UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES

Mathematical Sciences

Design of Factorial Experiments in Blocks and Stages

by

Emily Sarah Matthews

Thesis submitted for the degree of Doctor of Philosophy

July 2015

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL, HUMAN AND MATHEMATICAL SCIENCES

Mathematical Sciences

Doctor of Philosophy

DESIGN OF FACTORIAL EXPERIMENTS IN BLOCKS AND STAGES

by Emily Sarah Matthews

Factorial experiments are increasingly important in science and industry. In this thesis, we consider two types of factorial experiments; block designs with autocorrelated errors and multi-stage designs.

The design of D-optimal blocked experiments with autocorrelated errors, motivated by the manufacture of microstructured optical fibres, is discussed in Chapter 2. Autocorrelated errors extend the standard error assumptions for block designs to account for temporal or spatial ordering of the experimental units.

We found that D-optimal blocks designs with autocorrelated errors are, under specific model assumptions, robust to the misspecification of two unknown parameters, the autocorrelation parameter and the ratio of variance components. We also noted that these robust designs do not require specific treatment allocation or ordering.

In multi-stage experiments, treatments are applied to the same experimental unit at different stages, and responses are measured at the end of each stage. In Chapter 3, a compound Bayesian D-optimality objective function is used within a coordinate exchange algorithm to construct multi-stage factorial designs with different levels of restrictions on randomisation. In a multi-stage design, treatments are applied to the same experimental unit at different stages and responses are measured at the end of each stage. Comparison of efficiencies and the size and number of correlated columns demonstrated the limitations of a one number optimisation approach.

Chapter 4 demonstrates that frequentist variable selection methods for multi-stage split-plot designs rely on unreliable parameter estimates and highlights the benefits of Bayesian variable selection, which accounts for the uncertainty associated with unknown parameters. A Metropolis-Hastings within Gibbs sampling algorithm for the analysis of multivariate responses from supersaturated split-plot designs is presented in Chapter 4. The methodology in Chapters 3 and 4 is applied to the formulation and dissolution testing of a pharmaceutical product in Chapter 5.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Emily Sarah Matthews, declare that the thesis entitled "Design of Factorial Experiments in Blocks and Stages" and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- none of this work has been published before submission.

Signed:

Date:

Acknowledgements

Firstly I would like to thank my supervisor, Prof. David Woods, for his guidance and support throughout. I would also like to thank our collaborators at the Optoelectronics Research Centre (ORC) and GlaxoSmithKline (GSK) for their input, with particular thanks to Marco Petrovich and Natalie Wheeler at the ORC and Martin Owen, Mohammed Yahyah, Verity Fisher, Matthew Burke and Nena Mistry at GSK. This work was supported by studentship funding from the UK Engineering and Physical Sciences Research Council, Southampton Statistical Sciences Research Institute and Mathematical Sciences at the University of Southampton.

The support of my husband, Sean, and my family, Karen, Colin, Louise and Rob, has been invaluable, and I cannot thank them enough for their continued encouragement. I would also like to thank all the wonderful people I have had the pleasure of getting to know during my studies, with a special thank you to Maria, Sean and Izi for all they have done for me.

Chapter 1

Introduction

Factorial designed experiments where runs can be grouped with respect to some experimental feature or where treatments are applied in stages to the same experimental units are ubiquitous in industry. This thesis develops methodology for the design and analysis of factorial experiments with blocks and stages, and discusses the application of this methodology to two motivating examples from optoelectronics and chemistry.

This chapter defines factorial experiments with restricted randomisation and multiple stages (Section 1.1), discusses the motivating examples (Section 1.2), introduces the models used to analyse results from factorial experiments with restricted randomisation (Section 1.3), introduces the methods used to find optimal designs for these factorial experiments (Section 1.4), and gives a brief summary of the contents of the thesis (Section 1.5).

1.1 Introduction to Designed Factorial Experiments

Experimentation is widespread and not just limited to the areas where the term experiment is widely used, such as science and engineering. An experiment is defined as the process from which data are collected to answer a question of interest, where variables are controlled or changed in order to introduce variability in the response. The variables in an experiment that can be controlled or changed by the experimenter are referred to as factors, and factors can have different settings, or levels, such as high or low. Factor levels are often coded, with -1 used to denote the low level, 0 used to denote a middle level, and 1 used to denote the high level.

In experiments, a treatment, which is a combination of factor levels, is applied to an experimental unit, which is the subject or subdivision of material used in the experiment. The application of a treatment to an experimental unit is an experimental run. For example, if an experiment has six two-level factors, then one particular factorial treatment is $(1, 1, -1, 1, -1, 1)$.

Experiments can be used to gain understanding or improve a product or process through the comparison of treatments and the assessment of the impact of treatments on the experimental outcome, which is referred to as the response. The individual effect of a factor on the response, known as the main effect, is the difference in the average response when the level of that factor is changed. The estimated main effect of a two-level factor is

$$\bar{y}_H - \bar{y}_L, \tag{1.1}$$

where $\bar{y}_H$ is the average response when the factor level is 1 and $\bar{y}_L$ is the average response when the factor level is -1.

The level of a more than one factor can be altered in consecutive treatments in a factorial design. A factorial experiment allows the joint effects of two (or more) factors on the response, known as interactions, to be compared. Estimation of interactions is unique to factorial experiments as they require outputs from experimental runs where more than one factor level is changed between each treatment. The estimated two-factor interaction is defined as

$$\frac{1}{2}\left[(\bar{y}_{HH} - \bar{y}_{HL}) - (\bar{y}_{LH} - \bar{y}_{LL})\right], \tag{1.2}$$

where $\bar{y}_{HH}$ is the average response when the level of both factors is 1, $\bar{y}_{HL}$ and $\bar{y}_{LH}$ is the average response when the factor levels are -1 and 1, and $\bar{y}_{LL}$ is the average response when the level of both factors is -1. Higher order factorial effects can be defined similarly.

As most processes depend on several factors, the assessment of the joint effect of factors is important, hence factorial experiments are often used to gain understanding of, and improve, scientific and technological processes.The work in this thesis is motivated by two such processes: the manufacture of micro-structure optical fibres, and the formulation and dissolution testing of a pharmaceutical product, which are discussed in Sections 1.2.1 and 1.2.2, respectively.

An experiment can be designed to meet quantitative and qualitative objectives based on experimental resources and restrictions. Cox (1958) sets out five features of a good designed experiment; the absence of the systematic errors, intrinsic stability of the experimental material or process, suitable factor levels, the ability to be run without excessive difficulty or cost, and the ability to assess the uncertainty associated with the experiment. We have considered these five features when designing experiments in this thesis. Restrictions on the randomisation of the treatments affects both the difficulty or cost of running the experiment and the assessment of uncertainty. These points are discussed further in Sections 1.1.1 and 1.3.1, respectively.

The design matrix for a factorial experiment with f factors is the $n \times f$ matrix of treatments applied to experimental units in the n runs of the experiment. Throughout this thesis, we use $\mathcal{D}_{f,l,n}$ to define the set of all possible n run designs for f l-level factors, where the order of the rows is unimportant and there can be replicated treatments in the design.

A full factorial design $\mathbf{D}_{f,l,n}$ contains all l^f unique treatments for f l-level factors and is an element of $\mathcal{D}_{f,l,n}$. For example,

$$\mathbf{D}_{2,2,4} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \\ -1 & 1 \\ -1 & -1 \end{pmatrix} \in \mathcal{D}_{2,2,4}.$$

The parameters in the model, such as the linear mixed effects model discussed in Section 1.3, assumed for the responses from a full factorial experiment can be estimated independently. However, in this thesis we also consider fractional factorial designs, which can be regular or non-regular.

A fractional factorial design contains a subset of the l^f unique treatments in the full factorial design $\mathbf{D}_{f,l,n}$. The definition of regular and non-regular fractional factorial designs given by Wu and Hamada (2009) is dependent on the aliasing between parameter estimates in the model assumed for responses from the designed experiment. If two parameters are:

- *not aliased*, their influence on the response from the experiment can be estimated independently,
- *fully aliased*, their influence on the response from the experiment cannot be separated,
- *partially aliased*, their influence on the response is related, but some information regarding the influence of both parameters on the response can be estimated.

A fractional factorial design is:

- *regular* if each pair of factorial effects can either be estimated independently of each other or are fully aliased,
- *non-regular* if one or more pairs of factorial effects are partially aliased.

A number of the experiments in this work are non-regular fractional factorial designs with restricted randomisation.

1.1.1 Restricted Randomisation and Blocking

An important principle of the statistical design of experiments is randomisation of the order in which treatments are applied in the experiment. Randomisation is used to reduce the impact of uncontrolled differences between experimental units on the analysis of the experiment. However, complete randomisation of the runs is not possible for some experiments, for example due to physical constraints, and hence a restricted randomisation must be applied.

Designs with restricted randomisation are particularly common in industry, as there are often constraints on the run order due to restrictions on experimentation, such as how often a factor level can be changed, or the number of runs that can be performed each day. The aim of an optimal design with restricted randomisation is to minimise the impact of the restrictions on the precision and accuracy of the experimental output.

Block designs, which are common in industry and the focus of Chapter 2, have restricted randomisation. In a block design, experimental units that are anticipated to give similar responses if the same treatment is applied are grouped together into a set called a block. For example, the experimental units from the same batch of raw material or the units whose treatments are applied by the same technician could be grouped together.

Treatments in a block design are allocated within blocks and cannot be swapped between blocks without changing the properties of the design. However the order of these blocks and the allocation of treatments to units within a block can be randomised. Further information regarding block designs can be found in Section 2.2 of Chapter 2 and in literature such as Goos and Vandebroek (2001) and Goos (2002).

Another example of designs with restricted randomisation are split-plot designs, which group runs based on the levels of particular factor and are studied in Chapters 3 to 5 of this thesis. Split-plot experiments have two types of factors, whole- and sub-plot factors. The experimental runs are organised into whole-plots according to the levels of whole-plot factors, which are difficult or costly to change in the experiment. A common example of a whole-plot factor is temperature, as it is often difficult to quickly adjust the temperature of experimental equipment such as furnaces. The other factors in the experiment are referred to as sub-plot factors, and these can be varied as often as required in the experiment.

Treatments in split-plot designs are allocated to whole-plots based on the levels of the whole-plot factors and therefore cannot be swapped between whole-plots. However the order of the whole-plots and the allocation of levels of sub-plot factors within whole-plots can be randomised. Further information on split-plot designs can be found in Section 3.2.2 of Chapter 3 and in literature such as Box and Jones (1992), Miller (1997) and Goos and Gilmour (2012).

The presence of groups or blocks in an experiment needs to be considered when designing and analysing experiments, even though their influence is not usually of interest, as ignoring them can bias the analysis obtained from the experiment and reduce the precision of conclusions made from the experiment (for example through loss of power in hypothesis tests). The example in Section 3.2.4 in Chapter 3 demonstrates the impact of restrictions on randomisation on the estimation of model parameters.

1.1.2 Multi-stage Designs

In literature such as Freeman (1959), Trinca and Gilmour (2001) and Brien et al. (2011), multi-stage experiments are described as experiments that are conducted in distinct time intervals. In this thesis, we define a multi-stage experiment as an experiment that uses the same experimental unit in multiple stages; with different sets of treatments applied, and a different response recorded, at each stage. Stages in our definition can refer to time intervals or distinct manufacturing processes.

While each stage has a separate response, the experiment is designed for factors from all stages as it is assumed that the responses observed at Stage $s+1$ is affected by the factors in all the previous stages, Stages $1, \ldots, s$, as well as the factors in Stage $s+1$. This is similar to the definition of partition designs given by Perry et al. (2001, 2002) and Perry et al. (2007), which is discussed in Section 3.3 of Chapter 3.

1.2 Motivating Examples

In this section we discuss the two examples motivating the work in this thesis: the manufacture of microstructured optical fibres, and the formulation and dissolution testing of a pharmaceutical product. These examples result from collaborations with the Optoelectronics Research Centre (ORC) at the University of Southampton and GlaxoSmithKline (GSK), respectively.

1.2.1 Manufacture of Microstructured Optical Fibres

Microstructured optical fibres are popular in current optoelectronics research. Microstructured optical fibres have a hollow core. Light can be transmitted through a hollow core faster than it can through a core of glass capillaries, which is used in traditional optical fibres. Hence microstructured optical fibres outperform traditional fibres with respect to data transmission, and the hollow core in microstructured optical fibres also makes them useful for other applications such as gas sensing. The manufacture and properties of microstructured fibres is discussed by Poletti et al. (2013).

The manufacture of microstructured optical fibres requires two processes. These processes are performed in distinct, but ordered, time intervals.

- *Process 1 - cane manufacture*: A stack of thin glass capillaries with a hollow core encased in a larger jacket of glass, called a preform, is drawn (heated in a furnace and then stretched) into a cane. A one metre preform becomes a cane of approximately ten metres. The cane is cleaned to remove any impurities in the glass that may cause problems in Process 2.
- *Process 2 - fibre manufacture*: The cane from Process 1 is drawn into a fibre. One metre of cane can be drawn into 3 to 20 kilometres of fibre, depending on the stability of the process and the thickness of the fibre.

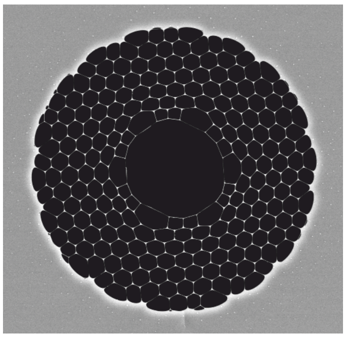


Figure 1.1: Cross section of a microstructured optical fibre, showing the hollow core and the surrounding glass capillaries.

Figure 1.1 is a cross section of a microstructured fibre. The large black circle in the centre of this cross section is the hollow core. The smaller circles, which are arranged around the hollow core, are the glass capillaries which have been stretched during Process 1 and 2. The grey background is the glass jacket, which has been drawn in Process 1 and 2 (not fully pictured).

The temperature at which the cane and the preform are drawn, the speed at which the preform and cane is fed into the machine which draws the cane and fibre (respectively), the speed at which the cane and preform are drawn, and the pressure in the core of the cane and preform can be varied. If certain factor ranges are chosen for the feed rate, draw speed and pressure, then the temperature will not be varied in order to ensure the experiment is stable.

Pressure in the hollow core is the key factor when manufacturing microstructured optical fibres. The pressure ensures the hollow core of the cane is expanded in the preform without causing the external structure to collapse, and also maintains the hollow core as the preform is drawn into a fibre.

The quality of the cane can be assessed after Process 1 by examining the structure of a section of the cane under a microscope. The quality of the cane can also be assessed by taking an X-ray of the cane, which is a non-destructive method suggested by Sandoghci et al. (2014). The transmission of light through the fibre is the output from Process 2.

As the manufacture of microstructured fibres relies on two processes, and distinct responses are measured at the end of each process, this experiment could be described as a two-stage split-plot design, as discussed in Chapter 3, where the cane manufacture factors are whole-plot factors.

If the factors used for manufacturing the cane are not thought to have a significant impact on the response, the cane used in fibre manufacture could be a blocking factor. This initial simplification of the process motivates the work in Chapter 2. It is appropriate to assume that lengths of fibres which are drawn after each other are more similar than those drawn further apart, where the relationship between responses for two lengths of fibres diminishes as the distance between them increases. Block designs suitable for this relationship between lengths of fibres are considered in Chapter 2.

1.2.2 Formulation and Dissolution Testing of a Pharmaceutical Product

The active pharmaceutical ingredient (API) is the chemical compound in a pharmaceutical product which treats a specific ailment or disease. The formulation of a pharmaceutical product containing a specific API is the focus of the work in Chapters 3 to 5. This formulation requires the application of six two-level factors to an experimental unit across two-stages; five factors in Stage 1 and one factor in Stage 2, with a response at the end of each stage.

The factors applied in Stage 1 are assumed to have an impact on the response from Stage 2. If the first two Stage 1 factors are varied too often, the cost of the experiment will be unacceptable. However the output from Stage 1 may be able to be re-batched, and hence two-stage split- and strip-plot designs (see Arnouts et al., 2010 and Section 3.2.3 in Chapter 3 for further detail on strip-plot designs) are discussed in Chapter 3.

Dissolution testing is used by pharmaceutical companies researching in drug development to assess the performance of a pharmaceutical product. The measurement of the rate of dissolution of a pharmaceutical product in different pH media is representative of the dissolution of the product in different parts of the human body. Different pharmaceutical products have different dissolution testing requirements, depending on where the product is supposed to dissolve and enter the bloodstream. Dissolution testing is therefore destructive, as the product cannot be reformed once it is subjected to the different media.

The second stage response, which is measured once all six factors are applied, is the results from dissolution testing of the pharmaceutical product. The analysis of the responses from dissolution testing is a regulatory requirement and is therefore an area of particular interest for pharmaceutical companies.

The identification of factors that influence dissolution testing, and settings of these factors that have a high probability of meeting specification required for responses from dissolution testing are also important as they help optimise the formulation process and identify new areas of the design space for experimentation. Therefore the analysis of this example motivates the work in Chapter 4, and the design and analysis of this specific experiment is discussed in Chapter 5.

1.3 Introduction to Linear Mixed Effect Models

1.3.1 Linear Mixed Effect Models

Once an output is measured from a factorial experiment with restricted randomisation, a linear mixed model can be used to approximate the relationship between the factors and the response. Linear mixed effect models are commonplace in literature for experiments with restricted randomisation, see for example Morris (2011, Chapter 10), as they allow for the correlation between the responses from units in the same block or whole-plot. Responses from treatments in the same block or grouping are assumed to be more similar than those in different groups, and this relationship is modelled in the correlation structure for the response.

Linear mixed effect models allow relationships between the response and the factors involving main effects, interactions and polynomial terms to be fitted. These models are suitable for the motivating examples considered in this thesis. However other more complex, non-linear, relationships, as discussed in Davidian and Giltinan (1995), may be more suitable for other experiments, dependent on the nature of the responses.

A linear mixed model for responses from a designed factorial experiment with restrictions on randomisation with n runs organised into b groups is given by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{1.3}$$

where $\mathbf{Y}$ is the $n \times 1$ vector of responses, $\mathbf{X}$ is the $n \times p$ model matrix, p is the number of parameters in the model, $\boldsymbol{\beta}$ is the $p \times 1$ vector of unknown parameters, $\mathbf{Z}$ is a $n \times b$ matrix showing the assignment of runs to particular groups, $\boldsymbol{\gamma}$ is a $b \times 1$ vector of random group effects and $\boldsymbol{\epsilon}$ is the $n \times 1$ vector of random errors with mean $\mathbf{0}$ and variance-covariance matrix $\boldsymbol{\Sigma}$.

Let the ith, $i = 1, \ldots, n$, design point in the design matrix, $\mathbf{D}$, be $\mathbf{x}_i$. Then the ith row of the model matrix, $\mathbf{X}$, is $f(\mathbf{x}_i)$ where f is the function that gives the model expansion of a treatment $\mathbf{x}_i$. Assume, for example, that an experiment has two factors and a model containing the intercept, the linear effects for the two factors and the product of the two factors is fitted, then $\boldsymbol{\beta}$=(β_0, β_1, β_2, β_{12}) and $f(\mathbf{x}_i)$=$(1, x_1, x_2, x_{12})$=(1, 1,−1, −1) when $\mathbf{x}_i$=(1,−1). When the factor levels in $\mathbf{D}$ are coded as $-1, 1$, then the main effect (1.1) of factor i is given by $2\beta_i$ and the interaction (1.2) between factors i and j is $2\beta_{ij}$.

The (i, j)th entry, $i = 1, \ldots, n$, $j = 1, \ldots, b$, of $\mathbf{Z}$, z_{ij}, indicates which group the ith run of the experiment belongs to. If the ith run of the experiment is in the jth group then $z_{ij} = 1$, otherwise $z_{ij} = 0$. We assume the groups in the experiment are representative of a wider population of groups, which implies that $\boldsymbol{\gamma}$ is a random effect and enables predictions to be made using the results from this experiment.

It is usually assumed that both $\boldsymbol{\gamma}$ and $\boldsymbol{\epsilon}$ are independently normally distributed with mean $\mathbf{0}$ and variance-covariance matrices $\sigma_\gamma^2\mathbf{I}_b$ and $\sigma_\epsilon^2\mathbf{I}_n$ respectively. It is also usually assumed that there is more variability in responses from different groups than between responses in the same groups, as units within the same group are assumed to be more similar than those in different groups. This implies that the inter (between) group variance σ_γ^2 is larger than the intra (within) group variance σ_ϵ^2. As both $\boldsymbol{\gamma}$ and $\boldsymbol{\epsilon}$ are normally distributed, $\mathbf{Y}$ is a normally distributed random variable with mean

$$\begin{aligned} E(\mathbf{Y}) &= E(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}) \\ &= E(\mathbf{X}\boldsymbol{\beta}) + E(\mathbf{Z}\boldsymbol{\gamma}) + E(\boldsymbol{\epsilon}) \\ &= \mathbf{X}\boldsymbol{\beta}, \end{aligned} \tag{1.4}$$

and variance-covariance matrix

$$\begin{aligned} \mathbf{V} &= \text{var}(\mathbf{Y}) = \text{var}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\gamma} + \boldsymbol{\epsilon}) \\ &= \text{var}(\mathbf{Z}\boldsymbol{\gamma}) + \text{var}(\boldsymbol{\epsilon}) \\ &= \mathbf{Z}\text{var}(\boldsymbol{\gamma})\mathbf{Z}^T + \sigma_\epsilon^2\mathbf{I}_n \\ &= \mathbf{Z}(\sigma_\gamma^2)\mathbf{Z}^T + \sigma_\epsilon^2\mathbf{I}_n \\ &= \sigma_\gamma^2\mathbf{Z}\mathbf{Z}^T + \sigma_\epsilon^2\mathbf{I}_n. \end{aligned} \tag{1.5}$$

Note that even though the random terms in the model (1.3), $\boldsymbol{\gamma}$ and $\boldsymbol{\epsilon}$ have no correlation structure, the responses from the same group have a correlation structure defined through $\mathbf{Z}\mathbf{Z}^T$ in (1.5). In Chapter 2 we adjust the assumption that var($\boldsymbol{\epsilon}$)= $\sigma_\epsilon^2\mathbf{I}_n$ to allow for a dependency between experimental units in the same group, which also affects the structure of (1.5); see Section 2.3 for further detail.

The vector of group effects, $\boldsymbol{\gamma}$, is often not of interest to the experimenter, but is included in the model to account for the presence of the groups. Ignoring this grouping factor could introduce bias into the experiment. Therefore, as $\boldsymbol{\gamma}$ does not have to be estimated independently, we could marginalise (1.3) and include the group terms in a new error term $\boldsymbol{\epsilon}^*$, to get the linear model

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}^*, \tag{1.6}$$

where $\boldsymbol{\epsilon}^*$ is the $n \times 1$ vector of normally distributed random errors with mean $\mathbf{0}$ and variance-covariance matrix (1.5). If there are no restrictions on randomisation then the n runs could be described as coming from n individual groups, hence $\mathbf{Z} = \mathbf{I}_n$, and responses from these designed experiments, which are referred to as completely randomised experiments, can be modelled using (1.6) where (1.5) is $(\sigma_\gamma^2 + \sigma_\epsilon^2)\mathbf{I}_n$.

The number of rows, n, and columns, p, of the model matrix $\mathbf{X}$, influences our ability to estimate components of the variance-covariance matrix (1.5), as $n - p$ is the degrees of freedom with which we can estimate the variance of the responses $\mathbf{Y}$. If $p = n$, the designed experiment from which the responses are measured, $\mathbf{D}$, is saturated and there are no degrees of freedom with which to estimate the variance components of the responses. If $p > n$ then the design is supersaturated, and the variance components cannot be reliably estimated. We discuss the properties of non-saturated and saturated designs in Chapter 2 and design and analyse supersaturated designs in Chapters 3 through 5.

1.3.2 Generalised Least Squares and Maximum Likelihood

The estimate of the fixed model parameter $\boldsymbol{\beta}$ in (1.6) is often of interest to experimenters, as these parameters indicate the individual and joint effects of the factors in the experiment. To find the estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ for (1.6) we minimise the function

$$(\mathbf{y} - \mathbf{X}\boldsymbol{\beta})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{X}\boldsymbol{\beta}) = \mathbf{y}^T\mathbf{V}^{-1}\mathbf{y} - 2\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} + \boldsymbol{\beta}^T\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}, \tag{1.7}$$

which is equivalent to finding $\hat{\boldsymbol{\beta}}$ such that

$$-\mathbf{X}^T\mathbf{V}^{-1}\mathbf{y} + (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})\hat{\boldsymbol{\beta}} = 0. \tag{1.8}$$

Hence the generalised least squares (GLS) estimator of $\boldsymbol{\beta}$ is given by

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y}. \tag{1.9}$$

The variance of this estimator is

$$\begin{aligned}
\mathrm{var}(\hat{\boldsymbol{\beta}}) &= \mathrm{var}((\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{Y}) \\
&= (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathrm{var}(\mathbf{Y})[(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}]^T \\
&= (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{V}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} \\
&= (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1} \\
&= (\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}.
\end{aligned} \tag{1.10}$$

The generalised least square estimator of $\boldsymbol{\beta}$ is unbiased, that is $E(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, and is equivalent to the maximum likelihood estimator. The likelihood is the joint probability density function for the responses considered as a function of the model parameters. Therefore the likelihood for $\mathbf{Y}$ in (1.6)

$$L(\boldsymbol{\beta}|\mathbf{Y}) = (2\pi)^{-n/2}|\mathbf{V}|^{-1/2} \exp\left[-\frac{1}{2}\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right)^T \mathbf{V}^{-1}\left(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\right)\right], \tag{1.11}$$

as $\mathbf{Y} \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \mathbf{V})$. The maximum likelihood estimator (MLE) is the estimator that maximises the likelihood, which is equivalent to finding the $\hat{\boldsymbol{\beta}}$ such that

$$\frac{\partial}{\partial \boldsymbol{\beta}} L(\hat{\boldsymbol{\beta}}|\mathbf{Y}) = \mathbf{0}, \tag{1.12}$$

which is equivalent to

$$\frac{\partial}{\partial \boldsymbol{\beta}} \ln L(\hat{\boldsymbol{\beta}}|\mathbf{Y}) = \mathbf{0}, \tag{1.13}$$

where $\ln L(\boldsymbol{\beta}|\mathbf{Y})$ is referred to as the log likelihood. As (1.7) is proportional to the log of (1.11), the GLS estimator (1.9) is therefore the solution to (1.12) and (1.13).

The variance of the MLE for the linear model is the inverse of the Fisher information matrix,

$$\mathcal{I}(\boldsymbol{\beta}) = E\left[\left(\frac{\partial^2}{\partial \boldsymbol{\beta}^2} \ln f(\boldsymbol{\beta}|\mathbf{Y})\right) |\boldsymbol{\beta}\right]. \tag{1.14}$$

evaluated at $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}$. The second derivative with respect to $\boldsymbol{\beta}$ of the log of (1.11) is $(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})$ and therefore (1.10) is the inverse of $\mathcal{I}(\boldsymbol{\beta})$.

The parameter estimators and the variance of the estimators depend on the model matrix $\mathbf{X}$ and the variance-covariance matrix $\mathbf{V}$, which is assumed to be known. Usually, when designing experiments, prior knowledge about the variance components is assumed. Once responses have been observed, $\mathbf{V}$ can be estimated by either maximising the likelihood (1.11) with respect to $\mathbf{V}$ or by using restricted maximum likelihood (REML, Patterson and Thompson, 1971; Harville, 1977). REML requires the transformation of the response to remove the influence of the other model parameters followed by the maximisation of the likelihood for these transformed responses.

Ordinary least squares (OLS) is a special case of GLS, where the responses $\mathbf{Y}$ are assumed to have equal variances and zero covariances. Hence, as discussed in Section 1.3.1, $\mathbf{V} = (\sigma_\gamma^2 + \sigma_\epsilon^2)\mathbf{I}_n$. The OLS estimator of $\boldsymbol{\beta}$,

$$\hat{\boldsymbol{\beta}} = (\sigma_\gamma^2 + \sigma_\epsilon^2)^{-1}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}, \tag{1.15}$$

is also unbiased and identical to the MLE. The variance of the OLS estimator is

$$\text{var}(\hat{\boldsymbol{\beta}}) = (\sigma_\gamma^2 + \sigma_\epsilon^2)^{-1}(\mathbf{X}^T\mathbf{X})^{-1}. \tag{1.16}$$

1.4 Design Optimality and Selection

The dependence of the variance-covariance matrix of $\hat{\boldsymbol{\beta}}$ on the model matrix $\mathbf{X}$ motivates a number of the design selection criteria. The variance of this estimator is the inverse of the Fisher information matrix,

$$\mathcal{I}(\boldsymbol{\beta}) = \mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}. \tag{1.17}$$

The aim of many experiments is to minimise the variability of the estimators. Optimality criteria and associated objective functions are discussed in Sections 1.4.1 and 1.4.2.

We aim to find designs which optimise an objective function over the set of possible designs $\mathcal{D}_{f,l,n}$. However, it can be computationally difficult or impossible to consider all designs as $|\mathcal{D}_{f,l,n}|$ can be large even when f and l are relatively small; for example $|\mathcal{D}_{3,2,6}| = 262144$. Therefore, the design selection algorithms discussed in Sections 1.4.3 and 1.4.4 do not consider every possible design, and hence can find designs which do not necessarily maximise the objective function over the whole of $\mathcal{D}_{f,l,n}$.

We present two different algorithms for design selection in this section, the coordinate exchange algorithm (Meyer and Nachtsheim (1995), Section 1.4.3) and interchange algorithm (Atkinson et al. (2007, Chapter 12), Section 1.4.4). The coordinate exchange algorithm swaps each factor level for each treatment in the design iteratively and the interchange algorithm swaps the position of treatments in a design. Both methods calculate improvements in a specific objective function. The coordinate exchange algorithm is used to find optimal designs in Chapters 2, 3 and 5, and the interchange algorithm is used to find optimal designs in Chapter 2.

These algorithms are referred to as 'greedy' algorithms, as they only accept moves which improve the value of the objective function. This could potentially lead to finding local, rather than, global maxima, although running the coordinate exchange algorithm for multiple random starts and running the interchange algorithm for multiple starting row swaps attempts to combat this problem. Stochastic algorithms such as simulated annealing (Aarts and van Laarhoven, 1989; Brooks and Morgan, 1995) accept suggested regressive changes to a design with a certain probability, and therefore allow sub-optimal moves to be made to escape local optima.

1.4.1 Design Optimality Criteria

A D-optimal design minimises the volume of the confidence ellipsoid for $\boldsymbol{\beta}$, where a confidence ellipsoid is an p dimensional extension of a confidence interval. This volume is inversely proportional to the determinant of the information matrix (1.17), and hence the objective function for D-optimality is

$$\phi_D = |\mathcal{I}(\boldsymbol{\beta})|. \tag{1.18}$$

A design $\mathbf{D}^*$ such that

$$\mathbf{D}^* = \underset{\mathbf{D} \in \mathcal{D}_{f,l,n}}{\arg\max}\, \phi_D \tag{1.19}$$

is a D-optimal design.

In Chapters 2 and 3 we compare designs using relative D-efficiency. Let $\mathbf{D}_1 \in \mathcal{D}_{f,l,n_1}$ and $\mathbf{D}_2 \in \mathcal{D}_{f,l,n_2}$ be the two design matrices we wish to compare, then

$$\left(\frac{n_2}{n_1}\right) \times \left(\frac{\phi_D(\mathbf{D}_1)}{\phi_D(\mathbf{D}_2)}\right)^{\frac{1}{p}} \times 100. \tag{1.20}$$

is the relative D-efficiency of these two designs.

D-optimality is particularly appropriate when the aim of the experiment is to gain scientific understanding through estimation of $\boldsymbol{\beta}$. This is the aim in both the motivating examples discussed in Section 1.2, hence this thesis focuses on D-optimal designs. D-optimality is also popular in the literature for a variety of other reasons: it is known to perform well with respect to other criteria, it is invariant to the scale or coding of the factor levels, and has powerful update formulae that speed up the code for design selection algorithms (Goos, 2002).

Other optimal designs discussed in the literature include:

- D_s-optimal designs, which are optimal for estimating a particular subset of the model parameters, referred to as the parameters of interest.
- A-optimal designs, which minimise the sum or average variance of the estimators of the fixed effect parameters.
- G-optimal designs, which minimise the maximum variance of predicted responses.
- V- (Q-, I-, or IV-) optimal designs, which minimise the average variances of predicted responses.

1.4.2 Bayesian Design Optimality Criteria

Bayesian methods allow the assimilation of prior knowledge, which is often specific to the application at hand. Using Bayes theorem (see Section 4.2.1 of Chapter 4 for more detail), the distribution of an unknown parameter $\boldsymbol{\theta}$ after data $\mathbf{y}$ has been observed is given by

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto L(\boldsymbol{\theta}|\mathbf{y})p(\boldsymbol{\theta}), \tag{1.21}$$

where $L(\boldsymbol{\theta}|\mathbf{y})$ is the likelihood of $\mathbf{y}$ given $\boldsymbol{\theta}$ and $p(\boldsymbol{\theta})$ is the prior distribution which represents our beliefs regarding $\boldsymbol{\theta}$ prior to $\mathbf{y}$ being observed.

Assuming that (1.3) models the responses, the variance covariance matrix for the responses is $(\sigma_\gamma^2 + \sigma_\epsilon^2)\mathbf{I}_n$ and $p(\boldsymbol{\beta}, (\sigma_\gamma^2 + \sigma_\epsilon^2))$ has a normal inverse gamma distribution, then the posterior variance covariance matrix for $\boldsymbol{\beta}$ is proportional to $(\mathbf{X}^T\mathbf{X} + \mathbf{R})^{-1}$ (O'Hagan and Forster, 2004, Chapter 11), where $\mathbf{R}^{-1}$ is proportional to the prior variance-covariance matrix for $\boldsymbol{\beta}$. Therefore, maximising

$$\phi_{BD} = |\mathbf{X}^T\mathbf{X} + \mathbf{R}| \tag{1.22}$$

is one way of obtaining a design to provide the most information regarding $\boldsymbol{\beta}$. The inclusion of $\mathbf{R}$ in (1.22) represents the prior knowledge assumed regarding the unknown parameters $\boldsymbol{\beta}$.

A Bayesian D-optimal design maximises (1.22) over $\mathcal{D}_{f,l,n}$. The objective function (1.22) can also be derived using the approach from Spezzaferri (1988) or by maximising the expected gain in Shannon information, see Chaloner and Verdinelli (1995) for further information. Chaloner and Verdinelli (1995) also presented utility functions for the derivation of objective functions for other Bayesian optimality criteria, such as Bayesian A-optimality.

We use a criterion based on (1.22) to find supersaturated designs in Chapters 3 and 5. The use of Bayesian optimality for supersaturated designs is advocated by authors such as Jones et al. (2008) because the inclusion of the prior precision matrix, $\mathbf{R}$, in the criterion regularises the information matrix and overcomes the problem of singular information matrices for supersaturated designs.

1.4.3 Design Selection: The Coordinate Exchange Algorithm

The coordinate exchange algorithm (Meyer and Nachtsheim, 1995) has been modified to find designs for experiments with restricted randomisation by authors such as Jones and Goos (2007) and Arnouts et al. (2010). Let $x_{i,j}$ be the jth element of $\mathbf{x}_i$, with $\mathbf{x}_i$ being the ith row of a design matrix, and let $\mathbf{D}_s \in \mathcal{D}_{f,l,n}$ be a starting design. Assume the aim is to maximise the objective function $\phi(\mathbf{D})$, then the coordinate exchange algorithm for two-level factors has the following general steps:

1. Set $\mathbf{D}_{1,0} = \mathbf{D}_s$ and calculate $\phi_s = \phi(\mathbf{D}_{1,0})$.
2. For $i = 1, \dots, n$, and $j = 1, \dots, f$:
    (a) Calculate $\phi_1 = \phi(\mathbf{D}_{i,j-1})$.
    (b) Let $\dot{x}_{i,j}$ be the (i,j)th element of $\mathbf{D}_{i,j-1}$.
    (c) Let $\mathbf{D}_{i,j}$ be equivalent to $\mathbf{D}_{i,j-1}$, but with (i,j)th element $x_{i,j} = -\dot{x}_{i,j}$.
    (d) Calculate $\phi_2 = \phi(\mathbf{D}_{i,j})$.
    (e) If $\phi_1 > \phi_2$, let $\mathbf{D}_{i,j} = \mathbf{D}_{i,j-1}$, otherwise, keep the swap and leave $\mathbf{D}_{i,j}$ and $x_{i,j}$ unchanged from (c).
3. Calculate $\phi_E = \phi(\mathbf{D}_{n,f})$.
4. If $\phi_S < \phi_E$, repeat from step 2 with $\mathbf{D}_{1,0} = \mathbf{D}_{n,f}$. Otherwise, stop the algorithm and return $\mathbf{D}_{1,0}$ as the design which maximises ϕ.

The coordinate exchange algorithm finds the design which maximises ϕ from a given starting design by swapping each factor level for each treatment in the design until no further improvement can be made.

To attempt to escape local optima, a subset of q starting designs from $\mathcal{D}_{f,l,n}$ are selected at random and the coordinate exchange algorithm is run for each of these designs. We define the design or designs found using the coordinate exchange algorithm for q starting designs with the largest value of ϕ as optimal for the optimality criterion relating to ϕ. We use a modified form of this algorithm in to find designs in Chapters 2 and 3.

1.4.4 Design Selection: The Interchange Algorithm

When the correlation structure is dependent on the order of treatments, the order of treatments in $\mathbf{D}$ is important and hence needs to be considered when finding optimal designs. The interchange algorithm (Atkinson et al., 2007, Chapter 12) searches for improvements for a optimal completely randomised design with respect to an objective function ϕ by swapping the order in which the treatments are applied.

Let $\mathbf{D}_s \in \mathcal{D}_{f,l,n}$ be an n-run starting completely randomised design with design points $\mathbf{x}_i$, $i = 1, \dots, n$, let $\mathcal{P} = \{(1,2), (1,3), \dots, (n-1,n)\}$ be the set of row indexes for all pairs of treatments in a design, with Qth element $\mathbf{P}_Q = (P_{Q1}, P_{Q2})$, and assume the aim is to maximise the objective function $\phi = \phi(\mathbf{D})$. Then, the interchange algorithm has the following general steps:

1. Set $\mathbf{D}_0 = \mathbf{D}_s$ and calculate $\phi_S = \phi(\mathbf{D}_0)$.

2. For $Q = 1, \dots, |\mathcal{P}|$:

   (a) Calculate $\phi_1 = \phi(\mathbf{D}_{Q-1})$.

   (b) Let $\mathbf{D}_Q$ be the design matrix where treatments $\mathbf{x}_{P_{Q1}}$ and $\mathbf{x}_{P_{Q2}}$ in $\mathbf{D}_{Q-1}$ are swapped.

   (c) Calculate $\phi_2 = \phi(\mathbf{D}_Q)$.

   (d) If $\phi_1 > \phi_2$ let $\mathbf{D}_Q = \mathbf{D}_{Q-1}$. Otherwise keep the swap and leave $\mathbf{D}_Q$ unchanged from (b).

3. Calculate $\phi_E = \phi(\mathbf{D}_{|\mathcal{P}|})$.

4. If $\phi_S < \phi_E$, repeat from step 2 with $\mathbf{D}_0 = \mathbf{D}_{|\mathcal{P}|}$. Otherwise stop the algorithm and return $\mathbf{D}_0$ as the design which maximises ϕ.

Therefore, the interchange algorithm finds the ordering of a given design which maximises ϕ by swapping pairs of design points until no further improvement can be made.

The interchange algorithm is run for q random permutations of the elements in $\mathcal{P}$, and the designs which maximise ϕ from the set of q final designs is chosen as the best design. We use this algorithm to allocate treatments for block designs with autoregressive intrablock errors in Chapter 2. This algorithm finds the run order that maximises ϕ and hence the run order in the resulting experiment should not be randomised.

1.5 Overview of Thesis

The main results in this thesis are given in Chapters 2 through 5. The aim of this thesis is to develop and assess methods to find and analyse designs with restricted randomisation in both blocks and stages.

In Chapter 2, we present results for block designs for linear mixed effect models where the intrablock errors are assumed to follow an autoregressive process of order one (AR(1) process). A block experiment with AR(1) intrablock errors is a special type of block experiment with a natural ordering of the runs within groups. Investigation into these designs is motivated by the manufacture of microstructured optical fibres discussed in Section 1.2.

The properties of saturated and non-saturated block designs with autocorrelated errors which maximise the D-optimality objective function found using two computer algorithms are compared. The robustness of these designs to misspecification of both the autoregressive parameter and the relative magnitude of the interblock and intrablock variances is assessed using D-efficiency. Finally, the treatment selection and allocation of designs with the same objective function value is discussed.

In Chapter 3 we discuss the theory of multi-tiered and multi-stage designs and provide our definition of multi-stage designs, which is extended from the partition design literature (Perry et al., 2001, 2002, 2007). We then present computer algorithms to find optimal multi-stage designs for compound Bayesian D-optimality with different restrictions on randomisation. The results presented in this chapter focus on the two-stage optimal designs suitable for the formulation of a pharmaceutical product discussed in Section 1.2. A model for the response from each stage, as well as cumulative models for the responses, which may be supersaturated, are considered, hence a compound Bayesian D-optimality objective function is used.

The correlation between the columns of the model matrix for two-stage optimal designs with different restrictions on randomisation is assessed to evaluate what information can be retrieved from the experiment about individual factors. Designs with good projectivity properties are discussed, and a comparison is made between designs based on known designs with good projectivity properties and the optimal two-stage completely randomised designs.

In Chapter 4, we discuss Bayesian variable selection and motivate our use of a computationally intensive Bayesian methods for selecting influential factors for split-plot designs through the analysis of simulated data using all subsets regression and the global and local search algorithm presented by Tan and Wu (2013).

We then present our method, which employs the Markov chain Monte Carlo sampling methods of Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990) and Metropolis-Hastings rejection sampling (Metropolis et al., 1953; Hastings, 1970). We assess the effectiveness of this method via simulations, using the posterior probability of parameters being active and the sampled parameter distributions for multivariate responses from the optimal two-stage split-plot experiment found in Chapter 3.

In Chapter 5, we show how the methodology in Chapters 3 and 4 can be applied to the pre-clinical formulation and dissolution testing of a pharmaceutical product discussed in Section 1.2. We also discuss how both a grid search and the efficient global optimisation (EGO) algorithm from Jones et al. (1998) can be used to optimise the probability of dissolution testing responses being meeting specification. Using these methods, we identify and analyse treatments with high predicted probabilities of meeting specification.

Finally, in Chapter 6, we present some brief conclusions and possibilities for future work.

Chapter 2

Block Designs for Mixed Effect Models and Autoregressive Intrablock Errors

Block designs organise the runs of an experiment into homogeneous groups based on some feature of experimentation, such as the batch of experimental material used or the technician running the experiment. Block designs are common in many areas of science and industry, as they provide more accurate and precise conclusions for experiments where restrictions on randomisation are induced through features of the experiment.

Mixed models, as introduced in Section 1.3.1 of Chapter 1, can be used to analyse the responses from block experiments, and they allow the response from unobserved blocks to be inferred. Here it is assumed that the blocks constitute a random sample from a population of blocks. The linear mixed effects model for the analysis of block designs is introduced in Section 2.2.

We consider an alternative to the usual exchangeable correlation structure and assume that the ordering of units within a block may have an influence on the response. Experimental units can be ordered in time and space. For example, the yields from neighbouring plots could be ordered based on the location of the plots, or the responses from a sequence of manufacturing processes could be ordered based on the order of the processes in time. The correlation between ordered responses is usually assumed to be positive as positive correlation is more widely applicable. The effect of ordering is modelled by assuming an autoregressive process of order 1 (an AR(1) process) for the intrablock (within-block) errors, as discussed in Section 2.3.

The order of runs in a block design can only be randomised within blocks; swapping two treatments from blocks will change the properties of the design. Therefore, there are two key choices for block designs, the choice of treatments and the allocation of these treatments to blocks. The coordinate exchange algorithm (Section 1.4.3) chooses and allocates treatments, and the interchange algorithm (Section 1.4.4) allocates treatments from an optimal design to blocks. In this chapter, we use both of these algorithms to find block designs with autoregressive intrablock errors which maximise the D-optimality objective function, (1.18) in Section 1.4.1.

When finding block designs for autoregressive intrablock errors, the autocorrelation parameter and the ratio of interblock (between-block) and intrablock (within-block) errors are unknown. However, both of these parameters are required in the D-optimality objective function used to find designs in this work. Therefore, we consider the robustness of the block designs found to the values of these parameters in Section 2.4.2. We also discuss the structure of D-efficient designs in Section 2.4.3, and use relative efficiencies to consider the performance of the coordinate exchange and interchange algorithms in Section 2.5.

2.1 Motivation and Aim of Work

The motivation for the work in this chapter arises from a collaboration with the Optoelectronics Research Centre (ORC) at the University of Southampton, as discussed in Section 1.2.1. The ORC aim to use experimentation to find which factor settings produce a microstructured optical fibre with the best light transmission properties. As discussed in Section 1.2.1, the manufacture of microstructured optical fibres requires two processes, and the second process is the motivation for the work in this chapter.

Fibre manufacture could be described as a block design, where the blocking factor is the cane which is drawn into fibre, and the experimental unit is the section of fibre to which a factorial treatment is applied in each run of experimentation. For example, four sections of fibre from four different canes could be used to create 16 experimental units. The effect of the cane on the response is not of interest, hence random block effects (as discussed in Section 2.2) are used to enable prediction of the properties of new canes.

Either three or four factor settings can be varied during the fibre manufacture process. The speed at which the cane is fed into the machine which draws it into a fibre, the speed at which the fibre is drawn within this machine, and the pressure at the core of the fibre are always varied in fibre manufacture. The temperature of the furnace used to heat the cane so it can be drawn into a fibre may or may not be varied in the experiment, depending on the range of the other factors.

If the same treatment was used to draw two fibres from the same cane, then we would assume that the light transmission properties of these fibres would be more similar than those measured for two fibres drawn using the same treatment from different canes. This supports the use of a block design for fibre manufacture, as an underlying assumption of block designs is that the responses from repeated treatments in the same block are more similar than responses from repeated treatments in different blocks.

The responses for the lengths of fibres that are drawn after each other are assumed to be positively correlated. The responses for all the fibres drawn from the same cane are assumed to have a correlation which decays as the distance between the lengths of fibre increases. Therefore, it seems appropriate to assume that the intrablock errors follow an autoregressive process, as discussed in Section 2.3.

The aim of the work in this chapter is to find designs which maximise the D-optimality objective function (1.18) which are appropriate for the manufacture of microstructured optical fibres. We use (1.18) as we want to gain scientific information about the fixed effect parameters, which model the impact of the factors on the response. As experimentation is costly and time consuming, 12 or 16 runs designs with three or four two-level factors would be practically feasible.

We use the coordinate exchange (Section 1.4.3) and interchange (Section 1.4.4) algorithms to find block designs with autoregressive intrablock errors which maximise (1.18). In Section 2.4.2 we discuss the robustness of these designs to misspecification of the ratio of the inter- and intrablock variance and the autocorrelation parameter, which are unknown prior to experimentation. In Section 2.4.3 we consider the structure of efficient designs. In Section 2.5 we use D-efficiency to compare the designs found using these two algorithms and discuss the importance of using algorithms to allocate treatments to blocks.

2.2 A Mixed Model with Random Block Effects for Analysing Block Designs

As introduced in Section 1.3.1 of Chapter 1, a linear mixed model for the analysis of responses from a block design with $n = bk$ runs arranged in b blocks of size k is given by (1.3), where $\mathbf{Z}$ is the $n \times b$ matrix which represents the allocation of the runs to blocks, $\boldsymbol{\gamma}$ is the $b \times 1$ vector of block effects and $\boldsymbol{\epsilon}$ is the $n \times 1$ vector of random within-block errors. If the ith run of an experiment, $i = 1, \ldots, n$, is in the jth block, $j = 1, \ldots, b$, then the (i, j)th element of $\mathbf{Z}$ will be 1, otherwise it is 0.

Block effects can be fixed, with $\boldsymbol{\gamma}$ as a $b \times 1$ vector of fixed values, or random, with $\boldsymbol{\gamma}$ as a $b \times 1$ vector of values drawn from some distribution. We use random block effects in this work as we assume that the blocks used in an experiment are a random sample from a hypothetical population of "all possible" blocks. This allows the results from the experiment to be generalised to future blocks and therefore enables inferences to be made.

Fixed block effects provide no basis for comparing the blocks within the experiment to a larger population of blocks. Hence, when the block effects are assumed to be fixed, the data from the experiment cannot be used to make predictions, and inferences more generally, about blocks other than those in the experiment. The use of fixed or random block effects is discussed in further detail by a range of authors such as Goos (2002, Section 2.3.2), Morris (2011, Chapter 8) and Goos and Jones (2011, Section 8.3.1).

The two random variables in (1.3), $\boldsymbol{\gamma}$ and $\boldsymbol{\epsilon}$, are assumed to be independently normally distributed, where $\boldsymbol{\gamma}$ ∼N$(\mathbf{0}_b, \sigma_\gamma^2 \mathbf{I}_b)$ and $\boldsymbol{\epsilon} \sim$ N$(\mathbf{0}_n, \sigma_\epsilon^2 \mathbf{P}_n)$ when $\mathbf{0}_b$ and $\mathbf{0}_n$ are the $b \times 1$ and $n \times 1$ vectors, respectively, with each element as 0. The variation between blocks effects, σ_γ^2, is identical for all the blocks as we assume that the responses from different blocks have the same variability. The $n \times n$ correlation matrix $\mathbf{P}_n$ describes the assumed relationship between the responses within blocks.

When $\boldsymbol{\gamma}$ ∼N$(\mathbf{0}_b, \sigma_\gamma^2 \mathbf{I}_b)$ and $\boldsymbol{\epsilon} \sim$ N$(\mathbf{0}_n, \sigma_\epsilon^2 \mathbf{P}_n)$, the variance-covariance matrix, $\mathbf{V}$, of $\mathbf{Y}$ in (1.3) is

$$\begin{aligned}
\mathbf{V} &= \text{var}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}) \\
&= \text{var}(\mathbf{Z}\boldsymbol{\gamma}) + \text{var}(\boldsymbol{\epsilon}) \\
&= \mathbf{Z}\sigma_\gamma^2 \mathbf{I}_b \mathbf{Z}^T + \sigma_\epsilon^2 \mathbf{P}_n \\
&= \sigma_\epsilon^2 \left(\frac{\sigma_\gamma^2}{\sigma_\epsilon^2} \mathbf{Z}\mathbf{Z}^T + \mathbf{P}_n \right) \\
&= \sigma_\epsilon^2 \left(\eta \mathbf{Z}\mathbf{Z}^T + \mathbf{P}_n \right) \\
&= \sigma_\epsilon^2 \left(\eta \mathbf{I}_b \otimes \mathbf{J}_k + \mathbf{P}_n \right), \qquad (2.1)
\end{aligned}$$

where $\eta = \sigma_\gamma^2 / \sigma_\epsilon^2$ is the relative magnitude of the interblock (between-block) and intrablock (within-block) variance components, $\mathbf{J}_k$ is the $k \times k$ matrix with one as every element and $\otimes$ is the Kronecker product.

When it is assumed that the order in which treatments within a block are applied can be randomised, then $\mathbf{P}_n = \mathbf{I}_n$ and the responses are said to have an exchangeable error structure. The $\mathbf{V}$ matrix for this error structure is

$$\begin{pmatrix}
\sigma_\gamma^2 \mathbf{J}_k + \sigma_\epsilon^2 \mathbf{I}_k & \mathbf{0}_{kk} & \ldots & \mathbf{0}_{kk} \\
\mathbf{0}_{kk} & \sigma_\gamma^2 \mathbf{J}_k + \sigma_\epsilon^2 \mathbf{I}_k & \ldots & \mathbf{0}_{kk} \\
\vdots & \vdots & \ddots & \vdots \\
\mathbf{0}_{kk} & \mathbf{0}_{kk} & \ldots & \sigma_\gamma^2 \mathbf{J}_k + \sigma_\epsilon^2 \mathbf{I}_k
\end{pmatrix}, \qquad (2.2)$$

where $\mathbf{0}_{kk}$ is the $k \times k$ matrix with each element as 0. The off-diagonal submatrices of (2.2) are $\mathbf{0}_{kk}$ as two observations from different blocks are independent. However, as two observations from the same block are correlated, the diagonal sub-matrices in

(2.2) are

$$\sigma_\gamma^2 \mathbf{J}_k + \sigma_\epsilon^2 \mathbf{I}_k = \begin{pmatrix} \sigma_\gamma^2 + \sigma_\epsilon^2 & \sigma_\gamma^2 & \dots & \sigma_\gamma^2 \\ \sigma_\gamma^2 & \sigma_\gamma^2 + \sigma_\epsilon^2 & \dots & \sigma_\gamma^2 \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_\gamma^2 & \sigma_\gamma^2 & \dots & \sigma_\gamma^2 + \sigma_\epsilon^2 \end{pmatrix}, \tag{2.3}$$

or, on substitution of $\eta = \sigma_\gamma^2/\sigma_\epsilon^2$,

$$\sigma_\gamma^2 \mathbf{J}_k + \sigma_\epsilon^2 \mathbf{I}_k = \sigma_\epsilon^2 \begin{pmatrix} \eta + 1 & \eta & \dots & \eta \\ \eta & \eta + 1 & \dots & \eta \\ \vdots & \vdots & \ddots & \vdots \\ \eta & \eta & \dots & \eta + 1 \end{pmatrix}. \tag{2.4}$$

2.3 Autoregressive Errors

Assume that there is a one dimensional structure, for example arising from a spatial or temporal dependency between the experimental units within blocks, which provides an implicit ordering of the units within the block. Then, the structure of the correlation matrix $\mathbf{P}_n$ in (2.2) will need to be adapted to ensure that the experiment is designed, and responses from the experiment are analysed, to account for this intrablock correlation.

Autoregressive processes, as discussed by Box et al. (2008, Chapter 2) and Fuller (1996, Chapter 2), can be used to describe a linear relationship between two ordered observations, such as the yields for two adjacent sections of a field or the responses for two experimental units that have treatments applied to them consecutively. An autoregressive process of order c, an AR(c) process, is defined as

$$d_t = \sum_{j=1}^{c} \phi_j d_{t-j} + a_t, \tag{2.5}$$

where d_t is the current observation, $d_{t-1}, \dots, d_{t-c}$ are the past observations, $\phi_1, \dots, \phi_c$ are the autoregressive parameters, $a_t \overset{iid}{\sim} N(0, \sigma_a^2)$ is a noise variable with constant variance, which is assumed to be independent of previous observations, $t - c \geq 1$, and $t = 2, \dots, n$.

We use autoregressive processes to extend the standard error structure for block designs. In this chapter, we assume that the responses for experimental units which are "closest" together in space or time have the strongest relationship, and that this relationship decays as the distance between experimental units increases. Therefore, we assume that intrablock errors follow an AR(1) process, which is (2.5) for $c = 1$.

Let $\epsilon_{j,h}$ be the error for run h ($h = 1, \ldots, k$) in the jth ($j = 1, \ldots, b$) block. If the intrablock errors are assumed to follow a stationary AR(1) process then, following Pantula and Pollock (1985), for $h = 2, \ldots, k$

$$\epsilon_{j,h} = \rho\epsilon_{j,h-1} + e_{j,h}, \tag{2.6}$$

where $|\rho| < 1$, $e_{j,h} \overset{iid}{\sim} \mathrm{N}(0, \sigma_\epsilon^2)$. It is assumed that $\epsilon_{j,1} \overset{iid}{\sim} \mathrm{N}\left(0, \sigma_\epsilon^2/(1-\rho^2)\right)$, which is derived from the limit of the variance of $\epsilon_{j,h}$ as $h \to -\infty$, and relies on the assumption that $\epsilon_{j,1}$ is bounded and $|\rho| < 1$.

When the intrablock errors are assumed to follow (2.6), then the correlation matrix $\mathbf{P}_n$ in (2.1) is given by

$$\mathbf{P}_n = \frac{1}{1-\rho^2}\left(\mathbf{I}_b \otimes \boldsymbol{\psi}\right), \tag{2.7}$$

where $\boldsymbol{\psi}$ is the $k \times k$ matrix

$$\boldsymbol{\psi} = \begin{pmatrix} 1 & \rho & \ldots & \rho^{k-1} \\ \rho & 1 & \ldots & \rho^{k-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho^{k-1} & \rho^{k-2} & \ldots & 1 \end{pmatrix}. \tag{2.8}$$

Therefore, the variance-covariance matrix when the intrablock errors follow an AR(1) process is

$$\mathbf{V} = \mathbf{I}_b \otimes \left(\sigma_\gamma^2 \mathbf{J}_k + \frac{\sigma_\epsilon^2}{1-\rho^2}\boldsymbol{\psi}\right). \tag{2.9}$$

As in equation (2.2), the off-diagonal sub-matrices for (2.9) are $\mathbf{0}_k$, as the runs from different blocks are assumed to be independent. We note that equations (2.2) and (2.9) are equivalent for $\rho = 0$ (and $\boldsymbol{\psi} = \mathbf{I}_k$).

Other error structures, such as nearest neighbour correlation, also account for an implicit one dimensional relationship between the responses within blocks. Nearest neighbour designs assume that treatments applied to one experimental unit have some constant residual effect on neighbouring units, where units can be neighbours in space or time. Nearest neighbour block designs therefore assume that responses in the same block have the same constant correlation, which implies that $\mathbf{P}_n = \mathbf{I}_b \otimes \boldsymbol{\psi}$ and

$$\boldsymbol{\psi} = \begin{pmatrix} 1 & \rho & 0 & \dots & 0 & 0 & 0 \\ \rho & 1 & \rho & \dots & 0 & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \dots & \rho & 1 & \rho \\ 0 & 0 & 0 & \dots & 0 & \rho & 1 \end{pmatrix}.$$

Autocorrelated errors assume that treatments applied to one experimental unit have a residual effect that decays as the spatial or temporal distance between the neighbouring units increases. It is more appropriate to assume autocorrelated, and not nearest neighbour, intrablock correlation for the motivating example described in Section 2.1.

2.4 Study of Robustness of Block Designs to Misspecification of Correlation Structure

In this section, we perform a study to assess:

- The importance of the allocation of runs in saturated and unsaturated block designs with AR(1) intrablock errors (Section 2.4.1).
- The robustness of block designs with AR(1) intrablock errors which maximise (1.18) to misspecification of ρ and η using D-efficiency, (1.20) in Section 1.4.1 (Section 2.4.1).
- The structure of robust designs (Section 2.4.3).

The designs in this study which maximise (1.18) were found using either the coordinate exchange algorithm (Section 1.4.3) or the interchange algorithm (Section 1.4.4), where $\mathbf{V}$ is as given by (2.9), and all combinations of $\rho \in \rho^*$, $\rho^* = \{0, 0.25, 0.5, 0.75\}$ and $\eta \in \eta^*$, $\eta^* = \{0, 2.5, 5, 7.5, 10\}$ were considered. We used random starting designs for the coordinate exchange algorithm. The D-optimal completely randomised design, which is found using the coordinate exchange algorithm for $\rho = \eta = 0$, was used as the starting design for the interchange algorithm.

We considered four experiments in this study, whose responses are all assumed to be modelled using (1.3), with variance-covariance matrix (2.9). Table 2.1 gives $\mathcal{D}_{f,l,n}$, b, k, and the elements of $\boldsymbol{\beta}$ in (1.3) for these four experiments. In Section 2.4.1, we compare designs for Experiment 1 and 2 to assess the importance of run order on saturated (Experiment 1) and unsaturated designs (Experiment 2). We use Experiments 2 to 4 to assess the robustness of 16 (Experiment 2) and 12 run (Experiment 3 and 4) unsaturated designs to the misspecification of ρ and η in Section 2.4.2. The number of factors, factor levels and runs considered in these experiments reflect the potential number of two-level factors in the experiment discussed in Section 2.1.

	Experiment			
	1	2	3	4
$\mathcal{D}_{f,l,n}$	$\mathcal{D}_{3,2,8}$	$\mathcal{D}_{3,2,16}$	$\mathcal{D}_{4,2,12}$	$\mathcal{D}_{4,2,12}$
b	2	4	3	3
k	4	4	4	4
$\boldsymbol{\beta}$	$\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\\ \beta_{12}\\ \beta_{13}\\ \beta_{23}\\ \beta_{123}\end{pmatrix}$	$\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\\ \beta_{12}\\ \beta_{13}\\ \beta_{23}\\ \beta_{123}\end{pmatrix}$	$\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\\ \beta_4\end{pmatrix}$	$\begin{pmatrix}\beta_0\\ \beta_1\\ \beta_2\\ \beta_3\\ \beta_4\\ \beta_{12}\\ \beta_{13}\\ \beta_{14}\\ \beta_{23}\\ \beta_{24}\\ \beta_{34}\end{pmatrix}$

Table 2.1: $\mathcal{D}_{f,l,n}$, b, k, and elements of $\boldsymbol{\beta}$ for experiments compared in Section 2.4 whose responses are modelled using (1.3) and variance covariance matrices are given by (2.9).

2.4.1 A Comparison of Run Allocation for Saturated and Unsaturated Designs

In this section, we will compare the allocation of runs in saturated and unsaturated block designs with AR(1) correlated intrablock errors. We do this using two experiments which are suitable for the motivation of this chapter, Experiment 1, which is a saturated eight run design with two blocks of size four, and Experiment 2, which is an unsaturated sixteen run design with four blocks of size four. Both these experiments are for three two-level factors, and can only include treatments from Table 2.2.

Treatment	x_1	x_2	x_3
1	1	1	1
2	1	1	-1
3	1	-1	1
4	1	-1	-1
5	-1	1	1
6	-1	1	-1
7	-1	-1	1
8	-1	-1	-1

Table 2.2: Treatments for Experiments 1 and 2.

The correlation structure and treatment allocation for the design is enforced through $\mathbf{V}$. The D-optimality objective function, (1.18), for saturated designs can be written as

$$\phi_D = |\mathbf{V}^{-1}||\mathbf{X}^T\mathbf{X}| \tag{2.10}$$

because the model matrix $\mathbf{X}$ is a $p \times p$ matrix. The efficiency (1.20) for saturated designs is independent of $\mathbf{V}$, hence saturated designs are not dependent on the correlation structure assumed for the errors in the design or the allocation of treatments to blocks (Goos, 2002, pg. 110-111). Therefore, (1.20) is 100% when any two saturated designs with the same treatments are compared, as every allocation of the same treatments to blocks is equally efficient.

The designs for Experiment 1 found using the coordinate exchange and interchange algorithm $\forall \rho \in \rho^*$ and $\forall \eta \in \eta^*$ are saturated, as $n = p = 8$, and they all have the eight treatments given in Table 2.2 without replication. As expected, all pairs of the saturated designs for Experiment 1 have 100% D-efficiency, and the completely randomised design, found using $\eta = \rho = 0$, is optimal for $\forall \rho \in \rho^*_{-0}$ and $\forall \eta \in \eta^*_{-0}$ where $\rho^*_{-0} = \rho^*/0 = \{0.25, 0.5, 0.75\}$ and $\eta^*_{-0} = \eta^*/0 = \{2.5, 5, 7.5, 10\}$.

The optimality criterion for unsaturated designs cannot be written as (2.10), hence we expect the allocation of treatments and, when AR(1) correlated errors are assumed, the order of treatments within blocks to impact on the optimality and efficiency of these designs. The designs which maximise (1.18) for Experiment 1 and 2 both contain all the treatments given in Table 2.2. The designs for Experiment 1 have one replicate of these treatments, and the designs for Experiment 2 have two replicates of these treatments.

Table 2.3 gives the per-run relative D-efficiencies for the saturated designs for Experiment 1 and unsaturated designs for Experiment 2 found using both the coordinate exchange and interchange algorithm when $\rho \in \rho^*$ and $\eta \in \eta^*$. The relative D-efficiencies in Table 2.3 are calculated using (1.20), where (1.18) for the saturated designs is the numerator of (1.20) and (1.18) for the unsaturated designs is the denominator of (1.20).

The results for both algorithms are reported in Table 2.3 as they were identical, hence the designs for Experiment 2 found using the coordinate exchange and interchange algorithm will have 100% relative D-efficiency. This is discussed in further detail in Section 2.5.

η \ ρ	0	0.25	0.5	0.75
0	100.00	97.66	90.68	78.79
2.5	85.26	82.08	77.81	71.09
5	79.31	76.40	72.71	67.20
7.5	75.78	73.01	69.39	64.65
10	73.30	70.63	67.18	62.77

Table 2.3: Relative per-run D-efficiencies, (1.20), of D-optimal eight and sixteen run designs for Experiment 1 and Experiment 2, respectively, found using the coordinate exchange and interchange algorithm for ρ given by the column heading and η given by the row heading (%, 2dp).

The relative efficiencies in Table 2.3 decrease as η and ρ increase, therefore, for the specific examples considered in this study, the allocation and ordering of treatments to blocks in unsaturated designs becomes more important as η and ρ increase, as expected. It would be interesting to see if this result holds for the comparison of other saturated and unsaturated designs with different numbers of runs.

2.4.2 Robustness to Misspecification of ρ and η

The autocorrelation parameter, ρ, and the ratio η of inter- and intrablock variance are unknown prior to experimentation. However, both of these parameters are required to calculate the objective function (1.18). Identifying whether the designs found are robust to misspecification of these parameters is important, as the ρ and η assumed when designing the experiment may differ from the true ρ and η.

The D-efficiency for all pairs of designs found using both the coordinate exchange and interchange algorithm for Experiments 2, 3 and 4 when $\rho \in \rho^*_{-0}$ and $\eta \in \eta^*_{-0}$, is approximately 100%. Therefore, when the design is found assuming it is blocked and has some autocorrelated errors, the design is robust to misspecification of ρ and η for the values considered in this study.

The assessment of the robustness of the D-optimal completely randomised design, which is found using the coordinate exchange algorithm for $\rho = \eta = 0$, to blocking and correlation is a common theme in the literature discussed in Section 2.6. We use the relative D-efficiency of the design found using the coordinate exchange algorithm for $\rho = \eta = 0$ to consider the robustness of the completely randomised design to blocking and correlation.

Tables 2.4 and 2.7 give the D-efficiencies of a single random ordering of the treatments in the D-optimal completely randomised design relative to the blocked and correlated designs found using the coordinate exchange and interchange algorithms for Experiments 2 and 4, respectively. Tables 2.5 and 2.6 give the D-efficiencies of a single random ordering of the treatments in the D-optimal completely randomised design relative to the blocked and correlated designs found using the coordinate exchange and interchange algorithms, respectively, for Experiment 3.

ρ η	0	0.25	0.5	0.75
0	100.00	99.00	94.86	87.05
2.5	79.88	81.44	82.41	81.87
5	77.38	79.04	80.22	80.33
7.5	76.37	78.07	79.29	79.59
10	75.83	77.54	78.79	79.16

Table 2.4: D-efficiencies, (1.20), for the D-optimal completely randomised design for Experiment 2 relative to the design for Experiment 2 found using the coordinate exchange and interchange algorithms which maximises (1.18) for ρ given by the column heading and η given by the row heading (%, 2dp).

ρ η	0	0.25	0.5	0.75
0	100.00	82.48	67.77	55.68
2.5	76.17	66.13	58.46	51.96
5	72.18	63.07	56.25	50.68
7.5	70.44	61.72	55.24	50.03
10	69.47	60.97	54.66	49.23

Table 2.5: D-efficiencies, (1.20), for the D-optimal completely randomised design for Experiment 3 relative to the design for Experiment 3 found using the coordinate exchange algorithm which maximises (1.18) for ρ given by the column heading and η given by the row heading (%, 2dp).

ρ η	0	0.25	0.5	0.75
0	100.00	87.45	75.96	64.83
2.5	76.17	70.50	65.42	60.38
5	72.26	67.27	62.93	58.85
7.5	70.57	65.85	61.79	58.08
10	69.62	65.05	61.14	57.61

Table 2.6: D-efficiencies, (1.20), for the D-optimal completely randomised design for Experiment 3 relative to the design for Experiment 3 found using the interchange algorithm which maximises (1.18) for ρ given by the column heading and η given by the row heading (%, 2dp).

η \ ρ	0	0.25	0.5	0.75
0	100.00	97.91	96.21	94.89
2.5	95.00	94.75	94.55	94.33
5	94.51	94.37	94.27	94.17
7.5	94.32	94.23	94.16	94.10
10	94.22	94.15	94.10	94.06

Table 2.7: D-efficiencies, (1.20), for the D-optimal completely randomised design for Experiment 4 relative to the design for Experiment 4 found using the coordinate exchange and interchange algorithms which maximises (1.18) for ρ given by the column heading and η given by the row heading (%, 2dp).

The efficiencies in Tables 2.4 to 2.7 are calculated using (1.18) for the D-optimal completely randomised design in the numerator, and (1.18) for the design found using the specified algorithm for the ρ given by the column heading and the η given by the row heading. The single random ordering of the completely randomised design used to calculate these efficiencies is the ordering returned by the coordinate exchange algorithm. The variance-covariance matrix (2.9) for the ρ given by the column heading and the η given by the row heading is used in both the numerator and denominator.

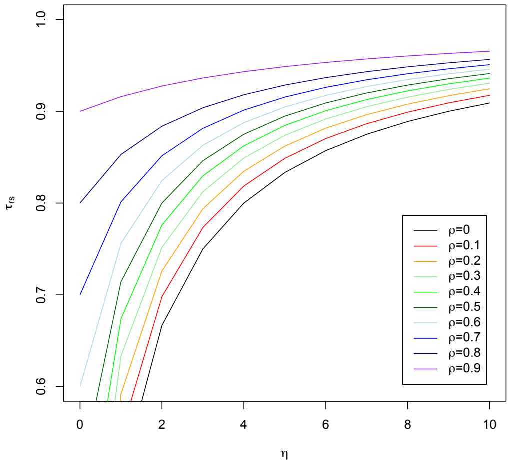
Firstly, we note that the efficiencies for both algorithms are identical for Experiments 2 (Table 2.4) and 4 (Table 2.7). Also, we note that the efficiencies in Tables 2.4 to 2.7 decreases as ρ and η increase. Finally, the efficiencies in Table 2.5 and 2.6 are lower than those in Table 2.7, hence the efficiencies increase as $n - p$ decreases. All of these results are discussed in further detail in Section 2.5.

The robustness of designs may depend on the value of the intrablock correlation,

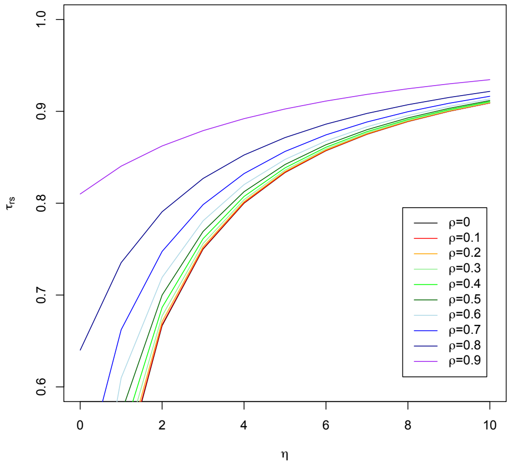
$$\tau_{rs} = \frac{(1-\rho^2)\eta + \rho^{|r-s|}}{(1-\rho^2)\eta + 1}, \tag{2.11}$$

where r, $s \in \{1, \ldots, k\}$ represent the positions of the two runs within a block. Figure 2.1 shows τ_{rs} as a function η for $\rho \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9\}$ when two runs are nearest neighbours (Figure 2.1a, $|r-s| = 1$) or are separated by another treatment (Figure 2.1b, $|r-s| = 2$). The value of τ_{rs} increases as η increases and the difference between τ_{rs} for different ρ decreases as η increases.

It may be difficult to identify the differences in the designs when the intrablock correlations are high and do not have large difference, therefore we expect the designs to increase in robustness with respect to ρ as η increases and the correlation between the different values becomes more similar. For our designs, when $\rho \in \rho^*$ varies and $\eta \in \eta^*$ is fixed, $\tau_{rs} > 0.78$ (2dp), for $|r-s| = 1, 2, 3$, and the maximum difference between τ_{rs} for different $\rho \in \rho^*$ was 17%. Therefore, the similarities in efficiency seen for Experiments 2, 3 and 4 when $\rho, \eta > 0$ are unsurprising. Increasing the size of the blocks would reduce the intrablock correlation and may therefore produce designs which have lower relative efficiencies.



(a)



(b)

Figure 2.1: Intra-block correlation, τ_{rs}, as a function of η for ρ = 0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, when (a) $|r-s|$ =1 and (b) $|r-s|$ =2.

2.4.3 Structure of Robust Designs

As mentioned in Section 2.4.2, the relative efficiencies of any two designs found using both the coordinate exchange and interchange algorithm when $\rho \in \rho^*_{-0}$ and $\eta \in \eta^*_{-0}$ is approximately 100%. There are some designs for Experiments 2, 3 and 4 which have an efficiency of exactly 100 %, and hence identical values for (1.18). In this section, we investigate whether two designs with the same value of (1.18) also have structural equivalences, such as treatment allocation to blocks and ordering within blocks.

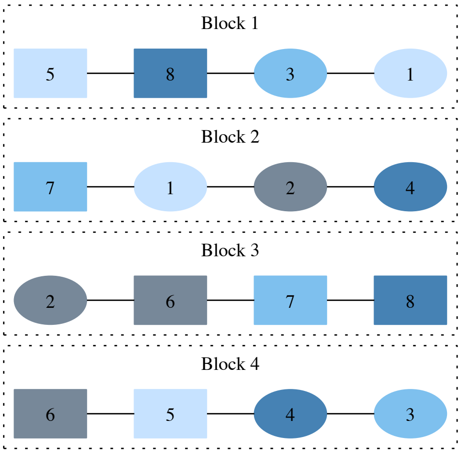


Figure 2.2: Allocation of treatments in Table 2.2 to blocks for the D-optimal 16 run design for Experiment 2 found using the coordinate exchange algorithm for $\eta = 10, \rho = 0.25$.

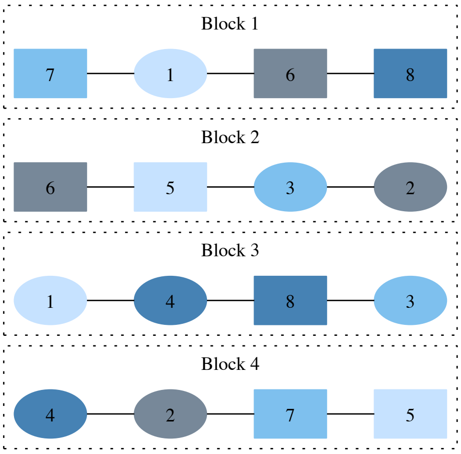


Figure 2.3: Allocation of treatments in Table 2.2 to blocks for the D-optimal 16 run design for Experiment 2 found using the coordinate exchange algorithm for $\eta = 10, \rho = 0.75$.

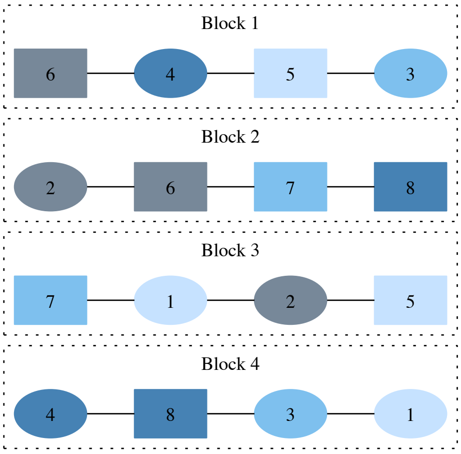


Figure 2.4: Allocation of treatments in Table 2.2 to blocks for the D-optimal 16 run design for Experiment 2 found using the interchange algorithm for $\eta = 10, \rho = 0.25$.

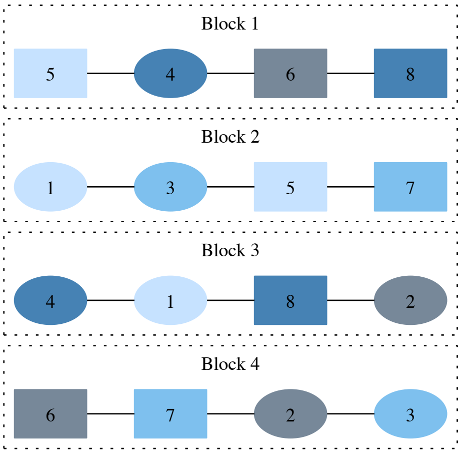


Figure 2.5: Allocation of treatments in Table 2.2 to blocks for the D-optimal 16 run design for Experiment 2 found using the interchange algorithm for $\eta = 10, \rho = 0.75$.

The designs for Experiment 2 all contain two replicates of the treatments in Table 2.2, regardless of their (1.18) value. However, designs with equal (1.18) values found using both the coordinate exchange and interchange algorithm for Experiment 2 do not necessarily have the same treatment allocation to blocks or ordering within blocks.

For example, the designs found using the coordinate exchange algorithm for $\eta = 10$ and $\rho = 0.25, 0.75$ have 100% relative D-efficiency, however, as seen in Figures 2.2 and 2.3, they do not have the same treatment allocation or ordering. Similarly, as seen by comparing Figures 2.4 and 2.5, the designs found using the interchange algorithm for $\eta = 10$ and $\rho = 0.25, 0.75$ have 100% relative D-efficiency but do not have the share treatment allocation or ordering.

The designs for Experiment 3 found using the coordinate exchange algorithm have an interesting structure. All the designs for $\rho \in \rho^*_{-0}$ and $\eta \in \eta^*_{-0}$ have eight distinct treatments from Table 2.8, four of which are repeated at the start and end of each block, as exemplified by Figures 2.6 and 2.7. This is an unusual structure, and it would be interesting to see if block designs with AR(1) intrablock errors found using the coordinate exchange algorithm which maximise (1.18) for different n, when the model containing the main effects is assumed for the response, all have this structure.

Treatment	x_1	x_2	x_3	x_4
1	1	1	1	1
2	1	1	1	-1
3	1	1	-1	1
4	1	1	-1	-1
5	1	-1	1	1
6	1	-1	1	-1
7	1	-1	-1	1
8	1	-1	-1	-1
9	-1	1	1	1
10	-1	1	1	-1
11	-1	1	-1	1
12	-1	1	-1	-1
13	-1	-1	1	1
14	-1	-1	1	-1
15	-1	-1	-1	1
16	-1	-1	-1	-1

Table 2.8: Treatments for 12 run designs.

Even though all the designs for Experiment 3 found using the coordinate exchange algorithm have this structure, we notice that they do not necessarily share the same treatment allocation or ordering. There are some designs which have 100% relative efficiency, identical treatments, treatment allocation and ordering. For example, Figure 2.13 shows the treatment allocation and ordering for two designs, the designs for Experiment 3 found using the coordinate exchange algorithm for $\eta = 5$, $\rho = 0.5, 0.75$, which have the same value of (1.18).

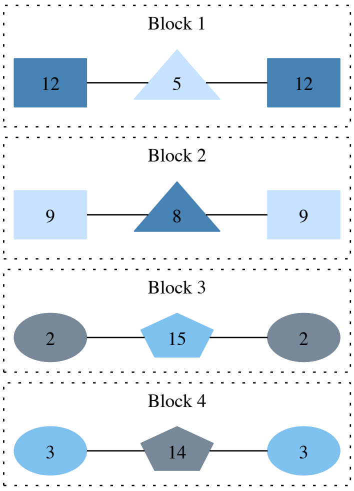


Figure 2.6: Allocation of treatments in Table 2.8 to blocks for the D-optimal 12 run design for Experiment 3 found using the coordinate exchange algorithm for $\eta = 5, \rho = 0.25$.

