (2\pi)^{-\frac{1}{2}} \sigma_w^{-1} \exp[-\frac{1}{2\sigma_w^2}(Y^* - \mathbf{x}^{*T}\boldsymbol{\beta})^T(Y^* - \mathbf{x}^{*T}\boldsymbol{\beta})] \\
&\times (2\pi)^{-\frac{n+p+1}{2}} \frac{\delta^\alpha}{\Gamma(\alpha)} \sigma_w^{-(n+p+2\alpha+2)} \tau^{-p} \mid \mathbf{V}\mid^{-\frac{1}{2}} \\
&\times \exp\{-\frac{1}{2\sigma_w^2}[(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\beta}^T\boldsymbol{K}_\tau\boldsymbol{\beta} + 2\delta]\} d\boldsymbol{\beta} d\sigma_w^2 \\
&= (2\pi)^{-\frac{n+p+2}{2}} \frac{\delta^\alpha}{\Gamma(\alpha)} \int_{\sigma_w^2}\int_{\boldsymbol{\beta}} \sigma_w^{-(n+p+2\alpha+3)} \tau^{-p} \mid \mathbf{V}\mid^{-\frac{1}{2}} \exp\{-\frac{\Omega}{2\sigma_w^2}\}\, d\boldsymbol{\beta} d\sigma_w^2.
\end{aligned}$$

Let $u = \Omega/2\sigma_w^2$. Then

$$\begin{aligned}
p(Y^*|M, \mathbf{x}^*, \mathbf{Y}, \boldsymbol{\rho}) &= (2\pi)^{-\frac{n+p+2}{2}} \frac{\delta^\alpha}{\Gamma(\alpha)} \tau^{-p} \mid \mathbf{V}\mid^{-\frac{1}{2}} \\
&\times \int_{\boldsymbol{\beta}} (\frac{\Omega}{2})^{-(n+p+2\alpha+1)/2} \left[\int_0^\infty u^{(n+p+2\alpha-1)/2} \exp(-u) du\right] d\boldsymbol{\beta}.
\end{aligned}$$

The inner integral inside the square brackets is the gamma function, $\Gamma(\frac{n+p+2\alpha+1}{2})$, which is a constant.

Let

$$\mathbf{U} = (\boldsymbol{K}_\tau + \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X} + \mathbf{x}^*\mathbf{x}^{*T})^{-1}(\hat{\boldsymbol{\beta}}^T(\boldsymbol{K}_\tau + \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}) + Y^*\mathbf{x}^{*T})^T, \qquad (3.9)$$

which is a $(p+1)\times 1$ vector. Then we can rewrite $\boldsymbol{\Omega}$ as

$$\boldsymbol{\Omega} = \quad W + (\boldsymbol{\beta}-\mathbf{U})^T(\boldsymbol{K}_\tau + \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X} + \mathbf{x}^*\mathbf{x}^{*T})(\boldsymbol{\beta}-\mathbf{U}), \tag{3.10}$$

where

$$\begin{aligned}
W &= (Y^* - \mathbf{x}^{*T}\boldsymbol{\beta})^T(Y^* - \mathbf{x}^{*T}\boldsymbol{\beta}) \; + S + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X} + \boldsymbol{K}_\tau)(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}) \\
&\quad - (\boldsymbol{\beta}-\mathbf{U})^T(\boldsymbol{K}_\tau + \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X} + \mathbf{x}^*\mathbf{x}^{*T})(\boldsymbol{\beta}-\mathbf{U}) \\
&\quad = S + (Y^*)^2 + \hat{\boldsymbol{\beta}}^T(\boldsymbol{K}_\tau + \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})\hat{\boldsymbol{\beta}} \\
&\quad - (\hat{\boldsymbol{\beta}}^T(\boldsymbol{K}_\tau + \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}) + Y^*\mathbf{x}_i^{*T})(\boldsymbol{K}_\tau + \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X} + \mathbf{x}^*\mathbf{x}^{*T})^{-1} \\
&\quad (\hat{\boldsymbol{\beta}}^T(\boldsymbol{K}_\tau + \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}) + Y^*\mathbf{x}^{*T})^T
\end{aligned}$$

is a scalar. Thus the posterior predictive density becomes

$$\begin{aligned}
&p(Y^*|M,\mathbf{x}^*,\mathbf{Y},\boldsymbol{\rho}) \\
&\propto \tau^{-p}\,|\mathbf{V}|^{-\frac{1}{2}}\Gamma(\frac{n+p+2\alpha+1}{2}) \times \\
&\int_{\boldsymbol{\beta}} [W + (\boldsymbol{\beta}-\mathbf{U})^T(\boldsymbol{K}_\tau + \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X} + \mathbf{x}^*\mathbf{x}^{*T})(\boldsymbol{\beta}-\mathbf{U})]^{-(n+p+2\alpha+1)/2} d\boldsymbol{\beta} \\
&\propto \tau^{-p}\,|\mathbf{V}|^{-\frac{1}{2}}\Gamma(\frac{n+p+2\alpha+1}{2}) \times \frac{1}{W^{(n+p+2\alpha+1)/2}} \\
&\int_{\boldsymbol{\beta}} [1 + (\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^*)^T(\boldsymbol{K}_\tau + \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X} + \mathbf{x}^*\mathbf{x}^{*T})(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}^*)/W]^{-(n+p+2\alpha+1)/2} d\boldsymbol{\beta}
\end{aligned}$$

The integrand in (3.11) is a multivariate t-distribution (Press, 1982). It follows that

$$p(Y^*|M,\mathbf{x}^*,\mathbf{Y},\boldsymbol{\rho}) \propto \frac{1}{W^{(n+2\alpha)/2}}. \tag{3.11}$$

Now, if we write

$$\mu_y^* = \mathbf{x}^{*T}\hat{\boldsymbol{\beta}}$$

and

$$\sigma_y^{*2} = \frac{S}{n+2\alpha-1}[1+\mathbf{x}^{*T}(\boldsymbol{K}_\tau+\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{x}^*],$$

then, based on Sherman-Morrison identity, W can be written as

$$\begin{aligned} W &= S+\frac{(Y^*-\mathbf{x}^*\hat{\boldsymbol{\beta}})^2}{1+\mathbf{x}^{*T}(\boldsymbol{K}_\tau+\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{x}^*} \\ &= S[1+\frac{1}{n+2\alpha-1}\times\frac{(Y^*-\mu_y^*)^2}{\sigma_y^{*2}}]. \end{aligned}$$

Now, we have

$$\begin{aligned} p(Y^*|M,\mathbf{x}^*,\mathbf{Y},\boldsymbol{\rho}) &\propto \frac{1}{W^{(n+2\alpha)/2}} \\ &\propto (S[1+\frac{1}{n+2\alpha-1}\times\frac{(Y^*-\mu_y^*)^2}{\sigma_y^{*2}}])^{-(n+2\alpha)/2} \\ &\propto [1+\frac{1}{n+2\alpha-1}\times\frac{(Y^*-\mu_y^*)^2}{\sigma_y^{*2}}]^{-(n+2\alpha)/2}, \end{aligned}$$

which indicates $Y^*|M,\mathbf{x}^*,\mathbf{Y},\boldsymbol{\rho}$ follows Student's t-distribution with mean μ_y^* and variance σ_y^{*2}, that is,

$$\begin{aligned} Y^*|M,\mathbf{x}^*,\mathbf{Y},\boldsymbol{\rho} &\sim t_{n+2\alpha-1}(\mu_y^*,\sigma_y^{*2}) \\ &\sim t_{n+2\alpha-1}(\mathbf{x}^{*T}\hat{\boldsymbol{\beta}},\frac{S}{n+2\alpha-1}[1+\mathbf{x}^{*T}(\boldsymbol{K}_\tau+\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{x}^*]) \end{aligned}$$

with the degrees of freedom $n+2\alpha-1$. This completes the proof.

Since the cumulative distribution function of a Student's t random variable can be calculated through the regularized incomplete beta function, the cumula-

tive posterior predictive density of the response given M, $\mathbf{x}^*$ and the data can be determined by the following formula in Johnson et al. (1995):

$$\Pr(\frac{Y^*-\mu_y^*}{\sigma_y^*} < t|M,\mathbf{x}^*,\mathbf{Y},\boldsymbol{\rho}) = \frac{1}{2}[1+\mathbf{I}_{t^2/(\nu+t^2)}(\frac{1}{2},\frac{\nu}{2})], \tag{3.12}$$

where $t > 0$, $\mathbf{I}_{t^2/(\nu+t^2)}(\frac{1}{2},\frac{\nu}{2})$ is the regularized incomplete beta function (Abramowitz and Stegun, 1965), and $\nu = n+2\alpha-1$.

In our example, for simplicity, we assume that $\alpha \to 0$ and $\delta \to 0$ in Theorem (3.3.1) and (3.12). In this case, the prior for σ_w^2 given the model M is proportional to $1/\sigma_w^2$.

When $\boldsymbol{\rho}$ is unknown, the marginal likelihood

$$\begin{aligned} L(\mathbf{Y}|M) &= \int_{\boldsymbol{\rho}}\int_{\sigma_w^2}\int_{\boldsymbol{\beta}} L(\mathbf{Y}|M,\sigma_w^2,\boldsymbol{\beta},\boldsymbol{\rho})p(\boldsymbol{\beta}|M,\sigma_w^2,\boldsymbol{\rho})p(\sigma_w^2|M,\boldsymbol{\rho})p(\boldsymbol{\rho})d\boldsymbol{\beta}d\sigma_w^2 d\boldsymbol{\rho} \\ &= \frac{(2\delta)^\alpha\Gamma(\frac{n+2\alpha-1}{2})}{\sqrt{2}\Gamma(\alpha)}\pi^{-\frac{n}{2}}\int_{\boldsymbol{\rho}}\tau^{-p}|\mathbf{V}|^{-\frac{1}{2}}|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}+\boldsymbol{K}_\tau|^{-\frac{1}{2}}S^{-\frac{n+2\alpha-1}{2}}p(\boldsymbol{\rho})d\boldsymbol{\rho}. \end{aligned}$$

where $p(\boldsymbol{\rho})$ is the prior probability density function of $\boldsymbol{\rho}$, and the posterior predictive density of the response Y^* given M, $\mathbf{x}^*$ and the data is

$$\begin{aligned} p(Y^*|M,\mathbf{x}^*,\mathbf{Y}) &\propto \int_{\boldsymbol{\rho}}\int_{\sigma_w^2}\int_{\boldsymbol{\beta}} L(Y^*|M,\mathbf{x}^*,\mathbf{Y},\sigma_w^2,\boldsymbol{\beta})p(\sigma_w^2,\boldsymbol{\beta},\boldsymbol{\rho}|M,\mathbf{Y})d\boldsymbol{\beta}d\sigma_w^2 d\boldsymbol{\rho}, \\ &\propto \int_{\boldsymbol{\rho}}[1+\frac{1}{n+2\alpha-1}\times\frac{(Y^*-\mu_y^*)^2}{\sigma_y^{*2}}\,]^{-(n+2\alpha)/2}p(\boldsymbol{\rho})d\boldsymbol{\rho} \end{aligned}$$

We assume that $\boldsymbol{\rho}$ follows a Dirichlet distribution with concentration parameters $\boldsymbol{\alpha} = (\alpha_1,\ldots,\alpha_s)$, following the suggestion of Lin and Yang (2022). Its

probability density function is:

$$p(\boldsymbol{\rho}) = \frac{1}{\mathrm{B}(\boldsymbol{\alpha})} \prod_{l=1}^{s} \rho_l^{\alpha_l - 1}, \tag{3.13}$$

where $\rho_l > 0$ for $l = 1, \ldots, s$, $\sum_{l=1}^{s} \rho_l = 1$, and $\mathrm{B}(\boldsymbol{\alpha}) = \frac{\prod_{l=1}^{s} \Gamma(\alpha_l)}{\Gamma(\sum_{l=1}^{s} \alpha_l)}$ is the multivariate beta function. Different choices of $\boldsymbol{\alpha}$ are considered in Lin and Yang (2022).

3.4 Posterior model probability

The posterior probability of M given the data can be calculated by Bayes' theorem,

$$P(M|\mathbf{Y}) = \frac{L(\mathbf{Y}|M)P(M)}{\sum L(\mathbf{Y}|M)P(M)}. \tag{3.14}$$

Since the denominator of (3.14) is a constant, by (2.1) and (3.9), the posterior probability of M_i given the data is

$$P(M_i|\mathbf{Y}, \boldsymbol{\rho}) \propto \quad \pi^{f_i}(1-\pi)^{m-f_i} \tau^{-t_i} |\mathbf{V}|^{-\frac{1}{2}} |\mathbf{X}_i^T \mathbf{V}^{-1} \mathbf{X}_i + \boldsymbol{K}_{\tau i}|^{-\frac{1}{2}} S_i^{-\frac{n+2\alpha-1}{2}},$$

where t_i is the number of terms in model M_i excluding the constant term, $\mathbf{X}_i$ is the model matrix,

$$\boldsymbol{K}_{\tau i} = \frac{1}{\tau^2} \begin{pmatrix} 0 & \mathbf{0}_{t_i}' \\ \mathbf{0}_{t_i} & \mathbf{I}_{t_i} \end{pmatrix},$$

and

$$S_i = \mathbf{Y}^T\mathbf{V}^{-1}\mathbf{Y} - \mathbf{Y}^T\mathbf{V}^{-1}\mathbf{X}_i(\mathbf{X}_i^T\mathbf{V}^{-1}\mathbf{X}_i + \boldsymbol{K}_{\tau i})^{-1}\mathbf{X}_i^T\mathbf{V}^{-1}\mathbf{Y} + 2\delta.$$

When $\boldsymbol{\rho}$ is unknown, the posterior probability of M_i given the data is

$$P(M_i|\mathbf{Y}) \propto \pi^{f_i}(1-\pi)^{m-f_i}\tau^{-t_i}\int_{\boldsymbol{\rho}}|\mathbf{V}|^{-\frac{1}{2}}|\mathbf{X}_i^T\mathbf{V}^{-1}\mathbf{X}_i + \boldsymbol{K}_{\tau i}|^{-\frac{1}{2}}S_i^{-\frac{n+2\alpha-1}{2}}p(\boldsymbol{\rho})d\boldsymbol{\rho}.$$

3.5 Optimization procedure

To facilitate the Markov Chain Monte Carlo (MCMC) sampling on the real line, we need to transform the constrained variance components $\boldsymbol{\rho} = (\rho_1, \rho_2, \rho_3)$ to unconstrained parameters $\boldsymbol{\theta} = (\theta_1, \theta_2)$ (Aitchison and Shen, 1980):

$$\rho_1 = \frac{e^{\theta_1}}{1+e^{\theta_1}+e^{\theta_2}}, \quad \rho_2 = \frac{e^{\theta_2}}{1+e^{\theta_1}+e^{\theta_2}}, \quad \rho_3 = 1-\rho_1-\rho_2 = \frac{1}{1+e^{\theta_1}+e^{\theta_2}}.$$

The determinant of Jacobian matrix $\mathbf{J}$ of the transformation from (θ_1, θ_2) to (ρ_1, ρ_2) is:

$$\det(\mathbf{J}) = \exp\left[\theta_1 + \theta_2 - 3\log(1+e^{\theta_1}+e^{\theta_2})\right]. \tag{3.15}$$

To show it, let $D = 1+e^{\theta_1}+e^{\theta_2}$. Then:

$$\rho_1 = \frac{e^{\theta_1}}{D}, \quad \rho_2 = \frac{e^{\theta_2}}{D}, \quad \rho_3 = \frac{1}{D}. \tag{3.16}$$

The Jacobian matrix for the transformation is:

$$\mathbf{J} = \begin{bmatrix} \rho_1(\rho_2+\rho_3) & -\rho_1\rho_2 \\ -\rho_1\rho_2 & \rho_2(\rho_1+\rho_3) \end{bmatrix}.$$

Taking the determinant:

$$\det(\mathbf{J}) = \rho_1\rho_2\rho_3(\rho_1+\rho_2+\rho_3) = \rho_1\rho_2\rho_3,$$

since $\rho_1+\rho_2+\rho_3 = 1$. By equation (3.16):

$$\det(\mathbf{J}) = \frac{e^{\theta_1}}{D}\cdot\frac{e^{\theta_2}}{D}\cdot\frac{1}{D} = \frac{e^{\theta_1+\theta_2}}{D^3}.$$

Therefore:

$$\log\left(\det(\mathbf{J})\right) = \theta_1+\theta_2-3\log\left(1+e^{\theta_1}+e^{\theta_2}\right),$$

which is equivalent to (3.15).

The optimization includes the following calculations:

Step 1. Compute the log marginal likelihood $L(\mathbf{Y}|M_i)$ in (3.13) using the bridge sampling. Using the method of Gronau et al. (2017), we treat the integrand in (3.13) as the posterior distribution and use a multivariate normal distribution as the proposal distribution. Samples of $\boldsymbol{\theta}$ are generated from the multivariate normal distribution and then transformed to $\boldsymbol{\rho}$. We estimated the marginal likelihood of each candidate model using an iterative bridge sampling algorithm in Meng and Wong (1996). In Example 1, we use 20000 iterations, 5000 burn-in samples, a thinning interval of 5 and a single chain.

Step 2. Compute $P(M_i|\mathbf{Y})$ using (3.15) and $L(\mathbf{Y}|M_i)$ obtained in *Step 1* for each model. Then select the candidate models that account for 95% of the cumulative posterior probability.

Step 3. For each candidate model M_i and a regressor $\mathbf{x}^*$, compute $\Pr(L \leq Y^* \leq U|M_i, \mathbf{x}^*, \mathbf{Y})$. Now

$$\Pr(L \leq Y^* \leq U|M_i, \mathbf{x}^*, \mathbf{Y}) = \int_{\boldsymbol{\rho}} \Pr(L \leq Y^* \leq U|M_i, \mathbf{x}^*, \mathbf{Y}, \boldsymbol{\rho}) p(\boldsymbol{\rho}) d\boldsymbol{\rho},$$

where $\Pr(L \leq Y^* \leq U|M_i, \mathbf{x}^*, \mathbf{Y}, \boldsymbol{\rho})$ can be calculated using (3.12). Samples obtained in *Step 1* using MCMC are used to estimate the integral.

Step 4. Compute the model-averaged probability $\Pr(L \leq Y^* \leq U|\mathbf{x}^*, \mathbf{Y})$ using (3.3) based on the posterior model probabilities and the probabilities obtained in *Step 2* and *Step 3*, respectively.

