UNIVERSITY OF SOUTHAMPTON

FACULTY OF SOCIAL, HUMAN AND MATHMATICAL SCIENCES

Mathematical Sciences

Statistical Analysis of Data from Experiments Subject to Restricted Randomisation

by

Sadiah Aljeddani

Thesis submitted for the degree of Doctor of Philosophy

July 2018

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL, HUMAN AND MATHMATICAL SCIENCES
Mathematical Sciences

Doctor of Philosophy

STATISTICAL ANALYSIS OF DATA FROM EXPERIMENTS SUBJECT TO RESTRICTED RANDOMISATION

by Sadiah Aljeddani

The selection of the best subset of variables, which will have a strong effect on an outcome of interest, is fundamental when avoiding overfitting in statistical modelling. However, when there are many variables, it is computationally difficult to find this best subset. The difficulties of variable selection would be more complex when designs are with restricted randomisation. This work aims to fill the gap of variable selection and model estimation for data from experiments subject to restricted randomisation by developing new methods for variable selection and model estimation using frequentist analysis and Bayesian analysis for experiments subject to restricted randomisation. Frequentist and Bayesian analysis methods are used to carry out a comparative study with respect to their performance in variable selection and model estimation. As a representative of frequentist analysis, the Penalised Generalised Least Square (PGLS) estimator is used in which a single shrinkage parameter is applied to all regression effects. Furthermore, as two different strata in split-plot design are existed, the PGLS approach is extended to perform variable selection and model estimation simultaneously in the context of split-plot design. The Penalised Generalised Least Squares for Split-Plot Design estimator (PGLS-SPD) is utilized, in which two shrinkage parameters are applied, one for the subplot effects and the other for the whole-plot effects. As a representative of Bayesian analysis, the Stochastic Search Variable Selection (SSVS) technique is used. This performs variable selection and model estimation simultaneously where the variance of all active factors will be sampled from one posterior distribution. As two different strata in split-plot design are existed, the SSVS approach to perform Bayesian variable selection is extended for the analysis of data from restricted randomised experiments by introducing the Stochastic Search Variable Selection for Split-Plot Design (SSVS-SPD) in which the variances of the active subplot and whole-plot factors are sampled from two different posterior distributions. The usefulness of frequentist and Bayesian approaches are demonstrated using two practical examples, and their properties are studied in simulation studies. The result of the comparative study of frequentist analysis and Bayesian analysis supports the utilization of SSVS-SPD method for the statistical analysis of data from experiments subject to restricted randomisation.

Contents

List of Figures

List of Tables

Declaration of Authorship

I, Sadiah Aljeddani , declare that the thesis entitled *Statistical Analysis of Data from Experiments Subject to Restricted Randomisation* 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

Foremost, I offer my sincerest gratitude to my supervisors Dr. Kalliopi Mylona and Prof. David Woods for their motivation, enthusiasm, and immense knowledge. Their guidance helped me in all the time of research and completing this thesis. I also express my gratitude to my parents for their unlimited support and to my husband who encouraged me to complete this work.

Chapter 1

Introduction

Two fundamental goals in statistical learning are to ensure high prediction accuracy, and to discover significant variables. One can select a model from a set of candidate models by analysing their data in such away as to balance the two previous goals. Ensuring the selected model has the highest prediction accuracy will allow the model to be used for future prediction. Parallel to this, selecting the simplest model by avoiding uninformative variables, which often do not influence the response variable, is also crucial to enhance scientific analysis. By removing the non-informative or non-active variables, the predictive ability of models can be improved and parsimoniously describe the relationship between the informative, or active variables and the response variable. Variable selection issue refers to obtaining an adequate subset of variables for the model. The subject of variable selection in linear regression analysis is a remarkable subject. The experimenter initially may be uncertain about the most influential structure of the model. It might be unclear that whether all of the variables should be included to the model or if only some of them have significant effects on the response variable. Therefore, variable selection procedure builds a regression model with appropriate subset of variables. In the literature, many variable selection methods have been proposed. A general review of some of these methods has been provided in this chapter.

The aim of this project is to develop both a frequentist methodology and a Bayesian methodology for selecting the active variables for data produced by experiments with

restricted randomisation, where the split-plot design is a particular example of such experiments. Furthermore, we aim to carry out a comparison between frequentist analysis and Bayesian analysis with respect to variable selection and model estimation.
In Section 1.1 of this chapter, an introduction to the design of experiments will be provided. Motivating examples underlying this work will be discussed in Section 1.2. The model which has been used to analyse results from split-plot experiments will be introduced in Section 1.3. In Section 1.4, introduction to frequentist variable selection methods will be discussed. Also, an introduction to Bayesian variable selection methods will be given in Section 1.5. Finally, in Section 1.6, an overview of this work will be provided.