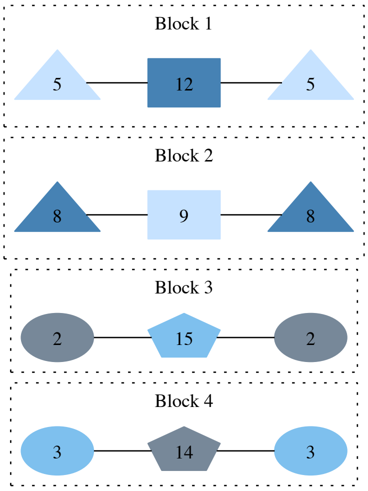


Figure 2.7: Allocation of treatments in Table 2.8 to blocks for the D-optimal 12 run design for Experiment 3 found using the coordinate exchange algorithm for $\eta = 5, \rho = 0.5, 0.75$.

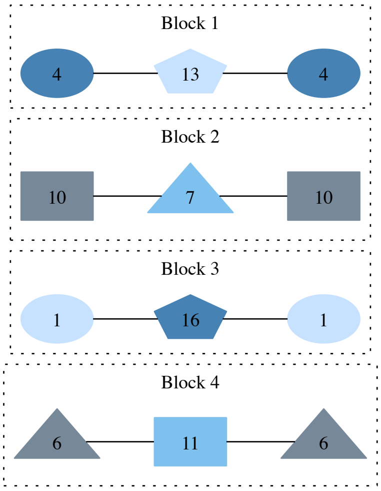


Figure 2.8: Allocation of treatments in Table 2.8 to blocks for the D-optimal 12 run design for Experiment 3 found using the coordinate exchange algorithm for $\eta = 7.5, \rho = 0.25$.

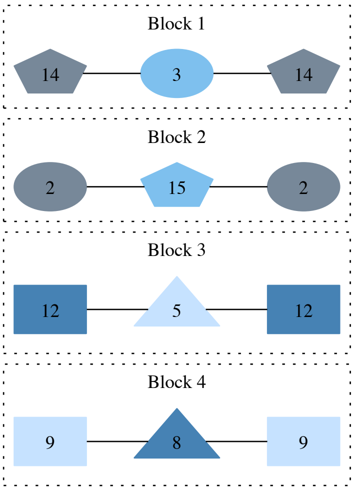


Figure 2.9: Allocation of treatments in Table 2.8 to blocks for the D-optimal 12 run design for Experiment 3 found using the coordinate exchange algorithm for $\eta = 7.5, \rho = 0.75$.

Other designs for Experiment 3 found using the coordinate exchange algorithm which have the same value of (1.18) do not have identical treatments, treatment ordering or allocation. For example, as shown in Figures 2.6 and 2.7, the designs for $\eta = 5$, $\rho = 0.25, 0.5, 0.75$, which have 100% relative efficiency, have identical treatments but different allocations and ordering of these treatments, and, as shown in Figures 2.8 and 2.9, the designs for $\eta = 7.5, \rho = 0.25, 0.75$, which have the same value of (1.18), have completely different subsets of the full factorial, treatment allocation and ordering.

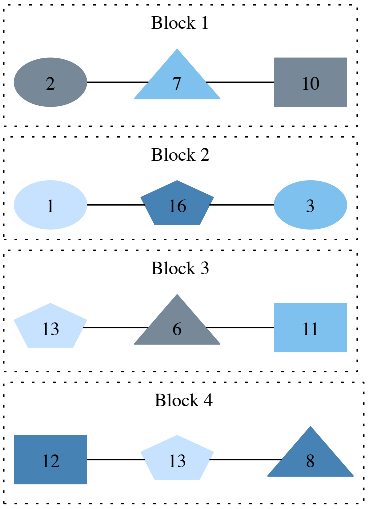


Figure 2.10: Allocation of treatments in Table 2.8 to blocks for the D-optimal 12 run design for Experiment 3 found using the interchange algorithm for $\eta = 10, \rho = 0.25, 0.5, 0.75$.

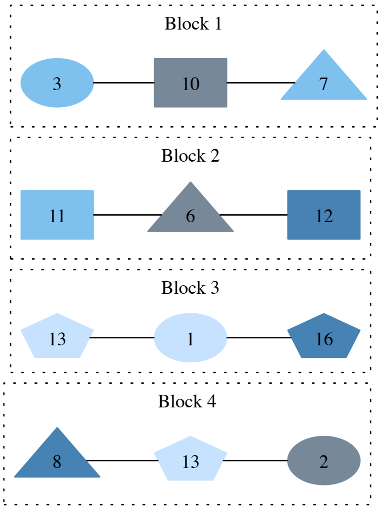


Figure 2.11: Allocation of treatments in Table 2.8 to blocks for the D-optimal 12 run design for Experiment 3 found using the interchange algorithm for $\eta = 7.5, \rho = 0.5$.

The designs with 100% relative D-efficiency found using the interchange algorithm for Experiment 3 will have the same treatments, as the interchange algorithm allocates the D-optimal design (found using the coordinate exchange algorithm for $\rho = \eta = 0$) to blocks. However, these designs do not have the same pattern as the designs found using the coordinate exchange algorithm, and do not necessarily have the same treatment ordering or allocation to blocks. Figure 2.10 shows the allocation of the treatments in Table 2.8 for the design for Experiment 3 found using $\eta = 10$ and $\rho = 0.25, 0.5, 0.75$, which all have the same value of (1.18). However, through comparison of Figures 2.10 and 2.11, we note that the designs for $\eta = 10$ and $\rho = 0.25, 0.5, 0.75$ and $\eta = 7.5$ and $\rho = 0.5$, which also have the same value of (1.18), do not have the same allocation or ordering of treatments.

Designs for Experiment 4 found using the coordinate exchange algorithm with identical (1.18) values do not use the same subset of the treatments in the full factorial, or have the same treatment allocation and ordering within blocks. For example, the D-optimal designs found using the coordinate exchange algorithm for $\eta = 10$, $\rho = 0.25, 0.75$ have identical (1.18) values but, as can be seen from comparing Figures 2.12 and 2.13, they do not have identical treatments, treatment allocation or treatment order.

Designs for Experiment 4 found using the interchange exchange algorithm with identical (1.18) values do not have the same treatment allocation and ordering within blocks. For example, the design found using the interchange exchange algorithm for $\eta = 5$, $\rho = 0.25$ has the same value of (1.18), but not the same treatment allocation and ordering, as the design found using the interchange algorithm for $\eta = 5$, $\rho = 0.5$. This can be seen by comparing Figures 2.14 and 2.15.

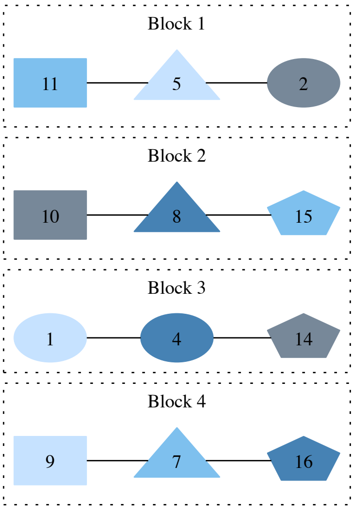


Figure 2.12: Allocation of treatments in Table 2.8 to blocks for the D-optimal 12 run design for Experiment 4 found using the coordinate exchange algorithm for $\eta = 10, \rho = 0.25$.

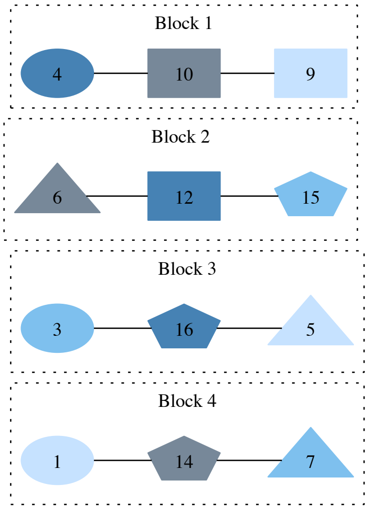


Figure 2.13: Allocation of treatments in Table 2.8 to blocks for the D-optimal 12 run design for Experiment 4 found using the coordinate exchange algorithm for $\eta = 10, \rho = 0.75$.

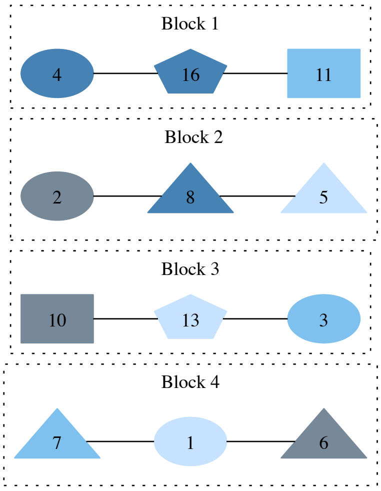


Figure 2.14: Allocation of treatments in Table 2.8 to blocks for the D-optimal 12 run design for Experiment 4 found using the interchange algorithm for $\eta = 5, \rho = 0.25$.

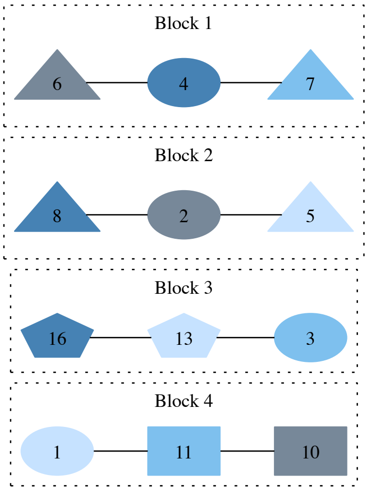


Figure 2.15: Allocation of treatments in Table 2.8 to blocks for the D-optimal 12 run design for Experiment 4 found using the interchange algorithm for $\eta = 5, \rho = 0.5$.

Therefore, we note that designs with 100% relative D-efficiency, and hence the same value of (1.18), do not necessarily have the same structure. The designs for Experiment 2 found using both the coordinate exchange and interchange algorithm have the same set of treatments. However designs for Experiment 2 with the same value of (1.18) did not have the same treatment allocation or ordering of treatments within blocks. The designs for Experiment 3 found using the coordinate exchange algorithm all had blocks with repeated treatments, however designs with 100% relative efficiency did not necessarily share treatments or have the same treatment allocation or ordering. The designs for Experiment 3 found using the interchange algorithm and the designs for Experiment 4 found using both algorithms which are 100% efficient with respect to each other also do not necessarily have the same treatment allocation or ordering.

2.5 Performance of the Coordinate Exchange and Interchange Algorithm

In this work, we used the coordinate exchange algorithm to find designs which maximise (1.18), and the interchange algorithm to allocate the D-optimal completely randomised design to blocks assuming AR(1) correlated errors. In this section, we want to compare the designs found using these two algorithms, as the interchange algorithm is more computationally efficient when finding designs for multiple ρ and η than the coordinate exchange algorithm. We also consider the performance of these algorithms by comparing the designs found using the algorithm to random re-orderings of the treatments in the D-optimal completely randomised design.

The relative efficiencies of designs for Experiments 2, 3 and 4 when $\rho \in \rho^*_{-0}$ and $\eta \in \eta^*_{-0}$ found using the coordinate exchange and interchange algorithm, where (1.18) for the design found using the interchange algorithm is the numerator of (1.20) and (1.18) for the design found using the coordinate exchange algorithm is the denominator of (1.20), are in the interval [85, 100]%, where only certain designs for Experiment 2 had 100% efficiency. There is therefore a gain in efficiency for using the coordinate exchange and not the interchange algorithm, despite the additional computational expense. This also suggests selecting the treatments and allocating them to blocks, instead of assigning a D-optimal completely randomised design to blocks, has some benefit.

Figure 2.3 in Section 2.4.3 shows the allocation to blocks and order of treatments within blocks for the design for Experiment 2 found using the coordinate exchange algorithm when $\eta = 10, \rho = 0.75$, and Figure 2.5 in Section 2.4.3 shows the allocation to blocks and order of treatments within blocks for the design for Experiment 2 found using the interchange algorithm when $\eta = 10$, $\rho = 0.75$. These two designs have 100% relative D-efficiency however they do not have the same treatments allocated to blocks or order of treatments within blocks. This is a representative example, as patterns in the treatment allocation or ordering could not be identified for the other designs found using different algorithms which have the same value of (1.18).

We assess the performance of both the coordinate exchange and interchange algorithm using the D-efficiency of random re-orderings of the treatments in the D-optimal completely randomised design relative to the designs found using the algorithm, which maximise (1.18). If these relative D-efficiencies are high, then the benefit gained by using an algorithm to allocate treatments to blocks is low.

Tables 2.9 and 2.12 give the interquartile ranges of the D-efficiencies for 1,000 random re-orderings of the D-optimal completely randomised design relative to the designs found using the coordinate exchange and interchange algorithms for Experiments 2 and 4, respectively. Tables 2.10 and 2.11 give the interquartile ranges of the D-efficiencies for 1,000 random re-orderings of the D-optimal completely randomised design relative to the designs found using the coordinate exchange and interchange algorithms, respectively, for Experiment 3. Note that (1.18) for the random re-ordering of treatments in the completely randomised designs is the numerator in (1.20) and (1.18) for the D-optimal design for the ρ and η value given by the column and row heading, respectively, is the denominator in (1.20).

We report the efficiencies for both algorithms in Tables 2.9 and 2.12 as they are identical. We also notice that the efficiencies in Tables 2.10 and 2.11 are very similar. Hence, the improvement which can be made by using an algorithm to allocate treatments are identical for Experiments 2 and 4, and very similar for Experiment 3.

ρ η	0	0.25	0.5	0.75
0	[100.00, 100.00]	[95.89,99.48]	[89.70,97.59]	[82.07,94.53]
2.5	[86.90,94.26]	[84.10,94.29]	[82.50,94.24]	[79.60,93.16]
5	[85.19,93.66]	[82.84,93.72]	[81.62,93.69]	[78.85,92.73]
7.5	[84.50,93.43]	[82.29,93.51]	[80.87,93.20]	[78.46,92.55]
10	[84.12,93.31]	[82.00,93.39]	[80.57,93.05]	[78.32,92.45]

Table 2.9: Interquartile range of D-efficiencies for random re-orderings of treatments in the D-optimal completely randomised design for Experiment 2 relative to the design found using the coordinate exchange and interchange algorithms which maximise (1.18) for the η and ρ given by the row and column headings, respectively (%, 2dp).

ρ η	0	0.25	0.5	0.75
0	[100.00, 100.00]	[80.26,84.53]	[65.34, 72.66]	[54.72,64.54]
2.5	[76.17,87.67]	[66.05,77.32]	[58.67,69.66]	[52.56,63.79]
5	[72.90,86.28]	[64.09,76.36]	[57.42,69.04]	[51.99,63.57]
7.5	[71.82,85.72]	[63.36,76.04]	[56.90,68.86]	[51.67,63.46]
10	[71.24,85.43]	[62.87,75.85]	[56.60,68.72]	[51.50,63.37]

Table 2.10: Interquartile range of D-efficiencies for random re-orderings of treatments in the D-optimal completely randomised design for Experiment 3 relative to the design found using the coordinate exchange algorithm which maximise (1.18) for the η and ρ given by the row and column headings, respectively (%, 2dp).

ρ η	0	0.25	0.5	0.75
0	[100.00, 100.00]	[85.11,89.63]	[73.23,81.43]	[63.71,75.14]
2.5	[76.17,87.67]	[70.42,82.43]	[65.65,77.96]	[61.07,74.11]
5	[72.99,86.38]	[68.36,81.45]	[64.24,77.24]	[60.38,74.11]
7.5	[71.95,85.88]	[67.60,81.13]	[63.66,77.03]	[59.99,73.67]
10	[71.39,85.62]	[67.08,80.93]	[63.31,76.87]	[59.78,73.56]

Table 2.11: Interquartile range of D-efficiencies for random re-orderings of treatments in the D-optimal completely randomised design for Experiment 3 relative to the design found using the interchange algorithm which maximise (1.18) for the η and ρ given by the row and column headings, respectively (%, 2dp).

η \ ρ	0	0.25	0.5	0.75
0	[100.00, 100.00]	[96.75,98.05]	[93.21,96.21]	[88.52,94.53]
2.5	[87.29,95.00]	[86.73,94.65]	[86.11,94.39]	[85.07,94.17]
5	[84.97,94.51]	[84.65,94.31]	[84.33,94.18]	[83.83,94.07]
7.5	[83.93,94.32]	[83.70,94.19]	[83.49,94.09]	[83.19,94.02]
10	[83.33,94.22]	[83.16,94.12]	[83.00,94.05]	[82.79,94.00]

Table 2.12: Interquartile range of D-efficiencies for random re-orderings of treatments in the D-optimal completely randomised design for Experiment 4 relative to the design found using the coordinate exchange and interchange algorithms which maximise (1.18) for the η and ρ given by the row and column headings, respectively (%, 2dp).

We notice that for all three experiments and both algorithms, the efficiencies in Tables 2.9 to 2.12 decrease as η and ρ increase, hence the importance of using an algorithm to allocate treatments to blocks and order treatments within blocks increases as ρ and η increases.

The efficiencies in Table 2.12 are higher than those in Tables 2.10 and 2.11, which suggests that as the design approaches saturation (as p approaches n), the benefit, with respect to D-efficiency, of using an algorithm to allocate treatments to blocks and order treatments to blocks for designs with autocorrelated intrablock errors decreases.

Therefore, we note that there is some benefit, with respect to D-efficiency, in using the more computationally expensive coordinate exchange algorithm over the interchange algorithm, and that designs with the same value of (1.18) for these two algorithms do not have the same treatment allocation or ordering. Also, we note that there is a benefit from using an algorithm to allocate and order treatments, and this benefit increases as ρ, η and $n - p$ increase.

2.6 Further Reading: Design and Analysis of Experiments with Correlated Errors

In this section we present some of the key literature regarding the design and analysis of experiments for a variety of different correlation structures with and without blocking factors. We also discuss some of the algorithms used to find designs with correlated errors, a number of which depend on classical, rather than optimal, design ideas. Classical designs have standard structures which can be defined without the use of a computer, and include fractional factorials. Optimal designs select, allocate and order treatments so a particular objective function is optimised.

The literature in this section considers nearest neighbour (NN), moving average (MA) and autoregressive integrated moving average (ARIMA) correlation structures for experiments with and without blocks. Autoregressive correlation structures, such as the AR(1) intrablock error structure we have assumed in this chapter, are also considered in certain literature.

The literature in this section includes first order NN designs (NN1) and second order nearest neighbour designs (NN2). The order of a NN design indicates the number of neighbours effected by the residual effect from applying a treatment to a particular experimental unit. The correlation between neighbours is assumed to be constant, and does not decay as the distance between the experimental units increases for all NN designs with order greater than or equal to two.

The majority of the work on designs for NN correlation has focused on experiments where the experimental units are arranged in rows and columns, such as an experiment where the experimental units are sections of a field, which is subdivided into rows from north to south and columns from east to west. An important consideration in designs where the experimental units are arranged into rows and columns is row-column balancing. A design is said to be row-balanced if each treatment is applied to neighbouring experimental units in rows an equal number of times and is column-balanced if each treatment is applied to neighbouring experimental units in columns an equal number of times. A design is nearest neighbour balanced if it is balanced in both rows and columns.

MA processes can be seen as extensions of NN processes, as they assume that the responses for a certain number of experimental units close to the unit to which the current treatment is being applied are correlated. However, unlike NN correlation, MA processes do not assume this correlation is constant. The most common MA process discussed in the literature for experiments with correlated errors is the moving average process of order 1 (MA(1) process),

$$\epsilon_i = \alpha_i + \theta\alpha_{i-1}, \tag{2.12}$$

where ϵ_i is the error for the ith run, α_i and α_{i-1} are random error terms from a given (usually normal) distribution and θ denotes the relationship between the error from the experimental unit the treatment is currently being applied to and its closest neighbour.

AR processes differ to MA processes as they assume that the correlation between the response for all experimental units is non-zero, and decays as the spatial or temporal distance between experimental units increases. ARIMA processes combine AR and MA processes, and assume that the response from the experimental units which are close together have a stronger, fixed, correlation than those further away, but also allows the correlation for those further away to decay. The ARIMA process of order 1

(ARIMA(0,1,1)) is

$$\epsilon_i = \epsilon_{i-1} + \alpha_i + \theta\alpha_{i-1}, \tag{2.13}$$

where α_i, α_{i-1} and θ are as defined in (2.12) and ϵ_{i-1} is the error for the response from the closest experimental unit. ARIMA(0,1,1) is a combination of an AR(1) and MA(1) process.

Optimal or Efficient Designs for Experiments with Correlated Errors

Williams (1952) was the first author to suggest methods other than randomisation to neutralise correlation between observations in an experiment and provided combinatorial methods, based on neighbour balance and the maximum likelihood equations, for the design and analysis of experiments with AR(1) and AR(2) correlated errors. Williams (1952) also showed that variances can be underestimated if the autoregressive nature of the errors is ignored. There is a wealth of literature for experiments with correlated errors, with and without blocking, that can be linked back to the work of Williams (1952).

NN correlation is the simplest correlation structure that allows for some spatial or temporal dependency, and the design of NN experiments and the analysis of responses when NN correlation is assumed is popular in literature such as Freeman (1979), Kiefer and Wynn (1981) and Morgan and Chakravarti (1988). Freeman (1979) discussed the use of complete Latin squares as balanced two-dimensional nearest neighbour designs, using the specific example of plant breeding as motivation. Complete Latin squares are Latin squares where all pairs of treatments are only adjacent once in all rows and columns. For further detail regarding the use Latin squares for experimental design see, for example, Bailey (2008, Chapters 6 and 9) and Montgomery (2012, Section 4.2).

The often cited paper by Kiefer and Wynn (1981) also advocated the use of Latin squares to construct NN1 designs, and defined conditions for weak universal optimality. The weak universal optimality criterion presented in Kiefer and Wynn (1981) includes A- and E-optimality, but does not include D-optimality, and has then been used by other authors such as Morgan and Chakravarti (1988).

Kiefer and Wynn (1981) found designs with and without fixed block effects, which were robust to misspecifying a process with NN correlation structure as uncorrelated, using classical design ideas. They identified classical combinatorial designs (such as Latin squares) which are optimal with respect to their universal optimality criteria assuming the errors are uncorrelated, and then selected designs from this set which are optimal for NN1 correlation.

Morgan and Chakravarti (1988) extended the work of Kiefer and Wynn (1981) and found robust block designs for both NN1 and NN2 correlation. Morgan and Chakravarti

(1988) also provided conditions for balanced incomplete block designs (BIBDs) to be universally optimum for NN1 and NN2 correlation structures using type II optimality defined by Takeuchi (1961). BIBDs are block designs where all treatments do not appear in each block, but each pair of treatment occurs in blocks λ times, when $\lambda > 1$. A catalogue of BIBDs was given by Fisher and Yates (1963). The objective function for type II optimality minimises the maximum variance of estimated treatment contrasts.

Morgan and Chakravarti (1988) assumed that the nearest neighbour correlation would be small, and hence their primary aim was to find designs which protect against unexpected correlation rather than assume this type of correlation exists in the experiment. They used this aim to justify their use of ordinary rather than generalised least squares when estimating pairwise treatments contrasts.

Other correlation structures such as AR(1), MA(1) and ARIMA(0,1,1), have also been considered by a number of authors. Kiefer (1961) proved optimality results for designs without blocks given in the paper by Williams (1952) for multiple criteria, including D-optimality. The work of both Williams (1952) and Kiefer (1961) was extended in the often cited paper by Kiefer and Wynn (1984), who found designs without blocks for AR(1) correlated errors for treatment comparisons with respect to a criteria based on A-optimality.

Berenblut and Webb (1974) used a similar approach to that given in the paper by Kiefer and Wynn (1981), as they found designs without blocks which maximise the D-optimality objective function without correlation (for $\mathbf{V} = (\sigma_\gamma^2 + \sigma_\epsilon^2)\mathbf{I}_n$) and then selected the designs from this set which minimise the D-optimality objective function for AR(1) correlated errors. They compared the efficiency of their designs to those found by Williams (1952), and stated that a design is robust if information regarding the parameters can be extracted efficiently from the results irrespective of the error structure. In Section 2.4.2, we considered the robustness of block designs which maximise the D-optimality objective function for errors which follow an AR(1) process assuming two non-zero ρ, and did not use the two-step approach to identify designs discussed by Berenblut and Webb (1974).

Martin et al. (1998a,b) discussed the properties of optimal designs for factors without blocks for three different types of correlated errors; AR(1), MA(1) and ARIMA(0,1,1). Martin et al. (1998a) showed that designs for two-level factors without blocks found using the algorithm in Cheng and Steinberg (1991) , which are D-optimal for AR(1) correlation, are also often almost A- and E-optimal for MA(1) and ARIMA(0,1,1) correlation. Cheng and Steinberg (1991) suggested that their designs were optimal due to the number of level changes between factors, however this was shown to be false by Martin et al. (1998a).

Martin et al. (1998b) extended the theoretical results presented in Martin et al. (1998a). Martin et al. (1998b) found designs for multi-level factors without blocks for AR(1), MA(1) and ARIMA(0,1,1) correlated errors. Multi-level factors are any factors with more than two levels. Martin et al. (1998b) presented the properties required for a multi-level design to be D- and A-optimal, which include good neighbour balance and a small number of self adjacencies for all lags. The number of self adjacencies at lag g is the number of times two levels occur in two treatments which are applied g runs apart. Both Martin et al. (1998a) and Martin et al. (1998b) considered main effects models for responses from designs without blocks, whereas we also considered interactions in our models, as discussed in Section 2.4.

Kunert (1987) showed that balanced NN designs from Gill and Shuka (1985) are universally optimal block designs when the responses are assumed to have AR(1) errors with positive ρ. The universal optimality criteria presented in Kunert (1987) includes D-optimality as a special case. Kunert (1987) also showed that neighbour balanced BIBDs have promising optimality results for positive ρ and AR(1) correlated errors. In this chapter, we focused on using coordinate exchange and interchange algorithms to find designs which maximise (1.18), instead of using classical designs such as BIBDs and Latin squares.

Jin and Morgan (2008) proved properties required for saturated block designs with AR(1) correlated errors to be optimal for a range of optimality criteria, including D-optimality, by extending the work of Chakrabarti (1963); Bapat and Dey (1991) and Bagchi and Bagchi (2001). While we also considered a saturated block design with AR(1) correlated errors in Section 2.4, we used factorial effects in the model assumed for the responses whereas Jin and Morgan (2008) considered pairwise treatment comparisons. Therefore, the properties defined in Jin and Morgan (2008) are not necessarily appropriate for the work in this chapter.

More general correlation structure results are given by Bischoff (1992), who derived the conditions required for a design without blocks to be D-optimal-invariant with respect to two general and unknown correlation structures. Bischoff (1992) stated that a design is D-optimal-invariant when it has the same D-optimality objective function value for two different correlation structures and the best linear unbiased estimates (BLUE) of the fixed effect parameter $\boldsymbol{\beta}$ in the linear model fitted to responses from the design is invariant to the change in the correlation structure. The BLUE of $\boldsymbol{\beta}$, $\tilde{\boldsymbol{\beta}}$, can be expressed as a linear function of $\mathbf{Y}$, $\mathrm{E}(\tilde{\boldsymbol{\beta}})=\boldsymbol{\beta}$, and it has the smallest variance among all unbiased linear estimators of $\boldsymbol{\beta}$. Whilst we considered invariance with respect to the D-optimality objective function in Section 2.4.2 and 2.5, we did not assess the impact of adjusting ρ on the specific estimates of $\boldsymbol{\beta}$.

Algorithms for Finding Optimal Designs for Experiments with Correlated Errors

Algorithms for identifying optimal designs for correlated observations either rely on reordering the runs in classical designs or point exchange algorithms. A point exchange algorithm selects treatments from a set of candidate treatments, and orders these treatments using an interchange procedure. The algorithms presented in the papers by Constantine (1989) and Garroi et al. (2009) select the run order of classical designs such as Latin squires and BIBDs. The algorithms by authors such as Jones and Eccleston (1980) and Chan and Eccleston (2003) find optimal designs using exchange and interchange procedures.

In this work we found block designs with AR(1) intrablock errors which maximise (1.18) using both a coordinate exchange algorithm (Section 1.4.3 in Chapter 1) and interchange algorithm (Section 1.4.4 in Chapter 1). The coordinate exchange algorithm, which is based on the algorithm by Meyer and Nachtsheim (1995), chooses the treatments and, due to the structure of the variance-covariance matrix for responses with autoregressive intrablock errors, also implicitly chooses the order of the runs within blocks. The interchange algorithm, based on the algorithm described by Atkinson et al. (2007), does not choose treatments but allocates and orders the treatments in a completely randomised designs using the assumption of AR(1) intrablock errors.

Constantine (1989) presented a method of identifying D-efficient designs without blocks for nearest neighbour correlation. The designs found using this method are D-efficient when compared to the design for experimental units with no correlation. The method uses Hadamard matrices as the starting design and relies on reordering the columns of the Hadamard matrices and multiplying rows of the matrices formed by this reordering to identify designs with a high D-efficiency. Hadamard matrices are discussed and tabulated by Hedayat and Wallis (1978). Constantine (1989) concluded that, based on the designs found from their algorithm, designs with multiple level changes are efficient when the NN correlation is assumed to be positive.

The results discussed by Cheng and Steinberg (1991) complimented the results found by Constantine (1989). Cheng and Steinberg (1991) began by demonstrating that reordering full and fractional factorials so the maximal number of level changes occur between each run produces the most efficient designs without blocks for errors which follow AR(1) processes and time series trend models. Time series trend models (Spezzaferri, 1988) replace the intercept in the model assumed for the response with a vector whose elements give the level of time trend expected for each observation.

Both Constantine (1989) and Cheng and Steinberg (1991) computed the D-efficiency of all their designs in comparison to the design with no correlation. In Section 2.4.2, we compared the D-efficiency for designs for different, non-zero, values of ρ, as well as considering the D-efficiency of designs for $\rho = 0$ relative to designs for $\rho > 0$. Cheng and Steinberg (1991) presented a reverse fold-over algorithm based on the work of Cheng (1985) and Coster and Cheng (1988), which finds the run order with the maximum number of level changes for two-level fractional factorial and full factorial designs.

Garroi et al. (2009) presented a variable-neighbourhood search algorithm to identify D-optimal run orders of central composite designs without blocks for experiments with AR(1) correlated errors. Central composite designs are a combination of a full or fractional factorial (depending on the number of runs in the experiment), centre points and axial points. For an experiment with f factors, there will be $2f$ axial points, where the rth pair of axial points have the levels $\pm\alpha$ for factor r and 0 for all other factors and α is chosen by the experimenter. For example, the 1st pair of axial points for an experiment with 4 factors will be (α, 0,0,0) and (-α, 0, 0, 0). Common choices for α include 1 and $\sqrt[4]{f}$. Garroi et al. (2009) used three-level factors as they included second order (polynomial) terms in their model for the responses, whereas our designs have two-level factors as we only considered first order and interaction terms in our models.

The variable-neighbourhood search algorithm in the paper by Garroi et al. (2009) considers six different types of perturbations that can be made by the design, which includes swaps, moves and relabelling of factors. The algorithm uses a steepest-ascent move strategy; that is, all solutions for that neighbourhood are generated and then the D-optimal perturbation is selected before moving to the next neighbourhood.

Garroi et al. (2009) used their algorithm to find designs which they showed to be robust to misspecification of the autocorrelation parameter ρ. They also demonstrated the loss in D-efficiency which occurs when it is assumed that $\boldsymbol{\beta}$ should be estimated using ordinary, and not generalised, least squares. The factorial part of the D-optimal designs for AR(1) correlation found by Garroi et al. (2009) exhibited features similar to those obtained by Constantine (1989), Cheng and Steinberg (1991) and Martin et al. (1998a), as the number of level changes between runs in the designs found by Garroi et al. (2009) is large, but not necessarily maximised.

Jones and Eccleston (1980) presented an algorithm to find optimal block designs for exchangeable correlation structure, with the aim of estimating comparative treatment effects. Unlike the algorithms of Constantine (1989), Cheng and Steinberg (1991) and Garroi et al. (2009), which rely on classical designs, Jones and Eccleston (1980) used random starting designs and candidate sets of potential treatments.

The algorithm given by Jones and Eccleston (1980) has an exchange and an interchange procedure. The exchange procedure deletes the weakest observation in the current design and replaces it with the strongest observation from the candidate set. The weakest and strongest observations are chosen based on their impact on a modified form of A-optimality. The exchange procedure is repeated until no further improvements can be made. The interchange procedure swaps treatments in the design until the sum of the weighted variance of the treatment contrasts is maximised. In this work, using the coordinate exchange algorithm allows us to identify the treatments and assign them to blocks using one procedure instead of two.

Satpati et al. (2007) extended the work of Jones and Eccleston (1980) and Zergaw (1989) by presenting an exchange-interchange algorithm to find efficient block designs with NN1 and AR(1) correlation for the estimation of comparative treatment effects. The exchange procedure used by Satpati et al. (2007) is the same as Jones and Eccleston (1980). However, the interchange procedure in the paper by Satpati et al. (2007) calculates the criterion value for all pairwise swaps and then maintains the swap that gives the most significant improvement in the objective function, and repeats this process until no further improvement is possible.

Satpati et al. (2007) found A- and D-optimal designs for pairwise treatment contrasts, not factorial effects, and assessed the robustness of their designs using A- and D-efficiency. They found designs which are robust to misspecification of the fixed values of the correlation parameters that they consider, [-0.5, 0.5] in steps of 0.05 for NN1 and [-0.95, 0.95] in steps of 0.05 for AR(1), and produced a catalogue of these designs.

Tack and Vandebroek (2002) presented an exchange algorithm for block designs with time dependent observations. However, rather than considering the time trend in the error structure, Tack and Vandebroek (2002) included time as an additional fixed effect in the model fitted to the responses from the experiment. They also included a fixed effect to represent the cost associated with running each treatment and the cost of changing the level of each factor in a treatment.

The exchange algorithm in the paper by Tack and Vandebroek (2002), finds D-optimal designs for the model with fixed time and cost effect and extends the point exchange algorithms of Atkinson and Donev (1989, 1996). The algorithm iteratively selects, with replacement, treatments from a course grid of candidate points until no further improvement can be made with respect to the D-optimality objective function. This differs from the algorithms in Jones and Eccleston (1980) and Satpati et al. (2007), as it has a single exchange procedure rather than a two-step exchange-interchange procedure.

Elliot et al. (1999) and Chan and Eccleston (2003) presented stochastic search algorithms based on simulated annealing. Algorithms based on simulated annealing (Aarts and van Laarhoven, 1989) accept or reject proposed changes to the current optimal design with some probability. The potential for accepting sub-optimal moves, where the objective function value has not improved, means that the algorithm can move away from local optima.

The algorithm in the paper by Elliot et al. (1999) is very general and can be used to find factorial designs with or without blocks for a range of optimality criteria and correlation structures. This algorithm optimises random initial designs using an annealing routine followed by a steepest descent routine. The simulated annealing routine swaps two randomly chosen runs in the current design with some probability. The steepest descent routine considers all possible pairwise exchanges of treatments in the design found from the annealing routine, and keeps any exchanges that improve the value of the objective function.

Chan and Eccleston (2003) provided algorithms to find NN balanced designs based on simulated annealing and tabu search (Glover, 1989; Gill, 1990). These algorithms enable designs with NN balance, which cannot be found using the combinatorial methods discussed in Chan and Eccleston (1998), to be identified. The algorithms rely on the minimisation of two objective functions proposed by Chan and Eccleston (2003), which are dependent on the number of times treatments are repeated in rows and columns, and the square of the number of times a treatments are neighbours. For NN balance, the first objective function should be zero, however as this is not possible for all designs and so partial neighbour balance can be achieved by minimising this objective function.

Analysis of Experiments with AR(1) Errors

Pantula and Pollock (1985) focused on the analysis of longitudinal studies, where the responses from n individuals are recorded at t_i, $i = 1, \dots, n$ consecutive time points. Pantula and Pollock (1985) assumed that the results for each individual have errors which follow a stationary AR(1) process. If we use the individual as a blocking factor, and times points as units within blocks, the linear model considered by Pantula and Pollock (1985) is identical to (1.3) with errors given by (2.6), which is used in this chapter.

Pantula and Pollock (1985) presented methods for estimating the fixed effect parameter $\boldsymbol{\beta}$, the between block variation σ^2_γ, the within-block variation σ^2_ϵ, and the autoregressive parameter ρ. They extended the approach discussed in the paper by Fuller and Battese (1973), and applied a transformation to the correlated errors to obtain generalised least square estimates of $\boldsymbol{\beta}$ and consistent estimators for σ^2_γ and σ^2_ϵ in (2.4), and ρ in (2.6). An estimator $\hat{\theta}$ is consistent if it converges in probability to the true parameter θ as the number of experimental runs tends to infinity.

Pantula and Pollock (1985) followed the approach of Andersen et al. (1981) to obtain a method of moments estimator for ρ. In earlier work, Azzalini (1984) established the method of moments estimators for the case where t_i is constant for all n individuals, which is more closely related to the experiment described in Section 2.1. The results in the paper by Pantula and Pollock (1985) were extended by Schaalje et al. (1991) to more complex mixed models with several random effects that vary over time.

2.7 Discussion

To meet the aims of our collaboration with the ORC (Sections 1.2.1 and 2.1), we used the coordinate exchange and interchange algorithms (Section 1.4.3 and 1.4.4 in Chapter 1) to find block designs which maximise the D-optimality objective function (1.18) for autoregressive intrablock errors. We assessed how saturated and unsaturated designs compare in terms of relative D-efficiency (1.20), investigated how robust designs are to misspecification of the autoregressive parameter ρ and the ratio of variances η with respect to D-efficiency, and compared the treatment selection, allocation to blocks and order within blocks for designs with the same value of (1.18). We also compared the design found using the coordinate exchange and interchange algorithm, and assessed the performance of these algorithms.

We use random block effects in the models for the responses and the D-optimality objective function. D-optimality is used when the aim of the experiment is to gain scientific understanding about $\boldsymbol{\beta}$. This is important for the application of the work in this chapter, as the ORC wish to find out which factors are most significant in the manufacture of fibres. Random block effects allow us to make predictions for unobserved blocks using the results from the experiment. Also, due to the time and cost involved in producing fibres, the ability to extend the results found and predict for future fibres would be beneficial for identifying new factor settings for future experiments.

There are two parameters, ρ and η, which are unknown prior to experimentation and may affect design performance. It would be beneficial if the D-optimal designs found are not reliant on these parameters. We found designs which maximise (1.18) for $\rho^* = \{0, 0.25, 0.5, 0.75\}$ and $\eta^* = \{0, 2.5, 5, 7.5, 10\}$. In Section 2.4.1, we found that the importance of allocation and ordering of treatments increased as $\rho \in \rho^*$ and $\eta \in \eta^*$ increased.

In Section 2.4.2, we found that designs found using both algorithms for $\rho \in \rho^*_{-0}$ and $\eta \in \eta^*_{-0}$, where $\rho^*_{-0} = \{0.25, 0.5, 0.75\}$ and $\eta^*_{-0} = \{2.5, 5, 7.5, 10\}$, are robust to misspecification. We noted, however, that there may be a relationship between the intrablock correlation and robustness. Therefore, a potential future work is to find designs with larger blocks and see if, as assumed, these designs are less robust to misspecification of ρ and η.

In Section 2.4.3, we used examples to prompt our discussion on the structure of designs with 100% D-efficiency, and concluded that designs with the same value of (1.18) do not necessarily have the same treatments, allocation of treatments to blocks or ordering of treatments within blocks. We did note, however, that the designs for Experiment 3 found using the coordinate exchange algorithm all had treatments which were repeated at the start and end of each block. Therefore, the work in Section 2.4.3 could be extended by finding designs with different run sizes which maximise (1.18) when the main effects model is assumed for the response and identifying whether block designs with autoregressive errors have repeated treatments at the start and end of blocks when main effects models are assumed for the response.

Finally, in Section 2.5, we found that the coordinate exchange algorithm finds designs with higher value of (1.18) than the interchange algorithm. We also noted that there is a benefit with respect to efficiency from allocating treatments using an algorithm instead of considering random allocations of the D-optimal completely randomised design, and this benefit increases as ρ, η and $n - p$ increase.

The work in this chapter could be extended by considering other objective functions, a different range of ρ and η, and different algorithms for finding designs. D-optimality is an estimation based criterion that is popular in literature. However, as noted from Section 2.6, A-optimality is also popular in the literature for designs with correlated errors. In future work, we could find designs for these two estimation based criteria and compare them with respect to efficiency, and treatment allocation and ordering.

The current range of ρ and η values is wide, and the number of values considered is quite small. The number and range of ρ and η values considered could be increased, for example we could consider $\eta^* = \{0, 0.5, 1, \dots, 10\}$ and $\rho^* = \{0, 0.05, 0.1, \dots, 0.95\}$, so the robustness of these designs could be considered with more generality. Alternatively, the experimental evidence from the ORC could be used to estimate η and ρ, and a range based on the confidence intervals for the estimates of these parameters could be considered, to make the designs more suitable for the manufacture of microstructured fibres.

A stochastic search algorithm, such as the simulated annealing algorithm (Aarts and van Laarhoven, 1989), could be used to find optimal block designs. Stochastic search algorithms accept moves based on a certain probability, and therefore can accept moves which do not maximise the objective function in order to escape local optima. The designs found using this algorithm could be compared to the designs found using the coordinate exchange algorithm.

Chapter 3

Optimal Designs for Multi-Stage Experiments

Multi-stage experiments are prevalent in industry and science. They use the same experimental units in multiple stages of experimentation, with distinct responses measured at each stage. Two models are fitted to each distinct response; a model which relates the response from the sth stage to the sth stage factors, and a "cumulative" model which relates the response from the sth stage to the factors from stages $s, (s-1), (s-2), \ldots, 2, 1$. The use of cumulative models means that the different stages of experimentation have to be designed in conjunction with each other.

In this chapter we use a compound Bayesian D-optimality objective function within a coordinate exchange algorithm to find multi-stage designs tailored to estimation of the parameters in all the models considered. The motivation for this work is the manufacture of optical fibres and the formulation of a pharmaceutical product, as discussed in Section 1.2 of Chapter 1, and the methodology presented in this chapter is used to find a multi-stage design for the experiment discussed in the case study in Chapter 5.

3.1 Definitions and Motivation

3.1.1 Definition of Multi-Stage Experiments

A multi-stage experiment is an experiment that uses the same experimental unit in multiple stages. A sub-treatment is applied to the experimental unit at each run in each stage, and a distinct response is recorded at the end of each stage. A sub-treatment is the combination of factor levels that is applied to the experimental unit at a particular stage, and a treatment is the combination of factor levels that is applied to the experimental unit across all stages of experimentation.

This definition of a stage is the same as the definition of a partition given by Perry et al. (2001, 2002) and discussed in Section 3.3. However, our multi-stage experiments are not partition experiments as we measure a response for each stage at the end of each stage and not at the end of the experiment as a whole. Multi-stage designs are the set of treatments and replications applied in a multi-stage experiment.

While each stage may have separate responses and sub-treatments, the stages are designed in unison as it is assumed that the responses observed at stage s are influenced by the factors varied at stages $1, 2, \ldots, (s-1), s$, and not just by the factors varied in the sub-treatments applied in stage s. However, it is important to note that the impact of varying the factors in stage s on the stage s response may still be of interest. Therefore, two models should be considered for the response from stage s; a model which describes the impact of the stage s factors on the stage s response and a "cumulative" model which describes the impact of the stage $1, 2, \ldots, s$ factors on the stage s response. The consideration of multiple models naturally leads to the use of the compound optimality criterion, as compound criterion allow multiple experimental objectives to be considered when designing experiments. Compound criterion are discussed further in Section 3.4.

3.1.2 Comparison of Multi-Stage and Multi-Tiered Experiments

Multi-tiered experiments are prevalent in literature, and sometimes multi-tiered experiments are defined as multi-stage experiments. However, this definition of multi-stage differs to that presented in Section 3.1.1, and the designs for multi-stage experiments presented in Section 3.6 are not necessarily also multi-tiered. Multi-tiered experiments, which involve multiple randomisations of treatments, were discussed and defined by Brien (1983), Brien and Payne (1999) and Brien and Bailey (2006). Treatments are randomised with limitations, based on the properties of certain factors, and tiers are defined as the sets into which the treatments can be organised, based on their randomisations. For example, split-plot experiments are two-tiered experiments, with one tier relating to the randomisation of whole-plots and the other relating to the randomisation of runs within whole-plots. Strip-plot experiments could also be described as two-tiered experiments, with one tier relating to the allocation of batches to rows and the other relating to the allocation of batches to columns. Split- and strip-plot experiments are discussed in Section 3.2.

We note that, as long as multiple randomisations are used in the experiment, multi-phase and multi-stratum experiments can also be classed as multi-tiered experiments. Multi-phase designs (as discussed by Brien et al., 2011) are used in the literature for experiments when a field phase is performed after a laboratory phase. A field phase is where initial results are collected, for example by applying treatments to wheat in an agricultural setting. A laboratory phase relates to any stage where further processing, measurement or testing is performed, for example measuring the quality of wheat sampled from the field trial in a controlled laboratory environment. Multi-stratum designs (as discussed by authors such as Bailey, 1991 and Trinca and Gilmour, 2001) are multi-tiered designs as each stratum is a tier, where a stratum is a group of treatments based on their randomisation. All multi-tiered designs have a single response measured after all the factors across all tiers have been applied.

Certain literature provides an alternative definition for multi-stage experiments. In Freeman (1959), Trinca and Gilmour (2001) and Brien et al. (2011), multi-stage experiments are defined as experiments that are conducted in distinct time intervals and are described as multi-tiered by Brien and Bailey (2006) if they involve multiple randomisations across these distinct time intervals. The key difference between this definition of multi-stage and the definition that we present in Section 3.1.1 is that we assume a response is measured at the end of each stage.

The treatments in a multi-stage experiment which adheres to the definition in Section 3.1.1 can be randomised with one randomisation if the experiment has no limitations on resource or complicated features. We refer to these multi-stage single-tiered experiments without restrictions on randomisation as completely randomised multi-stage experiments. Multi-stage multi-tiered experiments occur when there are restrictions on the randomisation of the treatments in a multi-stage experiments. Restrictions can occur due to; (i) hard to change factors, (ii) batching of experimental units and (iii) nesting. In Section 3.4 we discuss three different types of optimal designs for multi-stage experiments; (i) multi-stage completely randomised, (ii) multi-stage split-plot and (iii) multi-stage strip-plot designs.

3.1.3 Motivation and Aim of Work

The manufacture of optical fibres and the formulation of a pharmaceutical product, as discussed in Sections 1.2.1 and 1.2.2, respectively, of Chapter 1, motivates the work in this chapter. For both examples, the factors in the experiment are applied in two stages and a response is measured at the end of each stage. Randomisation of some of the factors in each experiment is restricted, as they are either the factors used to manufacture the cane (for optical fibre manufacture) or will increase the cost of the experiment if varied too often (for the formulation of a pharmaceutical product). When formulating the pharmaceutical product, the experimental units may be able to be re-batched after the certain factors have been applied, however this is not possible for the manufacture of optical fibres. Therefore, a two-stage split-plot design is applicable for both examples and a two-stage strip-plot design may be applicable for the formulation of a pharmaceutical product.

We assume that a response is measured at the end of each stage of the experiment and that three models are of interest; the model which relates the factors from Stage 1 to the Stage 1 response, the model which relates the factors from Stage 2 to the Stage 2 response and the model which relates the factors from Stages 1 and 2 to the Stage 2 response (the cumulative model). The aim of the experiment is gain as much information about the parameters which relate the factor to the responses by minimising the volume of the confidence ellipsoid for $\boldsymbol{\beta}$ for a weighted combination of these models, where the weights indicate the relative importance of each of these models to the experimenters. We therefore use the compound Bayesian D-optimality objective function from Section 3.4.2 to find optimal two-stage designs. We use Bayesian D-optimality as it enables us to include our level of prior belief for $\boldsymbol{\beta}$, and also allows us to consider unsaturated, saturated and supersaturated designs, which may occur due to the number of terms in the different models considered.

In this chapter we introduce a coordinate exchange algorithm (Meyer and Nachtsheim, 1995), to find optimal two-stage designs which are appropriate for the formulation of a pharmaceutical product, as this work will also be applied to a case study in Chapter 5. Three types of 12 and 16 run designs for six two-level factors will be considered:

1. two-stage completely randomised designs.

2. two-stage split-plot designs (Section 3.2.2). Factors 1 and 2 in Stage 1 are hard-to-change, and Factors 3, 4, 5 and 6 in Stage 2 are easy-to-change.

3. two-stage strip-plot designs (Section 3.2.3). Factors 1 and 2 in Stage 1 are row factors, and Factors 3, 4, 5 and 6 in Stage 2 are column factors.

In Section 3.6, we use the correlation between columns in the model matrices for these designs, as discussed in Section 3.5.2, to assess these designs. Correlation between columns in the model matrix will impact the estimated model parameters through variance inflation and bias, and the higher the correlation, the more inflated the variance and bias is. Parameters which relate to columns which are correlated will not be able to estimated independently, and will therefore be aliased.

In Section 3.6.1 we compare designs found using the coordinate exchange algorithm with random starting designs to designs found using the coordinate exchange algorithm with starting designs chosen to have good projection properties. The projectivity of a design measures how the design performs when only a subset of factors from the design are considered in a model for a particular response, which is appropriate for multi-stage designs as models are fitted to the responses at the end of each stage.

3.2 Introduction to Multi-Tiered Designs: Split-Plot and Strip-Plot Designs

In this section, we define the single-stage versions of the multi-stage experiments we consider in Section 3.6, and use an example to indicate the impact of restrictions on randomisation.

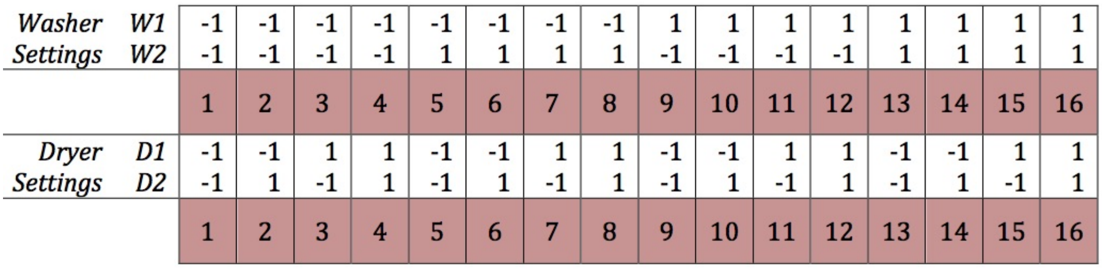
3.2.1 A Simple Example: Washing and Drying Cloths

Suppose a manufacturer of household appliances wants to find methods to reduce the wrinkling of laundry (Miller, 1997). They have 16 cloths (which are the experimental units) and wish to investigate the impact of four two-level factors: two which apply to washing machine (washer) settings and two which apply to tumble dryer (dryer) settings. The response from this experiment is the wrinkling of a cloth sample.

The $2^4 = 16$ possible treatments which can be applied in the process are listed in Table 3.1. The washer factors, or washer settings, are labelled $W1$ and $W2$ and the dryer factors, or dryer settings, are labelled $D1$ and $D2$. For example, $W1$ and $W2$ could be (washer) temperature and spin speed and $D1$ and $D2$ could be (dryer) temperature and time. The application of the treatments in Table 3.1 to cloths is dependent on the amount of experimental resource.

$W1$	$W2$	$D1$	$D2$
-1	-1	-1	-1
-1	-1	-1	1
-1	-1	1	-1
-1	-1	1	1
-1	1	-1	-1
-1	1	-1	1
-1	1	1	-1
-1	1	1	1
1	-1	-1	-1
1	-1	-1	1
1	-1	1	-1
1	-1	1	1
1	1	-1	-1
1	1	-1	1
1	1	1	-1
1	1	1	1

Table 3.1: The 16 treatments for the washing and drying cloths example.



Washer	W1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1	1	1	1	1
Settings	W2	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	1	1	1	1
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
Dryer	D1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1
Settings	D2	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1
		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

Figure 3.1: Allocation of cloths to washers and dryers for a completely randomised design.

A completely randomised design can be used to test the effect of $W1$, $W2$, $D1$ and $D2$ on the wrinkling of each of the 16 cloths if the experimenters can use 16 washers and 16 dryers. The cloths, labelled 1 to 16 in Figure 3.1, can be randomly allocated to the individual washers and dryers in any order. The linear model, (1.6) where (1.5) is $(\sigma_\gamma^2 + \sigma_\epsilon^2)\mathbf{I}_n$ in Section 1.3.1 of Chapter 1 would be used to analyse the data collected from this design.

The experiment described in Figure 3.1 could also be performed by resetting the same washer and dryer 16 times. However, there could be restrictions on the number of machines or the amount of time available, and therefore other types of designs with restricted randomisation, such as split-plot and strip-plot designs, which are discussed in Sections 3.2.2 and 3.2.3 respectively, would need to be considered for this experiment.

3.2.2 Split-Plot Designs

Split-plot designs, as discussed in Section 1.3.1 in Chapter 1, have restricted randomisation based on the levels of hard-to-change factors, which are factors that are either complex or expensive to adjust. Groups of experimental runs based on the level of the hard-to-change factors are referred to as whole-plots. The individual runs within whole-plots are referred to as sub-plots.

Example

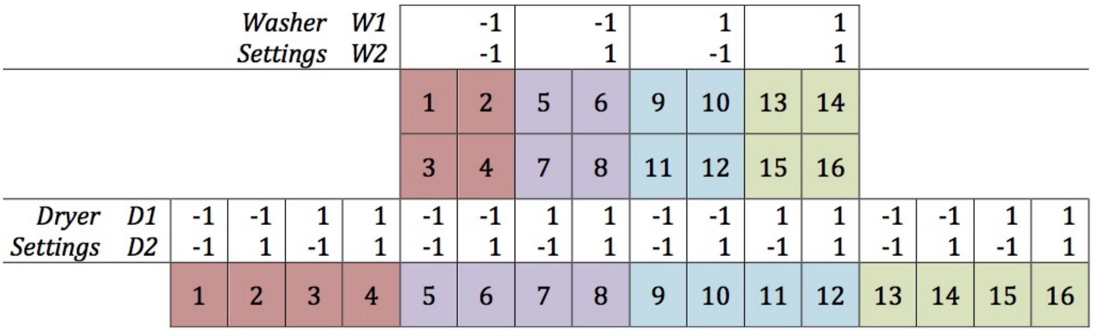


Figure 3.2: Allocation of cloths to washer and dryers for a split-plot design.

Assume that there is resource for four washing machines and 16 dryers in the experiment described in Section 3.2.1, and the experimenters wish to apply all 16 treatments in Table 3.1 to cloths in the experiment. This restriction on experimental resource would require four cloths to be washed in each washing machine, however all 16 cloths could be dried separately.

The split-plot design illustrated in Figure 3.2 would allow the experiment with these restrictions to be performed, where the four different colours represent the washing machine used. The cloths are randomly allocated to four washing machines, one for each pairwise combination of the two levels of $W1$ and $W2$, as long as there are four cloths in each. The four cloths from each washer are randomly allocated to four dryers with settings $(D1, D2)$=$(-1,-1)$, $(-1,1)$, $(1,-1)$ and $(1,1)$.

A Mixed Model for Analysing Split-Plot Designs

The model used to analyse a n run split-plot design with n_w whole-plots and n_s sub-plots in each whole-plot ($n = n_w \times n_s$) has the same form as (1.3) in Section 1.3.1, with $b = n_w$. The columns in $\mathbf{X}$ relating to the whole-plot factors are perfectly correlated with the columns of $\mathbf{Z}$, as the levels of the whole-plot factors are constant within each whole-plot. The variance-covariance matrix $\mathbf{V}$ for $\mathbf{Y}$ is given by (1.5) in Section 1.3.1. We note that the variance and covariances of terms within $\hat{\boldsymbol{\beta}}$, (1.9) in Section 1.3.2, relating to the whole-plot factors will be larger than in the completely randomised case, as demonstrated in Section 3.2.4. It is common when finding optimal split-plot designs to assume there is more variability between whole-plots than between sub-plots, hence $\sigma^2_\gamma > \sigma^2_\epsilon$ and $\eta = \sigma^2_\gamma/\sigma^2_\epsilon > 1$.

3.2.3 Strip-Plot Designs

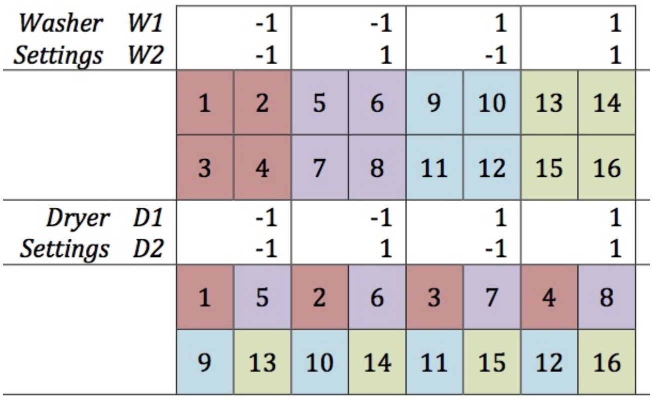
Strip-plot designs are another type of two-tiered design with restricted randomisation. Strip-plot designs arise naturally when there are two steps to a process where experimental units are batched in the first step and re-batched or regrouped at the second step. Factors applied to the original batches of experimental units are referred to as row factors, and treatments for the re-batched experimental units are referred to as column factors. Strip-plot designs are also known as row-column designs.

Re-batching means that the factor levels are reset less often in strip-plot designs than in split-plot designs with an equivalent number of hard-to-change and row factors. However, this reduction in experimental cost has the disadvantage of a more complicated mixed model for analysing strip-plot designs.

Example

Assume that only four washer and four dryers are available but the experimenters want to apply all 16 treatments in Table 3.1 to the experiment. The strip-plot design in Figure 3.3 would allow all 16 treatments to be applied as each cloth is washed in one of four washing machines which is set at one of four washer settings, and dried in one of four dryers which is set at one of four dryer settings. The colours in Figure 3.3 represent which washing machine is used to wash the cloth.

At the first step, washing, the 16 cloths are split into four batches of four and these four batches are assigned to one of the four washing machines at random. The row factors in this split-plot design are therefore the washing factors, $W1$ and $W2$. At the second step, drying, the cloths are re-batched so that each batch contains a cloth which was washed in each washer in step 1. These batches of cloths are then randomised to dryers, so a cloth from each washer is dried in each dryer and the column factors for this strip-plot design are $D1$ and $D2$.



Washer W1	-1		-1		1		1	
Settings W2	-1		1		-1		1	
	1	2	5	6	9	10	13	14
	3	4	7	8	11	12	15	16
Dryer D1	-1		-1		1		1	
Settings D2	-1		1		-1		1	
	1	5	2	6	3	7	4	8
	9	13	10	14	11	15	12	16

Figure 3.3: Allocation of cloths to washers and dryers for a strip-plot design.

A Mixed Model for Analysing Strip-Plot Designs

The mixed model for analysing responses from a strip-plot design in n runs with n_r rows and n_c columns ($n = n_r \times n_c$) is given by

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_\gamma\boldsymbol{\gamma} + \mathbf{Z}_\delta\boldsymbol{\delta} + \boldsymbol{\epsilon}, \tag{3.1}$$

where $\mathbf{Z}_\gamma$ is the $n \times n_r$ matrix representing the allocation of treatments to rows, $\boldsymbol{\gamma}$ is the $n_r \times 1$ vector of random row effects, $\mathbf{Z}_\delta$ is the $n \times n_c$ matrix representing the allocation of treatments to columns, $\boldsymbol{\delta}$ is the $n_c \times 1$ vector of random column effects and $\boldsymbol{\epsilon}$ is the $n \times 1$ vector of random errors for the runs within the rows and columns of the experiment. If the ith run, $i = 1, \ldots, n$, is in the jth row, $j = 1, \ldots, n_r$, and kth column, $k = 1, \ldots, n_c$, of an experiment, then the (i, j)th element of $\mathbf{Z}_\gamma$ is 1 and the (i, k)th element of $\mathbf{Z}_\delta$ is 1. The row and column factor columns of $\mathbf{X}$ will be confounded with columns in $\mathbf{Z}_\gamma$ and $\mathbf{Z}_\delta$, as the levels of the row factors are constant within rows and the levels of the column factors are constant within columns.

In this work, we assume that $\boldsymbol{\beta}$ contains the p parameters of interest and the three random effects $\boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{0}, \sigma_r^2\mathbf{I}_{n_r})$, $\boldsymbol{\delta} \sim \mathrm{N}(\mathbf{0}, \sigma_c^2\mathbf{I}_{n_c})$ and $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}, \sigma_\epsilon^2\mathbf{I}_n)$ are independently distributed where σ_r^2, σ_c^2 and σ_ϵ^2 are the constant between row, between column and within row and column covariances, respectively. The variance-covariance matrix for $\mathbf{Y}$ in (3.1) is given by

$$\begin{aligned}\mathbf{V} &= \text{var}(\mathbf{Y}) \\ &= \text{var}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}_\gamma\boldsymbol{\gamma} + \mathbf{Z}_\delta\boldsymbol{\delta} + \boldsymbol{\epsilon}) \\ &= \text{var}(\mathbf{Z}_\gamma\boldsymbol{\gamma} + \mathbf{Z}_\delta\boldsymbol{\delta} + \boldsymbol{\epsilon}) \\ &= \sigma_r^2\mathbf{Z}_\gamma\mathbf{Z}_\gamma^T + \sigma_c^2\mathbf{Z}_\delta\mathbf{Z}_\delta^T + \sigma_\epsilon^2\mathbf{I}_n \\ &= \sigma_\epsilon^2\left(\eta_1\mathbf{Z}_\gamma\mathbf{Z}_\gamma^T + \eta_2\mathbf{Z}_\delta\mathbf{Z}_\delta^T + \mathbf{I}_n\right), \end{aligned} \qquad (3.2)$$

where $\eta_1 = \sigma_r^2/\sigma_\epsilon^2$ and $\eta_2 = \sigma_c^2/\sigma_\epsilon^2$ are the relative magnitudes of the row and column variance components, respectively, compared to the within row-column variance. As shown for the example in Section 3.2.4, the restrictions on randomisation within strip-plot experiment affect the parameters for the main effect and pairwise product of the row and column factors.

3.2.4 Impact of Restrictions on Randomisations Example: Washing and Drying Cloths

In this section, we use the example designs discussed in Sections 3.2.1, 3.2.2 and 3.2.3 to demonstrate the impact of restrictions on randomisation on the variance of the generalised least square estimator of $\hat{\boldsymbol{\beta}}$, (1.10) in Section 1.3.2. Table 3.2 gives the variances for the parameters for β_{W1} , β_{W2}, β_{D1}, β_{D2}, $\beta_{W1,W2}$ and $\beta_{D1,D2}$, which are the diagonal elements of (1.10) for the model for these parameters, for varying η for the split-plot design and η_1, η_2 for the strip-plot design.

Estimator	$\hat{\beta}_{W1}$	$\hat{\beta}_{W2}$	$\hat{\beta}_{D1}$	$\hat{\beta}_{D2}$	$\hat{\beta}_{W1,W2}$	$\hat{\beta}_{D1,D2}$
Completely Randomised	0.0625	0.0625	0.0625	0.0625	0.0625	0.0625
Split-Plot, $\eta = 2$	0.5625	0.5625	0.0625	0.0625	0.5625	0.0625
Split-Plot, $\eta = 5$	1.3125	1.3125	0.0625	0.0625	1.3125	0.0625
Split-Plot, $\eta = 10$	2.5625	2.5625	0.0625	0.0625	2.5625	0.0625
Strip-Plot,$\eta_1 = 2, \eta_2 = 2$	0.5625	0.5625	0.5625	0.5625	0.5625	0.5625
Strip-Plot,$\eta_1 = 2, \eta_2 = 5$	0.5625	0.5625	1.3125	1.3125	0.5625	1.3125
Strip-Plot, $\eta_1 = 2, \eta_2 = 10$	0.5625	0.5625	2.5625	2.5625	0.5625	2.5625
Strip-Plot,$\eta_1 = 5, \eta_2 = 2$	1.3125	1.3125	0.5625	0.5625	1.3125	0.5625
Strip-Plot,$\eta_1 = 5, \eta_2 = 5$	1.3125	1.3125	1.3125	1.3125	1.3125	1.3125
Strip-Plot, $\eta_1 = 5, \eta_2 = 10$	1.3125	1.3125	2.5625	2.5625	1.3125	2.5625
Strip-Plot,$\eta_1 = 10, \eta_2 = 2$	2.5625	2.5625	0.5625	0.5625	2.5625	0.5625
Strip-Plot,$\eta_1 = 10, \eta_2 = 5$	2.5625	2.5625	1.3125	1.3125	2.5625	1.3125
Strip-Plot, $\eta_1 = 10, \eta_2 = 10$	2.5625	2.5625	2.5625	2.5625	2.5625	2.5625

Table 3.2: Variance, from (1.10), of fixed effect estimators, (1.9), in Section 1.3.2 of Chapter 1, for the example designs in Section 3.2.1, 3.2.2 and 3.2.3.

Firstly, we note that the variances for $\hat{\beta}_{W1}$, $\hat{\beta}_{W2}$ and $\hat{\beta}_{W1,W2}$, the main effects and product of the washing factors, in Table 3.2 are higher for the split-plot design and the strip-plot design than for the completely randomised design. This shows how the variances for parameters relating to whole-plot (split-plot) or row (strip-plot) factors are inflated by the restrictions on randomisation in split-plot and strip-plot designs. Also, we note that the variances for $\hat{\beta}_{D1}$, $\hat{\beta}_{D2}$ and $\hat{\beta}_{D1,D2}$, the main effects and product of the drying, or column, factors, are higher for the strip-plot design than for the split-plot and completely randomised design. This shows the impact of the additional restrictions on randomisation for strip-plot designs.

It is also interesting to note from Table 3.2 that the variance for $\hat{\beta}_{W1}$, $\hat{\beta}_{W2}$ and $\hat{\beta}_{W1,W2}$ increases as the ratio between the whole-plot and sub-plot variance, η, increases for split-plot designs and as the ratio of the row and within row-column variance, η_1, increases for strip-plot designs. The variance of $\hat{\beta}_{D1}$, $\hat{\beta}_{D2}$ and $\hat{\beta}_{D1,D2}$ increases as the ratio of the column and within-row-column variance, η_2 increases for strip-plot designs.

3.2.5 Further Reading

There is a range of literature on multi-tier designs. Brien and Bailey (2006) defined and discussed six types of treatment randomisation in multi-tier designs and provided numerous examples, including the wine tasting example seen in the papers by Brien (1983) and Brien and Payne (1999). Brien (1983) discussed the construction of ANOVA tables for multi-tiered experiments and Brien and Payne (1999) discussed how the algorithm in Genstat, based on algorithms by Wilkinson (1970) and Payne and Wilkinson (1977), can be used to determine the randomisation structure of multi-tier experiments. Brien and Bailey (2006) gave an indication of how to formulate a randomisation-based mixed model for data analysis and discussed the notion of inter-tier interactions, which are the interactions between factors in different tiers of the experiment.

Bingham et al. (2008) built on the work by Brien and Bailey (2006) and developed a general method for constructing fractional factorial designs by considering first the randomisation and then the treatment structure of a design. The construction of response surface designs for multi-stratum experiments was discussed by Trinca and Gilmour (2001). Strata for randomised experiments were defined and discussed by Bailey (1991) using results from group theory. Trinca and Gilmour (2001) constructed multi-stratum designs stratum by stratum, so that the factors in the current stratum are nearly orthogonal to factors in the higher strata and hence parameters relating to these factors can be estimated independently.

Two-tier designs were first described by McIntyre (1955), who was also the first to incorporate randomisation in each phase of a design. This idea was extended by Brien et al. (2011), who considered analysis based on both tiers, when a response is measured at the end of experimentation. There is a significant amount of literature on split-plot and strip-plot designs, which are two-tiered designs. Box and Jones (1992) provided an overview of split- and strip- plot designs, and variations, as well discussing analysis of such designs. Miller (1997) and Stein (1999) discussed the design of fractional factorial strip- and split- plot designs (respectively). Milliken et al. (1998) discussed strip-plot designs for two-step processes and demonstrated the analysis of such designs using a response surface model, hence linking to Trinca and Gilmour (2001). Other examples of strip-plot designs can be seen in the papers by Vivacqua and Bisgaard (2004, 2009). Goos and Gilmour (2012) discussed the analysis of split-plot and other multi-stratum designs and used Hasse diagrams to visualise the structure, randomisation, stratum and degrees of freedom for main effects, interactions and variance components.

3.3 Multi-Stage Experiments - Partition Designs

Partition experiments, whose construction and analysis were discussed by Perry et al. (2001, 2002, 2007) and Pieracci et al. (2010), are similar to the multi-stage experiments defined in Section 3.1.1. In a partition design, the same experimental unit is used across multiple stages of experimentation, which are referred to as partitions, and sub-treatments are applied to experimental units at each run in each partition. A response is assumed to exist for each partition, however these responses are not measured until the end of the experiment, after the all the partitions of experimentation are completed.

An experiment with Q partitions will have Q responses for these Q partitions measured after the sub-treatments in partition Q have been applied. This differs from our definition of multi-stage experiments, as we assume that the response for each stage is measured after all the sub-treatments for that stage are applied. Therefore, if an experiment has S stages, the response for the sth stage, $s = 1, 2, \ldots, S$, is measured after the sub-treatments at stage s are applied.

The authors of the partition design literature assume that the response from partition $q-1$ is still affected by the factors applied in partitions $1, 2, \ldots, q-2$ and is not affected by the factors applied in partition q, however they expect the factors from partition $q-1$ to be more influential than those in partitions $1, 2, \ldots, q-2$.

3.3.1 Design and Analysis of Partition Experiments for First Order Models

Perry et al. (2001) and Pieracci et al. (2010) considered the design and analysis of partition experiments when first order models of the responses are assumed, whereas Perry et al. (2002) and Perry et al. (2007) considered the design and analysis of partition experiments when second order models of the responses are assumed. Throughout this thesis we assume that first order models are appropriate for the response, and hence we only consider two-level factors in our designs. Therefore, we will only discuss the design and analysis methods for first order models in this section.

Assume that there are k_q factors in partition q, $q = 1, ..., Q$, then a first order model for the ith response, $i = 1, \ldots, n$ in the qth partition in terms of the factors in partition q is given by

$$Y_{qi} = \beta_{0q} + \sum_{l=1}^{k_q} \beta_{ql} x_{qli} + \sum_{l<m} \beta_{qlm} x_{qli} x_{qmi} + \epsilon_{qi}, \tag{3.3}$$

where β_{01} is the intercept for the qth partition, β_{ql} and β_{qlm}, $l, m = 1, \ldots, k_q$ and $l < m$, are the fixed effect parameters for the qth partition, x_{qli} is the level of the lth factor applied in the ith run for the qth partition and ϵ_{qi} is the random error for the ith run of the qth partition. The model (3.3) can be written in matrix form as

$$\mathbf{Y}_q = \beta_{0q} \mathbf{1}_n + \boldsymbol{\beta}_q \mathbf{X}_q + \boldsymbol{\epsilon}_q, \tag{3.4}$$

where $\mathbf{Y}_q = (Y_{q1}, \ldots, Y_{qn})^T$, $\mathbf{1}_n$ is the $n \times 1$ vector of ones, $\boldsymbol{\beta}_q = (\beta_{q1}, \ldots, \beta_{qk_q}, \beta_{q11}, \ldots, \beta_{qk_qk_q})^T$ is the vector of fixed effects parameters, $\mathbf{X}_q$ is the model matrix for the sub-treatments in partition q, and $\boldsymbol{\epsilon}_q = (\epsilon_{q1}, \ldots, \epsilon_{qn})^T$.

When it is assumed that a model of the form (3.3) describes the responses from the experiment, Perry et al. (2001) found the design with $n = k + 1$ runs, where $k = \sum_{q=1}^{Q} k_q$, using the following steps:

1. Let $q_1^*, \ldots, q_Q^*$ be the partitions ordered with respect to the number of factors, where q_1^* is the partition with the largest number of factors and q_Q^* is the partition with the smallest number of factors.

2. For partition q_1^*, find the largest possible regular fractional factorial design available such that the number of runs is less than $k + 1$. For more detail regarding fractional factorial designs see, for example, Box and Hunter (1961a,b).

3. Select the sub-treatments for partition q_2^* using defining relations between factors in partition q_2^* and the factors in partition q_1^*. These defining relations should be selected so that effects of interest can be estimated.

4. Repeat 3 for $q_3^*, \ldots, q_Q^*$.

5. Additional runs are added so the design has $n = k + 1$ runs by maximising

$$\frac{|\mathbf{X}^T \mathbf{X}|}{n} \tag{3.5}$$

when $\mathbf{X}$ is the model matrix for the design considering all Q partitions (which the authors refer to as D-efficiency), and ensuring the design has near equal occurrences of the levels for each factor.

We believe the authors set $n = k + 1$ so that a model containing the main effects from all Q partitions can be fitted to the response from partition Q. Therefore we assume that this cumulative model is considered in step 5 of the algorithm, as otherwise $p > n$ and hence $|\mathbf{X}^T\mathbf{X}| = 0$.

The projectivity of partition designs, that is the optimality or efficiency of designs produced using subsets of the columns of the original design, could be an important consideration. Partition designs consider models fitted to a subset of the factors in the design, however projectivity is not discussed by Perry et al. (2001, 2002, 2007) and Pieracci et al. (2010).

Perry et al. (2001) used the following steps to analyse the responses from partition experiments:

1. Find the fitted parameters for observed responses $\mathbf{y}_q = (y_{q1}, \ldots, y_{qn})$,

$$\hat{\mathbf{y}}_q = \hat{\beta}_{0q}\mathbf{1}_n + \hat{\boldsymbol{\beta}}_q\mathbf{X}_q, \tag{3.6}$$

   where $\hat{\beta}_{0q}$ be the fitted intercept for the qth partition, $\hat{\boldsymbol{\beta}}_q$ is the vector of fitted fixed effect parameters for the qth partition and $\mathbf{X}_q$ is the model matrix for the sub-treatments in the qth partition.

2. Select the significant fitted fixed effects based on model selection criteria such as R^2, adjR^2, C_p (see Kutner et al., 2004, Chapter 3) and the PRESS statistics (Allen, 1974).

3. Find the partition intercept response

$$\hat{\mathbf{y}}_{PIq} = \mathbf{y}_q - \hat{\mathbf{y}}_q, \tag{3.7}$$

   where $\hat{\mathbf{y}}_q$ is the fitted response from (3.6) in step 1.

4. For $t \neq j$ fit

$$\hat{\mathbf{y}}_{PIq} = \hat{\omega}_{0t}\mathbf{1}_n + \hat{\boldsymbol{\omega}}_t\mathbf{X}_t, \tag{3.8}$$

   where $\hat{\mathbf{y}}_{PIq}$ is calculated in (3.7), $\hat{\omega}_{0t}$ is the fitted intercept for partition t, $\hat{\boldsymbol{\omega}}_t$ is the matrix of fitted effects for partition t, and $\mathbf{X}_t$ is the model matrix for the sub-treatments in partition t.

5. Select the significant effects from (3.8) using the model selection criteria used in step 2.

6. Fit the final model for the union of significant terms selected in steps 2 and 5 to $\mathbf{y}_q$. There may be terms from all Q partitions in this final model.

This method of analysis reduces the model selection problem by only ever considering subsets of the factors. The partition intercept response is the residuals found after fitting the model to the factors in partition q. Hence fitting a model to the partition intercept response attempts to assess whether these residuals are influenced by the factors in other partitions, and includes these known sources of variation in the final model.

Although not explicitly mentioned by Perry et al. (2001), Pieracci et al. (2010) noted that the defining relations used when selecting the design will have an impact on the analysis of response from partition design. The defining relations used in Pieracci et al. (2010), for example, create a complex aliasing structure between the parameters in the model assumed for the responses from the design, and hence the analysis of the results described in this paper would require some thought.

Hamada and Wu (1992) discussed the analysis of designs with complex aliasing structures and used a procedure that depends on the assumption that only lower order interactions will be non-negligible and the principle of effect heredity, where an interaction is only included in a model if the main effects of the factors in that interaction are also in the model. In Chapter 4, we present a method which can be used to analyse split-plot designs with complex aliasing.

3.3.2 Partition Design Example

We now use the washing and drying example (Miller, 1997) to illustrate the construction and analysis of a partition design for first order models. Assume there are two, two-level, washing factors, $W1$ and $W2$, two, two-level, drying factors, $D1$ and $D2$, and the experiment has two partitions; washing and drying. There are $n = k + 1 = 5$ washing machines and dryers that can be used in this experiment, and there are no restrictions on randomisation, unlike the examples discussed in Sections 3.2.2 and 3.2.3. The wrinkling of the cloth due to washing and the dryness of the cloth due to drying is tested at the end of the experiment. A schematic for this experiment is shown in Figure 3.4.

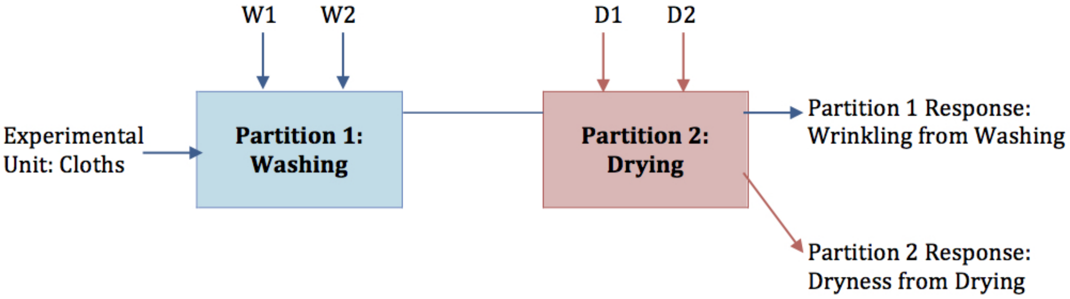

W1
W2
D1
D2
Experimental Unit: Cloths
Partition 1: Washing
Partition 2: Drying
Partition 1 Response: Wrinkling from Washing
Partition 2 Response: Dryness from Drying

Figure 3.4: A partition design for the washing and drying cloths experiment.

Assume the following models for the factors in the two partitions:

$$\mathbf{Y}_1 = \beta_{01}\mathbf{1}_n + \boldsymbol{\beta}_1\mathbf{X}_1 + \boldsymbol{\epsilon}_1, \text{ and} \tag{3.9}$$

$$\mathbf{Y}_2 = \beta_{02}\mathbf{1}_n + \boldsymbol{\beta}_2\mathbf{X}_2 + \boldsymbol{\epsilon}_2, \tag{3.10}$$

where $\mathbf{Y}_1 = (Y_{11}, \ldots, Y_{1n})^T$ and $\mathbf{Y}_2 = (Y_{21}, \ldots, Y_{2n})^T$, respectively, are the wrinkling from washing (partition 1) and dryness from drying (partition 2), β_{01} and β_{02} are the intercepts for washing and drying (respectively), $\boldsymbol{\beta}_1 = (\beta_{W1}, \ldots, \beta_{W1W2})^T$ and $\boldsymbol{\beta}_2 = (\beta_{D1}, \ldots, \beta_{D1D2})^T$ are the fixed effects from washing and drying (respectively), $\mathbf{X}_1$ and $\mathbf{X}_2$ are the model matrices for washing and drying (respectively), and $\boldsymbol{\epsilon}_1$ and $\boldsymbol{\epsilon}_2$ are the random errors associated with wrinkling from washing and dryness from drying, which are independently distributed with common variance.

When the responses can be analysed using (3.9) and (3.10), the experiment is designed using the following steps:

1. The treatments in the 2^2 factorial are used for the washing partition.

2. $D1$ is aliased with $W1 \times W2$, and hence β_{D1} and β_{W1W2} cannot be estimated simultaneously. We assumed that β_{W1W2} is not considered to be a key effect, and therefore this aliasing is acceptable.

3. $D2$ is chosen so that all four combinations of $D1$ and $D2$, $(D1, D2)$=$(-1, -1)$, $(1, -1)$, $(-1, 1)$, $(1, 1)$, are present in the design.

4. An extra row is added so that the design has $n = k+1$ runs, maximises (3.5) and ensures near equal occurrence of the factor levels for each factor.