Step 5. To find the factor levels that maximize $\Pr(L \leq Y^* \leq U|\mathbf{x}^*, \mathbf{Y})$, (a) A level combination of hard-to-change factors is selected first; (b) 2000 level combinations of easy-to-change factors are randomly selected within the range of the factors. Compute $\Pr(L \leq Y^* \leq U|\mathbf{x}^*, \mathbf{Y})$ for each combination of factor levels. For the contour plots in Figure 3.1 and the simulation study in Section 3.6.2.1, we consider 1000 level combinations to reduce computational cost; (c) Repeat *Step (a)* and *Step (b)* until all the level combinations of hard-to-change factors are considered; (d) The factor levels that maximize $\Pr(L \leq Y^* \leq U|\mathbf{x}^*, \mathbf{Y})$ are considered as the optimal factor levels.

3.6 Example: Staggered-level design with simulated data

Consider the 28-run I-optimal staggered-level design constructed by Arnouts and Goos (2015) in Table 4. It consists of two hard-to-change factors, w and s, and two easy-to-change factors, t_1 and t_2. w is the class-1 hard-to-change factor that is most difficult to reset and s is the class-2 hard-to change factor that is less difficult to reset.

The data in Table 1 is generated from the multivariate normal distribution with $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = 1$ and the mean

$$E(Y) = 80 + 11.38w + 4.20s + 12.30t_1 - 11.35t_2 + 2.64w^2 - 3.30st_1.$$

Since w and s are hard-to-change factors, we assume that they have three levels $-1, 0$, and 1, while t_1 and t_2 are considered quantitative variables which can be set to any values between -1 and 1.

Since the staggered-level design is a three-stratum design, $s = 3$. The indicator matrices U_1 and U_2 in the model (3.1) assign each run to the corresponding

whole-plot and subplot unit, respectively. In this example,

$$\mathbf{U}_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \ldots & & & & & & \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}_{28\times 7}$$

and

$$\mathbf{U}_2 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ \ldots & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}_{28\times 8}$$

Assume that the experimenter would like to find the factor values which maximize the probability that the new response Y^* is between 74.4 (30% quantile) and 81.6 (45% quantile), that is, maximize $\Pr(74.4 < Y^* < 81.6|\mathbf{x}^*, \mathbf{Y})$. To conduct the optimization, three parameters, τ, π, and α, need to be chosen for model averaging. Since it is suggested that τ should be greater than or equal to $3\sigma_y$ (Lin, 2018) and the standard deviation σ_y of the data is 3.13, we choose $\tau = 10$. The parameter π in 2.1 is assumed to be 0.5, that is, the probability that a factor is active is $\pi = 0.5$. The hyperparameter α in 3.13 is set to $(5,5,5)$. Sensitivity to the choices of the parameters will be investigated by considering alternative values and examining their impact on the model-averaged posterior probability z later in the example.

w	s	t_1	t_2	Y	w	s	t_1	t_2	Y
0	0	0	0	81.51	1	0	1	1	95.07
0	0	1	−1	104.87	1	0	−1	−1	92.75
0	1	0	−1	95.89	0	0	−1	0	69.15
0	1	1	1	83.40	0	0	0	0	80.30
1	1	1	−1	118.80	0	1	−1	1	64.40
1	1	−1	0	88.80	0	1	1	0	92.57
1	0	0	−1	106.36	−1	1	−1	−1	79.06
1	0	0	1	82.14	−1	1	0	1	64.41
−1	0	−1	1	47.65	−1	−1	−1	0	52.08
−1	0	1	0	83.65	−1	−1	1	−1	95.21
−1	−1	1	1	68.67	0	−1	−1	−1	74.19
−1	−1	0	−1	76.22	0	−1	0	1	67.61
1	−1	−1	1	61.93	0	0	0	0	82.17
1	−1	1	0	101.44	0	0	1	1	83.04

Table 3.1: The 28-run I-optimal staggered-level design (Arnouts and Goos, 2015) the simulated data.

3.6.1 Selection of optimal factor levels by individual models

First we find the factor levels that maximize $\Pr(74.4 < Y^* < 81.6 | M_i, \mathbf{x}^*, \mathbf{Y})$ based on individual models. For a given model, the data and an observation, Equation (3.12) can be used to compute the probability.

The following three models are considered:

1. The full quadratic model: We obtain that when the factor levels are $w = 0, s = 0, t_1 = 0.4208$, and $t_2 = 0.7473$, the probability reaches the maximum value of 0.8865.

2. The model includes the effects w, s, t_1, t_2, w^2, and st_1: At the factor levels of $w = 0, s = 0, t_1 = -0.2811$, and $t_2 = -0.0312$, the maximum probability obtained is 0.8602.

3. The model consists of w, s, t_1 and t_2: The maximum probability achieved is 0.6493 at the factor levels $w = 0, s = 0, t_1 = -0.0972$, and $t_2 = 0.2394$.

The results indicate that variability in prediction probabilities across different models can lead to unstable prediction results. Averaging the predictive density of the response over all competing models can help reduce variability in prediction results and provide a more reliable estimate of the future response.

3.6.2 Selection of optimal factor levels by Bayesian model averaging

In this experiment, there are four factors, resulting in $2^{14} = 16384$ models. First, we select the model using the strong effect heredity principle:

(1) if an interaction term is in the model, then the corresponding main effects must be included in the model;
(2) if a quadratic effect is in the model, then the corresponding main effect must be included in the model.
It yields 1337 candidate models. After calculating the posterior probabilities of the 1337 models using Equation (3.15), the top 9 models that account for 95% of the posterior probability are chosen as competing models.

Table 3.2 lists the 9 competing models and the posterior probability for each model is given in column 3. The first model has a substantially larger posterior probability (0.7491) than the others. The true model is ranked second with posterior probability 0.0926, while the remaining models have probabilities between 0.0122 and 0.0431. Notably, all models include the terms w, s, t_1, t_2, and st_1, which are present in the true model.

No.	Model	$P(M_i\|\mathbf{Y})$
1	w, s, t_1, t_2, st_1	0.7491
2	$w, s, t_1, t_2, w^2, st_1$	0.0926
3	$w, s, t_1, t_2, s^2, st_1$	0.0431
4	$w, s, t_1, t_2, t_2^2, st_1$	0.0389
5	$w, s, t_1, t_2, t_1^2, st_1$	0.0194
6	w, s, t_1, t_2, ws, st_1	0.0191
7	$w, s, t_1, t_2, st_1, t_1t_2$	0.0132
8	$w, s, t_1, t_2, wt_2, st_1$	0.0124
9	$w, s, t_1, t_2, wt_1, st_1$	0.0122

Table 3.2: The posterior probabilities for the 9 competing models.

Using the Bayesian model averaging approach based on the 9 competing models, the optimal factor levels that maximize $\Pr(74.4 < Y^* < 81.6|\mathbf{x}^*, \mathbf{Y})$ are

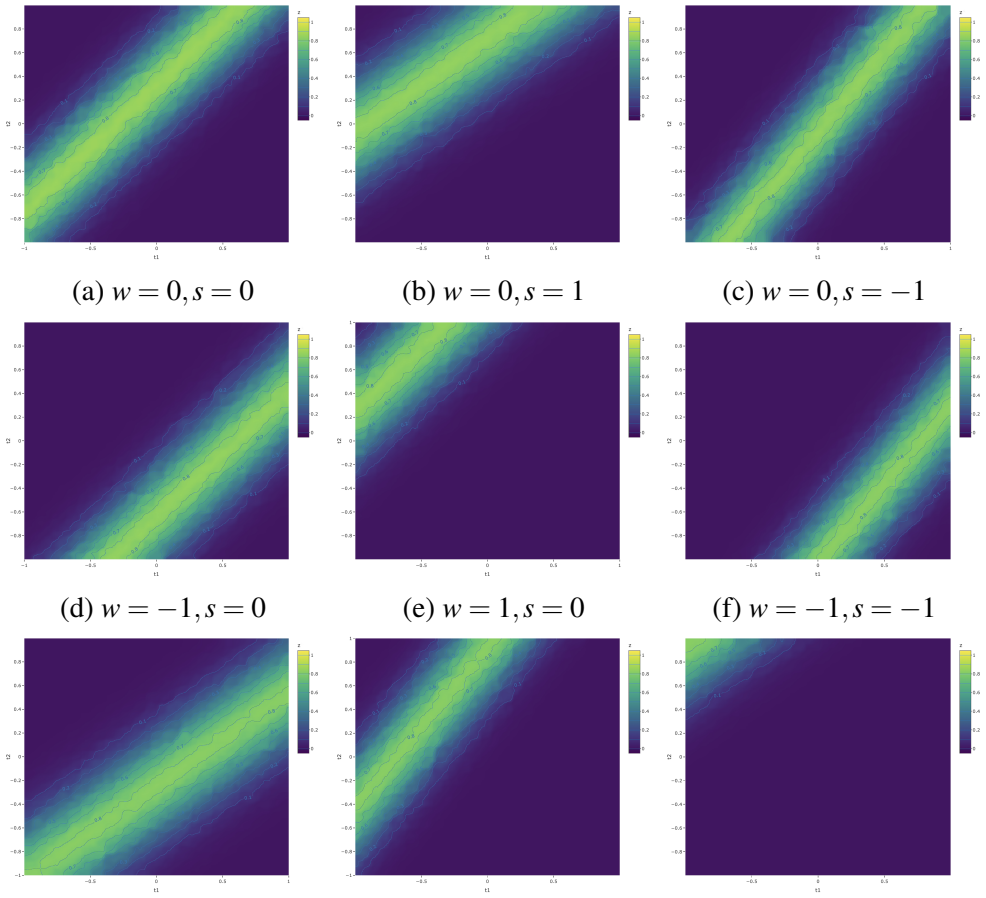
found to be $w = 0$, $s = 0$, $t_1 = 0.0571$, and $t_2 = 0.4003$ with the maximum model-averaged posterior probability 0.8595.

Since w and s are hard-to-change factors, we are interested in the optimal factor levels of t_1 and t_2 for given w and s. Table 3.3 presents the optimal factor levels of t_1 and t_2 for each level combination (w,s). The z^* in the last column represents the maximum model-averaged posterior probability $\Pr(74.4 < Y^* < 81.6|\mathbf{x}^*, \mathbf{Y})$. Needless to say, the probability achieves the maximum 0.8595 when the levels of both hard-to-change factors are 0, that is, $(w,s) = (0,0)$, which coincides with the result obtained in Section 3.6.1, that is, the maximum probability occurs at $(w,s) = (0,0)$. The second maximum probability 0.8417 is obtained when (w,s) is fixed at $(0,1)$. In this case, the optimal levels of t_1 and t_2 are -0.1086 and 0.6300, respectively. The smallest maximum model-averaged posterior probability 0.8171 happens when $(w,s) = (1,1)$ and $t_1 = -0.9188$ and $t_2 = 0.9963$.

Table 3.3 provides important information for optimizing a response under different levels of hard-to-change factors. By carefully examining the optimal factor levels and their associated posterior probabilities, we can gain a better understanding of how to improve the system's performance and maximize its potential for success.

w	s	t_1	t_2	z^*
0	0	0.0571	0.4003	0.8595
0	1	−0.1086	0.6300	0.8417
0	−1	−0.1547	−0.2751	0.8416
−1	0	0.1696	−0.4454	0.8403
1	0	−0.4783	0.8267	0.8383
−1	−1	0.4323	−0.4579	0.8248
−1	1	−0.0046	−0.2444	0.8215
1	−1	−0.2839	0.5450	0.8211
1	1	−0.9188	0.9963	0.8171

Table 3.3: Optimal levels of t_1 and t_2 for each level combination (w, s).



(a) $w = 0, s = 0$ (b) $w = 0, s = 1$ (c) $w = 0, s = -1$

(d) $w = -1, s = 0$ (e) $w = 1, s = 0$ (f) $w = -1, s = -1$

(g) $w = -1, s = 1$ (h) $w = 1, s = -1$ (i) $w = 1, s = 1$

Figure 3.1: Contour plots of the model-averaged posterior probability z for each level combination (w, s).

Figure 3.1 displays the contour plots of the model-averaged posterior probability z for each level combination (w, s). Higher probabilities are distributed along a linear area due to the linear regression model employed in our simulation. In particular, higher values of z (specifically, highlighted in yellow) in plot (a), where $(w, s) = (0, 0)$, are much more prevalent compared to other instances. The maximum model-averaged posterior probability z^* also reaches its peak in plot (a), as shown in Table 3.3. Table 3.3 lists the maximum model-averaged posterior

probability z^* values, yet Figure 3.1 reveals a cluster of neighbouring data points achieving similarly high performance. Similar performance of z can also be seen in most of the other plots. When $(w,s) = (1,1)$, the contour plot in (i) indicates that z is very rare to reach a good performance, which agrees with the conclusion in Table 3.3 that the probability reaches maximum when t_1 is close to -1 and t_2 is close to 1.

Our results show that the optimal factor levels with $(w,s) = (0,0)$ has a higher probability to reach large model-averaged posterior probability than other combinations. This result suggests that this particular combination is likely to produce the desired outcome and should be given careful consideration when optimizing the system's performance.

3.6.2.1 Simulation Study

We conduct a simulation study to evaluate the performance of the Bayesian model averaging approach against the AIC and BIC for optimizing the probability that the predicted response is between 74.4 and 81.61. The top model in Table 3.2 is fitted using the data in Table 3.1, and the fitted model is treated as the true model. The simulation study proceeds as follows for each replicate. First, we generate the response data from the true model using the I-optimal staggered-level design in Table 3.1. Second, we apply the MCMC sampling procedure in Section 3.5 to estimate marginal likelihoods and posterior model probabilities. Third, we compute AIC and BIC by maximizing the likelihood for each candidate model, allowing us to select the best single model under these criteria. After selecting a model by AIC or BIC, we determine the factor levels that maximize

the probability that the predicted response is between 74.4 and 81.61 under the selected model. For each simulated data set, we then evaluate the probability at the factor levels selected by Bayesian model averaging, AIC, and BIC. The corresponding maximum probability (denoted as "True optimum") under the true model is also calculated as a benchmark. The resulting probabilities across the 150 replicates are shown in Figure 3.2.



Figure 3.2: Comparison of Bayesian model averaging, AIC, and BIC over 150 simulation replicates.

Among the three methods, Figure 3.2 indicates that Bayesian model averaging performs closest to the true optimum overall. Its probabilities are typically high and relatively stable across replicates, with fewer large drops. In contrast, AIC and BIC tend to yield smaller probabilities and show greater variability, including sharp declines, such as the replicates 72 and 131. These declines indicate that AIC and BIC can sometimes choose a model that leads to poor factor levels, resulting in a significantly lower probability in that replicate.

The boxplot in Figure 3.3 summarizes the distribution of the maximum probability that the predicted response is between 74.4 and 81.61 across 150

simulation replicates for the three methods Bayesian model averaging, AIC, and BIC. The true optimum obtained from the true model is also plotted for comparison. Its distribution is tightly concentrated at a high value (or close to 1), with almost no spread, reflecting the maximum achievable probability when the true model is known. The Bayesian model averaging box is centered close to the true optimum, with a relatively high median and a comparatively small interquartile range. On the other hand, both AIC and BIC show larger spread with longer lower whiskers and multiple low outliers. Although Bayesian model averaging has some low outliers, they are less extreme and less frequent than those of AIC and BIC. The results of Figure 3.2 and Figure 3.3 indicates that Bayesian model averaging produces consistently high probabilities and is more robust to sampling variability, whereas AIC and BIC can perform well in many replicates but may occasionally exhibit severe underperformance.

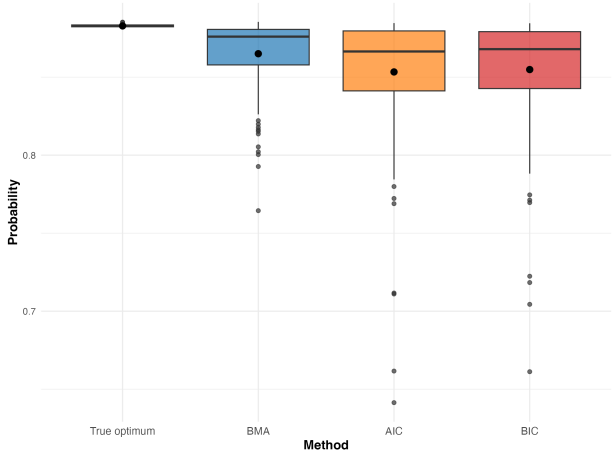


Figure 3.3: Boxplot comparison of Bayesian model averaging, AIC, and BIC over 150 simulation replicates.

3.6.3 Sensitivity analysis of parameters

We conduct the sensitivity analysis for the parameters π, τ, and $\boldsymbol{\alpha}$ and highlights the impact of the parameters on the optimal factor levels and the corresponding maximum model-averaged posterior probability z^*. Table 3.4 lists the optimal factor levels and the corresponding z^* values for $\pi = 0.5$, $\boldsymbol{\alpha} = (5,5,5)$, and $\tau = 1,5,10,15,20$. It shows that the optimal levels for w and s are 0 for all choices of τ, indicating that the operating condition for hard-to-change factors is robust to τ. z^* increases from 0.0982 at $\tau = 1$ to 0.6240 at $\tau = 5$, and then to 0.8595, 0.9347, and 0.9612 at $\tau = 10$, 15, and 20, respectively. Meanwhile, the optimal easy-to-change factor levels change noticeably from $\tau = 1$ to $\tau = 5$, but essentially converge for $\tau \geq 10$ with $t_1 = 0.0571$ and $t_2 = 0.4003$. In summary, τ has a strong effect on z^*, but the optimal factor levels become stable once τ is greater than 10.

τ	w	s	t_1	t_2	z^*
1	0	0	0.4942	−0.2429	0.0982
5	0	0	0.1111	0.4056	0.6240
10	0	0	0.0571	0.4003	0.8595
15	0	0	0.0571	0.4003	0.9347
20	0	0	0.0571	0.4003	0.9612

Table 3.4: Sensitivity analysis for τ.

The z^* values for $\pi = 0.25, 0.5$, and 0.75, with $\boldsymbol{\alpha} = (5,5,5)$ and $\tau = 10$, are listed in Table 3.5. The results again show that the optimal levels for the hard-to-change factors w and s are 0 for all choices of π. The optimal levels for the easy-to-change factors t_1 and t_2 are the same when $\pi = 0.25$ and $\pi = 0.5$,

with corresponding z^* values of 0.8637 and 0.8595, respectively, which are very close. In contrast, when $\pi = 0.75$, the optimal levels for t_1 and t_2 differ, and the resulting z^* value decreases to 0.8367. Thus, the optimal factor levels for the easy-to-change factors are fairly stable for moderate values of π, but may shift when π strongly favors larger models. Since z^* decreases as π increases, the model-averaged posterior probability is sensitive to the choice of π.