1.1 The Design of Experiments

This section introduces designed experiments and experiments with restricted randomisation. "An experiment is a planned intervention undertaken to observe the effects of one or more explanatory variables, often called "factors", on a response variable" (Peck et al., 2015). The basic goal of the experiment is to find out the relation between the explanatory and response variables. A "treatment" is a combination of values for the explanatory variables. Also, "an individual or a group of individuals used in the experiment can be defined as an experimental unit" (Peck et al., 2015). Moreover, what constitutes the experimental unit is determined by how the treatments are assigned. The design of an experiment is the overall plan for conducting an experiment where a good design minimises ambiguity in the interpretation of the results (Peck et al., 2015). Responses from the same experimental units can vary considerably. Obviously, variations can occur if the same experimental unit is experimented on using different values for the variables. However, even if the same values are used for the variables, variations can be produced by a large number of unknown sources, such as extraneous factors related to the environment, and variations in the experimental material. Such variability is unavoidable and inherent in the very process of experimentation, and is referred to as experimental error.

1.1.1 Completely Randomised Design and Restricted Randomised Design

Assigning the treatments or factors to be tested to the experimental units is technically known as randomisation, and it reduces the possibility of systematic error. Randomisation allows each treatment or factor in a study to have an equal chance of being in any experimental unit. It also ensures that the experimenter is not biased or partial in any way towards assigning treatments. However, this is not always possible, and hence we have to consider experiments with restricted randomisation. An experiment where all the treatments can be randomised completely by chance without restriction is referred to as a completely randomised design (CRD). In many industrial experiments, randomisation of the experimental units must be restricted to produce the expected assignment according to economical and time issues. For example, the experimenter may want to constrain the change of factor level during the experiment. In the following section, we introduce the split-plot design as an example of the restricted randomised design.

1.1.2 Split-Plot Design

According to Jones and Nachtsheim (2009), a split-plot experiment is a blocked experiment, where the blocks themselves serve as experimental units for a subset of the factors. Thus, there are two levels of experimental units. The blocks are referred to as whole plots, while the experimental units within blocks are called split plots, split units, or subplots. Corresponding to the two levels of experimental units are two levels of randomisation. One randomisation is conducted to determine the assignment of block-level treatments to whole plots. Then, as always in a blocked experiment, a randomisation of treatments to split-plot experimental units occurs within each block or whole plot.
The split-plot design was firstly invented by Fisher (1925) for a field experiment. A split-plot design involves two or more factors (for example A and B), but the experimental units receiving factor A will be of a different size than those receiving factor B. Factor A becomes a "whole-plot" factor, with its levels randomly assigned to some larger experimental unit (e.g. agricultural plots); levels of factor B are randomly assigned to some

smaller experimental unit within the whole plots (e.g., plants within plots). Three good examples of split-plot designs can be found in the article: "How to Recognize a Split Plot Experiment" (Kowalski and Potcner, 2003). In the statistical analysis of split-plot designs, the presence of two different sizes of strata must be taken into account. This leads to correlated observations within same plot.

There has been a lot of interest in experiments run in split-plot or other multistratum structures, they are discussed by authors such as Fisher (1925), Anbari and Lucas (1994), Miller (1997), and Jones and Nachtsheim (2009). The split-plot experiments have some factors which are more difficult to set than others. Hard-to-set (whole-plot) factors are blocked and are held constant, while the easy-to-set (subplot) factors can vary within the whole-plots. "All information on the whole-plot factors appears in the main plot stratum, so inferences on their effects depend mainly on the variance between main plots, whereas most information on the subplot factors and their interactions with the whole-plot factors appears in the subplot stratum, so inferences on their effects depend mainly on the variance between subplots" (Gilmour and Goos, 2009). In split-plot designs, observations belonging to the same whole-plot are correlated because of the restriction on the randomisation. For model estimation in this case, generalised least squares, and not ordinary least squares, should be used, and two variance components must be estimated (Goos, 2012). Further discussion on the mixed effects models for split-plot designs is given in Section 1.3.

1.2 Motivating Examples

Two examples motivated this work, the wind tunnel experiment and the freeze-dried coffee experiment. This section introduces these two motivating examples.

1.2.1 The Wind Tunnel Experiment

The design and the response variable for the first motivating example can be found in Simpson et al. (2004). The aim of Simpson et al's paper is the design and analysis of the wind tunnel experiment. The process of automobile wind tunnel testing has the primary

Factor	Low	Center	High
Front Ride Height	-0.5 in.	0 in.	+0.5 in.
Rear Ride Height	-1.0 in.	0 in.	+1.0 in.
Yaw Angle	-3 deg.	-1 deg.	+1 deg.
Grille Tape	0% cover	50% cover	100% cover