W1	W2	D1	D2
1	1	1	1
−1	1	−1	1
1	−1	−1	−1
−1	−1	1	−1
−1	−1	1	1

Table 3.3: Partition design for washing and drying cloths when first order models are assumed for the responses.

The design for first order models is given in Table 3.3. The aliasing relationships for this design are $D1 = W1 \times W2$ and $W1 = W2 \times D1$, which mean that β_{D1} and β_{W1W2}, and β_{W1} and β_{W2D1} cannot be estimated independently. This aliasing will need to be considered when analysing the responses from this experiment.

The analysis of the responses from the partition design in Table 3.3 is performed using the following steps:

1. Find the fitted parameters for observed responses,

$$\hat{\mathbf{y}}_1 = \hat{\beta}_{01}\mathbf{1}_n + \hat{\boldsymbol{\beta}}_1\mathbf{X}_1, \tag{3.11}$$

$$\hat{\mathbf{y}}_2 = \hat{\beta}_{02}\mathbf{1}_n + \hat{\boldsymbol{\beta}}_2\mathbf{X}_2, \tag{3.12}$$

   where $q = 1$ relates to washing, $q = 2$ relates to drying, $\hat{\mathbf{y}}_q$ is the observed response for partition q, $\hat{\beta}_{0q}$ be the fitted intercept for partition q, $\hat{\boldsymbol{\beta}}_q$ is the matrix of fitted fixed effect parameters for the qth partition and $\mathbf{X}_q$ is the model matrix for the sub-treatments in the qth partition.

2. Select the significant fitted fixed effects based on certain model selection criteria.

3. Find the partition intercept responses

$$\hat{\mathbf{y}}_{PI1} = \mathbf{y}_1 - \hat{\mathbf{y}}_1, \tag{3.13}$$

$$\hat{\mathbf{y}}_{PI2} = \mathbf{y}_2 - \hat{\mathbf{y}}_2, \tag{3.14}$$

   where $\hat{\mathbf{y}}_q$, $q = 1, 2$, are the fitted responses from (3.11) and (3.12) in step 1.

4. Fit

$$\hat{\mathbf{y}}_{PI1} = \hat{\omega}_{02}\mathbf{1}_n + \hat{\boldsymbol{\omega}}_2\mathbf{X}_2, \tag{3.15}$$

$$\hat{\mathbf{y}}_{PI2} = \hat{\omega}_{01}\mathbf{1}_n + \hat{\boldsymbol{\omega}}_1\mathbf{X}_1, \tag{3.16}$$

   where $\hat{\mathbf{y}}_{PI1}$ and $\hat{\mathbf{y}}_{PI2}$ are calculated in (3.13) and (3.14), respectively, $\hat{\omega}_{0t}$ is the fitted intercept for partition $t = 1, 2$, $\hat{\boldsymbol{\omega}}_t$ is the matrix of fitted effects for partition t, and $\mathbf{X}_t$ is the model matrix for the sub-treatments in partition t.

5. Select the significant effects from (3.15) and (3.16) based on the model selection criteria used in step 2.

6. Fit the final model for the union of significant terms selected in steps 2 and 5 to $\mathbf{y}_q$, $q = 1, 2$. There may be terms from both partitions in this final model.

3.4 Cumulative Models and Compound Bayesian D-optimality for Multi-Stage Experiments

Recall from our definition of multi-stage experiments from Section 3.1.1, that we assume two models are fitted to the response from each stage of the experiment; a model which relates the response from the current stage to the factors applied in the current stage, and a cumulative model which relates the response from the current stage and all the factors applied in previous stages. In Section 3.4.1 we discuss cumulative models in greater detail and describe the models required for multi-stage completely randomised, split-plot and strip-plot designs.

Compound criterion enable multiple experimental objectives to be considered when designing experiments. As we want to find designs which provide as much information about the parameters in each model considered, we require a compound criterion which calculates the volume of the confidence ellipsoid for the parameter vector for each model. The objective function for the compound criterion which meets the aims of our experiment is presented in Section 3.4.2.

3.4.1 Cumulative Models

Two models are considered for the response at stage $s = 2, \ldots, S$; the model for the factors in stage s and the cumulative model for the factors in stages $1, 2, \ldots, (s-1), s$. Note that for the first stage, $s = 1$, there is only one model as there are no previous stages to consider in a cumulative model. Therefore, $m = 2S - 1$ models will be considered for a S-stage design.

These m models can take on various forms, depending on the restrictions on randomisation assumed for the experiment. We note that the sub-treatments for each stage are not randomised individually, hence the restrictions are applied to the randomisation of the treatments as a whole. A multi-stage completely randomised design has no restrictions on randomisation, and (1.6) with $\mathbf{V} = (\sigma_\gamma^2 + \sigma_\epsilon^2)\mathbf{I}_n$ is used to analyse the responses for all $m = 2S - 1$ models.

A multi-stage split-plot design has some hard-to-change factors, and therefore (1.3) with variance given by (1.5) is used to analyse the responses in the $m^* \leq m$ models that include the parameters which relate to just whole-plot and or whole-plot and sub-plot factors, as the correlation for responses from different whole plots has to be accounted for. Any models that only include parameters relating to sub-plot factors can be analysed using (1.6) with $\mathbf{V} = (\sigma_\gamma^2 + \sigma_\epsilon^2)\mathbf{I}_n$, as the sub-plots are allocated to whole-plots using a single randomisation. For an example of the three models considered for the two responses from a two-stage split-plot design with two whole-plot and three sub-plot factors in the first stage and one sub-plot factor in the second stage, see Table 3.6 in Section 3.6.

A multi-stage strip-plot design has experimental units that can be re-batched after a certain number of factors have been applied. Strip-plot designs have a more complicated randomisation structure than split-plot designs, and the analysis of the S responses from multi-stage strip-plot designs is also more complicated. Only the final cumulative model, which relates the response from stage S to the factors in stages $1, 2, \ldots, S$, contains all the row and column factors, hence this is the only model that will be analysed using (3.1) with variance given by (3.2). The other models used to analyse the responses depend on what factors are considered in these models.

Any models which include parameters relating to a subset of the row factors and column factors can be analysed using (1.3) with variance given by (1.5), as the row factors can be thought of as hard-to-change factors and the column factors can be thought of as sub-plot factors. Any models which only include parameters relating to row or column factors can also be analysed using (1.3) with (1.5), as the correlation between responses in rows or columns has to be accounted for. Table 3.6 in Section 3.6 gives details of the three models considered for a two-stage strip-plot design with two row factors and three column factors in the first stage and one column factor in the second stage.

3.4.2 Bayesian D-Optimality for Multi-Stage Experiments

As discussed in Section 3.4.1, $m = 2S - 1$ models are fitted to the S responses from a S-stage design. It is important that we consider the performance of the design with respect to each of these m models, as a design which is optimal for one of these models may be very poor for another model. Compound optimality criteria, as discussed by authors such as Atkinson and Cox (1974), Läuter (1974), Atkinson and Bogacka (1997) and Atkinson et al. (2007, Chapter 21), allow a set of models to be considered when optimising the design, where the importance of each of the models is determined by a weight.

As the number of parameters, p_l, in model $l = 1, \ldots, m$ can be greater than n, and one aim of this experiment is to gain scientific knowledge about these parameters, we consider the Bayesian D-optimality of the multi-stage design with respect to each of these l models. The objective function for Bayesian D-optimality is

$$|\mathbf{X}_l^T \mathbf{V}_l^{-1} \mathbf{X}_l + \mathbf{R}_l|, \tag{3.17}$$

where $\mathbf{X}_l$ is the model matrix for model l, $\mathbf{V}_l$ is the variance-covariance matrix for model l and $\mathbf{R}_l$ is the prior precision matrix of an appropriate prior distribution of $\boldsymbol{\beta}_l$ for model l, so $\boldsymbol{\beta}_l$ is distributed with mean $\boldsymbol{\mu}_l$ and variance $\mathbf{R}_l^{-1}$.

In this work, we place an informative prior on all parameters in $\boldsymbol{\beta}$ apart from the intercept, hence $\mathbf{R}_l = \mathbf{I}_{p_l} - (\mathbf{e}_{p_l,1}\mathbf{e}_{p_l,1}^T)$, where $\mathbf{e}_{n,j}$ is the jth column of $\mathbf{I}_n$, in Table 3.6. The robustness of the designs found to different $\mathbf{R}_l$ matrices could be considered in future work. Also, the impact of increasing the elements of $\mathbf{R}_l$ relating to factors which are assumed to be more influential on the response could be assessed.

Assume that the Bayesian D-optimum design for model l, that is the design which maximises (3.17) for all $\mathbf{D} \in \mathcal{D}_{f,l,n}$, is $\mathbf{D}_{l*}$, and that the model matrix and variance-covariance matrix relating to $\mathbf{D}_{l*}$ for model l are $\mathbf{X}_{l*}$ and $\mathbf{V}_{l*}$, respectively. Then using the definition of a compound criterion given by Atkinson et al. (2007, Chapter 21), we want to find a design which maximises

$$\Phi = \prod_{l=1}^{m} \left[100 \left(\frac{|\mathbf{X}_l^T \mathbf{V}_l^{-1} \mathbf{X}_l + \mathbf{R}_l|}{|\mathbf{X}_{l*}^T \mathbf{V}_{l*}^{-1} \mathbf{X}_{l*} + \mathbf{R}_l|} \right)^{\frac{1}{p_l}} \right]^{w_l}, \tag{3.18}$$

where w_l, $0 < w_l \leq 1$, $\sum_l^m w_l = 1$ is the weight which demonstrates the importance of model l. Maximising (3.18) is equivalent to maximising

$$\begin{aligned} \log(\Phi) &= \sum_{l=1}^{m} w_l \log \left[100 \left(\frac{|\mathbf{X}_l^T \mathbf{V}_l^{-1} \mathbf{X}_l + \mathbf{R}_l|}{|\mathbf{X}_{l*}^T \mathbf{V}_{l*}^{-1} \mathbf{X}_{l*} + \mathbf{R}_l|} \right)^{\frac{1}{p_l}} \right] \\ &= \sum_{l=1}^{m} w_l \log(100) + \sum_{l=1}^{m} \frac{w_l}{p_l} \log \left(|\mathbf{X}_l^T \mathbf{V}_l^{-1} \mathbf{X}_l + \mathbf{R}_l| \right) \\ &\quad - \sum_{l=1}^{m} \frac{w_l}{p_l} \log \left(|\mathbf{X}_{l*}^T \mathbf{V}_{l*}^{-1} \mathbf{X}_{l*} + \mathbf{R}_l| \right) \\ &= \sum_{l=1}^{m} w_l \log(100) + \phi_D - \sum_{l=1}^{m} \frac{w_l}{p_l} \log \left(|\mathbf{X}_{l*}^T \mathbf{V}_{l*}^{-1} \mathbf{X}_{l*} + \mathbf{R}_l| \right). \end{aligned} \tag{3.19}$$

We note that the maximisation of (3.19) with respect to $\mathbf{X}_l$ is only dependent on ϕ_D as $\sum_{l=1}^m w_l \log(100)$ and $\sum_{l=1}^m \frac{w_l}{p_l} \log \left(|\mathbf{X}_{l*}^T \mathbf{V}_{l*}^{-1} \mathbf{X}_{l*} + \mathbf{R}_l| \right)$ are fixed. Therefore, our optimal S-stage designs maximise

$$\phi_D = \sum_{l=1}^{m} \frac{w_l}{p_l} \log |\mathbf{X}_l \mathbf{V}_l^{-1} \mathbf{X}_l + \mathbf{R}_l|. \tag{3.20}$$

.

3.5 Methods for Finding and Assessing Designs

The designs discussed in Section 3.6 are found using coordinate exchange algorithms which are extensions of the algorithm presented in Section 1.4.3 of Chapter 1. We assess the designs in Section 3.6 using the correlation between columns of the model matrices for the m models considered for the responses from the experiment, as high column correlations will affect our ability to accurately and precisely estimate the parameters in the models due to inflation of variance and bias. The matrix of the correlation between columns of a model matrix $\mathbf{X}$, $\mathbf{C}$, is presented and discussed in Section 3.5.2. We also use the relative efficiency to compare designs in Section 3.6.2, and this is defined in Section 3.5.3.

3.5.1 Coordinate Exchange Algorithm for Multi-Stage Designs

The coordinate exchange algorithm given in Section 1.4.3 where ϕ is the compound Bayesian D-optimality objective function (3.20), can be used when there are no restrictions on randomisation to find multi-stage completely randomised designs. However, adjustments are required when there are restrictions on randomisation.

Consider split-plot designs from $\mathcal{D}_{f,l,n}$ with n_w whole plots, n_s sub-plots (so $n = n_w n_s$), f columns which are ordered so that the first f_w columns give the levels of the whole plot factors and the remaining $f_s = f - f_w$ columns give the levels of the sub-plot factors and (i, j)th element $x_{i,j}$ for $i = 1, \dots, n$ and $j = 1, \dots, f$. Let N^k be the set of indices for the n_s treatments in whole-plot $k = 1, \dots, n_w$ and N^k_h be the hth, $h = 1, \dots, n_s$ element of N^k. Then we use the following coordinate exchange algorithm to find multi-stage split-plot designs, where the aim is to maximise to the compound Bayesian D-optimality objective function (3.20) and $\mathbf{D}_s \in \mathcal{D}_{f,l,n}$ is a starting split-plot design with n_w whole plots, n_s sub-plots, f_w whole plot factors and f_s sub-plot factors:

1. Set $\mathbf{D}_{1,0} = \mathbf{D}_s$ and calculate $\phi_s = \phi_D(\mathbf{D}_s)$.

2. For $k = 1, \dots, n_w$:

   (a) For $j = 1, \dots, f_w$:

      i. Calculate $\phi_1 = \phi_D(\mathbf{D}_{k,j-1})$.

      ii. Let $x^*_{i,j}$ be the (i, j)th element of $\mathbf{D}_{k,j-1}$ $\forall i \in N^k$.

      iii. Let $\mathbf{D}_{k,j}$ be equivalent to $\mathbf{D}_{k,j-1}$ but with (i, j)th elements $x_{i,j} = -x^*_{i,j}$ $\forall i \in N^k$.

      iv. Calculate $\phi_2 = \phi_D(\mathbf{D}_{k,j})$.

      v. If $\phi_1 > \phi_2$, let $x_{i,j} = x^*_{i,j}$ $\forall i \in N^k$, otherwise keep the swap and leave $x_{i,j}$ unchanged from (iii).

   (b) Rename $\mathbf{D}_{k,f_w}$ as $\mathbf{D}_{1,f_w}$.

   (c) For $j = f_w + 1, \dots, f$, and $h = 1, \dots, n_s$:

      i. Calculate $\phi_1 = \phi_D(\mathbf{D}_{h,j-1})$.

      ii. Let $x^*_{i,j}$ be the (i, j)th element of $\mathbf{D}_{k,j-1}$ for $i = N^k_h$.

iii. Let $\mathbf{D}_{h,j}$ be equivalent to $\mathbf{D}_{h,j-1}$ but with (i,j)th element $x_{i,j} = -x^*_{i,j}$ for $i = N^k_h$.

iv. Calculate $\phi_2 = \phi_D(\mathbf{D}_{h,j})$.

v. If $\phi_1 > \phi_2$, let $x_{i,j} = x^*_{i,j}$ for $i = N^k_h$, otherwise keep the swap and leave $x_{i,j}$ unchanged from (iii).

(d) If $k \leq n_w - 1$, restart from (a) with $\mathbf{D}_{n_s,f}$ as $\mathbf{D}_{k+1,0}$. If $k = n_w$, stop and let $\mathbf{D}_{n_s,f}$ be $\mathbf{D}_{n_w,f}$.

3. Calculate $\phi_E = \phi_D(\mathbf{D}n_w, f)$.

4. If $\phi_S < \phi_E$, repeat from 2 with $\mathbf{D}_{1,0} = \mathbf{D}_{n_w,f}$. Otherwise, stop the algorithm and return $\mathbf{D}_{1,0}$ as the design which maximises ϕ_D.

Consider strip-plot designs from $\mathcal{D}_{f,l,n}$ with n_r rows, n_c columns (so $n = n_r n_c$), f_r row factors and $f_c = f - f_w$ column factors, and (i,j)th element $x_{i,j}$ for $i = 1, \dots, n$ and $j = 1, \dots, f$. Let N^k be the set of indices for the n_c treatments in row $k = 1, \dots, n_r$ and N^h be the set of indices for the n_r treatments in column $h = 1, \dots, n_c$. Then we use the following coordinate exchange algorithm to find multi-stage strip-plot designs, where the aim is to maximise to the compound Bayesian D-optimality objective function (3.20) and $\mathbf{D}_s \in \mathcal{D}_{f,l,n}$ is a starting strip-plot design with n_r rows, n_c columns, f_r row factors and f_c column factors:

1. Set $\mathbf{D}_{1,0} = \mathbf{D}_s$ and calculate $\phi_s = \phi_D(\mathbf{D}_s)$.

2. For $k = 1, \dots, n_r$ and $j = 1, \dots, f_r$:

   (a) Calculate $\phi_1 = \phi_D(\mathbf{D}_{k,j-1})$.

   (b) Let $x^*_{i,j}$ be the (i,j)th element of $\mathbf{D}_{k,j-1}$ $\forall i \in N^k$.

   (c) Let $\mathbf{D}_{k,j}$ be equivalent to $\mathbf{D}_{k,j-1}$ but with (i,j)th elements $x_{i,j} = -x^*_{i,j}$ $\forall i \in N^k$.

   (d) Calculate $\phi_2 = \phi_D(\mathbf{D}_{k,j})$.

   (e) If $\phi_1 > \phi_2$, let $x_{i,j} = x^*_{i,j}$ $\forall i \in N^k$, otherwise keep the swap and leave $x_{i,j}$ unchanged from (c).

3. Rename $\mathbf{D}_{n_r,f_r}$ as $\mathbf{D}_{1,f_r}$.

4. For $h = 1, \dots, n_c$ and $j = f_r + 1, \dots, f_c$:

   (a) Calculate $\phi_1 = \phi_D(\mathbf{D}_{h,j-1})$.

   (b) Let $x^*_{i,j}$ be the (i,j)th elements of $\mathbf{D}_{h,j-1}$ $\forall i \in N^h$.

   (c) Let $\mathbf{D}_{h,j}$ be equivalent to $\mathbf{D}_{h,j-1}$ but with (i,j)th elements $x_{i,j} = -x^*_{i,j}$ $\forall i \in N^h$.

(d) Calculate $\phi_2 = \phi_D(\mathbf{D}_{h,j})$.

(e) If $\phi_1 > \phi_2$, let $x_{i,j} = x^*_{i,j}$ $\forall i \in N^h$, otherwise keep the swap and leave $x_{i,j}$ unchanged from (c).

5. Calculate $\phi_E = \phi_D(\mathbf{D}n_c, f)$.

6. If $\phi_S < \phi_E$, repeat from 2 with $\mathbf{D}_{1,0} = \mathbf{D}_{n_c,f}$. Otherwise, stop the algorithm and return $\mathbf{D}_{1,0}$ as the design which maximises ϕ_D.

There are two different types of starting designs used in this work:

1. **Random starting designs**: For an optimal multi-stage completely randomised design, any selection of n treatments from the l^f possible treatments for a l-level f factor full factorial experiment which ensure $\mathbf{X}^T\mathbf{X}+\mathbf{R}$ is non-singular can be used as a starting design. The starting designs for a coordinate exchange algorithm used to find an optimal multi-stage split-plot design are also combinations of these l^f possible treatments with non-singular $\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X}+\mathbf{R}$, where $\mathbf{V}$ is (1.5), however they have the added restriction that the levels of the f_w whole-plot factors must be constant in groups of size n_s. Similarly, the treatments in the starting designs for an optimal multi-stage strip-plot design must have constant levels of the f_r rows in blocks of size n_c and $\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X} + \mathbf{R}$, where $\mathbf{V}$ is (3.2), must be non-singular, but with the added restriction that the n_c column treatments must be the same in each of the n_r rows.

2. **Plackett-Burman and Hall based starting designs**: These starting designs are all possible subsets of f columns from the 12 run Plackett-Burman design (Plackett and Burman, 1946, Table 3.4) and 16 run Hall III design (Hall, 1961, Table 3.5). The 12 run Plackett-Burman and 16 run Hall III design are considered as they have been shown to have good projection properties, see Cheng (2006), Loeppky et al. (2007) and Section 3.6 for further discussion. In this work, we use these starting designs to find designs with no restrictions on randomisation.

Factor Run	1	2	3	4	5	6	7	8	9	10	11
1	1	1	−1	1	1	1	−1	−1	−1	1	−1
2	1	−1	1	1	1	−1	−1	−1	1	−1	1
3	−1	1	1	1	−1	−1	−1	1	−1	1	1
4	1	1	1	−1	−1	−1	1	−1	1	1	−1
5	1	1	−1	−1	−1	1	−1	1	1	−1	1
6	1	−1	−1	−1	1	−1	1	1	−1	1	1
7	−1	−1	−1	1	−1	1	1	−1	1	1	1
8	−1	−1	1	−1	1	1	−1	1	1	1	−1
9	−1	1	−1	1	1	−1	1	1	1	−1	−1
10	1	−1	1	1	−1	1	1	1	−1	−1	−1
11	−1	1	1	−1	1	1	1	−1	−1	−1	1
12	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1	−1

Table 3.4: The 12 run Plackett-Burman design for 11 factors.

Factor Run	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1	1	1	1	1	1	1	−1	−1	-1	−1	−1	−1	−1	−1
3	1	1	1	−1	−1	−1	−1	1	1	1	1	−1	−1	−1	−1
4	1	1	1	−1	−1	−1	−1	−1	−1	−1	−1	1	1	1	1
5	1	−1	−1	1	1	−1	−1	1	1	−1	−1	1	1	−1	−1
6	1	−1	−1	1	1	−1	−1	−1	−1	1	1	−1	−1	1	1
7	1	−1	−1	−1	−1	1	1	1	1	−1	−1	−1	−1	1	1
8	1	−1	−1	−1	−1	1	1	−1	−1	1	1	1	1	−1	−1
9	−1	1	−1	1	−1	1	−1	1	−1	1	−1	1	−1	1	−1
10	−1	1	−1	1	−1	1	−1	−1	1	−1	1	−1	1	−1	1
11	−1	1	−1	−1	1	−1	1	1	−1	−1	1	1	−1	−1	1
12	−1	1	−1	−1	1	−1	1	−1	1	1	−1	−1	1	1	−1
13	−1	−1	1	1	−1	−1	1	1	−1	−1	1	−1	1	1	−1
14	−1	−1	1	1	−1	−1	1	−1	1	1	−1	1	−1	−1	1
15	−1	−1	1	−1	1	1	−1	1	−1	1	−1	−1	1	−1	1
16	−1	−1	1	−1	1	1	−1	−1	1	−1	1	1	−1	1	−1

Table 3.5: The 16 run Hall III design for 15 factors.

3.5.2 Between Column Correlation for a Model Matrix

One method we use to assess and compare S-stage designs is the level of correlation between the columns in the model matrices for the $m = 2S - 1$ models considered, because the correlation between the columns will impact on the variance and bias of the parameter estimates related to these columns. The matrix of column correlations for the model matrix $\mathbf{X}_l$, $l = 1, \ldots, m$, is

$$\mathbf{C}_l = \frac{1}{n}(\mathbf{X}_l^T \mathbf{X}_l), \tag{3.21}$$

and the (i, j)th element of $\mathbf{C}_l$ is c_{lij}. If:

- $c_{lij} = 0$ the ith and jth column of $\mathbf{X}_l$ are not correlated.
- $c_{lij} \in [-1, 0)$ or $(0, 1]$ (so $-1 \geq c_{ij} < 0$ or $0 < c_{ij} \leq 1$) the ith and j column of $\mathbf{X}_l$ are correlated.

We prefer smaller to larger correlation between columns in $\mathbf{X}_l$ as the variance and bias of parameter estimators related to correlated columns of $\mathbf{X}$ will be inflated, and the higher the correlation the more inflated the variance and higher the bias. If two columns are correlated, then the parameters relating to these two columns cannot be estimated independently, and are therefore aliased.

In general, when comparing two designs, if one design has fewer non-zero entries in $\mathbf{C}_l$ for model l than another, this design is preferred as fewer parameters will have inflated variances and biases. However, if the range of c_{lij} in $\mathbf{C}_l$ for two designs is significantly different, then the design with smaller c_{lij} values may be preferred even if $\mathbf{C}_l$ has more non-zero entries. This is because the inflation of the variances and biases for the parameters relating to the columns will be less severe, and it is easier to identify the effect of parameters with smaller variances and bias.

Hence, a model matrix with more columns which have a small correlation may enable some information regarding the parameters which relate to the correlated columns to recovered, whereas a model matrix with fewer columns with larger correlation may mean that very little or no information regarding the parameters which relate to the columns can be recovered. The balance between the amount of correlated columns and the level of correlation needs to be carefully considered when comparing designs.

3.5.3 Relative Efficiencies

We can assess the relative performance of two designs using their relative efficiency. The relative efficiency of two S-stage designs found using the compound Bayesian D-

optimality criterion (3.20) is

$$\text{Eff}_D = \prod_{l=1}^{m} \left[100 * \left(\frac{|\mathbf{X}_{1,l}^T \mathbf{V}_l^{-1} \mathbf{X}_{1,l} + \mathbf{R}_l|}{|\mathbf{X}_{2,l}^T \mathbf{V}_l^{-1} \mathbf{X}_{2,l} + \mathbf{R}_l|} \right)^{\frac{1}{p_l}} \right]^{w_l}, \tag{3.22}$$

where $\mathbf{X}_{1,l}$ and $\mathbf{X}_{2,l}$ are the model matrices for the two optimal designs, $\mathbf{D}_1$ and $\mathbf{D}_2$, for models $l = 1, \ldots, m$, when $m = 2S - 1$. An efficiency of close to 100% suggests that $\mathbf{D}_1$ performs well with respect $\mathbf{D}_2$.

3.6 Study of the Impact of Restricted Randomisation and Benefits of Projectivity on Optimal Multi-Stage Designs

In this section we use the methods discussed in Section 3.5 to find and assess different multi-stage designs. There are two comparative studies performed in this section. In Section 3.6.1, we assess optimal two-stage completely randomised, split-plot and strip-plot designs found using the coordinate exchange algorithms in Section 3.5.1 with random starting designs. Each design has six factors, five of which are applied in the first stage and one of which is applied in the second stage, hence they are appropriate for the formulation of pharmaceutical products discussed in Sections 1.2.2 and 3.1.3. When finding two-stage split-plot designs, it is assumed that the first two factors are hard-to-change, therefore the design has $f_w = 2$ and $f_s = 4$. When finding two-stage strip-plot designs, the first two factors are assumed to be row factors and the other four factors are assumed to be column factors, so $f_r = 2$ and $f_c = 4$.

We consider two different run sizes, $n = 12, 16$, as the scientists formulating the pharmaceutical products have resource for between 12 and 16 runs. For both the 12 and 16 run two-stage split-plot designs, we assume that $n_w = 4$ and $\eta = 10$ and, for the 12 and 16 run two-stage strip-plot designs we assumed that $n_r = 4$ and $\eta_1 = \eta_2 = 10$. The number of whole-plots and rows were set based on the amount of times the experimenters were willing to reset the first two factors in the experiments, and the ratios of variances, η, η_1, η_2, were chosen to represent our belief that the variation between whole-plots, rows and columns, will much larger than variation within whole-plots, rows and columns.

We assume that all these two-stage designs have two responses which are measured after all the sub-treatments for each stage are applied. Three models are then fitted to these responses; (i) the model relating the response from stage one to the five factors in Stage 1, (ii) the model relating the response from Stage 2 to the single factor in Stage 2, and (iii) the cumulative model relating the response from Stage 2 to all six factors in Stages 1 and 2. Each model includes the intercept, the main effect and the pairwise products for all the factors in the sub-treatments or treatments considered. This means that $m = 3$ and (p_1, p_2, p_3)=(16,2,22). Following a discussion with the scientists formulating the pharmaceutical product regarding the relative importance they placed on the three models, we use the weights $\mathbf{w} = (w_1, w_2, w_3)$=(0.7, 0.1, 0.2).

As discussed in Section 3.4.1, the form of these models depends on the assumed restrictions on the randomisation. Table 3.6 details the models, variance-covariance matrices and $\mathbf{R}$ matrices considered when using (3.20) in the coordinate exchange algorithms in Section 3.5.1 to find optimal two-stage designs.

	Completely Randomised	Split-Plot	Strip-Plot
Model 1	(1.6)	(1.3)	(1.3)
	for Factors 1 to 5 and Stage 1 response		
Model 2	(1.6)	(1.6)	(1.3)
	for Factor 6 and Stage 2 response		
Model 3	(1.6)	(1.3)	(3.1)
	for Factors 1 to 6 and Stage 2 response		
$\mathbf{V}_1$	$(\sigma_\epsilon^2 + \sigma_\gamma^2)\mathbf{I}_n$	(1.5)	(1.5)
$\mathbf{V}_2$	$(\sigma_\epsilon^2 + \sigma_\gamma^2)\mathbf{I}_n$	$(\sigma_\epsilon^2 + \sigma_\gamma^2)\mathbf{I}_n$	(1.5)
$\mathbf{V}_3$	$(\sigma_\epsilon^2 + \sigma_\gamma^2)\mathbf{I}_n$	(1.5)	(3.2)

$\mathbf{R}_1$	$\mathbf{I}_{p_1} - (\mathbf{e}_{p_1,1}\mathbf{e}_{p_1,1}^T)$
	if $n = 12$
	$\mathbf{0}_{p_1}$
	if $n = 16$		

$\mathbf{R}_2$	$\mathbf{0}_{p_2}$		
$\mathbf{R}_3$	$\mathbf{I}_{p_3} - (\mathbf{e}_{p_3,1}\mathbf{e}_{p_3,1}^T)$		

Table 3.6: Table of models, variance-covariance matrices, $\mathbf{V}_l$ and prior precision matrices, $\mathbf{R}_l$ $(l = 1, 2, 3)$, for the two-stage designs found using the coordinate exchange algorithms in Section 3.5.1 and the objective function (3.20). Note that $\mathbf{e}_{n,j}$ is the jth column of $\mathbf{I}_n$.

The three multi-stage designs for $n = 12$ and the three multi-stage designs for $n = 16$ found for the models in Table 3.6 are compared using the column correlation matrix discussed in Section 3.5.2. This comparison is made to establish what impact the number of runs and the restrictions have on the correlation between columns of model matrices, and hence the variance and bias related to the fixed effect parameters for these models.

We can also use correlation between columns of model matrices as a method of assessing the projectivity of these designs. A design is said to have good projection properties when sub-designs created from subsets of the columns of the design have desirable properties, hence we consider projectivity in Section 3.6.1 as we assess the performance for these two-stage designs for three models which are based on subsets of the factors in the final design.

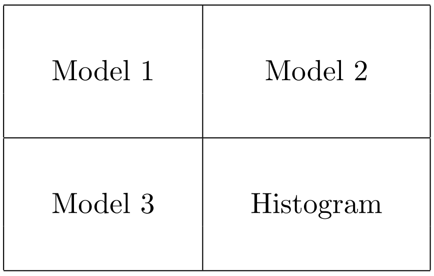
We consider projectivity in further detail in Section 3.6.2, as we compare designs with good projection properties to the optimal two-stage completely randomised designs. The designs with good projection properties considered are the 12 run Plackett-Burman design, given in Table 3.4, and the 16 run Hall III design, given in Table 3.5, which were identified as having good projection properties by Cheng (2006) and Loeppky et al. (2007), respectively.

In Section 3.6.2, we begin by using relative efficiency, as defined in Section 3.5.3, to compare the optimal combination of six columns of the Plackett-Burman design and the 16 run Hall III design to the optimal two-stage completely randomised designs found using random starting designs. This comparison is made to assess how well designs with good projection properties perform with respect to optimal designs found using a coordinate exchange algorithm. If these designs have high efficiencies, then it may be that tables of optimal multi-stage could be produced using the columns of these designs and hence could be identified without computation.

We also use both the column correlation matrices and relative efficiency to compare the optimal two-stage completely randomised designs found and analysed in Section 3.6.1 to designs found using the coordinate exchange algorithm with randomly selected subsets of the columns in the Plackett-Burman and Hall III designs as starting designs, to asses whether designs with good projectivity properties make good starting designs. Using subsets of the columns from designs with good projectivity properties as starting designs would reduce the number of possible starting designs, and could therefore have computational benefits.

Loeppky et al. (2007) used projection estimation capacity (PEC) to assess the projectivity of designs, where the design with the highest PEC over a range of k, $k = 1, \ldots, f$, has good projectivity properties. The PEC is $p_k = p_k(\mathbf{D})/t_k$, where $p_j(\mathbf{D})$ is the number of models containing the main effects and two factor interactions of k factors that can be estimated from $\mathbf{D}$ and t_k is the number of possible models with the main effects and two factor interactions of k factors that could be considered for a design with f columns. In Section 3.6.2, we use relative D-efficiency to assess the projectivity of the designs found using the coordinate exchange algorithm with the Plackett-Burman and Hall III designs as starting designs.

The column correlation matrices for the three models considered for the two responses are presented as heat maps in both Sections 3.6.1 and 3.6.2. Figure 3.5 is a schematic showing the arrangement of these column correlation matrices, and Table 3.7 gives the relationship between the axis labels in the correlation matrix heat maps and the columns of the model matrix.



Model 1	Model 2
Model 3	Histogram

Figure 3.5: Schematic for the arrangement of the figures containing the column correlation matrix heat maps and histogram of the correlation for the three models considered for the two responses from the two-stage designs.

Axis Label	Model 1	Model 2	Model 3
1	$\mathbf{1}$	$\mathbf{1}$	$\mathbf{1}$
2	$\mathbf{f}_1$	$\mathbf{f}_6$	$\mathbf{f}_1$
3	$\mathbf{f}_2$		$\mathbf{f}_2$
4	$\mathbf{f}_3$		$\mathbf{f}_3$
5	$\mathbf{f}_4$		$\mathbf{f}_4$
6	$\mathbf{f}_5$		$\mathbf{f}_5$
7	$\mathbf{f}_1\mathbf{f}_2$		$\mathbf{f}_6$
8	$\mathbf{f}_1\mathbf{f}_3$		$\mathbf{f}_1\mathbf{f}_2$
9	$\mathbf{f}_1\mathbf{f}_4$		$\mathbf{f}_1\mathbf{f}_3$
10	$\mathbf{f}_1\mathbf{f}_5$		$\mathbf{f}_1\mathbf{f}_4$
11	$\mathbf{f}_2\mathbf{f}_3$		$\mathbf{f}_1\mathbf{f}_5$
12	$\mathbf{f}_2\mathbf{f}_4$		$\mathbf{f}_1\mathbf{f}_6$
13	$\mathbf{f}_2\mathbf{f}_5$		$\mathbf{f}_2\mathbf{f}_3$
14	$\mathbf{f}_3\mathbf{f}_4$		$\mathbf{f}_2\mathbf{f}_4$
15	$\mathbf{f}_3\mathbf{f}_5$		$\mathbf{f}_2\mathbf{f}_5$
16	$\mathbf{f}_4\mathbf{f}_5$		$\mathbf{f}_2\mathbf{f}_6$
17			$\mathbf{f}_3\mathbf{f}_4$
18			$\mathbf{f}_3\mathbf{f}_5$
19			$\mathbf{f}_3\mathbf{f}_6$
20			$\mathbf{f}_4\mathbf{f}_5$
21			$\mathbf{f}_4\mathbf{f}_6$
22			$\mathbf{f}_5\mathbf{f}_6$

Table 3.7: Relationship between axis labels and columns of the model matrix for the correlation matrix heat maps. Any missing values represent where the axis labels end in the heat maps.

In Figure 3.5, Table 3.7, and throughout the rest of this section; (i) Model 1 is the model which relates the Stage 1 responses to the Stage 1 factors, (ii) Model 2 is the model which relates the Stage 2 responses to the Stage 2 factors and (iii) Model 3 is the cumulative model which relates the Stage 2 responses to all six factors in both stages.

As seen in Figure 3.5 and throughout Sections 3.6.1 and 3.6.2, there is a histogram in all figures containing the correlation matrix heat maps. This histogram shows the frequency of correlations between columns in $\mathbf{C}_3$, the column correlation matrix for model 3.

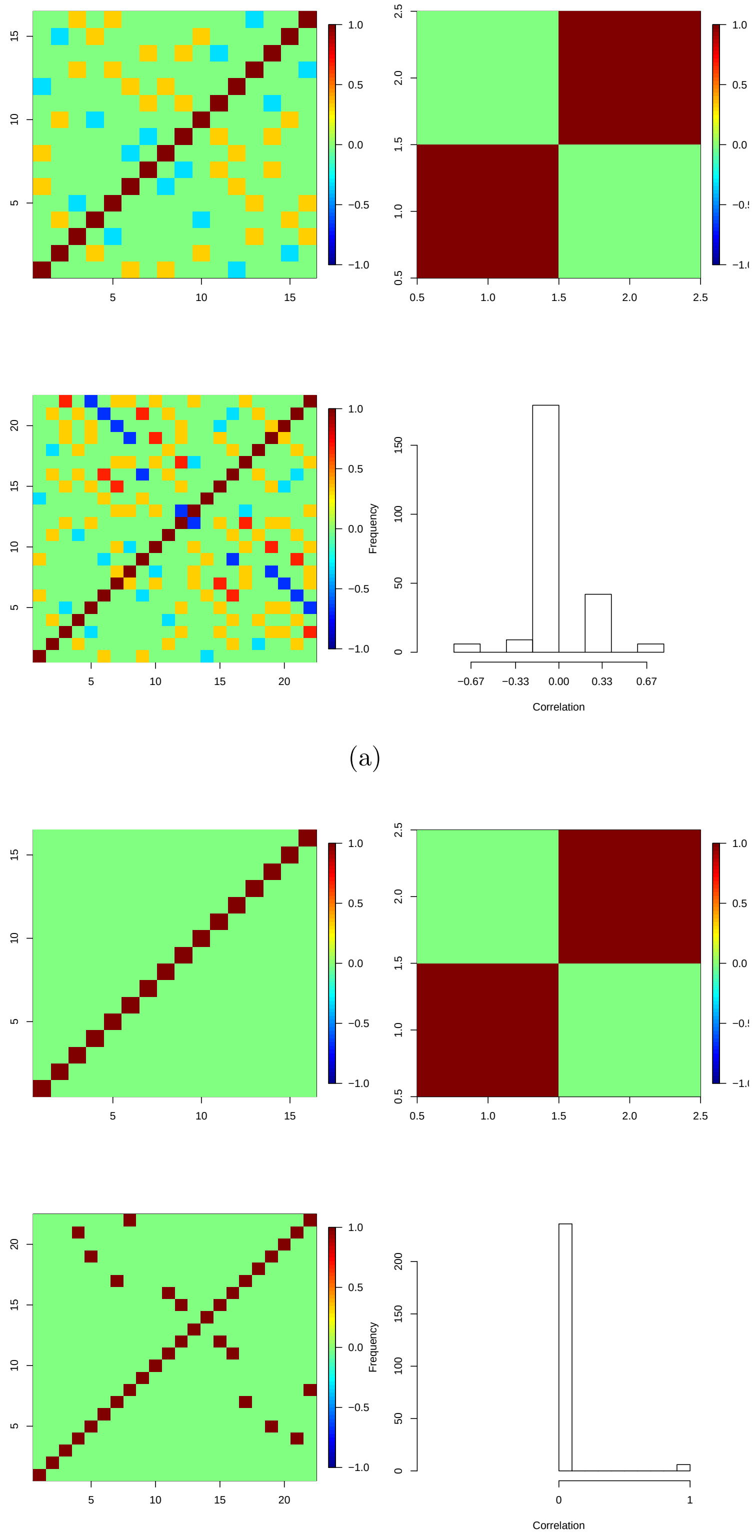
In Table 3.7, $\mathbf{1}$ is a $n \times 1$ column of ones, $\mathbf{f}_i$ is the ith column, $i = 1 \ldots, f$, of the design matrix $\mathbf{D} \in \mathcal{D}_{f,l,n}$, and $\mathbf{f}_i\mathbf{f}_j$ is the column of the model matrix $\mathbf{X}$ created by element wise multiplication of $\mathbf{f}_i$ and $\mathbf{f}_j$ (the i and jth column of the design matrix $\mathbf{D}$).

3.6.1 Study of the Impact of Run Size and Restrictions on Randomisation on Optimal Multi-Stage Designs

In this section, we are going to compare the column correlation matrices, (3.21) in Section 3.5.2, for optimal two-stage designs with different restrictions on randomisation, which are found using the coordinate exchange algorithms in Section 3.5.1 with ϕ_D from Section 3.4.2 and random starting designs. Throughout this section we refer to the optimal two-stage completely randomised design with 12 runs as $\mathbf{D}_{CRD12}$, the optimal two-stage completely randomised designs with 16 runs as $\mathbf{D}_{CRD16}$, the optimal two-stage split-plot design with 12 runs as $\mathbf{D}_{SLPD12}$, the optimal two-stage split-plot design with 16 runs as $\mathbf{D}_{SLPD16}$, the optimal two-stage strip-plot design with 12 runs as $\mathbf{D}_{STPD12}$ and the optimal two-stage strip-plot design with 16 runs as $\mathbf{D}_{STPD16}$. Note that there is a diagonal line of dark red squires going from the bottom left to the top right corners of every column correlation matrix heat map in Figures 3.6, 3.7 and 3.8, as every column has a correlation of 1 with itself.

We note from comparison of Figures 3.6a, 3.7a and 3.8a to Figures 3.6b, 3.7b and 3.8b that there are more correlated columns in the model matrices for 12 run designs than in the model matrices for 16 run designs. When a model is fitted to an n run design, and we assume that estimating the variance components is not of interest, we can estimate at most n parameters. This result was expected, as the number of correlated columns in a model matrix with p columns must increase as n decreases, and the rank of the model matrix decreases.

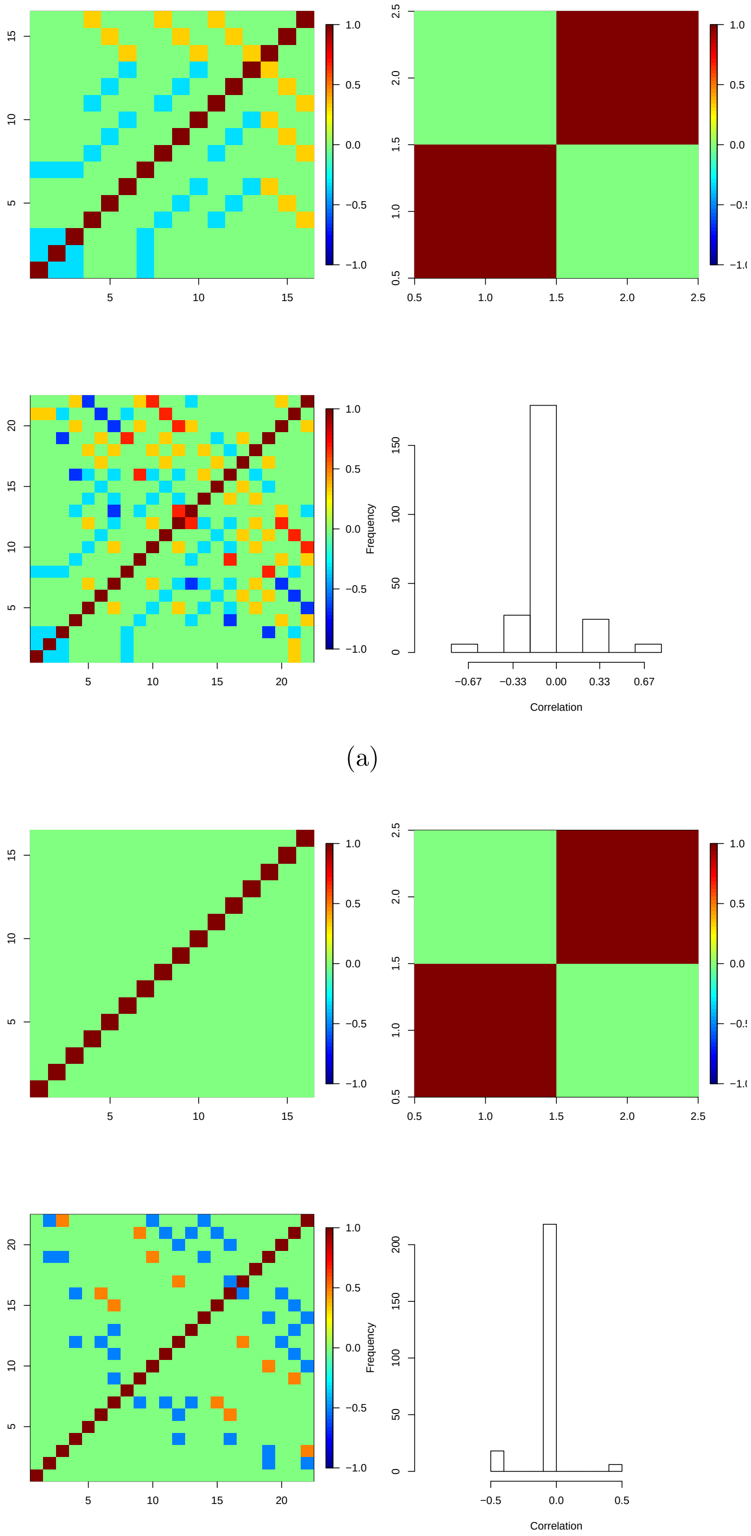
As seen in Figure 3.6b and 3.7b, the columns in the matrices for Model 1 for $\mathbf{D}_{CRD16}$ and $\mathbf{D}_{SLPD16}$ all have a correlation of zero. Therefore all the parameters in this model can be estimated independently, and only the columns involving Factor 6 ($\mathbf{f}_6$, $\mathbf{f}_{16}$, $\mathbf{f}_{26}$, $\mathbf{f}_{36}$, $\mathbf{f}_{46}$ and $\mathbf{f}_{56}$) are correlated in Model 3. This means that the bias and variance of the parameters relating to the single second stage factor, Factor 6, in Model 3 will be inflated, hence estimating these parameters will be more difficult.



(a)

(b)

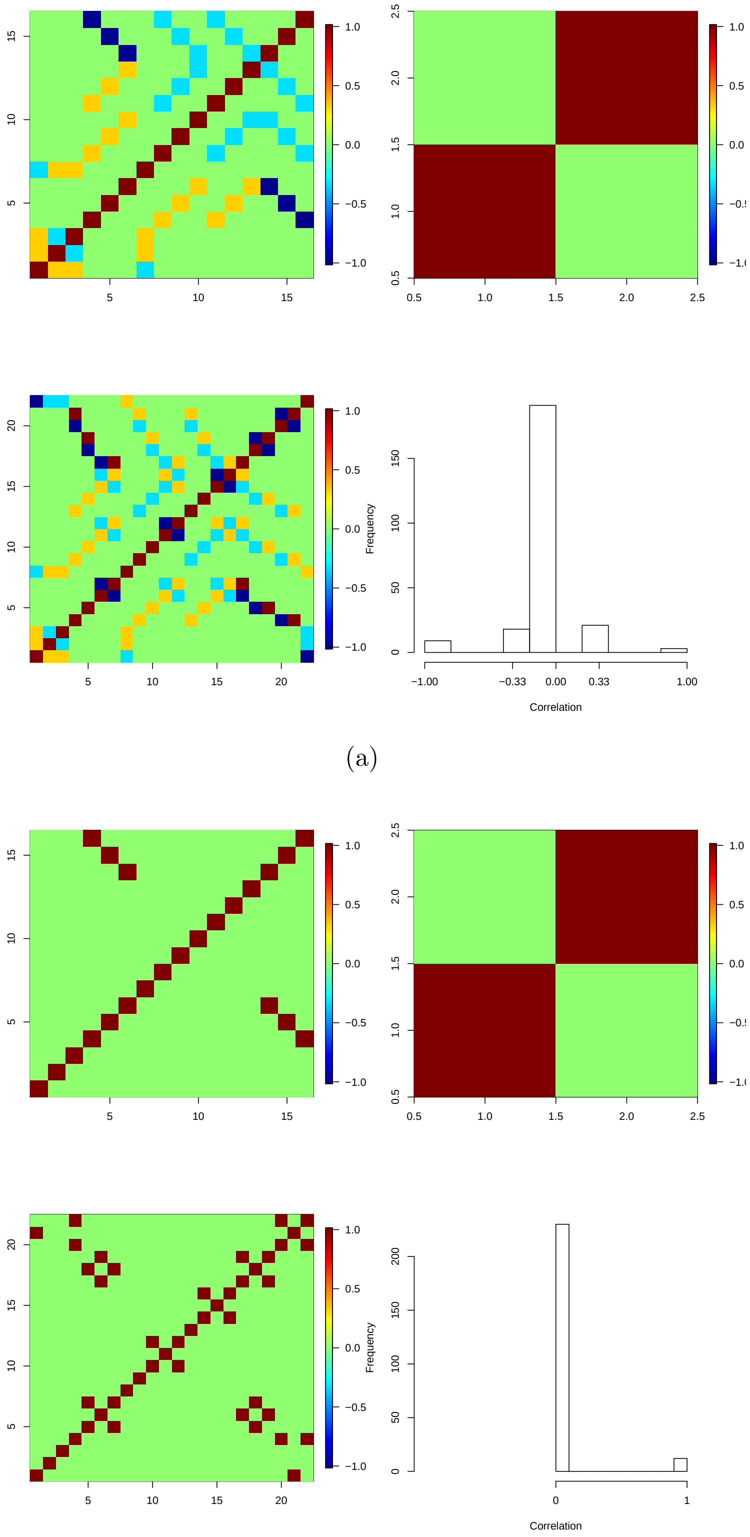
Figure 3.6: Heat maps depicting the column correlation matrices for the three models fitted to responses and histogram of correlations between columns of the model matrix for Model 3 for (a) $\mathbf{D}_{CRD12}$ and (b) $\mathbf{D}_{CRD16}$.



(a)

(b)

Figure 3.7: Heat maps depicting the column correlation matrices for the three models fitted to responses and histogram of correlations between columns of the model matrix for Model 3 for (a) $\mathbf{D}_{SLPD12}$ and (b) $\mathbf{D}_{SLPD16}$.



(a)

(b)

Figure 3.8: Heat maps depicting the column correlation matrices for the three models fitted to responses and histogram of correlations between columns of the model matrix for Model 3 for (a) $\mathbf{D}_{SRPD12}$ and (b) $\mathbf{D}_{SRPD16}$.

We note from comparison of the heat maps for the column correlations and histograms in Figures 3.6a and 3.7a, that the number of correlated terms and the correlations for $\mathbf{D}_{CRD12}$ and $\mathbf{D}_{SLPD12}$ are the same. There are 63 pairs of columns in the model matrix for Model 3 for both $\mathbf{D}_{CRD12}$ and $\mathbf{D}_{SLPD12}$ that are correlated, and for both designs this correlation is between -0.67 and 0.67. The restrictions on randomisation in $\mathbf{D}_{SLPD12}$ have affected which columns are correlated in the model matrices for these designs, as a number of the correlated columns involve the whole-plot factors. The relative efficiency (3.22) of $\mathbf{D}_{SLPD12}$ and $\mathbf{D}_{CRD12}$ when $\mathbf{V}_l = (\sigma_\epsilon^2 + \sigma_\gamma^2)\mathbf{I}_n$, $l = 1, 2, 3$, is 99.95%. Therefore, the aliasing structure and the value of (3.20) when $\mathbf{V}_l = (\sigma_\epsilon^2 + \sigma_\gamma^2)\mathbf{I}_n$ is very similar for $\mathbf{D}_{SLPD12}$ and $\mathbf{D}_{CRD12}$, even though these designs are found for two different randomisation assumptions.

The contrast between Figures 3.6b and 3.7b is interesting to note, as there are more correlated columns for Model 3 in Figure 3.7b than in Figure 3.6b, however the columns in the matrix for Model 3 in Figure 3.6b are fully correlated. Therefore, some of the parameters in Model 3 for the Stage 2 response from $\mathbf{D}_{CRD16}$ cannot be independently estimated. The relative efficiency (3.22) of $\mathbf{D}_{SLPD16}$ and $\mathbf{D}_{CRD16}$ when $\mathbf{V}_l = (\sigma_\epsilon^2 + \sigma_\gamma^2)\mathbf{I}_n$, $l = 1, 2, 3$, is 100%. Therefore, $\mathbf{D}_{SLPD16}$ maximises (3.20) when $\mathbf{V}_l = (\sigma_\epsilon^2 + \sigma_\gamma^2)\mathbf{I}_n$ and information from all the parameters in (1.6) for the Stage 2 responses can be recovered. Hence, we may choose to run $\mathbf{D}_{SLPD16}$, which is the optimal design found assuming two of the factors are hard-to-change, as a completely randomised design experiment over $\mathbf{D}_{CRD16}$.

Figures 3.7b and 3.8b also have an interesting comparison, as there are fewer correlated columns in the model matrices for $\mathbf{D}_{STPD16}$ than in $\mathbf{D}_{SLPD16}$, even though there are two restrictions on randomisation in strip-plot designs and one in split-plot designs. However, it is important to note from Figure 3.7b that none of the columns in the model matrices for $\mathbf{D}_{SLPD16}$ are fully correlated, whereas we can see from Figure 3.8b that all the columns in the model matrices for $\mathbf{D}_{STPD16}$ are fully correlated.

As stated in Section 3.5.2, the larger the correlation between columns, the more inflated and biased the parameter estimates relating to those columns will be in the model. We note, therefore, that the impact of increasing the number of restrictions on randomisation from one to two is not to increase the amount of correlated columns, but increase the severity of the correlation between columns and hence inflate the variances and biases related to these parameters.

The results of the comparisons of Figures 3.6a and 3.7a, Figures 3.6b and 3.7b and Figures 3.7b and 3.8b were unexpected, as we assumed that the number of correlated terms would increase as the restrictions on randomisation increase. These results demonstrate the limitations of a “one number” optimisation approach, as all of these designs were found to maximise the compound Bayesian D-optimality criterion (3.20), which does not consider the correlation between the columns in the model matrix. Therefore, the designs found could be improved by including a check of the correlation in the coordinate exchange algorithm. For example, each swap could be assessed to see if it improves (3.20) and also reduces either the level of correlation between the columns in the model matrices, or the number of columns which are correlated.

3.6.2 Comparison of D-optimal Designs to Designs with Desirable Projection Properties

In this Section we use relative efficiency (3.22) to compare $\mathbf{D}_{CRD12}$ and $\mathbf{D}_{CRD16}$, the two-stage 12 and 16 optimal completely randomised designs, respectively, found using the coordinate exchange algorithm with random starts, to:

- randomly selected subsets of six columns from the 12 run Plackett-Burman (Table 3.4) and the subset of six columns from the 16 run Hall III (Table 3.5) design which was identified by Loeppky et al. (2007) as having maximum PEC for $k = 6$, and
- $\mathbf{D}_{PB}$, the 12 run optimal two-stage completely randomised design found using the coordinate exchange algorithm with (3.20) and all possible subsets of six columns from the 12 run Plackett-Burman Design as starting designs (Table 3.4), and $\mathbf{D}_H$, the 16 run optimal two-stage completely randomised design found using the coordinate exchange algorithm with (3.20) and all possible subsets of six columns from the 16 run Hall III design as starting designs (Table 3.5).

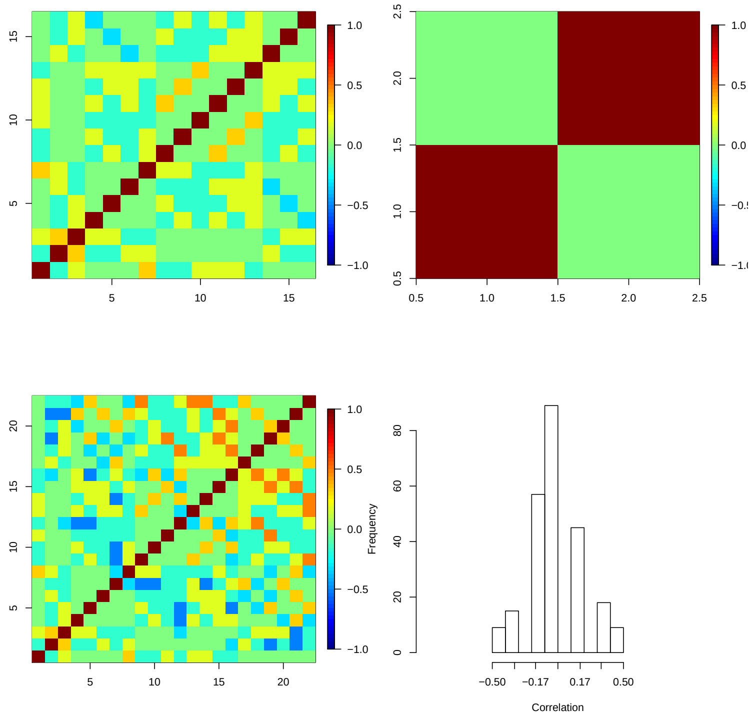


Figure 3.9: Heat maps depicting the column correlation matrices for the three models fitted to responses and histogram of correlations between columns of the model for Model 3 for $\mathbf{D}_{PB}$.

The efficiency of 396 of the 462 possible combinations of six columns of the Plackett-Burman design (Table 3.4) compared to $\mathbf{D}_{CRD12}$ is 95.58% (2dp), and the remaining 66 possible combinations of six columns of the Plackett-Burman design have a relative efficiency with respect $\mathbf{D}_{CRD12}$ of 91.43%. These efficiencies are quite high, hence these randomly selected columns of the Plackett-Burman design do perform relatively well when compared to $\mathbf{D}_{CRD12}$. However, these efficiencies are not high enough to suggest optimal designs can be found from this design without computation.

The relative efficiency of $\mathbf{D}_{PB}$ compared to $\mathbf{D}_{CRD12}$ is 99.98%. The relative efficiencies of the projections of $\mathbf{D}_{PB}$ and $\mathbf{D}_{CRD12}$ for the three models considered for the two responses assumed for this design, which can be calculated using (3.22) with $w_{l^*} = 1$ for the appropriate model, are (100%, 100%, 99.93%). Therefore, $\mathbf{D}_{PB}$ has good projectivity properties, when projectivity is measured using relative efficiency, when compared to $\mathbf{D}_{CRD12}$.

Comparison of Figures 3.6a and 3.9 shows the difference in which columns are correlated in the model matrices for $\mathbf{D}_{CRD12}$ and $\mathbf{D}_{PB}$, respectively, and the level of correlation for the correlated columns. We also note from comparison of Figures 3.6a and 3.9 that there are significantly more correlated columns in the model matrices for $\mathbf{D}_{PB}$ (153 of the columns are correlated in the model matrix for Model 3) than there are for $\mathbf{D}_{CRD12}$ (63 of the columns are correlated in the model matrix for Model 3), however the range of the correlation for the columns in the model matrices for $\mathbf{D}_{PB}$ is smaller (the correlations are all between -0.5 and 0.5) than the correlation for the columns in the model matrices for $\mathbf{D}_{CRD12}$ (the correlations are all between -0.67 and 0.67).

Deciding between $\mathbf{D}_{CRD12}$ and $\mathbf{D}_{PB}$ is difficult. However, we would recommend $\mathbf{D}_{CRD12}$ over $\mathbf{D}_{PB}$, as the difference in the maximum and minimum correlation is only 0.17 and number of correlated columns is much higher (there are 90 more correlated columns in the model matrix for $\mathbf{D}_{PB}$ than in $\mathbf{D}_{CRD12}$), and hence the number of parameters with inflated variances and bias of the parameter estimates in the models fitted to responses will be higher. Also, the value of (3.20) for $\mathbf{D}_{CRD12}$ is slighter higher than the value of (3.20) for $\mathbf{D}_{PB}$.

Columns 1, 2, 3, 4, 8, 10 and 12 of the Hall III design (Table 3.5) are identified as forming the best design with respect to PEC for a projection onto six factors by Loeppky et al. (2007), and this design had a relative D-efficiency (3.22) of 85.73% (2dp) when compared to $\mathbf{D}_{CRD16}$. This is quite low, and suggests that finding an multi-stage optimal design for this example has significant benefits over selecting a design with good projection properties.

The relative D-efficiency of $\mathbf{D}_H$ compared to $\mathbf{D}_{CRD16}$ is 100%. The relative efficiencies of the projections of $\mathbf{D}_H$ and $\mathbf{D}_{CRD16}$ for the three models are (100%, 100%, 100%). Therefore, $\mathbf{D}_H$ also has excellent projectivity properties, with respect to relative efficiency, when compared to $\mathbf{D}_{CRD16}$. However, comparison of Figures 3.6b and 3.10 shows that, even though the values of (3.20) for $\mathbf{D}_{CRD16}$ and $\mathbf{D}_H$ when $\mathbf{w}$ in (3.20) is (0.7, 0.1, 0.2), (1,0,0), (0,1,0) and (0,0,1) are the same, the pattern of correlation between columns in the model matrices for these designs are not the same.

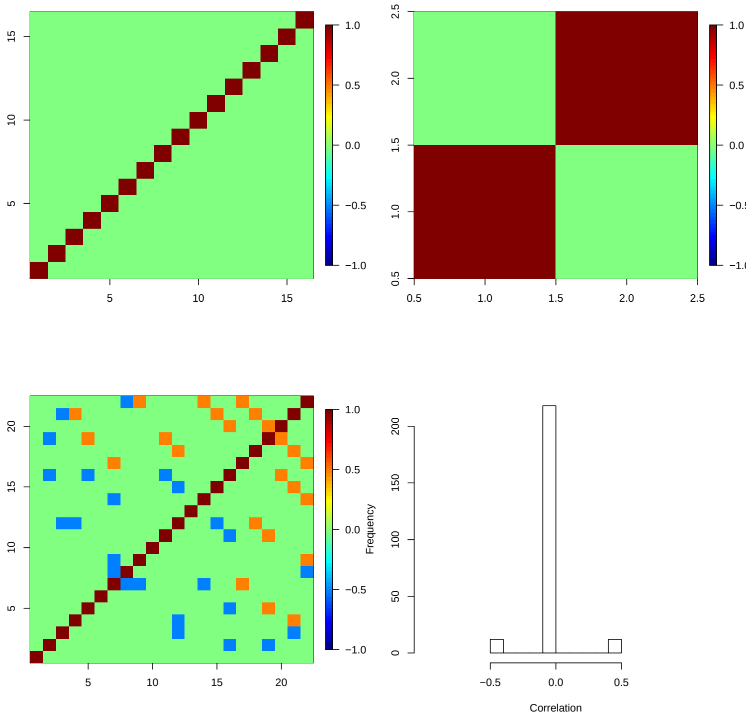


Figure 3.10: Heat maps depicting the column correlation matrices for the three models fitted to responses and histogram of correlations between columns of the model for Model 3 for $\mathbf{D}_H$.

Notice that the columns in the model matrices for Models 1 and 2 are independent for both $\mathbf{D}_{CRD16}$ and $\mathbf{D}_H$, but the correlations between the columns in the model matrix for Model 3 for $\mathbf{D}_H$ are either -0.5 or 0.5, whereas the correlations between columns in the model matrix for Model 3 for $\mathbf{D}_{CRD16}$ are either 0 or 1. We also note that $\mathbf{D}_H$ has 24 correlated columns whereas $\mathbf{D}_{CRD16}$ has only 6. However, the significant difference in the range of the correlated columns is important, as it means that the parameters which are correlated in $\mathbf{D}_{CRD16}$ will have variances and biases that are significantly more inflated than the parameters which are correlated in $\mathbf{D}_H$. Therefore, $\mathbf{D}_H$ may be preferable when compared to $\mathbf{D}_{CRD16}$ even though the model matrices for Model 3 for $\mathbf{D}_H$ has more correlated columns, however further assessment of this variance and bias inflation would need to be considered. Once again, this shows the importance of considering both efficiency and correlation when assessing models.

These results demonstrate the limitations of using only (3.20) to optimise designs when the correlation between columns in the model matrices, and hence the aliasing between parameters in the models assumed for the responses from the experiment, is also important. Designs which are equivalent under (3.20) are not necessarily equivalent with respect to (3.21). Therefore, as we have seen through our comparison of $\mathbf{D}_{PB}$ and

$\mathbf{D}_{CRD12}$, and $\mathbf{D}_H$ and $\mathbf{D}_{CRD16}$, comparisons between designs which have 100% relative efficiency can be complicated.

3.7 Discussion

The main focus of this chapter is the design and assessment of multi-stage designs with restricted randomisation, as defined in Section 3.1.1, which are appropriate for the motivation given in Section 3.1.3. Optimal multi-stage designs for (3.20), the compound Bayesian D-optimality objective function derived in Section 3.4.2, are found using coordinate exchange algorithms presented in Section 3.5.1. Our definition of multi-stage designs is most similar to the definition of partition designs provided by Perry et al. (2001, 2002, 2007) and Pieracci et al. (2010).

The correlation between columns in the model matrices for the various models considered for multi-stage designs is introduced as a method of assessing the optimal multi-stage designs in Section 3.5.2, and we noted that correlation between columns of the model matrix inflates the bias and variance of the parameter estimates related to those columns. Also, parameters relating to the columns which are correlated can not be estimated independently, and will therefore be aliased.

In Section 3.6.1, we used the column correlation matrices to compare two-stage designs with different restrictions on randomisation which are appropriate for our motivating example. We noted that, as expected, the 16 run designs have fewer correlated columns than the 12 run designs. We also noted that only columns relating to factor 6 are correlated in the matrix for Model 3 for $\mathbf{D}_{CRD16}$ and $\mathbf{D}_{SLPD16}$ (the optimal 16 run two-stage completely randomised and split-plot design, respectively).

The results from Section 3.6.1 highlighted the limitations of using a "one number" approach to optimisation, as the relative efficiencies and the comparison of correlation matrices for $\mathbf{D}_{SLPD16}$ and $\mathbf{D}_{CRD16}$ suggested that using $\mathbf{D}_{SLPD16}$, the optimal design found assuming two of the factors are hard-to-change, as a completely randomised design is preferable to using $\mathbf{D}_{CRD16}$. We also found that the number of correlated columns does not necessarily increase as more restrictions on randomisation are applied.

In Section 3.6.2, we compare our optimal completely randomised designs to designs identified by Loeppky et al. (2007) as having good projectivity properties using relative efficiency, as defined in Section 3.5.3, and column correlation matrices. Projectivity is important in this work, as we fit models to different subsets of the model space and our compound criterion ensures the designs perform well, with respect to Bayesian D-optimality, in each of these subsets.

In Section 3.6.2 we found that there is some benefit, with respect to the compound Bayesian D-optimality objective function, in finding optimal designs using the coordinate exchange algorithm over selecting columns from a design with good projectivity properties. We also found that $\mathbf{D}_{PB}$ and $\mathbf{D}_H$ (the optimal designs found using the coordinate exchange algorithm with all possible combinations of six columns from the Plackett-Burman and Hall III designs, respectively, as starting designs) have good projectivity over the three models considered in this chapter, when projectivity is assessed using the relative efficiency of these design compared to $\mathbf{D}_{CRD12}$ and $\mathbf{D}_{CRD16}$, respectively.

The limitations of a one number approach to optimisation were also seen in Section 3.6.2, as designs which had 100% efficiency with respect to the compound criterion, and over each projection, did not necessarily have the same correlated columns. The number of correlated columns is important, as we want to be able to gain as much information about as many parameters as possible, and the correlation between columns indicates the amount of information that can be gained about each parameter.

The work in this chapter could be extended in a number of different ways. The number of factors and the number of whole-plot and row factors considered in the designs in this chapter are specific to the formulation of a pharmaceutical product. However we could look at what affect grouping factors into different stages or increasing the number of stages has on the correlation matrices.

Also, we have only considered set values of the tuning parameters in (3.20), and hence an assessment the robustness of designs with respect to $\mathbf{w}$ and $\mathbf{R}_l$, $l = 1, \ldots, m$, would extend the current results. We assess the robustness of specific designs with respect to $\mathbf{w}$ in Section 5.2.1 of Chapter 5, but we have not performed a more general assessment of robustness.

We could also consider a different optimality criteria, such as a compound pseudo-Bayesian version of A-optimality, and compare the correlation matrices for the optimal designs found using compound Bayesian D- and A-optimality. We could also assess whether the optimal designs found using compound Bayesian D-optimality perform well with respect to compound Bayesian A-optimality.

In Sections 3.6.1 and 3.6.2, we suggested that considering the correlation matrix as well as the compound Bayesian D-optimality objective function in the coordinate exchange algorithm could help make the comparison between designs easier. We could implement this extension by optimising the number of correlated terms and some summary of the correlated terms, such as the mode, as well as (3.20).

Another possible extension is to consider alternative algorithms for finding optimal designs, as the coordinate exchange algorithm is a “greedy algorithm” which can stuck at local optima. Stochastic algorithms, such as simulated annealing algorithm where moves are accepted or rejected with some probability (Brooks and Morgan, 1995), could be considered as they accept sub-optimal moves and hence allow us to escape from local optima, and the designs found using these two algorithms could be compared.

Chapter 4

Bayesian Variable Selection for Supersaturated Split-Plot Experiments

The identification of influential factors is often important, particularly in supersaturated experiments where there are more terms in the model fitted to the response than runs in the experiment. We refer to substantially non-zero terms in the model as active. There are a variety of different variable selection methods for identifying these active terms for different types of experiments. In this chapter, we focus on variable selection methods when mixed models for multivariate responses are used to model data from split-plot experiments, which are appropriate for our motivating example (Section 4.1).

Bayesian variable selection is the focus of this chapter. Bayesian methodology allows the use of prior knowledge, and ensures that the uncertainty associated with parameters and models before experimentation is propagated through to estimated parameters and predicted responses. In Section 4.2, we describe the Bayesian framework, introduce Markov chain Monte Carlo (MCMC) sampling and Bayesian variable selection.

In Section 4.3, we provide motivation for our Bayesian variable selection method by assessing the performance of example frequentist (Section 4.3.2) and Bayesian (Section 4.3.3) variable selection methods for simulated data. In Section 4.3.2, we find that the frequentist method for analysing split-plot experiments is limited for supersaturated split-plot designs, as frequentist analysis relies on estimates of the variance components, and these cannot be accurately calculated due to a lack of degrees of freedom. These limitations were also identified by Gilmour and Goos (2009). Placing a prior distribution on the variance components in the mixed model supplements the data with available prior information, mitigating the estimation problems.

Advances in computing have made Bayesian methods easier to implement. Hence, Bayesian methodology has become more popular. MCMC methods can be used to address the complexity and multi-dimensionality of the likelihoods, posterior and conditional distributions associated with Bayesian modelling of responses. MCMC sampling methods produce a dependent sample from a given distribution via a Markov chain with the correct stationary distribution.

Our algorithm, as discussed in Section 4.4, combines two MCMC methods; Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990) and Metropolis-Hastings rejection sampling (Metropolis et al., 1953; Hastings, 1970). The Gibbs sampler uses samples drawn from lower dimensional conditional distributions to create dependent samples from non-normalised high dimensional posterior distributions. Metropolis-Hastings sampling is a form of rejection sampling, where a proposed new parameter value is either accepted or rejected with some probability.

The algorithm presented in Section 4.4.4 relies on the extension of the linear mixed model and prior distributions from Tan and Wu (2013) to multivariate responses. The linear mixed model for multivariate responses is presented in Section 4.4.1. The prior distributions suitable for variable selection are specified in 4.4.2.

In Section 4.4.3 and Appendix E we show how the conditional distributions for the parameters in the model assuming multivariate responses, which depend on the extended prior distributions, are calculated. The Metropolis-Hastings within Gibbs Sampling algorithm also relies on an extension of the joint sampling approach by Geweke (1996) to multivariate responses from a split-plot designed experiment, as discussed in Section 4.4.3 and Appendix F.

In Section 4.5, we demonstrate how samples from our Metropolis-Hastings within Gibbs sampling algorithm can be used to perform variable selection, and establish how well the Metropolis-Hastings within Gibbs sampling algorithm performs for simulated data when the variability in the data is large compared to the size of the active terms, and vice versa. The methodology in this chapter is applied to the responses from dissolution testing of a pharmaceutical product formulated by GSK in Chapter 5.

4.1 Motivation and Aim of Work

The motivation for this work is the dissolution testing of a pharmaceutical product formulated using designs discussed in Chapter 3, which was introduced in Section 1.2.2 in Chapter 1. After further discussions with the scientists regarding the experimental process, we confirmed that a sixteen run two-stage split-plot design would be appropriate for this experiment. It was anticipated that two responses will be measured; one after the sub-treatments involving Factors 1 to 5 have been applied (Stage 1) and another after the sub-treatments involving Factor 6 have been applied (Stage 2). The response from Stage 1 is assumed to be some quality control measure and the response from Stage 2 is the result of dissolution testing of the pharmaceutical product (which is discussed in further detail Section 5.1).

As this experiment is an initial screening experiment, the scientists wish to identify which factors should be used in future experimentation. Therefore, the main aim of this chapter is to motivate and introduce an appropriate method of variable selection for the responses from this experiment, where the variable selection method also allows for estimation of the terms in the model. For an experiment with n runs, a $n \times 2$ response matrix is measured for dissolution testing, therefore we require a variable selection method for multivariate responses from linear mixed models which are appropriate for split-plot designs.

4.2 Introduction to Bayesian Methodology

In this section we overview the Bayesian methodology applied in this chapter. In Section 4.2.1 we define the prior and posterior distribution and introduce Bayes theorem. In Section 4.2.2 we discuss MCMC methods and introduce Metropolis-Hastings and Gibbs sampling. Finally, in Section 4.2.3 we introduce the key focus of this chapter, Bayesian variable selection.

4.2.1 Bayesian Framework

The basis of Bayesian inference is updating a prior distribution for an uncertain quantity to a posterior distribution via Bayes theorem. The prior distribution encapsulates our uncertainty prior to observing the data, and the posterior distribution summarises our uncertainty after observing the data. Let $f(\mathbf{y}|\boldsymbol{\theta})$ be the likelihood function given the parameter $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_p)^T \in \Theta$ for the data $\mathbf{y} = (y_1, y_2, \dots, y_n)^T$, and let $p(\boldsymbol{\theta})$ be a prior distribution for $\boldsymbol{\theta}$. Using Bayes theorem, the posterior density of $\boldsymbol{\theta}$ given $\mathbf{y}$ is

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{\int_{\Theta} f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}}. \tag{4.1}$$

The denominator of (4.1) is the marginal likelihood of $\mathbf{y}$ and is independent of $\boldsymbol{\theta}$, hence it can be treated as a constant with respect to $\boldsymbol{\theta}$ and

$$p(\boldsymbol{\theta}|\mathbf{y}) \propto f(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}). \tag{4.2}$$

Equation (4.2) gives the unnormalised posterior density. For further detail on Bayes theorem and Bayesian inference in general, see O'Hagan and Forster (2004).

The choice of prior distribution is subjective, and is not always obvious. Prior distributions can be informative or noninformative, proper or improper. Informative prior distributions rely on some knowledge about the parameter, and are often chosen to have a particular mean, mode or variance based on the experimenters' knowledge or past data (for more discussion regarding informative prior distributions, see Section 2.4 of Gelman et al., 2004). Noninformative prior distributions are used when the information regarding a parameter available prior to experimentation is weak.

The prior distribution can have an impact on the posterior, and noninformative priors are often favoured as they have minimal impact on the posterior distribution. However, the choice of noninformative prior has to be carefully considered (see Sections 3.27 to 3.33 of O'Hagan and Forster, 2004 and Section 2.9 of Gelman et al., 2004). Two types of noninformative prior distributions are often discussed, proper and improper. A proper distribution has a density that either integrates to 1, or can be normalised to integrate to 1. The integral of the density for an improper distribution is not finite.

Improper prior distributions are often described as being indicative of no knowledge of the parameter prior to experimentation. However O'Hagan and Forster (2004) explained that improper prior distributions actually indicate that the knowledge regarding a particular parameter prior to experimentation is weak, but not non-existent. Both O'Hagan and Forster (2004) and Gelman et al. (2004) encouraged caution when using improper prior distributions, as they can lead to improper posterior distributions which can create computational difficulties.

Many prior distributions which are improper and noninformative in one parametrisation are not so under another, and this can lead to misleading results. Jeffreys (1946) introduced a type of improper and noninformative prior distribution which is invariant under transformation. Jeffreys prior distribution is popular in literature due to this property of invariance, which holds for a number of transformations, but even these priors are inconsistent for some examples.

When the prior and posterior distributions have the same distributional form, for example, both the prior and posterior have a beta or gamma distribution with different hyperparameters, then the prior distribution is referred to as a conjugate prior distribution. Conjugate prior distributions, where they exist, make the calculation of the posterior distribution more straightforward.

4.2.2 Markov chain Monte Carlo Methods

The complex models and high dimensional posterior distributions considered in practical Bayesian modelling often makes analytic derivation of the posterior density impossible. For example, if $p(\boldsymbol{\theta}|\mathbf{y})$ in (4.2) is not a known density, direct sampling from the distribution will not be possible. Markov chain Monte Carlo (MCMC) methods aid Bayesian analysis as they allow generation of dependent samples from arbitrary joint probability distributions. They also allow the approximation of the posterior distribution and calculation of integrals such as the expectation.

MCMC methods sample from Markov chains, where the conditional distribution of a parameter sampled at iteration q is only dependent on the parameter sampled at iteration $q-1$ and not on the parameter sampled at iterations $1,\ldots,q-2$. The stationary distribution of a Markov chain is the distribution which the Markov chain converges to after a certain number of iterations. MCMC methods rely on Markov chains which have the required posterior distribution as their stationary distribution, and it is important that enough samples are drawn so that the distribution of the sampled parameters is similar enough to the required stationary distribution.

Metropolis-Hastings Sampling

Metropolis-Hastings sampling (Metropolis et al., 1953; Hastings, 1970) is a MCMC method based on rejection sampling that is commonly used for sampling from conditional distributions of unknown distributional form. Rejection sampling methods accepted or reject a proposed move in the parameter space with some probability.

Let $f(\theta|\mathbf{y})$ be the conditional distribution from which we wish to sample, $\theta^{(q)}$ be the current parameter value and $\pi(\theta)$ be the proposal density. A Metropolis-Hastings algorithm for sampling $\theta^{(q)}$ has the following steps:

1. Sample a candidate value, $\theta_*^{(q+1)}$, from a proposal distribution with density $\pi(\theta)$.
2. Calculate the acceptance probability

$$\alpha\left(\theta^{(q)},\theta_*^{(q+1)}\right)=\min\left\{1,\frac{f\left(\theta_*^{(q+1)}|\mathbf{y}\right)\pi\left(\theta^{(q)}\right)}{f\left(\theta^{(q)}|\mathbf{y}\right)\pi\left(\theta_*^{(q+1)}\right)}\right\}. \tag{4.3}$$

3. Sample u from $U(0,1)$.
4. If $u<\alpha\left(\theta^{(q)},\theta_*^{(q+1)}\right)$ then $\theta^{(q+1)}=\theta_*^{(q+1)}$ otherwise $\theta^{(q+1)}=\theta^{(q)}$.

Note that the acceptance probability, (4.3) in step 2, is equal to

$$\alpha(\theta^{(q)},\theta_*^{(q+1)})=\min\left\{1,\frac{f\left(\mathbf{y}|\theta_*^{(q+1)}\right)}{f\left(\mathbf{y}|\theta^{(q)}\right)}\right\}, \tag{4.4}$$

when the proposal distribution is the prior distribution, $\pi(\theta)=p(\theta)$, and the likelihood is $f(\mathbf{y}|\theta)$.

Gibbs Sampling

Gibbs sampling (Geman and Geman, 1984; Gelfand and Smith, 1990) is a special case of Metropolis-Hastings rejection sampling which draws dependent samples from the posterior distribution of parameters using the lower dimension conditional distributions, which have known form. The proposal density in Gibbs sampling is the conditional distribution of the current parameter, $\theta_j^{(q)}$, hence $\pi(\theta_j^{(q)}) = f(\theta_j^{(q)}|\boldsymbol{\theta}_{-j}^{(q)}, \mathbf{y})$, where $\boldsymbol{\theta}_{-j}^{(q)}$ is the current value for the other elements of $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$ and $j = 1, \ldots, p$.