π	w	s	t_1	t_2	z^*
0.25	0	0	0.0571	0.4003	0.8637
0.50	0	0	0.0571	0.4003	0.8595
0.75	0	0	−0.3380	−0.0422	0.8367

Table 3.5: Sensitivity analysis for π.

Table 3.6 shows how the optimal factor settings and the z^* values change as $\boldsymbol{\alpha}$ varies, assuming $\pi = 0.5$ and $\tau = 10$. For every $\boldsymbol{\alpha}$, the optimal factor levels for hard-to-change factors remain $w = 0$ and $s = 0$. It indicates the whole-plot factor levels are robust to $\boldsymbol{\alpha}$. For easy-to-change factors, when $\boldsymbol{\alpha} = (1,1,1), (5,5,5)$, and $(10,10,10)$, the optimal factor levels are $(t_1, t_2) = (0.0571, 0.4003)$. When $\boldsymbol{\alpha}$ increases to $(15,15,15)$ and $(20,20,20)$, the optimum changes to $(t_1, t_2) = (-0.0102, 0.3173)$. Thus, $\boldsymbol{\alpha}$ has little effect on the easy-to-change factor levels until it becomes relatively large. The value of z^* increases from 0.8204 ($\boldsymbol{\alpha} = (1,1,1)$) to 0.8595 ($\boldsymbol{\alpha} = (5,5,5)$) and 0.8798 ($\boldsymbol{\alpha} = (10,10,10)$), then increases only slightly to 0.8841 and 0.8856 at $\boldsymbol{\alpha} = (15,15,15)$ and (20, 20, 20), respectively. This suggests that increasing $\boldsymbol{\alpha}$ beyond about 10 produces only small gains in z^*.

$\boldsymbol{\alpha}$	w	s	t_1	t_2	z^*
(1, 1, 1)	0	0	0.0571	0.4003	0.8204
(5, 5, 5)	0	0	0.0571	0.4003	0.8595
(10, 10, 10)	0	0	0.0571	0.4003	0.8798
(15, 15, 15)	0	0	−0.0102	0.3173	0.8841
(20, 20, 20)	0	0	−0.0102	0.3173	0.8856

Table 3.6: Sensitivity analysis for $\boldsymbol{\alpha}$.

In summary, the sensitivity analysis results indicate that the optimal hard-to-change settings remain highly robust across all values of τ, π, and $\boldsymbol{\alpha}$, whereas z^* exhibits greater sensitivity, especially with respect to τ.

In our program, convergence is assessed by acceptance rates (30%-40%) and effective sample sizes (1500-2500) from a 2-dimensional random walk Metropolis-Hastings sampler, and scalability is good because the MCMC state space remains fixed at two dimensions regardless of model complexity.

Chapter 4

Response Optimization on Data with a Multistratum Structure Under Non-normal Assumptions of Random Effects and Errors

4.1 Introduction

In practice, there is usually enough data to check the normality of random errors ϵ. However, since the number of whole plots or blocks in each stratum is typically small, it is difficult to assess the normality of the random stratum effects γ_l. Additionally, γ_l and ϵ may not follow normal distributions. In this chapter, we investigate the response optimization when the random effects and errors do not follow normal distributions. In such case, Theorem 3.3.1 does not hold, and

Equation (3.12) cannot be used to compute the cumulative posterior predictive density of Y^*. In Chapter 3, we define $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_s)$ so that we can derive the posterior predictive distribution under normal assumptions of random effects and errors. In this chapter, since random effects and errors are not normally distributed, the posterior distribution may not be conjugate. For simplicity, we consider the variances of $\boldsymbol{\gamma}_l$ and $\boldsymbol{\epsilon}$ directly, rather than modeling $\boldsymbol{\rho}$.

We will use Stan to fit the model. A Stan model is a program to define a statistical model through its log density or a Bayesian posterior distribution. It is widely used for fitting complex Bayesian models and generating posterior predictive quantities. In a Stan program, the unknown quantities to be estimated are specified and the likelihood and prior distributions are declared. Stan fits the model using gradient-based inference methods, most commonly Hamitonian Monte Carlo. After fitting the model in Stan, we use bridge sampling to estimate the marginal likelihood. The Bayesian model averaging approach will be used to find the optimal factor levels that maximize the probability that a future response falls within a targeted range.

Section 4.2 starts with a practical example. The corresponding model and prior assumptions are described. Section 4.3 details the optimization procedure implemented in Stan step by step. Model selection results are presented in Section 4.4. Section 4.5 conducts the response optimization using the Bayesian model averaging. Graphical analysis is performed for the model-averaged posterior probability. A model-robustness analysis in Section 4.6 evaluates the performance of Bayesian model averaging relative to individual models, and Section 4.7 provides a simulation study comparing Bayesian model averaging

with AIC and BIC. Sensitivity analysis for the prior parameter π is conducted.

4.2 A motivating example and prior assumptions

A study was conducted by J. M. Pardo at the University of Reading, UK, to investigate the impact of various process factors on the retention of volatile compounds during the freeze-drying of coffee (Gilmour and Goos, 2009). It involves five factors: the pressure in the drying chamber x_1, the heating temperature x_2, the initial solids content in the coffee solution x_3, the thickness of the slab freeze-dried as a batch x_4, and the freezing rate x_5, respectively. Due to time constraints, only 30 experiments could be conducted. The pressure in the drying chamber was considered as the whole-plot factor since it was controlled manually and was hard to change the settings. Thus, the study utilized a 30-run split-plot design with one whole-plot factor x_1 and four subplot factors x_2, x_3, x_4, and x_5.

Table 4.1 lists the split-plot design and the data, where the first column is the whole plot number (wp) and the last column is the response, the drying rate Y. The objective is to find the optimal factor levels that results in the maximum probability that the drying rate is within given bounds. Suppose that the experimenter is interested in having the drying rate between 60 and 85 and would like to find the factor levels that maximize $\Pr(60 < Y^* < 85|\mathbf{x}^*, \mathbf{Y})$. Since x_1 is a whole-plot factor, we consider three possible levels $-1, 0$, and 1. The subplot factors are considered as quantitative variables ranging from -1 to 1.

wp	x_1	x_2	x_3	x_4	x_5	Y	wp	x_1	x_2	x_3	x_4	x_5	Y
1	1	0	0	0	1	66.000	4	1	0	0	−1	0	66.934
1	1	0	0	1	0	66.094	4	1	1	0	0	0	79.220
1	1	−1	0	0	0	57.848	4	1	0	0	0	−1	65.203
1	1	0	0	0	0	66.000	4	1	0	−1	0	0	73.835
1	1	0	1	0	0	51.871	4	1	0	0	0	0	67.941
2	0	0	0	0	0	70.884	5	−1	0	0	0	0	69.184
2	0	−1	1	−1	1	56.763	5	−1	1	1	−1	1	85.379
2	0	1	1	1	−1	62.423	5	−1	1	−1	1	−1	74.300
2	0	1	−1	−1	−1	83.570	5	−1	−1	1	1	−1	50.360
2	0	−1	−1	1	1	65.191	5	−1	−1	−1	−1	1	60.266
3	−1	0	0	0	0	71.379	6	0	1	−1	1	1	89.160
3	−1	1	1	1	1	97.931	6	0	0	0	0	0	68.500
3	−1	−1	1	−1	−1	54.947	6	0	1	1	−1	−1	75.570
3	−1	−1	−1	1	−1	61.704	6	0	−1	1	1	1	56.470
3	−1	1	−1	−1	1	80.410	6	0	−1	−1	−1	−1	68.388

Table 4.1: The design and the data for the freeze-drying experiment.

Since the experiment used a split-plot design, the model (3.1) is simplified as

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

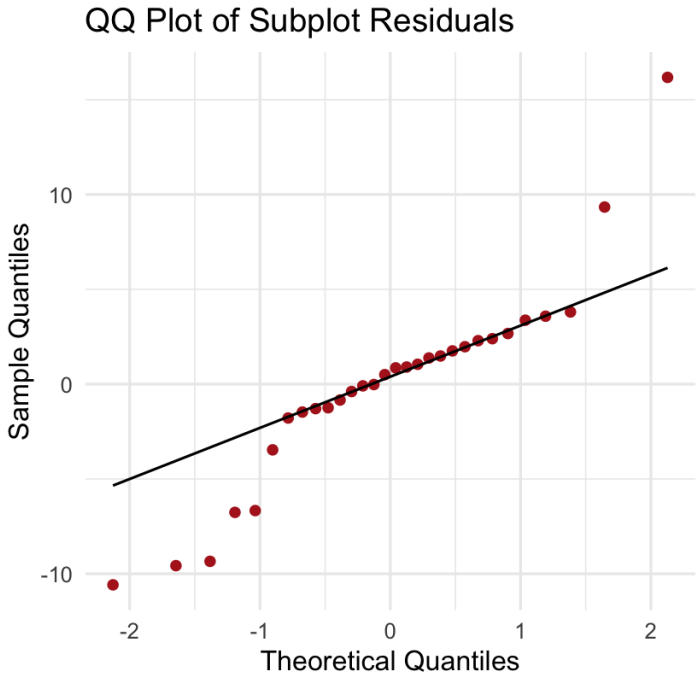
where $\mathbf{Y}$ is 30×1 vector of the responses, $\mathbf{X}$ is the model matrix, $\boldsymbol{\beta}$ is the $(p+1) \times 1$ vector of effects of the model terms, p is the number of terms in the model excluding the constant term, $\mathbf{U}$ is the $n \times 6$ indicator matrix for $\boldsymbol{\gamma}$, $\boldsymbol{\gamma}$ is the 6×1 vector of the random effects for the whole-plot stratum, and $\boldsymbol{\epsilon}$ is the 30×1 vector of random subplot errors.

From the design description, the 30 runs are divided equally among the 6 whole plots (5 runs per whole plot). Therefore $\mathbf{U}_1$ is a 30×6 block-diagonal indicator matrix as:

$$\mathbf{U}_1 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \\ \ldots & & & & & \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}_{30\times 6}$$



(a)



(b)

Figure 4.1: QQ plots of whole plot and subplot errors.

Figure 4.1 displays the quantile-quantile (QQ) plots for both the whole-plot errors and subplot errors, which are used to assess the normality assumption of the model residuals. The plots reveal significant deviations from the expected straight line, indicating that neither the whole-plot nor the subplot errors follow a normal distribution. Thus, the method in Chapter 3 cannot be applied to this example.

Following the suggestion by Gelman et al. (2013), Peterson (2023) proposed to model the random stratum effects with a heavy-tailed t distribution with 4 degrees of freedom, reasoning that this approach can accommodate occasional large deviations when the true underlying random effects exhibit heavy tails, while remaining conservative otherwise. To accommodate possible outlying values in the whole-plot errors, we assume that whole-plot errors follow a t distribution with mean 0 and 3 degrees of freedom. In addition, Figure 4.1(b) indicates that the subplot errors follow a heavy-tailed distribution. Following Peterson (2023), we assume that subplot errors follow a t distribution with mean 0 and ν degrees of freedom, where ν controls the heaviness of the tails. We assume that ν is an unknown model parameter and we estimate it from the data. Peterson (2023) used the half-t prior with 3 degrees of freedom for variance components following the suggestion by Chatfield (2018). We adopt the same assumption for the variance components of γ and ϵ. In addition, a weakly informative prior $\beta \sim N(0, 10)$ is applied, allowing the data to dominate and providing enough regulation to stabilize computation.

4.3 Optimization procedure

Under the assumption that the random effects follow a t distribution, it is difficult to derive the posterior predictive distribution of the response. Our optimization includes the following steps:

Step 1. Implement and fit the Stan models using the R package `rstan`. We assign a half-t prior with 3 degrees of freedom and scale parameter 2.5 (Peterson, 2023) to the variance components of γ and ϵ, and a gamma prior to ν with parameters 2 and 10.

Step 2. Compute $L(\mathbf{Y}|M_i)$ for a given model M_i. We employ the R package `bridgesampling` to compute the marginal likelihood using posterior samples obtained from Stan. Due to the small run size in Example 2, we use 30,000 iterations, 20,000 burn-in samples, and 4 chains.

Step 3. Compute $P(M_i|\mathbf{Y})$ using a prior model probability and the $L(\mathbf{Y}|M_i)$ obtained from *Step 2* and select the candidate models.

Step 4. Compute $\Pr(L \leq Y^* \leq U|M_i, \mathbf{x}^*, \mathbf{Y})$ for each candidate models M_i using the Stan samples obtained in *Step 1*.

Step 5. After model selection, the Bayesian model averaging approach is employed to identify the optimal factor levels that maximize $\Pr(L \leq Y^* \leq U|\mathbf{x}^*, \mathbf{Y})$. For a given $\mathbf{x}^*$, the model-averaged posterior probability can be obtained using (3.3). To find the factor levels that maximize this probability, we proceed as follows: (a) Select a level of x_1 (the hard-to-change factor); (b) Compute $\Pr(L \leq Y^* \leq U|\mathbf{x}^*, \mathbf{Y})$ for each combination of factor levels. For each model, $\gamma = (\gamma_1, \gamma_2, \gamma_3, \gamma_4, \gamma_5, \gamma_6)$ is estimated in Stan. For each level of x_1, there

are two estimates corresponding to the two whole plots in the design in Table 4.1. For example, when $x_1 = 1$, there are two estimates, γ_1 and γ_4, corresponding to whole plot 1 and whole plot 4, respectively. To ensure that level combinations of the easy-to-change factors are sampled equally from each whole plot, we generate 600 level combinations within the factor ranges using Latin Hypercube Sampling, with 100 combinations drawn from each whole plot. For the contour plots in Figure 4.2, we generate 1000 combinations from each whole plot to produce clearer graphs; (c) Repeat (a) and (b) until all the levels of x_1 have been considered; (d) The factor levels that maximize $\Pr(L \leq Y^* \leq U|\mathbf{x}^*, \mathbf{Y})$ are considered as the optimal settings.

4.4 Model selection

Since this experiment involves 5 factors, there are 2^{20} possible models. After using the strong effect heredity principle, 38619 models were left. Then, the posterior probabilities of these models are calculated. Table 4.2 provides the top 10 selected models with their posterior probabilities $P(M_i|\mathbf{Y})$ when $\pi = 0.5$. The posterior probability of the highest-ranked model is only 0.0020. Moreover, 3,589 models are needed to accumulate 50% of the posterior model probability. Together, these results indicate an extremely diffuse posterior over the model space, with no single model decisively favoured by the data. Such behaviour may be attributable to the t distributed errors. The heavy-tailed likelihood down weights extreme residuals, therefore reducing sensitivity to moderate model misspecification (O'Hagan, 1979; Gelman et al., 2013). As a result,

posterior mass may be spread across many nearly equivalent models rather than concentrating on a single dominant model (Hoeting et al., 1999).

No.	Model	$P(M_i\|\mathbf{Y})$
1	$x_1, x_2, x_3, x_4, x_5, x_1^2, x_3^2, x_4^2, x_5^2, x_1x_2, x_1x_3, x_1x_4, x_3x_5$	0.0020
2	$x_1, x_2, x_3, x_4, x_5, x_1^2, x_2^2, x_3^2, x_4^2, x_5^2, x_1x_2, x_1x_3, x_1x_4, x_3x_5$	0.0017
3	$x_1, x_3, x_4, x_5, x_1^2, x_3^2, x_4^2, x_5^2, x_1x_3, x_1x_4, x_3x_5$	0.0016
4	$x_1, x_2, x_3, x_4, x_5, x_1^2, x_2^2, x_3^2, x_4^2, x_5^2, x_1x_2, x_1x_3, x_1x_4, x_2x_3, x_3x_5$	0.0015
5	$x_1, x_3, x_4, x_5, x_1^2, x_4^2, x_5^2, x_1x_3, x_1x_4, x_3x_5$	0.0014
6	$x_1, x_3, x_4, x_5, x_1^2, x_3^2, x_4^2, x_5^2, x_1x_3, x_1x_4, x_1x_5, x_3x_5$	0.0014
7	$x_1, x_2, x_3, x_4, x_5, x_1^2, x_2^2, x_3^2, x_4^2, x_5^2, x_1x_2, x_1x_3, x_1x_4, x_1x_5, x_3x_5$	0.0014
8	$x_1, x_2, x_3, x_4, x_5, x_1^2, x_2^2, x_3^2, x_4^2, x_5^2, x_1x_2, x_1x_3, x_1x_4, x_3x_4, x_3x_5$	0.0013
9	$x_1, x_2, x_3, x_4, x_5, x_1^2, x_3^2, x_4^2, x_5^2, x_1x_2, x_1x_3, x_1x_4, x_2x_3, x_3x_5$	0.0013
10	$x_1, x_2, x_3, x_4, x_5, x_1^2, x_2^2, x_4^2, x_5^2, x_1x_2, x_1x_3, x_1x_4, x_3x_5$	0.0013

Table 4.2: Top 10 models and their posterior probabilities.

Due to computational constraints, we cannot include too many models in the model-averaging procedure. To determine an appropriate number of models, we computed the model-averaged posterior probability using the top 100, 200, and 500 models, obtaining values of 0.7971, 0.7946, and 0.7978, respectively. The similarity of the results suggests that using 100 models is good enough for model averaging. Thus, we use the top 100 candidate models in this example, except in Section 4.7, where the model space is small and we instead use the set of top models that account for 95% of the posterior probability.