The acceptance probability in step 2 of Metropolis-Hastings sampling when the proposal density is the conditional distribution is

$$\alpha\left(\theta_j^{(q)}, \theta_{j*}^{(q+1)}\right) = \min\left\{1, \frac{f\left(\theta_{j*}^{(q+1)}|\boldsymbol{\theta}_{-j}^{(q)}, \mathbf{y}\right) f\left(\theta_j^{(q)}|\boldsymbol{\theta}_{-j}^{(q)}, \mathbf{y}\right)}{f\left(\theta_j^{(q)}|\boldsymbol{\theta}_{-j}^{(q)}, \mathbf{y}\right) f\left(\theta_{j*}^{(q+1)}|\boldsymbol{\theta}_{-j}^{(q)}, \mathbf{y}\right)}\right\} = 1. \tag{4.5}$$

Therefore, every proposed sample from the conditional distribution is accepted, and at each iteration of the Gibbs sampling algorithm $\theta_j^{(q)}$ is sampled from $f(\theta_j^{(q)}|\boldsymbol{\theta}_{-j}^{(q)}, \mathbf{y})$, so

$$\begin{aligned}
\theta_1^{(q)} &\sim f\left(\theta_1|\theta_2^{(q-1)}, \theta_3^{(q-1)}, \ldots, \theta_p^{(q-1)}, \mathbf{y}\right) \\
\theta_2^{(q)} &\sim f\left(\theta_2|\theta_1^{(q)}, \theta_3^{(q-1)}, \ldots, \theta_p^{(q-1)}, \mathbf{y}\right) \\
&\vdots \\
\theta_p^{(q)} &\sim f\left(\theta_p|\theta_1^{(q)}, \theta_2^{(q)}, \ldots, \theta_{p-1}^{(q)}, \mathbf{y}\right).
\end{aligned}$$

Diagnostics

The performance of MCMC algorithms can be assessed in many ways, as discussed by, for example, Schafer (1997). The main focus of MCMC algorithm diagnostics is an assessment of whether the Markov chains in MCMC algorithms have converged to their stationary distribution. In Appendix C we use trace plots and autocorrelation function (ACF) plots to assess the performance of our Metropolis-Hastings within Gibbs sampling algorithm.

Trace plots plot the values of the parameters sampled using Markov chains against the iteration number, and use a line to join the values for successive samples. Long term trends in this plot for a given parameter, such as an upward or downward drift or shifts in the mean of the sample, suggest that successive sampled parameters are highly correlated and indicate that the Markov chain has not converged to the correct stationery distribution.

The correlation between iterative samples is assessed using ACF plots, which show the correlation against the lag, or distance, between sampled parameters. If the Markov chain for a give parameter converges to the required stationary distribution, then the correlations in the ACF plot will be low. The larger the correlation in the ACF plots, the fewer independent samples from the posterior the chain represents. Further detail on both trace and ACF plots are given in Appendix C.1.

4.2.3 Bayesian Variable Selection

Assume that models with up to p_{max} parameters are fitted to responses from a design with l-level factors. Then $n_m = l^{p_{max}}$ possible models of size $1, \ldots, p_{max}$ could be fitted to these responses. Each of these models, $M = 1, \ldots, n_m$, will have a prior probability, $p(M)$, of being correct, and the responses $\mathbf{y}$ will have a marginal likelihood conditional on the model, $p(\mathbf{y}|M)$. Therefore, using (4.2),

$$p(M|\mathbf{y}) \propto p(\mathbf{y}|M)p(M), \tag{4.6}$$

is the posterior probability of model M being appropriate given the responses $\mathbf{y}$.

Hence, in principal, the posterior probability for each of the n_m models could be calculated. The model which maximises $p(M|\mathbf{y})$ could then be selected as the model which best describes the relationship between the response and the factors in the experiment. The parameters in this model would then be used to identify which terms are active.

However, the likelihood $p(\mathbf{y}|M)$ is often not available in closed form and calculating (4.6) for n_m models, even when p_{max} is relatively small, is computationally expensive. Hence, MCMC methods are used to sample $p(M|\mathbf{y})$ and perform Bayesian variable selection.

The stochastic search variable selection (SSVS) algorithm discussed by George and McCulloch (1993, 1997) is a popular method of Bayesian variable selection. An indicator vector defining the activity of each model term is used in the SSVS algorithm. Gibbs samples from the posterior distribution of this indicator vector are used for variable selection. Variable selection for supersaturated designs and models with multivariate responses using the SSVS algorithm has been considered by authors such as Chipman et al. (2001) and Brown et al. (1998). However, we consider a different MCMC algorithm in Section 4.4 as our prior distribution for the fixed effects in the mixed model for multivariate responses from split-plot designs differs to that considered in the SSVS algorithm.

Other approaches for modelling multivariate responses are presented in the papers by Ng (2010) and Overstall and Woods (2015). Ng (2010) considered a Bayesian decision theoretic approach (Berger, 1985) to model multivariate responses and find optimal variable settings from completely randomised designs. Overstall and Woods (2015) used both parametric and non-parametric models to predict outputs for multivariate responses from computer models.

4.3 Motivation for Bayesian Variable Selection

In this section, we use the results of a frequentist and Bayesian variable selection method for simulated univariate data to motivate our derivation of a Metropolis-Hastings within Gibbs sampling algorithm (Section 4.4.4) to perform variable selection for multivariate responses from split-plot experiments, applied in Section 4.5 and Chapter 5.

4.3.1 Simulated Responses

We assess the performance of the two methods presented in Section 4.3.2 and 4.3.3 using univariate simulated responses for $\mathbf{D}_{SLP16}$, the sixteen run two-stage split-plot experiment from Chapter 3, which, as discussed in Section 4.1, is appropriate for the experiment motivating this work. This sixteen run experiment has six factors, two whole-plot and four sub-plot factors, which are applied in two-stages. The two whole-plot factors and three of the sub-plot factors are applied in Stage 1, and a single sub-plot factor is applied in Stage 2.

We note that we use univariate $(n \times 1)$ instead of multivariate $(n \times r)$ responses, as both the methods discussed in this section are for univariate data, and we want to consider a simple motivating example at this stage to motivate our use of the Metropolis-Hastings within Gibbs sampling algorithm (which can be used for univariate or multivariate responses).

Let $\mathbf{f}_j$, $j = 1, \ldots, 6$, be the jth column of the design matrix $\mathbf{D}_{SLPD16}$, then the models for the two simulated responses are:

- The model where the terms for $\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3, \mathbf{f}_4, \mathbf{f}_3\mathbf{f}_4, \mathbf{f}_3\mathbf{f}_5$ are active for the response from **Stage 1**. These simulated responses are generated using (1.3) where $\boldsymbol{\beta} = (4.80, 4.77, -3.73, -4.93, -4.83, 6.73)^T$, $\mathbf{X}$ is the model matrix corresponding to these terms, $\boldsymbol{\gamma}$ is a single sample from $\mathrm{N}(\mathbf{0}_4, 10\mathbf{I}_4)$ and $\boldsymbol{\epsilon}$ is a single sample from $\mathrm{N}(\mathbf{0}_{16}, \mathbf{I}_{16})$.
- The model where the terms for $\mathbf{f}_2, \mathbf{f}_4, \mathbf{f}_6$ are active for the response from **Stage 2**. These simulated responses are generated using (1.3) where $\boldsymbol{\beta} = (5.04, 5.48, -4.93)^T$, $\mathbf{X}$ is the model matrix corresponding to these terms, $\boldsymbol{\gamma}$ is a single sample from $\mathrm{N}(\mathbf{0}_4, 10\mathbf{I}_4)$ and $\boldsymbol{\epsilon}$ is a single sample from $\mathrm{N}(\mathbf{0}_{16}, \mathbf{I}_{16})$.

4.3.2 Frequentist Variable Selection via All Subsets Regression

All Subsets Regression

All subsets regression is a simple, but computationally expensive, method of variable selection. In all subsets regression, all models up to a certain size are fitted and the best model is selected with respect to some model selection criterion.

For an experiment with n runs and a model for the responses with v variance components, all subsets regression has the following steps:

1. For $j = 1, \ldots, n - v$:

   (a) Fit all models including j parameters to the response, $\mathbf{y}$.

   (b) Select the fitted model with j parameters which optimises the model selection criterion.

2. Select the fitted model from the $n - v$ models identified in Step 1(b) which optimises the model selection criterion $\forall j = 1, \ldots, n - v$.

The active variables correspond to the parameters included in the model selected in Step 2. The model selection criteria used in Step 1(b) rely on the estimates of the variance components, therefore only models with $p \leq n - v$ terms can be fitted.

Penalised Model Selection Criteria

Penalised model selection criteria for linear mixed models, such as (1.3), are functions of the log likelihood, evaluated at the maximum likelihood estimates, with terms which penalise over fitting. The penalties in these criteria are required to adjust for the fact that models with more terms will always maximise the likelihood, even if the additional terms in the model are just describing the noise in the response.

The maximised log likelihood for the model fitted to a split-plot design with n runs (discussed in Section 1.3 of Chapter 1 and Section 3.2.2 of Chapter 3) is

$$\ln \hat{L} = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln(|\hat{\mathbf{V}}|) - \frac{1}{2}(\mathbf{y} - \hat{\boldsymbol{\mu}})^T \hat{\mathbf{V}}^{-1}(\mathbf{y} - \hat{\boldsymbol{\mu}}) \tag{4.7}$$

where $\mathbf{y}$ is the $n \times 1$ vector of observed responses, $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ is the maximum likelihood estimate of the mean, $\hat{\boldsymbol{\beta}}$ is the $p \times 1$ maximum likelihood estimate of the vector of fixed effect parameters $\boldsymbol{\beta}$ for $\mathbf{y}$, $\hat{\mathbf{V}} = \hat{\sigma}^2_\gamma \mathbf{Z}\mathbf{Z}^T + \hat{\sigma}^2_\epsilon \mathbf{I}_n$ is the maximum likelihood estimate of the variance-covariance matrix, $\hat{\sigma}^2_\gamma$ is the maximum likelihood estimate of the whole-plot variance for $\mathbf{y}$ and $\hat{\sigma}^2_\epsilon$ is the maximum likelihood estimate of the sub-plot variance for $\mathbf{y}$.

Using the maximised log likelihood, (4.7), the following penalised model selection criterion can be defined:

1. *Akaike information criterion (AIC)*: The AIC,

$$-2\ln \hat{L} + 2p, \tag{4.8}$$

was derived by Akaike (1973) using an extension of the maximum likelihood approach.

2. *Bayesian information criterion (BIC)*: The BIC,

$$-2\ln \hat{L} + p\ln(n), \tag{4.9}$$

was derived from the large sample limits of Bayes estimators by Schwarz (1978), which differs from the method used by Akaike (1973) to derive the AIC.

3. *Penalised Akaike information criterion (pAIC)*: The pAIC,

$$-2\ln \hat{L} + 2p\left(\frac{n}{n-p-1}\right), \tag{4.10}$$

was proposed by Hurvich and Tsai (1989), as they noted that the AIC selects more parameters than necessary when p and n are similar in size. The penalty for this criterion is maximised when $p = n$ and decreases as $n > p$. Hence for small n, the pAIC tends to select smaller subsets of predictors than the AIC, but it performs in a similar way to the AIC for large n.

4. *Modified Akaike information criterion (mAIC)*: The mAIC,

$$-2\ln \hat{L} + 2p^2, \tag{4.11}$$

was proposed by Pan (2001) as a way of selecting models for correlated responses, however, it can be used for models for uncorrelated observations. The penalty for mAIC, $2p^2$, is more severe than the penalty for BIC, $p\ln(n)$, and pAIC, $2pn/(n-p-1)$ when n is small. The mAIC may therefore under-fit the models, that is, select models with fewer active terms than the true model.

The model which minimises AIC, BIC, pAIC or mAIC is preferred. When analysing the simulated data using all subsets regression, we compare the results for BIC, pAIC and mAIC to assess the sensitivity of results to choice of penalty.

Results for Simulated Responses

We performed all subsets regression on the simulated responses for Stage 1 and Stage 2 using the BIC (4.9), pAIC (4.10) and mAIC (4.11). The penalty for overfitting in mAIC is the most severe, and the penalty for pAIC is more severe than the penalty for BIC. Therefore, we may expect there to be more overfitting when all subsets regression with BIC is used when compared to the other two criteria. We considered two maximum numbers of parameters in this method, 10 and 12, both of which are larger than the number of terms in the true model but still allow some degrees of freedom for estimation of the variance components.

Recall from the discussion of Figure 3.7b in Chapter 3 that the model matrix for the model fitted to the Stage 1 responses, which is referred to as Model 1 in Section 3.6.1, does not have any correlated columns. The model matrix for the cumulative model fitted to the Stage 2 responses, which is referred to as Model 3 in Section 3.6.1, does have correlated columns. Hence, it is more complicated to select the correct model for the Stage 2 responses than the Stage 1 responses.

The all subsets regression results for the simulated responses from Stage 1 and 2 are given in Tables 4.1 and 4.2, respectively. Firstly, we note that the model found in Step 1(b) of all subsets regression when $p = 6$ and $p = 3$ for the simulated responses from Stage 1 and Stage 2, respectively, is the correct model. Hence, all subsets regression identifies the correct model when p is known. However, as p is unknown prior to experimentation, we need to assess the final model found using all subsets regression.

Criterion	p^*	Correct model when $p = 6$?	Correct Terms in Final Model	Additional Terms in Final Model
BIC	10	Yes	6 of 6	5
pAIC	10	Yes	6 of 6	3
mAIC	10	Yes	1 of 6	0
BIC	12	Yes	6 of 6	7
pAIC	12	Yes	6 of 6	3
mAIC	12	Yes	1 of 6	0

Table 4.1: Models selected using all subsets regression for various model selection criteria and maximum model size when (1.3) with $\boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{0}_{n_w}, 10\mathbf{I}_{n_w})$ is fitted to the simulated responses from Stage 1 of the optimal sixteen run two-stage split-plot experiment from Chapter 3.

Criterion	p^*	Correct model when $p = 3$?	Correct Terms in Final Model	Additional Terms in Final Model
BIC	10	Yes	2 of 3	9
pAIC	10	Yes	3 of 3	5
mAIC	10	Yes	3 of 3	1
BIC	12	Yes	3 of 3	10
pAIC	12	Yes	3 of 3	9
mAIC	12	Yes	3 of 3	1

Table 4.2: Models selected using all subsets regression for various model selection criteria and maximum model size when (1.3) with $\boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{0}_{n_w}, 10\mathbf{I}_{n_w})$ is fitted to the simulated responses from Stage 2 of the optimal sixteen run two-stage split-plot experiment from Chapter 3.

We note from Tables 4.1 and 4.2 that the number of additional terms in the final model, that is the model which minimises the criteria, decreases from BIC to pAIC to mAIC and increases from $p^* = 10$ to $p^* = 12$ for both stages. This shows the impact of having a more severe penalty. It also suggests that as we introduce more possible active terms into all subsets regression there is more over-fitting.

The contrast between the results for Stage 1 and Stage 2 for mAIC in Tables 4.1 and 4.2 is interesting. We note from these tables that all subsets regression fails to select the correct model for the response from Stage 1, but performs extremely well for the response from Stage 2. This was unexpected, as none the columns in the model matrix for Model 1 are correlated and some of the columns in the model matrix for Model 3 are correlated.

The penalty for the true model for Stage 1 is 72, whereas the penalty for the true model for Stage 2 is 18. This large difference in penalties could be the reason for these results. Also, there are fewer active effects in the true model for Stage 2 compared to the true model for Stage 1, but the same number of experimental runs for both models. A future area of research could be to assess the impact of the penalty function further.

Tables 4.3 and 4.4 give the maximum likelihood estimates of σ_γ^2 and σ_ϵ^2 for the best models of each size found in all subsets regression. These estimates are calculated using the function `lme` in the `lmer` package in `R`, which relies on REML and GLS estimation (as discussed in Section 1.3.2 of Chapter 1).

p	$\hat{\sigma}_\epsilon^2$	$\hat{\sigma}_\gamma^2$
1	112.69	7.8942×10^{-8}
2	71.377	5.1745×10^{-8}
3	24.003	45.146
4	0.51354	51.019
5	0.51354	1.4035
6	**0.27559**	**1.4630**
7	0.10814	1.5049
8	0.055581	1.5180
9	0.081108	2.2018×10^{-12}
10	0.041686	2.1471×10^{-12}
11	0.026367	2.5714×10^{-12}
12	1.4741×10^{-30}	4.3389

Table 4.3: Estimates of the variance components for the models of size p which minimise BIC, pAIC and mAIC for the simulated Stage 1 response from the optimal 16-run two-stage split-plot experiment from Chapter 3. The estimates of the variance components for the true model are highlighted in bold.

p	$\hat{\sigma}^2_\epsilon$	$\hat{\sigma}^2_\gamma$
1	23.780	2.7516
2	1.6750	8.2780
3	**0.77283**	**8.5035**
4	0.11080	8.6690
5	0.021064	8.6915
6	0.010751	8.6940
7	0.0047456	8.6955
8	0.015798	2.9142×10^{-12}
9	0.00046304	8.5530
10	0.00012549	8.5531
11	4.9809×10^{-6}	8.5258
12	1.1405×10^{-30}	1.5536

Table 4.4: Estimates of the variance components for the models of size p which minimise BIC, pAIC and mAIC for simulated Stage 2 response from the optimal 16-run two-stage split-plot experiment from Chapter 3. The estimates of the variance components for the true model are highlighted in bold.

We notice that a number of the estimates for σ^2_ϵ in Tables 4.3 and 4.4 are close to zero and not approximately equal to one, which is the fixed value of σ^2_ϵ which was used when generating the data. We also note that some of the estimates for σ^2_γ are very small. The results in Tables 4.3 and 4.4 suggest that the functions in R cannot find accurate estimates the variance(s) due to the lack of degrees of freedom in the model, and supports the findings of Gilmour and Goos (2009).

We do note, however that the estimates of the variance components for the true model for the simulated Stage 2 response, which are highlighted in bold in Table 4.4, are not too different to true values of $\sigma^2_\epsilon = 1$ and $\sigma^2_\gamma = 10$. These estimates are closer to the true values than those for the simulated Stage 1 response, which are highlighted in bold in Table 4.3. This result may be because the true model for Stage 2 has 13 degrees of freedom with which to estimate the variance components, whereas the true model for Stage 1 has 10 degrees of freedom.

4.3.3 Bayesian Variable Selection via Global and Local Search Algorithm

Global and Local Search Algorithm

Tan and Wu (2013) presented two Bayesian approaches for variable selection for split-plot experiments: the forward selection algorithm, which can be used when the columns of the model matrix considered are not correlated, and the global and local search algorithm, which can be used for designs when the columns of the model matrix considered are either correlated or uncorrelated.

In this chapter we focus on the global and local search algorithm, as it can be used to identify the models for responses from both unsaturated and supersaturated split-plot designs. The global and local search algorithm performs a global search on a diverse set of starting points, followed by a local search on models with posterior probability of being the true model above a certain threshold. This algorithm identifies a final model with high posterior probability of being the true model.

Tan and Wu (2013) extended the conjugate hierarchical model for SSVS given by George and McCulloch (1997) to account for the hierarchical error structure. The linear mixed model fitted to data from a split-plot experiment with $n = n_w n_s$ responses, n_w whole plots and n_s subplots presented by Tan and Wu (2013) is

$$\mathbf{y} \sim N(\beta_0 \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}, \sigma^2 \mathbf{V}(\psi)) \tag{4.12}$$

where $\mathbf{y}$ is the $n \times 1$ vector of random responses, β_0 is the intercept, $\mathbf{1}$ is a $n \times 1$ vector of ones, $\mathbf{X}$ is the $n \times p$ design matrix, $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed effect parameters, $\sigma^2 = \sigma^2_\gamma + \sigma^2_\epsilon$ is the sum of the whole-plot and subplot variance components and

$$\mathbf{V}(\psi) = \mathbf{I}_{n_w} \otimes \begin{pmatrix} 1 & \psi & \dots & \psi \\ \psi & 1 & \dots & \psi \\ \vdots & \vdots & \ddots & \vdots \\ \psi & \psi & \dots & 1 \end{pmatrix}, \tag{4.13}$$

where $\psi = \sigma^2_\gamma / \sigma^2$.

As the focus of the algorithm is variable selection, Tan and Wu (2013) follow the work of George and McCulloch (1993) and introduce an indicator vector $\boldsymbol{\delta} = (\delta_1, \dots, \delta_p)$, where

$$\delta_j = \begin{cases} 1 & \text{if } \beta_j \neq 0 \\ 0 & \text{if } \beta_j = 0 \end{cases}, j = 1 \dots, p, \tag{4.14}$$

.

When $\delta_j = 1$ the term is assumed to be active and is included in the model and when $\delta_j = 0$ the term is assumed to be non-active and is not included in the model.

Following George and McCulloch (1997), Tan and Wu (2013) assumed that $\boldsymbol{\beta}|\sigma^2, \boldsymbol{\delta}, c \sim$ N$(\mathbf{0}_p, \sigma^2 \mathbf{S}_{\boldsymbol{\delta},c})$, where $\mathbf{S}_{\boldsymbol{\delta},c}$ is a diagonal matrix with jth diagonal element of $cI(\delta_j = 1) + dI(\delta_j = 0)$, $j = 1, \dots, p$. Note that c is given a prior distribution and d is assumed to be a fixed small non-negative number. Tan and Wu (2013) also followed the approach of George and McCulloch (1997) with their assumption that $p(\beta_0) \propto 1$, $p(\sigma^2)$ is $IG(\nu/2, \nu\lambda/2)$ and $p(\delta_j)$ is Bernoulli(ρ_a).

The objective of variable selection can be achieved by finding the indicator vector that has the highest posterior probability,

$$
\begin{aligned}
p(\boldsymbol{\delta}|\beta_0, \boldsymbol{\beta}, \sigma^2, c, \psi, \mathbf{y}) &= \int_0^\infty \int_{\mathcal{B}} \int_0^\infty \sum_c^C \int_0^\infty p(\beta_0, \boldsymbol{\beta}, \sigma^2, \boldsymbol{\delta}, c, \psi|\mathbf{y}) \mathrm{d}\beta_0 \mathrm{d}\boldsymbol{\beta} \mathrm{d}\sigma^2 \mathrm{d}c \mathrm{d}\psi \\
&= \sum_c^C \int_0^\infty p(\boldsymbol{\delta}, c, \psi|\mathbf{y}) \mathrm{d}c \mathrm{d}\psi, \qquad (4.15)
\end{aligned}
$$

where $\beta_0, \boldsymbol{\beta}, \sigma^2$ can be integrated out analytically because of the choice of prior used, $C = \{\frac{1}{4}, \frac{9}{1}, 1, 4, 9, 16, 25\}$ and $\mathcal{B}$ is set of all possible $\boldsymbol{\beta}$ vectors.

Using the calculations from Tan and Wu (2013), we note that

$$
p(\boldsymbol{\delta}, c, \psi|\mathbf{y}) \propto [\nu\lambda + RSS_{\boldsymbol{\delta}}]^{-\frac{n-1+\nu}{2}} |\mathbf{X}_{\boldsymbol{\delta}} \mathbf{S}_{\delta,c} \mathbf{X}_{\boldsymbol{\delta}}^T + \mathbf{V}(\psi)|^{-\frac{1}{2}} (\mathbf{1}^T \mathbf{V}(\psi)^{-1} \mathbf{1})^{-\frac{1}{2}} p(\boldsymbol{\delta}) p(\psi) p(c), \qquad (4.16)
$$

where $RSS_{\boldsymbol{\delta}} = (\mathbf{y} - \bar{\mathbf{y}})^T (\mathbf{V}(\psi) + \mathbf{X}_{\boldsymbol{\delta}} \mathbf{S}_{\boldsymbol{\delta},c} \mathbf{X}_{\boldsymbol{\delta}}^T)^{-1} (\mathbf{y} - \bar{\mathbf{y}})$, $\bar{\mathbf{y}} = \mathbf{1}\bar{y}$, $\bar{y} = \sum_{i=1}^n y_i/n$, $\mathbf{X}_{\boldsymbol{\delta}}$ is the model matrix for the model including the terms which are indicated as active by $\boldsymbol{\delta}$, $\mathbf{1}^T \mathbf{V}(\psi)^{-1} \mathbf{1} = n/[1 + \{(n_s - 1)\psi\}]$. The prior distribution for $\boldsymbol{\delta}$ is

$$
p(\boldsymbol{\delta}) = \rho_a^{\sum_{j=1}^p \delta_j} (1 - \rho_a)^{p - \sum_{j=1}^p \delta_j}
$$

,

where ρ_a is the prior probability that jth term is active. The prior distribution for ψ is $p(\psi) \sim \beta(\nu, \lambda)$, and the prior distribution for c is

$$
p(c) = \begin{cases} \frac{1}{7} & c \in C \\ 0 & \text{otherwise} \end{cases} .
$$

As we assume that ψ has a beta prior and c has a uniform prior, we can use Gauss-Jacobi quadrature (see Appendix A for further detail) to approximate (4.15) as

$$
\begin{aligned}
&\sum_c^C \sum_g^{n_a} \frac{w(\psi)_g}{7} [\nu\lambda + RSS_{\boldsymbol{\delta}}]^{-\frac{n-1+\nu}{2}} \left| \mathbf{X}_{\boldsymbol{\delta}} \mathbf{S}_{\delta,c} \mathbf{X}_{\boldsymbol{\delta}}^T + \mathbf{V}\left(\frac{a(\psi)_g + 1}{2}\right) \right|^{-\frac{1}{2}} \\
&\times \left(\mathbf{1}^T \mathbf{V} \left(\frac{a(\psi)_g + 1}{2}\right)^{-1} \mathbf{1} \right)^{-\frac{1}{2}} \rho_a^{\sum_{j=1}^p \delta_j} (1 - \rho_a)^{p - \sum_{j=1}^p \delta_j}, \qquad (4.17)
\end{aligned}
$$

where $a(\psi)_g$ and $w(\psi)_g$ are the $g = 1, \ldots, n_a$ abscissa and corresponding weights for the Gauss-Jacobi quadrature.

As the normalisation constant for (4.17) is unknown, it cannot be directly computed. To avoid this issue, Tan and Wu (2013) use the log-posterior odds ratio

$$o(\boldsymbol{\delta}) = \log\left(\frac{p(\boldsymbol{\delta}|\mathbf{y})}{p(\mathbf{0}_p|\mathbf{y})}\right) = \log(p(\boldsymbol{\delta}|\mathbf{y})) - \log(p(\mathbf{0}_p|\mathbf{y})), \tag{4.18}$$

as maximising this ratio is equivalent to maximising $p(\boldsymbol{\delta}|\mathbf{y})$, and does not require the normalising constant. Substituting (4.17) into (4.18) gives

$$\begin{aligned} o(\boldsymbol{\delta}) \quad \propto \quad & \log\left[\sum_{c}^{C}\sum_{g}^{n_a}\frac{w(\psi)_g}{7}[\nu\lambda + RSS_{\boldsymbol{\delta}}]^{-\frac{n-1+\nu}{2}}\left|\mathbf{X}_{\boldsymbol{\delta}}\mathbf{S}_{\delta,c}\mathbf{X}_{\boldsymbol{\delta}}^T + \mathbf{V}\left(\frac{a(\psi)_g+1}{2}\right)\right|^{-\frac{1}{2}}\right. \\ & \left.\times\left(\mathbf{1}^T\mathbf{V}\left(\frac{a(\psi)_g+1}{2}\right)^{-1}\mathbf{1}\right)^{-\frac{1}{2}}\rho_a^{\sum_{j=1}^{p}\delta_j}(1-\rho_a)^{p-\sum_{j=1}^{p}\delta_j}\right] \\ & -\log\left[\sum_{c}^{C}\sum_{g}^{n_a}\frac{w(\psi)_g}{7}[\nu\lambda + RSS_{\mathbf{0}_p}]^{-\frac{n-1+\nu}{2}}\left|\mathbf{V}\left(\frac{a(\psi)_g+1}{2}\right)\right|^{-\frac{1}{2}}\right. \\ & \left.\times\left(\mathbf{1}^T\mathbf{V}\left(\frac{a(\psi)_g+1}{2}\right)^{-1}\mathbf{1}\right)^{-\frac{1}{2}}(1-\rho_a)^p\right], \end{aligned} \tag{4.19}$$

where $RSS_{\mathbf{0}_p} = (\mathbf{y}-\bar{\mathbf{y}})^T(\mathbf{V}((a(\psi)_g+1)/2))^{-1}(\mathbf{y}-\bar{\mathbf{y}})$. We find n_a by evaluating (4.19) for a fixed $\boldsymbol{\delta}$ for a range of n_a values, and selecting the n_a where the difference between successive (4.19) values as n_a increases is small. We say that (4.19) stabilises for this n_a value.

The global and local search algorithm requires a starting set of indicator vectors, Δ_{start}, and Tan and Wu (2013) recommended using the rows of a maximin design with n_{start} rows and p_{max} column, where p_{max} is the largest number of terms that we wish to consider in the model for $\mathbf{y}$ and is therefore the dimension of $\boldsymbol{\delta}$. Tan and Wu (2013) recommended maximin designs as they found that the global and local search does not consistently give good models for randomly selected starting indicator vectors.

For a suitable its (such as $its = 1000$), we can use the following algorithm to find a maximin design with n_{start} rows (for further detail see Santner et al., 2003, Chapter 5):

1. For $q = 1, \ldots, its$:

   (a) Randomly sample $\mathbf{D}_q$ from $\mathcal{D}_{p_{max},2,n_{start}}$, where the ith row of $\mathbf{D}_q$ is $\boldsymbol{\delta}_{i,q} = (x_{q,i,1}, \ldots, x_{q,i,p_{max}})$, $i = 1, \ldots, n_{start}$ and $\delta_{q,i,j} \in \{0, 1\}$ $\forall j = 1, \ldots, p_{max}$.

   (b) Find $d_q^* = \min_{\forall i=1,\ldots,n_{start}, j=1,\ldots,p_{max}} d(\mathbf{D}_q)$, where $d(\mathbf{D}_q) = (x_{q,i_1,j} - x_{q,i_2,j})^2$ for $i_1 \neq i_2, i_1, i_2 \in \{1, \ldots, n_{start}\}$

2. Find $\mathbf{D}_{mm} = \arg\max_{\forall q=1,\ldots,its} d_q^*$, the maximin design.

The rows of $\mathbf{D}_{mm}$, $\boldsymbol{\delta}_i$, $i = 1, \ldots, n_{start}$, are used as the starting values for $\boldsymbol{\delta}$ in the global and local search algorithm.

Global Search

The global search swaps the levels of each element in the initial $\boldsymbol{\delta}$ to identify the indicator vector that maximises (4.19), which is the indicator vector for the highest posterior probability (HPP) model.

1. Let $\Delta^* = \emptyset$
2. For $i = 1, \ldots, n_{start}$.
    - (a) Set $\boldsymbol{\delta} = \mathbf{x}_i$ and $\zeta = -\infty$.
    - (b)
        - i. Obtain $\boldsymbol{\delta}^j$, $\forall j \in \{1, \ldots, p_{max}\}$ by switching the jth element of $\boldsymbol{\delta}$ from 0 to 1, or 1 to 0.
        - ii. Find $j^* = \arg\max_{\forall j} o(\boldsymbol{\delta}^j)$, where $o(\boldsymbol{\delta})$ is (4.19).
        - iii. If $o(\boldsymbol{\delta}^{j^*}) \leq \zeta$ stop and let $\boldsymbol{\delta}^* = \boldsymbol{\delta}$. Otherwise set $\boldsymbol{\delta} = \boldsymbol{\delta}^{j^*}$ and $\zeta = o(\boldsymbol{\delta}^{j^*})$ and repeat from i.
        - iv. Add $\boldsymbol{\delta}^*$ to Δ^*.
3. Find the indicator vector for the HPP model, $\boldsymbol{\delta}^{HPP} = \arg\max_{\forall \boldsymbol{\delta} \in \Delta^*} o(\boldsymbol{\delta})$.

Local Search

The local search performs swaps on the (HPP) model indicator vector and identifies the $\boldsymbol{\delta}$ which maximises the posterior density of $\boldsymbol{\delta}$.

1. For $t = 1, 2, 3, 4$
    - (a) Let $\boldsymbol{\delta} \in \Delta^{MACV}$ if and only if $\boldsymbol{\delta} \in \Delta^*$ and $o(\boldsymbol{\delta}) > MACV$, where $MACV = o(\boldsymbol{\delta}^{HPP}) - \ln(50t)$.
    - (b) Let $\mathcal{I} = \mathcal{J} = \Delta^{MACV}$ be ordered sets. If a new item is added to $\mathcal{I}$ or$\mathcal{J}$, it is the last element of the set.
    - (c) For $k = 1, \ldots, |\Delta^{MACV}|$
        - i. Let $\boldsymbol{\delta}$ be the kth element of Δ^{MACV}.
        - ii. For $j = 1, \ldots, p_{max}$
            - A. Obtain $\boldsymbol{\delta}^j$ by switching the jth element of $\boldsymbol{\delta}$ from 0 to 1, or 1 to 0.

B. If $\boldsymbol{\delta}^j \notin \mathcal{I}$ and $o(\boldsymbol{\delta}^j) \geq MACV$, add $\boldsymbol{\delta}^j$ to $\mathcal{I}$ and $\mathcal{J}$.

C. Remove the first element of $\mathcal{J}$. If $|\mathcal{J}| = 0$ or $|\mathcal{I}| \geq 10^4$ then stop and let $\Delta^{t,j} = \mathcal{I}$. Otherwise let $\boldsymbol{\delta}$ be the first element of $\mathcal{J}$ and repeat from A.

2. For t=1, 2, 3, 4

   (a) For $j = 1, \ldots, p$

      i. Let $\boldsymbol{\delta} \in \Delta^t$ if and only if $\boldsymbol{\delta} \in \Delta^{t,j}$ and $\boldsymbol{\delta} = \max q(\boldsymbol{\delta})$, where

$$q(\boldsymbol{\delta}) = \frac{\exp o(\boldsymbol{\delta})}{\sum_{\boldsymbol{\delta} \in \Delta^{t,j}} \exp o(\boldsymbol{\delta})}, \tag{4.20}$$

      is an estimate of the posterior density of $\boldsymbol{\delta}$.

   (b) Find $\boldsymbol{\delta}^t = \arg\max_{\forall \boldsymbol{\delta} \in \Delta^t} q(\boldsymbol{\delta})$.

3. If $\boldsymbol{\delta}^{t_1} = \boldsymbol{\delta}^{t_2}\ \forall t_1, t_2 \in \{1, 2, 3, 4\} \cap t_1 \neq t_2$ then the algorithm has worked correctly, and any $\boldsymbol{\delta}^t$, $t = 1, 2, 3, 4$, is the indicator vector for the final model. If $\boldsymbol{\delta}^{t_1} \neq \boldsymbol{\delta}^{t_2}$ for some $t_1 \neq t_2$ for some $t_1 \neq t_2$, $t_1, t_2 \in \{1, 2, 3, 4\}$, then the algorithm has not worked and should be re-run for a $\mathbf{D}_{mm}$ with larger n_{start}.

Results for Simulated Data

We used the global and local search algorithm from Tan and Wu (2013) to perform Bayesian variable selection for the simulated responses from Stage 1 and Stage 2 discussed in Section 4.3.1. The global and local search algorithm requires us to specify d, ν, λ, ρ_a, the number of abscissa for Gauss-Jacobi quadrature, n_a, the size of the maximin design n_{start}, the maximum model size p_{max}. For this simulated data, we found that $n_a = 7$ is the smallest value for which (4.19) stabilises, and set $p_{max} = 21$ as this is the total number of main effects and pairwise product terms for the columns of the design matrix for our experiment. We consider $n_{start} = 100, 200, 300$, where Tan and Wu (2013) suggested that $n_{start} = 100$ should be sufficient, and set $d = \nu = \lambda = 0$ and $\rho_a = 0.25$, as suggested by Tan and Wu (2013).

Two models are found in the global and local search algorithm; the highest posterior probability (HPP) model, which is found after the global search, and the final model which is found after a local searches around the HPP model. Both of these models for three different maximin starting designs are presented in Tables 4.5 and 4.6.

The results in Tables 4.5 and 4.6 are very promising, as the final model for Stage 1 is the correct model with no over-fitting when a 100, 200 and 300 run maximin design is used and the final model for Stage 2 when a 300 run maximin design is used is the correct model with no overfitting, which is better than the model found using all subsets regression and the mAIC criterion in Tables 4.1 to 4.2.

n_{start}	HPP Model		Final Model	
	Correct Terms	Additional Terms	Correct Terms	Additional Terms
100	6 of 6	3	6	0
200	5 of 6	2	6	0
300	6 of 6	1	6	0

Table 4.5: Model selection results for the global and local search algorithm when (1.3) with $\boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{0}_{n_w}, 10\mathbf{I}_{n_w})$ is fitted to simulated data for Stage 1 of the optimal 16-run two-stage split-plot experiment found in Chapter 3.

n_{start}	HPP Model		Final Model	
	Correct Terms	Additional Terms	Correct Terms	Additional Terms
100	3 of 3	7	3 of 3	3
200	3 of 3	5	3 of 3	1
300	3 of 3	3	3 of 3	0

Table 4.6: Model selection results for the global and local search algorithm when (1.3) with $\boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{0}_{n_w}, 10\mathbf{I}_{n_w})$ is fitted to simulated data for Stage 2 of the optimal 16-run two-stage split-plot experiment found in Chapter 3.

Notice the difference between the overfitting in the HPP model and the final model, and hence the importance of performing the local search as well as the global search. Also, note that the algorithm finds the correct final model for Stage 1 with a smaller maximin design than for Stage 2. This is expected, as the model matrix for Stage 1 has no correlated columns, whereas the model matrix for Stage 2 has some correlated columns, and therefore the correct model should be more difficult to find.

4.3.4 Conclusions

The simulation results in Sections 4.3.2 and 4.3.3 demonstrate some of the issues associated with frequentist variable selection for models fitted to responses from saturated and supersaturated split-plot experiments. Frequentist variable selection methods struggle to find the correct model as the model selection criterion rely on estimates of the variance components, which are difficult to estimate.

Gilmour and Goos (2009) advocated the use of Bayesian variable selection methods as the use of prior information can help overcome the problem of estimating the variance components. The results for our simulated data support this conclusion as the global and local search algorithm found the correct model with less overfitting than all subsets regression.

Therefore, in Section 4.4, we present a method of performing Bayesian variable selection using samples from the Metropolis Hastings within Gibbs sampling algorithm. We use this method as we can use the samples from this algorithm for variable selection, parameter estimation and prediction, whereas the global and local search can only be used for variable selection and parameter estimation would have to be performed separately. Also, as mentioned in Section 4.3.1, the responses from the experiment motivating this work are multivariate and the methodology presented in Tan and Wu (2013) is for univariate data.

We note that the main focus of the criterion used to find the design, (3.20) from Section 3.4.2 of Chapter 3, is the estimation of the fixed effects and not the variance components. The criterion introduced in the recent work by Mylona et al. (2014) considered both fixed effect and variance component estimation. In Chapter 6, we extend the criterion presented by Mylona et al. (2014) to supersaturated multi-stage designs, find an optimal design for this extended criterion using the coordinate exchange algorithm from Section 3.5.1 and assess whether using designs for this criterion improves the estimates of the variance components found and enables frequentist analysis methods to be used.

4.4 Bayesian Variable Selection for Multivariate Linear Mixed Models

Following the discussion in Section 4.3.4, we introduce the Metropolis-Hastings within Gibbs sampling algorithm (Section 4.4.4). We will use the samples from this algorithm to perform variable selection for responses analysed using linear mixed models for supersaturated split-plot experiments. The conditional distributions (Section 4.4.3) used in the algorithm rely on the multivariate linear mixed model (Section 4.4.1) and extensions of the prior distributions from Tan and Wu (2013) (Section 4.4.2). Our choice of prior distribution for the fixed effect parameters also requires the joint sampling approach from Geweke (1996) to be extended to multivariate responses from linear mixed models.

The extensions of the prior distributions presented in Section 4.4.2 could also be used to extend the global and local search algorithm presented in Section 4.3.3 to multivariate responses from linear mixed models. However we chose to use a Metropolis-Hastings within Gibbs sampling algorithm as it allows us to perform variable selection, estimation and prediction, whereas additional calculations are required to perform estimation and prediction using the results from the global and local search algorithm.

4.4.1 Linear Mixed Model for Multivariate Responses

The extension of (1.3) in Section 1.3.1 of Chapter 1, the linear mixed effects model for a $n \times 1$ vector of responses from a split-plot experiment with n_w whole-plots and n_s sub-plots per whole-plot, to a split plot experiment with a $(n_w n_s) \times r$ response matrix of $\mathbf{Y}$ is

$$\mathbf{Y} = \mathbf{1}_n \boldsymbol{\beta}_0^T + \mathbf{X}\mathbf{B} + \mathbf{Z}\boldsymbol{\Gamma} + \mathbf{E}, \tag{4.21}$$

where $\mathbf{1}_n$ is the $n \times 1$ vector of ones, $n = n_w n_s$, $\boldsymbol{\beta}_0^T = (\beta_{01}, \beta_{02}, \ldots, \beta_{0r})$ is the $1 \times r$ vector of intercepts, $\mathbf{X}$ is the $n \times p$ model matrix with each column corresponding to a main effect or interaction parameter,

$$\mathbf{B} = \begin{pmatrix} \beta_{11} & \beta_{12} & \ldots & \beta_{1r} \\ \beta_{21} & \beta_{22} & \ldots & \beta_{2r} \\ \vdots & \vdots & \ldots & \vdots \\ \beta_{p1} & \beta_{p2} & \ldots & \beta_{pr} \end{pmatrix}$$

is the $p \times r$ vector of fixed effect parameters, $\mathbf{Z}$ is the $n \times n_w$ indicator matrix with (i, j)th element equal to 1 if the ith run of the experiment is in whole-plot j, $\boldsymbol{\Gamma}$ is the $n_w \times r$ random effect matrix and $\mathbf{E}$ is the $n \times r$ random error matrix.

In this work we assume that $\boldsymbol{\Gamma} \sim \text{MN}(\mathbf{0}_{n_w r}, \phi\mathbf{I}_{n_w}, \boldsymbol{\Sigma})$ and $\mathbf{E} \sim \text{MN}(\mathbf{0}_{nr}, (1-\phi)\mathbf{I}_n, \boldsymbol{\Sigma})$ are independent and matrix normally distributed, where $\mathbf{0}_{n_w r}$ and $\mathbf{0}_{nr}$ are the $n_w \times r$ and $n \times r$ zero mean matrices, $\phi\mathbf{I}_{n_w}$ and $(1-\phi)\mathbf{I}_n$ are $n_w \times n_w$ and $n \times n$ between-row scale matrices, and $\boldsymbol{\Sigma}$ is a $r \times r$ between-column scale matrix. Note that $0 < \phi \leq 1$ controls the relative scale of $\boldsymbol{\Gamma}$ and $\mathbf{E}$. Also, the between column scale matrix is the same for both $\boldsymbol{\Gamma}$ and $\mathbf{E}$, hence we are assuming that the responses have similar impacts on both the random effect $\boldsymbol{\Gamma}$ and the random errors $\mathbf{E}$.

Using the results in Section D.1 of Appendix D, the following marginal distribution can be derived:

$$\mathbf{Y}|\boldsymbol{\beta}_0, \mathbf{B}, \boldsymbol{\Sigma}, \boldsymbol{\delta}, c, \phi \sim \text{MN}(\mathbf{1}_n\boldsymbol{\beta}_0^T + \mathbf{X}\mathbf{B}, \mathbf{V}(\phi), \boldsymbol{\Sigma}) \tag{4.22}$$

where

$$\mathbf{V}(\phi) = \mathbf{I}_{n_w} \otimes (\phi\mathbf{J}_{n_s} + (1-\phi)\mathbf{I}_{n_s}) \tag{4.23}$$

is the symmetric $n \times n$ scale matrix for the rows of $\mathbf{Y}$ and $\mathbf{J}_{n_s}$ is the $n_s \times n_s$ matrix of ones.

Let vec($\mathbf{Y}$) be the $nr \times 1$ vector of column-stacked entries of $\mathbf{Y}$. Then, using the results from Section D.1,

$$\mathrm{vec}(\mathbf{Y})|\boldsymbol{\beta}_0, \mathbf{B}, \boldsymbol{\Sigma}, \boldsymbol{\delta}, c, \phi \sim \mathrm{N}(\mathrm{vec}(\mathbf{1}_n\boldsymbol{\beta}_0^T + \mathbf{XB}), \boldsymbol{\Sigma} \otimes \mathbf{V}(\phi)). \tag{4.24}$$

We note that (4.22) when $r = 1$ is equivalent to $\mathbf{Y} \sim \mathrm{N}(\mathbf{1}_n\beta_0^T + \mathbf{XB}, \sigma^2\mathbf{V}(\psi))$ which is the distribution for the responses assumed by Tan and Wu (2013).

4.4.2 Prior Distributions

We now specify the prior distributions suitable for variable selection. We follow the prior distributions used by Tan and Wu (2013), extending them where necessary to the multivariate case. Following the assumptions given in Tan and Wu (2013), the prior distribution for the intercept is

$$p(\boldsymbol{\beta}_0) \propto 1. \tag{4.25}$$

The prior distribution for the elements of the indicator vector $\boldsymbol{\delta}$ is

$$p(\delta_j) = \begin{cases} \rho_a & \text{if } \delta_j = 1 \\ 1 - \rho_a & \text{if } \delta_j = 0 \end{cases}, \tag{4.26}$$

where ρ_a is the prior probability that the parameter $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jr})$, which is the jth row or $\mathbf{B}$, is non-zero (active). In this work, we use the same prior probability as Tan and Wu (2013), and set $\rho_a = 0.25$.

Note that for the motivating example in this chapter it is appropriate to assume that if the term β_{jR^*} is active for the R^*th response, then it is active for all r responses. Similarly, if β_{jR^*} is not active for the R^*th response, then it is not active for all r responses. It is non-trivial to extend this assumption, and assume that every β_{jR}, $j = 1, \ldots, p$, $R = 1, \ldots, r$ can be active or non-active independently, and this is an area of potential future work.

This definition of active requires the use of a "spike-and-slab" type prior distribution for the fixed effects, where $(\boldsymbol{\beta}_j|\boldsymbol{\Sigma}, c, \delta = 1) \sim \mathrm{N}(\mathbf{0}_r, c\boldsymbol{\Sigma})$ and $(\boldsymbol{\beta}_j|\boldsymbol{\Sigma}, c, \delta = 1) = \mathbf{0}_r$ when not active. Using the calculations in Appendix E, the prior density for $\boldsymbol{\beta}_j$ is

$$p(\boldsymbol{\beta}_j) \propto |c\boldsymbol{\Sigma}|^{-\frac{p_a}{2}} \exp\left(-\frac{1}{2}\mathrm{tr}\left((c\boldsymbol{\Sigma})^{-1}\mathbf{B}^T\mathbf{B}\right)\right). \tag{4.27}$$

where $p_a = \sum_j \delta_j$ is the number of active terms and δ_j is the jth element of the indicator vector $\boldsymbol{\delta}$. If $\boldsymbol{\beta}_j = \mathbf{0}_r$, $\delta_j = 0$, and if $\boldsymbol{\beta}_j \neq \mathbf{0}_r$, $\delta_j = 1$.

The prior distribution for the column covariance, as used by Overstall and Woods (2015), is

$$\mathbf{\Sigma} \sim \mathrm{IW}(\mathbf{0}_{rr}, -r + 1), \tag{4.28}$$

where $\mathrm{IW}(\mathbf{S}, d)$ is the inverse Wishart distribution with scale matrix $\mathbf{S}$ and degrees of freedom d, as discussed in Section D.2 of Appendix D.

The prior distribution for ϕ, which controls the relative scale of the two random terms in (4.21), is

$$\phi \sim \beta(a, b) \tag{4.29}$$

where $\beta(a, b)$ is a beta distribution with shape parameters $a, b \geq 0$. We consider two a, b values; $a = b = 2$ as given by Tan and Wu (2013), and $a = 11, b = 2$, where $\phi \sim \beta(2, 2)$ is referred to as Prior 1 and $\phi \sim \beta(11, 2)$ is referred to as Prior 2. The mode of a $\beta(a, b)$ distribution for ϕ is

$$\frac{a-1}{a+b-2}, \tag{4.30}$$

hence the mode for Prior 1 is 0.5 (equal variance-covariance matrices for $\mathrm{vec}(\mathbf{\Gamma})$ and $\mathrm{vec}(\mathbf{E})$) and the mode for Prior 2 is 0.91 (2dp, so the variance-covariance matrix for $\mathrm{vec}(\mathbf{\Gamma})$ equals 11 times the variance-covariance matrix for $\mathrm{vec}(\mathbf{E})$).

Finally, the prior distribution for c is

$$p(c) = \begin{cases} \frac{1}{7} & \text{if } c \in \{\frac{1}{4}, \frac{9}{16}, 1, 4, 9, 16, 25\} \\ 0 & \text{otherwise} \end{cases}. \tag{4.31}$$

For a large fixed c the elements of $c\mathbf{\Sigma}$ are large relative to the non-zero elements of $\mathbf{B}$, hence small active terms will be missed and the amount of active terms would be small. Similarly, for a small fixed c the elements of $c\mathbf{\Sigma}$ are small relative to non-zero elements of $\mathbf{B}$, hence there would be a large number of active terms. Therefore, the support for c in (4.31) was recommended by Tan and Wu (2013), as it covers both small ($c = (1/2)^2, (3/4)^2, 1$) and large ($c = 2^2, 3^2, 4^2, 5^2$) values of c and enables results from the MCMC algorithm to be averaged over models with a small and large number of active terms. Note that when $r = 1$, (4.25) to (4.31) are equivalent to the prior distributions presented by Tan and Wu (2013).

4.4.3 Full Conditional Distributions

We use the prior distributions presented in Section 4.4.2 to derive the full conditional distributions, from which we will sample in the Metropolis-Hastings within Gibbs sampling algorithm discussed in Section 4.4.4. The full conditional distributions for $\boldsymbol{\beta}_0, \boldsymbol{\Sigma}$ and c, as calculated in Appendix E, can be sampled from directly. The conditional distributions for these parameters are as follows:

$$\boldsymbol{\beta}_0|\mathbf{Y}, \mathbf{B}, \boldsymbol{\Sigma}, \boldsymbol{\delta}, c, \phi \sim \mathrm{N}(\hat{\boldsymbol{\beta}}_0, (\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n)^{-1}\boldsymbol{\Sigma}), \tag{4.32}$$

where $\hat{\boldsymbol{\beta}}_0 = (\mathbf{Y} - \mathbf{XB})^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n(\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n)^{-1}$,

$$\boldsymbol{\Sigma}|\mathbf{Y}, \boldsymbol{\beta}_0, \mathbf{B}, \boldsymbol{\delta}, c, \phi \sim \mathrm{IW}(\mathbf{S}^*, -m + 1 + p_a + n), \tag{4.33}$$

where $\mathbf{S}^* = \{(\mathbf{Y} - \mathbf{1}_n\boldsymbol{\beta}_0^T - \mathbf{XB})^T\mathbf{V}(\phi)^{-1}(\mathbf{Y} - \mathbf{1}_n\boldsymbol{\beta}_0^T - \mathbf{XB})\} + \{(\mathbf{B}^T\mathbf{B})/c\}$, and

$$p(c|\mathbf{Y}, \boldsymbol{\beta}_0, \mathbf{B}, \boldsymbol{\Sigma}, \boldsymbol{\delta}, \phi) = \begin{cases} \frac{1}{c^*}c^{-\frac{p_a}{2}} \exp\left\{-\frac{1}{2c}\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{B}^T\mathbf{B}\right)\right\} & \text{if } c \in C \\ 0 & \text{otherwise} \end{cases}, \tag{4.34}$$

where $c^* = \sum_{c \in C} c^{\frac{p_a}{2}} \exp\left\{-\frac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{B}^T\mathbf{B})\right\}$ and $C = \{\frac{1}{4}, \frac{9}{16}, 1, 4, 9, 16, 25\}$.

The full conditional distribution for ϕ is

$$p(\phi|\mathbf{Y}, \boldsymbol{\beta}_0, \mathbf{B}, \boldsymbol{\Sigma}, \boldsymbol{\delta}, c) \propto |\mathbf{V}(\phi)|^{-\frac{m}{2}}\phi^{a-1}(1-\phi)^{b-1}$$

$$\times \exp\left[-\frac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{1}\boldsymbol{\beta}_0^T - \mathbf{XB})^T\mathbf{V}(\phi)^{-1}(\mathbf{Y} - \mathbf{1}_n\boldsymbol{\beta}_0^T - \mathbf{XB})\right\}\right], \tag{4.35}$$

which is a non-standard distribution that cannot be sampled from directly. Therefore, Metropolis-Hastings sampling (Section 4.2.2) is used to sample from distribution (4.35). In this work, we use Prior 1, $\beta(2,2)$, and Prior 2, $\beta(11,2)$, as proposal distributions.

An MCMC chain for $p(\boldsymbol{\beta}_j|\mathbf{Y}, \boldsymbol{\beta}_0, \boldsymbol{\Sigma}, \boldsymbol{\delta}, c, \phi)$ is reducible and does not converge to the required stationery distribution, hence we cannot sample from the conditional distribution for $\boldsymbol{\beta}_j$ (George and McCulloch, 1997). This problem is created by the use of the spike-and-slab prior, and can be solved by either extending the approach of Geweke (1996), and sampling from the joint conditional distribution of δ_j and $\boldsymbol{\beta}_j$, or using a different prior distribution such as the mixture of normal priors suggested by Box and Meyer (1986) and applied by Gilmour and Goos (2009).

We use a spike-and-slab prior distribution in this work, as the mixture of normal prior distribution require the relative weight of the variance component of the distribution for active and non-active effects to be known or estimated, which would be difficult to elicit for our motivating example. Also, spike-and-slab prior distributions give clearer conclusions about whether an effect is active or not, which is important for our motivating example.

Therefore, we extend the joint sampling approach of Geweke (1996) and sample from the joint conditional distribution of $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jr})$ and δ_j, $j = 1, \ldots, p$. Note that, as discussed before, we assume that the term is active for all r responses, so all elements of $\boldsymbol{\beta}_j$ are non-zero if jth term is active, and that $\boldsymbol{\beta}_j^T = \mathbf{0}_r$ if the j term is not active.

As shown in Appendix F, the joint sampling of $\boldsymbol{\beta}_j$ and δ_j uses the conditional posterior probability that $\boldsymbol{\beta}_j = \mathbf{0}$,

$$\rho_j = \frac{1 - \rho_a}{(1 - \rho_a) + (\rho_a BF_j)}, \tag{4.36}$$

where ρ_a is the prior probability that parameter j is active (which is assumed to be constant for all j in this work, but relaxing this assumption could be an area for future work) and

$$BF_j = \frac{p(\boldsymbol{\beta}_j | \mathbf{Y}, \boldsymbol{\beta}_0, \boldsymbol{\Sigma}, \delta_j = 1, c, \phi)}{p(\boldsymbol{\beta}_j | \mathbf{Y}, \boldsymbol{\beta}_0, \boldsymbol{\Sigma}, \delta_j = 0, c, \phi)} = \left(\frac{|\boldsymbol{\Sigma}_*|}{|c\boldsymbol{\Sigma}|} \right)^{\frac{1}{2}} \exp \left(\frac{1}{2} \bar{\boldsymbol{\beta}}_j \boldsymbol{\Sigma}_*^{-1} \bar{\boldsymbol{\beta}}_j^T \right). \tag{4.37}$$

To jointly sample $\boldsymbol{\beta}_j$ and δ_j:

1. Sample u from $U(0, 1)$.
2. If $\rho_j > u$ then $\boldsymbol{\beta}_j^T = \mathbf{0}_r$ and $\delta_j = 0$. If $\rho_q \leq u$ then $\delta_j = 1$ and $\boldsymbol{\beta}_j^T \sim \mathrm{N}(\bar{\boldsymbol{\beta}}_j^T, \boldsymbol{\Sigma}_*)$.

To calculate (4.37), and hence sample $\boldsymbol{\beta}_j$ and δ_j,

$$\bar{\boldsymbol{\beta}}_j = \mathbf{b}_j \boldsymbol{\omega}^{-1} \boldsymbol{\Sigma}_*,$$

$$\mathbf{b}_j = \frac{\sum_{k=1}^{n_w} \mathbf{X}_{kj}^T \mathbf{V}(\phi)_k^{-1} \mathbf{Y}_{kj}}{\sum_{k=1}^{n_w} \mathbf{X}_{kj}^T \mathbf{V}(\phi)_k^{-1} \mathbf{X}_{kj}},$$

$$\boldsymbol{\Sigma}_* = (\boldsymbol{\omega}^{-1} + (c\boldsymbol{\Sigma})^{-1})^{-1},$$

and

$$\boldsymbol{\omega} = \frac{\boldsymbol{\Sigma}}{\sum_{k=1}^{n_w} \mathbf{X}_{kj}^T \mathbf{V}(\phi)_k^{-1} \mathbf{X}_{kj}},$$

are required, where $\mathbf{X}_{kj}$ is the $n_s \times 1$ vector of entries in the model matrix $\mathbf{X}$ relating to whole plot $k = 1, \ldots, n_w$ and parameter j,

$$\mathbf{V}(\phi)_k = \phi \mathbf{J}_{n_s} + (1 - \phi)\mathbf{I}_{n_s}$$

is the between row scale matrix for whole plot k and

$$\mathbf{Y}_{kj} = \mathbf{Y}_k - \sum_{l \neq j} \mathbf{X}_{kl} \boldsymbol{\beta}_l,$$

for $l = 1, \ldots, p$. These calculations are shown in Appendix F.

Note that when $r = 1$, the conditional distributions given in this section can be used to perform variable selection via MCMC algorithms for univariate responses from a mixed model likelihood for split-plot designs.

4.4.4 Metropolis-Hastings within Gibbs Sampling Algorithm

Our Metropolis-Hastings within Gibbs algorithm produces dependent samples from the posterior distributions of $\boldsymbol{\beta}_0$, $\mathbf{B}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\delta}$, c and ϕ. A set of initial values for the parameters, $\boldsymbol{\beta}_0^{(0)}$, $\mathbf{B}^{(0)}$, $\boldsymbol{\Sigma}^{(0)}$, $\boldsymbol{\delta}^{(0)}$, $c^{(0)}$ and $\phi^{(0)}$ are required.

The steps at the qth iteration, $q = 1, \ldots, its$ of our Metropolis-Hastings within Gibbs sampling algorithm are:

1. Sample $\boldsymbol{\beta}_0^{(q)}$ from (4.32) with $\mathbf{Y}$, $\mathbf{B}^{(q-1)}$, $\boldsymbol{\Sigma}^{(q-1)}$ and $\phi^{(q-1)}$.

2. For $j = 1, \ldots, p$:

    (a) Calculate $\rho_j^{(q)}$, (4.36), using (4.37) for $\mathbf{Y}$, $\boldsymbol{\beta}_j^{(q-1)}$, $\boldsymbol{\Sigma}^{(q-1)}$, $c^{(q-1)}$ and $\phi^{(q-1)}$.

    (b) Sample $u_j^{(q)}$ from $U(0, 1)$.

    (c) If $\rho_j^{(q)} > u_j^{(q)}$, let $(\boldsymbol{\beta}_j^T)^{(q)} = \mathbf{0}_r$ and $\delta_j^{(q)} = 0$. Otherwise, sample $(\boldsymbol{\beta}_j^T)^{(q)}$ from $N((\bar{\boldsymbol{\beta}}_j^T)^{(q)}, \boldsymbol{\Sigma}_*^{(q)})$ using $\mathbf{Y}$, $\boldsymbol{\beta}_j^{(q-1)}$, $\boldsymbol{\Sigma}^{(q-1)}$, $c^{(q-1)}$ and $\phi^{(q-1)}$ and let $\delta_j^{(q)} = 1$.

3. Sample $\boldsymbol{\Sigma}^{(q)}$ from (4.33) with $\mathbf{Y}$, $\boldsymbol{\beta}_0^{(q)}$, $\mathbf{B}^{(q)}$, $c^{(q-1)}$ and $\phi^{(q-1)}$.

4. (a) Sample $\phi_*^{(q)}$ from $\beta(a, b)$.

    (b) Calculate $\alpha^{(q)} = \min\left\{1, \frac{p\left(\mathbf{Y}|\boldsymbol{\beta}_0^{(q)}, \mathbf{B}^{(q)}, \boldsymbol{\Sigma}^{(q)}, \boldsymbol{\delta}^{(q)}, c^{(q-1)}, \phi_*^{(q)}\right)}{p\left(\mathbf{Y}|\boldsymbol{\beta}_0^{(q)}, \mathbf{B}^{(q)}, \boldsymbol{\Sigma}^{(q)}, \boldsymbol{\delta}^{(q)}, c^{(q-1)}, \phi^{(q-1)}\right)}\right\}$.

    (c) Sample $u^{(q)}$ from $U(0, 1)$.

    (d) If $\alpha^{(q)} > u^{(q)}$ then set $\phi^{(q)} = \phi_*^{(q)}$, otherwise set $\phi^{(q)} = \phi^{(q-1)}$.

5. Sample $c^{(q)}$ from the set C with probability mass function given by (4.34) with $c^{(q-1)}$, $\mathbf{\Sigma}^{(q)}$ and $\mathbf{B}^{(q)}$.

In Section 4.5 we use $its = 10,000$, $\boldsymbol{\beta}_0^{(0)} = \mathbf{0}_r$, $\mathbf{B}^{(0)} = \mathbf{0}_{pr}$, $\mathbf{\Sigma}^{(0)} = \mathbf{I}_r$, $\boldsymbol{\delta}^{(0)} = \mathbf{1}_p$, $\phi^{(0)} = 0.2$ and $c^{(0)} = 0.25$.

4.5 Analysis of Simulated Data

In this Section, we use simulated data to assess the performance of the Metropolis-Hastings within Gibbs sampling algorithm presented in Section 4.4.4. The simulated data used is appropriate for the motivating example discussed in Section 4.1, as we consider multivariate data for both stages of the split-plot experiment designed in Chapter 3. We discuss how we generated the simulated data in Section 4.5.1. In Section 4.5.2, we assess whether the samples from the Metropolis-Hastings within Gibbs sampling algorithm select the correct active terms for the saturated model for the experiment for Factors 1 to 5 and for the supersaturated model for responses from the experiment for Factors 1 to 6.

4.5.1 Simulated Responses

In this chapter, we use the same models used in Section 4.3.1, hence the active terms:

- for the **Stage 1** model are $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3, \boldsymbol{\beta}_4, \boldsymbol{\beta}_{34}$ and $\boldsymbol{\beta}_{35}$, which relate to columns $\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3, \mathbf{f}_4, \mathbf{f}_3\mathbf{f}_4$ and $\mathbf{f}_3\mathbf{f}_5$ of $\mathbf{X}$, respectively. This model is for the response measured after Factors 1 to 5 are applied, and consists of both main effects and interactions.

  The model matrix for the model containing all the main effects and pairwise column combinations for Stage 1 has no correlated columns (see Figure 3.7b in Section 3.6.1), so $\mathbf{f}_1, \mathbf{f}_2, \mathbf{f}_3, \mathbf{f}_4, \mathbf{f}_3\mathbf{f}_4$ and $\mathbf{f}_3\mathbf{f}_5$ are not correlated with any other columns for Stage 1.

- for the **Stage 2** model are $\boldsymbol{\beta}_2, \boldsymbol{\beta}_4$ and $\boldsymbol{\beta}_6$, which relate to columns $\mathbf{f}_2, \mathbf{f}_4$, and $\mathbf{f}_6$ of $\mathbf{X}$, respectively. This model is for the response measured after Factors 1 to 6 are applied, and consists of main effects from both Stages of the experiment.

  The model matrix for the model containing all the main effects and pairwise column combinations for Stage 2 has correlated columns. Using Figure 3.7b in Section 3.6.1, we note that $\mathbf{f}_2 = -0.5\mathbf{f}_2\mathbf{f}_6 = 0.5\mathbf{f}_5\mathbf{f}_6$ and $\mathbf{f}_5 = -0.5\mathbf{f}_1\mathbf{f}_3 = -0.5\mathbf{f}_1\mathbf{f}_5 = -0.5\mathbf{f}_2\mathbf{f}_3 = 0.5\mathbf{f}_2\mathbf{f}_5$. The parameters related to these columns will be biased and have inflated variance, therefore the true model will be more difficult to identify for this stage.

Notice that the active parameters are now vectors, as we are considering multivariate responses with 18 runs and 2 correlated columns for each stage, which are appropriate for the motivating example discussed in Section 4.1 and Section 5.1 of Chapter 5. For this study, we use the 16 run two-stage optimal split-plot design from Section 3.6.1 of Chapter 3, with two additional centre points. Let

- $\mathbf{X}_s$, $s = 1, 2$, be the $18 \times p_{a,s}$ model matrix for the response from stage $s = 1, 2$, where $p_{a,s}$ is the number of active terms in the model for Stage s ($p_{a,1} = 6, p_{a,2} = 3$).
- $P_1 = \{1, \dots, 15\}$ and $P_2 = \{1, \dots, 21\}$ be the set of indexes for the parameters in the models (without intercept) for Stage 1 and Stage 2, respectively.
- $P_{1A} = \{1, 2, 3, 4, 13, 14\}$ and $P_{1N} = P_1/P_{1A}$ be the set of indexes for the active and non-active parameters in the models for stage 1, respectively.
- $P_{2A} = \{2, 4, 6\}$ and $P_{2N} = P_2/P_{2A}$ be the set of indexes for the active and non-active parameters in the models for Stage 2, respectively.
- $\boldsymbol{\mu}_1 = (\mu_{11}, \mu_{12})^T$ and $\boldsymbol{\mu}_2 = (\mu_{21}, \mu_{22})^T$ where μ_{11} and μ_{12} are set to -5 or 5 with probability 0.5 and μ_{21} and μ_{22} are set to -20 or 20 with probability 0.5.
- $\phi_1 = 1/2$ and $\phi_2 = 10/11$ be the modes of Prior 1 and Prior 2.

	$\mathbf{Y}_{1s}$	$\mathbf{Y}_{2s}$
Generated from:	$\text{MN}(\mathbf{XB}^*, \mathbf{V}(\phi_1), \boldsymbol{\Sigma}_1)$	$\text{MN}(\mathbf{XB}^*, \mathbf{V}(\phi_2), \boldsymbol{\Sigma}_2)$
$\boldsymbol{\beta}_j^*$ for $j \in P_{sA}$ generated from:	$\text{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$	$\text{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1)$
$\boldsymbol{\beta}_j^*$ for $j \in P_{sN}$ is:	$\mathbf{0}_r$	$\mathbf{0}_r$

	$\mathbf{Y}_{3s}$	$\mathbf{Y}_{4s}$
Generated from:	$\text{MN}(\mathbf{XB}^*, \mathbf{V}(\phi_1), \boldsymbol{\Sigma}_1)$	$\text{MN}(\mathbf{XB}^*, \mathbf{V}(\phi_2), \boldsymbol{\Sigma}_2)$
$\boldsymbol{\beta}_j^*$ for $j \in P_{sA}$ generated from:	$\text{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1)$	$\text{N}(\boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1)$
$\boldsymbol{\beta}_j^*$ for $j \in P_{sN}$ is:	$\mathbf{0}_r$	$\mathbf{0}_r$

Table 4.7: The four 18×2 multivariate responses generated for models from Stage s, $s = 1, 2$.

There are four combinations of $\boldsymbol{\mu}_1$, $\boldsymbol{\mu}_2$, ϕ_1 and ϕ_2, therefore there are four responses generated for each stage, and details of these four responses are given in Table 4.7, where

$$\boldsymbol{\Sigma}_1 = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix},$$

and

$$\mathbf{\Sigma}_2 = \begin{pmatrix} 11 & 9.9 \\ 9.9 & 11 \end{pmatrix}.$$

4.5.2 Analysis of Multivariate Simulated Responses from a Split-Plot Design

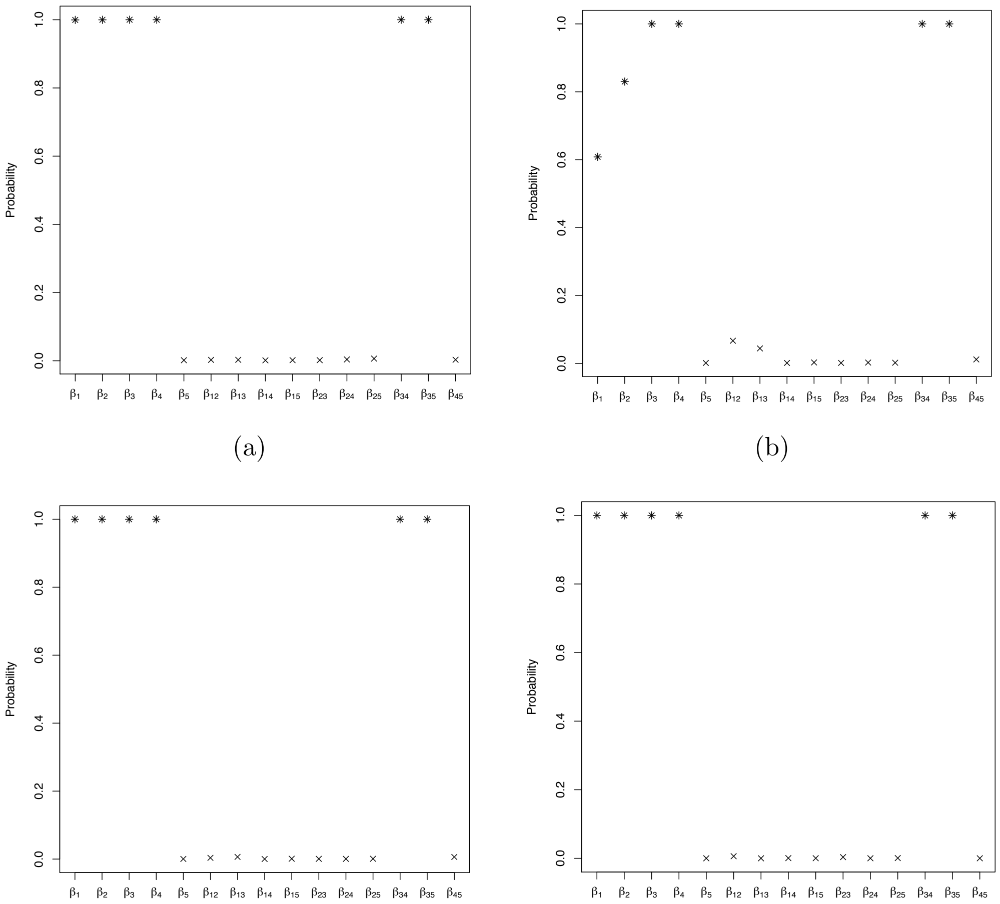
In this section, we will use the simulated data from Section 4.5.1 to assess the performance of the Metropolis-Hastings within Gibbs sampling algorithm as a method of variable selection and parameter estimation. Figures 4.1 and 4.2 give the approximate posterior probability of the terms in models for the simulated Stage 1 and 2 responses, respectively, being active for the four simulated responses. The posterior probability of parameter $\boldsymbol{\beta}_j$, $j = 1, \ldots, p$, being active is approximated by

$$\sum_{q=1}^{its} \frac{\delta_j^{(q)}}{its} \tag{4.38}$$

where $\delta_j^{(q)}$ is δ_j sampled at iteration q, $q = 1, \ldots, its$, of the Metropolis-Hastings within Gibbs sampling algorithm.

These figures show that, in general, the algorithm performs well as the parameters which were active when generating the responses have the highest posterior probability of being active. However, we note that when we use $\boldsymbol{\mu}_1$ and ϕ_2 (Figures 4.1(b) and 4.2(b)) the algorithm has difficultly correctly identifying the active terms because the active parameters, $\boldsymbol{\beta}_j$, $j \in P_{2A}$, are small compared to $\mathbf{\Sigma}$.

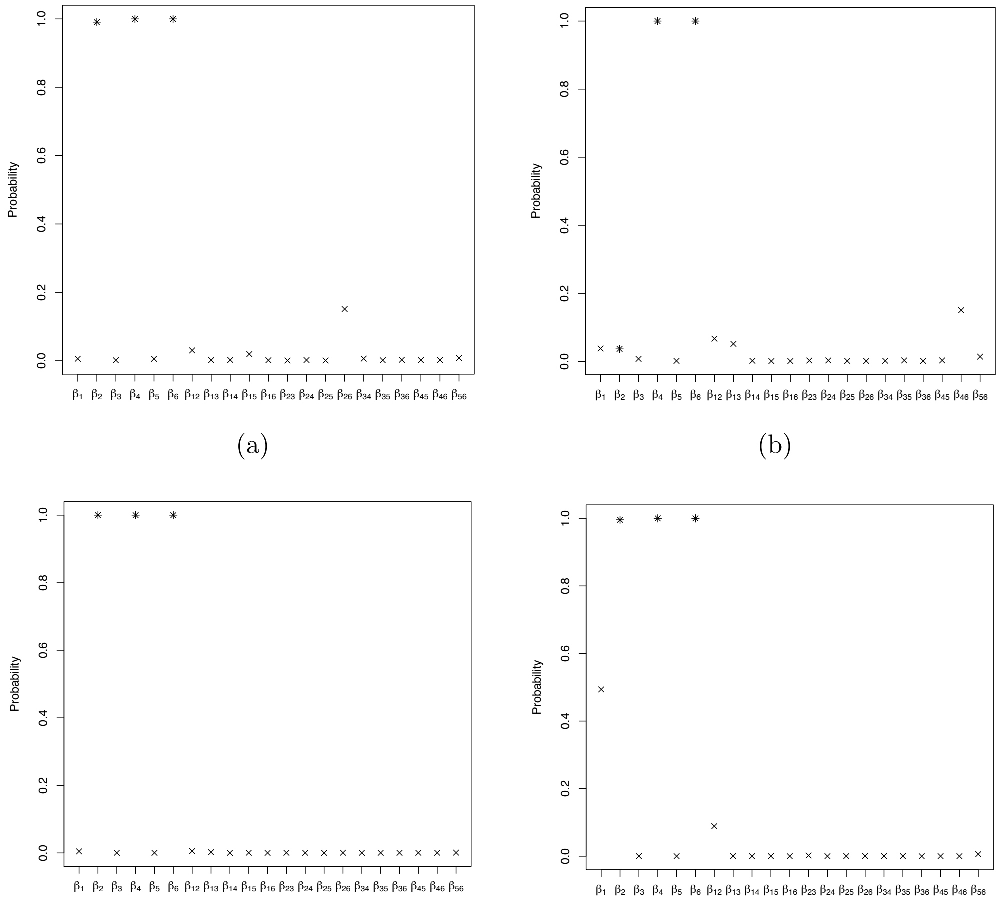
Through comparison of Figures 4.1(b) and 4.2(b), we also note that the active parameters still have high (greater than 0.8) approximated probabilities of being active for $\mathbf{Y}_{21}$ whereas one of the active parameters, $\boldsymbol{\beta}_2$, has a very small (less than 0.05) approximated probability of being active for $\mathbf{Y}_{22}$. This therefore shows not only the impact of large $\mathbf{\Sigma}$ in comparison to $\boldsymbol{\beta}_j$, $j \in P_{2A}$, but also the additional difficulties in identifying the active terms in the model for the response from the supersaturated experiment for Factors 1 to 6 when compared to identifying the active terms in the model for the response from the saturated experiment for Factors 1 to 5.



(a) (b)

(c) (d)

Figure 4.1: Approximate posterior probability of the terms in models fitted to (a) $\mathbf{Y}_{11}$, (b) $\mathbf{Y}_{21}$, (c) $\mathbf{Y}_{31}$ and (d) $\mathbf{Y}_{41}$ from Table 4.7 being active. Terms which are assumed to be active when generating the responses are denoted by $*$.



(a)

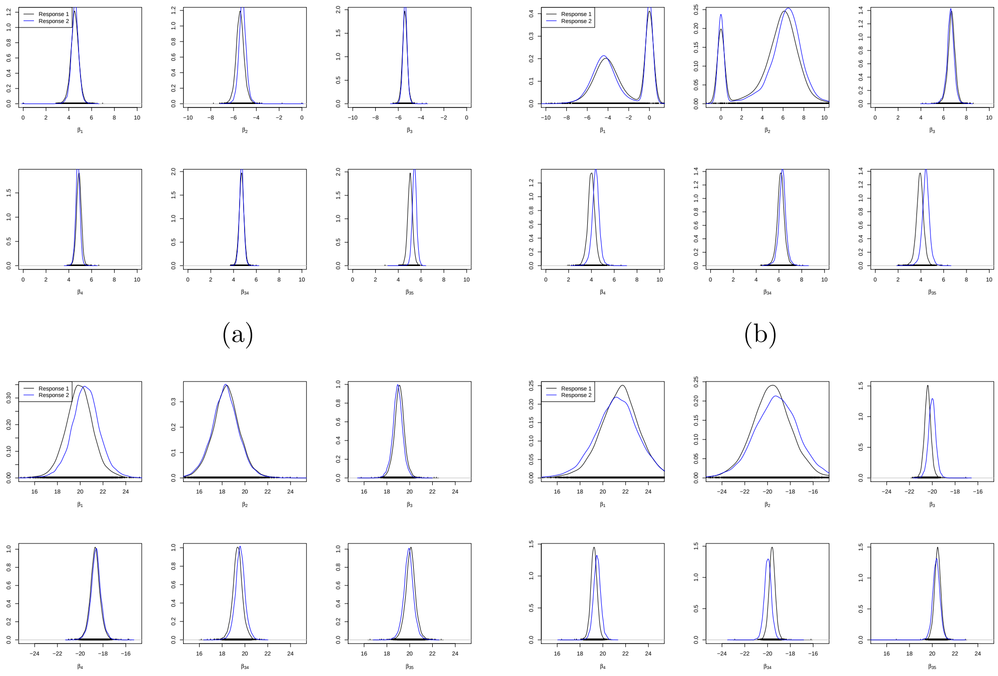
(b)

(c)

(d)

Figure 4.2: Approximate posterior probability of the terms in models fitted to (a) $\mathbf{Y}_{12}$, (b) $\mathbf{Y}_{22}$, (c) $\mathbf{Y}_{32}$ and (d) $\mathbf{Y}_{42}$ from Table 4.7 being active. Terms which are assumed to be active when generating the responses are denoted by $*$.

Figures 4.3 and 4.4 give the estimated marginal posterior densities for the active terms in the models used to simulate the responses from Stages 1 and 2, respectively. These marginal posterior densities, which are estimated using the samples of $\boldsymbol{\beta}_j$, $j \in P_{sA}$, $s = 1, 2$, from the sampling algorithm, all have substantial densities for non-zero values. When $\boldsymbol{\beta}_j$, $j \in P_{sA}$, $s = 1, 2$, is large compared to $\boldsymbol{\Sigma}$, the modes of these marginal densities are close to the true values, even with the more complex sampling required due to the spike-and-slab prior distribution.