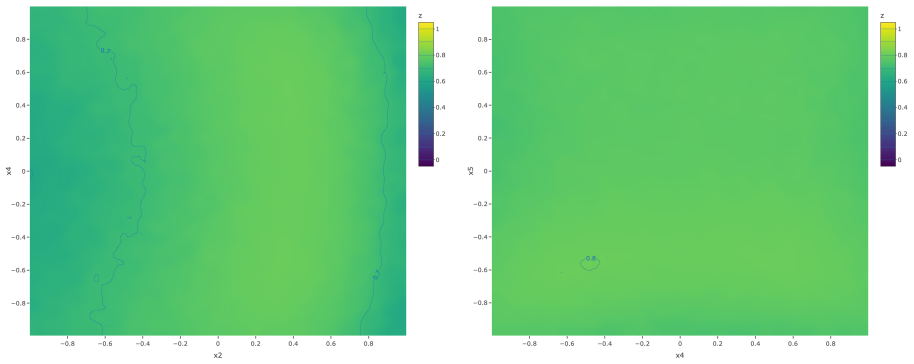
4.5 Optimal factor levels

We are interested in finding the optimal factor levels that result in the maximum probability that a predicted response falls between 60 and 85. By maximizing

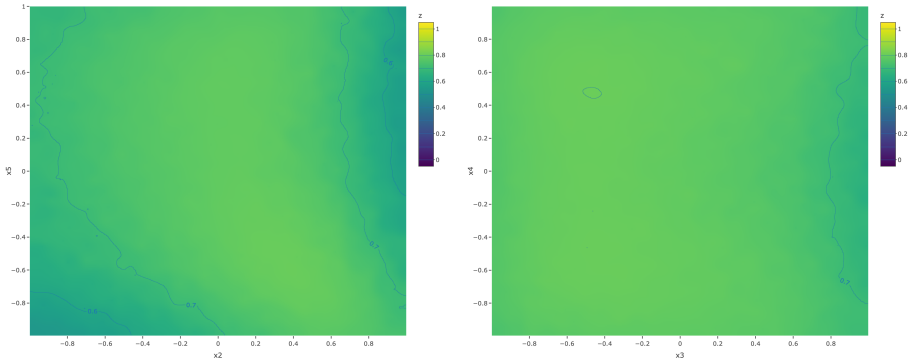
the probability in (3.3), we found that the optimal factor levels are

$$(x_1, x_2, x_3, x_4, x_5) = (0, 0.3471, -0.4020, -0.4085, -0.5686) \tag{4.1}$$

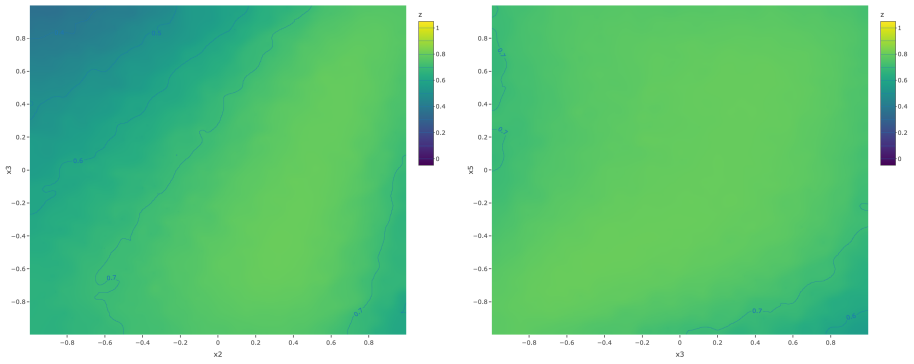
with the maximum model-averaged posterior probability $z^* = 0.7971$.



(a) $x_1 = 0, x_3 = -0.4020, x_5 = -0.5686$ (b) $x_1 = 0, x_2 = 0.3471, x_3 = -0.4020$



(c) $x_1 = 0, x_3 = -0.4020, x_4 = -0.4085$ (d) $x_1 = 0, x_2 = 0.3471, x_5 = -0.5686$



(e) $x_1 = 0, x_4 = -0.4085, x_5 = -0.5686$ (f) $x_1 = 0, x_2 = 0.3471, x_4 = -0.4085$

Figure 4.2: Contour plots of the model-averaged posterior probability z.

Figure 4.2 provides the contour plots of the model-averaged posterior probability z when the whole-plot factor is at its optimal level 0 and two of the four subplot factors are fixed at their optimal levels as specified in (4.1). The value of

z is relatively high when the colour of the area is close to yellow. For example, Figure 4.2(a) shows the contour plot when $x_1 = 0, x_3 = -0.4020, x_5 = -0.5686$ and x_2 and x_4 vary between -1 and 1. The plot indicates that when x_1, x_3, and x_5 are at their optimal levels, the maximum value of z approaches 0.8 when x_2 ranges approximately from 0 to 0.6, for all values of x_4, as indicated by the yellow region. Across all graphs, the surfaces are largely flat, with z remaining around $0.7 - 0.8$ over most of the design region, indicating a broad plateau and robustness of the optimized factor levels to changes in the subplot factors.

The most noticeable variation occurs in graphs (a) and (c), where x_2 varies and produces the clearest gradient, while x_4 and x_5 have little effect (contours are approximately vertical). When x_2 is fixed at its optimal value (graphs (b), (d), and (f)), the surfaces are nearly uniform, suggesting that x_3, x_4, and x_5 have weak influence near the optimum. Graph (e), which varies x_2 and x_3 with x_4 and x_5 fixed at their optimized values, shows a mild diagonal pattern, suggesting a small joint effect of x_2 and x_3. Overall, x_2 is the primary tuning factor among the subplot factors, whereas the remaining subplot factors can be selected based on practical considerations with minimal impact on the model-averaged posterior probability.

4.6 Model-robustness analysis

Table 4.3: Model-robustness analysis for the top 10 selected models

	Assumed model				
True Model	1	2	3	4	5
1	0.8589	0.8491	0.7800	0.8583	0.8391
2	0.8508	0.8816	0.6556	0.8707	0.8635
3	0.6382	0.6744	0.6752	0.5986	0.6100
4	0.8424	0.8732	0.6868	0.8608	0.8429
5	0.6487	0.6790	0.5918	0.5999	0.6748
6	0.6733	0.6841	0.6294	0.6825	0.6571
7	0.8441	0.8713	0.6665	0.7757	0.8505
8	0.8410	0.8574	0.6621	0.7613	0.8422
9	0.8093	0.8607	0.7450	0.7771	0.8337
10	0.8567	0.8766	0.6867	0.7836	0.8788
Mean	0.7863	0.8107	0.6779	0.7569	0.7893
SD	0.0931	0.0913	0.0535	0.1002	0.1002

Table 4.3: Model-robustness analysis for the top 10 selected models continued

	Assumed model					
True Model	6	7	8	9	10	BMA
1	0.8321	0.8397	0.8565	0.8601	0.8514	0.8510
2	0.8264	0.8507	0.8212	0.7329	0.8705	0.8652
3	0.6781	0.6736	0.6485	0.6083	0.5993	0.6575
4	0.8518	0.8697	0.8163	0.7327	0.8038	0.8416
5	0.6824	0.6773	0.6526	0.6600	0.6585	0.6076
6	0.7047	0.6419	0.6792	0.6438	0.6531	0.6546
7	0.8516	0.8708	0.8197	0.7351	0.8081	0.8447
8	0.8516	0.8416	0.8614	0.8480	0.8069	0.8441
9	0.8296	0.8496	0.8040	0.8511	0.8210	0.8421
10	0.8289	0.8844	0.8766	0.7434	0.8819	0.8746
Mean	0.7937	0.7999	0.7836	0.7415	0.7755	0.7883
SD	0.0736	0.0951	0.0886	0.0892	0.1004	0.1038

Let $(x_1, x_2, ..., x_5)$ be the optimal factor levels obtained by maximizing the probability that the predicted response is between 60 and 85 under an assumed model.

Table 4.3 presents the probability $P(60 < Y^* < 85|M_i, x_1, ..., x_5, \mathbf{Y})$, where M_i is the true model, for different combinations of the true model and the assumed model for the top 10 models in Table 4.2. For example, if the assumed model is the model 1, then the probability is maximized at the factor levels $(0, 0.0820, -0.2580, 0.5894, 0.4991)$ with the probability 0.8589. However, if the true model is actually the model 3 and the factor levels are used, which are obtained from the wrong model 1, then the probability is 0.6382, which is much smaller. In addition, when the model 3 is the true model, the probability is maximized at the factor levels $(1, 0.7430, -0.2118, -0.2631, 0.3967)$, where the level of the hard-to-change factor x_1 is 1, which is not consistent with the optimal level of x_1 when the model 1 is assumed to be the true model. The last column presents the probability $P(60 < Y^* < 85|M_i, x_1, ..., x_5, \mathbf{Y})$ for the true model M_i at the optimal factor levels $(x_1, x_2, ..., x_5)$ obtained from the Bayesian model averaging approach, where BMA represents the Bayesian model averaging. The model-averaged choice has a higher mean probability 0.7883 than the solutions obtained by assuming most individual models. However, it also has a relatively larger standard deviation due to Model 5. Thus, the model-averaged optimal factor levels provide a strong compromise across candidate true models, achieving very good mean performance, although a poor performance can still happen occasionally.

4.7 Simulation study

We conduct a simulation study to compare the Bayesian model averaging ap-

proach with AIC and BIC. Model 5 in Table 4.2 is fitted to the experimental data in Table 4.1, and the fitted model is treated as the true model. Due to computation constraints, we assume that the four main effects x_1, x_3, x_4, and x_5 in Model 5 are active. Under the strong effect heredity principle, any quadratic or interaction term can involve only these four factors. In addition, we restrict the total number of terms in the model to be between 8 and 12, which results in 837 candidate models. Using these 837 models as the model space, we conduct model selection and select the top models that account for 95% of the posterior probability. To compute AIC and BIC, bridge sampling is used to estimate the maximum log-likelihood for each model.

First, we simulate a data set from the true model using the design in Table 4.1. Then we obtain the optimal factor levels that maximize the probability that the predicted response is between 60 and 85 using Bayesian model averaging, AIC, and BIC, as well as the true model. Finally, we evaluate the probability under the true model at the factor levels obtained by each method. The results show that the probability reaches to approximately 0.999 for all methods. This might be due to our strong restriction of the model space. Because the true model lies within this constrained space and the candidate models are relatively similar, Bayesian model averaging, AIC, and BIC tend to select models that are close enough to yield almost the same maximum probability. Another reason might be that the interval is wide. To better distinguish the three methods, we consider a narrower interval and instead maximize the probability that the predicted response is between 65 and 70 in this simulation study. With the narrower interval the objective is more sensitive, so differences between methods become visible.



Figure 4.3: Comparison of Bayesian model averaging, AIC, and BIC over 20 simulation replicates.

We simulate 20 data sets using the design in Table 4.1. The results are presented in Figure 4.3. The true model consistently yields the highest probability that the predicted response is between 65 and 70 (green curve) across replicates, as expected. Bayesian model averaging typically stays reasonably close to the true optimum, but it shows a few noticeable drops such as replicates 10 and 16, indicating that sampling variability and model uncertainty can still move the selected factor levels away from the best region. AIC shows larger fluctuations than Bayesian model averaging and includes at least one sharp decline (around replicate 10). BIC sometimes matches the true optimum very closely, but it also exhibits the most severe failure around replicate 17.

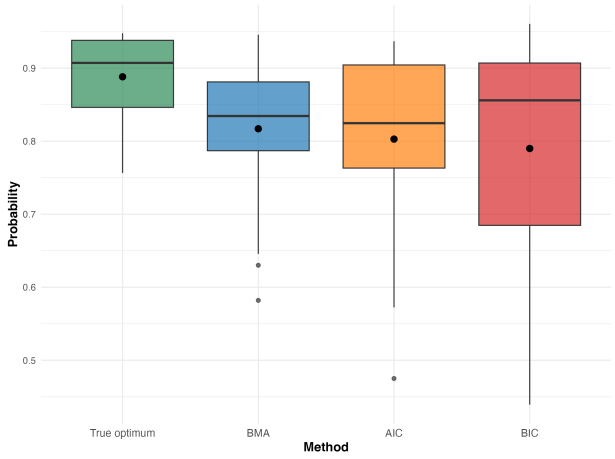


Figure 4.4: Boxplot comparison of Bayesian model averaging, AIC, and BIC over 20 simulation replicates.

Figure 4.4 is a boxplot summarizing the distribution of the maximum probability that the predicted response is between 65 and 70 across 20 simulation replicates. The true optimum obtained from the true model has the highest median, around 0.90, with relatively small spread. Bayesian model averaging has a median around 0.84 with a moderate interquartile range and a few lower outliers (around 0.6). This indicates fairly stable performance with occasional moderate underperformance. AIC has a similar median to Bayesian model averaging but a broader lower tail and a low outlier (around 0.47). Although BIC has a relatively high median, it has the largest spread and the lowest minimum (around 0.44). Figure 3.2 and Figure 3.3 highlight the risk of AIC and BIC methods since they can occasionally select a model that leads to severe under performance, whereas Bayesian model averaging is generally more stable and avoids the most extreme failures.

4.7.1 Sensitivity analysis of π

In this example, we assume that $\pi = 0.5$. Table 4.4 presents the the values of z^* for $\pi = 0.25, 0.5, 0.75$. While the optimal factor levels vary across these cases, the whole-plot factor x_1 changes from 0 for $\pi = 0.25$ and 0.50 to 1 for $\pi = 0.75$. In addition, z^* rises from $z^* = 0.6965$ at $\pi = 0.25$ to 0.7971 at $\pi = 0.5$, and further to 0.8332 at $\pi = 0.75$. These indicate that the achievable maximum model-averaged posterior probability and the optimal factor levels are sensitive to the model prior π. In this example, larger π (favoring larger mean models) yields higher z^* and shifts the optimum toward boundary settings for the whole-plot factor.

π	x_1	x_2	x_3	x_4	x_5	z^*
0.25	0	0.3022	-0.4983	-0.6032	-0.5665	0.6965
0.50	0	0.3471	-0.4020	-0.4085	-0.5686	0.7971
0.75	1	0.4009	0.0048	-0.1998	-0.0184	0.8332

Table 4.4: Sensitivity analysis for π.

In our program, convergence is monitored using Gelman-Robin $\hat{R}$ (< 1.01) diagnostics and effective sample sizes (ESS $>$ 10,000 for most parameters) from 4 parallel Stan chains, while scalability is computationally intensive due to the high-dimensional joint posterior but is mitigated through parallelization across chains and pre-screening of candidate models.

Chapter 5

Model selection using a generalized g-prior for data with a multistratum structure

5.1 Introduction

Zellner's g-prior is widely used in model comparison since it is computational convenient and accounts for the structure of the design matrix. Despite its convenience, the standard g-prior assumes independent and homoscedastic errors, limiting its direct applicability to complex data structures such as multistratum designs. In this paper, we propose a generalized g-prior for model selection within a Bayesian framework for data with a multistratum structure.

Section 5.2 introduces the generalized g-prior formulation that incorporates stratum-specific information and derives posterior parameter estimates. Section

5.3 develops the posterior model probabilities Section 5.4 presents four examples. For some examples, three effect size scenarios (large, moderate, small) are considered, and the performance of different g-prior choices is evaluated. Marginal posterior probabilities of individual effects are examined, and comparisons with AIC and BIC are provided.

5.2 Generalized g-prior and posterior parameter estimates

In this section, we introduce the generalized g-prior based on the model (3.1). For a given model M and $\boldsymbol{\rho}$, assume a non-informative prior for the total variance, that is,

$$\sigma_w^2 \mid M, \boldsymbol{\rho} \propto \frac{1}{\sigma_w^2}. \tag{5.1}$$

By adopting the idea of Zellner's g-prior (Zellner, 1986), the prior for $\boldsymbol{\beta}$ is assumed to be

$$\boldsymbol{\beta}|M, \sigma_w^2, \boldsymbol{\rho} \sim N(\tilde{\boldsymbol{\beta}}, g\sigma_w^2(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}), \tag{5.2}$$

where g is a scalar parameter that is to be chosen and can be interpreted as a measure of the information available in the prior relative to the sample. $\tilde{\boldsymbol{\beta}}$ is the prior mean of the regression coefficients of $\boldsymbol{\beta}$. It is constructed by centering the prior on the null model and scaling the covariance structure by the observed information matrix, multiplied by a hyper parameter g.

The effects of the model terms are assumed to be normally distributed. By

(5.2), we get

$$
\begin{aligned}
&P(\boldsymbol{\beta} \mid M, \sigma_w^2, \boldsymbol{\rho}) \\
&= (2\pi)^{-\frac{p+1}{2}} \det\big(g\sigma_w^2(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\big)^{-\frac{1}{2}} \exp\Big[-\frac{1}{2g\sigma_w^2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\Big] \\
&= (2\pi)^{-\frac{p+1}{2}} g^{-\frac{p+1}{2}} \sigma_w^{-(p+1)} |\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|^{\frac{1}{2}} \exp\Big[-\frac{1}{2g\sigma_w^2}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})\Big].
\end{aligned}
\tag{5.3}
$$

Since $\mathbf{Y} \sim N(\mathbf{X}\boldsymbol{\beta}, \sigma_w^2\mathbf{V})$, the likelihood of M is

$$
L(\mathbf{Y}|M, \sigma_w^2, \boldsymbol{\beta}, \boldsymbol{\rho}) = (2\pi)^{-\frac{n}{2}} \sigma_w^{-n} \, |\mathbf{V}|^{-\frac{1}{2}} \, \exp[-\frac{1}{2\sigma_w^2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})].
$$

Then the joint distribution of σ_w^2 and $\boldsymbol{\beta}$ given M, $\mathbf{Y}$ and $\boldsymbol{\rho}$ is

$$
\begin{aligned}
&P(\sigma_w^2, \boldsymbol{\beta}|M, \mathbf{Y}, \boldsymbol{\rho}) \\
&= L(\mathbf{Y}|M, \sigma_w^2, \boldsymbol{\beta}, \boldsymbol{\rho}) P(\boldsymbol{\beta}|M, \sigma_w^2, \boldsymbol{\rho}) P(\sigma_w^2|M, \boldsymbol{\rho}) \\
&\propto \sigma_w^{-(n+p+3)} g^{-\frac{p+1}{2}} |\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|^{\frac{1}{2}} \, |\mathbf{V}|^{-\frac{1}{2}} \\
&\exp\{-\frac{1}{2\sigma_w^2}[(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) \, + \frac{1}{g}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})]\}.
\end{aligned}
$$

Let

$$
\begin{aligned}
S =& \mathbf{Y}^T\mathbf{V}^{-1}\mathbf{Y} + \frac{1}{1+g}\tilde{\boldsymbol{\beta}}^T\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\tilde{\boldsymbol{\beta}} - \frac{2}{1+g}\mathbf{Y}^T\mathbf{V}^{-1}\mathbf{X}\tilde{\boldsymbol{\beta}} \\
&- \frac{g}{1+g}\mathbf{Y}^T\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y}
\end{aligned}
$$

and

$$
\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1} \, \mathbf{Y},
$$

which is the Bayesian estimator for $\boldsymbol{\beta}$ (Lin, 2018). One can easily check that

$$
\begin{aligned}
&(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) \ +\frac{1}{g}(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}) \\
&= S+\frac{1+g}{g}(\boldsymbol{\beta}-\frac{1}{1+g}\tilde{\boldsymbol{\beta}}-\frac{g}{1+g}\hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})(\boldsymbol{\beta}-\frac{1}{1+g}\tilde{\boldsymbol{\beta}}-\frac{g}{1+g}\hat{\boldsymbol{\beta}}).
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&P(\sigma_w^2,\boldsymbol{\beta}|M,\mathbf{Y},\boldsymbol{\rho}) \\
&\propto \sigma_w^{-(n+p+3)}g^{-\frac{p+1}{2}}|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|^{\frac{1}{2}}\ |\mathbf{V}|^{-\frac{1}{2}} \\
&\exp\{-\frac{1}{2\sigma_w^2}[S+\frac{1+g}{g}(\boldsymbol{\beta}-\frac{1}{1+g}\tilde{\boldsymbol{\beta}}-\frac{g}{1+g}\hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})(\boldsymbol{\beta}-\frac{1}{1+g}\tilde{\boldsymbol{\beta}}-\frac{g}{1+g}\hat{\boldsymbol{\beta}})]\}.
\end{aligned}
$$

By integrating over $\boldsymbol{\beta}$, we obtain that

$$
\begin{aligned}
& P(\sigma_w^2|M,\mathbf{Y},\boldsymbol{\rho}) \\
\propto\ & (1+g)^{-\frac{p+1}{2}}|\mathbf{V}|^{-\frac{1}{2}}\ \sigma_w^{-(n+1)}\exp[-\frac{1}{2\sigma_w^2}S]\ \int_{\boldsymbol{\beta}}\sigma_w^{-(p+1)}(\frac{1+g}{g})^{\frac{p+1}{2}}|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|^{\frac{1}{2}} \\
& \exp\{-\frac{1+g}{2g\sigma_w^2}[(\boldsymbol{\beta}-\frac{1}{1+g}\tilde{\boldsymbol{\beta}}-\frac{g}{1+g}\hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})(\boldsymbol{\beta}-\frac{1}{1+g}\tilde{\boldsymbol{\beta}}-\frac{g}{1+g}\hat{\boldsymbol{\beta}})]\}d\boldsymbol{\beta},
\end{aligned}
\tag{5.4}
$$

where inside the inner integral in (5.4) is the multivariate normal distribution for $\boldsymbol{\beta}$ with mean $\frac{1}{1+g}\tilde{\boldsymbol{\beta}}+\frac{g}{1+g}\hat{\boldsymbol{\beta}}$ and variance $\frac{g\,\sigma_w^2}{1+g}\left(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\right)^{-1}$, thus, is integrated to 1. It follows that

$$
\begin{aligned}
&P(\sigma_w^2|M,\mathbf{Y},\boldsymbol{\rho}) \\
&\propto (1+g)^{-\frac{p+1}{2}}|\mathbf{V}|^{-\frac{1}{2}}\ \sigma_w^{-(n+2)}\exp[-\frac{1}{2\sigma_w^2}S] \\
&\propto (1+g)^{-\frac{p+1}{2}}|\mathbf{V}|^{-\frac{1}{2}}\ S^{-\frac{n+2}{2}}(\frac{\sigma_w^2}{S})^{-\frac{n+2}{2}}\exp[-\frac{1}{2\sigma_w^2}S] \\
&\propto (\frac{\sigma_w^2}{S})^{-\frac{n+2}{2}}\exp[-\frac{1}{2\sigma_w^2}S].
\end{aligned}
$$

Note that (5.5) is the inverse Chi-squared distribution for $\frac{\sigma_w^2}{S}$ with degrees of freedom n, that is,

$$
\frac{\sigma_w^2}{S}|M,\mathbf{Y},\boldsymbol{\rho} \sim \quad inv-\chi_n^2.
$$

Similarly,

$$
\begin{aligned}
&P(\boldsymbol{\beta}|M,\mathbf{Y},\boldsymbol{\rho}) \\
\propto\ & g^{-\frac{p+1}{2}}|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|^{\frac{1}{2}}|\mathbf{V}|^{-\frac{1}{2}}\ \int_{\sigma_w^2}\sigma_w^{-(n+p+3)} \\
&\exp\{-\frac{1}{2\sigma_w^2}[\frac{1+g}{g}(\boldsymbol{\beta}-\frac{1}{1+g}\tilde{\boldsymbol{\beta}}-\frac{g}{1+g}\hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})(\boldsymbol{\beta}-\frac{1}{1+g}\tilde{\boldsymbol{\beta}}-\frac{g}{1+g}\hat{\boldsymbol{\beta}})+S]\}d\sigma_w^2.
\end{aligned}
$$

Assume

$$
a=\frac{1+g}{g}(\boldsymbol{\beta}-\frac{1}{1+g}\tilde{\boldsymbol{\beta}}-\frac{g}{1+g}\hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})(\boldsymbol{\beta}-\frac{1}{1+g}\tilde{\boldsymbol{\beta}}-\frac{g}{1+g}\hat{\boldsymbol{\beta}})+S
$$

and

$$
x=\sigma_w^2/a.
$$

Then,

$$
\begin{aligned}
& P(\boldsymbol{\beta}|M,\mathbf{Y},\boldsymbol{\rho}) \\
\propto\ & g^{-\frac{p+1}{2}}|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|^{\frac{1}{2}}|\mathbf{V}|^{-\frac{1}{2}}a^{-\frac{n+p+1}{2}} \int_x x^{-\frac{n+p+1}{2}-1}\exp\{-\frac{1}{2x}\}dx \quad (5.5)
\end{aligned}
$$

where in (5.5) the integrand is an inverse chi-square distribution with $\nu = n+p+1$ degrees of freedom. Thus,

$$
\begin{aligned}
& P(\boldsymbol{\beta}|M,\mathbf{Y},\boldsymbol{\rho}) \\
\propto\ & g^{-\frac{p+1}{2}}|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|^{\frac{1}{2}}|\mathbf{V}|^{-\frac{1}{2}}a^{-\frac{n+p+1}{2}} \\
\propto\ & g^{-\frac{p+1}{2}}|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|^{\frac{1}{2}}|\mathbf{V}|^{-\frac{1}{2}} \\
& [\frac{1+g}{g}(\boldsymbol{\beta}-\frac{1}{1+g}\tilde{\boldsymbol{\beta}}-\frac{g}{1+g}\hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})(\boldsymbol{\beta}-\frac{1}{1+g}\tilde{\boldsymbol{\beta}}-\frac{g}{1+g}\hat{\boldsymbol{\beta}})+S]^{-\frac{n+p+1}{2}} \\
\propto\ & g^{-\frac{p+1}{2}}|\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|^{\frac{1}{2}}|\mathbf{V}|^{-\frac{1}{2}}S^{-\frac{n+p+1}{2}} \\
& [1+\frac{1}{n}(\boldsymbol{\beta}-\frac{1}{1+g}\tilde{\boldsymbol{\beta}}-\frac{g}{1+g}\hat{\boldsymbol{\beta}})^T[\frac{Sg}{n(1+g)}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}]^{-1}(\boldsymbol{\beta}-\frac{1}{1+g}\tilde{\boldsymbol{\beta}}-\frac{g}{1+g}\hat{\boldsymbol{\beta}})]^{-\frac{n+p+1}{2}},
\end{aligned}
$$

which implies that $\boldsymbol{\beta} \mid M,\mathbf{Y},\boldsymbol{\rho}$ has a multivariate t-distribution with n degrees of freedom, and the mean is $\frac{1}{1+g}\tilde{\boldsymbol{\beta}}+\frac{g}{1+g}\hat{\boldsymbol{\beta}}$ and the covariance matrix is $\frac{Sg}{(1+g)(n-2)}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$, that is,

$$
\boldsymbol{\beta}|M,\mathbf{Y},\boldsymbol{\rho} \sim MVT(\frac{1}{1+g}\tilde{\boldsymbol{\beta}}+\frac{g}{1+g}\hat{\boldsymbol{\beta}},\frac{Sg}{(1+g)(n-2)}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}).
$$

In addition, by (5.1) and (5.2), the conditional distribution of $\boldsymbol{\beta}$ given M, σ_w^2,

$\mathbf{Y}$ and $\boldsymbol{\rho}$ is

$$
\begin{aligned}
&P(\boldsymbol{\beta}|M,\sigma_w^2,\mathbf{Y},\boldsymbol{\rho}) \\
&\propto \sigma_w^{-(n+p+1)} g^{-\frac{p+1}{2}} |\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|^{\frac{1}{2}}\ |\mathbf{V}|^{-\frac{1}{2}} \\
&\exp\{-\frac{1}{2\sigma_w^2}[S+\frac{1+g}{g}(\boldsymbol{\beta}-\frac{1}{1+g}\tilde{\boldsymbol{\beta}}-\frac{g}{1+g}\hat{\boldsymbol{\beta}})^T(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})(\boldsymbol{\beta}-\frac{1}{1+g}\tilde{\boldsymbol{\beta}}-\frac{g}{1+g}\hat{\boldsymbol{\beta}})]\} \\
&\propto \sigma_w^{-(n+p+1)} g^{-\frac{p+1}{2}} |\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|^{\frac{1}{2}}\ |\mathbf{V}|^{-\frac{1}{2}} \exp(-\frac{S}{2\sigma_w^2}) \\
&\exp\{-\frac{1}{2}[(\boldsymbol{\beta}-\frac{1}{1+g}\tilde{\boldsymbol{\beta}}-\frac{g}{1+g}\hat{\boldsymbol{\beta}})^T[(\frac{\sigma_w^2 g}{1+g}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}]^{-1}(\boldsymbol{\beta}-\frac{1}{1+g}\tilde{\boldsymbol{\beta}}-\frac{g}{1+g}\hat{\boldsymbol{\beta}})]\},
\end{aligned}
$$

which indicates that $\boldsymbol{\beta}$ follows the multivariate normal distribution with mean $\frac{1}{1+g}\tilde{\boldsymbol{\beta}}+\frac{g}{1+g}\hat{\boldsymbol{\beta}}$ and variance $\frac{\sigma_w^2 g}{1+g}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$, that is,

$$\boldsymbol{\beta}|M,\sigma_w^2,\mathbf{Y},\boldsymbol{\rho} \sim MVN(\frac{1}{1+g}\tilde{\boldsymbol{\beta}}+\frac{g}{1+g}\hat{\boldsymbol{\beta}},\frac{\sigma_w^2 g}{1+g}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}).$$

5.3 Posterior model probabilities

To conduct model selection, we need to compute the posterior model probabilities for the candidate models. Under the model M and $\boldsymbol{\rho}$, the marginal likelihood of

M is

$$
\begin{aligned}
&L(\mathbf{Y} \mid M, \boldsymbol{\rho}) \\
&\propto \int_{\sigma_w^2}\int_{\boldsymbol{\beta}} L(\mathbf{Y} \mid M, \sigma_w^2, \boldsymbol{\beta}, \boldsymbol{\rho}) P(\boldsymbol{\beta} \mid M, \sigma_w^2, \boldsymbol{\rho}) P(\sigma_w^2 \mid M, \boldsymbol{\rho})\, d\boldsymbol{\beta}\, d\sigma_w^2 \\
&\propto \int_{\sigma_w^2}\int_{\boldsymbol{\beta}} (1+g)^{-\frac{p+1}{2}} |\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|^{\frac{1}{2}} |\mathbf{V}|^{-\frac{1}{2}} \sigma_w^{-(n+2)} \exp\left[-\frac{1}{2\sigma_w^2} S\right] \sigma_w^{-(p+1)} \left(\frac{1+g}{g}\right)^{\frac{p+1}{2}} \\
&\qquad \exp\left\{-\frac{1+g}{2g\sigma_w^2}\left[(\boldsymbol{\beta} - \tfrac{1}{1+g}\tilde{\boldsymbol{\beta}} - \tfrac{g}{1+g}\hat{\boldsymbol{\beta}})^T (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})(\boldsymbol{\beta} - \tfrac{1}{1+g}\tilde{\boldsymbol{\beta}} - \tfrac{g}{1+g}\hat{\boldsymbol{\beta}})\right]\right\} d\boldsymbol{\beta}\, d\sigma_w^2, \\
&\propto \int_{\sigma_w^2} (1+g)^{-\frac{p+1}{2}} |\mathbf{V}|^{-\frac{1}{2}} \sigma_w^{-(n+2)} \exp\left[-\frac{1}{2\sigma_w^2} S\right] \left\{\int_{\boldsymbol{\beta}} \sigma_w^{-(p+1)} \left(\frac{1+g}{g}\right)^{\frac{p+1}{2}} |\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}|^{\frac{1}{2}}\right. \\
&\qquad \left.\exp\left\{-\frac{1+g}{2g\sigma_w^2}\left[(\boldsymbol{\beta} - \tfrac{1}{1+g}\tilde{\boldsymbol{\beta}} - \tfrac{g}{1+g}\hat{\boldsymbol{\beta}})^T (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})(\boldsymbol{\beta} - \tfrac{1}{1+g}\tilde{\boldsymbol{\beta}} - \tfrac{g}{1+g}\hat{\boldsymbol{\beta}})\right]\right\} d\boldsymbol{\beta}\right\} d\sigma_w^2, \\
&\propto \int_{\sigma_w^2} (1+g)^{-\frac{p+1}{2}} |\mathbf{V}|^{-\frac{1}{2}} \sigma_w^{-(n+2)} \exp\left[-\frac{1}{2\sigma_w^2} S\right] d\sigma_w^2 \qquad (5.6) \\
&\propto (1+g)^{-\frac{p+1}{2}} |\mathbf{V}|^{-\frac{1}{2}} S^{-\frac{n+2}{2}} \int_{\sigma_w^2} \left(\frac{\sigma_w^2}{S}\right)^{-\frac{n+2}{2}} \exp\left[-\frac{1}{2\sigma_w^2} S\right] d\sigma_w^2, \qquad (5.7) \\
&\propto (1+g)^{-\frac{p+1}{2}} |\mathbf{V}|^{-\frac{1}{2}} S^{-\frac{n}{2}}.
\end{aligned}
$$

Inside the inner integral in (5.6) is the multivariate normal distribution with mean $\frac{1}{1+g}\tilde{\boldsymbol{\beta}} + \frac{g}{1+g}\hat{\boldsymbol{\beta}}$ and variance $\frac{g\sigma_w^2}{1+g}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}$ and the expression within the integral in (5.7) constitutes an inverse Chi-squared distribution with n degrees of freedom, thus, both integrals evaluate to constants.

When $\boldsymbol{\rho}$ is unknown, the marginal likelihood of M is

$$
L(\mathbf{Y}|M) \propto \int_{\boldsymbol{\rho}} (1+g)^{-\frac{p+1}{2}} |\mathbf{V}|^{-\frac{1}{2}} S^{-\frac{n}{2}} p(\boldsymbol{\rho}) d\boldsymbol{\rho}, \qquad (5.8)
$$

where $p(\boldsymbol{\rho})$ is the density function of $\boldsymbol{\rho}$. For real-world datasets, the parameter $\boldsymbol{\rho}$ is usually unknown. Following Lin and Yang (2022), we assume that $\boldsymbol{\rho}$ follows

a Dirichlet distribution with concentration parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_s)$. The integral in Equation (5.8) can be evaluated by the function "integrate" in R (See the R code provided in Lin and Yang (2022)).

Let M_i be a model in the model space and π_j, $j = 1, 2, \ldots, m$, be the probability that a factor j is active. Assume that the prior probabilities of active factors are independent, the prior for the model M_i (Box and Meyer, 1993) is defined as

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

If $\pi_j = \pi$ for $j = 1, 2, \cdots, m$, then

$$P(M) = \pi^f (1 - \pi)^{m-f},$$

where f is the number of factors in the model M. The posterior probability of M_i for the i_{th} model given the data can be calculated by Bayes' theorem,

$$P(M_i|\mathbf{Y}) = \frac{L(\mathbf{Y}|M_i)P(M_i)}{\sum L(\mathbf{Y}|M_i)P(M_i)}.$$

Thus, by (5.8), the posterior probability of M_i given the data is

$$P(M_i|\mathbf{Y}, g) \propto \prod_{j \in M_i} (\pi_j) \prod_{j' \notin M_i} (1 - \pi_{j'}) \int_{\boldsymbol{\rho}} (1+g)^{-\frac{p+1}{2}} |\mathbf{V}|^{-\frac{1}{2}} S^{-\frac{n}{2}} p(\boldsymbol{\rho}) d\boldsymbol{\rho}. \quad (5.9)$$

When g is unknown, the posterior probability of M_i given the data is

$$P(M_i|\mathbf{Y}) \propto \prod_{j\in M_i}(\pi_j)\prod_{j'\notin M_i}(1-\pi_{j'})\int_g\int_{\boldsymbol{\rho}}(1+g)^{-\frac{p+1}{2}}|\mathbf{V}|^{-\frac{1}{2}}S^{-\frac{n}{2}}p(\boldsymbol{\rho})\pi(g)d\boldsymbol{\rho}dg, \tag{5.10}$$

where $\pi(g)$ is a prior of g. We will conduct the model selection based on the posterior model probability in (5.9) or (5.10).

5.4 Examples

To validate the proposed methodology, we present three examples in this section. Using the posterior model probabilities developed in Section 3, we perform model selection and discuss various choices of g.

The choice of g is critical and has been extensively studied in the literature. Kass and Wasserman (1995) proposed the unit information prior, specified by $g = n$, which produces Bayes factors asymptotically equivalent to the Bayesian Information Criterion (BIC). Alternatively, Foster and George (1994) developed priors for model selection based on the Risk Inflation Criterion (RIC), from a minimax perspective, recommended setting $g = q^2$, where q is the number of potential predictors.

Empirical Bayes (EB) approach estimate a separate g for each model (Liang et al., 2008). Based on the marginal likelihood $P(\mathbf{Y}|M,g)$ in (5.8), the empirical Bayes g-prior for a model M is:

$$\hat{g}_{EB} = \arg\max_g P(\mathbf{Y}|M,g).$$

If a model has large effect size, then $\hat{g}_{EB}$ will be large, while for the small effect size, it will be small. Due to the fact that the original hyper-g prior lacked consistency under the null model, Liang et al. (2008) proposed a modification which is known as the hyper-g/n prior and defined as:

$$\pi_1(g) = \frac{a-2}{2n}\left(1+\frac{g}{n}\right)^{-\frac{a}{2}}, \quad g > 0,$$

where $2 < a \leq 4$ is suggested. This prior is a variant of the hyper-g prior, scaled by n to ensure asymptotic consistency. We use $a = 3$ in our examples. We also consider the case $a = 4$, for which the results are similar and therefore not shown here.

In this paper, we also consider a prior $\pi_2(g) = g^{-2}\exp(-g^{-1})$, which is an inverse gamma distribution with shape 1 and scale 1. This is motivated by Marin and Robert (2007) who considered a prior $\pi_3(g) = g^{-3/4}$, which can be considered as a degenerate inverse gamma distribution with shape $-\frac{1}{4}$ and scale 0 (Bové and Held, 2011). When $\pi_1(g)$, $\pi_2(g)$, and $\pi_3(g)$ are used, the posterior probability for each model will be calculated by (5.10).

In Bayesian analysis, the marginal posterior probability for effect j is defined (Box and Meyer, 1993) as

$$P_j = \sum_{M_i\text{: effect } j \text{ is included in } M_i} P(M_i \mid \mathbf{Y}),$$

where the sum is taken over all candidate models M_i that include the effect j. An effect is considered important if it appears in a few models with high posterior probabilities or in many models with relatively low posterior probabilities.