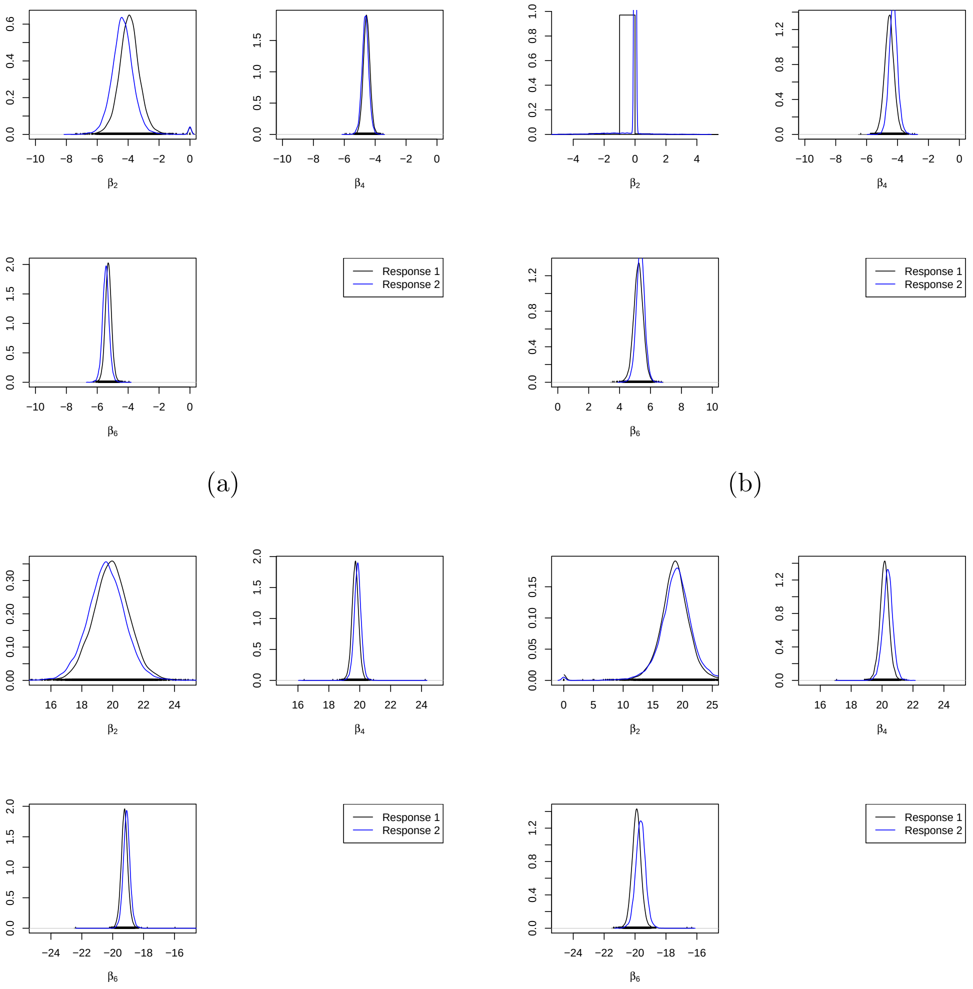
(a)

(b)

(c)

(d)

Figure 4.3: Marginal posterior density plots estimated from the samples of the terms in the model for the response which were active when simulating (a) $\mathbf{Y}_{11}$, (b) $\mathbf{Y}_{21}$, (c) $\mathbf{Y}_{31}$ and (d) $\mathbf{Y}_{41}$ in Table 4.7.



(a)

(b)

(c)

(d)

Figure 4.4: Marginal posterior density plots estimated from the samples of the terms in the model for the response which were active when simulating (a) $\mathbf{Y}_{12}$, (b) $\mathbf{Y}_{22}$, (c) $\mathbf{Y}_{32}$ and (d) $\mathbf{Y}_{42}$ in Table 4.7.

Comparing Figures 4.1 and 4.2 and Figures 4.3 and 4.4, respectively, shows the relationship between δ_j and $\boldsymbol{\beta}_j$, which are jointly sampled in the Metropolis-Hastings within Gibbs sampling algorithm. Notice that when the approximated probability of the parameter being active, (4.38), which relies on the samples of δ_j, is low, then there is a spike in the marginal posterior densities for $\boldsymbol{\beta}_j^T$ at $\mathbf{0}_2$, as $\delta_j^{(q)} = 0 \implies \left(\boldsymbol{\beta}_j^{(q)}\right)^T = \mathbf{0}_2$. This can be clearly seen when comparing the approximated probabilities of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ being active in Figure 4.1(b) and the estimated posterior densities for $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ in Figure 4.3(b). The estimated posterior density for $\boldsymbol{\beta}_2$ in Figure 4.4(b) is concentrated around zero, as over 95% of the sampled δ_2 are 0 and hence over 95% of the $\left(\boldsymbol{\beta}_2^{(q)}\right)^T = \mathbf{0}_2$.

The convergence properties of the samples from the Metropolis-Hastings within Gibbs sampling algorithm are assessed in Section C.2 of Appendix C using the trace and ACF plots discussed in Section 4.2.2 and Appendix C.1. We notice from this assessment that a number of the sampling chains have good convergence properties. However, there is some evidence to suggest that considering an alternative proposal distribution for ϕ would be beneficial.

4.6 Discussion

The focus of this chapter was Bayesian variable selection for multivariate responses from split-plot designs using samples from a Metropolis-Hastings within Gibbs sampling algorithm. In Section 4.3 we provided a comparison of a frequentist and Bayesian variable selection method. This provided evidence for our use of Bayesian variable selection, which is also supported by Gilmour and Goos (2009). However, instead of using the variable selection presented in Section 4.3.3, we used the Metropolis-Hastings within Gibbs sampling algorithm presented in Section 4.4.4, as samples from this algorithm can be used for parameter estimation as well as Bayesian variable selection.

The Metropolis-Hastings within Gibbs sampling algorithm given in Section 4.4.4 draws dependent samples from the full conditional distributions presented in Section 4.4.3, which rely on the linear mixed effects model for multivariate responses presented in Section 4.4.1, and the multivariate extensions of the prior distributions from Tan and Wu (2013) presented in 4.4.2. As we assumed that a spike-and-slab prior distribution is appropriate for the fixed effect terms in the linear mixed effects model, we jointly sampled the indicator vector and the fixed effect parameters within our sampler. This required the extension of the joint sampling approach of Geweke (1996) to multivariate responses from split-plot experiments (see Appendix F for further detail).

We then assessed the performance of the algorithm for simulated multivariate responses generated from a single model with fixed active parameters. We noted in Section 4.5.2 that there is an impact on the model selection and parameter estimation when the active terms are small relative to the column scale matrix. However, the posterior probabilities for the true active terms were still larger than the other terms for the majority of the simulated data. Also, the samples of the terms which were assumed to be active when generating the data could be used to estimate the correct distribution for these terms in the majority of cases.

The work in this chapter could be extended in a number of ways. A full simulation study, where simulated responses for a large number of randomly generated models, could be undertaken to further assess the performance of the Metropolis-Hastings within Gibbs sampling algorithm. Such an assessment would also enable us to gain insight into the performance of the Metropolis-Hastings within Gibbs sampling algorithm for a range of different models. For example, we could look at whether the performance is impacted when the number of terms in the model is close to the number or runs, or when there between stage interactions in the cumulative model for Stage 2.

We could see what impact the use of a mixture of normal prior distributions, as seen in the paper by Box and Meyer (1986) and Gilmour and Goos (2009), has on the effectiveness of the Metropolis-Hastings within Gibbs sampling algorithm. If mixture of normal prior distributions were used, algorithm in Section 4.4.4 could be modified and compared to an extended version of the SSVS algorithm by Brown et al. (1998).

We could relax the assumption that the random effect $\mathbf{\Gamma}$ and the random error $\mathbf{E}$ in (4.21) have the same between column scale matrix $\mathbf{\Sigma}$. For example, assume instead that $\mathbf{\Gamma} \sim \mathrm{MN}(\mathbf{0}_{n_w}, \phi\mathbf{I}_{n_w}, \mathbf{\Sigma}_1)$ and $\mathbf{E} \sim \mathrm{MN}(\mathbf{0}_n, (1-\phi)\mathbf{I}_n, \mathbf{\Sigma}_2)$. However, this assumption would mean that the variance of the vectorised response matrix, $\mathrm{vec}(\mathbf{Y})$, is equal to $\phi\mathbf{\Sigma}_1 \otimes \mathbf{Z}\mathbf{Z}^T + (1-\phi)\mathbf{\Sigma}_2 \otimes \mathbf{I}_n$. This variance of $\mathrm{vec}(\mathbf{Y})$ does not, in general, imply a matrix normal distribution for $\mathbf{Y}$ (see Appendix D.1 for more detail).

We could also relax our assumption that all r elements of $\boldsymbol{\beta}_j$ are active or non-active, as, whilst this assumption is suitable for the motivating example for this work, it may not be suitable for multivariate responses from all experiments. This would require the extension of the indicator vector to a matrix, and would require further, non-trivial, extension of the joint sampling approach in Section 4.4.3 and Appendix F.

Chapter 5

Industrial Case Study: Formulation and Dissolution Testing of a Pharmaceutical Product at GlaxoSmithKline

5.1 Introduction

In this Chapter, we apply the design and modelling methodology presented in Chapters 3 and 4 to an experiment to investigate formulation and dissolution testing of a pharmaceutical product performed by GlaxoSmithKline (GSK) (introduced in Section 1.2.2 in Chapter 1). This experiment has six controllable factors. The first two factors are hard-to-change, leading to restrictions on randomisation and the experiment being run as a split-plot.

The aim of this experiment is to study formulation of an active pharmaceutical ingredient (API) in a capsule. The API is a chemical compound that is used to treat a specific ailment or disease. The API is coated onto a core bead, and this bead is then coated with two further sustainable release layers. These beads are then placed in a capsule. The API is released once the capsule containing it has opened and the sustainable release layers have broken down. The rate of dissolution is measured by the amount of API dissolved over time.

The pharmaceutical product will be formulated in two stages, where the first stage has five controllable factors and the second stage has a single factor, see Figure 5.1. The five factors in Stage 1 are of more scientific interest than the single Stage 2 factor, Factor 6.

It is assumed that a measure of the quality of the API formed will be measured after Factors 1 to 5 are applied in the first stage of experimentation, and that the outcome of dissolution testing will be measured after Factor 6 is applied in the second stage of experimentation. The results from dissolution testing of the pharmaceutical product, are considered to be more important than the quality of the API.

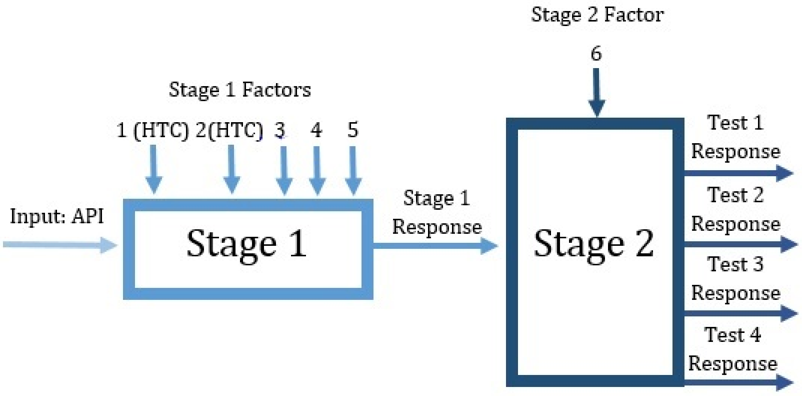


Figure 5.1: Schematic for the GSK experiment showing the two stages, six factors and four responses. Two of the Stage 1 factors are hard to change (HTC).

Three models will be fitted to these two responses; (i) the model for the first stage response with respect to the factors in the first stage, (ii) the model for the second stage response with respect to the factor in the second stage, and (iii) the cumulative model (as defined in Section 3.4.1) for the second stage response with respect to the factors in the first and second stage. This experiment therefore matches the definition of a multi-stage experiment given in Section 3.1.1 of Chapter 3.

The two-stage design for the formulation of this pharmaceutical product is found using the coordinate exchange algorithm given in Section 3.5.1 in Chapter 3, and is discussed in Section 5.2.2. There are resources for sixteen factorial points and three centre points. We use the compound Bayesian D-optimality objective function, (3.20) from Section 3.4 in the coordinate exchange algorithm. We use the same models and variance-covariance matrix structures for this design as given for the two-stage split-plot design in Table 3.6 in Chapter 3.

The dissolution of a pharmaceutical product in media is assessed via a dissolution test. The response for a dissolution test is a measure of the average difference between the dissolution of the pharmaceutical product and a reference product over time, referred to as the f_2 statistic:

$$f_2 = 50 \log \left[\frac{100}{t_d} \sum_{t=1}^{t_d} (D_t - D_t^*)^2 + 1 \right]^{-\frac{1}{2}}, \tag{5.1}$$

where D_t is the dissolution of the pharmaceutical product in the media at time t and D_t^* is the dissolution of the reference product in the media at time t, $t = 1, \dots, t_d$.

The f_2 is a value between 0 and 100, and a product is said to meet specification if

f_2 is greater than some threshold determined by regulators. If multiple references are considered, the f_2 values for different references from the same test will be correlated, as the same dissolution data is used to calculate the f_2 for each reference. Further detail regarding dissolution testing is given in Chow (2007, Chapter 11).

In this experiment, two reference profiles are considered for dissolution tests in four different media, and the f_2 statistic is calculated for each of these two references for each test for all 19 experimental runs. Therefore, this experiment has a bivariate response for each of the four tests, which are the four responses for Stage 2 given in Figure 5.1.

After initial analysis, which is discussed in Section 5.3.1, the posterior distribution of the multivariate (bivariate) response for each test are approximated using samples from the Metropolis-Hastings within Gibbs Sampling algorithm given in Section 4.4.4 of Chapter 4. In Section 5.3.2, we aim to identify the factors with a high posterior probability of being active. In Section 5.4.4, the parameters sampled using the Metropolis-Hastings within Gibbs Sampling algorithm are used in a grid search (Section 5.4.2) and with the efficient global optimisation (EGO, Section 5.4.3) algorithm (Jones et al., 1998) to find new treatments with a high posterior probability of passing specification for single and multiple dissolution tests.

5.2 Design of the Experiment

5.2.1 Robustness of the Compound Bayesian D-optimal Two-stage Split-plot Designs to the Model Weights

When designing the experiments in Chapter 3, we assumed that the vector of model weights, $\mathbf{w} = (w_1, w_2, w_3)$ in (3.20) in Section 3.5.1, was (0.7, 0.1, 0.2). This assumption was made using the information available at the time, and it produced a correlation structure that was suitable for this experiment. Only the column in the model matrix for Factor 6 and pairwise products of columns involving Factor 6, which is not of primary interest in the formulation but needs to be included in the experiment, were correlated.

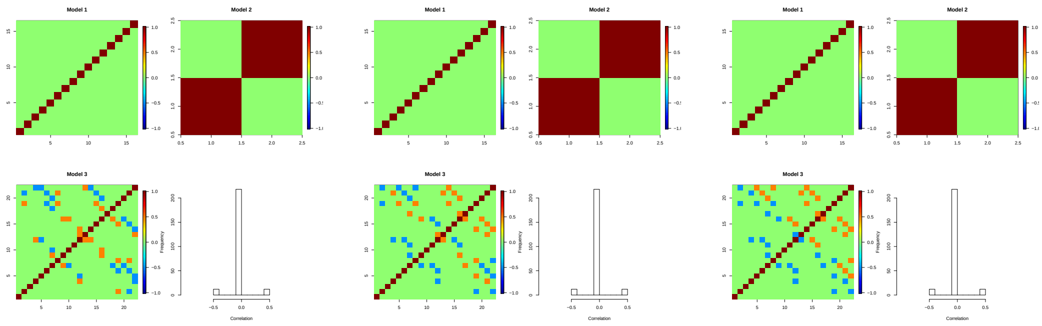
However, after the experiment was run, we found out that the result of the dissolution testing, which is the performed after Stage 2, is of greater importance than the measure of quality taken after Stage 1. Therefore, it seems appropriate to place more weight on the cumulative model for the dissolution testing, or Stage 2, response. Hence, we wish to assess the robustness of the designs for this experiment to different values of $\mathbf{w}$.

To do this, we find compound Bayesian D-optimal designs for different values of $\mathbf{w}$ and compare the correlation of the columns in the model matrices, (3.21), for the three models assumed for the two responses from this experiment. These three models are; (i) the model for the Stage 1 response with respect to the Stage 1 factors, (ii) the model for the Stage 2 response with respect to the Stage 2 factors, and (iii) the cumulative model for the Stage 2 response with respect to the Stage 1 and Stage 2 factors. The form of these models is given in Table 3.6 and the columns in the model matrices for these models are given in Table 3.7 in Section 3.6 of Chapter 3.

We used 30 different weights in total, which are labelled (1) to (30) in Table 5.1; 10 with a weight of 0 on Model 2 ($w_2 = 0$, weights (1) to (10)), 10 with a weight of 0.01 on Model 2 ($w_2 = 0.01$, weights (11) to (20)), and 10 with a weight of 0.1 on Model 2 (w_2=0.1, weights (21) to (30)). We chose to have the smallest weight on Model 2 as this is the model which relates the responses from dissolution testing to Factor 6, which is of less interest than the cumulative model.

w_1	w_3	w_3	w_3
0.01	0.99 (1)	0.98 (11)	0.89 (21)
0.05	0.95 (2)	0.94 (12)	0.85 (22)
0.1	0.9 (3)	0.89 (13)	0.8 (23)
0.2	0.8 (4)	0.79 (14)	0.7 (24)
0.3	0.7 (5)	0.69 (15)	0.6 (25)
0.4	0.6 (6)	0.59 (16)	0.5 (26)
0.5	0.5 (7)	0.49 (17)	0.4 (27)
0.6	0.4 (8)	0.39 (18)	0.3 (28)
0.7	0.3 (9)	0.29 (19)	0.2 (29)
0.8	0.2 (10)	0.19 (20)	0.1 (30)

Table 5.1: Table giving the weights, $\mathbf{w} = (w_1, w_2 = 1 - w_1 - w_3, w_3)$ used in the assessment of the robustness of compound Bayesian D-optimal two-stage split-plot designs to $\mathbf{w}$.



(a)

(b)

(c)

Figure 5.2: Heat map of column correlation matrices for the compound Bayesian D-optimal design for: (a) (5), (b) (15), and (c) (25) in Table 5.1.

When examining the column correlations for the 30 optimal designs, we note that they only differ with respect to which of the columns involving Factor 6 are correlated for Model 3. An example can be seen in Figure 5.2. All 30 optimal designs have the same value of the Bayesian D-optimality objective function, (3.17), for Model 3. Therefore, even though these designs have different correlated columns under Model 3, they all have the same number of correlated columns for Model 3, which all involve Factor 6. These designs also all have no correlated columns for Models 1 and 2, and have the same values of the compound Bayesian D-optimality objective function.

5.2.2 Compound Bayesian D-Optimal Two-Stage Split-Plot Design for Formulation

The sixteen run compound Bayesian D-optimal two-stage split-plot design used for this experiment is given in Table 5.2. The structure of the design reflects the restrictions on randomisation. The order of the whole plot-plots, and the order of runs within whole-plots, can be randomised in this design.

Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
1	−1	−1	−1	1	−1
1	−1	−1	1	−1	1
1	−1	1	−1	−1	1
1	−1	1	1	1	−1
−1	1	−1	1	−1	−1
−1	1	1	1	1	−1
−1	1	1	−1	−1	1
−1	1	−1	−1	1	1
−1	−1	1	−1	1	−1
−1	−1	−1	−1	−1	−1
−1	−1	1	1	−1	1
−1	−1	−1	1	1	1
1	1	−1	−1	−1	−1
1	1	1	1	−1	−1
1	1	−1	1	1	1
1	1	1	−1	1	1

Table 5.2: The compound Bayesian D-optimal two-stage split-plot design, found using the coordinate exchange algorithm in Section 3.5.1, for the pharmaceutical formulation experiment.

Figure 5.3 provides a heat map for the column correlation matrix ((3.21) in Section 3.6.1) for the three model matrices considered for this design. Notice that, as with the compound Bayesian D-optimal two-stage split-plot designs given in Section 3.6.1 and the designs discussed in Section 5.2.1, the columns in the matrix for Model 1 not correlated and the columns in the matrix for Model 3 have correlations in the range $(-1, 1)$.

The terms relating to the correlated columns will be aliased and their bias and variance will be inflated. As no columns in the matrix for Model 1 are correlated, and the columns in the matrix for Model 3 have correlation in $(-1, 1)$, only columns involving the Stage 2 factor (Factor 6) are correlated with other columns in the matrix for Model 3. This is not concerning for this particular application, as the effect of Factor 6 is not of key importance.

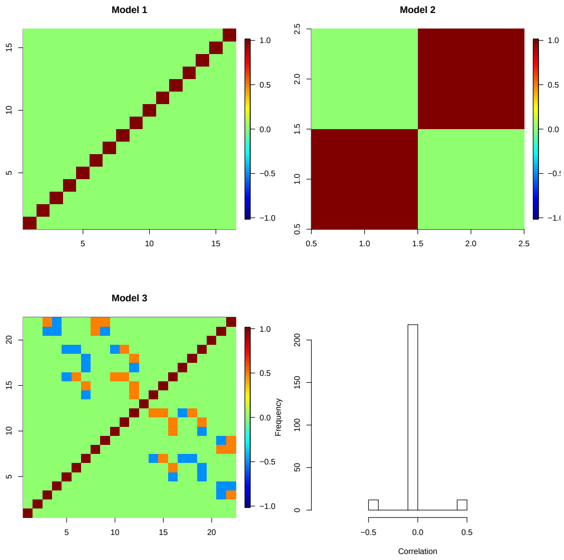


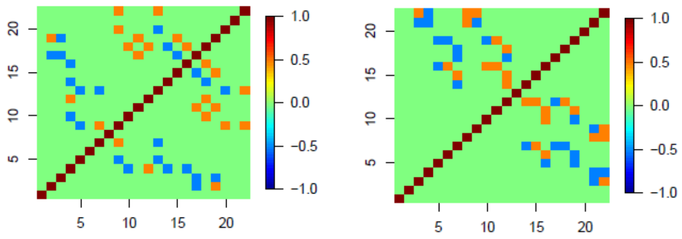
Figure 5.3: Heat map of column correlation matrices for the design in Table 5.2.

5.2.3 D-optimal Split-Plot Design for Formulation

When the experiment was performed, only the dissolution testing f_2 values were measured (the four responses from Stage 2 in Figure 5.1). Therefore, the experiment could have been designed as a single-stage split-plot design. To assess the impact of using a design for a two-stage experiment rather than a single-stage design, we find a D-optimal split-plot design and compare the column correlation matrix for Model 3 for the two-stage and single-stage design.

The single-stage D-optimal split-plot design was found using coordinate exchange algorithm in Section 3.5.1 with the Bayesian D-optimality objective function given by (3.17) when $\mathbf{X}_l$ is the model matrix for Model 3, $\mathbf{V}_l$ is (1.5) and $\mathbf{R}_l = \mathbf{I}_{22} - (\mathbf{e}_{22,1}\mathbf{e}_{22,1}^T)$, where $\mathbf{e}_{n,j}$ is the jth column of $\mathbf{I}_n$. This objective function is equivalent to using (3.20) with $\mathbf{w} = (0, 0, 1)$.

The column correlation matrices for Model 3 for the single-stage and two-stage split-plot designs are shown in Figure 5.4. Unlike the two-stage design, the single-stage design displays correlation between columns involving Factors 1 to 5, which were viewed by the scientist as a priori more important than Factor 6. Therefore, the two-stage design has some advantages over the single-stage design for this experiment, and was used to formulation pharmaceutical products which were then dissolution tested.



(a) (b)

Figure 5.4: Heat map of column correlation matrices for Model 3 for the optimal (a) single-stage and (b) two-stage split-plot designs.

5.3 Modelling of f_2 from Dissolution Testing

The pharmaceutical products which were formulated using the compound Bayesian D-optimal two-stage spit-plot design in Table 5.2 were then dissolution tested with respect to two reference products. In this section, we discuss the modelling of the output from dissolution testing, f_2 (5.1), for these pharmaceutical products.

In Section 5.3.1 we discuss the results of our initial analysis and the transformation applied to the f_2 values. In Section 5.3.2 we discuss the results of Bayesian variable selection using the samples from the Metropolis-Hastings within Gibbs sampling algorithm, and the assumption of correlation between the responses for the two tests is discussed in Section 5.3.3.

5.3.1 Initial Analysis and Transformation

In our initial analysis of the dissolution testing data, we assessed the validity of our assumption that the residuals from our experiment are normally distributed using diagnostic plots, such as those discussed and presented in Appendix B.1. We noted from this initial analysis that a logit transformation of the data was required in order for the assumption of normality to be appropriate.

The logit transformation is particularly relevant in this case. The f_2 statistic has the range of 0 to 100, and the logit transformation preserves this range. There is still some evidence of lack of normality in the model assessment plots given in Appendix B.2, however the transformed data was considered suitable to continue the analysis.

For an observed response, y_{iR}, $i = 1, \ldots, n$, $R = 1, \ldots, r$, the logit-transformed response, y^L_{iR}, for a response with the range 0 to 100 is

$$y^L_{iR} = \ln\left(\frac{y_{iR}}{100 - y_{iR}}\right), \tag{5.2}$$

Hence, before analysing the data we use (5.2) to transform the observed f_2 values in the $n \times r$ matrix of responses $\mathbf{Y}$.

During this initial analysis, we also noted that one of the centre points had unusual results for all four tests. After discussing these unusual results with the formulation team, we discovered there was an issue with this product during formulation. We therefore decided to discount it from our analysis. Hence, we had a 18×2 matrix of responses for each of the four tests.

We used two prior distributions for the within whole-plot correlation parameter ϕ, as discussed in Section 4.4.2. Prior 1 assumes the variance covariance matrix of the vectorised random effect and random error from the linear mixed model described in Section 4.4.1 are equal, whereas Prior 2 assumes the variance covariance matrix of the vectorised random effect is larger than the variance covariance matrix of the vectorised random error.

The experiment in Table 5.2 in Section 5.2.2 is a screening experiment, and further experimentation should be performed based on the results of this experiment. A model for the responses from the experiment in Table 5.2 containing the main effects and two factor interactions for all six factors is supersaturated with respect to a 18 run experiment. Therefore, the aim of this experiment is to find the terms that are likely to be active and the factors of most importance (variable selection), and then perform further experiments with these factors.

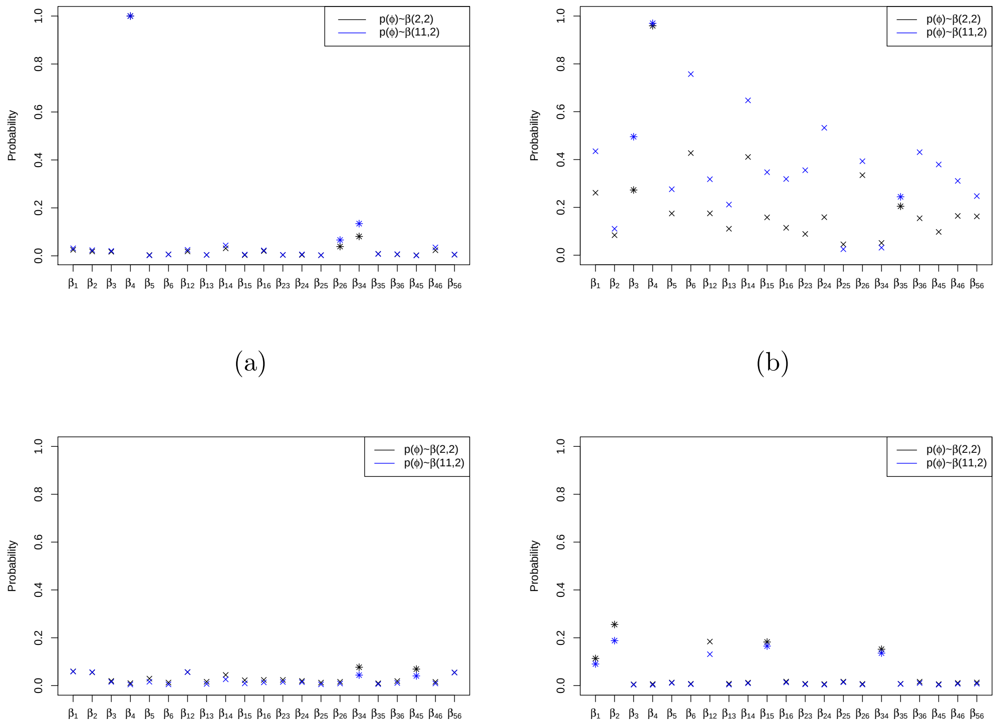
5.3.2 Variable Selection

This experiment is an initial, screening, experiment. Therefore, the identification of active terms in the models fitted to the dissolution testing data is particularly important, as the factors relating to active terms will be used in future experimentation. The active terms are identified using Bayesian variable selection, which we perform using samples of the indicator vector $\boldsymbol{\delta}$ from the Metropolis-Hastings within Gibbs sampling algorithm described in Section 4.4.4.

Recall that $\boldsymbol{\delta}$ is a $p \times 1$ vector, and the elements of $\boldsymbol{\delta}$, δ_j, $j = 1, \ldots, p$, are either 0, if the corresponding term $\boldsymbol{\beta}_j$ is not active ($\boldsymbol{\beta}_j = \mathbf{0}_r$), or 1, if $\boldsymbol{\beta}_j$ is active ($\boldsymbol{\beta}_j \neq \mathbf{0}_r$). The posterior probability of $\boldsymbol{\beta}_j$ being non-zero (or "active") can be approximated by (4.38) from Section 4.5.2. We note that we assume that the parameter is active for both correlated responses. An area of future work, as discussed in Section 4.6, is to assume that the parameter can be active for either correlated response independently.

Figure 5.5 shows the approximate posterior probabilities, calculated using (4.38), of the six main effects and fifteen two-way products (interactions) being active. The terms which have the highest posterior probabilities of being active when the responses are assumed to be uncorrelated are indicated by $*$ in this figure. We notice from Figures 5.5(a) and 5.5(b) that the main effect for Factor 4, $\boldsymbol{\beta}_4$, has a very high estimated posterior probability of being active for both Tests 1 and 2. The main effects for Factors 3 and 6, $\boldsymbol{\beta}_3$ and $\boldsymbol{\beta}_6$, and the interaction between Factors 1 and 2 and Factors

2 and 4, $\boldsymbol{\beta}_{14}$ and $\boldsymbol{\beta}_{24}$, also have a high estimated probability of being active for Test 2 (Figure 5.5(b)). All of the terms in Tests 3 and 4 have low estimated probabilities of being active, as seen in Figures 5.5(c) and 5.5(d).



(a)

(b)

(c)

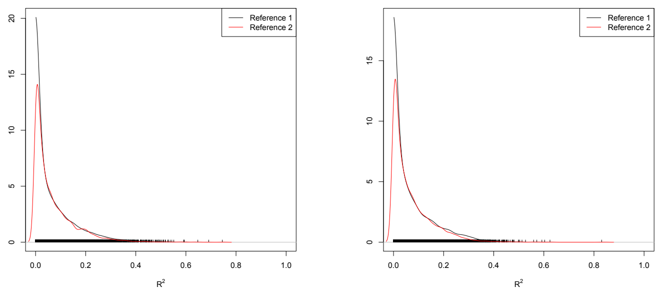
(d)

Figure 5.5: Approximate posterior probability of the six main effects and the fifteen two-way products (interactions) of the six factors in the model fitted to the logit transformed dissolution testing data from the design in Table 5.2 for Test (a) 1, (b) 2, (c) 3, and (d) 4 when both Prior 1 and Prior 2 are used. The parameters with the highest probability of being active when the f_2 values from the two references are assumed to be uncorrelated are indicated by $*$.

The approximate probabilities for the sampler with Prior 1 and Prior 2 are given in Figure 5.5. Note that the approximate posterior probabilities for both Prior 1 and Prior 2 are similar for Tests 1, 3 and 4, as shown in Figures 5.5(a), 5.5(c) and 5.5(d). However, there is some difference between the probabilities for the two prior distributions for Test 2, as shown in Figure 5.5(b).

For Test 2, the estimated posterior probabilities for Prior 2 are higher than the estimated posterior probabilities for Prior 1. These differences may be caused by the difference in the performance of the sampling algorithm for these two priors, which is discussed in detail in Appendix C.3.

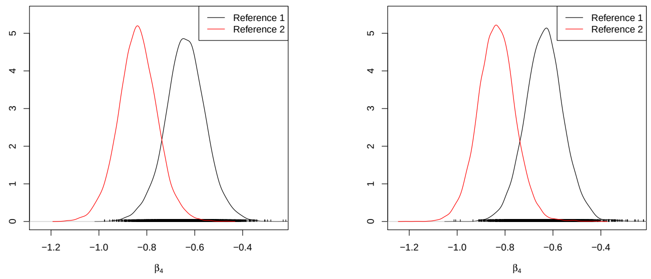
We note from Figure 5.5 that, in general, the terms with the highest posterior probability of being active when the response is assumed to be uncorrelated also have probabilities greater than the other parameters when the correlation between responses is considered. However when the correlation is accounted for, $\boldsymbol{\beta}_{35}$, which was active for uncorrelated responses, has a low probability of being active for Test 2, and other interaction terms such as $\boldsymbol{\beta}_{14}$ and $\boldsymbol{\beta}_{26}$ have higher probabilities (Figure 5.5(b)). We also note that $\boldsymbol{\beta}_{12}$, which was not active for uncorrelated responses has a slightly higher probability of being active in Test 4 (Figure 5.5(d)).



(a) (b)

Figure 5.6: Posterior density plots of the goodness-of-fit statistic R^2 for Test 3 for the sampler with (a) Prior 1 and (b) Prior 2.

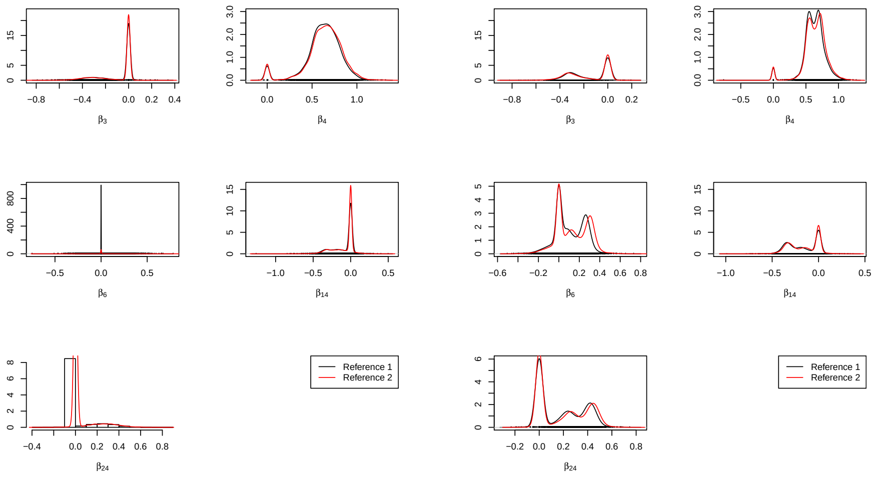
It is very difficult to assess which terms are active for Test 3 (Figure 5.5(c)), as the estimated posterior probability of being active for the majority of the parameters is close to zero. Figure 5.6 shows the posterior density of the goodness-of-fit statistic R^2 for Test 3 for both prior distributions. The mode of the posterior density for R^2 is very small for both prior distributions. Hence, the models fitted to the responses at each stage of the sampling algorithm are poor, and the variability in the model is not well explained by the terms in the models fitted to the responses.



(a) (b)

Figure 5.7: Marginal posterior density plots for $\boldsymbol{\beta}_4$ for Test 1 for (a) Prior 1 and (b) Prior 2.

Figures 5.7 and 5.8 are the posterior densities for parameters in Test 1 and 2, respectively, for which (4.38) is greater than 0.5 for both Prior 1 and 2. We note that only the marginal posterior density for β_{4R}, $R = 1, 2$, is unimodal in Test 1 (Figure 5.7), and has non-zero modes which dominate the zero mode for Test 2 (Figure 5.8). Hence, Factor 4 is a significant factor and should play a key role in future experimentation.



(a) (b)

Figure 5.8: Marginal conditional posterior density plots for $\boldsymbol{\beta}_3$, $\boldsymbol{\beta}_4$, $\boldsymbol{\beta}_6$, $\boldsymbol{\beta}_{14}$, $\boldsymbol{\beta}_{24}$ for Test 2 for (a) Prior 1 and (b) Prior 2.

Comparing Figure 5.8(a) and 5.8(b) explains why the probabilities for the sampler with Prior 2 are higher in Figure 5.5(b), as a high proportion of these densities are non-zero. We note that the difference in the performance of the sampling algorithm is the largest for the two prior distributions for ϕ for Test 2. This large difference appears to have affected the results. See Appendix C.3 for further detail.

As the variable selection results are impacted by the choice of prior distribution for ϕ for Test 2, we would need to consider how to select active terms for future experimentation. We could, for example, include all the terms that are selected using the sampler with Prior 1 and all the terms that are selected using the sampler with Prior 2 in future experimentation. Alternatively, we could consider the active terms which are common for both prior distributions.

Figure 5.9 shows the bivariate densities for $\boldsymbol{\beta}_4$ for Tests 1 and 2 for the sampler with Prior 1. We note that $\boldsymbol{\beta}_4$ has opposing signs for Test 1 and Test 2. Therefore, with all other factors held constant, the predicted response for Test 1 is maximised if Factor 4 is set to its low level, whereas the predicted response from Test 2 is maximised if Factor 4 is set to its high level. These results also hold for Prior 2 (not shown). This difference in signs creates a tension if the aim is to maximise f_2 for each test, as discussed in Section 5.4.

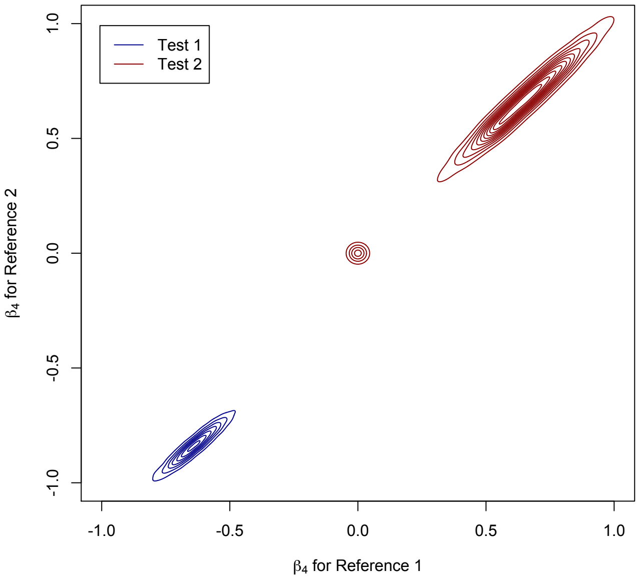


Figure 5.9: Bivariate density for $\boldsymbol{\beta}_4$ for Tests 1 and 2, for the sampler with Prior 1.

Samples from the posterior distribution of $\boldsymbol{\beta}_4$ are less variable for Test 1 than they are for Test 2. This is shown in Figures 5.7, 5.8 and 5.9 as all the densities for Test 1 have a smaller area than the densities for Test 2. We also note from these figures that $\boldsymbol{\beta}_4$ is never sampled as $\mathbf{0}_2$ for Test 1 (as the plots in Figure 5.7 for $\boldsymbol{\beta}_4$ have no peak at 0, and there is no mass at (0,0) in Figure 5.9 for Test 1), whereas $\boldsymbol{\beta}_4^T = \mathbf{0}_2$ for some samples for Test 2 (as the plots in Figure 5.8 for $\boldsymbol{\beta}_4$ has a peak at 0, and there is some mass at (0,0) in Figure 5.9 for Test 2).

The diagnostic plots presented in Appendix B.2 show that the fit of the posterior median of the predicted responses would need to be assessed in more detail in future experimentation. We note, however, that this experiment is meant as an initial screening study, and hence the aim is to identify the important factors and not find an accurate and precise predictive model.

5.3.3 Correlation

It is natural to assume that the two responses from each test are correlated, as the f_2 values for the two references are found using the same observed dissolution data, D_t, see (5.1).

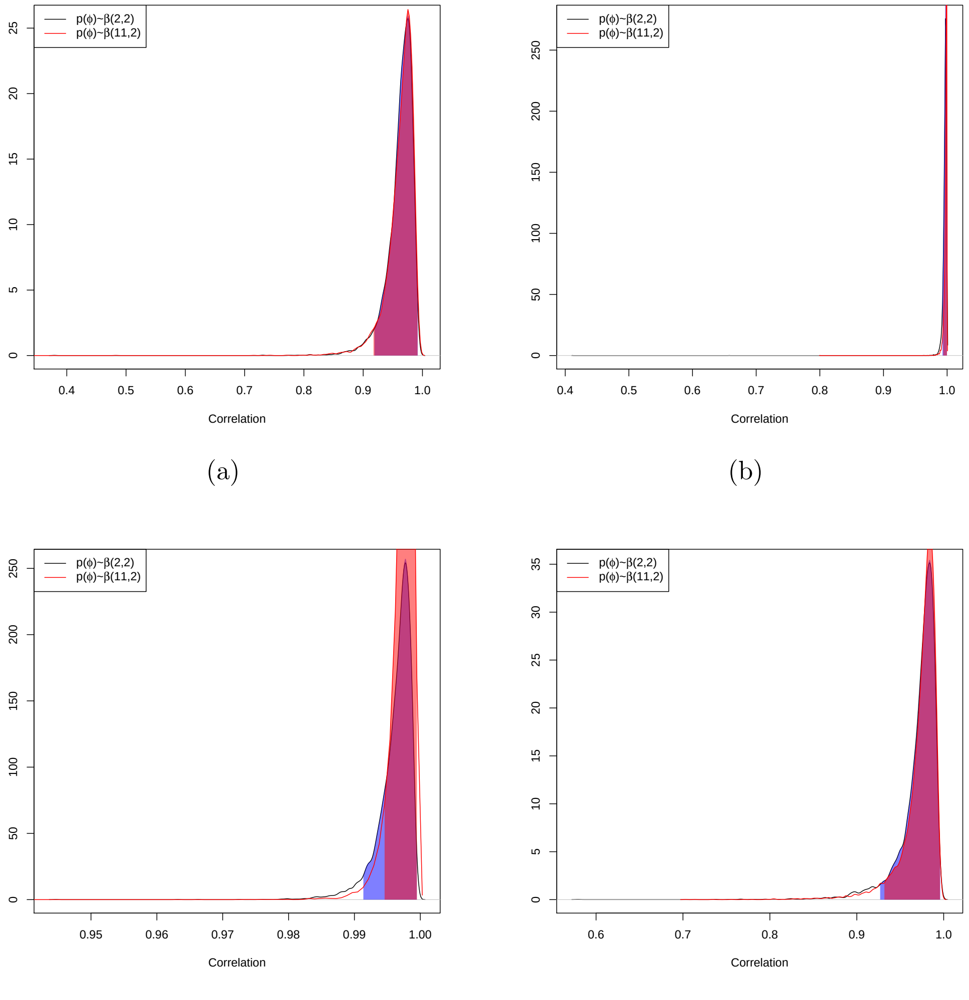
The correlation between the two random variables, Y_1 and Y_2, sampled from the bivariate matrix normal distribution is

$$\mathrm{Corr}(Y_1, Y_2) = \frac{\mathrm{Cov}(Y_1, Y_2)}{\sqrt{\mathrm{Var}(Y_1)\mathrm{Var}(Y_2)}} \tag{5.3}$$

where $\mathrm{Cov}(Y_1, Y_2)$ is the covariance between the two random variables and $\mathrm{Var}(Y_i)$, $i = 1, 2$, is the variance. Note that

$$\mathbf{\Sigma} = \begin{pmatrix} \mathrm{Var}(Y_1) & \mathrm{Cov}(Y_1, Y_2) \\ \mathrm{Cov}(Y_1, Y_2) & \mathrm{Var}(Y_2) \end{pmatrix}. \tag{5.4}$$

We can estimate (5.3) using the samples of $\mathbf{\Sigma}$. Figure 5.10 is the approximate posterior distribution between the f_2 for the two references found using the samples of $\mathbf{\Sigma}$ for all four dissolution tests, and Prior 1 and 2.



(a)

(b)

(c)

(d)

Figure 5.10: Posterior density for the correlation (5.3) between the columns of $\mathbf{Y}$ for Test: (a) 1, (b) 2, (c) 3, and (d) 4, for the sampler with Prior 1 and Prior 2. The shaded area is the 95% highest posterior density interval.

These densities all have very small 95% highest probability density intervals (shaded areas in the plots), centred on high correlation values, and modes which are close to 1. The α_{hpd}% highest posterior density interval is the area where α_{hpd}% of the sampled values lie. The results for the two ϕ prior distributions are very similar. Therefore, there is strong evidence that the responses from the two references for the same test are highly correlated, and this supports our modelling assumption that the active terms will be common to both references.

5.4 Predicted Probability of Meeting Specification

In this section, we discuss how we use the samples from the Metropolis-Hastings within Gibbs sampling algorithm to optimise the dissolution process, and identify new formulations from inside and outside the current experimental region that have a high probability of meeting specification for all tests. The probability of meeting specification is important, as the specification is set by the regulators. Therefore, the main aim of the experiment is to identify points which meet specification. The probability of meeting specification is a univariate summary of multivariate modelling, which is easier to optimise and visualise.

Let τ_d, be the fixed threshold for Test d, $d = 1, \dots, 4$, that is required by drug development regulators, and $y_{dR}(\mathbf{x})$, be the response for point $\mathbf{x} = (x_1, x_2, x_3, x_4, x_5, x_6)$ where x_f, $f = 1, \dots, 6$ is the level of factor f used in $\mathbf{x}$, for Test d and reference R, $R = 1, 2$. Then $\mathbf{x}$ meets specification for Test d and reference R if $y_{dR}(\mathbf{x}) \geq \tau_d$.

Let

$$I_{dR}(\mathbf{x}) = \begin{cases} 1 & \text{if } y_{dR}(\mathbf{x}) \geq \tau_d \\ 0 & \text{otherwise} \end{cases} \tag{5.5}$$

be the random variable which indicates whether the point $\mathbf{x}$ meets specification for Test d for reference R. The probability of meeting specification for Test d for reference R can be defined as $p_{dR}(\mathbf{x}) = P(I_{dR}(\mathbf{x}) = 1)$.

If tests are independent, the probability of $\mathbf{x}$ meeting specification for Tests $d_1, \dots, d_t$ is $p_{d_1 \dots d_t R}(\mathbf{x}) = p(I_{d_1 \dots d_t R}(\mathbf{x}) = 1)$, where

$$I_{d_1 \dots d_t R}(\mathbf{x}) = I_{d_1 R}(\mathbf{x}) \dots I d_t R(\mathbf{x}) = \begin{cases} 1 & \text{if } y_{d_1 R}(\mathbf{x}) \geq \tau_{d_1} \cap \dots \cap y_{d_t R}(\mathbf{x}) \geq \tau_{d_t} \\ 0 & \text{otherwise} \end{cases} \tag{5.6}$$

We can approximate $p_{dR}(\mathbf{x})$ and $p_{1 \dots d_t R}(\mathbf{x})$ for a new treatment $\mathbf{x}^*$ using predicted responses. The $(1 \times r)$ logit-transformed predicted response for $\mathbf{x}^*$ for References $1, \dots, r$ and Test d for the qth, $q = 1, \dots, its$, set of parameters sampled in the Metropolis-Hastings within Gibbs sampling algorithm is given by

$$\hat{\mathbf{y}}_d(\mathbf{x}^*)^{L,(q)} = (\hat{y}_{d1}(\mathbf{x}^*)^{L,(q)}, \dots, \hat{y}_{dr}(\mathbf{x}^*)^{L,(q)}) = f(\mathbf{x}^*)\mathbf{B}^{(q)} + \boldsymbol{\epsilon}^{(q)}, \tag{5.7}$$

where $\mathbf{B}^{(q)}$ is the qth sample of the $p \times r$ fixed effects matrix $\mathbf{B}$ from the algorithm and $\boldsymbol{\epsilon}^{(q)} \sim N(\mathbf{0}_r, \boldsymbol{\Sigma}^{(q)})$, when $\mathbf{0}_r$ is the $r \times r$ matrix with zero as every element and $\boldsymbol{\Sigma}^{(q)}$ is the qth sample of the $r \times r$ scale matrix $\boldsymbol{\Sigma}$ from the algorithm.

The posterior predicted response for Reference R, Test d and $\mathbf{x}^*$ for the qth MCMC sample $\hat{y}_{dR}(\mathbf{x}^*)^{(q)}$ is found by transforming $\hat{y}_{dR}(\mathbf{x}^*)^{L,(q)}$ using

$$\hat{y}_{dR}(\mathbf{x}^*)^{(q)} = \frac{100 \exp(\hat{y}_{dR}^{L,(q)}(\mathbf{x}^*))}{1 + \exp(\hat{y}_{dR}^{L,(q)}(\mathbf{x}^*))}. \tag{5.8}$$

Let

$$\psi_{dR}(\mathbf{x}^*)^{(q)} = \begin{cases} 1 & \text{if } \hat{y}_{dR}(\mathbf{x}^*)^{(q)} \geq \tau_d \\ 0 & \text{otherwise} \end{cases}, \tag{5.9}$$

and

$$\psi_{d_1 \dots d_t R}(\mathbf{x}^*)^{(q)} = \begin{cases} 1 & \text{if } \hat{y}_{d_1 R}(\mathbf{x}^*)^{(q)} \geq \tau_{d_1} \cap \cdots \cap \hat{y}_{d_t R}(\mathbf{x}^*)^{(q)} \geq \tau_{d_t} \\ 0 & \text{otherwise} \end{cases}, \tag{5.10}$$

then the approximate probability of treatment $\mathbf{x}^*$ meeting specification for Test d for reference R is

$$\hat{p}_{dR}(\mathbf{x}^*) = \frac{\sum_{q=1}^{its} \psi_{dk}(\mathbf{x}^*)^{(q)}}{its}, \tag{5.11}$$

and, more generally, the approximate probability of treatment $\mathbf{x}^*$ meeting specification Tests $d_1, \dots, d_t$ for reference R is

$$\hat{p}_{d_1 \dots d_t R}(\mathbf{x}^*) = \frac{\sum_{q=1}^{its} \psi_{d_1 \dots d_t R}(\mathbf{x}^*)^{(q)}}{its}. \tag{5.12}$$

In Section 5.4.1 we calculate $\hat{p}_{d_1 \dots d_t k}(\mathbf{x}^*)$ for two additional points added to the split plot experiment in Table 5.2 by the scientists during experimentation; (i) the modified centre point, $\mathbf{x}_{TP1} = (0, 0, -1, 0, 0, 0)$, which we refer to as Test Point 1, and (ii) modified $\mathbf{x}_{10}$, $\mathbf{x}_{TP2} = (-1, -1, -1, -0.5, -1, -1)$, which we refer to as Test Point 2. We also assess the performance of our model using these points.

We have noted in Section 5.3.2 that all terms have a low posterior probability of being active for Test 3. Therefore, we cannot identify influential factors for Test 3 and we do not have a convincing statistical model. Even though Test 3 is the least important test in the scientists' opinion, the pharmaceutical product still needs to meet specification for this test in order to meet regulatory requirements, therefore for completeness, we want find the point which is the solution to

$$\arg\max_{\mathbf{x}\in\mathcal{X}} \hat{p}_{1234R}(\mathbf{x}), \tag{5.13}$$

where $\mathcal{X} = [-1,1]^6 \subset \mathbb{R}^f$ is the set of all possible points for f factors with levels which lie in the range $[-1,1]$, and

$$\arg\max_{\mathbf{x}\in\mathcal{X}} \hat{p}_{124R}(\mathbf{x}), \tag{5.14}$$

when $R = 1, 2$.

We also consider optimisation of the probability for an extrapolated design region, and therefore we find the point which is the solution to

$$\arg\max_{\mathbf{x}\in\mathcal{X}^*} \hat{p}_{1234R}(\mathbf{x}), \tag{5.15}$$

where $\mathcal{X} \subset \mathcal{X}^* = [-2,2]^6 \subset \mathbb{R}^f$ is the set of all possible points for f factors with levels which lie in the range $[-2,2]$, and

$$\arg\max_{\mathbf{x}\in\mathcal{X}^*} \hat{p}_{124R}(\mathbf{x}). \tag{5.16}$$

In Sections 5.4.2 and 5.4.3 we discuss how we use a grid search and the Efficient Global Optimisation (EGO) algorithm (Jones et al., 1998), respectively, to optimise $\hat{p}_{d_1\ldots d_t R}(\mathbf{x})$ and solve (5.13) and (5.14). The results of these two methods of optimisation are discussed and compared in Section 5.4.4.

The calculations made in this section rely heavily on the assumptions used in the model, which need some further investigation based on the evidence in Appendix B.2. Expecting a supersaturated screening design to produce an accurate and precise predictive model is unrealistic, as estimating the scale matrix $\boldsymbol{\Sigma}$ is difficult. Also, screening designs are more suited for variable selection rather than prediction. However, there were significant time and resource constraints on this project, and our collaborators wanted to understand how the probability of meeting specification was influenced by the factor settings based on current knowledge.

5.4.1 Probability of Test Point 1 and Test Point 2 Meeting Specification

In addition to running the points in the split-plot design in Table 5.2, the experimenters also formulated pharmaceutical products for two additional test points. These test points were formed by adjusting points from the original design that had promising dissolution results. Test Point 1, $\mathbf{x}_{TP1}$=(0,0,$-$1,0,0,0), is an adjustment of the centre point $\mathbf{x}_{CP}$=(0,0,0,0,0,0). Test Point 2, $\mathbf{x}_{TP2}$=($-$1,$-$1,$-$1,$-$0.5,$-$1,$-$1), is an adjustment of run 10, $\mathbf{x}_{10}$=($-$1,$-$1,$-$1,$-$1,$-$1,$-$1).

The results of our modelling of the dissolution testing data in Section 5.3.2 suggests that these adjustments are reasonable. Recall from the analysis of Figures 5.5, 5.7, 5.8 and 5.9 that Factor 4 is influential for Test 1 and Test 2, and we recall from Figure 5.5(b) and 5.8 that the main effect Factor 3 has a high approximate posterior predicted probability of being active for Test 2.

Table 5.3 gives the approximate posterior probabilities, which are calculated using (5.12), of $\mathbf{x}_{CP}$ and $\mathbf{x}_{TP1}$ meeting specification for both references and the two prior distributions for ϕ. Similarly, Table 5.4 gives the approximate posterior probabilities of $\mathbf{x}_{10}$ and $\mathbf{x}_{TP2}$ meeting specification for both references for the samplers with Prior 1 and Prior 2.

Firstly, we note that the probabilities for the sampler with Prior 1 and Prior 2 in Tables 5.3 and 5.4 are very similar. Therefore, for the probability of meeting specification for these points appears to be robust to choice of these two prior distributions. A potential area for future work is consider the robustness of the probability to different prior distributions.

In Table 5.3 we see that $\hat{p}_{1234R}(\mathbf{x}_{TP1}) > \hat{p}_{1234R}(\mathbf{x}_{CP})$ and $\hat{p}_{124R}(\mathbf{x}_{TP1}) > \hat{p}_{124R}(\mathbf{x}_{CP})$ for $R = 1, 2$ and Prior 1 and 2. These approximate posterior probabilities have increased because $\hat{p}_{2R}(\mathbf{x}_{TP1}) > \hat{p}_{2R}(\mathbf{x}_{CP})$ for $R = 1, 2$ and Prior 1 and 2. This increase in $\hat{p}_{2k}(\mathbf{x}^*)$ was expected, as our analysis in Section 5.3.2 showed that the main effect for Factor 3, which is the only factor adjusted between $\mathbf{x}_{CP}$ and $\mathbf{x}_{TP1}$, has a high posterior predicted probability of being active for Test 2. As the main effect for Factor 3 has low posterior predicted probability of being active for Tests 1, 3 and 4, we expected $\hat{p}_{dR}(\mathbf{x}^*)$, $d = 1, 3, 4$ and $R = 1, 2$ to not be effected by changing the level of Factor 3.

Both centre points met specification for Tests 1, 3 and 4 but failed to meet specification Test 2 for both references. As can be seen from Table 5.3, our model predicts that $\mathbf{x}_{CP}$ has a low probability of passing Test 2 if it was repeated. When dissolution testing was performed on the pharmaceutical product formulated using $\mathbf{x}_{TP1}$, it only failed to meet specification for Test 1, Reference 2. The predicted probabilities from Table 5.3 suggest that repeat formulations of $\mathbf{x}_{TP1}$ would have a high probability of not meeting specification all four tests, Tests 1, 3, 4, or Test 2, for both references.

		$\mathbf{x}^* = \mathbf{x}_{CP}$	$\mathbf{x}^* = \mathbf{x}_{TP1}$
Prior 1	$\hat{p}_{12341}(\mathbf{x}^*)$	0.16	0.18
	$\hat{p}_{12342}(\mathbf{x}^*)$	0.17	0.19
	$\hat{p}_{1241}(\mathbf{x}^*)$	0.24	0.27
	$\hat{p}_{1242}(\mathbf{x}^*)$	0.26	0.29
	$\hat{p}_{11}(\mathbf{x}^*)$	0.77	0.77
	$\hat{p}_{12}(\mathbf{x}^*)$	0.83	0.83
	$\hat{p}_{21}(\mathbf{x}^*)$	0.38	0.42
	$\hat{p}_{22}(\mathbf{x}^*)$	0.39	0.44
	$\hat{p}_{31}(\mathbf{x}^*)$	0.68	0.67
	$\hat{p}_{32}(\mathbf{x}^*)$	0.68	0.67
	$\hat{p}_{41}(\mathbf{x}^*)$	0.81	0.82
	$\hat{p}_{42}(\mathbf{x}^*)$	0.77	0.79
Prior 2	$\hat{p}_{12341}(\mathbf{x}^*)$	0.12	0.17
	$\hat{p}_{12342}(\mathbf{x}^*)$	0.13	0.19
	$\hat{p}_{1241}(\mathbf{x}^*)$	0.19	0.28
	$\hat{p}_{1242}(\mathbf{x}^*)$	0.21	0.30
	$\hat{p}_{11}(\mathbf{x}^*)$	0.74	0.73
	$\hat{p}_{12}(\mathbf{x}^*)$	0.78	0.78
	$\hat{p}_{21}(\mathbf{x}^*)$	0.33	0.50
	$\hat{p}_{22}(\mathbf{x}^*)$	0.35	0.52
	$\hat{p}_{31}(\mathbf{x}^*)$	0.63	0.63
	$\hat{p}_{32}(\mathbf{x}^*)$	0.64	0.63
	$\hat{p}_{41}(\mathbf{x}^*)$	0.78	0.78
	$\hat{p}_{42}(\mathbf{x}^*)$	0.75	0.75

Table 5.3: Approximate posterior probabilities (5.12) of the centre point, $\mathbf{x}_{CP}$=(0,0,0,0,0,0), and Test Point 1, $\mathbf{x}_{TP1}$=(0,0,−1,0,0,0), meeting specification for the sampler with Prior 1 and Prior 2.

		$\mathbf{x}^* = \mathbf{x}_{10}$	$\mathbf{x}^* = \mathbf{x}_{TP2}$
Prior 1	$\hat{p}_{12341}(\mathbf{x}^*)$	0.06	0.11
	$\hat{p}_{12342}(\mathbf{x}^*)$	0.07	0.11
	$\hat{p}_{1241}(\mathbf{x}^*)$	0.10	0.16
	$\hat{p}_{1242}(\mathbf{x}^*)$	0.10	0.16
	$\hat{p}_{11}(\mathbf{x}^*)$	0.97	0.91
	$\hat{p}_{12}(\mathbf{x}^*)$	0.99	0.96
	$\hat{p}_{21}(\mathbf{x}^*)$	0.13	0.23
	$\hat{p}_{22}(\mathbf{x}^*)$	0.14	0.24
	$\hat{p}_{31}(\mathbf{x}^*)$	0.67	0.67
	$\hat{p}_{32}(\mathbf{x}^*)$	0.67	0.68
	$\hat{p}_{41}(\mathbf{x}^*)$	0.78	0.79
	$\hat{p}_{42}(\mathbf{x}^*)$	0.73	0.74
Prior 2	$\hat{p}_{12341}(\mathbf{x}^*)$	0.05	0.07
	$\hat{p}_{12342}(\mathbf{x}^*)$	0.05	0.08
	$\hat{p}_{1241}(\mathbf{x}^*)$	0.08	0.11
	$\hat{p}_{1242}(\mathbf{x}^*)$	0.09	0.12
	$\hat{p}_{11}(\mathbf{x}^*)$	0.94	0.85
	$\hat{p}_{12}(\mathbf{x}^*)$	0.97	0.91
	$\hat{p}_{21}(\mathbf{x}^*)$	0.12	0.18
	$\hat{p}_{22}(\mathbf{x}^*)$	0.13	0.19
	$\hat{p}_{31}(\mathbf{x}^*)$	0.63	0.63
	$\hat{p}_{32}(\mathbf{x}^*)$	0.63	0.63
	$\hat{p}_{41}(\mathbf{x}^*)$	0.75	0.76
	$\hat{p}_{42}(\mathbf{x}^*)$	0.71	0.72

Table 5.4: Approximate posterior probabilities (5.12) of run 10, $\mathbf{x}_{10}$=(−1,−1,−1,−1,−1,−1), and Test Point 2, $\mathbf{x}_{TP2}$=(−1,−1,−1,−0.5,−1,−1), meeting specification for the sampler with Prior 1 and Prior 2.

Comparison of the columns in Table 5.4 suggests that increasing Factor 4 in $\mathbf{x}_{10}$ from -1 to -0.5 leads to increased posterior probability of meeting specification for all the tests. It is unsurprising that Factor 4 should have such an impact on the probabilities, as $\boldsymbol{\beta}_4$ has a high posterior probability of being active for Test 1 and Test 2 (see Figures 5.5 (a),(b), 5.7 and 5.8).

We also note from Table 5.13 that $\hat{p}_{1R}(\mathbf{x}_{10}) > \hat{p}_{1R}(\mathbf{x}_{TP2})$, whereas $\hat{p}_{2R}(\mathbf{x}_{10}) < \hat{p}_{2R}(\mathbf{x}_{TP2})$. This is expected, as $\boldsymbol{\beta}_4$ is negative for Test 1, hence $\hat{y}_{1R}(\mathbf{x}^*)$ and $\hat{p}_{1R}(\mathbf{x}^*)$ will decrease when the level of Factor 4 is increased and all other factors are held constant. Similarly, $\boldsymbol{\beta}_4$ is positive for Test 2, so $\hat{y}_{2R}(\mathbf{x}^*)$ and $\hat{p}_{2R}(\mathbf{x}^*)$ will increase when the level of Factor 4 increases, when all other factors are held constant.

The pharmaceutical products which were formulated for $\mathbf{x}_{10}$ and $\mathbf{x}_{TP2}$ both failed to meet specification for Test 2 for both references. The predicted probabilities in Table

5.4 suggest that these points have a high probability of not meeting specification for Test 2, and all four tests, again if they were repeated.

Figures 5.11 and 5.12 are the approximate posterior densities for $\hat{y}_{dR}(\mathbf{x}_{TP1})$, when $d = 1, 2, 3, 4$ and $R = 1, 2$, for Prior 1 and 2, respectively. Similarly, Figures 5.13 and 5.14 are the approximate posterior predictive densities for $\hat{y}_{dR}(\mathbf{x}_{TP2})$, when $d = 1, 2, 3, 4$ and $R = 1, 2$, for Prior 1 and 2, respectively. The predicted responses, $\hat{y}_{dR}(\mathbf{x}^*)$, $\mathbf{x}^* = \mathbf{x}_{TP1}, \mathbf{x}_{TP2}$, are calculated using (5.7) and (5.8). We use these figures to assess the performance of our model by comparing the mode and spread of the densities in these figures to the observed f_2 values, which are plotted as a solid black (for Reference 1) and red (for Reference 2) points. The threshold value, which the observed f_2 has to be greater than in order for the point to meet specification, is given as a dotted line.

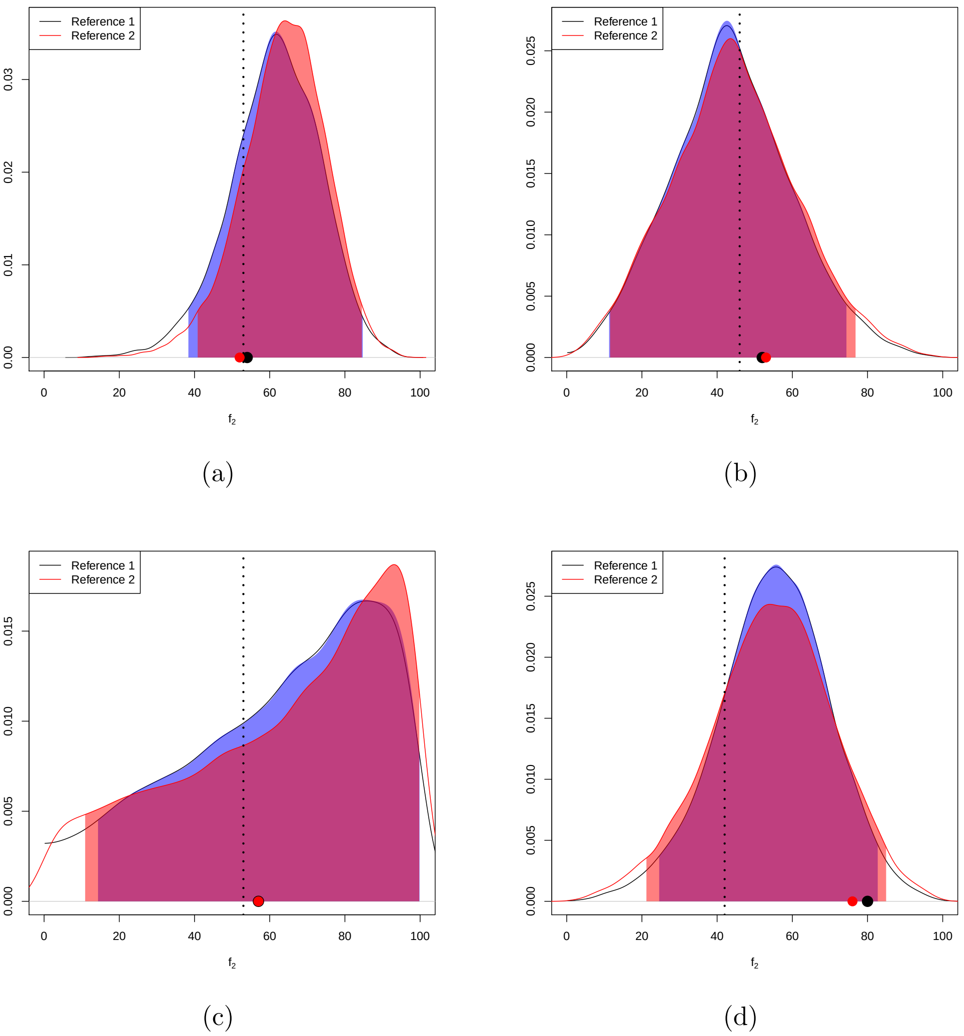
The posterior predictive densities in Figures 5.11 to 5.14 support the probabilities in Tables 5.11 and 5.13, as part of the 95% highest posterior density interval for all the densities is on the left of threshold. Therefore, a number of the predicted responses for $\mathbf{x}_{TP1}$ and $\mathbf{x}_{TP2}$ will not meet specification. The variability of the approximate predictive densities in Figures 5.11 to 5.14 also demonstrates the underlying uncertainty regarding the f_2 value for Test Points 1 and 2 in our model.

We note in Figures 5.11(b) and 5.12(b) that the modes of the approximate posterior predictive densities for $\hat{y}_{2R}(\mathbf{x}_{TP1})$ when $R = 1, 2$ are close to the threshold for the sampler for both Prior 1 and 2, which explains why $\hat{p}_{2R}(\mathbf{x}_{TP1})$ in Table 5.3 for Prior 1 and 2 is in the range [0.42, 0.52].

The mode and a large proportion of the 95% posterior density intervals for the approximate posterior densities of $\hat{y}_{2R}(\mathbf{x}_{TP2})$ are to the left of the threshold in Figures 5.13(b) and 5.14(b), which explains the low $\hat{p}_{2R}(\mathbf{x}_{TP2})$ seen in Table 5.4.

We note that the observed f_2 values for $\mathbf{x}_{TP1}$ for Test 2 in Figures 5.11(b) and 5.12(b), and Test 4 in Figures 5.11(d) and 5.12(d), are greater than the mode of the approximate posterior predictive density. Therefore, the model is likely to underestimate $y_{2R}(\mathbf{x}_{TP1})$ and $y_{4R}(\mathbf{x}_{TP1})$ for $R = 1, 2$. The observed f_2 values for $\mathbf{x}_{TP1}$ for Test 1 in Figures 5.11(a) and 5.12(a), and Test 3 in Figures 5.11(c) and 5.12(c), are greater than the mode of the approximate posterior predictive density. Hence, the model is likely to overestimate $y_{1R}(\mathbf{x}_{TP1})$ and $y_{3R}(\mathbf{x}_{TP1})$ for $R = 1, 2$.

The observed f_2 values for $\mathbf{x}_{TP2}$ for all tests, both references and the sampler with Prior 1 and 2 are close to the modes of the approximate posterior predictive densities, and therefore the estimates for $y_{dR}(\mathbf{x}_{TP2})$, $R = 1, 2$, $d = 1, 2, 3, 4$ are centred around more appropriate values than the estimates for $y_{dR}(\mathbf{x}_{TP1})$. However, it is important to note that the observed values for both $\mathbf{x}_{TP1}$ and $\mathbf{x}_{TP2}$ lie within the 95% highest posterior density intervals, hence providing some validity for the model.



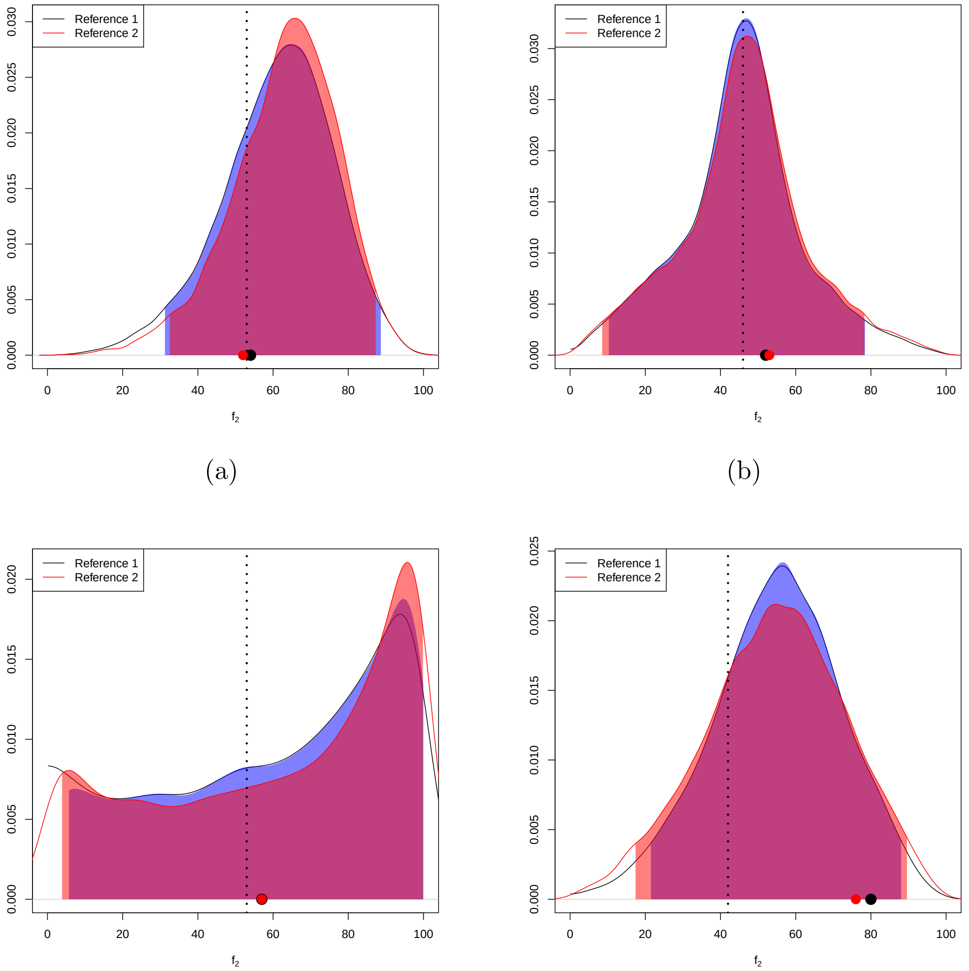
(a)

(b)

(c)

(d)

Figure 5.11: Approximate posterior predictive density for (a) $\hat{y}_{11}(\mathbf{x}_{TP1})$ and $\hat{y}_{12}(\mathbf{x}_{TP1})$, (b) $\hat{y}_{21}(\mathbf{x}_{TP1})$ and $\hat{y}_{32}(\mathbf{x}_{TP1})$, (c) $\hat{y}_{31}(\mathbf{x}_{TP1})$ and $\hat{y}_{32}(\mathbf{x}_{TP1})$, and (d) $\hat{y}_{41}(\mathbf{x}_{TP1})$ and $\hat{y}_{42}(\mathbf{x}_{TP1})$, when $\mathbf{x}_{TP1}$=(0,0,−1,0,0,0), for the sampler with Prior 1. The shaded area is the 95% posterior density interval. The dotted line is the threshold which the predicted response needs to be greater than to meet specification. The black and red solid points indicate the observed f_2 values for $\mathbf{x}_{TP1}$ for Reference 1 and 2, respectively.



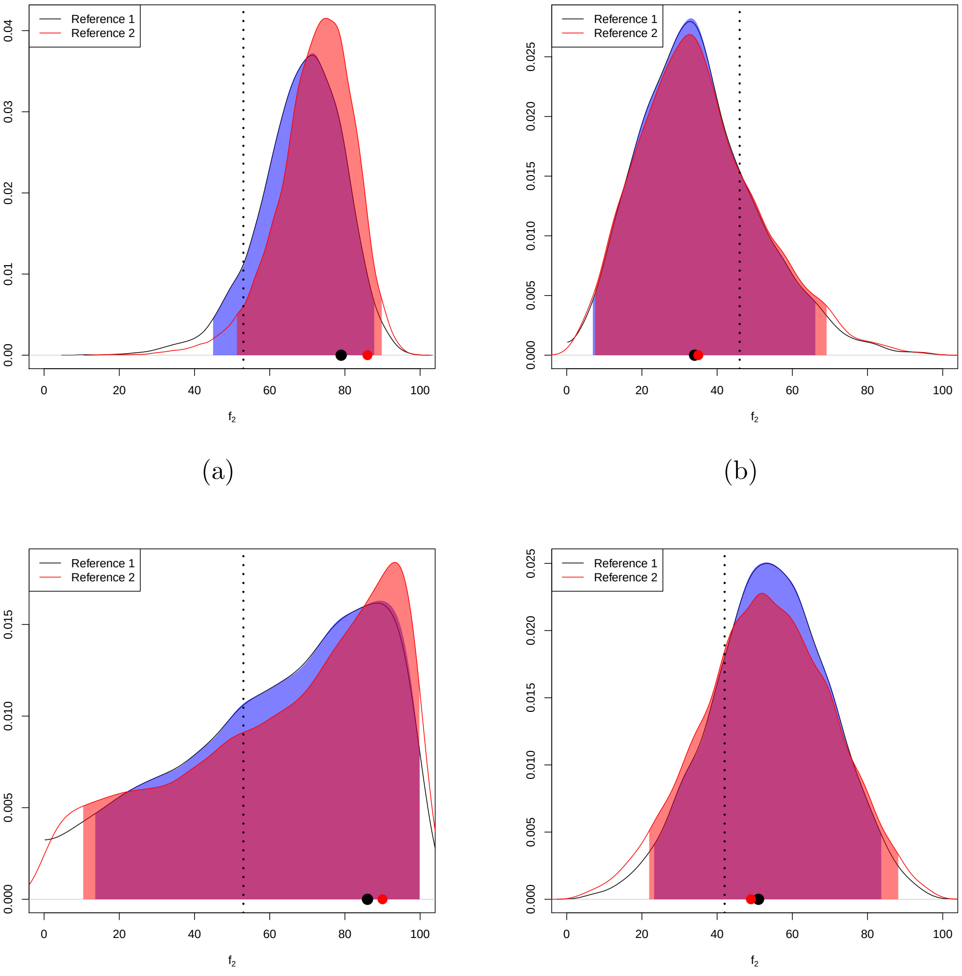
(a)

(b)

(c)

(d)

Figure 5.12: Approximate posterior predictive density for (a) $\hat{y}_{11}(\mathbf{x}_{TP1})$ and $\hat{y}_{12}(\mathbf{x}_{TP1})$, (b) $\hat{y}_{21}(\mathbf{x}_{TP1})$ and $\hat{y}_{32}(\mathbf{x}_{TP1})$, (c) $\hat{y}_{31}(\mathbf{x}_{TP1})$ and $\hat{y}_{32}(\mathbf{x}_{TP1})$, and (d) $\hat{y}_{41}(\mathbf{x}_{TP1})$ and $\hat{y}_{42}(\mathbf{x}_{TP1})$, when $\mathbf{x}_{TP1}$=(0,0,−1,0,0,0) for the sampler with Prior 2. The shaded area is the 95% posterior density interval. The dotted line is the threshold which the predicted response needs to be greater than to meet specification. The black and red solid points indicate the observed f_2 values for $\mathbf{x}_{TP1}$ for Reference 1 and 2, respectively.



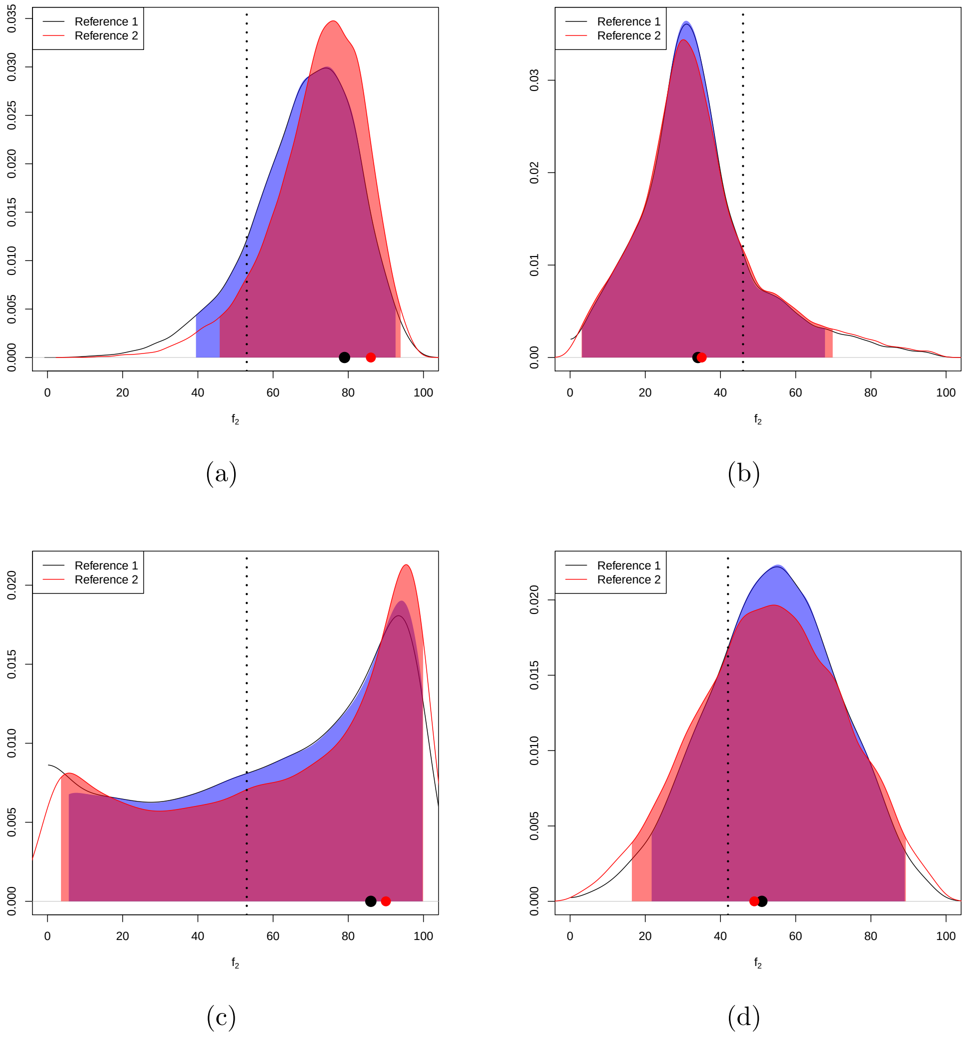
(a)

(b)

(c)

(d)

Figure 5.13: Approximate posterior predictive density for (a) $\hat{y}_{11}(\mathbf{x}_{TP2})$ and $\hat{y}_{12}(\mathbf{x}_{TP2})$, (b) $\hat{y}_{21}(\mathbf{x}_{TP2})$ and $\hat{y}_{32}(\mathbf{x}_{TP2})$, (c) $\hat{y}_{31}(\mathbf{x}_{TP2})$ and $\hat{y}_{32}(\mathbf{x}_{TP2})$, and (d) $\hat{y}_{41}(\mathbf{x}_{TP2})$ and $\hat{y}_{42}(\mathbf{x}_{TP2})$, when $\mathbf{x}_{TP2}$=(−1,−1,−1,−0.5,−1,−1) for the sampler with Prior 1. The shaded area is the 95% posterior density interval. The dotted line is the threshold which the predicted response needs to be greater than to meet specification. The black and red solid points indicate the observed f_2 values for $\mathbf{x}_{TP2}$ for Reference 1 and 2, respectively.