5.4.1 Example 1

In this example, we use a practical response surface experiment considered in Vining et al. (2005) and Xiao et al. (2025) to illustrate our method. The experiment involves two hard-to-change (whole-plot) factors, x_1 and x_2 (furnace temperatures at zones 1, and 2, respectively), and two easy-to-change (subplot) factors, x_3 (binder amount in the formulation) and x_4 (grinding speed of the batch). To investigate the effects of four factors on the strength of a ceramic pipe, Xiao et al. (2025) constructed a 48-run two-stage Bayesian *I-I* optimal split-plot design with 12 whole plots, as shown in Table 5.1. We will consider three cases of model effect sizes:

Case 1: large effect size

$$E(y_1) = 100 + 20.38x_1 + 27.20x_2 + 21.30x_3 + 25.11x_4 - 23.64x_2^2 + 21.35x_4^2 - 29.328x_1x_3 + 27.09x_2x_4.$$

Case 2: moderate effect size

$$E(y_2) = 100 + 11.38x_1 + 20.20x_2 + 12.30x_3 + 18.23x_4 - 12.64x_2^2 + 11.35x_4^2 - 20.98x_1x_3 + 17.33x_2x_4.$$

Case 3: small effect size

$$E(y_3) = 100 + 7.38x_1 + 10.20x_2 + 8.30x_3 + 10.87x_4 - 7.64x_2^2 + 7.35x_4^2 - 10.83x_1x_3 + 9.63x_2x_4.$$

The simulated responses y_1, y_2, and y_3, generated from multivariate normal distributions with $\sigma_1^2 = \sigma_2^2 = \sigma^2 = 1$, are presented in Table 5.1. The column labeled "wp" indicates the whole plot number. Since this experiment involves 4 factors, there are 2^{14} possible models. Using the strong effect heredity principle, that is, (1) if an interaction term is in the model, the corresponding main effects must also be included, and (2) if a quadratic effect is included, the corresponding main effect must also be included, we identify 1,337 potential models.

Table 5.1: 48-run Bayesian *I*-*I* optimal split-plot design with simulated responses

wp	x_1	x_2	x_3	x_4	y_1	y_2	y_3
1	0	0	0	0	100.17	100.17	100.17
1	0	0	−1	1	123.81	115.93	108.57
1	0	0	1	−1	118.53	106.41	105.77
1	0	0	0	0	102.29	102.29	102.29
2	1	−1	0	0	70.12	79.12	90.12
2	1	−1	−1	1	96.29	98.83	100.02
2	1	−1	1	−1	85.26	80.73	93.54
2	1	−1	1	1	80.91	82.13	95.62
3	0	1	0	1	178.07	155.43	131.37
3	0	1	0	−1	72.85	83.49	89.55
3	0	1	−1	0	82.28	95.28	94.28
3	0	1	1	0	125.52	120.52	111.52
4	0	−1	0	1	70.05	80.93	92.27
4	0	−1	−1	−1	50.81	64.93	79.59
4	0	−1	1	0	70.81	79.81	90.81
4	0	−1	0	0	49.52	67.52	82.52
5	−1	0	1	0	132.12	123.78	113.63
5	−1	0	−1	0	29.34	55.69	73.84
5	−1	0	0	1	126.62	118.74	111.38
5	−1	0	0	−1	77.50	83.38	90.74
6	1	0	−1	−1	124.83	113.37	106.58
6	1	0	1	1	158.75	132.22	123.01
6	1	0	0	0	122.05	113.05	109.05
6	1	0	0	0	121.42	112.42	108.42

Table 5.1: 48-run Bayesian *I*-*I* optimal split-plot design with simulated responses continued

wp	x_1	x_2	x_3	x_4	y_1	y_2	y_3
7	−1	−1	1	−1	101.83	98.60	99.11
7	−1	−1	−1	0	−22.42	21.93	55.08
7	−1	−1	1	1	97.38	99.91	101.10
7	−1	−1	0	1	46.55	66.43	81.77
8	1	1	−1	0	131.92	127.57	112.42
8	1	1	1	1	190.25	157.96	136.05
8	1	1	−1	1	205.95	174.96	140.75
8	1	1	1	−1	84.72	85.71	93.92
9	0	0	0	0	99.28	99.28	99.28
9	0	0	−1	0	76.74	85.74	89.74
9	0	0	0	0	99.68	99.68	99.68
9	0	0	0	1	144.59	127.71	116.35
10	−1	1	1	1	207.38	176.40	142.19
10	−1	1	−1	1	103.15	106.86	100.95
10	−1	1	−1	−1	3.04	40.03	64.24
10	−1	1	0	0	83.30	96.30	95.30
11	0	0	−1	0	80.33	89.33	93.33
11	0	0	1	0	122.98	113.98	109.98
11	0	0	0	−1	96.51	93.39	96.75
11	0	0	0	0	100.51	100.51	100.51
12	0	0	−1	0	81.37	90.37	94.37
12	0	0	0	−1	99.47	96.35	99.71
12	0	0	0	0	100.54	100.54	100.54
12	0	0	1	1	169.70	143.82	128.46

Table 5.2 presents the selected models from the 1,337 potential models using different choices of g-priors when $\boldsymbol{\alpha} = (1,1)$ and $\pi = 0.5$. It is observed that the true model is selected for all the three cases when $\pi_i(g)$, $i = 1,2,3$, are used as g-priors. However, although the empirical Bayes g-prior $\hat{g}_{EB}$ selects the true model under small effect sizes, it tends to ignore the quadratic effect x_2^2 when the effect sizes are moderate or large.

When $g = n = 48$, the proposed model selection method can identify the true model when effect sizes are large. It fails to select the quadratic effect x_2^2 when models have small or moderate effect sizes. Since main effects, quadratic effects, and two-factor interactions are potential predictors, $q = 14$. When $g = q^2 = 196$, the proposed model selection method can identify the true model when the effect sizes are large or moderate. However, the quadratic effect x_2^2, which is included in the true model, can not be identified when the effect size is small. This indicates that the proposed model selection method when g is fixed may have limited power to identify quadratic effects when effect sizes are relatively small.

Although the true model is not the top-selected model for small effect sizes when a fixed g-prior is applied, it is ranked second in some cases. For example, Table 5.3 presents the top five models selected for Case 3 when $g = 196$. It shows that the true model is ranked second, with a posterior model probability of 0.1937, which is much larger than that of the next three models. We also considered other values of $\boldsymbol{\alpha}$ from (0.25, 0.25) to (100, 100) and found that the true model can be selected in some cases. In particular, the true model is identified when $\boldsymbol{\alpha}$ is increased to $\boldsymbol{\alpha} = (7,7)$ for $g = 48$ in Case 2, whereas an increase to $\boldsymbol{\alpha} = (4,4)$ is required for $g = 196$ in Case 3. For $\hat{g}_{EB}$, the true

model is identified only when α is increased to $(9,9)$ in Case 1 and decreased to $(0.5,0.5)$ in Case 3.

We also considered the AIC and BIC model selection methods. The selected models are summarized in Table 5.2. Both methods select the same model for each of the three cases. However, for Cases 1 and 3, additional effects not in the true model are selected. For example, for both cases, x_2x_3 and x_3^2, neither of which is in the true model, are selected. For Case 2, the effect x_2^2, which is part of the true model, is not selected, while the additional effects x_2x_3, x_1^2, and x_3^2 are included. This indicates that using AIC and BIC may result in overfitting, with selected models including extra, non-significant terms rather than identifying the true model. Therefore, the proposed Bayesian model selection method using g-priors, especially the integrated g-priors $\pi_i(g)$, $i = 1,2,3$, perform better than the AIC and BIC methods in this problem.

Table 5.2: Selected models using different methods for Case 1, Case 2, and Case 3

Methods	Case 1	Case 2	Case 3
$\pi_i(g)$, $i = 1,2,3$	$x_1,x_2,x_3,x_4,x_1x_3,$ x_2x_4,x_2^2,x_4^2	$x_1,x_2,x_3,x_4,x_1x_3,$ x_2x_4,x_2^2,x_4^2	$x_1,x_2,x_3,x_4,x_1x_3,$ x_2x_4,x_2^2,x_4^2
$\hat{g}_{EB}$	$x_1,x_2,x_3,x_4,x_1x_3,$ x_2x_4,x_4^2	$x_1,x_2,x_3,x_4,x_1x_3,$ x_2x_4,x_4^2	$x_1,x_2,x_3,x_4,x_1x_3,$ x_2x_4,x_2^2,x_4^2
$g = 48$	$x_1,x_2,x_3,x_4,x_1x_3,$ x_2x_4,x_2^2,x_4^2	$x_1,x_2,x_3,x_4,x_1x_3,$ x_2x_4,x_4^2	$x_1,x_2,x_3,x_4,x_1x_3,$ x_2x_4,x_4^2
$g = 196$	$x_1,x_2,x_3,x_4,x_1x_3,$ x_2x_4,x_2^2,x_4^2	$x_1,x_2,x_3,x_4,x_1x_3,$ x_2x_4,x_2^2,x_4^2	$x_1,x_2,x_3,x_4,x_1x_3,$ x_2x_4,x_4^2
AIC, BIC	$x_1,x_2,x_3,x_4,x_1x_3,$ $x_2x_3,x_2x_4,x_2^2,x_3^2,x_4^2$	$x_1,x_2,x_3,x_4,x_1x_3,$ $x_2x_3,x_2x_4,x_1^2,x_3^2,x_4^2$	$x_1,x_2,x_3,x_4,x_1x_3,$ $x_2x_4,x_3x_4,x_1^2,x_2^2,x_3^2,x_4^2$

Table 5.3: The top 5 potential models for Case 3 when $g = 196$

Model	$P(M\|\mathbf{Y})$
$x_1, x_2, x_3, x_4, x_4^2, x_1x_3, x_2x_4$	0.3969
$x_1, x_2, x_3, x_4, x_2^2, x_4^2, x_1x_3, x_2x_4$	0.1937
$x_1, x_2, x_3, x_4, x_4^2, x_1x_3, x_1x_4, x_2x_4$	0.0534
$x_1, x_2, x_3, x_4, x_4^2, x_1x_3, x_2x_4, x_3x_4$	0.0418
$x_1, x_2, x_3, x_4, x_1^2, x_4^2, x_1x_3, x_2x_4$	0.0347

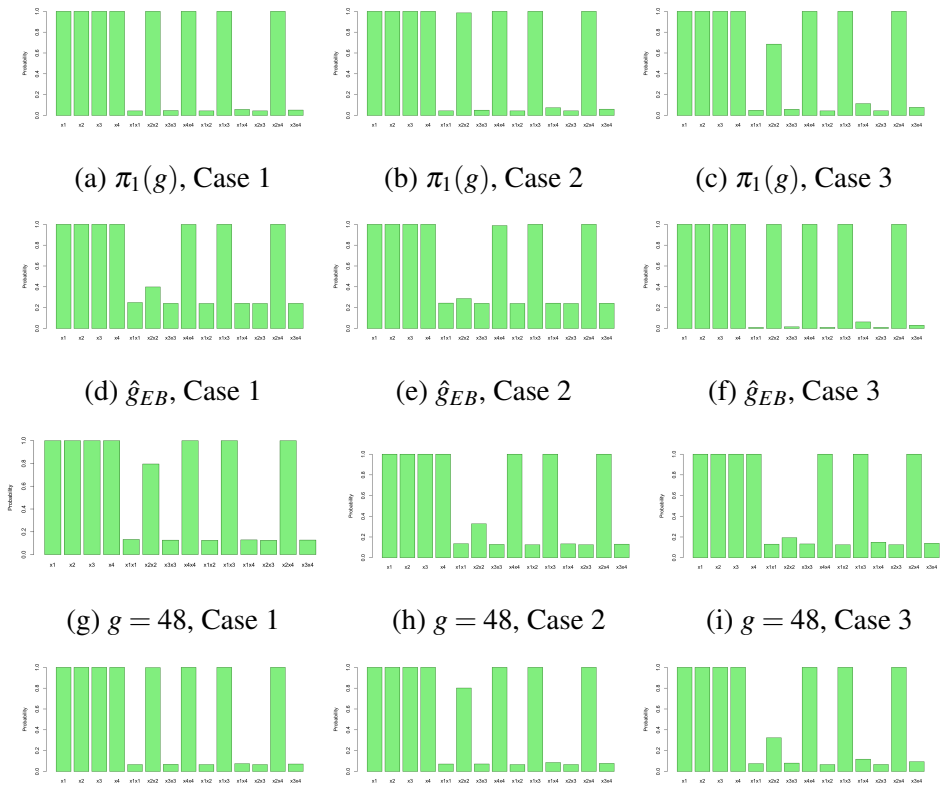
Figure 5.1 displays the marginal posterior probabilities of the effects for the design in Table 5.1 across all three cases using the g-priors discussed in this section, except for $\pi_2(g)$ and $\pi_3(g)$ whose graphs are similar to those obtained using $\pi_1(g)$.

Clearly, when the prior $\pi_1(g)$ (and likewise $\pi_2(g)$ and $\pi_3(g)$) is used, the marginal posterior probabilities of all effects in the true model are either equal to 1 or very close to 1, except for x_2^2 in Case 3, which has a probability of 0.685 but is still not small. On the other hand, when the empirical Bayes g-prior $\hat{g}_{EB}$ is used, all effects in the true model have marginal posterior probabilities equal to or close to 1 for large effect sizes; however, for small and moderate effect sizes, the marginal posterior probability of the effect x_2^2 is small.

When the prior $g = 48$ and $g = 196$ are applied, the marginal posterior probabilities of those effects are either equal to 1 or close to 1 although the marginal posterior probability of the effect x_2^2 is small in Cases 2 and 3 when $g = 48$ and Case 3 when $g = 196$. In particular, for Cases 1, 2, and 3, x_2^2 has larger marginal posterior probabilities 0.998, 0.801, and 0.324, respectively, when $g = 196$ than when $g = 48$, for which the corresponding probabilities are 0.795, 0.328, and 0.193, respectively. This indicates that although the

proposed Bayesian model selection method may not have sufficient power to select quadratic terms when a fixed g-prior is used, increasing the value of g may improve the power.

Based on our results, for this experiment we recommend using the priors $\pi_i(g), i = 1,2,3$, for the Bayesian model selection to identify active effects. This is expected to results in accurate parameter estimation and effective response prediction for the strength of the ceramic pipe.



(a) $\pi_1(g)$, Case 1 (b) $\pi_1(g)$, Case 2 (c) $\pi_1(g)$, Case 3

(d) $\hat{g}_{EB}$, Case 1 (e) $\hat{g}_{EB}$, Case 2 (f) $\hat{g}_{EB}$, Case 3

(g) $g = 48$, Case 1 (h) $g = 48$, Case 2 (i) $g = 48$, Case 3

(j) $g = 196$, Case 1 (k) $g = 196$, Case 2 (l) $g = 196$, Case 3

Figure 5.1: Marginal posterior probabilities of the effects for the design in Table 5.1

5.4.2 Example 2

Consider the 20-run *D*-optimal staggered-level design with four factors in Table 5.4 (Table 9 in Arnouts and Goos (2015)), where x_1 and x_2 are factors whose levels are very hard-to-change and hard-to-change, respectively, and x_3 and x_4 are factors whose levels are easy-to-change. Three model effect sizes will be considered:

Case 1: large effect size

$$E(y_1) = 70 + 15.38x_1 + 23.20x_2 + 21.30x_3 - 25.3x_4 - 19.64x_1^2 + 22.35x_1x_2.$$

Case 2: moderate effect size

$$E(y_2) = 70 + 11.38x_1 + 20.20x_2 + 12.30x_3 - 20.3x_4 - 12.64x_1^2 + 11.35x_1x_2.$$

Case 3: small effect size

$$E(y_3) = 70 + 7.38x_1 + 9.20x_2 + 8.30x_3 - 10.3x_4 - 5.64x_1^2 + 6.35x_1x_2.$$

Responses are generated from multivariate normal distributions with $\sigma_1^2 = \sigma_2^2 = \sigma^2 = 1$. The simulated responses y_1, y_2, and y_3 are presented in the last three columns in Table 5.4. By applying the strong effect heredity principle, we select 1,337 candidate models.

Table 5.4: 20-run D-optimal staggered-level design with simulated responses

Whole plot	Subplot	x_1	x_2	x_3	x_4	y_1	y_2	y_3
1	1	0	0	0	0	69.36	69.36	69.36
1	1	0	0	−1	−1	73.82	77.82	71.82
1	2	0	−1	1	0	67.75	61.75	68.75
1	2	0	−1	−1	1	1.15	18.15	43.15
2	2	−1	−1	1	1	29.07	28.07	51.07
2	2	−1	−1	0	−1	57.47	55.47	62.47
2	3	−1	1	1	1	32.23	47.23	58.23
2	3	−1	1	−1	1	−9.91	23.09	42.09
3	3	1	1	1	1	110.21	95.21	88.21
3	3	1	1	0	−1	139.42	123.42	100.42
3	4	1	−1	−1	−1	24.77	45.77	58.77
3	4	1	−1	0	1	−3.78	18.22	47.22
4	4	−1	−1	−1	0	11.68	23.68	44.68
4	4	−1	−1	1	−1	79.60	68.60	71.60
4	5	−1	1	1	1	29.86	44.86	55.86
4	5	−1	1	1	−1	82.13	87.13	78.13
5	5	1	1	1	−1	156.14	131.14	104.14
5	5	1	1	−1	1	62.09	65.09	66.09
5	6	1	0	1	−1	109.96	98.96	87.96
5	6	1	0	−1	0	42.46	54.46	61.46

Table 5.5 displays the selected models for the three cases with different choices of g-priors when $\boldsymbol{\alpha} = (1,1,1)$ and $\pi = 0.5$. Our results show that the proposed Bayesian model selection method works well in all three cases when the empirical Bayes g-prior $\hat{g}_{EB}$, $\pi_1(g)$, $\pi_2(g)$, and $\pi_3(g)$ are utilized, since the true model is selected in all the cases. It indicates that the proposed method is robust to the integrated g-priors. On the other hand, when $g = n = 20$ and $g = q^2 = 196$, the top-ranked selected models include all the main effects and the interaction term x_1x_2 for all three cases. The term x_1^2, which is included in the true model, was selected only for models with large and moderate effect sizes when $g = 196$. However, our results show that the true model is ranked second

for all three cases when $g = 48$ and for Case 3 when $g = 196$. We have further considered increasing the values of $\boldsymbol{\alpha}$ from $(2,2,2)$ to $(20,20,20)$, however, the true model still cannot be identified.

Table 5.5 also lists the selected models using the AIC and BIC methods. The models selected by the AIC method across the three cases are identical and include the extra terms, x_2x_4 and x_2^2, compared with the true model. Similarly, the models selected by the BIC method for the three cases are also the same but include one additional term, x_2x_4. This again indicates that these selection criteria have a tendency to overfit the model.

Table 5.5: Selected models using different methods for Case 1, Case 2, and Case 3

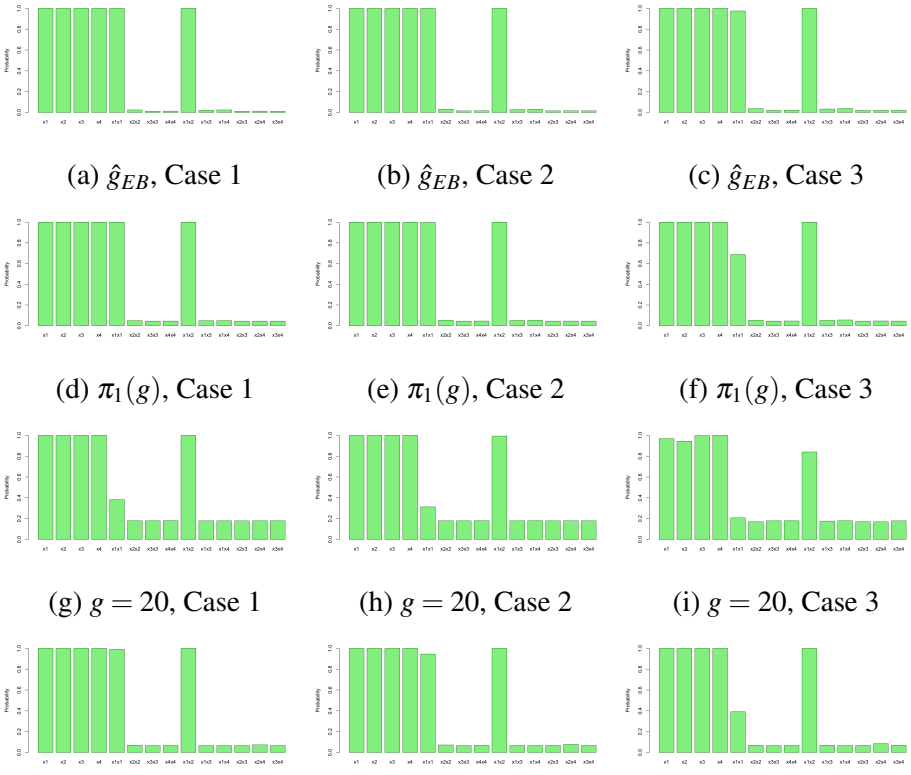
Methods	Case 1	Case 2	Case 3
$\hat{g}_{EB}$, $\pi_i(g)$, $i = 1,2,3$	$x_1, x_2, x_3, x_4, x_1x_2, x_1^2$	$x_1, x_2, x_3, x_4, x_1x_2, x_1^2$	$x_1, x_2, x_3, x_4, x_1x_2, x_1^2$
$g = 20$	$x_1, x_2, x_3, x_4, x_1x_2$	$x_1, x_2, x_3, x_4, x_1x_2$	$x_1, x_2, x_3, x_4, x_1x_2$
$g = 196$	$x_1, x_2, x_3, x_4, x_1x_2, x_1^2$	$x_1, x_2, x_3, x_4, x_1x_2, x_1^2$	$x_1, x_2, x_3, x_4, x_1x_2$
AIC	$x_1, x_2, x_3, x_4, x_1x_2, x_2x_4, x_1^2, x_2^2$	$x_1, x_2, x_3, x_4, x_1x_2, x_2x_4, x_1^2, x_2^2$	$x_1, x_2, x_3, x_4, x_1x_2, x_2x_4, x_1^2, x_2^2$
BIC	$x_1, x_2, x_3, x_4, x_1x_2, x_2x_4, x_1^2$	$x_1, x_2, x_3, x_4, x_1x_2, x_2x_4, x_1^2$	$x_1, x_2, x_3, x_4, x_1x_2, x_2x_4, x_1^2$

Figure 5.2 displays the marginal posterior probabilities over the candidate models for each effect in all three cases when the empirical Bayes, $\pi_1(g)$, $g = 20$, and $g = 196$ are used as g-priors. The marginal posterior probabilities of each effect obtained using $\pi_2(g)$ and $\pi_3(g)$ are similar to those obtained using $\pi_1(g)$. Therefore, their plots are not included in Figure 5.2. Our results show that when the empirical Bayes and $\pi_1(g)$ priors are used, the marginal posterior probabilities

of all main effects, the interaction effect x_1x_2, and the quadratic term x_1^2 are equal to 1 in all three cases, except for x_1^2, whose marginal posterior probabilities in Case 3 are 0.975 and 0.688, respectively, which are still relatively high. This further demonstrates that the integrated g-priors are robust for selecting the true model.

When $g = 20$, the results show that for large and moderate effect sizes, the marginal posterior probabilities of all main effects and the interaction effect x_1x_2 are approximately equal to 1, whereas the marginal posterior probabilities of x_1^2 are 0.383 and 0.313, respectively. For the small effect size, the marginal posterior probabilities of these effects are smaller: for the main effects, they range from 0.943 to 1, while for x_1x_2 and x_1^2 they are 0.840 and 0.208, respectively. Note that in all the cases, x_1^2 has relatively small marginal posterior probabilities and thus cannot be identified in the model selection.

When $g = 196$, the marginal posterior probabilities of all effects in the true model are equal to 1, except for x_1^2 under moderate and small effects sizes, for which the probabilities are 0.945 and 0.392, respectively. These values are larger in all the corresponding cases than those obtained when $g = 20$. This once again indicates that when a fixed g-prior is used, the proposed Bayesian model selection method may not have enough power to select quadratic terms, but the power can be improved when the value of g increases.



(a) $\hat{g}_{EB}$, Case 1 (b) $\hat{g}_{EB}$, Case 2 (c) $\hat{g}_{EB}$, Case 3

(d) $\pi_1(g)$, Case 1 (e) $\pi_1(g)$, Case 2 (f) $\pi_1(g)$, Case 3

(g) $g = 20$, Case 1 (h) $g = 20$, Case 2 (i) $g = 20$, Case 3

(j) $g = 196$, Case 1 (k) $g = 196$, Case 2 (l) $g = 196$, Case 3

Figure 5.2: Marginal posterior probabilities of the effects for the design in Table 5.4

5.4.3 Example 3

Consider the 27-run D-optimal split-plot design in nine whole plots with one whole-plot factor w and two subplot factors s_1 and s_2 in Table 5.6 (Table 3 in Goos and Vandebroek (2003)). Three models are considered:

Case 1: large effect size

$$E(y_1) = 100 + 20.38x_1 + 27.20x_2 + 21.30x_3 - 23.64x_1^2 + 21.35x_1x_2.$$

Case 2: moderate effect size

$$E(y_2) = 100 + 11.38x_1 + 20.20x_2 + 12.30x_3 - 12.64x_1^2 + 11.35x_1x_2.$$

Case 3: small effect size

$$E(y_3) = 100 + 7.38x_1 + 10.20x_2 + 8.30x_3 - 7.64x_1^2 + 7.35x_1x_2.$$

For each model, data are generated from multivariate normal distributions with the variances $\sigma_1^2 = \sigma^2 = 1$. The obtained data are presented in Table 5.6.

Table 5.6: 27-run *D*-optimal split-plot design with simulated responses

Whole plot	w	s_1	s_2	y_1	y_2	y_3
1	−1	1	0	61.78	84.78	87.78
1	−1	−1	−1	30.85	56.85	75.85
1	−1	−1	1	70.84	78.84	89.84
2	−1	1	1	81.51	95.51	94.51
2	−1	−1	1	70.18	78.18	89.18
2	−1	0	−1	33.77	62.77	75.77
3	−1	1	−1	41.62	73.62	80.62
3	−1	1	1	82.94	96.94	95.94
3	−1	−1	0	49.52	66.52	81.52
4	0	1	−1	105.78	107.78	101.78
4	0	0	0	99.98	99.98	99.98
4	0	−1	−1	52.29	68.29	82.29
5	0	1	0	128.99	121.99	111.99
5	0	0	1	121.67	112.67	108.67
5	0	−1	−1	52.51	68.51	82.51
6	1	0	1	120.69	113.69	110.69
6	1	1	−1	125.42	119.42	110.42
6	1	−1	−1	29.31	57.31	76.31
7	1	1	0	143.64	128.64	115.64
7	1	0	−1	74.06	85.06	90.06
7	1	−1	1	69.05	79.05	90.05
8	1	−1	0	49.38	68.38	83.38
8	1	1	1	167.38	143.38	126.38
8	1	1	−1	124.20	118.20	109.20
9	1	0	−1	76.84	87.84	92.84
9	1	−1	1	70.82	80.82	91.82
9	1	1	1	167.26	143.26	126.26

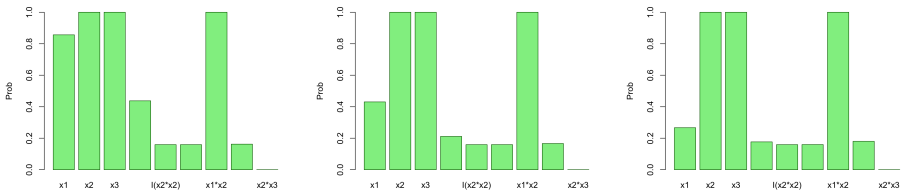
Since the split-plot design involves three factors and a model can contain main effects, two-factor interaction effects, and quadratic effects, there are $2^9 = 512$ potential models. We will compute the posterior probability for each model assuming $\boldsymbol{\alpha} = (1,1)$ and $\pi = 0.5$. Table 5.7 lists the selected models for each case with different choices of g and AIC and BIC.

The empirical Bayes estimator $\hat{g}_{EB}$ demonstrates strong performance, as it captures the true model for all cases. $\pi_1(g)$ (where $a = 3,4$), $\pi_2(g)$ and $\pi_3(g)$ identify the true model for large and moderate effect sizes, but not for small effect size. Figure 5.3 presents the marginal posterior probabilities for the nine candidate terms in three cases using $\pi_1(g)$ with $a = 3$. It indicates that Case 1 and Case 2 successfully select all the terms in the true model. However, the marginal posterior probability for the term x_1^2 is lower in Case 2 than in Case 1. On the other hand, Case 3 fails to select the true model, since the marginal posterior probability of x_1^2 is very small. As shown in Table 5.8, which presents the top five selected models for Case 3 when $\pi_1(g)$ with $a = 3$ is used, the true model is ranked second with a relatively smaller posterior probability.

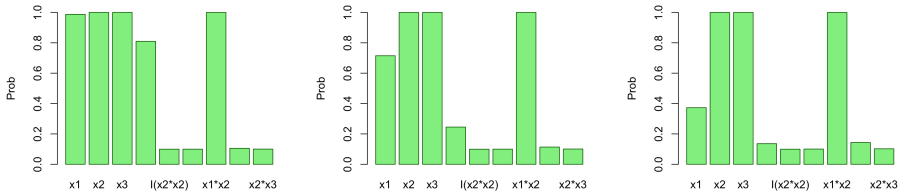
The results also show that $g = n = 27$ and $g = p^2 = 81$ are not able to identify the correct model, especially for moderate and small effect sizes for all the cases except for Case 1 when $g = 81$. However, after considering different values of $\boldsymbol{\alpha}$, we find that increasing the values of $\boldsymbol{\alpha}$ may increase the likelihood of selecting the true model. Therefore, for moderate and small effect sizes, we suggest setting a larger values of $\boldsymbol{\alpha}$ compared to the case of the large effect size. For AIC and BIC, they miss some terms in Case 1, while they select extra terms in Cases 2 and 3, indicating that they do not perform better than $\hat{g}_{EB}$ and $\pi_i(g)$, $i = 1,2,3$.

Table 5.7: Selected models for Case 1, Case 2, and Case 3 with different choices of g

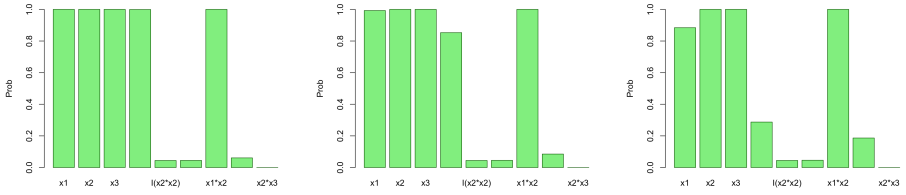
g prior	Case 1	Case 2	Case 3
$\hat{g}_{EB}$	$x_1, x_2, x_3, x_1^2, x_1x_2$	$x_1, x_2, x_3, x_1^2, x_1x_2$	$x_1, x_2, x_3, x_1^2, x_1x_2$
$\pi_i(g)$, $i = 1,2,3$	$x_1, x_2, x_3, x_1^2, x_1x_2$	$x_1, x_2, x_3, x_1^2, x_1x_2$	x_1, x_2, x_3, x_1x_2
$g = 27$	x_1, x_2, x_3, x_1x_2	x_2, x_3, x_1x_2	x_2, x_3, x_1x_2
$g = 81$	$x_1, x_2, x_3, x_1^2, x_1x_2$	x_1, x_2, x_3, x_1x_2	x_2, x_3, x_1x_2
AIC, *BIC*	x_1, x_2, x_3, x_1x_2	$x_1, x_2, x_3, x_1^2, x_3^2, x_1x_2, x_2x_3$	$x_1, x_2, x_3, x_1^2, x_2^2, x_3^2, x_2x_3$



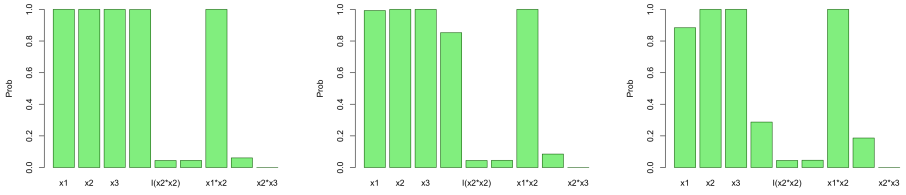
(a) $g = 27$, Case 1 (b) $g = 27$, Case 2 (c) $g = 27$, Case 3



(d) $g = 81$, Case 1 (e) $g = 81$, Case 2 (f) $g = 81$, Case 3

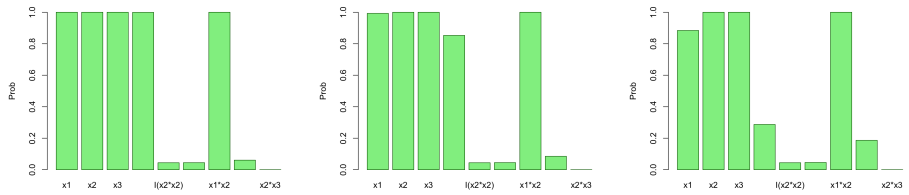


(g) $\hat{g}_{EB}$, Case 1 (h) $\hat{g}_{EB}$, Case 2 (i) $\hat{g}_{EB}$, Case 3

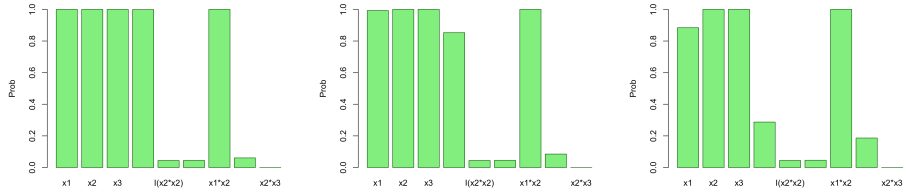


(j) $\pi_1(g)$, Case 1 (k) $\pi_1(g)$, Case 2 (l) $\pi_1(g)$, Case 3

Figure 5.3: (a) Marginal posterior probabilities of the effects for the design in Table 5.6



(m) $\pi_2(g)$, Case 1 (n) $\pi_2(g)$, Case 2 (o) $\pi_2(g)$, Case 3



(p) $\pi_3(g)$, Case 1 (q) $\pi_3(g)$, Case 2 (r) $\pi_3(g)$, Case 3

Figure 5.3: (b) Marginal posterior probabilities of the effects for the design in Table 5.6

Table 5.8: The top 5 potential models for Case 3 using $\pi_1(g)$ with $a = 3$

Model	$P(M\|\mathbf{Y})$
x_1, x_2, x_3, x_1x_2	0.4271
$x_1, x_2, x_3, x_1^2, x_1x_2$	0.1967
$x_1, x_2, x_3, x_1x_2, x_1x_3$	0.0983
x_2, x_3, x_1x_2	0.0750
$x_1, x_2, x_3, x_1^2, x_1x_2, x_1x_3$	0.0446

Table 5.9: Selected models for different choices of $\boldsymbol{\alpha}$ for Case 3

$\boldsymbol{\alpha}$	$\pi_1(g)$	$\pi_2(g)$	$\pi_3(g)$
$(2,2)$, $(3,3)$	x_1, x_2, x_3, x_1x_2	$x_1, x_2, x_3, x_1^2, x_1x_2$	x_1, x_2, x_3, x_1x_2
$(4,4)$, $(5,5)$	$x_1, x_2, x_3, x_1^2, x_1x_2$	$x_1, x_2, x_3, x_1^2, x_1x_2$	$x_1, x_2, x_3, x_1^2, x_1x_2$

However, when the value of $\boldsymbol{\alpha}$ changes, it is possible to select the true model in Case 3. Table 5.9 lists the selected models with $\boldsymbol{\alpha} = (2,2)$, $(3,3)$, $(4,4)$, and

$(5,5)$ for Case 3. It is observed that the true model is selected for all the choices of $\boldsymbol{\alpha}$ when $\pi_2(g)$ is used and only for $\boldsymbol{\alpha} = (4,4)$ and $(5,5)$ when $\pi_1(g)$ and $\pi_3(g)$ are applied. It indicates that increasing the values of $\boldsymbol{\alpha}$ improves the accuracy of model selection.

Chapter 6

Conclusion and future work

This thesis has developed Bayesian methodologies for model selection and response optimization in multistratum designs. The proposed approaches account for model uncertainty and complex error structures, providing robust and effective tools for industrial experiments.

We propose two Bayesian predictive approaches to study the response optimization problem when the data has a multistratum structure. To reduce model uncertainty, we apply the Bayesian model-averaging approach to optimize the response. In Chapter 3, error terms are assumed to be normally distributed. Under appropriate prior assumptions, we show that the posterior predictive density of a new response follows a Student's t-distribution. We develop R code to calculate the marginal likelihood using bridge sampling. In Chapter 4, error terms are assumed to follow t distributions. Stan and bridge sampling are used to estimate the model parameters and marginal likelihood. Simulation studies and robustness analyses suggest that the Bayesian model averaging approach performs more

effectively and stably than AIC and BIC.