(a)

(b)

(c)

(d)

Figure 5.14: Approximate posterior predictive density for (a) $\hat{y}_{11}(\mathbf{x}_{TP2})$ and $\hat{y}_{12}(\mathbf{x}_{TP2})$, (b) $\hat{y}_{21}(\mathbf{x}_{TP2})$ and $\hat{y}_{32}(\mathbf{x}_{TP2})$, (c) $\hat{y}_{31}(\mathbf{x}_{TP2})$ and $\hat{y}_{32}(\mathbf{x}_{TP2})$, and (d) $\hat{y}_{41}(\mathbf{x}_{TP2})$ and $\hat{y}_{42}(\mathbf{x}_{TP2})$, when $\mathbf{x}_{TP2}$=(−1,−1,−1,−0.5,−1,−1) for the sampler with Prior 2. The shaded area is the 95% posterior density interval. The dotted line is the threshold which the predicted response needs to be greater than to meet specification. The black and red solid points indicate the observed f_2 values for $\mathbf{x}_{TP2}$ for Reference 1 and 2, respectively.

5.4.2 Grid Search

The point, $\mathbf{x}^*$, which maximises (5.12) for Reference R, $R = 1, \ldots, r$ and Tests $d_1, \ldots, d_t$, can be found using a grid search and subsequent Nelder-Mead optimisation (Nelder and Mead, 1965). This procedure has the following steps, where $\mathbf{D} \in \mathcal{D}_{f,l,f^l}$ is the full factorial design with f l-level factors and $n = f^l$ runs, and $\mathbf{x}_i$ is the ith row of $\mathbf{D}$:

1. For $i = 1, \ldots, n$ and $q = 1, \ldots, its$
   (a) Calculate $\hat{y}^{(q)}_{d_1 k}(\mathbf{x}_i), \ldots, \hat{y}^{(q)}_{d_1 k}(\mathbf{x}_i)$ using (5.7) and the transformation (5.8).
   (b) Calculate (5.10).
2. Calculate $\hat{p}_{d_1 \ldots d_t k}(\mathbf{x}_i)$ using (5.12).
3. Find $\hat{p}_{GS} = \max_{\forall i} \hat{p}_{d_1 \ldots d_t k}(\mathbf{x}_i)$.
4. Find $\mathbf{x}_{GS} = \arg\max_{\mathbf{x}_i \forall i} \hat{p}_{d_1 \ldots d_t k}(\mathbf{x}_i)$.
5. Use $\mathbf{x}_{GS}$ as the starting value for a local optimisation using the Nelder-Mead algorithm (Nelder and Mead, 1965, the default optimisation method for `optim` in `R`), which optimises (5.12) for treatments over a continuous range of factor levels. Call the treatment and probability found using the Nelder-Mead algorithm $\mathbf{x}_{NM}$ and $\hat{p}_{NM}$, respectively.
6. If $\hat{p}_{GS} = \hat{p}_{NM}$, then $\mathbf{x}_{GS}$ and $\mathbf{x}_{NM}$ may be identical. If they are not, then either both $\mathbf{x}_{GS}$ and $\mathbf{x}_{NM}$ should be run, or one could be chosen based on some other feature of experimentation. Otherwise, $\mathbf{x}_{NM}$ is the treatment which optimises (5.12).

This method is very slow as it requires q calculations of (5.7) for each of the n treatments in $\mathbf{D}$. The samples used in (5.7) can be thinned to make computation faster. In this chapter, we thinned our $its = 10,000$ samples to 500 by selecting every 20th sample.

An alternative method of speeding up computation would be to reduce l in order to reduce n. Our experiment had six factors, so $f = 6$, and we let $l = 9$, where the levels were (-1, -0.75, -0.5, -0.25, 0, 025, 0.5, 0.75, 1), as this was computationally feasible.

5.4.3 Efficient Global Optimisation (EGO) Algorithm

As discussed in Section 5.4.2, it is computationally expensive to calculate (5.12), as it requires q predicted responses, (5.7), to be sampled for each treatment. One method of speeding up the optimisation of (5.12) is to reduce the number of points it is calculated for. Computer experiments provide a suitable, more computationally efficient, method of optimising (5.12).

Computer experiments assume that the form of a computationally expensive function, such as (5.12), is unknown, and this unknown function is sometimes referred to as a black-box function. Computer experiments also assume that we only have a certain number of observed outputs from this black-box function. These observed outputs are used to build a surrogate model for the unknown black-box function. The function can then be maximised using this surrogate model. Further discussion on computer experiments is given by Fang et al. (2006).

In this section, we introduce the efficient global optimisation (EGO) algorithm (Jones et al., 1998), which builds a surrogate model for (5.12) using a computer experiment, and uses this surrogate model to identify the point which optimises (5.12). The EGO algorithm requires:

1. **Space-filling Designs**: A space-filling design optimises the location of the points within the boundaries of the current design region (which are the maximum and minimum factor levels) using certain criterion. The criterion used to determine the location of the points can be stochastic, so the points are randomly generated, or deterministic, such as a distance metric. Further detail on space filling designs are given in Fang et al. (2006, Part II). In this work we use a randomly sampled Latin-hypercube design, which is a generalisation of a Latin square to more than two dimensions. Latin-hypercube designs are discussed in McKay et al. (1979), Stein (1987) and Fang et al. (2006).

   Jones et al. (1998) suggested having at least $10f$ points in the initial randomly sampled Latin-hypercube, with adjustments such that the inter-point spacing $1/(n-1)$ is a finite-decimal. For $f = 6$ Jones et al. (1998) suggested using 65 points as then the inter-point spacing is 1/64=0.015625.

2. **Kriging Modelling**: Kriging is a method of point interpolation, or smoothing, that uses Gaussian processes. A Gaussian process is a correlated stochastic process which can be defined using a mean function and correlation function. Any realisation from a Gaussian process has a multivariate normal distribution. Gaussian processes are discussed in detail in Rasmussen and Williams (2006).

   In this section, we assume that

$$p^{-1}_{d_1 \dots d_t R}(\mathbf{x}_i) = \mu + \epsilon(\mathbf{x}_i) \tag{5.17}$$

   where $p^{-1}_{d_1 \dots d_t R}(\mathbf{x}_i)$ is reciprocal of $p_{d_1 \dots d_t R}(\mathbf{x}_i)$, μ is the average of $p^{-1}_{d_1 \dots d_t R}(\mathbf{x}_i)$, and $\epsilon(\mathbf{x}_i) \sim N(0, \sigma_g^2)$. If we assume that $\epsilon \sim GP(0, \sigma_g^2 \kappa(\mathbf{x}_i, \mathbf{x}_{i^*}))$, then two realisations from ϵ, $\epsilon(\mathbf{x}_i)$ and $\epsilon(\mathbf{x}_{i^*})$, $i, i^* \in \{1, \dots, n\}$, are dependent. In this chapter, we assume that

$$\text{corr}\{\epsilon(\mathbf{x}_i), \epsilon(\mathbf{x}_{i^*})\} = \kappa(\mathbf{x}_i, \mathbf{x}_{i^*}), \tag{5.18}$$

where

$$\kappa(\mathbf{x}_i, \mathbf{x}_{i^*}) = \exp\left(-\sum_{j=1}^{f} \theta_{1j}|x_{if} - x_{i^*f}|^{\theta_2}\right), \tag{5.19}$$

when x_{if} is the fth element of $\mathbf{x}_i$, $\theta_{1j} \geq 0$, $j = 1, \ldots, f$, and $\theta_2 \in \{1, 2\}$. We use the `km` function in the package `DiceKriging` in `R` to fit the Kriging model, which optimises θ_{1j}, $j = 1, \ldots, f$, and sets $\theta_2 = 2$.

The Kriging model can also be used to predict the response for new, unobserved, points. Let $\hat{\mathbf{p}}^{-1}_{d_1 \ldots d_t R}(\mathbf{X}) = (\hat{p}^{-1}_{d_1 \ldots d_t R}(\mathbf{x}_1), \ldots, \hat{p}^{-1}_{d_1 \ldots d_t R}(\mathbf{x}_n))^T$ be a vector of the inverse probabilities predicted using (5.17) for $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)^T$, and $\hat{p}^{-1}_{d_1 \ldots d_t R}(\mathbf{x}^*)$ be the probability we want to predict. Following the results in Fang et al. (2006) and Rasmussen and Williams (2006),

$$\begin{pmatrix} \hat{p}^{-1}_{d_1 \ldots d_t R}(\mathbf{x}^*) \\ \hat{\mathbf{p}}^{-1}_{d_1 \ldots d_t R}(\mathbf{X}) \end{pmatrix} \sim N\left(\begin{pmatrix} \mu \\ \mu \mathbf{1}_n \end{pmatrix}, \begin{pmatrix} \sigma_g^2 & \boldsymbol{\sigma}_{12} \\ \boldsymbol{\sigma}_{21} & \boldsymbol{\sigma}_{22} \end{pmatrix}\right), \tag{5.20}$$

where μ and σ_g^2 are the scalar mean and variance, respectively, of the distribution for $\hat{p}^{-1}_{d_1 \ldots d_t R}(\mathbf{x}^*)$, $\mu \mathbf{1}_n$ and $\boldsymbol{\sigma}_{22}$ are the $n \times 1$ mean vector and $n \times n$ variance-covariance matrix, respectively, of the distribution for $\hat{\mathbf{p}}^{-1}_{d_1 \ldots d_t R}(\mathbf{X})$, $\boldsymbol{\sigma}_{12}$ is the $1 \times n$ vector of covariances between the response at the new point and the response at the existing points, and $\boldsymbol{\sigma}_{12} = \boldsymbol{\sigma}_{21}^T$.

The mean and variance of the predicted inverse probability for the new treatment $\mathbf{x}^*$ given $\hat{\mathbf{p}}^{-1}_{d_1 \ldots d_t R}(\mathbf{X})$ are

$$\mathrm{E}(\hat{p}^{-1}_{d_1 \ldots d_t R}(\mathbf{x}^*)|\hat{\mathbf{p}}^{-1}_{d_1 \ldots d_t R}(\mathbf{X})) = \mu + \boldsymbol{\sigma}_{12}\boldsymbol{\sigma}_{22}^{-1}(\hat{\mathbf{p}}^{-1}_{d_1 \ldots d_t R}(\mathbf{X}) - \mu \mathbf{1}_n) \tag{5.21}$$

$$\mathrm{Var}(\hat{p}^{-1}_{d_1 \ldots d_t R}(\mathbf{x}^*)|\hat{\mathbf{p}}^{-1}_{d_1 \ldots d_t R}(\mathbf{X})) = \sigma_g^2 - \boldsymbol{\sigma}_{12}\boldsymbol{\sigma}_{22}^{-1}\boldsymbol{\sigma}_{21} \tag{5.22}$$

respectively. In order to estimate (5.21) and (5.22), μ_1, $\boldsymbol{\mu}_2$, σ_{11}^2, $\boldsymbol{\sigma}_{12}$ and $\boldsymbol{\sigma}_{22}$ will need to be estimated. Note that (5.21) is the best linear unbiased predictor of $\hat{p}^{-1}_{d_1 \ldots d_t R}(\mathbf{x}^*)$ and (5.22) is the mean squared error for the best linear unbiased predictor of $\hat{p}^{-1}_{d_1 \ldots d_t R}(\mathbf{x}^*)$ (Fang et al., 2006, Section 5.4.1).

3. **Sequential Design via Expected Improvement**: The EGO algorithm uses expected improvement to add points to the initial space-filling design, and hence converge to the point which maximises the response from the unknown function using the surrogate model. Expected improvement balances the objectives of exploration, which is achieved by evaluating the computationally expensive function

for new points, and exploitation, or using the samples from the computationally expensive function that we already have to reduce computation time.

In the EGO algorithm, we wish to add the $\mathbf{x}^*$ which maximises the expected improvement,

$$\left(\min_{\forall i=1,\ldots,n} \hat{p}^{-1}_{d_1\ldots d_t R}(\mathbf{x}_i) - \tilde{p}^{-1}_{d_1\ldots d_t R}(\mathbf{x}^*)\right) \Phi \left(\frac{\min_{\forall i=1,\ldots,n} \hat{p}^{-1}_{d_1\ldots d_t R}(\mathbf{x}_i) - \tilde{p}^{-1}_{d_1\ldots d_t R}(\mathbf{x}^*)}{\mathrm{sd}(\mathbf{x}^*)}\right)$$

$$+\mathrm{sd}(\mathbf{x}^*)\Phi\left(\frac{\min_{\forall i=1,\ldots,n} \hat{p}^{-1}_{d_1\ldots d_t R}(\mathbf{x}_i) - \tilde{p}^{-1}_{d_1\ldots d_t R}(\mathbf{x}^*)}{\mathrm{sd}(\mathbf{x}^*)}\right), \tag{5.23}$$

where Φ and ϕ are the standard normal cumulative density and probability density functions, respectively, $\hat{p}^{-1}_{d_1\ldots d_t R}(\mathbf{x}_i)$ is the reciprocal of (5.12) found using the surrogate Gaussian process model (5.17), $\tilde{p}^{-1}_{d_1\ldots d_t R}(\mathbf{x}^*)$ is (5.21) and $\mathrm{sd}(\mathbf{x}^*)$ is the square root of (5.22).

The EGO algorithm has the following steps:

1. Let $\mathbf{D}_0$ be a randomly sampled Latin-hypercube with n runs, where $n \geq 10f$ and $1/(n-1)$ is a finite decimal.

2. Let $v = 0$.

3. Let n be the number of rows in $\mathbf{D}_v$.

4. Calculate (5.7) for $q = 1, \ldots, its$ and use these predicted responses to calculate $\hat{p}^{-1}_{d_1\ldots d_t R}(\mathbf{x}_i)$ $\forall i = 1, \ldots, n$ using (5.12).

5. Fit (5.17) to the predicted responses from step 4 to estimate (5.21) and (5.22).

6. Find a new treatment $\mathbf{x}^*$ that maximises (5.23).

7. Add $\mathbf{x}^*$ to $\mathbf{D}_v$, call the new design $\mathbf{D}_{v+1}$.

8. If (5.23) is greater than $0.01 \min_{\forall i=1,\ldots,n} \hat{p}^{-1}_{d_1\ldots d_t R}(\mathbf{x}_i)$, let $v = v + 1$ and repeat from step 3. Otherwise, stop the algorithm.

The final point in this algorithm is the point which minimises the inverse of (5.12), and therefore maximises (5.12). This algorithm is significantly faster than the grid search method, but relies heavily on interpolation between observed points using Kriging.

5.4.4 Maximising the Posterior Predicted Probability of Meeting Specification

In this section, we use a grid search and the EGO algorithm to find points from inside and outside the current experimental region which maximise the posterior predicted probability of meeting specification, and therefore solve the maximisation problems given in (5.13) through (5.16).

Recall that when solving (5.13) and (5.14), we are maximising $\hat{p}_{124R}(\mathbf{x})$ and $\hat{p}_{1234R}(\mathbf{x})$, $R = 1, 2$, in the region $\mathcal{X}$, which is our current experimental region where factor levels have the range $[-1, 1]$. When solving (5.15) and (5.16), we are maximising $\hat{p}_{124R}(\mathbf{x})$ and $\hat{p}_{1234R}(\mathbf{x})$ in $\mathcal{X}^*$, which is the expanded experimental region where factor levels have the range $[-2, 2]$.

Optimal Point 1, $\mathbf{x}_1^*$=(0.27, 0.5, -1.01, 0.25, 0.84, 1) is the point inside the current region of experimentation which solves (5.13) for the grid search. Although this point was found using the grid search for Tests 1 to 4, we noticed that $\hat{p}_{124R}(\mathbf{x}_1^*)$ was close to the maximum for Tests 1, 2 and 4, for $R = 1, 2$. We also note that an equivalent optimal point, $\mathbf{x}_{1.1}^*$, was found using the EGO, where $\hat{p}_{124R}(\mathbf{x}_1^*) \approx \hat{p}_{124R}(\mathbf{x}_{1.1}^*)$ and $\hat{p}_{1234R}(\mathbf{x}_1^*) \approx \hat{p}_{1234R}(\mathbf{x}_{1.1}^*)$, $R = 1, 2$. We consider Optimal Point 1, and not $\mathbf{x}_{1.1}^*$, as a pharmaceutical product was formulated for this point.

The level of Factor 3 in $\mathbf{x}_1^*$ is less than -1, therefore $\mathbf{x}_1^*$ is in $\mathcal{X}^*$ and not $\mathcal{X}$. However, as all other factor levels are in the set $[-1, 1]$, $\mathbf{x}_1^*$ is close to the boundary of $\mathcal{X}$, and we are not concerned about the extrapolation of the model required for this point. The results from Section 5.3.2 and 5.4.1 suggest that increasing the level of Factor 3 in $\mathbf{x}_1^*$ from -1.01 to -1 will cause a small decrease in $\hat{y}_{2R}(\mathbf{x}_1^*)$, $R = 1, 2$, and hence a small decrease in $\hat{p}_{2R}(\mathbf{x}_1^*)$, $\hat{p}_{124R}(\mathbf{x}_1^*)$, and $\hat{p}_{1234R}(\mathbf{x}_1^*)$.

Optimal Point 1 was found during the Nelder-Mead optimisation within the grid search. Nelder-Mead optimisation is an unconstrained optimisation method that can lead to factor levels being found outside boundaries. The constrained optimisation algorithm given by Byrd et al. (1995) could be used to avoid this potential problem, as upper and lower limits on the levels of the factors can be set when optimising using the algorithm of Byrd et al. (1995). However, in this work we use Nelder-Mead optimisation as it is faster than Byrd et al. (1995).

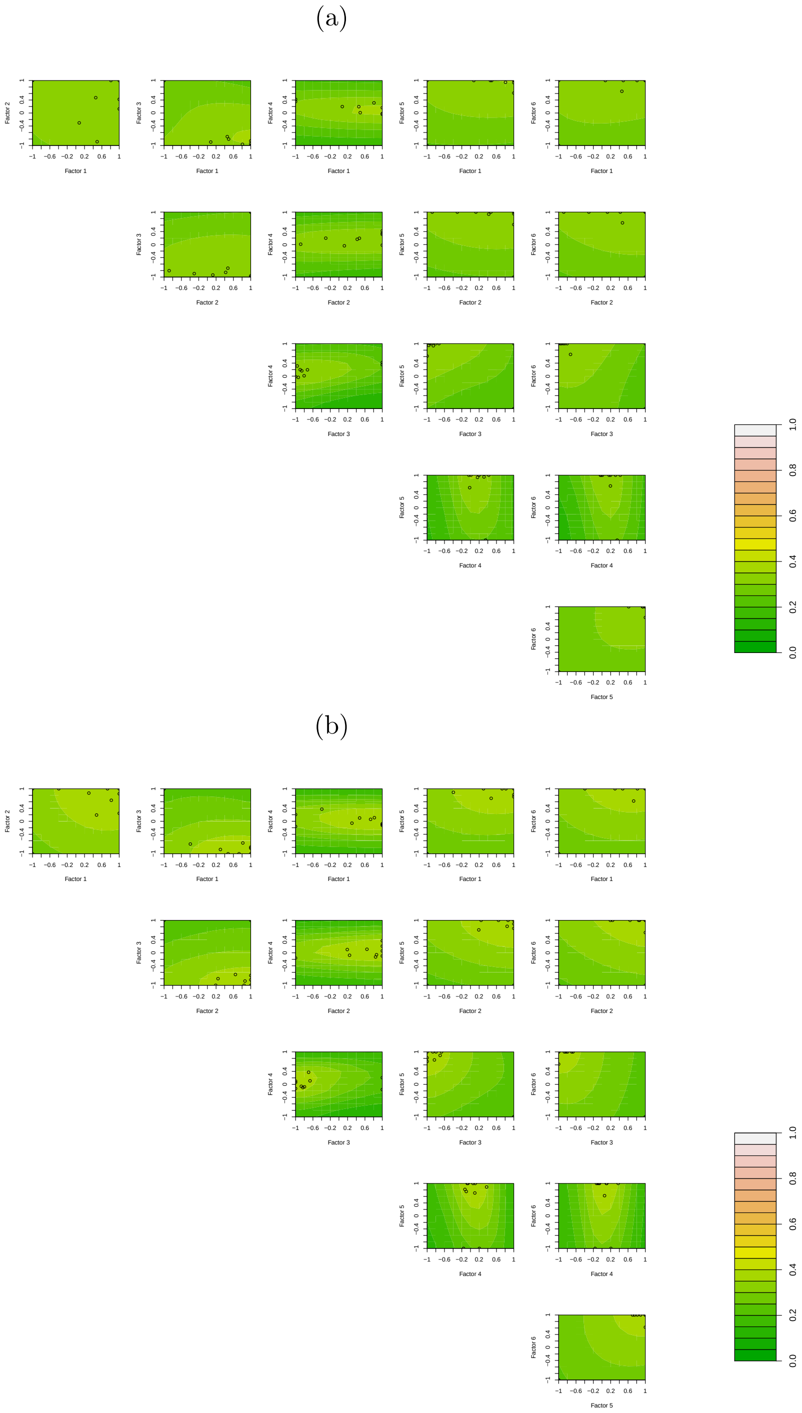
Table 5.5 gives $\hat{p}_{d_1,\ldots,d_tR}(\mathbf{x}_1^*)$ and $\hat{p}_{dR}(\mathbf{x}_1^*)$, when $d_1 \ldots d_t = 123, 1234$, $d = 1, 2, 3, 4$ and $R = 1, 2$, for the sampler with Prior 1 and Prior 2. We notice that $\hat{p}_{124R}(\mathbf{x}_1^*)$ and $\hat{p}_{1234R}(\mathbf{x}_1^*)$ for $R = 1, 2$ and both Prior 1 and Prior 2 in Table 5.5 are low. This is because $\hat{p}_{dR}(\mathbf{x}_1^*)$, $d = 1, 2, 3, 4$, $R = 1, 2$ are low for both Prior 1 and 2.

The pharmaceutical product formulated for treatment $\mathbf{x}_1^*$ failed to meet specification for all four tests, which was expected following the low predicted probabilities in Table 5.5. Although no repeats of the dissolution testing for this treatment were made to enable further assessment, our model suggests that repeated formulations of $\mathbf{x}_1^*$ would not meet specification for all four tests.

		$\mathbf{x}^* = \mathbf{x}_1^*$
Prior 1	$\hat{p}_{12341}(\mathbf{x}^*)$	0.24
	$\hat{p}_{12342}(\mathbf{x}^*)$	0.25
	$\hat{p}_{1241}(\mathbf{x}^*)$	0.36
	$\hat{p}_{1242}(\mathbf{x}^*)$	0.37
	$\hat{p}_{11}(\mathbf{x}^*)$	0.67
	$\hat{p}_{12}(\mathbf{x}^*)$	0.71
	$\hat{p}_{21}(\mathbf{x}^*)$	0.64
	$\hat{p}_{22}(\mathbf{x}^*)$	0.66
	$\hat{p}_{31}(\mathbf{x}^*)$	0.67
	$\hat{p}_{32}(\mathbf{x}^*)$	0.67
	$\hat{p}_{41}(\mathbf{x}^*)$	0.83
	$\hat{p}_{42}(\mathbf{x}^*)$	0.80
Prior 2	$\hat{p}_{12341}(\mathbf{x}^*)$	0.25
	$\hat{p}_{12342}(\mathbf{x}^*)$	0.25
	$\hat{p}_{1241}(\mathbf{x}^*)$	0.39
	$\hat{p}_{1242}(\mathbf{x}^*)$	0.40
	$\hat{p}_{11}(\mathbf{x}^*)$	0.64
	$\hat{p}_{12}(\mathbf{x}^*)$	0.66
	$\hat{p}_{21}(\mathbf{x}^*)$	0.77
	$\hat{p}_{22}(\mathbf{x}^*)$	0.79
	$\hat{p}_{31}(\mathbf{x}^*)$	0.63
	$\hat{p}_{32}(\mathbf{x}^*)$	0.64
	$\hat{p}_{41}(\mathbf{x}^*)$	0.79
	$\hat{p}_{42}(\mathbf{x}^*)$	0.77

Table 5.5: Approximate posterior probabilities (5.12) of Optimal Point 1, $\mathbf{x}_1^*$=(0.27,0.50,$-$1.01,0.25,0.84,1.00), meeting specification for the sampler with Prior 1 and Prior 2.

Figure 5.15 shows two-dimensional projections of $\hat{p}_{124R}(\mathbf{x}_1^*)$ in the locality of $\mathbf{x}_1^*$, which is estimated using Kriging, where, for each plot, the four factor not being varied are held fixed at their values in $\mathbf{x}_1^*$. The plots in Figure 5.15 show the low posterior probability of passing these three tests in the locality of $\mathbf{x}_1^*$.



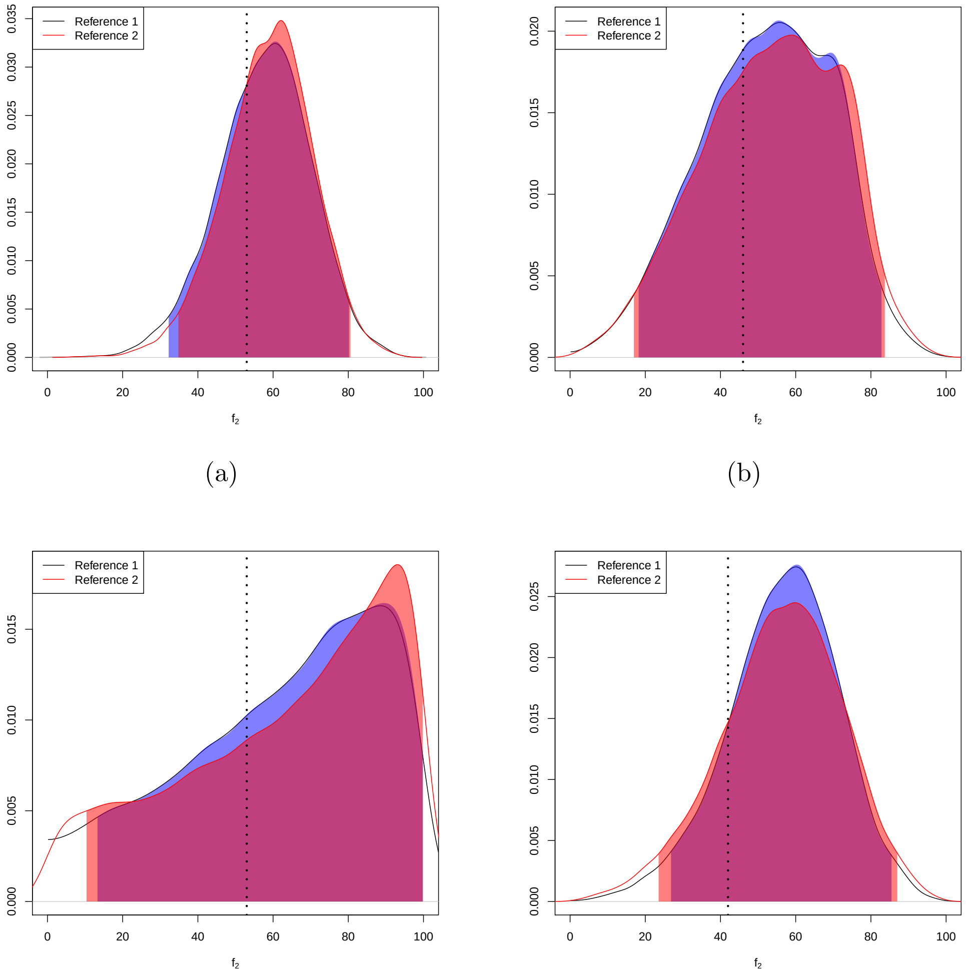
(a)

(b)

Figure 5.15: Contour plots for the approximate posterior probability (5.12) of meeting specification for Tests 1, 2 and 4 in the locality of $\mathbf{x}_1^*$=$(0.27, 0.50, -1.01, 0.25, 0.84, 1)$ using Kriging and the results of the EGO algorithm (Section 5.4.3) for: (a) Reference 1 and (b) Reference 2. The black circles are the levels for the treatments which maximise the expected improvement in the EGO algorithm. In each plot, the four factors not varied are set to their values in $\mathbf{x}_1^*$.

The posterior predicted probability surfaces around the point $\mathbf{x}_1^*$ using the grid search and Kriging are very similar. However, the surface using the results from the grid search is less smooth as the EGO algorithm uses a smoother to predict (not shown). The plots for $\hat{p}_{1234R}(\mathbf{x}_1^*)$ for both the grid search and the EGO algorithm (not shown) have a similar shape but lower probabilities.

The key role played by Factor 4 can be seen in Figure 5.15. Recall that $\boldsymbol{\beta}_4$, the main effect for Factor 4, has a high posterior probability of being non-zero for both Tests 1 and 2 (Figures 5.5(a) and 5.5(b)), and $\boldsymbol{\beta}_4$ has a high posterior probability of being negative for Test 1 and positive for Test 2 (Figure 5.9). This creates a conflict when it is required to maximise $\hat{y}_{1R}(\mathbf{x})$ and $\hat{y}_{2R}(\mathbf{x})$, $R = 1, 2$. This dichotomy leads to a narrow ridge of higher probability around x_{14}^*, which is clearly seen in the third column of Figure 5.15.



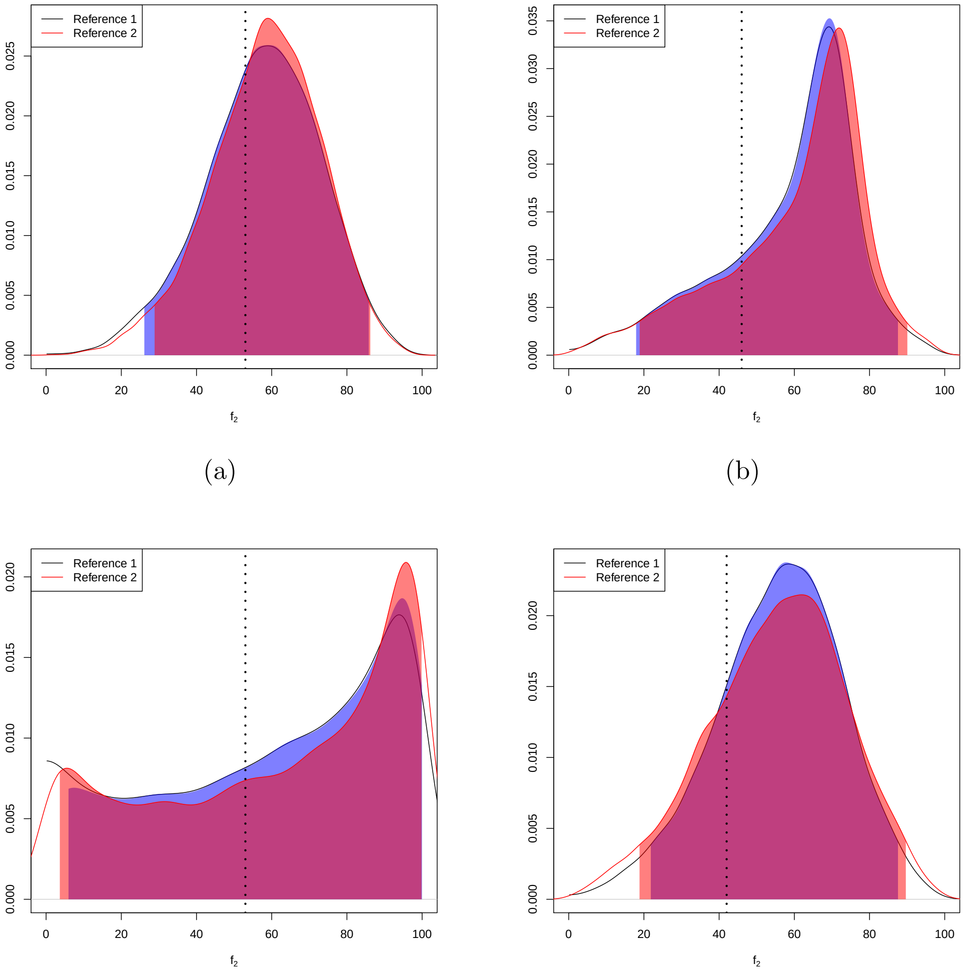
(a)

(b)

(c)

(d)

Figure 5.16: Approximate posterior predictive density for (a) $\hat{y}_{11}(\mathbf{x}_1^*)$ and $\hat{y}_{12}(\mathbf{x}_1^*)$, (b) $\hat{y}_{21}(\mathbf{x}_1^*)$ and $\hat{y}_{32}(\mathbf{x}_1^*)$, (c) $\hat{y}_{31}(\mathbf{x}_1^*)$ and $\hat{y}_{32}(\mathbf{x}_1^*)$, and (d) $\hat{y}_{41}(\mathbf{x}_1^*)$ and $\hat{y}_{42}(\mathbf{x}_1^*)$, when $\mathbf{x}_1^*$=(0.27,0.50,−1.01, 0.25, 0.84, 1.00), for sampler with Prior 1. The shaded area is the 95% posterior density interval. The dotted line is the threshold which the predicted response needs to be greater than to meet specification.



(a)

(b)

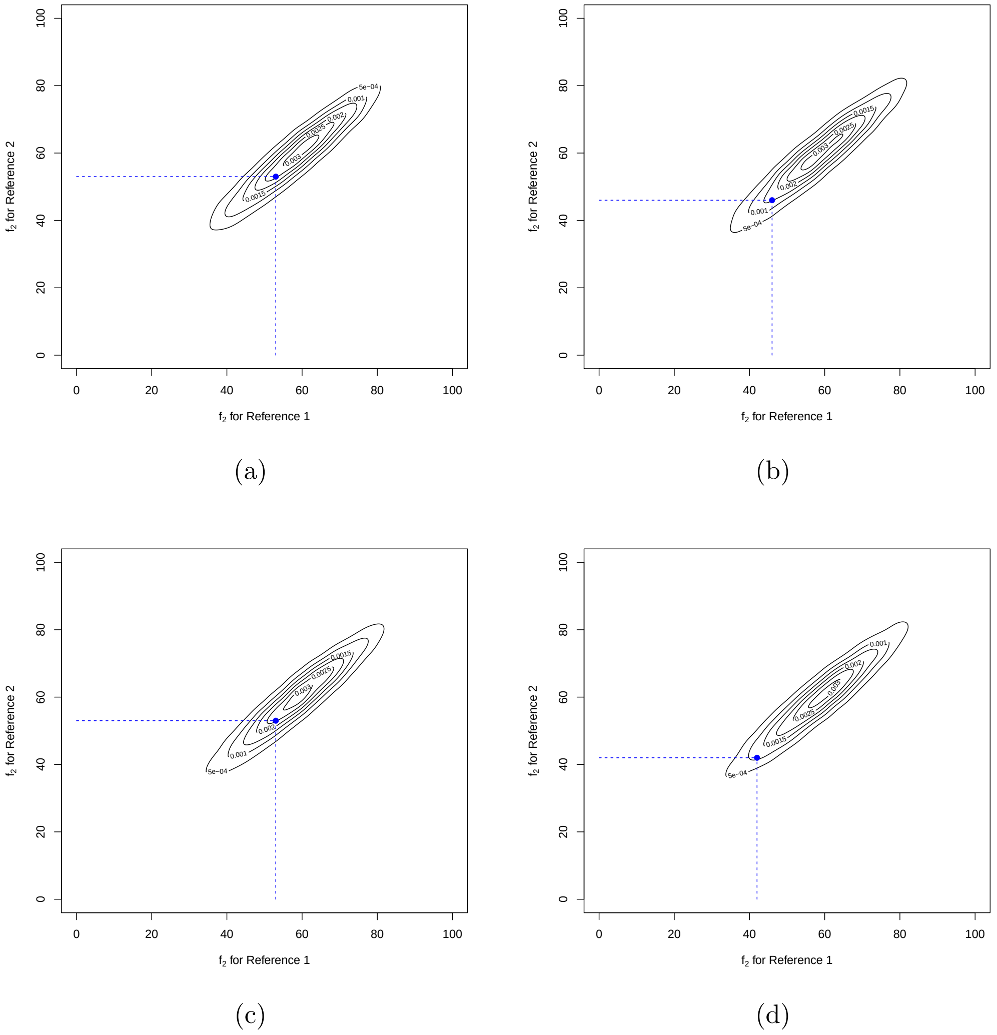
(c)

(d)

Figure 5.17: Approximate posterior predictive density for (a) $\hat{y}_{11}(\mathbf{x}_1^*)$ and $\hat{y}_{12}(\mathbf{x}_1^*)$, (b) $\hat{y}_{21}(\mathbf{x}_1^*)$ and $\hat{y}_{32}(\mathbf{x}_1^*)$, (c) $\hat{y}_{31}(\mathbf{x}_1^*)$ and $\hat{y}_{32}(\mathbf{x}_1^*)$, and (d) $\hat{y}_{41}(\mathbf{x}_1^*)$ and $\hat{y}_{42}(\mathbf{x}_1^*)$, when $\mathbf{x}_1^*$=(0.27,0.50,−1.01, 0.25, 0.84, 1.00), for the sampler with Prior 2. The shaded area is the 95% posterior density interval. The dotted line is the threshold which the predicted response needs to be greater than to meet specification.

Figures 5.16 and 5.17 give the approximate posterior predictive density of $\hat{y}_{dR}(\mathbf{x}_1^*)$, $d = 1, 2, 3, 4$, $R = 1, 2$ for the sampler with Prior 1 and Prior 2, respectively. The threshold for meeting specification for each test is given in these figures as dotted black line. The shaded areas in these figures are the 95% highest posterior density intervals for each reference.

We note that a significant proportion of the densities in Figures 5.16 and 5.17 are to the left of the threshold. This explains why the posterior predicted probabilities of meeting specification for various tests are low. We also note that there is a significant amount of uncertainty in the predicted responses, demonstrated by the width of the highest posterior density intervals.



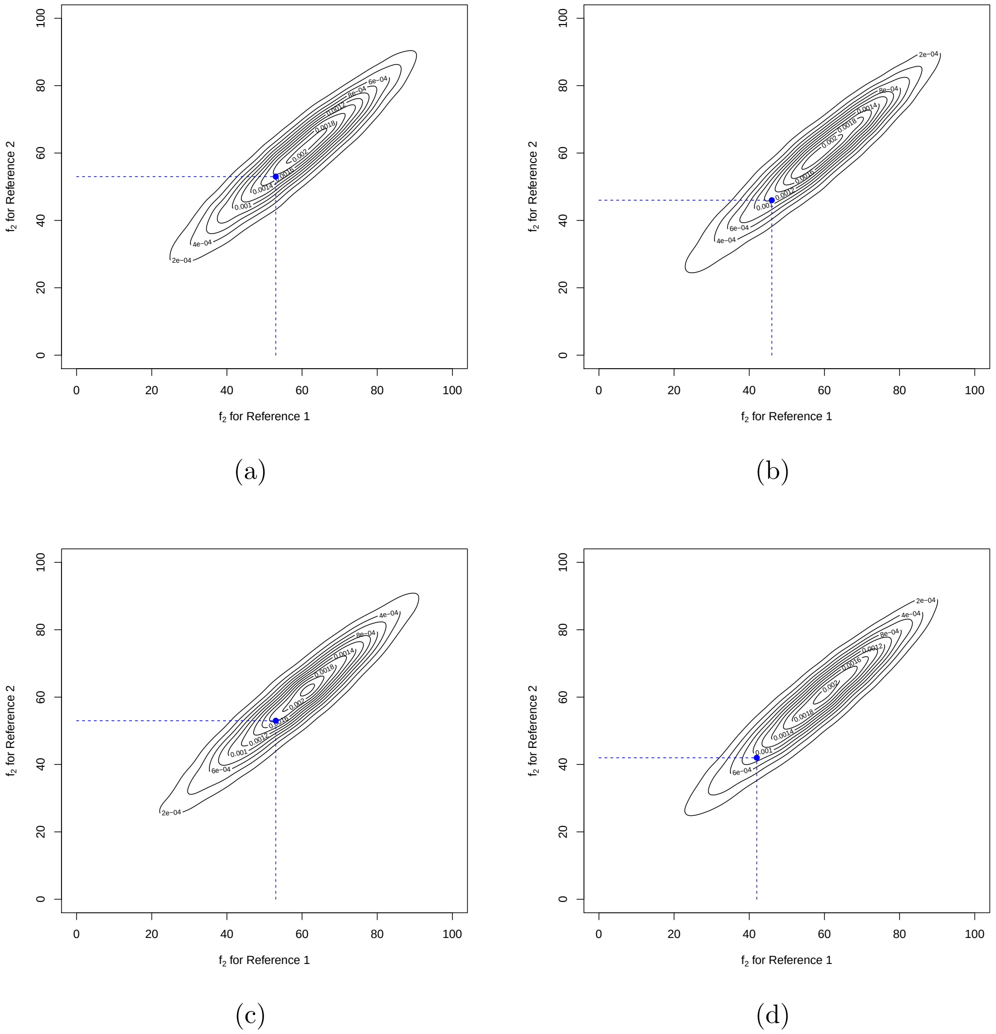
(a)

(b)

(c)

(d)

Figure 5.18: Approximate bivariate posterior predictive density for (a) $\hat{\mathbf{y}}_1(\mathbf{x}_1^*)$, (b) $\hat{\mathbf{y}}_2(\mathbf{x}_1^*)$, (c) $\hat{\mathbf{y}}_3(\mathbf{x}_1^*)$, and (d) $\hat{\mathbf{y}}_4(\mathbf{x}_1^*)$, when $\mathbf{x}_1^*$=(0.27,0.50,−1.01, 0.25, 0.84, 1.00) for the sampler with Prior 1. Any predicted f_2 values greater than the blue dotted lines pass the dissolution test.



(a)

(b)

(c)

(d)

Figure 5.19: Approximate bivariate posterior predictive density for (a) $\hat{\mathbf{y}}_1(\mathbf{x}_1^*)$, (b) $\hat{\mathbf{y}}_2(\mathbf{x}_1^*)$, (c) $\hat{\mathbf{y}}_3(\mathbf{x}_1^*)$, and (d) $\hat{\mathbf{y}}_4(\mathbf{x}_1^*)$, when $\mathbf{x}_1^*$=(0.27,0.50,−1.01, 0.25, 0.84, 1.00) for the sampler with Prior 2. Any predicted f_2 values greater than the blue dotted lines pass the dissolution test.

Figures 5.18 and 5.19 are contour plots for the approximate bivariate posterior predictive densities for $\hat{\mathbf{y}}_d(\mathbf{x}_1^*) = (\hat{y}_{d1}(\mathbf{x}_1^*), \hat{y}_{d2}(\mathbf{x}_1^*))$ $d = 1, 2, 3, 4$, for the sampler with Prior 1 and 2, respectively. The blue point is the threshold, and the dotted blue lines are the thresholds for the two references. We note that there is a significant proportion of the density that lies in the bottom left hand corner of these plots which is enclosed by the blue dotted lines. This provides further explanation for the low probabilities of passing various tests.

The experimenters were keen to explore outside the current region of experimentation, as the probabilities in Figure 5.15 increase to the maximum at the boundaries of $\mathcal{X}$, and therefore the model suggests that the probabilities of passing tests may be maximised outside the current experimental region. Ideally, further experimentation would be used to choose this point, for example a form of hill climbing or steepest ascent (Mee and

Xiao, 2008; Edwards and Fuerte, 2011). However, due to time and resource constraints the current model was used to indicate which points outside the experimental region may be promising.

The point found by grid search which maximises (5.15) and (5.16) is $\mathbf{x}^*_{2.1}$=(0.72, 1.36, -2.16, -0.32, 2.40, 1.41). However, we were then informed by the experimenters that Factor 6 could not be set to 1.41. A more preferable level for Factor 6, when considering the levels of the other factors in $\mathbf{x}^*_{2.1}$, would be 0.5. We therefore compare the posterior predicted probabilities of $\mathbf{x}^*_{2.1}$=(0.72, 1.36, -2.16, -0.32, 2.40, 1.41) and $\mathbf{x}^*_{2.2}$=(0.72, 1.36, -2.16, -0.32, 2.40, 0.5).

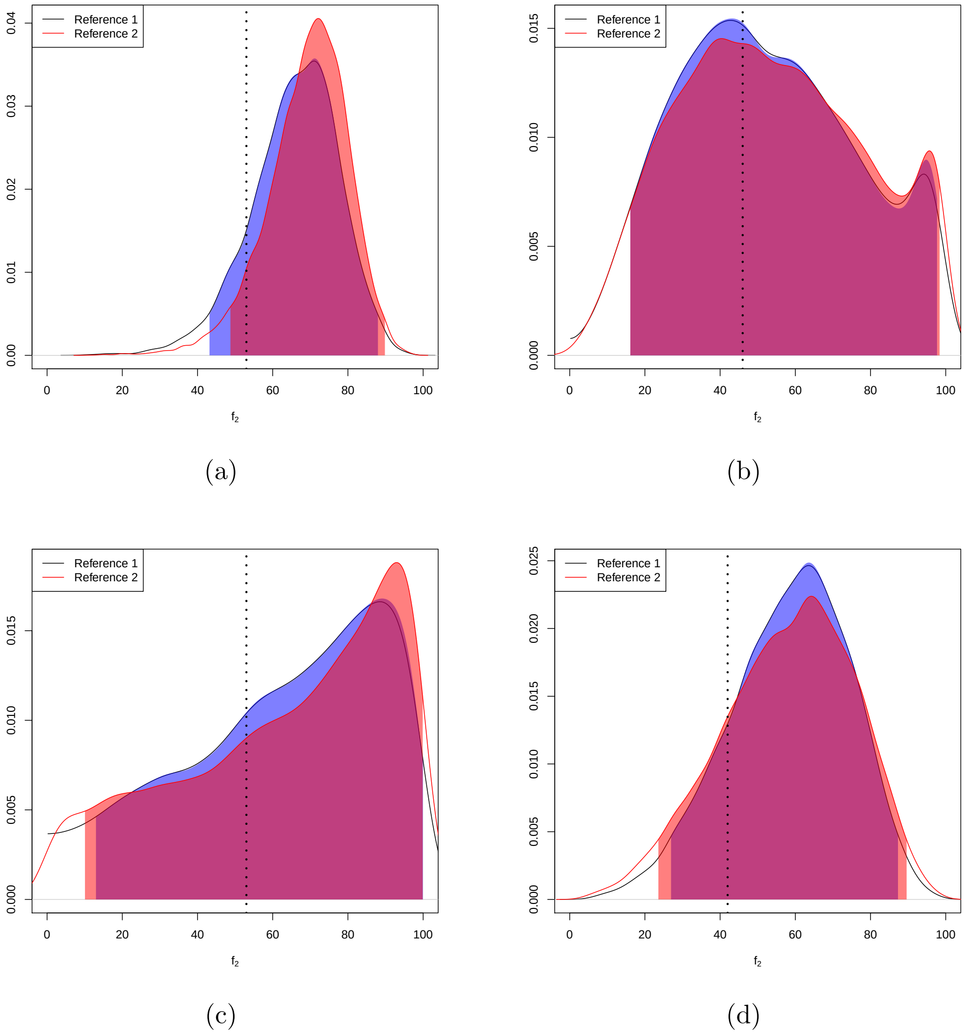
		$\mathbf{x}^* = \mathbf{x}^*_{2.1}$	$\mathbf{x}^* = \mathbf{x}^*_{2.2}$
Prior 1	$\hat{p}_{12341}(\mathbf{x}^*)$	0.29	0.27
	$\hat{p}_{12342}(\mathbf{x}^*)$	0.31	0.29
	$\hat{p}_{1241}(\mathbf{x}^*)$	0.44	0.42
	$\hat{p}_{1242}(\mathbf{x}^*)$	0.46	0.44
	$\hat{p}_{11}(\mathbf{x}^*)$	0.87	0.87
	$\hat{p}_{12}(\mathbf{x}^*)$	0.93	0.93
	$\hat{p}_{21}(\mathbf{x}^*)$	0.60	0.58
	$\hat{p}_{22}(\mathbf{x}^*)$	0.62	0.60
	$\hat{p}_{31}(\mathbf{x}^*)$	0.67	0.66
	$\hat{p}_{32}(\mathbf{x}^*)$	0.67	0.67
	$\hat{p}_{41}(\mathbf{x}^*)$	0.84	0.84
	$\hat{p}_{42}(\mathbf{x}^*)$	0.81	0.81
Prior 2	$\hat{p}_{12341}(\mathbf{x}^*)$	0.32	0.31
	$\hat{p}_{12342}(\mathbf{x}^*)$	0.33	0.32
	$\hat{p}_{1241}(\mathbf{x}^*)$	0.50	0.48
	$\hat{p}_{1242}(\mathbf{x}^*)$	0.53	0.51
	$\hat{p}_{11}(\mathbf{x}^*)$	0.83	0.82
	$\hat{p}_{12}(\mathbf{x}^*)$	0.88	0.87
	$\hat{p}_{21}(\mathbf{x}^*)$	0.76	0.74
	$\hat{p}_{22}(\mathbf{x}^*)$	0.77	0.75
	$\hat{p}_{31}(\mathbf{x}^*)$	0.62	0.62
	$\hat{p}_{32}(\mathbf{x}^*)$	0.63	0.62
	$\hat{p}_{41}(\mathbf{x}^*)$	0.80	0.80
	$\hat{p}_{42}(\mathbf{x}^*)$	0.77	0.77

Table 5.6: Approximate posterior probabilities (5.12) of Optimal Point 2.1, $\mathbf{x}^*_{2.1}$=(0.72,1.36,−2.16,−0.32,2.40,1.41), and Optimal Point 2.2, $\mathbf{x}_{2.2}$=(0.72,1.36,−2.16,−0.32,2.40,0.50), meeting specification for the sampler with Prior 1 and Prior 2.

Table 5.6 gives $\hat{p}_{d_1 \ldots d_t R}(\mathbf{x}^*)$ and $\hat{p}_{dR}(\mathbf{x}^*)$ when $\mathbf{x} = \mathbf{x}^*_{2.1}, \mathbf{x}^*_{2.2}$, $d_1 \ldots d_t = 124, 1234$, $d = 1, 2, 3, 4$ and $R = 1, 2$ for the sampler with Prior 1 and Prior 2. We note from

Table 5.6 that changing the level of Factor 6 does not have a significant impact on the probabilities of meeting specification. This is to be expected, as Factor 6 was not identified as an important factor in the analysis in Section 5.3.2.

A pharmaceutical product for $\mathbf{x}^*_{2.2}$ was formulated and dissolution tested, but did not meet specification for all four tests. This was not surprising, as the probability of $\mathbf{x}_{2.2}$ meeting specification for each test individually and all four tests was low. Our model also suggests that $\mathbf{x}^*_{2.2}$ would fail to meet specification if it was repeated, however as repeated measures were not taken, further assessment cannot be made.



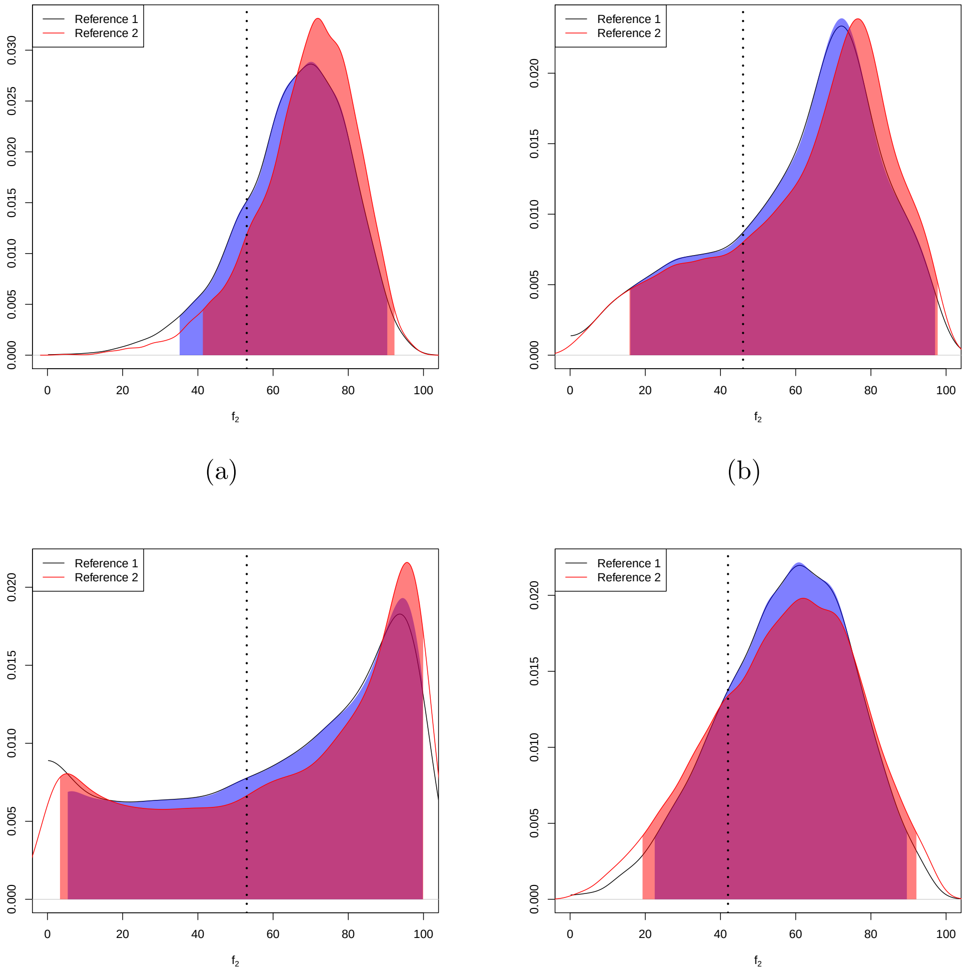
(a)

(b)

(c)

(d)

Figure 5.20: Approximate posterior predictive density for (a) $\hat{y}_{11}(\mathbf{x}^*_{2.2})$ and $\hat{y}_{12}(\mathbf{x}^*_{2.2})$, (b) $\hat{y}_{21}(\mathbf{x}^*_{2.2})$ and $\hat{y}_{32}(\mathbf{x}^*_{2.2})$, (c) $\hat{y}_{31}(\mathbf{x}^*_{2.2})$ and $\hat{y}_{32}(\mathbf{x}^*_{2.2})$, and (d) $\hat{y}_{41}(\mathbf{x}^*_{2.2})$ and $\hat{y}_{42}(\mathbf{x}^*_{2.2})$, when $\mathbf{x}^*_{2.2}$=(0.72,1.36,−2.16, −0.32, 2.40, 0.50), for the sampler with Prior 1. The shaded area is the 95% posterior density interval. The dotted line is the threshold which the prior needs to be greater than to meet specification.



(a)

(b)

(c)

(d)

Figure 5.21: Approximate posterior predictive density for (a) $\hat{y}_{11}(\mathbf{x}^*_{2.2})$ and $\hat{y}_{12}(\mathbf{x}^*_{2.2})$, (b) $\hat{y}_{21}(\mathbf{x}^*_{2.2})$ and $\hat{y}_{32}(\mathbf{x}^*_{2.2})$, (c) $\hat{y}_{31}(\mathbf{x}^*_{2.2})$ and $\hat{y}_{32}(\mathbf{x}^*_{2.2})$, and (d) $\hat{y}_{41}(\mathbf{x}^*_{2.2})$ and $\hat{y}_{42}(\mathbf{x}^*_{2.2})$, when $\mathbf{x}^*_{2.2}$=(0.72,1.36,−2.16, −0.32, 2.40, 0.50), for the sampler with Prior 2. The shaded area is the 95% posterior density interval. The dotted line is the threshold which the predicted response needs to be greater than to meet specification.

Figures 5.20 and 5.21 give the approximate posterior predictive density of $\hat{y}_{dR}(\mathbf{x}^*_{2.2})$, $d = 1, 2, 3, 4$, $R = 1, 2$ for the sampler with Prior 1 and 2, respectively. These figures provide additional evidence for the low probabilities in Table 5.6, as a large proportion of both the density and highest posterior density intervals are less than the threshold. There is also large variability in the predictions.

5.5 Discussion

In this chapter, we discussed the design of an experiment for the formulation of a pharmaceutical product, the modelling of the f_2 values from dissolution testing and the prediction of points which maximise the probability of passing specification. Using initial information, we described the experiment as a two-stage split-plot design with five factors in the first stage, two of which were hard to change, and one factor in the second stage. Due to experimental constraints, our experiment could only have sixteen factorial runs.

We used the coordinate exchange algorithm and compound Bayesian D-optimality criterion to find the sixteen run two-stage split-plot design given in Table 5.2, which is suitable for the formulation of this pharmaceutical product. We noted in Section 5.2.1 that the number of correlated columns in the matrices for the models fitted to two-stage split-plot designs is not affected by the weight matrix in the compound Bayesian D-optimality objective function. We chose the design in Table 5.2, as the terms which were correlated were preferable.

The final experiment was not two-stage, however, as a response was not measured after the first stage. We therefore also found a sixteen run single-stage split-plot design, and compared this design to the optimal two-stage design using the correlation matrix for the columns in the matrix for the cumulative model in Section 5.2.3. The single-stage split-plot design had correlation between the columns involving Factors 1 to 5, whereas the two-stage split-plot design only had correlation between columns involving Factor 6. As the scientists' prior assumption was that Factors 1 to 5 are more important than Factor 6, the two-stage split-plot design was preferred to the single-stage split-plot design. Therefore, the two-stage split-plot design in Table 5.2 was used to formulate pharmaceutical products.

Dissolution testing was performed after the pharmaceutical products were formulated. The pharmaceutical products were dissolved in four different media, and f_2 values were calculated for two references. We therefore had bivariate data for each of the four dissolution tests for the sixteen points in the optimal two-stage split-plot design (Table 5.2) and two centre points.

In Section 5.3.2, Bayesian variable selection was used to identify the influential terms in the models fitted to the f_2 values for these four dissolution tests. During this analysis, we noted that the bivariate responses for Tests 1 and 2 are easier to model than the responses for Tests 3 and 4. Also, we noted that the main effect of Factor 4 has a high posterior probability of being active for Tests 1 and 2. However, the main effect for Factor 4 has opposing effect on these two tests, as it is negative for Test 1 and positive for Test 2.

The diagnostic plots in Appendix B.2 suggest that there are some problems with the fit of the models considered in this chapter. However, this was expected as this experiment is a screening experiment and is supersaturated, and the aim of this experiment was to identify the influential terms and not produce an accurate and precise model. Future experimentation, using the influential factors identified using Bayesian variable selection, would enable more detailed models to be made.

The posterior predicted responses can be used to approximate the probability of meeting specification, where a point $\mathbf{x}$ meets the specification for dissolution Test d and Reference R if $\hat{y}_{dR}(\mathbf{x}) \geq \tau_d$. We used both a grid search (Section 5.4.2) and the EGO algorithm (5.4.3) to find the point which optimises the probability of meeting specification for Tests 1,2,3 and 4 both inside and outside of the current experimental region.

The optimal points inside and outside of the current experimental region discussed in Section 5.4.4 had a low probability of meeting specification for Tests 1, 2, 3 and 4. Single formulations of these optimal points failed to meet specification for all four tests. The predicted probability surface showed a ridge for Factor 4, due to the conflict in signs for the main effect of this factor in Tests 1 and 2.

We could use the results from this chapter to perform further experimentation. Future work could therefore focus on extending and adapting existing methodology to exploring outside the current experimental region. Both Mee and Xiao (2008) and Edwards and Fuerte (2011) discuss methods of optimisation using steepest ascent or compromise ascent after screening for multiple response experiments. Mee and Xiao (2008) use Pareto optimality, and prove that only search directions that are convex combinations of paths of steepest ascent should be used to optimise multiple responses from screening experiments. Edwards and Fuerte (2011) use Bayesian reliabilities to identify compromise directions for exploration of design spaces for multiple responses. The approaches in these papers could be adapted and extended for our experiment.

Chapter 6

Conclusions and Future Work

6.1 Conclusions

The focus of this thesis is the design and analysis of factorial experiments in blocks and stages, using motivating examples from optoelectronic engineering and chemistry.

In Chapter 2 we found D-optimal block designs for linear mixed models with random block effects and autocorrelated errors appropriate for our motivating example of the manufacture of microstructured optical fibres. The designs were found using coordinate exchange and interchange algorithms

The designs found in Chapter 2 using both algorithms were robust to misspecification of the autocorrelation parameter, ρ, and the ratio of variances, η, which are unknown prior to experimentation. However, completely ignoring the correlation structure or blocks leads to a loss in efficiency. We also noted that designs found using the coordinate exchange algorithm had a higher D-optimality objective function value than designs found using the interchange algorithm, which assigns treatments from the D-optimal unblocked design to blocks. We also found a number of equivalent D-optimal designs, that is designs with the same value of the D-optimality objective function. These equivalent designs did not necessarily have the same treatments, allocation or ordering of treatments.

An example from chemistry, the formulation and dissolution testing of a pharmaceutical product, was the motivation for the work in Chapters 3 to 5. In Chapter 3 we discussed the design of optimal multi-stage designs with potentially restricted randomisation suitable for the formulation of a pharmaceutical product. We define a multi-stage experiment as an experiment which applies sub-treatments to the same experimental unit at multiple stages, and measures a distinct response after the sub-treatments are applied in each stage. We found optimal multi-stage designs for a compound Bayesian D-optimality criterion using the coordinate exchange algorithm with both random starting designs and designs with good projection properties (Cheng, 2006;

Loeppky et al., 2007).

In general, the model matrices for the three models fitted to the two responses for the optimal two-stage split- and strip-plot designs found in Chapter 3 had more correlated columns than two-stage completely randomised designs. Therefore, the variance and bias of more parameters are inflated in two-stage split- and strip-plot designs than in two-stage completely randomised designs. We also found that using the coordinate exchange algorithm and designs with good projection properties as starting designs can find two-stage completely randomised designs with efficiencies of approximately 100%, but can have significantly more correlated model matrix columns.

In Chapter 4, we presented a method of Bayesian variable selection for multivariate responses from supersaturated split-plot experiments, which was motivated by the dissolution testing of a pharmaceutical product. We motivate our Bayesian variable selection method through an initial comparison of frequentist and Bayesian analysis methods for simulated responses from the two-stage split-plot design found in Chapter 3. We found that the frequentist approach is affected by the difficulty of estimating variance components. Therefore, we concluded that Bayesian methodology should be used to analyse the supersaturated multi-stage split-plot experiment with multivariate responses, agreeing with the conclusions of Gilmour and Goos (2009).

In Chapter 4 we introduce a Metropolis-Hastings within Gibbs sampling algorithm, which generates dependent samples from the posterior distribution of the parameters in the linear mixed effects model for multivariate responses. Samples from the algorithm can be used to perform variable selection, estimate model parameters and predict responses. We demonstrated the performance of this algorithm, which extends the work of Geweke (1996) and Tan and Wu (2013), for data simulated for a supersaturated multi-stage split-plot design.

In Chapter 5, we presented a case study from our collaboration with GlaxoSmithKline, and applied the methodology developed in Chapters 3 and 4 to the formulation and dissolution testing of a pharmaceutical product. We also detailed how we located a formulation that maximised the estimated probability of passing particular specifications using both a grid search and the EGO algorithm (Jones et al., 1998). Both approaches use samples from the Metropolis-Hastings within Gibbs sampling algorithm introduced in Chapter 4. The work in Chapter 5 demonstrates the impact of using our methodology in industry.

6.2 Future Work

Possible extensions to the work in this thesis are discussed in Sections 2.7, 3.7, 4.6 and 5.5. In Section 6.2.1, we present a summary of some general extensions that could be considered. In Section 6.2.2, we consider a more specific extension which develops methodology from the recent literature.

6.2.1 General Extensions

Optimality Criteria

D-optimality is particularly appropriate when the aim of the experiment is to gain scientific understanding through estimation of the fixed effect parameters (Goos, 2002). As this is the aim of the collaborative research presented in this thesis, we have used the D-optimality objective function, (1.18), to find optimal block designs for autocorrelated errors in Chapter 2. We have also used a compound Bayesian D-optimality objective function, (3.20), to find optimal multi-stage designs in Chapter 3.

Alternative optimality criteria can be used to find designs for estimation of the fixed effect parameters. For example, A-optimality, which minimizes the average variance of the parameters estimates, is advocated by a number of authors. Gilmour and Trinca (2012) recommended using a new criterion based on A-, and not D-, optimality for the estimation of fixed effect parameters. The A-optimality objective function allows the terms in the model to be weighted according to their perceived importance prior to experimentation, whereas the D-optimality objective function does not.

Considering A-optimality would allow us to ascertain whether our D-optimal block and multi-stage designs perform well with respect to alternative criteria. It would also allow us to asses whether the objective function used in the coordinate exchange algorithm influences the treatment allocations and robustness to ρ and η seen for the block designs in Chapter 2, or alters the correlation between columns in the model matrices considered for the multi-stage designs in Chapter 3.

In Chapter 4 we found that frequentist variable selection methods are dependent on the estimates of the variance components. We considered that this may be because the compound criteria we used to find designs is based on D-optimality, and therefore the primary aim of these designs is the estimation of the fixed effects and not the variance components. We could find compound V-optimal designs, which minimise the predicted variance over the design region, and compare the designs found for fixed effect and variance component estimation. A new criterion for split-plot designs which balances the objectives of fixed effect and variance component estimation has been presented by Mylona et al. (2014). This criterion is discussed in detail in Section 6.2.2.

Range of Parameters

The range of ρ and η values considered when assessing the robustness of the block designs found in Chapter 2 to misspecification of ρ and η was quite wide, but the number of specific values considered was not very large. Also, only a small number of $\mathbf{w}$ vectors were considered when assessing the robustness of the two-stage split-plot designs to misspecification of $\mathbf{w}$ in (3.20) in Chapter 5.

These values and vectors were chosen in order to obtain general results within our computational boundaries. However, if we were able to obtain past experimental data for the block designs in Chapter 2, and used the results from the experiment in Chapter 5, we could use this prior information to estimate ρ and η, and select an appropriate $\mathbf{w}$ vector.

Algorithms for Design and Analysis

In Chapters 2, 3 and 5 we used a coordinate exchange algorithm to find optimal designs, as it can be easily adapted to different design structures and objective functions. However, the coordinate exchange algorithm is a 'greedy' algorithm, which only accepts moves which increase the objective function value and can therefore get stuck at local optima.

Alternative stochastic algorithms, such as simulated annealing (Aarts and van Laarhoven, 1989; Brooks and Morgan, 1995), accept or reject moves based on a certain probability. This use of an acceptance probability allows these algorithms to escape local optima. Hence, stochastic algorithms may find designs with higher objective function values. Using two (or more) algorithms to find optimal designs would allow the comparison of these designs based on properties such as objective function values, treatment allocation and correlation between columns in the model matrix.

In Chapters 4 and 5 we use samples from a Metropolis-Hastings with Gibbs sampling algorithm to perform variable selection, estimate model parameters and predict multivariate responses. We used this algorithm as both Gibbs and Metropolis-Hastings sampling are well known Bayesian methodologies which are popular in literature. However, there are alternative methods of variable selection using MCMC algorithms, such as the reversible jump MCMC algorithm proposed by Green (1995).

The reversible jump MCMC algorithm provides a method of sampling from the joint posterior of a model indicator and model parameters, without knowledge of the dimension of these vectors. Variable selection can therefore be performed using samples from this algorithm without having to know the maximum number of parameters, p, in the model, which required in our Metropolis-Hastings within Gibbs sampling algorithm. The reversible jump MCMC algorithm searches spaces for models with different numbers of parameters. However, the rules which define when jumps are made have to be determined, which adds extra complexity when initialising the algorithm.

Prior and Proposal Distributions

The prior distributions presented in Chapter 4 were multivariate extensions of those given by Tan and Wu (2013). However, other prior distributions may be more appropriate for other motivating examples. If alternative prior distributions were considered, the robustness of the variables selected, parameters estimated and responses predicted to these different prior distributions could be assessed.

A starting point for this future work could be to consider the mixture of normal prior distribution for fixed effect parameters discussed in Box and Meyer (1986) and Gilmour and Goos (2009). Using this prior would allow us to extend the stochastic search variable selection (SSVS) algorithm discussed by George and McCulloch (1993, 1997); Brown et al. (1998) and Chipman et al. (2001). We could then compare the variables selected, parameters estimated and responses predicted using samples from our Metropolis-Hastings within Gibbs sampling algorithm and the extended SSVS algorithm.

In Appendix C, we note that the effective sample size of the samples of correlation parameter ϕ, which is sampled using Metropolis-Hastings rejection sampling, is small and hence the current Metropolis-Hastings within Gibbs sampling algorithm is not very efficient. Therefore, considering a different proposal distribution for the Metropolis-Hastings step in the algorithm presented in Chapter 4 would be beneficial. Any alternative proposal distributions considered need to have the support [0,1], therefore possible alternatives include beta distributions with different shape and scale parameters and the uniform distribution. Proposal distributions which adapt as ϕ is sampled could also be considered.

6.2.2 Designs for Estimating Both Fixed Effects and Variance Components

In Chapter 4, the reliance of the frequentist variable selection method on the difficult to estimate variance components was given as the main reason for pursuing Bayesian variable selection methods. Estimating the variance components is important to frequentist variable selection, as the model selection criteria given in Section 4.3.2 of Chapter 4 depend on these estimates through the calculation of the maximised log likelihood (4.7).

The compound Bayesian D-optimality criterion used to find the two-stage split-plot design analysed in Chapter 4 only considers the estimation of the fixed effects and not the variance components. Therefore, it may be possible that the performance of these frequentist analysis methods could be improved by considering designs found using an optimality criterion which accounts for estimation of both the fixed effects and the variance components.

Mylona et al. (2014) introduced an objective function for finding split-plot designs with the aims of fixed effect and variance component estimation. In this section, we will use a compound version of this objective function to find a two-stage split-plot design. We will compare this design to the two-stage design found using compound Bayesian D-optimality objective function in Chapter 3. We will also assess the performance of the frequentist analysis method considered in Chapter 4, all subsets regression, for data simulated from this new design.

The objective function maximised by Mylona et al. (2014) is

$$\phi_{MGJ} = \frac{\alpha}{p} \log|\phi_D| + \frac{1-\alpha}{2} \log|\mathbf{N}| \tag{6.1}$$

where $0 < \alpha \leq 1$, ϕ_D is the D-optimality criteria, (1.18), and

$$\mathbf{N} = \frac{1}{2} \begin{pmatrix} \mathrm{tr}\left((\mathbf{PZZ}^T)^2\right) & \mathrm{tr}(\mathbf{P}^2\mathbf{ZZ}^T) \\ \mathrm{tr}(\mathbf{P}^2\mathbf{ZZ}^T) & \mathrm{tr}(\mathbf{P}^2) \end{pmatrix} \tag{6.2}$$

is a general expression for the information matrix for estimating the variance components using REML (Section 1.3.2, Patterson and Thompson, 1971; Harville, 1977). In (6.2), $\mathbf{P} = \mathbf{V}^{-1} - \mathbf{V}^{-1}\mathbf{X}(\mathbf{X}^T\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{V}^{-1}$, when $\mathbf{V}$ is the variance covariance matrix for responses modelled using the linear mixed effects model (1.3).

Mylona et al. (2014) recommended using a Bayesian approach, and placed a prior on $\rho = \sigma_\gamma^2/(\sigma_\epsilon^2 + \sigma_\gamma^2)$. The Bayesian composite objective function is

$$\Phi_{MGJ} = \int_0^1 \phi_{MGJ}(\rho)p(\rho)d\rho, \tag{6.3}$$

where $p(\rho)$ is the prior distribution for ρ. Mylona et al. (2014) assumed that $\rho \sim \beta(1,1) \equiv U(0,1)$. Hence $p(\rho) \propto 1$ and

$$\Phi_{MGJ} \propto \int_0^1 \phi_{MGJ}(\rho)d\rho. \tag{6.4}$$

Evaluation of (6.4) requires numerical approximation. Mylona et al. (2014) used Gauss-Jacobi quadrature (Appendix A) to approximate (6.4) as

$$\Phi_{MGJ} \approx \frac{1}{2}\sum_{g=1}^{n_a} w_g^{GJ} \phi_{MGJ}\left(\frac{a_g^{GJ}+1}{2}\right) \tag{6.5}$$

where the n_a abscissas a_g^{GJ}, and corresponding weights w_g^{GJ}, $g = 1, \ldots, n_a$, are obtained from the Jacobi polynomial. An appropriate n_a is chosen by evaluating (6.5) for a representative, fixed, design for multiple n_a values and selecting the smallest n_a at which (6.5) stabilises. The value at which n_a stabilises is the n_a where, as n_a increases, the difference in (6.5) is small.

The design considered in Section 4.3 of Chapter 4 is a two-stage split-plot design. The objective function (6.5) is for single-stage split-plot designs and therefore needs to be extended for multi-stage designs. The compound optimality objective function we use to find multi-stage split-plot designs with the objectives of fixed effect and variance component estimation is

$$D_{MGJ} = \sum_{l=1}^{m} \frac{w_l}{p_l} \Phi_l^*, \tag{6.6}$$

where

$$\Phi_l^* \approx \frac{1}{2} \sum_{g=1}^{n_a} w_g^{GJ} \phi_l^* \left(\frac{a_g^{GJ} + 1}{2} \right) \tag{6.7}$$

and

$$\phi_l^*(\rho) = \frac{\alpha}{p_l} \log |\mathbf{X}_l \mathbf{V}_\rho^{-1} \mathbf{X}_l + \mathbf{R}_l| + \frac{1-\alpha}{2} \log |\mathbf{N}_l|. \tag{6.8}$$

The objective function (6.6) is an extension of the compound Bayesian D-optimality objective function, (3.20), presented in Section 3.4.2 of Chapter 3. Recall that $\mathbf{X}_l$ is the model matrix for design $\mathbf{D}$ for model $l = 1, \ldots, m$, $\mathbf{V}_\rho$ is the variance covariance matrix for the responses from the experiment, which is dependent on ρ, $\mathbf{R}_l$ is the prior precision matrix for model l, p_l is the number of parameters in model l and $\mathbf{N}_l$ is (6.2) calculated using $\mathbf{X}_l$ and $\mathbf{V}_\rho$. We use the Bayesian D-optimality objective function in (6.6) as it allows us to consider our prior uncertainty about $\boldsymbol{\beta}$ and find supersaturated designs.

Note that (6.6) requires the choice of n_a and α. We found the smallest value of n_a for which (6.6) is stable is 12 for a two-stage split-plot design. In their paper, Mylona et al. (2014) suggested $\alpha \in [0.5, 0.75]$, and we set $\alpha = 0.75$ in order to give more weight to the estimation of the fixed effect parameters, which has been the primary aim throughout this thesis.

We used (6.6) with $n_a = 12$ and $\alpha = 0.75$ as the objective function, ϕ, in the coordinate exchange algorithm given in Section 3.5.1 of Chapter 3 to find a sixteen-run two-stage split-plot design, given in Table 6.1.

We note that the level of Factor 2 is constant in the design in Table 6.1, and the design also has two repeated whole plot treatments. The constant level of Factor 2 means the main effect for this factor is aliased with the interaction. The repeated whole-plot treatments will improve the estimation of the variance components, however they restrict the number of combination of the levels of Factors 1 and 2 considered and hence impact on our ability to estimate the parameters related to these factors.

Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
1	1	-1	1	-1	1
1	1	-1	-1	1	1
1	1	1	1	1	-1
1	1	1	-1	-1	-1
-1	1	-1	-1	1	-1
-1	1	1	-1	-1	1
-1	1	1	1	1	1
-1	1	-1	1	-1	-1
1	1	-1	1	1	1
1	1	1	1	-1	-1
1	1	-1	-1	-1	1
1	1	1	-1	1	-1
-1	1	1	1	-1	1
-1	1	-1	-1	-1	-1
-1	1	1	-1	1	1
-1	1	-1	1	1	-1

Table 6.1: 16 run two-stage split-plot design found using the coordinate exchange algorithm in Section 3.5.1 of Chapter 3 with (6.6).

Figure 6.1 is the heat map for the correlations between columns in the matrix for the model including the intercept, main effects and pairwise products of all six factors. The column correlation matrix, (3.21), is discussed in further detail in Section 3.5.2 of Chapter 3, and the axes labels in Figure 6.1 relate to the columns for Model 3 given in Table 3.7, Section 3.6 of Chapter 3.

We note from Figure 6.1 that all of the columns which are correlated in the model matrix for the design in Table 6.1 are fully correlated with each other. This means that the parameters estimates related to the correlated columns cannot be estimated independently. The defining relation, which is the relationship between the columns in the model matrix for the full 2^6 factorial design that can be used to select the sixteen rows in Table 6.1, is $\mathbf{1}_n = \mathbf{f}_2 = -\mathbf{f}_1\mathbf{f}_3\mathbf{f}_6$=$-\mathbf{f}_1\mathbf{f}_2\mathbf{f}_4\mathbf{f}_6$, where $\mathbf{1}_n$ is the $n \times 1$ vector of ones.

The columns correlated with $\mathbf{1}_n$ are referred to as words, and the number of columns included in these words is referred to as the length of the word. For example, $\mathbf{f}_2$ is a word of length one and $-\mathbf{f}_1\mathbf{f}_3\mathbf{f}_6$ is a word of length 3. The resolution of a design is one greater than the length of the shortest word in the defining relation. The design in Table 6.1 is therefore a resolution II design.

The design in Table 6.1 is therefore severely limited, as it is a resolution II design and has a main effect fully aliased with the intercept. We would not recommend this design for use in practise. However, we will investigate whether, as expected, the repeated whole-plots improve the estimates of the variance components, and therefore the results for the frequentist analysis method of all subsets regression, when compared to the results in Section 4.3.2 in Chapter 4.

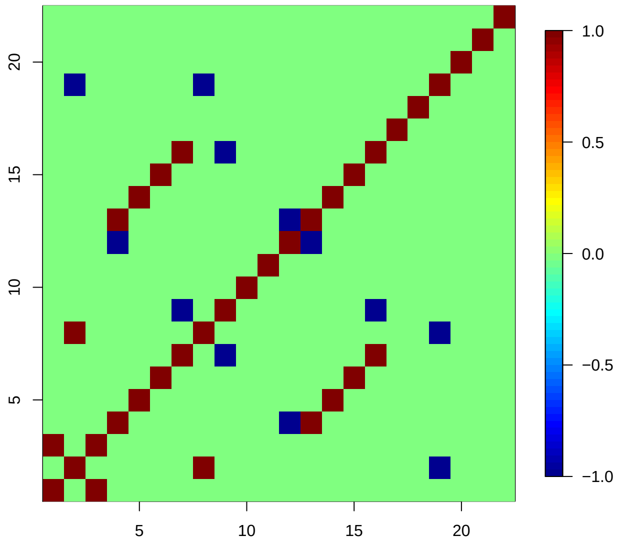


Figure 6.1: Heat map of column correlation matrix for Model 3 from Table 3.7 for the design in Table 6.1.

We used the same models as used in 4.3.2 to simulate responses, hence we assume that:

- The responses from **Stage 1** are generated using (1.3) with $\boldsymbol{\beta}$=(4.80, 4.77, -3.73, -4.93, 0, 0,0,0,0,0,0,0,-4.83,6.73,0)^T, where $\mathbf{X}$ is the model matrix including the columns for the main effects and all pairwise products of the five factors in stage one, $\boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{0}_4, 10\mathbf{I}_4)$ and $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}_{16}, \mathbf{I}_{16})$.