Sensitivity analysis show that, when errors are normally distributed in Chapter 3, the model-averaged posterior probability decreases as π increases, whereas the opposite pattern is observed when errors follow a Student's t-distribution in Chapter 4. In general, increasing π tends to favour larger mean models. Under normal errors, including more weak terms can increase parameter uncertainty and inflate predictive variance, therefore increasing predictive uncertainty. This effect can be amplified during the optimization procedure since if larger models yield greater predictive variability, the optimal factor levels chosen by the Bayesian model averaging approach may move to regions where the posterior predictive variance is larger or where competing models disagree more strongly, which reduces the model-averaged posterior predictive probability. In contrast, the t distribution is more robust to outliers and heavy-tail behaviour, and moving from smaller to larger mean models can improve the centering of the predictive mean while keeping predictive variation under control. As a result, the additional robustness may lead to an increase in the model-averaged posterior predictive probability as π increases.

In Chapter 5, we develop a generalized g-prior for data with a multistratum structure. Posterior parameter estimates and the posterior model probability are derived. The properties of this prior under different selections of g are investigated. Our results show that the integrated g-priors $\pi_1(g)$, $\pi_2(g)$ and $\pi_3(g)$ consistently exhibit strong power in identifying the true model, whereas the empirical Bayes estimator $\hat{g}_{EB}$ and the fixed g-priors have strong power to identify main effects and interaction effects but less power to identify quadratic

effects. However, when a fixed g-prior is applied, increasing the value of g may improve this power. In addition, changing the value of α may enhance model selection accuracy.

While our methodology demonstrates strong performance, it is subject to certain limitations. First, the approach developed in Chapter 3 relies on the assumption of normally distributed errors. Consequently, it may not be directly applicable to real-world datasets where this assumption is violated.

Second, in Chapter 4, we employ Stan and bridge sampling for each candidate model. Although this approach can handle non-normal errors, it is computationally intensive. Fitting 100 models with 4 chains and 30,000 iterations each requires hours to days of computation, even with parallelization. This makes the method impractical for extremely time-sensitive analyses or for applications requiring real-time feedback. Although the computational cost is justifiable for offline, high-stakes experiments, it remains a barrier for routine use.

Future Work

Multiple response surface optimization is a methodology used to optimize multiple responses simultaneously in a process or system. It involves the construction and optimization of response surface models for response variables, followed by the identification of optimal operating conditions that optimize all responses simultaneously. The posterior probability of a model and the posterior predictive distribution were addressed in Johnson (1987). Peterson (2004) introduced a posterior predictive approach which accounts for both the correlations between

response variables and the model parameter uncertainty. Based on the methodology, Miro-Quesada et al. (2004) studied the optimization problem for robust parameter design, which takes into account the noise variables. Ng (2010) applied the Bayesian model averaging method to the multiple response surface optimization with the consideration of the variability of the predictions. For a review on multiple response optimization using Bayesian predictive approach, see del Castillo and Reis (2020). However, the existing research concentrated only on completely randomized designs. We plan to consider the multiple response surface optimization for multistratum experiments.

We will derive the posterior probability for each model and find the posterior predictive distribution. The Occam's window algorithm will be used to select the competing models based on the posterior probabilities of each model. The objective of the optimization is to find the factor values that lead to a desirable response outcome. The Bayesian model averaging method will be used to account for the model uncertainty. We will illustrate our proposed method through examples and delve into various aspects such as model selection, model robustness analysis, and sensitivity analysis. The study is expected to present several challenges, including the need to derive the posterior predictive function, which involves intricate matrix calculations. Additionally, implementing Bayesian model analysis using R programming poses another hurdle.

Another direction for future work is the application of alternative response optimization criteria within the Bayesian model averaging framework to multistratum data. While this thesis focused on maximizing the predictive probability that a new response falls within a specified interval, other optimization objectives

are common in industrial practice, such as targeting a specific nominal value, minimizing predictive variance, or using a desirability function. Moreover, the literature on bridge sampling for complex hierarchical models is still evolving. Choosing alternative proposal distributions and bridge functions may help establish best practices for estimating marginal likelihoods in multistratum settings, leading to more stable and accurate model selection and, ultimately, more reliable response optimization.

While in Chapter 4, we assumed heavy-tailed t-distributions for the error terms, future work could explore skew-t distributions, mixture models, or Dirichlet process mixtures to handle data with both heavy tails and asymmetries. This would further relax the distributional assumptions and broaden applicability.

Contribution of Authors

Chapters 3, 4, and 5 are based on collaborative research that has been submitted to peer-reviewed venues. The specific contributions of all authors for each of these works are detailed below.

Chapter 3: Response Optimization on Data with a Multistratum Structure Under Normal Assumptions of Random Effects and Errors

Chapter 4: Response Optimization on Data with a Multistratum Structure Under Non-normal Assumptions of Random Effects and Errors

These two chapters are based on a paper co-authored with Dr. Po Yang and Dr. Chang-Yun Lin. As the first author, I was responsible for deriving the algorithm, designing the experimental methodology, implementing data sampling techniques, writing the statistical code, analyzing the results, and drafting the initial manuscript. My supervisor, Dr. Po Yang, provided overall guidance and supervision for the project and revised the manuscript for clarity and impact. Dr. Chang-Yun Lin contributed ideas.

Chapter 5: Model selection using a generalized g-prior for data with a multistratum structure

This chapter draws from a paper co-written with Dr. Po Yang. As the first author, I was responsible for developing the algorithm, crafting the experimental methodology, performing data sampling procedures, coding the statistical analyses, interpreting the results, and preparing the original draft of the manuscript. My supervisor, Dr. Po Yang, offered overarching direction and oversight for the project, and edited the paper to enhance its clarity and impact.

Artificial Intelligence Training Statement

No use for Artificial Intelligence Training

Except as otherwise permitted by any open licensing applied to this work, the author expressly reserves all rights to use of this work for Generative Artificial Intelligence training, the development of machine learning language models, and similar technologies. Where open licensing applies to this work, the author expresses their preference to opt out of the preceding uses.

Bibliography

Abramowitz, M. and I. A. Stegun (Eds.) (1965). *Handbook of Mathematical Functions*. New York: Dover Publications. (Cited on pages 40.)

Aitchison, J. and S. M. Shen (1980). Logistic-normal distributions: Some properties and uses. *Biometrika 67*(2), 261–272. (Cited on pages 42.)

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723. (Cited on pages 3.)

Arnouts, H. and P. Goos (2012). Staggered-level designs for experiments with more than one hard-to-change factor. *Technometrics 54*(4), 355–366. (Cited on pages 2.)

Arnouts, H. and P. Goos (2015). Staggered-level designs for response surface modeling. *Journal of Quality Technology 47*(2), 156–175. (Cited on pages 2, 27, 28, 45 and 97.)

Barnard, G. A. (1963). The logic of least squares. *Journal of the Royal Statistical Society: Series B (Methodological) 25*(1), 124–130. (Cited on pages 5.)

Bové, D. S. and L. Held (2011). Hyper-*g* priors for generalized linear models. *Bayesian Analysis 6*(3), 387–410. (Cited on pages 4 and 89.)

Box, G. E. P., J. S. Hunter, and W. G. Hunter (2005). *Statistics for Experimenters: Design, Innovation, and Discovery* (2 ed.). Hoboken, NJ: John Wiley & Sons. (Cited on pages 1, 24 and 25.)

Box, G. E. P. and R. D. Meyer (1993). Finding the active factors in fractionated screening experiments. *Journal of Quality Technology 25*(2), 94–105. (Cited on pages 13, 17 and 89.)

Brown, P. J., M. Vannucci, and T. Fearn (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 60*(3), 627–641. (Cited on pages 3.)

Carpenter, B., A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell (2017). Stan: A probabilistic programming language. *Journal of Statistical Software 76*(1), 1–32. (Cited on pages 7.)

Chatfield, M. J. (2018). *Uncertainty of variance estimators in analytical and process variability studies*. Ph. D. thesis, University of Southampton, UK. (Cited on pages 66.)

Cortes, L. A., J. R. Simpson, and P. A. Parker (2018). Response surface split-plot designs: A literature review. *Quality and Reliability Engineering International 34*(7), 1374–1389. (Cited on pages 2.)

del Castillo, E. and M. S. Reis (2020). Bayesian predictive optimization of multiple and profile response systems in the process industry: A review and extensions. *Chemometrics and Intelligent Laboratory Systems 206*, 104–121. (Cited on pages 7 and 112.)

Dellaportas, P., J. J. Forster, and I. Ntzoufras (2002). On Bayesian model and variable selection using MCMC. *Statistics and Computing 12*(1), 27–36. (Cited on pages 3.)

Fisher, R. A. (1925). Theory of statistical estimation. *Mathematical Proceedings of the Cambridge Philosophical Society 22*(5), 700–725. (Cited on pages 2 and 24.)

Foster, D. P. and E. I. George (1994). The risk inflation criterion for multiple regression. *The Annals of Statistics 22*, 1947–1975. (Cited on pages 88.)

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin (2013). *Bayesian Data Analysis* (3 ed.). Boca Raton, FL: Chapman & Hall/CRC. (Cited on pages 66 and 68.)

George, E. I. and R. E. McCulloch (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association 88*(423), 881–889. (Cited on pages 3.)

Gilmour, S. G. and P. Goos (2009). Analysis of data from non-orthogonal multistratum designs in industrial experiments. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 58*(4), 467–484. (Cited on pages 62.)

Gilmour, S. G., P. Goos, and H. Großmann (2026). Pure error REML for analyzing data from multi-stratum designs. *Computational Statistics & Data Analysis 218*, 108–322. (Cited on pages 3.)

Goos, P., K. Brickey, and Y. Chen (2024). A blocked staggered-level design for an experiment with two hard-to-change factors. *Journal of Quality Technology 56*(3), 177–194. (Cited on pages 2.)

Goos, P. and M. Vandebroek (2003). D-optimal split-plot designs with given numbers and sizes of whole plots. *Technometrics 45*(3), 235–245. (Cited on pages 101.)

Gronau, Q. R., A. Sarafoglou, D. Matzke, A. Ly, U. Boehm, M. Marsman, D. S. Leslie, J. J. Forster, E.-J. Wagenmakers, and H. Steingroever (2017). A tutorial on bridge sampling. *Journal of Mathematical Psychology 81*, 80–97. (Cited on pages 4, 20 and 43.)

Hanson, T. E., A. J. Branscum, and W. O. Johnson (2014). Informative g-priors for logistic regression. *Bayesian Analysis 9*(3), 597–612. (Cited on pages 4.)

Hoeting, J. A., D. Madigan, A. E. Raftery, and C. T. Volinsky (1999). Bayesian model averaging: A tutorial. *Statistical Science 14*(4), 382–417. (Cited on pages 6 and 69.)

Johnson, M. E. (1987). *Multivariate Statistical Simulation: A Guide to Selecting and Generating Continuous Multivariate Distributions*, Volume 192. New York: John Wiley & Sons. (Cited on pages 111.)

Johnson, N. L., S. Kotz, and N. Balakrishnan (1995). *Continuous Univariate Distributions* (2 ed.), Volume 2. New York: Wiley. (Cited on pages 19 and 40.)

Jones, B. and P. Goos (2007). A candidate-set-free algorithm for generating d-optimal split-plot designs. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 56*(3), 347–364. (Cited on pages 2.)

Jones, B. and P. Goos (2009). D-optimal design of split-split-plot experiments. *Biometrika 96*(1), 67–82. (Cited on pages 2.)

Jones, B. and P. Goos (2012). I-optimal versus d-optimal split-plot response surface designs. *Journal of Quality Technology 44*(2), 85–101. (Cited on pages 2.)

Jones, B. and C. J. Nachtsheim (2009). Split-plot designs: What, why, and how. *Journal of Quality Technology 41*(4), 340–361. (Cited on pages 2.)

Kang, G. and S. Jeong (2024). Model selection-based estimation for generalized additive models using mixtures of g-priors: Towards systematization. *Bayesian Analysis 41*, 1–31. Advance publication. (Cited on pages 5.)

Kass, R. E. and L. Wasserman (1995). A reference bayesian test for nested hypotheses and its relationship to the schwarz criterion. *Journal of the American Statistical Association 90*, 928–934. (Cited on pages 88.)

Khuri, A. I. and S. Mukhopadhyay (2010). Response surface methodology. *Wiley Interdisciplinary Reviews: Computational Statistics 2*(2), 128–149. (Cited on pages 5.)

Letsinger, J. D., R. H. Myers, and M. Lentner (1996). Response surface methods for bi-randomization structures. *Journal of Quality Technology 28*(4), 381–397. (Cited on pages 2.)

Li, Y. and M. A. Clyde (2018). Mixtures of g-priors in generalized linear models. *Journal of the American Statistical Association 113*(524), 1828–1845. (Cited on pages 4.)

Liang, F., R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger (2008). Mixtures of g-priors for bayesian variable selection. *Journal of the American Statistical Association 103*(481), 410–423. (Cited on pages 4, 88 and 89.)

Lin, C.-Y. (2018). Generalized bayesian d-criterion for single-stratum and multistratum designs. *Quality and Reliability Engineering International 34*(9), 1566–1574. (Cited on pages 82.)

Lin, C.-Y. and P. Yang (2019). Data-driven multistratum designs with the generalized bayesian *dd* criterion for highly uncertain models. *Computational Statistics & Data Analysis 138*, 222–238. (Cited on pages 3.)

Lin, C.-Y. and P. Yang (2022). Bayesian analysis and follow-up experiments for supersaturated multistratum designs. *Journal of Quality Technology 54*(5), 527–546. (Cited on pages 3, 40, 41, 86 and 87.)

Madigan, D. and A. E. Raftery (1994). Model selection and accounting for model uncertainty in graphical models using occam's window. *Journal of the American Statistical Association 89*(428), 1535–1546. (Cited on pages 3 and 5.)

Marin, J.-M. and C. P. Robert (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. New York: Springer. (Cited on pages 89.)

Meng, X.-L. and W. H. Wong (1996). Simulating ratios of normalizing constants via a simple identity: A theoretical exploration. *Statistica Sinica 6*, 831–860. (Cited on pages 4, 21 and 43.)

Miro-Quesada, G., E. del Castillo, and J. P. Peterson (2004). A bayesian approach for multiple response surface optimization in the presence of noise variables. *Journal of Applied Statistics 31*(3), 251–270. (Cited on pages 6 and 112.)

Montgomery, D. C. (2017). *Design and Analysis of Experiments* (9 ed.). Hoboken, NJ: John Wiley & Sons. (Cited on pages 2.)

Myers, R. H., D. C. Montgomery, and C. M. Anderson-Cook (2016). *Response Surface Methodology: Process and Product Optimization Using Designed Experiments* (4 ed.). Hoboken, NJ: John Wiley & Sons. (Cited on pages 5.)

Mylona, K., P. Goos, and B. Jones (2014). Optimal design of blocked and split-plot experiments for fixed effects and variance component estimation. *Technometrics 56*(2), 132–144. (Cited on pages 2.)

Ng, S. H. (2010). A bayesian model-averaging approach for multiple-response optimization. *Journal of Quality Technology 42*(1), 52–68. (Cited on pages 6 and 112.)

O'Hagan, A. (1979). On outlier rejection phenomena in bayes inference. *Journal of the Royal Statistical Society: Series B (Methodological) 41*(3), 358–367. (Cited on pages 68.)

Parker, P. A., S. M. Kowalski, and G. G. Vining (2007). Construction of balanced equivalent estimation second-order split-plot designs. *Technometrics 49*(1), 56–65. (Cited on pages 2.)

Peterson, J. J. (2004). A bayesian reliability approach to multiple response surface optimization. *Journal of Quality Technology 36*(2), 139–153. (Cited on pages 6 and 111.)

Peterson, J. J. (2023). Response surfaces, blocking and split-plots: A predictive distribution case study. *Quality Engineering 35*(1), 172–191. (Cited on pages 7, 66 and 67.)

Press, S. J. (1982). *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference* (2 ed.). Malabar, FL: Robert E. Krieger. (Cited on pages 5, 17 and 38.)

Rajagopal, R. and E. del Castillo (2005). Model-robust process optimization using bayesian model averaging. *Technometrics 47*(2), 152–163. (Cited on pages 6.)

Rajagopal, R., E. del Castillo, and J. J. Peterson (2005). Model and distribution-robust process optimization with noise factors. *Journal of Quality Technology 37*(3), 210–222. Corrigendum: Journal of Quality Technology, 38(1), 83–84. (Cited on pages 6.)

Roberts, H. V. (1965). Probabilistic prediction. *Journal of the American Statistical Association 60*(309), 50–62. (Cited on pages 5.)

Robinson, T. J., C. M. Anderson-Cook, and M. S. Hamada (2009). Bayesian analysis of split-plot experiments with nonnormal responses for evaluating nonstandard performance criteria. *Technometrics 51*(1), 56–65. (Cited on pages 6.)

Robinson, T. J., A. L. Pintar, C. M. Anderson-Cook, and M. S. Hamada (2012). A bayesian approach to the analysis of split-plot combined and product arrays and optimization in robust parameter design. *Journal of Quality Technology 44*(4), 304–320. (Cited on pages 6.)

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464. (Cited on pages 3.)

Stan Development Team (2020). *RStan: The R interface to Stan*. Stan Development Team. R package version 2.21.2, `http://mc-stan.org/`. (Cited on pages 7.)

Trinca, L. A. and S. G. Gilmour (2001). Multistratum response surface designs. *Technometrics 43*(1), 25–33. (Cited on pages 2.)

Trinca, L. A. and S. G. Gilmour (2015). Improved split-plot and multistratum designs. *Technometrics 57*(2), 145–154. (Cited on pages 3.)

Vining, G. G., S. M. Kowalski, and D. C. Montgomery (2005). Response surface designs within a split-plot structure. *Journal of Quality Technology 37*(2), 115–129. (Cited on pages 90.)

Wu, C. F. J. and H. Zhang (2007). Sequential response surface methodology:

A review. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(4), 589–611. (Cited on pages 5.)

Xiao, Y., P. Yang, and N. Zou (2025). Bayesian sequential *i*-optimal designs for split-plot experiments under model uncertainty. *Journal of Quality Technology 57*(5), 440–459. (Cited on pages 2 and 90.)

Zellner, A. (1986). On assessing prior distributions and bayesian regression analysis with *g*-prior distributions. In P. K. Goel and A. Zellner (Eds.), *Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*, pp. 233–243. Amsterdam: North-Holland. (Cited on pages 4, 14 and 80.)