- The responses from **Stage 2** are generated using (1.3) with $\boldsymbol{\beta}$=(0, 5.04, 0, 5.48, 0, -4.25, 0,0,0,0,0,0,0,0,0,0,0,0,0,0,0)^T, where $\mathbf{X}$ is the model matrix including the columns for the main effects and all pairwise products of the six factors in stage one and two, $\boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{0}_4, 10\mathbf{I}_4)$ and $\boldsymbol{\epsilon} \sim \mathrm{N}(\mathbf{0}_{16}, \mathbf{I}_{16})$.

We note from Figure 6.1 that $\mathbf{f}_1 = \mathbf{f}_1\mathbf{f}_2 = -\mathbf{f}_3\mathbf{f}_6$, $\mathbf{f}_3 = -\mathbf{f}_1\mathbf{f}_6 = \mathbf{f}_2\mathbf{f}_3$, $\mathbf{f}_4 = \mathbf{f}_2\mathbf{f}_4$ and $\mathbf{f}_6 = -\mathbf{f}_1\mathbf{f}_3 = \mathbf{f}_2\mathbf{f}_6$. The terms relating to these columns cannot be independently estimated when performing all subsets regression. Aliased terms cannot be considered in the same model when fitting the linear mixed model (1.3) to the simulated data. This therefore reduces p^* used in the all subsets regression for the design in Table 6.1, as the maximum number of columns in the model which can be selected without including correlated columns is eight.

Also, the complete aliasing means that if the term relating to $\mathbf{f}_1$, for example, is selected in all subsets regression, then this is identical to selecting the term relating to $\mathbf{f}_1\mathbf{f}_2$ or $-\mathbf{f}_3\mathbf{f}_6$. Therefore, in the results given in Tables 6.2 and 6.3, we state we have found the correct model if the term used to simulate the data or any of the aliased terms are identified as active.

We also indicate the number of additional aliased terms are in the final model, for example if the full true model for the simulated response from Stage 2, there are three additional aliased terms which cannot be distinguished from those in this model; one for $\mathbf{f}_4$ and two for $\mathbf{f}_6$.

Criterion	Correct model for $p = 6$?	Correct Terms in Final Model	Additional Terms in Final Model	Additional Aliased Terms in Final Model
BIC	Yes	6 of 6	0	5
pAIC	Yes	5 of 6	0	3
mAIC	Yes	5 of 6	0	3

Table 6.2: Models selected using all subsets regression for various model selection criteria when (1.3) is assumed to model simulated responses from Stage 1 of the 16 run two-stage split-plot experiment in Table 6.1.

Criterion	Correct model for $p = 3$?	Correct Terms in Final Model	Additional Terms in Final Model	Additional Aliased Terms in Final Model
BIC	Yes	3 of 3	4	3
pAIC	Yes	3 of 3	2	2
mAIC	Yes	3 of 3	1	2

Table 6.3: Models selected using all subsets regression for various model selection criteria when (1.3) is assumed to model simulated responses from Stage 2 of the 16 run two-stage split-plot experiment in Table 6.1.

The final models in Table 6.2 (Stage 1) and Table 6.3 (Stage 2) have fewer additional terms than the models in Tables 4.1 and 4.2 in Section 4.3.2 of Chapter 4. Also, more of the correct terms are identified for mAIC for the simulated response from Stage 1, and the correct model is identified for simulated responses from Stage 2 for all three model selection criteria.

However, these results are impacted by the additional aliased parameters in the final model, as the aliasing between these terms means these will not be able to be distinguished from other terms. The correlation between columns for the model matrix for Model 3 for the two-stage optimal split-plot design, Figure 3.7b in Section 4.3.2, is in $(-1, 0)$ or $(0, 1)$, hence there is the potential to gain information about all the effects of interest for this design.

p	$\hat{\sigma}^2_\epsilon$	$\hat{\sigma}^2_\gamma$
1	56.88	14.11
2	20.76	22.34
3	0.75	27.39
4	0.75	9.04
5	0.65	9.07
6	**0.61**	**9.08**
7	0.59	9.09
8	0.58	9.09

Table 6.4: Estimates of the variance components (2dp) when the models of size p which maximise BIC, pAIC and mAIC for simulated responses for Stage 1 of the 16-run two-stage split-plot experiment given in Table 6.1. The estimates of the variance components for the true model are highlighted in bold.

p	$\hat{\sigma}^2_\epsilon$	$\hat{\sigma}^2_\gamma$
1	0.45	9.26
2	0.26	9.31
3	**0.14**	**9.34**
4	0.09	9.35
5	0.07	9.36
6	0.06	9.36
7	0.06	5.98
8	0.05	5.98

Table 6.5: Estimates of the variance components (2dp) when the models of size p which maximise BIC, pAIC and mAIC for simulated responses for Stage 2 of the 16-run two-stage split-plot experiment given in Table 6.1. The estimates of the variance components for the true model are highlighted in bold.

Tables 6.4 and 6.5 give the estimates of σ^2_ϵ and σ^2_γ from the models fitted to the simulated responses from Stage 1 and 2, respectively, for the experiment in Table 6.1. The majority of these estimates are an improvement on the estimates for the optimal two-stage split-plot design for (3.20), which were presented in Tables 4.3 and 4.4 in Section 4.3.2 of Chapter 4.

Notice that the estimates of σ^2_ϵ in both Tables 6.4 and 6.5 are larger than those in Tables 4.3 and 4.4, however they still underestimate the true value of $\sigma^2_\epsilon = 1$. The estimates of σ^2_γ in Tables 6.4 and 6.5 are closer than those in Tables 4.3 and 4.4 to the true value of $\sigma^2_\gamma = 10$.

The variance components for the model used to simulate the data are highlighted in bold in Tables 6.4 and 6.5. We notice that the estimates of the two variance components in Table 6.4 are larger, and closer to the true value, than the estimates of the variance components for the true model given in Table 4.3. Also, the estimate for σ^2_γ in Table

6.5 is larger, and closer to the true value that the estimate of σ_γ^2 in Table 4.4. However, the estimate for σ_ϵ^2 for the true model for Stage 2 is smaller in Table 6.5 than in Table 4.4.

Therefore, based on the results discussed in this section, a two-stage design found using (6.6) appears to have the potential to overcome the problems with frequentist analysis discussed in Section 4.3. However, the correlation of ± 1 between columns in the model matrix for the resolution II design in Table 6.1, as shown in Figure 6.1, makes this design unusable in practise. Hence, further research would be required to identify a criterion which finds supersaturated designs which balance the objectives of estimating fixed effects and variance components.

Although not shown here, we also found 16 run two-stage split-plot designs for (6.6) with $\alpha = 0.5, 0.55, 0.6, 0.65, 0.7$ using the coordinate exchange algorithm in Section 3.5.1. These were also all resolution II designs, and therefore had full aliasing between the mean, a main effect, and two-factor interactions.

We are currently researching supersaturated split-plot designs found using (6.6) with $l = 1$, and investigating the impact of α and $\mathbf{R}_1$ on the number, and size, of correlated columns. We are also considering whether using a criterion such as $E(s^2)$ optimality (Booth and Cox, 1962) instead of Bayesian D-optimality in (6.6) will help avoid highly correlated columns.

Appendix

A Gaussian Quadrature

Numerical quadrature is a method of numerically approximating definite integrals when there is no analytical solution. In quadrature, the integral is approximated by a weighted sum of the integrand evaluated at selected values in the domain of the integral, where the n_a values at which the integrand is evaluated are called the abscissas.

Gaussian quadrature is a class of quadrature techniques that are suited to certain types of integrals. An n_a-point Gaussian quadrature rule gives an exact result for polynomials of degree $2n_a - 1$ or less when certain, non-equally spaced, abscissas and weights are chosen. Other quadrature methods are suitable for different forms of integrands. For further detail regarding quadrature methods, see Ralston and Rabinowitz (2001).

Gauss-Jacobi quadrature approximates and integral of the form

$$\int_{-1}^{1} f(x)(1+x)^{\alpha}(1-x)^{\beta} dx \tag{A.1}$$

as

$$\sum_{g=1}^{n_a} w_g f(a_g), \tag{A.2}$$

where the abscissas a_g, $g = 1, \ldots, n_a$, are the roots of the Jacobi polynomial of degree n_a and

$$w_g = \frac{-2n_a + \alpha + \beta + 2}{n_a + \alpha + \beta + 1} \frac{\Gamma(n_a + \alpha + 1)\Gamma(n_a + \beta + 1)}{\Gamma(n_a + \alpha + \beta + 1)(n_a + 1)!} \frac{2^{\alpha+\beta}}{P'_{n_a}(a_i) P_{n_a+1}(a_i)} \tag{A.3}$$

are the weights when P_{n_a} is the Jacobi polynomial of degree n_a. For more details regarding Jacobi polynomials, see Szegö (1975).

In Sections 4.3.3 and 6.2.2, we use Gauss-Jacobi quadrature to approximate analytically intractable integrals for a function involving a correlation parameter, ρ, which is assumed to have a $\beta(\kappa, \lambda)$ prior,

$$p(\rho) = \frac{\Gamma(\kappa+\lambda)}{\Gamma(\kappa)\Gamma(\lambda)}\rho^{\kappa-1}(1-\rho)^{\lambda-1}, \tag{A.4}$$

where $\kappa, \lambda \geq 0$ are the two shape parameters. Assuming we have some likelihood function involving ρ, for example $f(\mathbf{y}|\rho)$, the marginal density of $\mathbf{y}$ has the form

$$\int_0^1 f(\mathbf{y}|\rho)p(\rho)d\rho = \int_0^1 f(\mathbf{y}|\rho)\frac{\Gamma(\kappa+\lambda)}{\Gamma(\kappa)\Gamma(\lambda)}\rho^{\kappa-1}(1-\rho)^{\lambda-1}d\rho. \tag{A.5}$$

Let $\rho = (\phi+1)/2$, then (A.5) is proportional to

$$\frac{1}{2}^{\kappa+\lambda-1}\int_{-1}^{1} f\left(\mathbf{y}\middle|\frac{\phi+1}{2}\right)(\phi+1)^{\kappa-1}(1-\phi)^{\lambda-1}d\phi, \tag{A.6}$$

which has the same form as (A.1) and can therefore be approximated by

$$\frac{1}{2}^{\kappa+\lambda-1}\sum_{g=1}^{n_a} w_g^{GJ} f\left(\mathbf{y}\middle|\frac{a_g^{GJ}+1}{2}\right), \tag{A.7}$$

where a_g^{GJ}, $g = 1, \ldots, n_a$, are the n_a abscissas and w_g^{GJ} are the n_a weights obtained from Gauss-Jacobi polynomials. These abscissas and weights can be found using statistical software, for example the function `gauss.quad` in the package `statmod` in `R`.

B Assessment of Models from Chapter 5

We have assumed that the errors are normally distributed with constant variance when fitting the linear mixed effect model, (4.21) from Section 4.4.1 of Chapter 4, to the responses from the GlaxoSmithKline (GSK) experiment in Chapter 5. In this appendix, we present model assessment plots for the models in Chapter 5 (Section B.2), which use the posterior median of each column of the sampled logit-transformed predicted responses, (5.7) from Section 5.4 of Chapter 5, as the fitted response, and therefore ignore the correlation between the columns.

Samples from the Metropolis-Hastings within Gibbs sampling algorithm (Section 4.4.4) are used to calculate (5.7). Both $\beta(2, 2)$ and $\beta(11, 2)$ are used as prior distributions for $\phi = \sigma_\gamma^2/(\sigma_\epsilon^2 + \sigma_\gamma^2)$ in the sampling algorithm, and throughout this appendix we refer to $\beta(2, 2)$ as Prior 1 and $\beta(11, 2)$ as Prior 2. The plots in Section B.2 are representative examples of the plots from Matthews (2015).

B.1 Model Assessment Plots

We use three plots to assess the models fitted to the responses in Chapter 5; the quantile-quantile (QQ) plots, plots of the posterior medians against the standardised residuals and plots of the factor levels against the standardised residuals.

The vector of residuals is $\mathbf{r}_{dR} = (r_{dR}(\mathbf{x}_1), \ldots, r_{dR}(\mathbf{x}_n))^T$, where $r_{dR}(\mathbf{x}_i)$ is the residual for treatment $\mathbf{x}_i$, $i = 1, \ldots, n$, for Test d and Reference R. Let $\tilde{y}_{dR}(\mathbf{x}_i)$ be the posterior median of the *its* samples of the logit-transformed predicted response for $\mathbf{x}_i$, (5.7) from Section 5.4, and $y_{dR}(\mathbf{x}_i)$ be the logit-transformed observed response for $\mathbf{x}_i$, then

$$r_{dR}(\mathbf{x}_i) = \frac{(\tilde{y}^L_{dR}(\mathbf{x}_i) - y^L_{dR}(\mathbf{x}_i))^2}{n}. \tag{A.8}$$

Similarly, the vector of standardised residuals is $\dot{\mathbf{r}}_{dR} = (\dot{r}_{dR}(\mathbf{x}_1), \ldots, \dot{r}_{dR}(\mathbf{x}_n))^T$, where $\dot{r}_{dR}(\mathbf{x}_i)$ is the standardised residual for treatment $\mathbf{x}_i$, $i = 1, \ldots, n$, for Test d and Reference R. Let $\tilde{\sigma}^2_{dR}$ be the posterior median of the Rth diagonal element of the *its* sampled scale matrices for Test d, $\mathbf{\Sigma}_d$, then

$$\dot{r}_{dR}(\mathbf{x}_i) = \frac{(\tilde{y}^L_{dR}(\mathbf{x}_i) - y^L_{dR}(\mathbf{x}_i))^2}{n\tilde{\sigma}^2_{dR}}. \tag{A.9}$$

QQ-plots are used to assess whether the residuals are approximately normally distributed. In a QQ-plot, the quantiles from the vector of residuals are plotted against the quantiles of an appropriate normal distribution. Hence, if the residuals are approximately normally distributed, the plotted points form a straight line. An example QQ plot is given in Figure A.1. Significant departures from a straight line suggest that a different distributional assumption should be considered for the residuals.

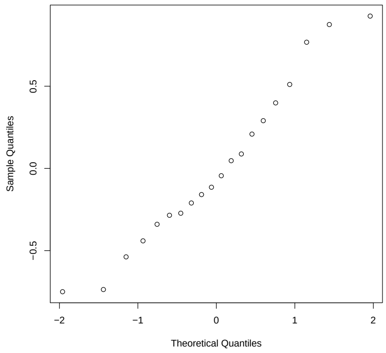


Figure A.1: An example QQ-plot, which plots the theoretical quantiles of the normal distribution which the errors are assumed to follow against the quantiles of the residuals (A.8) for all treatments, $\mathbf{x}_i$, $i = 1, \ldots, n$. These residuals are calculated using samples from the Metropolis-Hastings within Gibbs Sampling algorithm (Section 4.4.4) and the logit-transformed observed responses for Test 2, Reference 1 and Prior 2 from Chapter 5.

The plots of the posterior median against the standardised residuals are used to determine whether the assumption of a constant variance for the errors is appropriate. If the residuals have equal variances then the points on this plot will be a random scatter, however any sort of trend or fan suggests that there are unequal variances (heteroscedasticity), and hence alternative distributions for the error should be considered. Figure A.2 is an example posterior median against standardised residual plot.

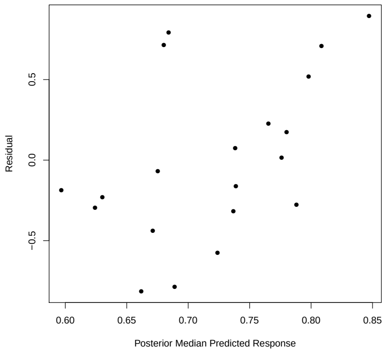


Figure A.2: An example plot of the posterior median of the predicted responses against the standardised residuals (A.9) for all treatments, $\mathbf{x}_i$, $i = 1, \ldots, n$. These standardised residuals are calculated using samples from the Metropolis-Hastings within Gibbs Sampling algorithm (Section 4.4.4) and the logit-transformed observed responses for Test 3, Reference 1 and Prior 2 from Chapter 5.

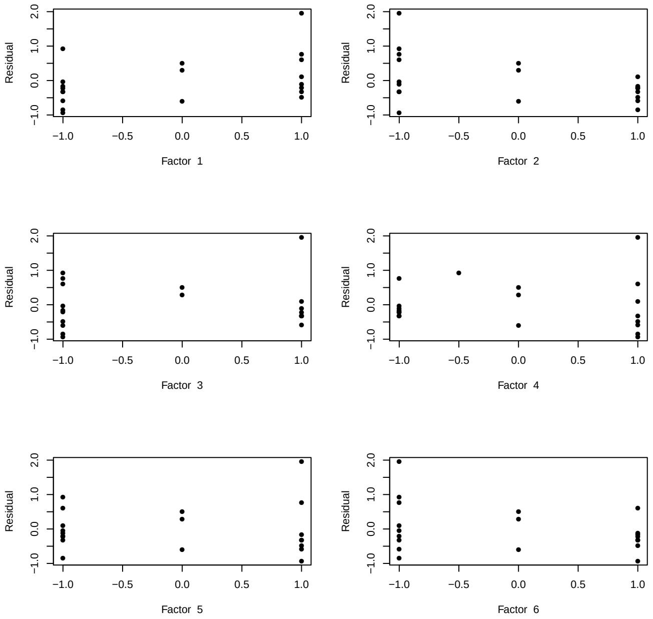


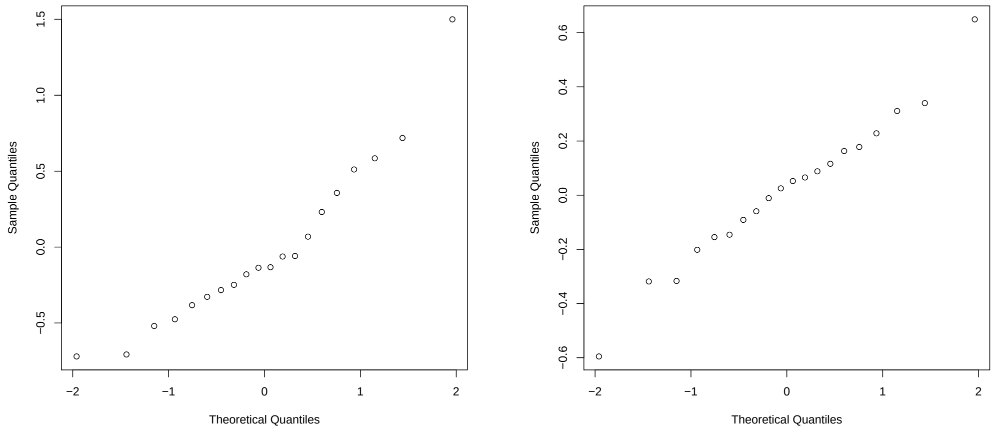
Figure A.3: An example plot of the factor levels against the standardised residuals, (A.9) for all treatments, $\mathbf{x}_i$, $i = 1, \ldots, n$. These standardised residuals are calculated using samples from the Metropolis-Hastings within Gibbs Sampling algorithm (Section 4.4.4) and the logit-transformed observed responses for Test 3, Reference 1 and Prior 2 from Chapter 5.

Plots of the standardised residuals against the factor levels are also used to determine whether the assumption of a constant variance for the errors is appropriate. If there is evidence of some relationship between the variability of the residuals and the factor levels, such as a difference in the spread of the points at different factor levels or a trend in the residuals as the factor levels change, the assumption of constant variance does not hold. An example plot of the factor levels against the standardised residuals is given in Figure A.3.

B.2 Assessment of Models from Chapter 5

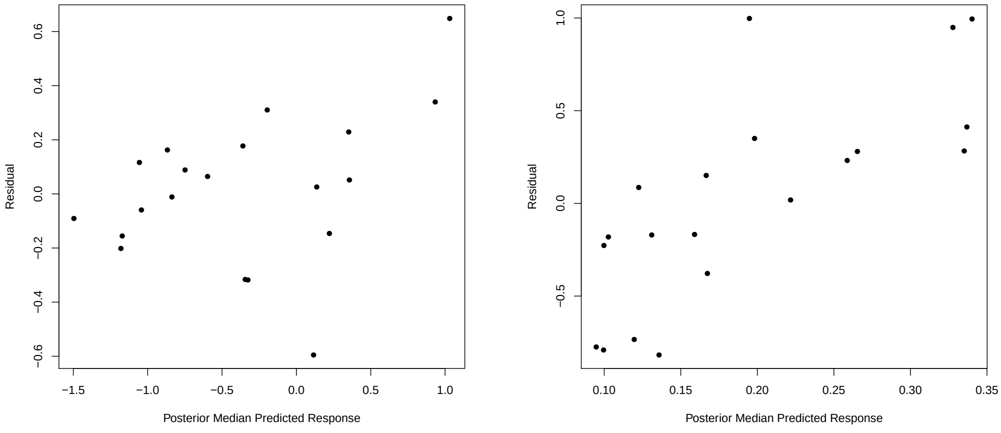
The model assessment plots for models fitted to the experimental data from Chapter 5 are given in Matthews (2015). In this section we discuss certain plots which were selected to be representative examples. As the supersaturated split-plot design for the experiment in Chapter 5 is a screening design, future experimentation will be undertaken for the influential factors identified in the analysis in Section 5.3 and we do not expect these initial results to show a model with a perfect fit.

Figure A.4 provides two example QQ-plots with plotted points which are close to a straight line. This suggests the assumption of normally distributed errors holds, however the effect of outliers, such as the outlier seen in Figure A.4(a), may need to be considered.



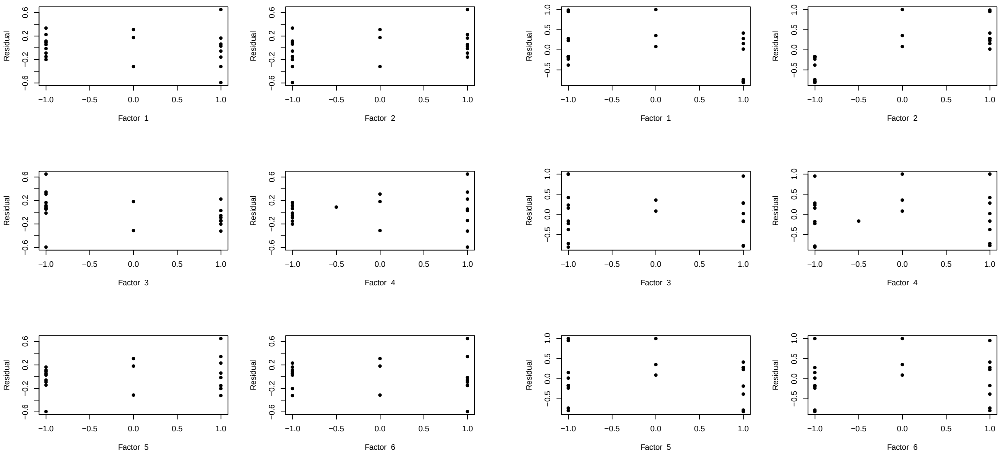
(a) (b)

Figure A.4: Example QQ-plots for models fitted to logit-transformed f_2 values for the two-stage supersaturated split-plot experiment in Chapter 5. These residuals are calculated using samples from the Metropolis-Hastings within Gibbs sampling algorithm (Section 4.4.4) and the logit-transformed observed responses for (a) Test 1, Reference 1 and Prior 2, and (b) Test 2, Reference 1 and Prior 2.



(a) (b)

Figure A.5: Example plots of the posterior median response against the standardised residuals (A.9) for models fitted to logit-transformed f_2 values for the two-stage supersaturated split-plot experiment in Chapter 5. These standardised residuals are calculated using samples from the Metropolis-Hastings within Gibbs sampling algorithm (Section 4.4.4) and the logit-transformed observed responses for (a) Test 2, Reference 1 and Prior 2, and (b) Test 4, Reference 2 and Prior 2.



(a) (b)

Figure A.6: Example plots of the factor levels against the standardised residuals (A.9) for models fitted to logit-transformed f_2 values for the two-stage supersaturated split-plot experiment in Chapter 5. These standardised residuals are calculated using samples from the Metropolis-Hastings within Gibbs sampling algorithm (Section 4.4.4) and the logit-transformed observed responses for (a) Test 2, Reference 1 and Prior 2, and (b) Test 4, Reference 2 and Prior 2.

Figures A.5 and A.6 provide two example model assessment plots for models with a poor fit. The other plots for the models considered in Chapter 5 are similar to Figures A.2 and A.3 in Section B.1. Figure A.5(a) shows a plot with some fanning, hence the assumption of constant variances may not hold, as the variability of the residuals increases as the response increases. Therefore, a stronger transformation may need to be applied to the responses from the experiment to ensure the assumption of constant variances holds.

The residuals in Figure A.5(b) have an upward trend, hence we are underfitting for some responses and overfitting for others. There is therefore some bias in the results, which will affect any modelling or prediction performed using these models.

Figure A.6 gives the factor level against standardised residual plots for the standardised residuals which were plotted against the posterior median in Figure A.5. There is some increase in variability of the residuals as the level of Factors 4 and 6 increases in Figure A.6(a), which matches the fanning seen in Figure A.5(a). The residuals are negative when Factor 2 is -1 and positive when Factor 2 is +1 in Figure A.6(b), hence this figure also gives evidence of bias.

C Assessment of MCMC Samples from Chapters 4 and 5

In this appendix, we use two plots; the trace plot and the autocorrelation function (ACF) plot, which are introduced in Section C.1, to investigate the convergence properties of the Markov chains formed by the samples in the Metropolis-Hastings within Gibbs sampling algorithm from Section 4.4.4 of Chapter 4 for the responses in Chapters 4 and 5. These plots are by no means the only methods for analysing the convergence of Markov Chain Monte Carlo (MCMC) sample, and alternatives are discussed and presented by a number of authors including Schafer (1997) and O'Hagan and Forster (2004, Chapter 8).

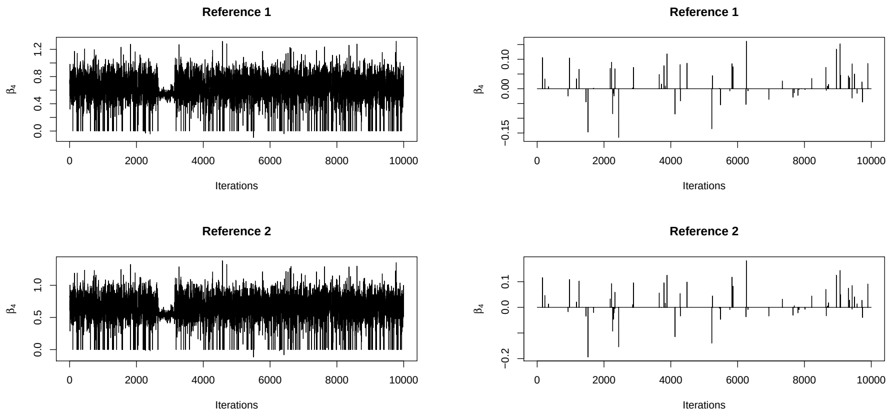
The plots in Section C.2 and C.3 are chosen to be representative examples of the plots in Matthews (2015). Recall that $\beta(2,2)$ and $\beta(11,2)$ are both used as prior distributions for $\phi = \sigma_\gamma^2/(\sigma_\epsilon^2+\sigma_\gamma^2)$ in Chapter 4 and 5 (see Section 4.4 for further detail). Throughout this appendix we refer to $\beta(2,2)$ as Prior 1 and $\beta(11,2)$ as Prior 2.

C.1 MCMC Assessment Plots

We use the trace and autocorrelation function (ACF) plot to assess whether the Markov chains from the Metropolis-Hastings within Gibbs sampling algorithm in Section 4.4.4 of Chapter 4 appear to have converged to their known conditional distributions. The trace plot is a line plot of the iterative samples from the conditional distribution, and a Markov chain which converges to the required uni-modal posterior distribution will not have any shifts in mode.

However, some of the conditional distributions we sample from are unusual, for example the conditional distribution for the fixed effect vector $\boldsymbol{\beta}_j$, $j = 1, \dots, p$, is sampled from different distributions with a certain probability. These conditional distributions will affect the Markov chains formed by sampling from them, therefore the trace plots for these parameters will be unusual.

Figure A.7(a) is an example trace plot for a fixed effect parameter $\boldsymbol{\beta}_j$, which is the jth row of $\mathbf{B}$ in (4.21) from Section 4.4.1 of Chapter 4, with a high posterior probability of being active. Therefore, this $\boldsymbol{\beta}_j$ is mainly sampled from a normal distribution. Figure A.7(b) is an example trace plot for a $\boldsymbol{\beta}_j$ which has a low posterior probability of being active and is therefore mainly sampled as $\mathbf{0}_r$.

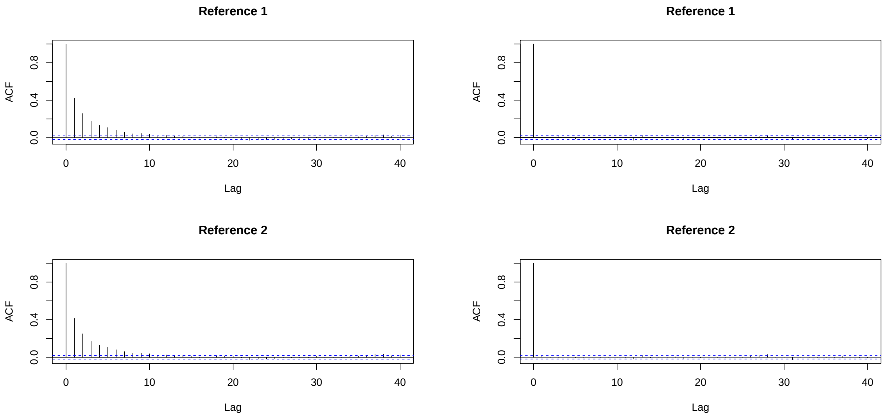


(a) (b)

Figure A.7: Example trace plots the fixed effect vector $\boldsymbol{\beta}_j$, $j = 1, \ldots, p$, which is sampled using the Metropolis-Hastings within Gibbs sampling algorithm, when $\boldsymbol{\beta}_j$ has a high posterior probability of being (a) active ($\boldsymbol{\beta}_4$ for Test 2 and Prior 1 from Chapter 5) and (b) non-active ($\boldsymbol{\beta}_4$ for Test 4 and Prior 1 from Chapter 5).

The trace plots for the indicator variable δ_j, $j = 1, \ldots, p$, and c are also unusual as these parameters are sampled from particular sets, $\delta_j \in \{0, 1\}$ and $c \in \{1/4, 9/16, 1, 4, 9, 16, 25\}$. Trace plots for these variables are not considered, as they are difficult to gain insight from.

The ACF plot shows the correlation between iterative samples against the distance, or lag, between iterative samples. When the Markov chain of sampled parameters converges to the required posterior distribution, the correlation between samples should be low throughout. Figure A.8 gives two examples ACF plots for Markov chains with good convergence properties.



(a) (b)

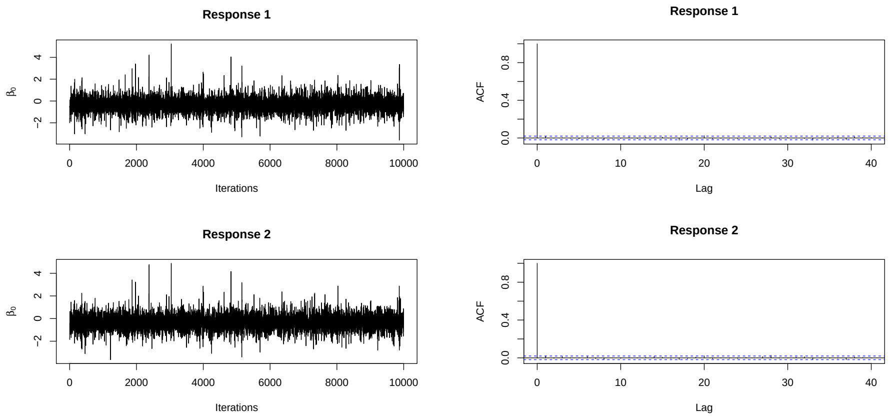
Figure A.8: Example ACF plots for the fixed effect vector $\boldsymbol{\beta}_j$, $j = 1, \ldots, p$, which is sampled using the Metropolis-Hastings within Gibbs sampling algorithm, when $\boldsymbol{\beta}_j$ has a high posterior probability of being (a) active ($\boldsymbol{\beta}_4$ for Test 2 and Prior 1 from Chapter 5) and (b) non-active ($\boldsymbol{\beta}_4$ for Test 4 and Prior 1 from Chapter 5).

C.2 Assessment of MCMC samples from Chapter 4

In this section we present example trace plots for each of the parameters sampled using the Metropolis-Hastings within Gibbs sampling algorithm given in Section 4.4.4 for the simulated responses from Table 4.7 in Section 4.5, which are chosen to be representative of the plots for all the parameters and simulated responses given in Matthews (2015).

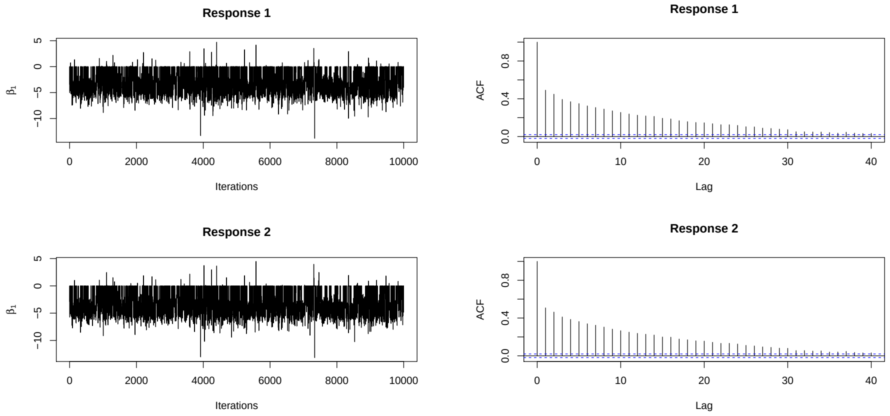
Figure A.9 is the example trace and ACF plot for the intercept parameter, $\boldsymbol{\beta}_0$. This figure suggests the Markov chains formed by sampling $\boldsymbol{\beta}_0$ from (4.32) (as given in Section 4.4.3 of Chapter 4) has good convergence properties, as there no shifts in the mean or trends in Figure A.9(a) and the correlations in Figure A.9(b) are low.

Figure A.10 is an example trace and ACF plots for Markov chains formed by sampling of $\boldsymbol{\beta}_j$, $j = 1, \ldots, p$ using the extended Geweke (1996) approach, which relies on (4.36) and is discussed in Section 4.4.3 of Chapter 4, in the Metropolis-Hastings within Gibbs Sampling algorithm, when $\boldsymbol{\beta}_j$ has a high posterior probability of being active. Similarly, Figure A.12 is an example trace and ACF plots for $\boldsymbol{\beta}_j$ when $\boldsymbol{\beta}_j$ a high posterior probability of being non-active. Figure A.10(a), is representative of a Markov chain with good convergence properties. However, even though the correlation between sampled decreases as the lag increases in Figure A.10(b), some of the correlations are quite high. Figure A.12(b) shows that the correlation for the non-active parameter samples are low, however Figure A.12(a) is unusual because of the high proportion of parameters sampled as zero, as discussed in Section C.1.



(a) (b)

Figure A.9: Example (a) trace and (b) ACF plot for the Markov chain formed by sampling $\boldsymbol{\beta}_0$ from (4.32) in the Metropolis-Hastings within Gibbs sampling algorithm (Section 4.4.4) from Chapter 4. These plots are for the samples of $\boldsymbol{\beta}_0$ for $\mathbf{Y}_{12}$ from Table 4.7.



(a) (b)

Figure A.10: Example (a) trace and (b) ACF plot for the Markov chain formed by sampling $\boldsymbol{\beta}_j$, $j = 1, \ldots, p$, using (4.36) in the Metropolis-Hastings within Gibbs sampling algorithm (Section 4.4.4) when $\boldsymbol{\beta}_j$ has a high posterior probability of being active for the simulated responses from Chapter 4. These plots are for the samples of $\boldsymbol{\beta}_1$ for $\mathbf{Y}_{21}$ from Table 4.7.

Figure A.11 is an example trace plot for a $\boldsymbol{\beta}_j$ which has a zero predicted posterior probability of being active, and is consistently sampled as $\mathbf{0}_r^T$. This occurs for three of the parameters sampled for the simulated data in Chapter 4, and, as the autocorrelation function cannot be calculated for these constant samples, the ACF plots for these parameters are not given in Matthews (2015).

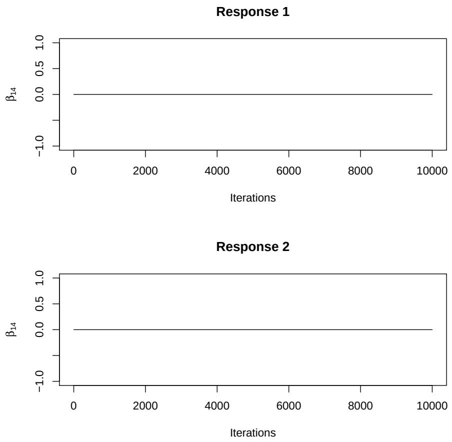
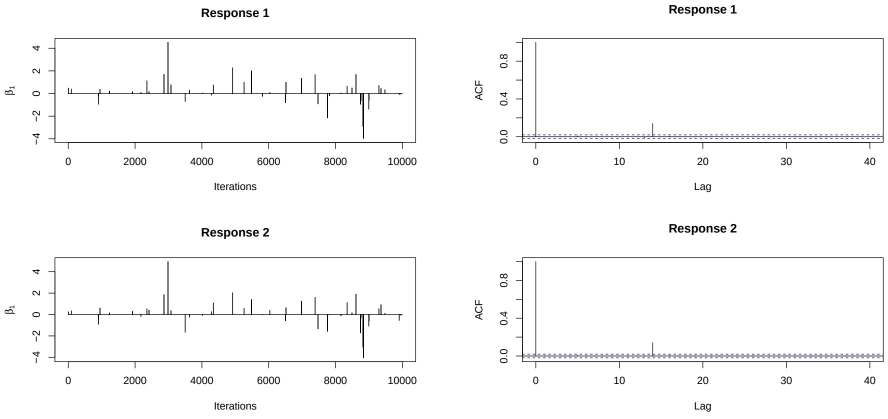


Figure A.11: Example trace plot for the Markov chain formed by sampling $\boldsymbol{\beta}_j$, $j = 1, \ldots, p$, using (4.36) in the Metropolis-Hastings within Gibbs sampling algorithm (Section 4.4.4) when $\boldsymbol{\beta}_j$ has a zero posterior probability of being active for the simulated responses from Chapter 4. These plots are for the samples of $\boldsymbol{\beta}_{14}$ for $\mathbf{Y}_{42}$ from Table 4.7.



(a) (b)

Figure A.12: Example (a) trace and (b) ACF plot for the Markov chain formed by sampling $\boldsymbol{\beta}_j$, $j = 1, \ldots, p$, using (4.36) in the Metropolis-Hastings within Gibbs sampling algorithm (Section 4.4.4) when $\boldsymbol{\beta}_j$ has a high posterior probability of not being active for the simulated responses from Chapter 4. These plots are for the samples of $\boldsymbol{\beta}_1$ for $\mathbf{Y}_{32}$ from Table 4.7.

Figure A.13 is an example ACF plot for c, the weighting of the scale matrix, $\boldsymbol{\Sigma}$, for the active fixed effect parameters, $\boldsymbol{\beta}_j$. This ACF plot has low correlations between samples, which decrease as lag increases, which suggest that the Markov Chain formed by sampling c from (4.34) (from Section 4.4.3 of Chapter 4) has good convergence properties.

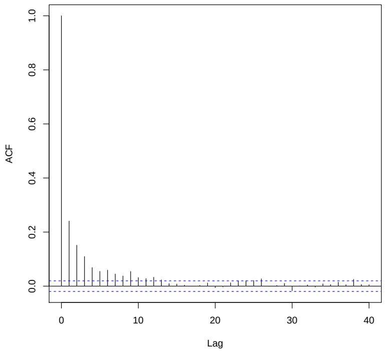


Figure A.13: Example ACF plot for the Markov chain formed by sampling c from (4.34) in the Metropolis-Hastings within Gibbs sampling algorithm (Section 4.4.4) for the simulated responses from Chapter 4. This plots is for the samples of c for $\mathbf{Y}_{41}$ from Table 4.7.

Figure A.14 is an example trace and ACF plot for parameter sampled using Metropolis-Hastings rejection sampling, ϕ, the between run correlation parameter. The trace plot (Figure A.14(a)) shows that the Markov chain formed by the samples of ϕ considers different ϕ values and does not get stuck at one proposed value of ϕ, however the ACF plot (Figure A.14(b)) shows that the autocorrelation decreases but is high at small lags.

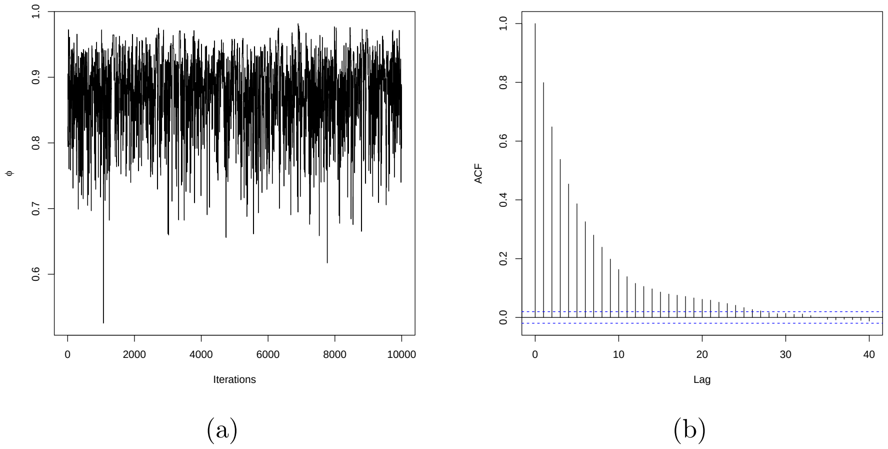


Figure A.14: Example (a) trace and (b) ACF plot for the Markov chain formed by sampling ϕ using Metropolis-Hastings sampling for (4.35) in the Metropolis-Hastings within Gibbs sampling algorithm (Section 4.4.4) for the simulated responses from Chapter 4. These plots are for the samples of ϕ for $\mathbf{Y}_{22}$ from Table 4.7.

We can use the effective sample size, which is an estimate of the number of independent samples from the posterior the chain represents, to assess the impact of the correlation on the samples of ϕ. The effective sample size is given by

$$\frac{its}{1+2\sum_{l=1}^{\infty}\rho_l} \tag{A.10}$$

where ρ_l is the autocorrelation at lag l, and its is the number of iterations of the Metropolis-Hastings within Gibbs sampling algorithm. The infinite sum $\sum_{l=1}^{\infty} \rho_l$ has to be estimated in order to get an estimate of (A.10), however the larger the autocorrelation, the smaller the effective sample size is.

The estimated effective sample sizes, found using `effectiveSize` in the library `coda` in `R`, for the Markov chains for ϕ considered in Chapter 4 are given in Table A.1. These estimated effective sample sizes are lower than the 10% of its that we would ideally want, where $its = 10000$, and show the impact the high autocorrelation is having on the number of equivalent independent samples we can obtain from the Markov chain for ϕ. To improves these effective samples sizes, we could increase its or consider an alternative proposal distribution for ϕ.

The acceptance rate can also be used to assess parameters which are sampled using Metropolis-Hastings rejection sampling. The acceptance rate is the ratio of the number of times the proposal is accepted over the total number of iterations, an an acceptance rate of between 0.1 and 0.4 is preferred. The acceptance rates of the Markov chains for ϕ considered in Chapter 4 are given in Table A.2, and are all either close to or between 0.1 and 0.4, as required.

The results for the simulated response $\mathbf{Y}_{32}$, which is the simulated response for a supersaturated designs when the mean of the active responses is assumed to be ±20 for Prior 1, are concerning, as the estimated effective sample size is very low (237) and the acceptance probability (0.09) is also very low. A more extensive simulation study which considers multiple simulated responses for this mean and prior distribution would help establish whether these settings always produce Markov chains with low effective sample sizes and acceptance probability, and try and establish what features of the experiment are causing these problems.

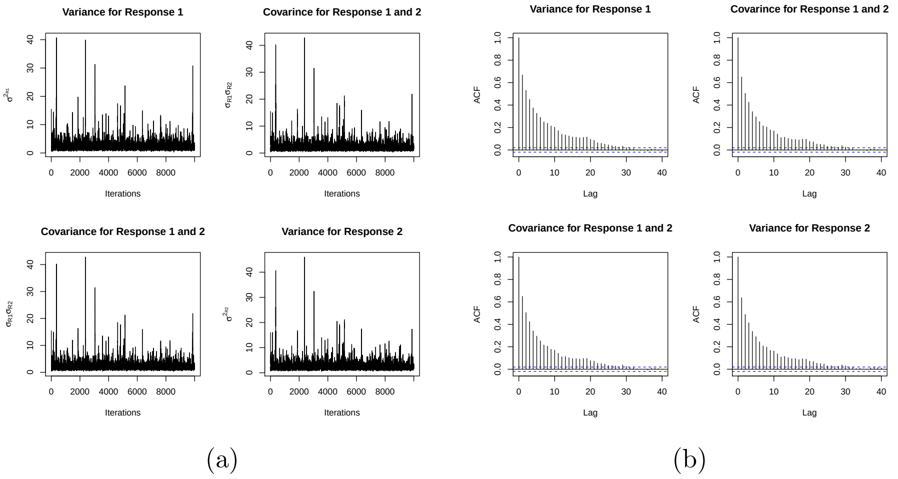
Response	Effective Sample Size
$\mathbf{Y}_{11}$	988
$\mathbf{Y}_{21}$	419
$\mathbf{Y}_{31}$	579
$\mathbf{Y}_{41}$	979
$\mathbf{Y}_{12}$	778
$\mathbf{Y}_{22}$	914
$\mathbf{Y}_{32}$	237
$\mathbf{Y}_{42}$	308

Table A.1: Estimated effective sample size, rounded to nearest whole number, of samples of ϕ for the simulated responses, given in Table 4.7 of Section 4.5.1, from Chapter 4 .

Response	Acceptance Rate
$\mathbf{Y}_{11}$	0.38
$\mathbf{Y}_{21}$	0.36
$\mathbf{Y}_{31}$	0.39
$\mathbf{Y}_{41}$	0.35
$\mathbf{Y}_{12}$	0.34
$\mathbf{Y}_{22}$	0.38
$\mathbf{Y}_{32}$	0.09
$\mathbf{Y}_{42}$	0.25

Table A.2: Acceptance rate, rounded to 2dp, of samples of ϕ for the simulated responses, given in Table 4.7 of Section 4.5.1, from Chapter 4.

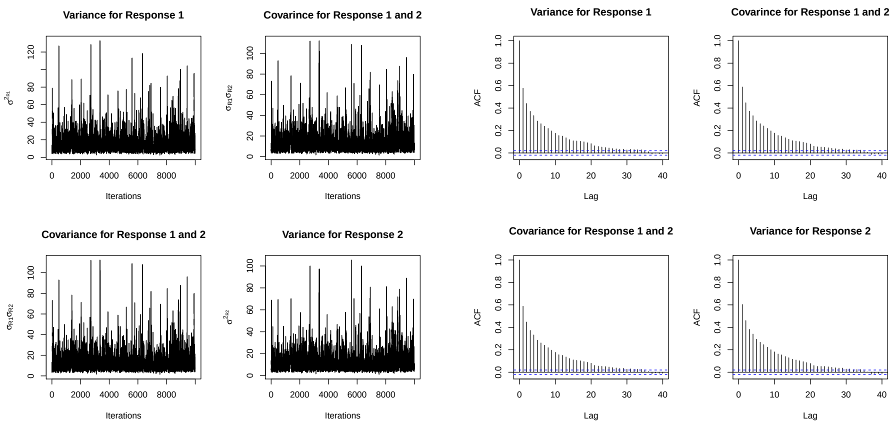
Figure A.15 is an example trace and ACF plot for the four elements of the scale matrix, $\mathbf{\Sigma}$ when Prior 1 is assumed. Similarly, Figure A.16 is an example trace and ACF plot for the four elements of the sampled $\mathbf{\Sigma}$ when Prior 2 is assumed. We notice that there are significant peaks in both trace plots, Figures A.15(a) and A.16(a), where the MCMC algorithm explores the tails of the distribution, and that the ACF plots, Figures A.15(b) and A.16(b), have relatively high correlations at low lags that decrease over time. The peaks in the trace plot do not represent significant shifts in mode, and are expected, as the sampler will consider models where a small number of fixed effect vectors, $\boldsymbol{\beta}_j$, are active. The elements of $\mathbf{\Sigma}$ will be large when there are a small number of active terms in the models, as the variability in the responses will be assumed to be random instead of being explained by the settings of the factors in the experiment.



(a)

(b)

Figure A.15: Example (a) trace and (b) ACF plot for the Markov chain formed by sampling $\mathbf{\Sigma}$ from (4.33) in the Metropolis-Hastings within Gibbs sampling algorithm (Section 4.4.4) with Prior 1 for the simulated responses from Chapter 4. These plots are for the samples of $\mathbf{\Sigma}$ for $\mathbf{Y}_{12}$ from Table 4.7.



(a) (b)

Figure A.16: Example (a) trace and (b) ACF plot for the Markov chain formed by sampling $\mathbf{\Sigma}$ from (4.33) in the Metropolis-Hastings within Gibbs sampling algorithm (Section 4.4.4) with Prior 2 for the simulated responses from Chapter 4. These plots are for the samples of $\mathbf{\Sigma}$ for $\mathbf{Y}_{22}$ from Table 4.7.

Response	Effective Sample Size
$\mathbf{Y}_{11}$	$\begin{pmatrix} 1605 & 1566 \\ 1566 & 1557 \end{pmatrix}$
$\mathbf{Y}_{21}$	$\begin{pmatrix} 428 & 433 \\ 433 & 433 \end{pmatrix}$
$\mathbf{Y}_{31}$	$\begin{pmatrix} 2394 & 2991 \\ 2991 & 2851 \end{pmatrix}$
$\mathbf{Y}_{41}$	$\begin{pmatrix} 1905 & 1874 \\ 1874 & 1649 \end{pmatrix}$
$\mathbf{Y}_{12}$	$\begin{pmatrix} 833 & 938 \\ 938 & 1020 \end{pmatrix}$
$\mathbf{Y}_{22}$	$\begin{pmatrix} 1043 & 1039 \\ 1039 & 1017 \end{pmatrix}$
$\mathbf{Y}_{32}$	$\begin{pmatrix} 1362 & 1383 \\ 1383 & 1164 \end{pmatrix}$
$\mathbf{Y}_{42}$	$\begin{pmatrix} 326 & 318 \\ 318 & 314 \end{pmatrix}$

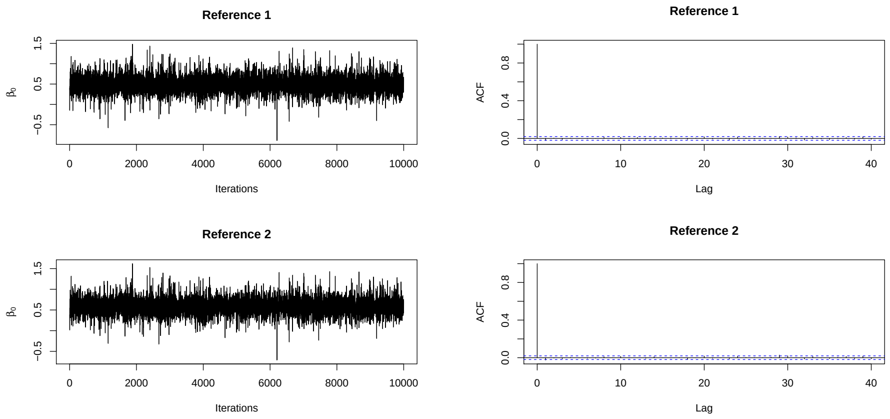
Table A.3: Estimated effective sample size, rounded to nearest whole number, of samples of the four elements of $\mathbf{\Sigma}$ for the simulated responses, given in Table 4.7 of Section 4.5.1, from Chapter 4 .

As the correlations in Figures A.15(b) and A.16(b) are relatively high, we chose to look at the estimated effective samples size for these Markov chains. Table A.3 gives the effective samples sizes for the four elements of the $\mathbf{\Sigma}$ for all the sampled parameters from Chapter 4. The effective sample sizes for the samples of $\mathbf{\Sigma}$ found using $\mathbf{Y}_{21}$ and $\mathbf{Y}_{42}$ are particularly low, however the effective sample sizes for the other simulated responses are either close to or greater than 1000, which is 10% of $its = 10000$. A more extensive simulation study could be performed to obtain a distribution of the sample sizes for responses for the specific settings given in Table 4.7, and therefore assess whether the low sample sizes for $\mathbf{Y}_{21}$ and $\mathbf{Y}_{42}$ are common and caused by our distributional assumptions or particular features of the experiment, such as the correlation between columns in the supersaturated design for Stage 2.

C.3 Assessment of MCMC samples from Chapter 5

In this section we present representative example trace and ACF plots for the parameters sampled using the Metropolis-Hastings within Gibbs sampling algorithm in Section 4.4.4 for responses from the experiment described in Section 5.1. The plots for all the parameters sampled for all four dissolution tests are given in Matthews (2015).

Figure A.17 is an example of the trace and ACF plots for the Markov chain formed by sampling $\boldsymbol{\beta}_0$ from (4.32). Figure A.17(a) is an example of a trace plot for a Markov chain with good convergence properties, as there are no trends or shifts in mean present. Similarly, Figure A.17(b) indicates that the Markov chain for $\boldsymbol{\beta}_0$ has good convergence properties as the correlation is low across all lags.



(a) (b)

Figure A.17: Example (a) trace and (b) ACF plot for the Markov chain formed by sampling $\boldsymbol{\beta}_0$ using (4.32) in from the Metropolis-Hastings within Gibbs sampling algorithm (Section 4.4.4) from Chapter 5. These plots are for the samples of $\boldsymbol{\beta}_0$ for Test 1 and Prior 1.

Figure A.18 is the trace and ACF plots for the Markov chain formed by sampling of $\boldsymbol{\beta}_j$, $j = 1, \ldots, p$, respectively using the extension of the joint sampling method from Geweke (1996), which is discussed in Section 4.4.3 and relies on (4.36), when $\boldsymbol{\beta}_j$ has a high posterior probability of being active. Similarly, Figure A.19 is the trace and ACF plots for samples of $\boldsymbol{\beta}_j$ when $\boldsymbol{\beta}_j$ has a high posterior probability of not being active.

Figures A.18(a) and A.19(a) are unusual, as there are sections of the chain where $\boldsymbol{\beta}_j$ is assumed to be active and normally distributed and other sections where $\boldsymbol{\beta}_j$ is assumed to not be active and hence sampled as $\mathbf{0}_r^T$. This is expected, and is discussed in Section C.1. The high level of correlation in Figure A.18(b) when compared to Figure A.19(b) is due to the number of indicator variables that are consecutively sampled as 1, which are jointly sampled with the fixed effect parameters, and will effect the effective sample size, (A.10), for these parameters. However, as the Metropolis-Hastings within Gibbs sampling algorithm identifies the correct parameters as having a high posterior probability of being active we are not too concerned about the effective sample size for this parameter.

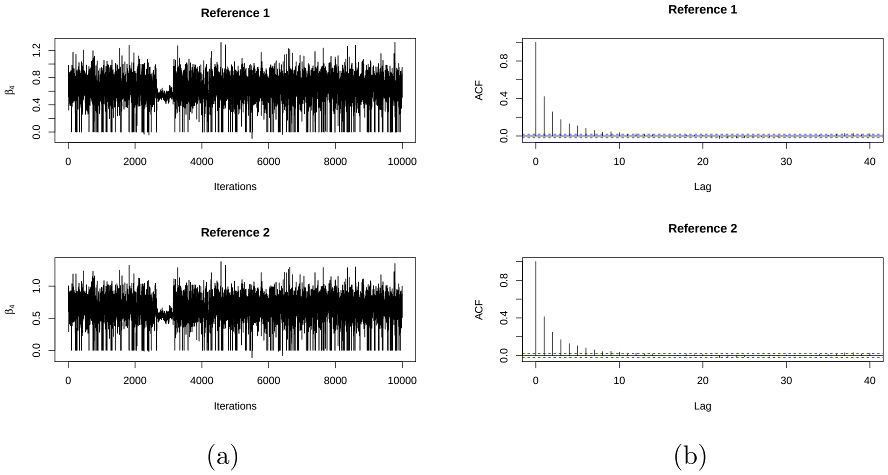


Figure A.18: Example (a) trace and (b) ACF plot for the Markov chain formed by sampling $\boldsymbol{\beta}_j$, $j = 1, \ldots, p$ using (4.36) in from the Metropolis-Hastings within Gibbs sampling algorithm (Section 4.4.4) for a when $\boldsymbol{\beta}_j$ has a high posterior probability of being active for the experimental responses from Chapter 5. These plots are for the samples of $\boldsymbol{\beta}_4$ for Test 2 and Prior 1.

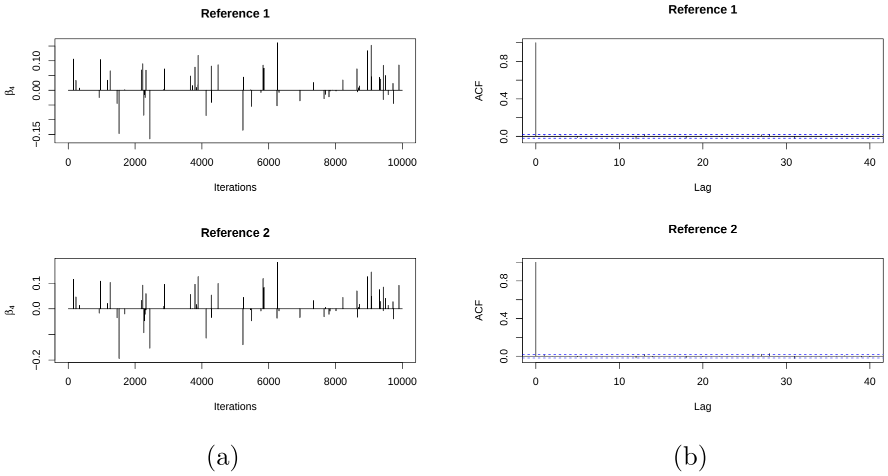
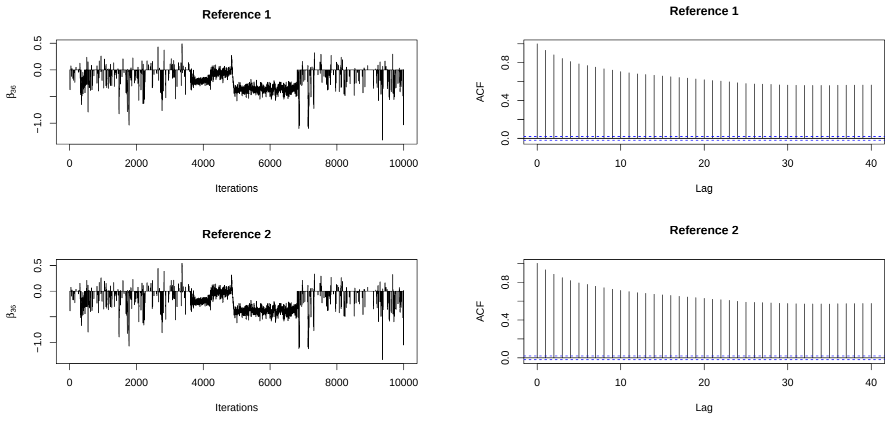


Figure A.19: Example (a) trace and (b) ACF plot for the Markov chains formed by sampling $\boldsymbol{\beta}_j$, $j = 1, \ldots, p$ using (4.36) in the Metropolis-Hastings within Gibbs sampling algorithm (Section 4.4.4) when $\boldsymbol{\beta}_j$ has a high posterior probability of not being active for the experimental responses from Chapter 5. These plots are for the samples of $\boldsymbol{\beta}_4$ for Test 4 and Prior 1.



(a) (b)

Figure A.20: Example unusual (a) trace and (b) ACF plot for the Markov chains formed by sampling $\boldsymbol{\beta}_j$, $j = 1, \ldots, p$ using (4.36) in the Metropolis-Hastings within Gibbs sampling algorithm (Section 4.4.4) for the experimental responses from Test 2 from Chapter 5. These plots are for the samples of $\boldsymbol{\beta}_{36}$ for Test 2 and Prior 2.

Figure A.20 is an example of some of the unusual trace and ACF plots seen for $\boldsymbol{\beta}_j$ for Test 2 in Matthews (2015). Notice that there is a shift in the mean of Markov chain for this parameter in Figure A.20(a), which occurs due to a prolonged period where $\boldsymbol{\beta}_j$ is sampled from the normal distribution and not as $\mathbf{0}_r^T$. The correlations in Figure A.20(b), the ACF plot for this $\boldsymbol{\beta}_j$, are significantly higher than we would like to see and would impact the effective sample size for these parameter.

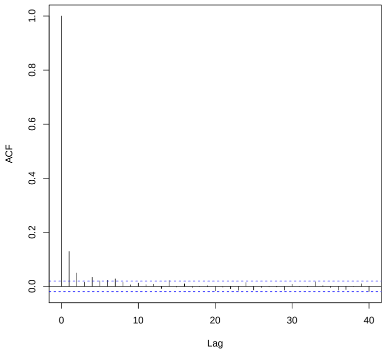
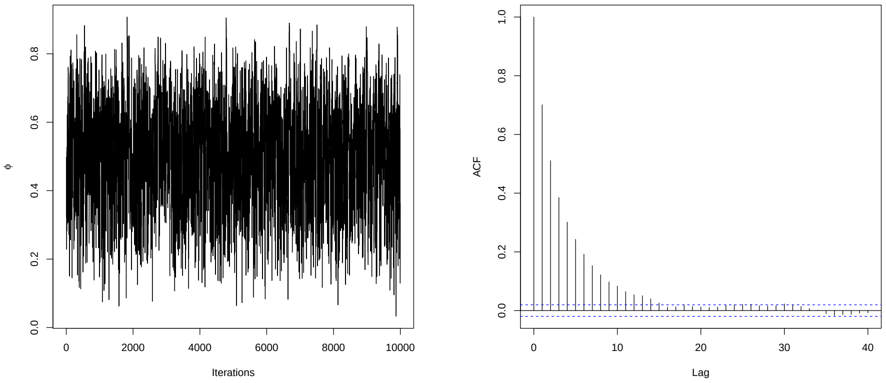


Figure A.21: Example ACF plot for the Markov chain formed by sampling c from (4.34) in the Metropolis-Hastings within Gibbs sampling algorithm (Section 4.4.4) for the experimental responses from Chapter 5. This plots is for the samples of c for Test 1 and Prior 1.

Figure A.21 is an example ACF plot for a Markov chain formed by sampling c from (4.34). As the correlations in this plot are low, and decrease as the lags increase, this Markov chain can be suggested to have good convergence properties.

Figure A.22 is an example trace and ACF plot for the Markov chain formed by sampling the correlation parameter ϕ using Metropolis-Hastings rejection sampling. The trace plot, Figure A.22(a), shows that the Metropolis-Hastings rejection sampling is finding values across the range and is not getting stuck at a particular proposed value. However, as discussed in Section C.2 for Chapter 4, the ACF plot, Figure A.22(b), does display some high correlations at small lags.



(a)

(b)

Figure A.22: Example (a) trace and (b) ACF plot for the Markov chain formed by sampling ϕ using Metropolis-Hastings sampling on (4.35) in the Metropolis-Hastings within Gibbs sampling algorithm (Section 4.4.4) for the experimental responses from Chapter 5. These plots are for the samples of ϕ for Test 1 and Prior 1.

As in Section C.2, we can use the estimated effective sample size and the acceptance rate to further assess the Markov chains for ϕ. The estimated effective sample sizes are given in Table A.4, and we note that they are particularly low for Test 2 and Prior 2, and Test 4 and Prior 1. In future work, an alternative proposal distribution should be considered, and an assessment of the acceptance rates for this proposal distribution should be made to see how it compares to the proposal distribution assumed in this work.

Response	Effective Sample Size
Test 1, Prior 1	1410
Test 1, Prior 2	581
Test 2, Prior 1	1342
Test 2, Prior 2	149
Test 3, Prior 1	466
Test 3, Prior 2	683
Test 4, Prior 1	255
Test 4, Prior 2	867

Table A.4: Estimated effective sample size, rounded to nearest whole number, of samples of ϕ for the responses from the experiment in Chapter 5.

The acceptance rates in Table A.5 are all acceptable, however we note that the Markov chains with low estimated effective sample size also have low acceptance rates. Again, future work could be done to assess the improvements that could be made by considering an alternative proposal distribution.

Response	Acceptance Rate
Test 1, Prior 1	0.48
Test 1, Prior 2	0.30
Test 2, Prior 1	0.38
Test 2, Prior 2	0.15
Test 3, Prior 1	0.30
Test 3, Prior 2	0.39
Test 4, Prior 1	0.15
Test 4, Prior 2	0.38

Table A.5: Acceptance rate, rounded to 2dp, of samples of ϕ for the responses from the experiment in Chapter 5.

Figure A.23 is the example trace and ACF plot of the Markov chain formed by sampling the scale matrix $\mathbf{\Sigma}$ from (4.33) when Prior 1 is assumed, and Figure A.24 is the trace and ACF plot for $\mathbf{\Sigma}$ when Prior 2 is assumed. Figures A.23 and A.24 are similar to Figures A.15 and A.16 in Section C.2, and the spikes in Figures A.23(a) and A.24(a) are expected as they occur when we consider samples with very few active $\boldsymbol{\beta}_j$.

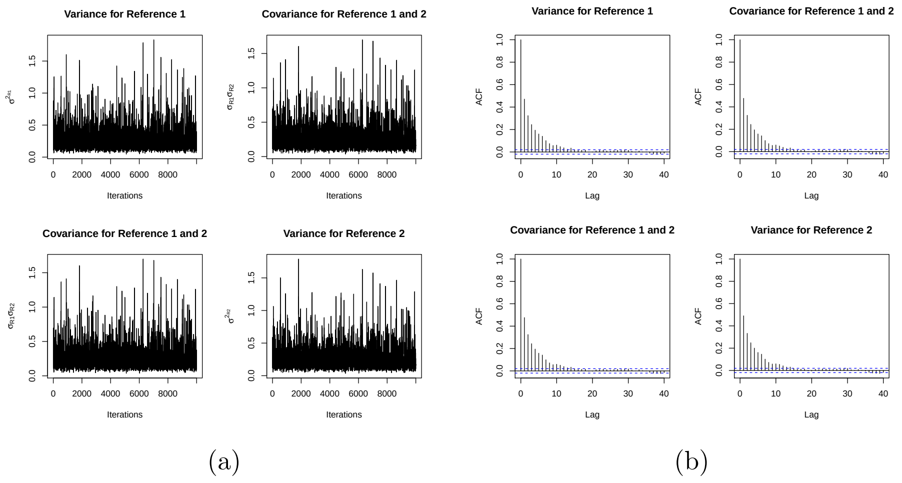
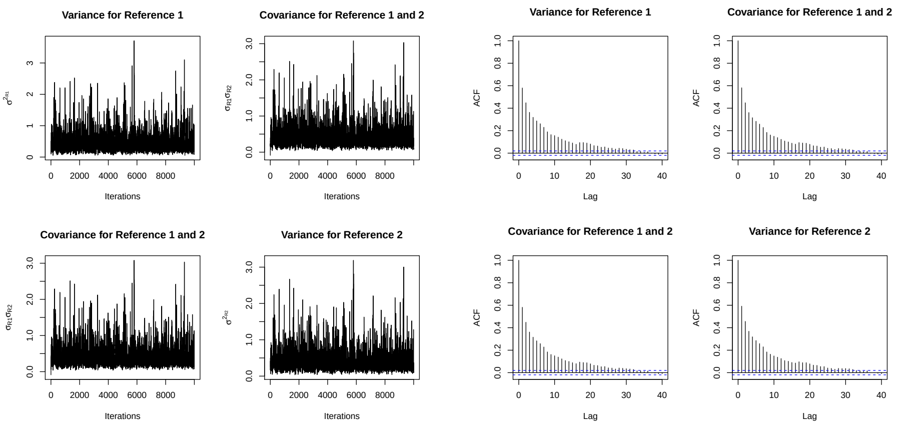


Figure A.23: Example (a) trace and (b) ACF plot for the Markov chain formed by sampling $\mathbf{\Sigma}$ from (4.33) in the Metropolis-Hastings within Gibbs sampling algorithm (Section 4.4.4) with Prior 1 for the experimental responses from Chapter 5. These plots are for the samples of $\boldsymbol{\beta}_{36}$ for Test 1 and Prior 1.



(a) (b)

Figure A.24: Example (a) trace and (b) ACF plot for the Markov chain formed by sampling $\mathbf{\Sigma}$ from (4.33) in the Metropolis-Hastings within Gibbs sampling algorithm (Section 4.4.4) with Prior 2 for the experimental responses from Chapter 5. These plots are for the samples of $\boldsymbol{\beta}_{36}$ for Test 1 and Prior 2.

Response	Effective Sample Size
Test 1, Prior 1	$\begin{pmatrix} 1976 & 1966 \\ 1966 & 1917 \end{pmatrix}$
Test 1, Prior 2	$\begin{pmatrix} 938 & 976 \\ 976 & 958 \end{pmatrix}$
Test 2, Prior 1	$\begin{pmatrix} 783 & 780 \\ 780 & 771 \end{pmatrix}$
Test 2, Prior 2	$\begin{pmatrix} 106 & 106 \\ 106 & 106 \end{pmatrix}$
Test 3, Prior 1	$\begin{pmatrix} 549 & 541 \\ 541 & 533 \end{pmatrix}$
Test 3, Prior 2	$\begin{pmatrix} 601 & 598 \\ 598 & 595 \end{pmatrix}$
Test 4, Prior 1	$\begin{pmatrix} 337 & 352 \\ 352 & 406 \end{pmatrix}$
Test 4, Prior 2	$\begin{pmatrix} 795 & 722 \\ 722 & 717 \end{pmatrix}$

Table A.6: Estimated effective sample size, rounded to nearest whole number, of samples of the four elements of $\mathbf{\Sigma}$ for the responses from the experiment in Chapter 5.

As in Section C.2, we use the effective sample size to assess the impact of the high correlation present in Figures A.23(b) and A.24(b). Table A.6 gives the effect sample sizes of the four elements of $\mathbf{\Sigma}$ for all the tests and prior distributions for ϕ considered in Chapter 5. As in Table A.4, Test 2 and Prior 2, and Test 4 with Prior 1 has low effective sample sizes, and in future work we should considering a different prior distribution for

$\mathbf{\Sigma}$ to see if we could reduce the correlation in the Markov chains for the elements of $\mathbf{\Sigma}$ and hence increase the effective sample sizes.

D Matrix Distributions

In this appendix we introduce two matrix distributions, the matrix normal and inverse Wishart distribution, which are used in the Metropolis-Hastings within Gibbs sampling algorithm in Section 4.4.4 of Chapter 4. We also provide proof of (4.22) from Section 4.4.1 of Chapter 4 using the properties of the matrix normal distribution.

D.1 Matrix Normal Distribution

If a $n \times r$ matrix $\mathbf{Y}$ is matrix normally distributed with mean $\boldsymbol{\mu}$, between row scale matrix $\mathbf{\Sigma}_1$ and between column scale matrix $\mathbf{\Sigma}_2$, then $\mathbf{Y}$ has probability density function

$$p(\mathbf{Y}|\boldsymbol{\mu}, \mathbf{\Sigma}_1, \mathbf{\Sigma}_2) = (2\pi)^{-\frac{nr}{2}} |\mathbf{\Sigma}_1|^{-\frac{r}{2}} |\mathbf{\Sigma}_2|^{-\frac{n}{2}} \exp\left[-\frac{1}{2}\text{tr}\left\{\mathbf{\Sigma}_2^{-1}(\mathbf{Y}-\boldsymbol{\mu})^T\mathbf{\Sigma}_1^{-1}(\mathbf{Y}-\boldsymbol{\mu})\right\}\right]. \tag{A.11}$$

Note that

$$\begin{aligned}
&(2\pi)^{-\frac{nr}{2}} |\mathbf{\Sigma}_1|^{-\frac{r}{2}} |\mathbf{\Sigma}_2|^{-\frac{n}{2}} \exp\left[-\tfrac{1}{2}\text{tr}\left\{\mathbf{\Sigma}_2^{-1}(\mathbf{Y}-\boldsymbol{\mu})^T\mathbf{\Sigma}_1^{-1}(\mathbf{Y}-\boldsymbol{\mu})\right\}\right] \\
&= (2\pi)^{-\frac{nr}{2}} |\mathbf{\Sigma}_1|^{-\frac{r}{2}} |\mathbf{\Sigma}_2|^{-\frac{n}{2}} \exp\left[-\tfrac{1}{2}\text{tr}\left\{(\mathbf{Y}-\boldsymbol{\mu})^T\mathbf{\Sigma}_1^{-1}(\mathbf{Y}-\boldsymbol{\mu})\mathbf{\Sigma}_2^{-1}\right\}\right] \\
&= (2\pi)^{-\frac{nr}{2}} |\mathbf{\Sigma}_1|^{-\frac{r}{2}} |\mathbf{\Sigma}_2|^{-\frac{n}{2}} \exp\left[-\tfrac{1}{2}\text{vec}\left(\mathbf{Y}-\boldsymbol{\mu}\right)^T \text{vec}\left\{\mathbf{\Sigma}_1^{-1}(\mathbf{Y}-\boldsymbol{\mu})\mathbf{\Sigma}_2^{-1}\right\}\right] \\
&= (2\pi)^{-\frac{nr}{2}} |\mathbf{\Sigma}_1|^{-\frac{r}{2}} |\mathbf{\Sigma}_2|^{-\frac{n}{2}} \exp\left[-\tfrac{1}{2}\left\{\text{vec}(\mathbf{Y})-\text{vec}(\boldsymbol{\mu})\right\}^T\left(\mathbf{\Sigma}_2^{-1}\otimes\mathbf{\Sigma}_1^{-1}\right)\left\{\text{vec}(\mathbf{Y})-\text{vec}(\boldsymbol{\mu})\right\}\right] \\
&= (2\pi)^{-\frac{nr}{2}} |\mathbf{\Sigma}_2\otimes\mathbf{\Sigma}_1|^{-\frac{1}{2}} \exp\left[-\tfrac{1}{2}\left\{\text{vec}(\mathbf{Y})-\text{vec}(\boldsymbol{\mu})\right\}^T\left(\mathbf{\Sigma}_2\otimes\mathbf{\Sigma}_1\right)^{-1}\left\{\text{vec}(\mathbf{Y})-\text{vec}(\boldsymbol{\mu})\right\}\right].
\end{aligned}$$

Hence,

$$\text{vec}(\mathbf{Y}) \sim \text{N}(\text{vec}(\boldsymbol{\mu}), \mathbf{\Sigma}_2\otimes\mathbf{\Sigma}_1), \tag{A.12}$$

where $\text{vec}(\mathbf{Y})$ is the $nr \times 1$ vector of column-stacked entries of $\mathbf{Y}$.

Derivation of (4.22) from Section 4.4.1 of Chapter 4

We want to show that if the $n \times r$ matrix of responses $\mathbf{Y}$ is assumed to follow the multivariate linear mixed effects model (4.21), where $\mathbf{\Gamma} \sim \text{MN}(\mathbf{0}_{n_w r}, \phi\mathbf{I}_{n_w}, \mathbf{\Sigma})$ and $\mathbf{E} \sim$

$\text{MN}(\mathbf{0}_{nr}, (1-\phi)\mathbf{I}_n, \mathbf{\Sigma})$ are independent and matrix normally distributed, then (4.22) holds.

As $\mathbf{E}$ and $\mathbf{\Gamma}$ are independent, if $\mathbf{Y}$ is matrix normally distributed then, using (A.12), the marginal distribution of $\text{vec}(\mathbf{Y})$ will be a normal distribution with

$$E\{\text{vec}(\mathbf{Y})\} = \text{vec}(\mathbf{XB}) + E\{\text{vec}(\mathbf{Z\Gamma})\} + E\{\text{vec}(\mathbf{E})\},$$

and

$$\text{Var}\{\text{vec}(\mathbf{Y})\} = \text{Var}\{\text{vec}(\mathbf{Z\Gamma})\} + \text{Var}\{\text{vec}(\mathbf{E})\}.$$

As $\mathbf{\Gamma} \sim \text{MN}(\mathbf{0}_{n_w r}, \phi\mathbf{I}_{n_w}, \mathbf{\Sigma})$, then

$$\mathbf{Z\Gamma} \sim \text{MN}(\mathbf{0}_{nr}, \phi\mathbf{ZZ}^T, \mathbf{\Sigma}),$$

hence, by (A.12),

$$\text{vec}(\mathbf{Z\Gamma}) \sim \text{N}(\mathbf{0}_{nr}, \mathbf{\Sigma} \otimes \phi\mathbf{ZZ}^T).$$

As $\mathbf{E} \sim \text{MN}(\mathbf{0}_{nr}, (1-\phi)\mathbf{I}_n, \mathbf{\Sigma})$, then by (A.12)

$$\text{vec}(\mathbf{E}) \sim \text{N}(\mathbf{0}_{nr}, \mathbf{\Sigma} \otimes (1-\phi)\mathbf{I}_n).$$

Therefore,

$$E\{\text{vec}(\mathbf{Y})\} = \text{vec}(\mathbf{XB})$$

and

$$\begin{aligned}
\text{Var}\{\text{vec}(\mathbf{Y})\} &= (\mathbf{\Sigma} \otimes \phi\mathbf{ZZ}^T) + (\mathbf{\Sigma} \otimes (1-\phi)\mathbf{I}_n) \\
&= \mathbf{\Sigma} \otimes (\phi\mathbf{ZZ}^T + (1-\phi)\mathbf{I}_n) \\
&= \mathbf{\Sigma} \otimes [\{\phi(\mathbf{I}_{n_w} \otimes \mathbf{J}_{n_s})\} + \{(1-\phi)(\mathbf{I}_{n_w} \otimes \mathbf{I}_{n_s})\}] \\
&= \mathbf{\Sigma} \otimes [\mathbf{I}_{n_w} \otimes \{\phi\mathbf{J}_{n_s} + (1-\phi)\mathbf{I}_{n_s}\}] \\
&= \mathbf{\Sigma} \otimes \mathbf{V}(\phi).
\end{aligned}$$

Hence,

$$\text{vec}(\mathbf{Y}) \sim \text{N}(\text{vec}(\mathbf{XB}), \boldsymbol{\Sigma} \otimes \mathbf{V}(\phi)),$$

and therefore, using (A.12), the marginal distribution of $\mathbf{Y}$ is $\text{MN}(\mathbf{XB}, \mathbf{V}(\phi), \boldsymbol{\Sigma})$, as required.

D.2 Inverse Wishart Distribution

If a $r \times r$ matrix $\boldsymbol{\Sigma}$ is from an inverse Wishart distribution with d degrees of freedom and scale matrix $\mathbf{S}$, then $\boldsymbol{\Sigma}$ has probability density function

$$p(\boldsymbol{\Sigma}|\mathbf{S}, d) = \frac{|\mathbf{S}|^{\frac{d}{2}}|\boldsymbol{\Sigma}|^{-\frac{d+r+1}{2}}}{2^{\frac{dr}{2}}\Gamma_r\left(\frac{d}{2}\right)} \exp\left\{-\frac{1}{2}\text{tr}(\mathbf{S}\boldsymbol{\Sigma}^{-1})\right\}. \tag{A.13}$$

E Conditional Distributions for Metropolis-Hastings within Gibbs Sampling Algorithm

In this Section we demonstrate how (4.27) is calculated (Section E.1), and show how (4.22) from Section 4.4.1 and the prior distributions from Section 4.4.2 can be used to find the full conditional distributions for $\boldsymbol{\beta}_0$ (Section E.2), $\boldsymbol{\Sigma}$ (Section E.3), ϕ (Section E.4) and c (Section E.5) which are given in Section 4.4.3 and sampled from in the Metropolis-Hastings within Gibbs sampling algorithm in Section 4.4.4 of Chapter 4. The majority of the calculations in this appendix use (4.2) from Section 4.2.1.

E.1 Prior Distribution for B

Let the fixed effects matrix, $\mathbf{B}$, from (4.21) be divided into two matrices, $\mathbf{B}_A$ and $\mathbf{B}_{NA}$; where $\mathbf{B}_A$ is the $p_A \times r$ matrix formed from the p_A rows of $\mathbf{B}$ where $\boldsymbol{\beta}_{j_A}$, $j_A = 1, \ldots, p_A$ is active, and $\mathbf{B}_{NA}$ is the $p_{NA} \times r$ matrix of formed from the p_{NA} rows of $\mathbf{B}$ where $\boldsymbol{\beta}_{j_{NA}}$, $j_{NA} = 1, \ldots, p_{NA}$ is not active. Recall that $\boldsymbol{\beta}_{j_A}^T$ are assumed to be independent with distribution $\text{N}(\mathbf{0}_r, c\boldsymbol{\Sigma})$, hence $\mathbf{B}_A \sim \text{MN}(\mathbf{0}_{p_A r}, \mathbf{I}_{p_A}, c\boldsymbol{\Sigma})$, and that $\boldsymbol{\beta}_{j_{NA}}^T = \mathbf{0}_r$, where $\mathbf{0}_r$ is the $1 \times r$ vector with every element as 0, hence $p(\mathbf{B}_{NA}|\boldsymbol{\Sigma}, c, \boldsymbol{\delta}) \propto 1$.

If we assume that $\mathbf{B}_A$ and $\mathbf{B}_{NA}$ are independent, then the prior distribution for $\mathbf{B}$ is proportional to

$$\begin{aligned} p(\mathbf{B}|\boldsymbol{\Sigma}, c, \boldsymbol{\delta}) &= p(\mathbf{B}_A|\boldsymbol{\Sigma}, c, \boldsymbol{\delta})p(\mathbf{B}_{NA}|\boldsymbol{\Sigma}, c, \boldsymbol{\delta}) \\ &\propto |c\boldsymbol{\Sigma}|^{-\frac{p_A}{2}}|I_{p_A}|^{-\frac{r}{2}} \exp\left[-\frac{1}{2}\text{tr}\{(c\boldsymbol{\Sigma})^{-1}\mathbf{B}_A^T\mathbf{I}_{p_A}\mathbf{B}_A\}\right]. \end{aligned} \tag{A.14}$$

We note that $\mathbf{B}_A^T\mathbf{I}_{p_A}\mathbf{B}_A = \mathbf{B}^T\mathbf{B}$ and $p_A = \sum_{j=1}^p \delta_j$, hence (A.14) can be written as

$$|c\boldsymbol{\Sigma}|^{-\frac{\sum_{j=1}^p \delta_j}{2}} \exp\left[-\frac{1}{2}\mathrm{tr}\{(c\boldsymbol{\Sigma})^{-1}\mathbf{B}^T\mathbf{B}\}\right]. \tag{A.15}$$

E.2 Conditional Distribution for $\boldsymbol{\beta}_0$

The prior distribution for the $r \times 1$ intercept vector $\boldsymbol{\beta}_0$ is $p(\boldsymbol{\beta}_0) \propto 1$. To find the conditional distribution for $\boldsymbol{\beta}_0$ we begin by rearranging the likelihood and then find the maximum likelihood estimate of $\boldsymbol{\beta}_0$. We know from (4.22) that

$$\begin{aligned}
p(\mathbf{Y}|\boldsymbol{\beta}_0,\dots) &\propto \exp\left[-\tfrac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}((\mathbf{Y}-\mathbf{XB})-\mathbf{1}_n\boldsymbol{\beta}_0^T)^T\mathbf{V}(\phi)^{-1}((\mathbf{Y}-\mathbf{XB})-\mathbf{1}_n\boldsymbol{\beta}_0^T)\right\}\right] \\
&= \exp\left[-\tfrac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(\dot{\mathbf{Y}}-\mathbf{1}_n\boldsymbol{\beta}_0^T)^T\mathbf{V}(\phi)^{-1}(\dot{\mathbf{Y}}-\mathbf{1}_n\boldsymbol{\beta}_0^T)\right\}\right] \\
&= \exp\left[-\tfrac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\left(\dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\dot{\mathbf{Y}} + \boldsymbol{\beta}_0\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n\boldsymbol{\beta}_0^T - \boldsymbol{\beta}_0\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\dot{\mathbf{Y}}\right.\right.\right. \\
&\qquad\left.\left.\left. -\dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n\boldsymbol{\beta}_0^T\right)\right\}\right] \\
&= \exp\left\{-\mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\dot{\mathbf{Y}}\right) - \mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n\boldsymbol{\beta}_0^T\right)\right. \\
&\qquad\left. +\mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\dot{\mathbf{Y}}\right) + \mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n\boldsymbol{\beta}_0^T\right)\right\},
\end{aligned}$$

where $\dot{\mathbf{Y}} = \mathbf{Y} - \mathbf{XB}$.

To find the maximum likelihood estimate for $\boldsymbol{\beta}_0$, we calculate

$$\frac{\partial}{\partial\boldsymbol{\beta}_0}\log(p(\mathbf{Y}|\boldsymbol{\beta}_0,\dots)) = \mathbf{0} \tag{A.16}$$

where

$$\begin{aligned}
\log\{p(\mathbf{Y}|\boldsymbol{\beta}_0,\dots)\} \propto &-\mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\dot{\mathbf{Y}}\right) - \mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n\boldsymbol{\beta}_0^T\right) \\
&+\mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\dot{\mathbf{Y}}\right) + \mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n\boldsymbol{\beta}_0^T\right).
\end{aligned}$$

As $\mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\dot{\mathbf{Y}}\right)$ is a constant with respect to $\boldsymbol{\beta}_0$,

$$\begin{aligned}
\tfrac{\partial}{\partial\boldsymbol{\beta}_0}\log\{p(\mathbf{Y}|\boldsymbol{\beta}_0,\dots)\} \propto &-\tfrac{\partial}{\partial\boldsymbol{\beta}_0}\left\{\mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n\boldsymbol{\beta}_0^T\right)\right\} \\
&+\tfrac{\partial}{\partial\boldsymbol{\beta}_0}\left\{\mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\dot{\mathbf{Y}}\right)\right\} + \tfrac{\partial}{\partial\boldsymbol{\beta}_0}\left\{\mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n\boldsymbol{\beta}_0^T\right)\right\}.
\end{aligned}$$

Calculating these partial derivatives individually gives,

$$-\tfrac{\partial}{\partial\boldsymbol{\beta}_0}\left\{\mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n\boldsymbol{\beta}_0^T\right)\right\} = -\tfrac{\partial}{\partial\boldsymbol{\beta}_0}\left\{\mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n\boldsymbol{\beta}_0^T\mathbf{I}_1\right)\right\}$$

$$
\begin{aligned}
&= -\tfrac{1}{2}\left\{(\boldsymbol{\Sigma}^{-1})^T(\mathbf{I}_1)^T\boldsymbol{\beta}_0\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n + \mathbf{I}_1\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n\right\}\\
&= -\tfrac{1}{2}\left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n + \boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n\right\}\\
&= -\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n.
\end{aligned}
$$

$$
\tfrac{\partial}{\partial\boldsymbol{\beta}_0}\left\{\mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\dot{\mathbf{Y}}\right)\right\} = \tfrac{1}{2}(\boldsymbol{\Sigma}^{-1})^T\dot{\mathbf{Y}}^T[\{\mathbf{V}(\phi)\}^{-1}]^T\mathbf{1}_n = \tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n.
$$

$$
\tfrac{\partial}{\partial\boldsymbol{\beta}_0}\left\{\mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n\boldsymbol{\beta}_0^T\right)\right\} = \tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n.
$$

Therefore,

$$
\begin{aligned}
\tfrac{\partial}{\partial\boldsymbol{\beta}_0}\log\{p(\mathbf{Y}|\boldsymbol{\beta}_0,\dots)\} = \mathbf{0} &\implies -\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\beta}}_0\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n + \boldsymbol{\Sigma}^{-1}\dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n = \mathbf{0}\\
&\implies \boldsymbol{\Sigma}^{-1}\left(-\hat{\boldsymbol{\beta}}_0\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n + \dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n\right) = \mathbf{0}\\
&\implies \hat{\boldsymbol{\beta}}_0\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n = \dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n\\
&\implies \hat{\boldsymbol{\beta}}_0 = \tfrac{\dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n}{\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n}.
\end{aligned}
$$

The conditional distribution for $\boldsymbol{\beta}_0$ is given by

$$
\begin{aligned}
&p(\boldsymbol{\beta}_0|\mathbf{Y},\dots) \propto p(\mathbf{Y}|\boldsymbol{\beta}_0,\dots)p(\boldsymbol{\beta}_0)\\
&\propto \exp\Big\{-\mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\dot{\mathbf{Y}}\right) - \mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n\boldsymbol{\beta}_0^T\right) + \mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\dot{\mathbf{Y}}\right)\\
&\qquad +\mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n\boldsymbol{\beta}_0^T\right)\Big\}\\
&= \exp\Big\{-\mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\dot{\mathbf{Y}}\right) - \mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n\boldsymbol{\beta}_0^T\right) + \mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\dot{\mathbf{Y}}\right)\\
&\qquad +\mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n\boldsymbol{\beta}_0^T\right)+\tfrac{1}{2}\left(\tfrac{\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\dot{\mathbf{Y}}}{\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n}\right)(\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n)\boldsymbol{\Sigma}^{-1}\left(\tfrac{\dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n}{\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n}\right)\\
&\qquad -\tfrac{1}{2}\left(\tfrac{\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\dot{\mathbf{Y}}}{\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n}\right)(\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n)\boldsymbol{\Sigma}^{-1}\left(\tfrac{\dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n}{\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n}\right)\Big\}\\
&= \exp\left\{-\mathrm{tr}\left(\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\dot{\mathbf{Y}}\right) + \tfrac{1}{2}\left(\tfrac{\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\dot{\mathbf{Y}}}{\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n}\right)(\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n)\boldsymbol{\Sigma}^{-1}\left(\tfrac{\dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n}{\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n}\right)\right\}\\
&\quad\times\exp\Big[-\mathrm{tr}\left\{\tfrac{1}{2}(\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n)\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0\boldsymbol{\beta}_0^T\right\} + \mathrm{tr}\left\{\tfrac{1}{2}(\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n)\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0\left(\tfrac{\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\dot{\mathbf{Y}}}{\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n}\right)\right\}\\
&\qquad +\mathrm{tr}\left\{\tfrac{1}{2}(\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n)\boldsymbol{\Sigma}^{-1}\left(\tfrac{\dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n}{\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n}\right)\boldsymbol{\beta}_0^T\right\}\\
&\qquad -\tfrac{1}{2}\left(\tfrac{\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\dot{\mathbf{Y}}}{\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n}\right)(\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n)\boldsymbol{\Sigma}^{-1}\left(\tfrac{\dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n}{\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n}\right)\Big]\\
&\propto \exp\Big[-\mathrm{tr}\left\{\tfrac{1}{2}(\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n)\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0\boldsymbol{\beta}_0^T\right\} + \mathrm{tr}\left\{\tfrac{1}{2}(\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n)\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0\hat{\boldsymbol{\beta}}_0^T\right\}
\end{aligned}
$$

$$
\begin{aligned}
&\qquad +\mathrm{tr}\left\{\tfrac{1}{2}(\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n)\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\beta}}_0\boldsymbol{\beta}_0^T\right\} - \tfrac{1}{2}\hat{\boldsymbol{\beta}}_0^T(\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n)\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\beta}}_0\Big]\\
&= \exp\Big[-\mathrm{tr}\left\{\tfrac{1}{2}\boldsymbol{\beta}_0^T(\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n)\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0\right\} + \mathrm{tr}\left\{\tfrac{1}{2}\hat{\boldsymbol{\beta}}_0^T(\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n)\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0\right\}\\
&\qquad +\mathrm{tr}\left\{\tfrac{1}{2}\boldsymbol{\beta}_0^T(\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n)\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\beta}}_0\right\} - \tfrac{1}{2}\hat{\boldsymbol{\beta}}_0^T(\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n)\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\beta}}_0\Big]\\
&= \exp\Big[-\tfrac{1}{2}\Big\{\boldsymbol{\beta}_0^T(\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n)\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0 - \boldsymbol{\beta}_0^T(\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n)\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\beta}}_0\\
&\qquad -\hat{\boldsymbol{\beta}}_0^T(\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n)\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_0 + \hat{\boldsymbol{\beta}}_0^T(\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n)\boldsymbol{\Sigma}^{-1}\hat{\boldsymbol{\beta}}_0\Big\}\Big]\\
&= \exp\left[-\tfrac{1}{2}\left\{\left(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}_0\right)^T(\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n)\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\beta}_0-\hat{\boldsymbol{\beta}}_0\right)\right\}\right],
\end{aligned}
$$

which is proportional to the normal density function of a normally distributed random vector with mean $\hat{\boldsymbol{\beta}}_0$ and variance-covariance matrix $(\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n)\boldsymbol{\Sigma}^{-1}$, therefore

$$
\boldsymbol{\beta}_0|\mathbf{Y},\mathbf{B},\boldsymbol{\Sigma},\boldsymbol{\delta},c,\phi \sim N\left(\hat{\boldsymbol{\beta}}_0, \frac{\boldsymbol{\Sigma}}{(\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n)}\right), \tag{A.17}
$$

where

$$
\hat{\boldsymbol{\beta}}_0 = \frac{\dot{\mathbf{Y}}^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n}{(\mathbf{1}_n^T\mathbf{V}(\phi)^{-1}\mathbf{1}_n)}. \tag{A.18}
$$

E.3 Conditional Distribution for $\boldsymbol{\Sigma}$

The prior distribution for the between column scale matrix $\boldsymbol{\Sigma}$ is $\mathrm{IW}(\mathbf{0}_{rr}, -r+1)$. This is the multivariate extension of the prior distribution given by Tan and Wu (2013) and is the prior distribution given by Overstall and Woods (2015). The prior distribution for $\mathbf{B}$ is dependent on $\boldsymbol{\Sigma}$, hence

$$
\begin{aligned}
&p(\boldsymbol{\Sigma}|\mathbf{Y},\dots) = p(\mathbf{Y}|\boldsymbol{\Sigma},\dots)p(\boldsymbol{\Sigma})p(\mathbf{B})\\
&\quad\propto |\boldsymbol{\Sigma}|^{-\frac{n}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}(r+1-r+1)}|\boldsymbol{\Sigma}|^{-\frac{\sum_{j=1}^{p}\delta_j}{2}}\exp\left\{-\tfrac{1}{2}\mathrm{tr}\left(\tfrac{1}{c}\boldsymbol{\Sigma}^{-1}\mathbf{B}^T\mathbf{B}\right)\right\}\times\\
&\qquad\exp\left[-\tfrac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(\mathbf{Y}-\mathbf{1}_n\boldsymbol{\beta}_0^T-\mathbf{X}\mathbf{B})^T\mathbf{V}(\phi)^{-1}(\mathbf{Y}-\mathbf{1}_n\boldsymbol{\beta}_0^T-\mathbf{X}\mathbf{B})\right\}\right]\\
&\quad= |\boldsymbol{\Sigma}|^{-\frac{1}{2}\left(r+1-r+1+n+\sum_{j=1}^{p}\delta_j\right)}\\
&\qquad\quad\times\exp\left[-\tfrac{1}{2}\mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}\left((\mathbf{Y}-\mathbf{1}_n\boldsymbol{\beta}_0^T-\mathbf{X}\mathbf{B})^T\mathbf{V}(\phi)^{-1}(\mathbf{Y}-\mathbf{1}_n\boldsymbol{\beta}_0^T-\mathbf{X}\mathbf{B})+\tfrac{\mathbf{B}^T\mathbf{B}}{c}\right)\right\}\right]\\
&\quad= |\boldsymbol{\Sigma}|^{-\frac{1}{2}\left\{r+1+\left(-r+1+\sum_{j=1}^{p}\delta_j+n\right)\right\}}\exp\left\{-\tfrac{1}{2}\mathrm{tr}(\boldsymbol{\Sigma}^{-1}\mathbf{S}^*)\right\},
\end{aligned}
$$

where

$$
\mathbf{S}^* = (\mathbf{Y}-\mathbf{1}_n\boldsymbol{\beta}_0^T-\mathbf{X}\mathbf{B})^T\mathbf{V}(\phi)^{-1}(\mathbf{Y}-\mathbf{1}_n\boldsymbol{\beta}_0^T-\mathbf{X}\mathbf{B})+\frac{\mathbf{B}^T\mathbf{B}}{c}, \tag{A.19}
$$

which is proportional to the density of a matrix which has an inverse Wishart distribution with scale matrix $\mathbf{S}^*$ and $-r+1+\sum_{j=1}^p \delta_j + n$ degrees of freedom, therefore

$$\mathbf{\Sigma}|\mathbf{Y}, \boldsymbol{\beta}_0, \mathbf{B}, \boldsymbol{\delta}, c, \phi \sim \text{IW}\left(\mathbf{S}^*, -r+1+\sum_{j=1}^p \delta_j + n\right).$$

E.4 Conditional Distribution for ϕ

The likelihood (4.22) is also dependent on ϕ, hence for general $a, b > 0$,

$$\begin{aligned} p(\phi|\mathbf{Y}, \boldsymbol{\beta}_0, \mathbf{B}, \mathbf{\Sigma}, \boldsymbol{\delta}, c) &\propto |\mathbf{V}(\phi)|^{-\frac{r}{2}} \phi^{a-1}(1-\phi)^{b-1} \\ &\times \exp\left[-\tfrac{1}{2}\text{tr}\left\{\mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{1}_n\boldsymbol{\beta}_0^T - \mathbf{XB})^T\mathbf{V}(\phi)^{-1}(\mathbf{Y} - \mathbf{1}_n\boldsymbol{\beta}_0^T - \mathbf{XB})\right\}\right]. \end{aligned}$$

This distribution cannot be sampled from directly, hence we have to use Metropolis-Hastings rejection sampling where ϕ^* is drawn from a $\beta(a, b)$ distribution and the probability of accepting or rejecting this proposal, α, is the minimum of 1 and

$$\frac{|\mathbf{V}(\phi^*)|^{-\frac{m}{2}} \exp\left[-\tfrac{1}{2}\text{tr}\left\{\mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{1}_n\boldsymbol{\beta}_0^T - \mathbf{XB})^T\mathbf{V}(\phi^*)^{-1}(\mathbf{Y} - \mathbf{1}_n\boldsymbol{\beta}_0^T - \mathbf{XB})\right\}\right]}{|\mathbf{V}(\phi)|^{-\frac{m}{2}} \exp\left[-\tfrac{1}{2}\text{tr}\left\{\mathbf{\Sigma}^{-1}(\mathbf{Y} - \mathbf{1}_n\boldsymbol{\beta}_0^T - \mathbf{XB})^T\mathbf{V}(\phi)^{-1}(\mathbf{Y} - \mathbf{1}_n\boldsymbol{\beta}_0^T - \mathbf{XB})\right\}\right]}. \tag{A.20}$$

E.5 Conditional Distribution for c

The prior distribution for c is a uniform prior distribution with support $C = \{1/4, 9/16, 1, 4, 9, 16, 25\}$ (Tan and Wu, 2013). The prior distribution for $\mathbf{B}$ is also dependent on c hence

$$\begin{aligned} p(c|\mathbf{Y}, \dots) &\propto p(\mathbf{B})p(c) \\ &\propto \begin{cases} c^{-\frac{\sum_{j=1}^p \delta_j}{2}} \exp\left\{-\frac{1}{2c}\text{tr}(\mathbf{\Sigma}^{-1}\mathbf{B}^T\mathbf{B})\right\} \frac{1}{7} & \text{if } c \in C \\ 0 & \text{otherwise} \end{cases} \end{aligned}$$

$$\begin{aligned} p\left(c = \tfrac{1}{4}|\mathbf{Y}, \dots\right) &\propto \tfrac{1}{7}\left(\tfrac{1}{4}\right)^{-\frac{\sum_{j=1}^p \delta_j}{2}} \exp\left\{-\tfrac{1}{2(1/4)}\text{tr}(\mathbf{\Sigma}^{-1}\mathbf{B}^T\mathbf{B})\right\} \\ &= \frac{2^{\sum_{j=1}^p \delta_j}}{7} \exp\left\{-2\text{tr}(\mathbf{\Sigma}^{-1}\mathbf{B}^T\mathbf{B})\right\} \\ p\left(c = \tfrac{9}{16}|\mathbf{Y}, \dots\right) &\propto \tfrac{1}{7}\left(\tfrac{9}{16}\right)^{-\frac{\sum_{j=1}^p \delta_j}{2}} \exp\left\{-\tfrac{1}{2(9/16)}\text{tr}(\mathbf{\Sigma}^{-1}\mathbf{B}^T\mathbf{B})\right\} \\ &= \tfrac{1}{7}\left(\tfrac{4}{3}\right)^{\sum_{j=1}^p \delta_j} \exp\left\{-\tfrac{8}{9}\text{tr}(\mathbf{\Sigma}^{-1}\mathbf{B}^T\mathbf{B})\right\} \end{aligned}$$

$$
\begin{aligned}
p\left(c=1|\mathbf{Y},\dots\right) &\propto \tfrac{1}{7}\left(1\right)^{-\frac{\sum_{j=1}^{p}\delta_j}{2}} \exp\left\{-\tfrac{1}{(2\times 1)}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{B}^T\mathbf{B})\right\}\\
&= \tfrac{1}{7}\exp\left\{-\tfrac{1}{2}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{B}^T\mathbf{B})\right\}\\
p\left(c=4|\mathbf{Y},\dots\right) &\propto \tfrac{1}{7}\left(4\right)^{-\frac{\sum_{j=1}^{p}\delta_j}{2}} \exp\left\{-\tfrac{1}{(2\times 4)}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{B}^T\mathbf{B})\right\}\\
&= \tfrac{1}{7}\left(\tfrac{1}{2}\right)^{\sum_{j=1}^{p}\delta_j}\exp\left\{-\tfrac{1}{8}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{B}^T\mathbf{B})\right\}\\
p\left(c=9|\mathbf{Y},\dots\right) &\propto \tfrac{1}{7}\left(9\right)^{-\frac{\sum_{j=1}^{p}\delta_j}{2}} \exp\left\{-\tfrac{1}{(2\times 9)}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{B}^T\mathbf{B})\right\}\\
&= \tfrac{1}{7}\left(\tfrac{1}{3}\right)^{\sum_{j=1}^{p}\delta_j}\exp\left\{-\tfrac{1}{18}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{B}^T\mathbf{B})\right\}\\
p\left(c=16|\mathbf{Y},\dots\right) &\propto \tfrac{1}{7}\left(16\right)^{-\frac{\sum_{j=1}^{p}\delta_j}{2}} \exp\left\{-\tfrac{1}{(2\times 16)}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{B}^T\mathbf{B})\right\}\\
&= \tfrac{1}{7}\left(\tfrac{1}{4}\right)^{\sum_{j=1}^{p}\delta_j}\exp\left\{-\tfrac{1}{32}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{B}^T\mathbf{B})\right\}\\
p\left(c=25|\mathbf{Y},\dots\right) &\propto \tfrac{1}{7}\left(25\right)^{-\frac{\sum_{j=1}^{p}\delta_j}{2}} \exp\left\{-\tfrac{1}{(2\times 25)}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{B}^T\mathbf{B})\right\}\\
&= \tfrac{1}{7}\left(\tfrac{1}{5}\right)^{\sum_{j=1}^{p}\delta_j}\exp\left\{-\tfrac{1}{50}\mathrm{tr}(\mathbf{\Sigma}^{-1}\mathbf{B}^T\mathbf{B})\right\}
\end{aligned}
$$

Therefore

$$
p(c|\mathbf{Y},\boldsymbol{\beta}_0,\mathbf{B},\mathbf{\Sigma},\boldsymbol{\delta},\phi) = \begin{cases} \dfrac{1}{\sum_{c\in C} c^{-\frac{\sum_{j=1}^{p}\delta_j}{2}}\exp\left\{-\frac{1}{2c}\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{B}^T\mathbf{B}\right)\right\}} & \\ \times c^{-\frac{\sum_{j=1}^{p}\delta_j}{2}}\exp\left\{-\frac{1}{2c}\mathrm{tr}\left(\mathbf{\Sigma}^{-1}\mathbf{B}^T\mathbf{B}\right)\right\} & \text{if } c\in C\\ 0 & \text{otherwise.} \end{cases} \tag{A.21}
$$

F Extension of Joint Sampling Approach by Geweke (1996)

As discussed in George and McCulloch (1997) and Section 4.4.3 of Chapter 4, we cannot sample directly from the conditional distribution of $\boldsymbol{\beta}_j$, $j = 1,\dots,p$, which is the jth row of the fixed effect matrix $\mathbf{B}$ in (4.21), as the Markov chain for the conditional distribution of $\boldsymbol{\beta}_j$ is reducible and does not converge to the required posterior distribution. Therefore, we extend the approach given by Geweke (1996) to multivariate responses from split-plot experiments and jointly sample the indicator variable δ_j and $\boldsymbol{\beta}_j$.

Let $\mathbf{Y}_k$ be the $n_s \times r$ response matrix for whole-plot k, $k = 1,\dots,n_w$. Then

$$
\mathbf{Y}_k \sim \mathrm{MN}(\mathbf{1}_{n_s}\boldsymbol{\beta}_0^T + \mathbf{X}_k\mathbf{B}, \mathbf{V}(\phi)_k, \mathbf{\Sigma}) \tag{A.22}
$$

where $\mathbf{1}_{n_s}$ is the $n_s \times 1$ vector of ones, $\boldsymbol{\beta}_0$ is the $r \times 1$ vector of intercepts, $\mathbf{X}_k$ is the $n_s \times p$ model matrix for whole-plot k, $\mathbf{B}$ is the $p \times r$ matrix of fixed effects,

$$\mathbf{V}(\phi)_k = \begin{pmatrix} 1 & \phi & \dots & \phi \\ \phi & 1 & \dots & \phi \\ \vdots & \vdots & \dots & \vdots \\ \phi & \phi & \dots & \phi \end{pmatrix} = \phi \mathbf{J}_{n_s} + (1-\phi)\mathbf{I}_{n_s} \tag{A.23}$$

is the $n_s \times n_s$ row scale matrix for $\mathbf{Y}_k$ and $\boldsymbol{\Sigma}$ is $r \times r$ the column scale matrix for $\mathbf{Y}_k$. As in the paper by Geweke (1996), we let

$$\mathbf{Y}_{kj} = \mathbf{Y}_k - \sum_{l \neq j} \mathbf{X}_{kl} \boldsymbol{\beta}_l \tag{A.24}$$

be the $n_s \times r$ response matrix for whole-plot k and the jth term $\boldsymbol{\beta}_j$, where $\mathbf{X}_{kl}$ is the $n_s \times 1$ column of $\mathbf{X}_k$ relating to the $1 \times r$ vector $\boldsymbol{\beta}_l$. Therefore,

$$\mathbf{Y}_{kj} \sim \mathrm{MN}(\mathbf{X}_{kj}\boldsymbol{\beta}_j, \mathbf{V}(\phi)_k, \boldsymbol{\Sigma}), \tag{A.25}$$

and the likelihood of $\mathbf{Y}_{kj}$, conditional on $\boldsymbol{\beta}_j$, is

$$p(\mathbf{Y}_{kj}|\boldsymbol{\beta}_j^T) = \exp\left[-\frac{1}{2}\sum_{k=1}^{n_w} \mathrm{tr}\left\{\boldsymbol{\Sigma}^{-1}(\mathbf{Y}_{kj} - \mathbf{X}_{kj}\boldsymbol{\beta}_j)^T \mathbf{V}(\phi)_k^{-1}(\mathbf{Y}_{kj} - \mathbf{X}_{kj}\boldsymbol{\beta}_j)\right\}\right]. \tag{A.26}$$

The conditional distribution of $\boldsymbol{\beta}_j$ is

$$p(\mathbf{Y}_{kj}|\boldsymbol{\beta}_j^T)p(\boldsymbol{\beta}_j^T) \tag{A.27}$$

where $p(\boldsymbol{\beta}_j^T)$ is the prior for $\boldsymbol{\beta}_j^T$. Recall that $p(\boldsymbol{\beta}_j^T = \mathbf{0}_r) \propto 1$, hence

$$p(\boldsymbol{\beta}_j^T = \mathbf{0}_r|\mathbf{Y}_{kj}) \propto \exp\left\{\frac{1}{2}\sum_{k=1}^{n_w} \mathrm{tr}\left(\boldsymbol{\Sigma}^{-1}\mathbf{Y}_{kj}^T \mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}\right)\right\}. \tag{A.28}$$

When $\boldsymbol{\beta}_j$ is assumed to be active, $\boldsymbol{\beta}_j^T \sim \mathrm{N}(\mathbf{0}_r, c\boldsymbol{\Sigma})$. Therefore,

$$\begin{aligned} p(\boldsymbol{\beta}_j^T \neq \mathbf{0}_r|\mathbf{Y}_{kj}) \propto {} & \exp\left[-\tfrac{1}{2}\mathrm{tr}\left\{\textstyle\sum_{k=1}^{n_w} \boldsymbol{\Sigma}^{-1}(\mathbf{Y}_{kj} - \mathbf{X}_{kj}\boldsymbol{\beta}_j)^T \mathbf{V}(\phi)_k^{-1}(\mathbf{Y}_{kj} - \mathbf{X}_{kj}\boldsymbol{\beta}_j)\right\}\right] |c\boldsymbol{\Sigma}|^{-\frac{1}{2}} \\ & \times \exp\left(-\tfrac{1}{2}\boldsymbol{\beta}_j(c\boldsymbol{\Sigma})^{-1}\boldsymbol{\beta}_j^T\right) \end{aligned}$$

$$\begin{aligned}
&= |c\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\Big[-\tfrac{1}{2}\mathrm{tr}\Big\{\boldsymbol{\Sigma}^{-1}\Big(\textstyle\sum_{k=1}^{n_w}\mathbf{Y}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj} + \sum_{k=1}^{n_w}\boldsymbol{\beta}_j^T\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\boldsymbol{\beta}_j \\
&\qquad\qquad - \textstyle\sum_{k=1}^{n_w}\mathbf{Y}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\boldsymbol{\beta}_j - \sum_{k=1}^{n_w}\boldsymbol{\beta}_j^T\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}\Big)\Big\}\Big] \\
&\quad \times \exp\left(-\tfrac{1}{2}\boldsymbol{\beta}_j(c\boldsymbol{\Sigma})^{-1}\boldsymbol{\beta}_j^T\right) \\
&= |c\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\Big\{-\mathrm{tr}\Big(\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\mathbf{Y}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}\Big) + \mathrm{tr}\Big(\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\mathbf{Y}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\boldsymbol{\beta}_j\Big) \\
&\qquad +\mathrm{tr}\Big(\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\boldsymbol{\beta}_j^T\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}\Big) - \mathrm{tr}\Big(\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\boldsymbol{\beta}_j^T\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\boldsymbol{\beta}_j\Big)\Big\} \\
&\quad \times \exp\left(-\tfrac{1}{2}\boldsymbol{\beta}_j(c\boldsymbol{\Sigma})^{-1}\boldsymbol{\beta}_j^T\right) \\
&= |c\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\Big\{-\mathrm{tr}\Big(\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\mathbf{Y}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}\Big) + \mathrm{tr}\Big(\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\mathbf{Y}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\boldsymbol{\beta}_j\Big) \\
&\qquad +\mathrm{tr}\Big(\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\mathbf{Y}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\boldsymbol{\beta}_j\Big) - \mathrm{tr}\Big(\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\boldsymbol{\beta}_j^T\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\boldsymbol{\beta}_j\Big)\Big\} \\
&\quad \times \exp\left(-\tfrac{1}{2}\boldsymbol{\beta}_j(c\boldsymbol{\Sigma})^{-1}\boldsymbol{\beta}_j^T\right) \\
&= |c\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\Big\{-\mathrm{tr}\Big(\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\mathbf{Y}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}\Big) + \mathrm{tr}\Big(\boldsymbol{\Sigma}^{-1}\boldsymbol{\beta}_j^T\textstyle\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}\Big) \\
&\qquad -\mathrm{tr}\Big(\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\boldsymbol{\beta}_j^T\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\boldsymbol{\beta}_j\Big)\Big\} \\
&\quad \times \exp\left(-\tfrac{1}{2}\boldsymbol{\beta}_j(c\boldsymbol{\Sigma})^{-1}\boldsymbol{\beta}_j^T\right) \\
&= |c\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\Big[-\mathrm{tr}\Big(\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\mathbf{Y}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}\Big) \\
&\qquad -\mathrm{tr}\Big(\tfrac{1}{2}\boldsymbol{\beta}_j\boldsymbol{\Sigma}^{-1}\Big(\textstyle\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\Big)\boldsymbol{\beta}_j^T\Big) \\
&\qquad +\mathrm{tr}\Big\{\Big(\boldsymbol{\Sigma}^{-1}\textstyle\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\Big)\boldsymbol{\beta}_j^T\Big(\frac{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}}{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}}\Big)\Big\} \\
&\qquad +\mathrm{tr}\Big\{\boldsymbol{\Sigma}^{-1}\Big(\frac{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}}{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}}\Big)^T\Big(\textstyle\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\Big)\Big(\frac{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}}{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}}\Big)\Big\} \\
&\qquad -\mathrm{tr}\Big\{\boldsymbol{\Sigma}^{-1}\Big(\frac{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}}{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}}\Big)^T\Big(\textstyle\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\Big)\Big(\frac{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}}{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}}\Big)\Big\}\Big] \\
&\quad \times \exp\left(-\tfrac{1}{2}\boldsymbol{\beta}_j(c\boldsymbol{\Sigma})^{-1}\boldsymbol{\beta}_j^T\right) \\
&= |c\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\Big[-\mathrm{tr}\Big(\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\mathbf{Y}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}\Big) \\
&\quad -\mathrm{tr}\Big(\tfrac{1}{2}\boldsymbol{\beta}_j\boldsymbol{\Sigma}^{-1}\Big(\textstyle\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\Big)\boldsymbol{\beta}_j^T\Big) \\
&\quad +\mathrm{tr}\Big\{\boldsymbol{\Sigma}^{-1}\Big(\textstyle\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\Big)\boldsymbol{\beta}_j^T\Big(\frac{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}}{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}}\Big)\Big\} \\
&\quad +\mathrm{tr}\Big\{\boldsymbol{\Sigma}^{-1}\textstyle\sum_{k=1}^{n_w}\mathbf{Y}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\Big(\frac{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}}{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}}\Big)\Big\} \\
&\quad -\mathrm{tr}\Big\{\frac{\boldsymbol{\Sigma}^{-1}}{2}\Big(\frac{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}}{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}}\Big)^T\Big(\textstyle\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\Big)\Big(\frac{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}}{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}}\Big)\Big\}
\end{aligned}$$

$$
\begin{aligned}
&\quad -\mathrm{tr}\Bigg\{\tfrac{1}{2}\boldsymbol{\Sigma}^{-1}\Big(\textstyle\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\Big)\Bigg(\dfrac{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}}{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}}\Bigg)^T\Bigg(\dfrac{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}}{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}}\Bigg)\Bigg\}\Bigg]\\
&\quad \times\exp\big(-\tfrac{1}{2}\boldsymbol{\beta}_j(c\boldsymbol{\Sigma})^{-1}\boldsymbol{\beta}_j^T\big)\\
&= |c\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\Big[-\mathrm{tr}\Big(\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\mathbf{Y}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}\Big)+\mathrm{tr}\Big(\boldsymbol{\Sigma}^{-1}\textstyle\sum_{k=1}^{n_w}\mathbf{Y}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\mathbf{b}_j\Big)\\
&\qquad\qquad\qquad -\mathrm{tr}\Big(\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\mathbf{b}_j^T\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\mathbf{b}_j\Big)\Big]\\
&\quad \times\exp\Big(\mathrm{tr}\big(\boldsymbol{\omega}^{-1}\boldsymbol{\beta}_j^T\mathbf{b}_j\big)-\mathrm{tr}\Big(\tfrac{1}{2}\boldsymbol{\omega}^{-1}\mathbf{b}_j^T\mathbf{b}_j\Big)-\mathrm{tr}\big(\tfrac{1}{2}\boldsymbol{\beta}_j\boldsymbol{\omega}^{-1}\boldsymbol{\beta}_j^T\big)-\tfrac{1}{2}\boldsymbol{\beta}_j(c\boldsymbol{\Sigma})^{-1}\boldsymbol{\beta}_j^T\Big)\\
&= |c\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\Big\{-\mathrm{tr}\Big(\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\mathbf{Y}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}\Big)+\mathrm{tr}\Big(\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\mathbf{Y}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\mathbf{b}_j\Big)\\
&\qquad\qquad\qquad +\mathrm{tr}\Big(\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}(\mathbf{X}_{kj}\mathbf{b}_j)^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}\Big)\\
&\qquad\qquad\qquad -\mathrm{tr}\Big(\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}(\mathbf{X}_{kj}\mathbf{b}_j)^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\mathbf{b}_j\Big)\Big\}\\
&\quad \times\exp\Big\{\mathrm{tr}\big(\mathbf{b}_j\boldsymbol{\omega}^{-1}\boldsymbol{\beta}_j^T\big)-\mathrm{tr}\Big(\tfrac{1}{2}\mathbf{b}_j\boldsymbol{\omega}^{-1}\mathbf{b}_j^T\Big)-\mathrm{tr}\big(\tfrac{1}{2}\boldsymbol{\beta}_j\boldsymbol{\omega}^{-1}\boldsymbol{\beta}_j^T\big)-\tfrac{1}{2}\boldsymbol{\beta}_j(c\boldsymbol{\Sigma})^{-1}\boldsymbol{\beta}_j^T\Big\}\\
&= |c\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\Big[\mathrm{tr}\Big\{-\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\Big(\mathbf{Y}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}-\mathbf{Y}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\mathbf{b}_j\\
&\qquad\qquad\qquad\quad -(\mathbf{X}_{kj}\mathbf{b}_j)^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}+(\mathbf{X}_{kj}\mathbf{b}_j)^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\mathbf{b}_j)\}]\\
&\quad \times\exp\Big(\boldsymbol{\beta}_j\boldsymbol{\omega}^{-1}\mathbf{b}_j^T-\tfrac{1}{2}\mathbf{b}_j\boldsymbol{\omega}^{-1}\mathbf{b}_j^T-\tfrac{1}{2}\boldsymbol{\beta}_j\boldsymbol{\omega}^{-1}\boldsymbol{\beta}_j^T-\tfrac{1}{2}\boldsymbol{\beta}_j(c\boldsymbol{\Sigma})^{-1}\boldsymbol{\beta}_j^T\Big)\\
&= |c\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\Big[\mathrm{tr}\Big\{-\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\left(\mathbf{Y}_{kj}-\mathbf{X}_{kj}\mathbf{b}_j\right)^T\mathbf{V}(\phi)_k^{-1}\left(\mathbf{Y}_{kj}-\mathbf{X}_{kj}\mathbf{b}_j\right)\Big\}\Big]\\
&\quad \times\exp\Big(\boldsymbol{\beta}_j\boldsymbol{\omega}^{-1}\mathbf{b}_j^T-\tfrac{1}{2}\boldsymbol{\beta}_j\boldsymbol{\omega}^{-1}\boldsymbol{\beta}_j^T-\tfrac{1}{2}\boldsymbol{\beta}_j(c\boldsymbol{\Sigma})^{-1}\boldsymbol{\beta}_j^T\Big)\exp\Big(-\tfrac{1}{2}\mathbf{b}_j\boldsymbol{\omega}^{-1}\mathbf{b}_j^T\Big)\\
&= |c\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\Big[\mathrm{tr}\Big\{-\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\left(\mathbf{Y}_{kj}-\mathbf{X}_{kj}\mathbf{b}_j\right)^T\mathbf{V}(\phi)_k^{-1}\left(\mathbf{Y}_{kj}-\mathbf{X}_{kj}\mathbf{b}_j\right)\Big\}\Big]\\
&\quad \times\exp\Big(\boldsymbol{\beta}_j\boldsymbol{\omega}^{-1}\mathbf{b}_j^T-\tfrac{1}{2}\boldsymbol{\beta}_j\left(\boldsymbol{\omega}^{-1}+(c\boldsymbol{\Sigma})^{-1}\right)\boldsymbol{\beta}_j^T\Big)\exp\Big(-\tfrac{1}{2}\mathbf{b}_j\boldsymbol{\omega}^{-1}\mathbf{b}_j^T\Big)\\
&= |c\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\Big[\mathrm{tr}\Big\{-\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\left(\mathbf{Y}_{kj}-\mathbf{X}_{kj}\mathbf{b}_j\right)^T\mathbf{V}(\phi)_k^{-1}\left(\mathbf{Y}_{kj}-\mathbf{X}_{kj}\mathbf{b}_j\right)\Big\}\Big]\exp\Big(-\tfrac{1}{2}\mathbf{b}_j\boldsymbol{\omega}^{-1}\mathbf{b}_j^T\Big)\\
&\quad \times\exp\Big[-\tfrac{1}{2}\boldsymbol{\beta}_j\left(\boldsymbol{\omega}^{-1}+(c\boldsymbol{\Sigma})^{-1}\right)\boldsymbol{\beta}_j^T+\boldsymbol{\beta}_j\left(\boldsymbol{\omega}^{-1}+(c\boldsymbol{\Sigma})^{-1}\right)\left(\boldsymbol{\omega}^{-1}+(c\boldsymbol{\Sigma})^{-1}\right)^{-1}\boldsymbol{\omega}^{-1}\mathbf{b}_j^T\\
&\quad -\tfrac{1}{2}\Big\{\mathbf{b}_j\boldsymbol{\omega}^{-1}\left(\boldsymbol{\omega}^{-1}+(c\boldsymbol{\Sigma})^{-1}\right)^{-1}\Big\}\left(\boldsymbol{\omega}^{-1}+(c\boldsymbol{\Sigma})^{-1}\right)\Big\{\mathbf{b}_j\boldsymbol{\omega}^{-1}\left(\boldsymbol{\omega}^{-1}+(c\boldsymbol{\Sigma})^{-1}\right)^{-1}\Big\}^T\\
&\quad +\tfrac{1}{2}\Big\{\mathbf{b}_j\boldsymbol{\omega}^{-1}\left(\boldsymbol{\omega}^{-1}+(c\boldsymbol{\Sigma})^{-1}\right)^{-1}\Big\}\left(\boldsymbol{\omega}^{-1}+(c\boldsymbol{\Sigma})^{-1}\right)\Big\{\mathbf{b}_j\boldsymbol{\omega}^{-1}\left(\boldsymbol{\omega}^{-1}+(c\boldsymbol{\Sigma})^{-1}\right)^{-1}\boldsymbol{\omega}^{-1}\mathbf{b}_j^T\Big\}^T\Big]\\
&= |c\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\Big[\mathrm{tr}\Big\{-\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\left(\mathbf{Y}_{kj}-\mathbf{X}_{kj}\mathbf{b}_j\right)^T\mathbf{V}(\phi)_k^{-1}\left(\mathbf{Y}_{kj}-\mathbf{X}_{kj}\mathbf{b}_j\right)\Big\}\Big]\\
&\quad \times\exp\Big\{\tfrac{1}{2}\left(\mathbf{b}_j\boldsymbol{\omega}^{-1}\boldsymbol{\Sigma}_*\right)\boldsymbol{\Sigma}_*^{-1}\left(\mathbf{b}_j\boldsymbol{\omega}^{-1}\boldsymbol{\Sigma}_*\right)^T-\tfrac{1}{2}\mathbf{b}_j\boldsymbol{\omega}^{-1}\mathbf{b}_j^T\Big\}\\
&\quad \times\exp\Big\{-\tfrac{1}{2}\boldsymbol{\beta}_j\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\beta}_j^T+\boldsymbol{\beta}_j\boldsymbol{\Sigma}_*^{-1}\left(\mathbf{b}_j\boldsymbol{\omega}^{-1}\boldsymbol{\Sigma}_*\right)^T-\tfrac{1}{2}\left(\mathbf{b}_j\boldsymbol{\omega}^{-1}\boldsymbol{\Sigma}_*\right)\boldsymbol{\Sigma}_*^{-1}\left(\mathbf{b}_j\boldsymbol{\omega}^{-1}\boldsymbol{\Sigma}_*\right)^T\Big\}\\
&= |c\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\Big[\mathrm{tr}\Big\{-\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\left(\mathbf{Y}_{kj}-\mathbf{X}_{kj}\mathbf{b}_j\right)^T\mathbf{V}(\phi)_k^{-1}\left(\mathbf{Y}_{kj}-\mathbf{X}_{kj}\mathbf{b}_j\right)\Big\}\Big]
\end{aligned}
$$

$$\times \exp\left(\tfrac{1}{2}\bar{\boldsymbol{\beta}}_j \boldsymbol{\Sigma}_*^{-1}\bar{\boldsymbol{\beta}}_j^T - \tfrac{1}{2}\mathbf{b}_j\boldsymbol{\omega}^{-1}\mathbf{b}_j^T - \tfrac{1}{2}\boldsymbol{\beta}_j\boldsymbol{\Sigma}_*^{-1}\boldsymbol{\beta}_j^T + \boldsymbol{\beta}_j\boldsymbol{\Sigma}_*^{-1}\bar{\boldsymbol{\beta}}_j^T - \tfrac{1}{2}\bar{\boldsymbol{\beta}}_j\boldsymbol{\Sigma}_*^{-1}\bar{\boldsymbol{\beta}}_j^T\right)$$

$$= |c\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left[\operatorname{tr}\left\{-\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\left(\mathbf{Y}_{kj} - \mathbf{X}_{kj}\mathbf{b}_j\right)^T \mathbf{V}(\phi)_k^{-1}\left(\mathbf{Y}_{kj} - \mathbf{X}_{kj}\mathbf{b}_j\right)\right\}\right]$$

$$\times \exp\left\{-\frac{1}{2}\left(\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}}_j\right)\boldsymbol{\Sigma}_*^{-1}\left(\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}}_j\right)^T\right\}\exp\left(\frac{1}{2}\bar{\boldsymbol{\beta}}_j\boldsymbol{\Sigma}_*^{-1}\bar{\boldsymbol{\beta}}_j^T - \frac{1}{2}\mathbf{b}_j\boldsymbol{\omega}^{-1}\mathbf{b}_j^T\right) \qquad (A.29)$$

$$\propto |\boldsymbol{\Sigma}_*|^{-\frac{1}{2}} \exp\left\{-\tfrac{1}{2}\left(\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}}_j\right)\boldsymbol{\Sigma}_*^{-1}\left(\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}}_j\right)^T\right\},$$

which is proportional to the density for a vector which is normally distributed with mean $\bar{\boldsymbol{\beta}}_j^T$ and variance $\boldsymbol{\Sigma}_*$. Therefore, $\boldsymbol{\beta}_j^T \sim \mathrm{N}(\bar{\boldsymbol{\beta}}_j^T, \boldsymbol{\Sigma}_*)$, conditional on $\boldsymbol{\beta}_j^T \neq \mathbf{0}_r$.

To jointly sample δ_j and $\boldsymbol{\beta}_j$, we need to calculate the conditional posterior probability that $\boldsymbol{\beta}_j^T = \mathbf{0}_r$, which is

$$\rho_j = \frac{1 - \rho_a}{(1 - \rho_a) - \rho_a BF}, \qquad (A.30)$$

where ρ_a is the prior probability that $\boldsymbol{\beta}_j^T \neq \mathbf{0}_r$ and BF is the Bayes Factor, which is the ratio of the conditional distribution for $\boldsymbol{\beta}_j^T = \mathbf{0}_r$ and $\boldsymbol{\beta}_j^T \neq \mathbf{0}_r$. To calculate BF, we first need to integrate out $\boldsymbol{\beta}_j$ from (A.29). Hence we calculate

$$\textstyle\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} |c\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left[\operatorname{tr}\left\{-\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\sum_{k=1}^{n_w}\left(\mathbf{Y}_{kj} - \mathbf{X}_{kj}\mathbf{b}_j\right)^T \mathbf{V}(\phi)_k^{-1}\left(\mathbf{Y}_{kj} - \mathbf{X}_{kj}\mathbf{b}_j\right)\right\}\right]$$

$$\times \exp\left(\tfrac{1}{2}\bar{\boldsymbol{\beta}}_j\boldsymbol{\Sigma}_*^{-1}\bar{\boldsymbol{\beta}}_j^T - \tfrac{1}{2}\mathbf{b}_j\boldsymbol{\omega}^{-1}\mathbf{b}_j^T\right)\exp\left\{-\tfrac{1}{2}\left(\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}}_j\right)\boldsymbol{\Sigma}_*^{-1}\left(\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}}_j\right)^T\right\} d\boldsymbol{\beta}_j$$

$$= |c\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\left(\tfrac{1}{2}\bar{\boldsymbol{\beta}}_j\boldsymbol{\Sigma}_*^{-1}\bar{\boldsymbol{\beta}}_j^T - \tfrac{1}{2}\mathbf{b}_j\boldsymbol{\omega}^{-1}\mathbf{b}_j^T\right)$$

$$\times \exp\left[\operatorname{tr}\left\{-\tfrac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\left(\mathbf{Y}_{kj} - \mathbf{X}_{kj}\mathbf{b}_j\right)^T \mathbf{V}(\phi)_k^{-1}\left(\mathbf{Y}_{kj} - \mathbf{X}_{kj}\mathbf{b}_j\right)\right\}\right]$$

$$\times |\boldsymbol{\Sigma}_*|^{\frac{1}{2}} \textstyle\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} |\boldsymbol{\Sigma}_*| \exp\left\{-\tfrac{1}{2}\left(\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}}_j\right)\boldsymbol{\Sigma}_*^{-1}\left(\boldsymbol{\beta}_j - \bar{\boldsymbol{\beta}}_j\right)^T\right\} d\boldsymbol{\beta}_j,$$

and the conditional density for $\boldsymbol{\beta}_j^T \neq \mathbf{0}_r$ is therefore

$$\left(\frac{|\boldsymbol{\Sigma}_*|}{|c\boldsymbol{\Sigma}|}\right)^{\frac{1}{2}} \exp\left[\operatorname{tr}\left\{-\frac{\boldsymbol{\Sigma}^{-1}}{2}\sum_{k=1}^{n_w}\left(\mathbf{Y}_{kj} - \mathbf{X}_{kj}\mathbf{b}_j\right)^T \mathbf{V}(\phi)_k^{-1}\left(\mathbf{Y}_{kj} - \mathbf{X}_{kj}\mathbf{b}_j\right)\right\}\right]$$

$$\times \exp\left(\frac{1}{2}\bar{\boldsymbol{\beta}}_j\boldsymbol{\Sigma}_*^{-1}\bar{\boldsymbol{\beta}}_j^T - \frac{1}{2}\mathbf{b}_j\boldsymbol{\omega}^{-1}\mathbf{b}_j^T\right). \qquad (A.31)$$

Using (A.28) and (A.31), the BF is

$$\mathrm{BF} = \frac{\left(\frac{|\boldsymbol{\Sigma}_*|}{|c\boldsymbol{\Sigma}|}\right)^{\frac{1}{2}} \exp\left[\operatorname{tr}\left\{-\frac{\boldsymbol{\Sigma}^{-1}}{2}\sum_{k=1}^{n_w}\left(\mathbf{Y}_{kj} - \mathbf{X}_{kj}\mathbf{b}_j\right)^T \mathbf{V}(\phi)_k^{-1}\left(\mathbf{Y}_{kj} - \mathbf{X}_{kj}\mathbf{b}_j\right)\right\}\right] \exp\left(\frac{1}{2}\bar{\boldsymbol{\beta}}_j\boldsymbol{\Sigma}_*^{-1}\bar{\boldsymbol{\beta}}_j^T - \frac{1}{2}\mathbf{b}_j\boldsymbol{\omega}^{-1}\mathbf{b}_j^T\right)}{\exp\left\{-\operatorname{tr}\left(\frac{\boldsymbol{\Sigma}^{-1}}{2}\sum_{k=1}^{n_w}\mathbf{Y}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}\right)\right\}}$$

$$
\begin{aligned}
&= \left(\frac{|\boldsymbol{\Sigma}_*|}{|c\boldsymbol{\Sigma}|}\right)^{\frac{1}{2}} \exp\left\{-\mathrm{tr}\left(\frac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\mathbf{Y}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}\right) + \mathrm{tr}\left(\frac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\mathbf{Y}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\mathbf{b}_j\right)\right.\\
&\qquad +\mathrm{tr}\left(\frac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\mathbf{b}_j^T\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}\right) - \mathrm{tr}\left(\frac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\mathbf{b}_j^T\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\mathbf{b}_j\right)\\
&\qquad \left.-\tfrac{1}{2}\mathbf{b}_j\boldsymbol{\omega}^{-1}\mathbf{b}_j^T + \mathrm{tr}\left(\frac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\mathbf{Y}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}\right)\right\} \exp\left(\tfrac{1}{2}\bar{\boldsymbol{\beta}}_j\boldsymbol{\Sigma}_*^{-1}\bar{\boldsymbol{\beta}}_j^T\right)\\
&= \left(\frac{|\boldsymbol{\Sigma}_*|}{|c\boldsymbol{\Sigma}|}\right)^{\frac{1}{2}} \exp\left[\mathrm{tr}\left(\frac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\mathbf{Y}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\mathbf{b}_j\right) + \mathrm{tr}\left(\textstyle\sum_{k=1}^{n_w}\mathbf{b}_j^T\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}\frac{\boldsymbol{\Sigma}^{-1}}{2}\right)\right.\\
&\qquad \left.-\mathrm{tr}\left\{\frac{\boldsymbol{\Sigma}^{-1}}{2}\mathbf{b}_j^T\left(\textstyle\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\right)\mathbf{b}_j^T\right\} - \tfrac{1}{2}\mathbf{b}_j\boldsymbol{\omega}^{-1}\mathbf{b}_j\right] \exp\left(\tfrac{1}{2}\bar{\boldsymbol{\beta}}_j\boldsymbol{\Sigma}_*^{-1}\bar{\boldsymbol{\beta}}_j^T\right)\\
&= \left(\frac{|\boldsymbol{\Sigma}_*|}{|c\boldsymbol{\Sigma}|}\right)^{\frac{1}{2}} \exp\left[\mathrm{tr}\left(\frac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\mathbf{Y}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\mathbf{b}_j\right) + \mathrm{tr}\left(\frac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\mathbf{Y}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\mathbf{b}_j\right)\right.\\
&\qquad \left.-\mathrm{tr}\left\{\mathbf{b}_j\left(\frac{\boldsymbol{\Sigma}^{-1}}{2}\textstyle\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\right)\mathbf{b}_j^T\right\} - \tfrac{1}{2}\mathbf{b}_j\boldsymbol{\omega}^{-1}\mathbf{b}_j^T\right] \exp\left(\tfrac{1}{2}\bar{\boldsymbol{\beta}}_j\boldsymbol{\Sigma}_*^{-1}\bar{\boldsymbol{\beta}}_j^T\right)\\
&= \left(\frac{|\boldsymbol{\Sigma}_*|}{|c\boldsymbol{\Sigma}|}\right)^{\frac{1}{2}} \exp\left[\tfrac{1}{2}\mathrm{tr}\left\{\left(\boldsymbol{\Sigma}^{-1}\textstyle\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\right)\left(\frac{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}}{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}}\right)^T\mathbf{b}_j\right\}\right.\\
&\qquad +\tfrac{1}{2}\mathrm{tr}\left\{\left(\boldsymbol{\Sigma}^{-1}\textstyle\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\right)\left(\frac{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}}{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}}\right)^T\mathbf{b}_j\right\}\\
&\qquad \left.-\tfrac{1}{2}\mathrm{tr}\left\{\mathbf{b}_j\left(\boldsymbol{\Sigma}^{-1}\textstyle\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\right)\mathbf{b}_j^T\right\} - \tfrac{1}{2}\mathbf{b}_j\boldsymbol{\omega}^{-1}\mathbf{b}_j^T\right] \exp\left(\tfrac{1}{2}\bar{\boldsymbol{\beta}}_j\boldsymbol{\Sigma}_*^{-1}\bar{\boldsymbol{\beta}}_j^T\right)\\
&= \left(\frac{|\boldsymbol{\Sigma}_*|}{|c\boldsymbol{\Sigma}|}\right)^{\frac{1}{2}} \exp\left[\tfrac{1}{2}\mathrm{tr}\left\{\mathbf{b}_j\left(\boldsymbol{\Sigma}^{-1}\textstyle\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\right)\mathbf{b}_j^T\right\}\right.\\
&\qquad +\tfrac{1}{2}\mathrm{tr}\left\{\mathbf{b}_j\left(\boldsymbol{\Sigma}^{-1}\textstyle\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\right)\mathbf{b}_j^T\right\}\\
&\qquad \left.-\tfrac{1}{2}\mathrm{tr}\left\{\mathbf{b}_j\left(\boldsymbol{\Sigma}^{-1}\textstyle\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}\right)\mathbf{b}_j^T\right\} - \tfrac{1}{2}\mathbf{b}_j\boldsymbol{\omega}^{-1}\mathbf{b}_j^T\right] \exp\left(\tfrac{1}{2}\bar{\boldsymbol{\beta}}_j\boldsymbol{\Sigma}_*^{-1}\bar{\boldsymbol{\beta}}_j^T\right)\\
&= \left(\frac{|\boldsymbol{\Sigma}_*|}{|c\boldsymbol{\Sigma}|}\right)^{\frac{1}{2}} \exp\left[\mathbf{b}_j\boldsymbol{\omega}^{-1}\mathbf{b}_j^T - \tfrac{1}{2}\mathbf{b}_j\boldsymbol{\omega}^{-1}\mathbf{b}_j^T - \tfrac{1}{2}\mathbf{b}_j\boldsymbol{\omega}^{-1}\mathbf{b}_j^T\right] \exp\left(\tfrac{1}{2}\bar{\boldsymbol{\beta}}_j\boldsymbol{\Sigma}_*^{-1}\bar{\boldsymbol{\beta}}_j^T\right)\\
&= \left(\frac{|\boldsymbol{\Sigma}_*|}{|c\boldsymbol{\Sigma}|}\right)^{\frac{1}{2}} \exp\left(\tfrac{1}{2}\bar{\boldsymbol{\beta}}_j\boldsymbol{\Sigma}_*^{-1}\bar{\boldsymbol{\beta}}_j^T\right).
\end{aligned}
$$

To jointly sample δ_j and $\boldsymbol{\beta}_j$, we

1. Sample u from $U(0,1)$.

2. If $\rho_j > u$ then $\delta_j = 0$ and $\boldsymbol{\beta}_j^T = \mathbf{0}_r$. Otherwise, $\delta_j = 1$ and $\boldsymbol{\beta}_j^T \sim \mathrm{N}\ (\bar{\boldsymbol{\beta}}_j^T, \boldsymbol{\Sigma}_*)$, where $\bar{\boldsymbol{\beta}}_j = \mathbf{b}_j\boldsymbol{\omega}^{-1}\boldsymbol{\Sigma}_*$ and $\boldsymbol{\Sigma}_* = ((\boldsymbol{\omega}^{-1}) + (c\boldsymbol{\Sigma})^{-1})^{-1}$ when

$$\mathbf{b}_j = \frac{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{Y}_{kj}}{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}}$$

and

$$\boldsymbol{\omega} = \frac{\boldsymbol{\Sigma}}{\sum_{k=1}^{n_w}\mathbf{X}_{kj}^T\mathbf{V}(\phi)_k^{-1}\mathbf{X}_{kj}}.$$

Bibliography

Aarts, E. H. L. and van Laarhoven, P. J. M. (1989) Simulated annealing: An introduction. *Statistica Neerlandica*, **43**, 31–52.

Akaike, H. (1973) Information theory and an extension of the maximum likelihood principle. In *Proceedings of the Second International Symposium on Information Theory* (eds. B. N. Petrov and F. Csaki), 267–281. Akademiai Kiado, Budapest.

Allen, D. M. (1974) The relationship between variable selection and data augmentation and a method for prediction. *Technometrics*, **16**, 125–127.

Andersen, A. H., Jensen, E. B. and Geert, S. (1981) Two-way analysis of variance with correlated errors. *International Statistical Review*, **49**, 153–167.

Arnouts, H., Goos, P. and Jones, B. (2010) Design and analysis of industrial strip-plot experiments. *Quality and Reliability Engineering International*, **26**, 127–136.

Atkinson, A. C. and Bogacka, B. (1997) Compound D- and D_s-optimum designs for determining the order of a chemical reaction. *Technometrics*, **39**, 347–356.

Atkinson, A. C. and Cox, D. R. (1974) Planning experiments for discriminating between models (with discussion). *Journal of the Royal Statistical Society, Series B*, **36**, 321–348.

Atkinson, A. C. and Donev, A. N. (1989) The construction of exact D-optimum experimental designs with application to blocking response surface designs. *Biometrika*, **76**, 515–526.

— (1996) Experimental designs optimally balanced for trend. *Technometrics*, **38**, 333–341.

Atkinson, A. C., Donev, A. N. and Tobias, R. D. (2007) *Optimum Experimental Designs, with SAS.* Oxford University Press, Oxford, second edn.

Azzalini, A. (1984) Estimation and hypothesis testing for collections of autoregressive time series. *Biometrika*, **71**, 85–90.

Bagchi, B. and Bagchi, S. (2001) Optimality of partial geometric designs. *Annals of Statistics*, **29**, 577–594.

Bailey, R. A. (1991) Strata for randomized experiments. *Journal of the Royal Statistical Society, Series B*, **53**, 27–78.

— (2008) *Design of Comparative Experiments.* Cambridge University Press, Cambridge.

Bapat, R. B. and Dey, A. (1991) Optimal block designs with minimal number of observations. *Statistics and Probability Letters*, **11**, 399–402.

Berenblut, I. I. and Webb, G. I. (1974) Experimental design in the presence of autocorrelated errors. *Biometrika*, **61**, 427–437.

Berger, J. O. (1985) *Statistical Decision Theory and Bayesian Analysis.* Springer-Verlag, New York, second edn.

Bingham, D., Sitter, R., Kelly, E., Moore, L. and Olivas, J. D. (2008) Factorial designs with multiple levels of randomization. *Statistica Sinica*, **18**, 493–513.

Bischoff, W. (1992) On exact D-optimal designs for regression models with correlated observations. *Annals of the Institute of Statistical Mathematics*, **44**, 229–238.

Booth, K. H. V. and Cox, D. R. (1962) Some systematic supersaturated designs. *Technometrics*, **4**, 489–495.

Box, G. E. P. and Hunter, J. S. (1961a) The 2^{k-p} fractional factorial designs, part I. *Technometrics*, **3**, 311–351.

— (1961b) The 2^{k-p} fractional factorial designs, part II. *Technometrics*, **3**, 449–458.

Box, G. E. P., Jenkins, G. M. and Reinsel, G. C. (2008) *Time Series Analysis: Forecasting and Control.* John Wiley and Sons, New York.

Box, G. E. P. and Jones, S. (1992) Split-plot designs for robust product experimentation. *Journal of Applied Statistics*, **19**, 3–26.

Box, G. E. P. and Meyer, R. D. (1986) An analysis for unreplicated fractional factorials. *Technometrics*, **28**, 11–18.

Brien, C. J. (1983) Analysis of variance tables based on experimental strutcture. *Biometrics*, **39**, 53–59.

Brien, C. J. and Bailey, R. A. (2006) Multiple randomizations. *Journal of the Royal Statistical Society, Series B*, **68**, 571–609.

Brien, C. J., Harch, B. D., Correll, R. L. and Bailey, R. A. (2011) Multiphase experiments with at least one later laboratory phase. I. orthogonal designs. *Journal of Agricultural, Biological and Environmental Statistics*, **16**, 422–450.

Brien, C. J. and Payne, R. W. (1999) Tiers, structure formulae and the analysis of complicated experiments. *Journal of the Royal Statistical Society, Series D*, **48**, 41–52.

Brooks, S. P. and Morgan, B. J. T. (1995) Optimization using simulated annealing. *Journal of the Royal Statistical Society, Series D*, **44**, 241–257.

Brown, P., Vannucci, M. and Fearn, T. (1998) Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society, Series B*, **68**, 627–641.

Byrd, R. H., Lu, P., Nocedal, J. and Zhu, C. (1995) A limited memory algorithm for bound constrained optimisation. *SIAM Journal on Scientific Computing*, **16**, 1190–1208.

Chakrabarti, M. C. (1963) On the C-matrix in design of experiments. *Journal of the Indian Statistical Association*, **1**, 8–23.

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

Chan, B. S. P. and Eccleston, J. A. (1998) On the construction of complete and partial nearest neighbour balanced designs. *Australasian Journal of Combinatorics*, **18**, 39–50.

— (2003) On the construction of nearest-neighbour balanced row-column designs. *Australian and New Zealand Journal of Statistics*, **45**, 97–106.

Cheng, C. S. (1985) Run orders of factorial designs. *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer*, **1**, 619–633.

Cheng, C.-S. (2006) Projection properties of factorial designs for factor screening. In *Screening; Methods for Experimentation in Industry, Drug Discovery and Genetics* (eds. A. Dean and S. Lewis), 156–168. Springer, New York.

Cheng, C.-S. and Steinberg, D. M. (1991) Trend robust two-level factorial designs. *Biometrika*, **78**, 325–336.

Chipman, H., George, E. I. and McCulloch, R. E. (2001) The practical implementation of Bayesian model selection. *IMS Lecture Notes - Monograph Series*, **38**, 65–134.

Chow, S.-C. (2007) *Statistical Designs and Analysis of Stability Studies.* Chapman and Hall /CRC, New York.

Chung, F. R. K. (1997) *CBMS Regional Conference Series in Mathematics, No. 92, Spectral Graph Theory.* American Mathematical Society, Rhode Island.

Constantine, G. M. (1989) Robust designs for serially correlated observations. *Biometrika*, **76**, 245–251.

Coster, D. C. and Cheng, C. S. (1988) Minimum cost trend-free run orders of fractional factorial designs. *Annals of Statistics*, **16**, 1188–1205.

Cox, D. R. (1958) *Planning of Experiments.* John Wiley and Sons, New York.

Davidian, M. and Giltinan, D. M. (1995) *Monographs on Statistics and Applied Probability 62, Nonlinear Models for Repeated Measurement Data.* Chapman and Hall, London.

Edwards, D. J. and Fuerte, J. N. (2011) Compromise ascent directions for multiple-response applications. *Quality and Reliability Engineering International*, **27**, 1107–1118.

Elliot, L. J., Eccleston, J. A. and Martin, R. J. (1999) An algorithm for the design of factorial experiments when the data are correlated. *Statistics and Computing*, **9**, 195–201.

Fang, K.-T., Li, R. and Sudjianto, A. (2006) *Design and Modelling for Computer Experiments.* Chapman and Hall /CRC, New York.

Fisher, R. A. and Yates, F. (1963) *Statistical Tables for Biological, Agricultural and Medical Research.* Oliver and Boyd, Edinburgh, 6th edn.

Freeman, G. H. (1959) The use of the same experimental material for more than one set of treatments. *Journal of the Royal Statistical Society, Series C*, **8**, 13–20.

— (1979) Some two-dimensional designs balanced for nearest neighbours. *Journal of the Royal Statistical Society, Series B*, **41**, 88–95.

Fuller, W. A. (1996) *Introduction to Statistical Time Series.* John Wiley and Sons, New York, second edn.

Fuller, W. A. and Battese, G. E. (1973) Transformations for estimation of linear models with nested-error structure. *Journal of the American Statistical Association*, **68**, 626–632.

Garroi, J.-J., Goos, P. and Sörensen, K. (2009) A variable-neighbourhood search algorithm for finding optimal run orders in the presence of serial correlation. *Journal of Statistical Planning and Inference*, **139**, 30–44.

Gelfand, A. E. and Smith, A. F. M. (1990) Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Gelman, A., Carlin, B., Stern, H. and Rubin, D. (2004) *Bayeian Data Analysis.* Chapman and Hall /CRC, New York, second edn.

Geman, S. and Geman, D. (1984) Stochastic relaxation, Gibbs distributions and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

George, E. I. and McCulloch, R. E. (1993) Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, **88**, 881–889.

— (1997) Approaches for Bayesian variable selection. *Statistica Sinica*, **7**, 339–374.

Geweke, J. (1996) Variable selection and model comparison in regression. In *Bayesian Statistics 5* (eds. J. Bernardo, J. Berger, A. Dawid and A. Smith). Clarendon Press, Oxford.

Gill, P. S. (1990) A Monte Carlo simulation study of analyses of block designs under correlated errors model. *Communications in Statistics - Simulation and Computation*, **19**, 175–188.

Gill, P. S. and Shuka, G. K. (1985) Efficiency of nearest neighbour balanced block designs for correlated observations. *Biometrika*, **72**, 539–544.

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

Gilmour, S. G. and Trinca, L. A. (2012) Optimum design of experiments for statistical inference. *Applied Statistics*, **61**, 1–25.

Glover, F. (1989) Tabu search - part I. *ORSA Journal on Computing*, **1**, 190–206.

Goos, P. (2002) *The Optimal Design of Blocked and Split-Plot Experiments.* Springer-Verlag, New York.

Goos, P. and Gilmour, S. G. (2012) A general strategy for analyzing data from split-plot and multistratum experimental designs. *Technometrics*, **54**, 340–354.

Goos, P. and Jones, B. (2011) *Optimal Design of Experiments: A Case Study Approach.* John Wiley and Sons, Chichester.

Goos, P. and Vandebroek, M. (2001) D-optimal response surface designs in the presence of random block effects. *Computational Statistics and Data Analysis*, **37**, 433–453.

Green, P. J. (1995) Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. *Biometrika*, **82**, 711–732.

Gupta, A. K. and Nagar, D. K. (2000) *Matrix Variate Distributions.* No. 104 in Monographs and Surveys in Pure and Applied Mathematics. Chapman and Hall /CRC, New York.

Hall, M. J. (1961) Hadamard matrices of order 16. *Research Summary No 36-10 Volume 1*, Jet Propulsion Laboratory, Pasadena, California.

Hamada, M. and Wu, C. F. J. (1992) Analysis of designed experiments with complex aliasing. *Journal of Quality Technology*, **24**, 130–137.

Harville, D. A. (1977) Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association*, **72**, 320–338.

Hastings, W. K. (1970) Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–109.

Hedayat, A. and Wallis, W. D. (1978) Hadamard matrices and their applications. *Annals of Statistics*, **6**, 1184–1238.

Hurvich, C. M. and Tsai, C. L. (1989) Regression and time series model selection in small samples. *Biometrika*, **76**, 297–307.

Jeffreys, H. (1946) An invariant form of the prior probability in estimation problems. *Proceedings of the Royal Society of London, Series A.*, **186**, 453–461.

Jin, B. and Morgan, J. P. (2008) Optimal saturated block designs when observations are correlated. *Journal of Statistical Planning and Inference*, **138**, 3299–3308.

Jones, B. and Eccleston, J. A. (1980) Exchange and interchange procedures to search for optimal designs. *Journal of the Royal Statistical Society, Series B*, **42**, 238–243.

Jones, B. and Goos, P. (2007) A candidate-set-free algorithm for generating *D*-optimal split-plot designs. *Journal of the Royal Statistical Society, Series C*, **56**, 347–364.

Jones, B., Lin, D. K. J. and Nachtsheim, C. J. (2008) Bayesian *D*-optimal supersaturated designs. *Journal of Statistical Planning and Inference*, **138**, 86–92.

Jones, D. R., Schonlau, M. and Welch, W. J. (1998) Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, **13**, 455–492.

Kiefer, J. (1961) Optimum experimental designs V, with applications to systematic and rotatable designs. *Proceedings of the 4th Berkeley Symposium on Mathematical Statistics and Probability*, **1**, 381–405.

Kiefer, J. and Wynn, H. P. (1981) Optimum balanced block and latin square designs for correlated observations. *The Annals of Statistics*, **9**, 737–757.

— (1984) Optimum and minimax exact treatment designs for one-dimensional autoregressive error processes. *Annals of Statistics*, **12**, 414–450.

Kunert, J. (1987) Neighbour balanced block designs for correlated errors. *Biometrika*, **74**, 717–724.

Kutner, M., Nachtsheim, C. J., Neter, J. and Li, W. (2004) *Applied Linear Statistical Models*. McGraw-Hill, New York, fifth edn.

Läuter, E. (1974) Experimental design in a class of models. *Mathematische Operationsforschung Statistik*, **5**, 379–398.

Loeppky, J. L., Sitter, R. R. and Tang, B. (2007) Nonregular designs with desirable projection properties. *Technometrics*, **49**, 454–467.

Martin, R. J., Eccleston, J. A. and Jones, G. (1998b) Some results on multi-level factorial designs with dependent observations. *Journal of Statistical Planning and Inference*, **73**, 91–111.

Martin, R. J., Jones, G. and Eccleston, J. A. (1998a) Some results on two-level factorial designs with dependent observations. *Journal of Statistical Planning and Inference*, **66**, 363–384.

Matthews, E. (2015) Supplementary materials for 'Design of Factorial Experiments in Blocks and Stages'. *Southampton e-Print no. 377763*, University of Southampton. URL http://eprints.soton.ac.uk/id/eprint/377763.

McIntyre, G. A. (1955) Design and analysis of two phase experiments. *Biometrics*, **11**, 324–334.

McKay, M. D., Beckman, R. J. and Conover, W. J. (1979) Comparison of three methods for selecting values of input variables in the analysis of output from computer code. *Technometrics*, **21**, 239–245.

Mee, R. W. and Xiao, J. R. (2008) Steepest ascent for multiple-response applications. *Technometrics*, **50**, 371–382.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. and Teller, E. (1953) Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087–1092.

Meyer, R. K. and Nachtsheim, C. J. (1995) The coordinate-exchange algorithm for constructing exact optimal experimental designs. *Technometrics*, **37**, 60–69.

Miller, A. (1997) Strip-plot configurations of fractional factorials. *Technometrics*, **39**, 153–161.

Milliken, G. A., Shi, X., Mendicino, M. and Vasudev, P. K. (1998) Strip-plot designs for two-step processes. *Quality and Reliability Engineering International*, **14**, 197–210.

Montgomery, D. C. (2012) *Design and Analysis of Experiments*. John Wiley and Sons, New York, eighth edn.

Morgan, J. P. and Chakravarti, I. M. (1988) Block designs for first and second order neighbor correlations. *Annals of Statistics*, **16**, 1206–1224.

Morris, M. D. (2011) *Design of Experiments: An Introduction Based on Linear Models*. Chapman and Hall/ CRC, New York.

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

Nelder, J. A. and Mead, R. (1965) A simplex method for function minimization. *Computer Journal*, **7**, 308–313.

Ng, S. H. (2010) A Bayesian model-averaging approach for multiple-response optimization. *Journal of Quality Technology*, **42**, 52–68.

O'Hagan, A. and Forster, J. J. (2004) *Kendall's Advanced Theory of Statistics, Volume 2B; Bayesian Inference*. Arnold, London,, second edn.

Overstall, A. M. and Woods, D. C. (2015) Multivariate emulation of computer simulators: model selection and diagnostics with application to a humanitarian relief model. *arXiv 1506.04489*. URL http://xxx.tau.ac.il/pdf/1506.04489v1.pdf.

Pan, W. (2001) Akaike's information criterion in generalized estimating equations. *Biometrics*, **57**, 120–125.

Pantula, S. G. and Pollock, K. H. (1985) Nested analysis of variance with autocorrelated errors. *Biometrics*, **41**, 909–920.

Patterson, H. D. and Thompson, R. (1971) Recovery of inter-block information when block sizes are unequal. *Biometrika*, **58**, 545–554.

Payne, R. W. and Wilkinson, G. N. (1977) A general algorithm for analysis of variance. *Journal of the Royal Statistical Society, Series C*, **26**, 251–260.

Perry, L. A., Montgomery, D. C. and Fowler, J. W. (2001) Partition experimental designs for sequential processes: Part I - first-order models. *Quality and Reliability Engineering International*, **17**, 429–438.

— (2002) Partition experimental designs for sequential processes: Part II - second-order models. *Quality and Reliability Engineering International*, **18**, 373–382.

— (2007) A partition experimental design for a sequential process with a large number of variables. *Quality and Reliability Engineering International*, **23**, 555–564.

Pieracci, J., Perry, L. and Conley, L. (2010) Using partition designs to enhance purification process understanding. *Biotechnology and Bioengineering*, **107**, 814–824.

Plackett, R. L. and Burman, J. P. (1946) The design of optimum multifactorial experiments. *Biometrika*, **33**, 305–325.

Poletti, F., Petrovich, M. N. and Richardson, D. J. (2013) Hollow-core photonic bandgap fibres: technology and applications. *Nanophotonics*, **2**, 315–340.

Ralston, A. and Rabinowitz, P. (2001) *A First Course in Numerical Analysis*. Dover Publications, New York, second edn.

Rasmussen, C. E. and Williams, C. K. I. (2006) *Gaussian Processes for Machine Learning*. Massachusetts Institute of Technology, Cambridge.

Sandoghci, S. R., Jasion, G. T., Wheeler, N. V., Jain, S., Lian, Z., Wooler, J. P., Boardman, R. P., Baddela, N., Chen, Y., Hayes, J., Fokoua, E. N., Bradley, T., Gray,

D. R., Mousavi, S. M., Petrovich, M. N., Poletti, F. and Richardson, D. J. (2014) X-ray tomography for structural analysis of microstructured and multimaterial optical fibers and preforms. *Optics Express*, **22**, 26181–26192.

Santner, T. J., Williams, B. J. and Notz, W. (2003) *The Design and Analysis of Computer Experiments.* Springer-Verlag, New York.

Satpati, S. K., Parsad, R. and Gupta, V. K. (2007) Efficient block designs for dependent observations - a computer-aided search. *Communications in Statistics - Theory and Methods*, **36**, 1187–1223.

Schaalje, B., Zhang, J., Pantula, S. G. and Pollock, K. H. (1991) Analysis of repeated-measurements data from randomized block experiments. *Biometrics*, **47**, 813–824.

Schafer, J. L. (1997) *Analysis of Incomplete Multivariate Data.* Chapman and Hall /CRC, New York.

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

Spezzaferri, F. (1988) Nonsequential designs for model descrimination. In *Bayesian Statistics 3* (eds. J. M. Bernado, M. H. DeGroot, D. V. Lindley and A. F. M. Smith). Clarendon Press, Oxford.

Stein, M. (1999) *Colloqium Publications Volume 23: Interpolation of Spatial Data: Some theory for Kriging.* New York: Springer-Verlag, New York.

Stein, M. L. (1987) Large sample properties of simulations using Latin hypercube sampling. *Technometrics*, **29**, 143–151.

Szegö, G. (1975) *Orthogonal Polynomials.* American Mathematical Society, Rhode Island, fourth edn.

Tack, L. and Vandebroek, M. (2002) Trend-resistant and cost-efficient block designs with fixed or random block effects. *Journal of Quality Technology*, **34**, 422–436.

Takeuchi, K. (1961) On the optimality of certain type of PBIB designs. *Reports of Statistical Applications Research, Union of Japanese Scientists and Engineers*, **8**, 140–145.

Tan, M. H. Y. and Wu, C. F. J. (2013) A Bayesian approach for model selection in fractionated split-plot experiments with applications in robust parameter design. *Technometrics*, **55**, 359–372.

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

Vivacqua, C. A. and Bisgaard, S. (2004) Strip-block experiments for process improvement and robustness. *Quality Engineering*, **16**, 495–200.

— (2009) Post-fractionated strip-block designs. *Technometrics*, **51**, 47–55.

Wilkinson, G. N. (1970) A general recursive procedure for analysis of variance. *Biometrika*, **57**, 19–46.

Williams, R. M. (1952) Experimental designs for serially correlated observations. *Biometrika*, **39**, 151–167.

Wu, C. F. J. and Hamada, M. S. (2009) *Experiments: Planning, Analysis and Optimization*, vol. 2nd. Wiley, New York.

Zergaw, G. (1989) A sequential method of constructing optimal block designs. *Australian Journal of Statistics*, **31**, 333–342.