Figure 1.1: 1997 Chevrolet Monte Carlo Winston Cup car (Simpson et al., 2004)

objective of characterising aerodynamic performance and developing improvements. The four relevant factors, as given by Simpson et al. (2004), are changes in yaw angle, changes in front ride height, changes in rear ride height, and the changes in grill tape in the NASCAR Winston Cup Chevrolet Monte Carlo stock car Figure 1.1. Four responses are of interest; y_1 (coefficient of lift-front), y_2 (coefficient of lift-rear), y_3 (coefficient of drag), and y_4 (lift over drag ratio). As (Simpson et al., 2004) studied the impacts of the experimental variables on the coefficient of drag y_3, we will discuss the analysis for y_3 by our approaches in this work to be compared with their analysis. The purpose of this experiment was to determine the impact of changes in yaw angle, ride height, and grille tape on the coefficients of lift and drag of the car. The magnitude of the interactions and quadratic effects are also of great interest, and are therefore included, along with the main effects, in the model. Details of these four responses and four factors in this experiment are given in Table 1.1 and Table 1.2 respectively. During the experiment, the tunnel had to be closed to make changes to the ride heights of each wheel. This time-consuming process necessitated experimentation with restrictions on randomisation, where the two ride height factors were hard-to-change. The two remaining factors, yaw angle and grille tape, were fairly easily changed because yaw angle is electronically automated and grille tape only requires covering or uncovering all or part of the cooling inlet with duct tape.

In this experiment, there were two randomisations, randomly selecting front ride height

Table 1.1: Wind tunnel car test responses

Response	Notation	Notes
Coefficient of lift-front	y_1	Measured to indicate downforce over the front axle
Coefficient of lift-rear	y_2	Measured to indicate downforce over the rear axle
Coefficient of drag	y_3	Measured to indicate the way the car passes through the surrounding air
Lift over drag ratio	y_4	Measured to indicate the ratio of downforce to drag

Table 1.2: Wind tunnel car test factors

Factor (label)	Type	Low level -1	Center level 0	High level 1
Front ride height	Hard-to-change (w_1)	-0.5 in	0 in	$+0.5$ in
Rear ride height	Hard-to-change (w_2)	-1.0 in	0 in	$+1.0$ in
Yaw angle	Easy-to-change (s_1)	$-3.0°$	$-1.0°$	$+1.0°$
Grille tape coverage	Easy-to-change (s_2)	0%	50%	100%

and rear ride height of the car and then randomly running the factor combinations in yaw angle and grille tape. This approach resulted in two error terms, one associated with the ride height factors and the other associated with the yaw angle and grille tape factors, as well as interactions between the ride height factors and yaw angle/grill tape. This experiment was therefore a split-plot design experiment, where the front and rear ride height were whole plot factors, and yaw angle and grille tape were subplot factors. As in Table 1.2, the four experimental variables are denoted by w_1, w_2, s_1, and s_2. These were each included at three levels. The design of this experiment is shown in Table 1.3.

1.2.2 The Freeze-Dried Coffee Experiment

The experiment was described by Gilmour and Trinca (2000), to investigate the effects of several process factors on the retention of volatile compound in the freeze-drying of coffee. This experiment aimed to study the aroma retention in the freeze-drying of the coffee. Each experimental run involved four steps:

- preparation of the coffee solution;
- addition of the maker volatile compounds;
- the freeze-drying;
- analysis of volatile compounds.

Run	WP	s_1	s_2	w_1	w_2	y_3
1	8	0	0	-1	-1	0.382
2	8	1	-1	-1	-1	0.380
3	8	-1	-1	-1	-1	0.402
4	8	1	1	-1	-1	0.369
5	8	-1	1	-1	-1	0.394
6	2	1	-1	-1	1	0.393
7	2	-1	1	-1	1	0.41
8	2	-1	-1	-1	1	0.419
9	2	1	1	-1	1	0.386
10	2	0	0	-1	1	0.400
11	3	1	-1	1	1	0.414
12	3	0	0	1	1	0.421
13	3	1	1	1	1	0.405
14	3	-1	-1	1	1	0.435
15	3	-1	1	1	1	0.428
16	5	0	0	1	-1	0.400
17	5	1	1	1	-1	0.383
18	5	-1	-1	1	-1	0.418
19	5	-1	1	1	-1	0.408
20	5	1	-1	1	-1	0.399
21	7	1	-1	0	0	0.394
22	7	0	0	0	0	0.401
23	7	-1	-1	0	0	0.420
24	7	1	1	0	0	0.384
25	7	-1	1	0	0	0.409
26	6	0	0	1	1	0.419
27	6	-1	-1	1	1	0.436
28	6	-1	1	1	1	0.426
29	6	1	1	1	1	0.403
30	6	1	-1	1	1	0.412
31	4	-1	-1	-1	1	0.420
32	4	-1	1	-1	1	0.412
33	4	1	1	-1	1	0.387
34	4	1	-1	-1	1	0.394
35	4	0	0	-1	1	0.404
36	9	1	1	-1	-1	0.365
37	9	0	0	-1	-1	0.383
38	9	-1	1	-1	-1	0.391
39	9	1	-1	-1	-1	0.382
40	9	-1	-1	-1	-1	0.400
41	1	1	1	1	-1	0.385
42	1	1	-1	1	-1	0.398
43	1	0	0	1	-1	0.401
44	1	-1	-1	1	-1	0.416
45	1	-1	1	1	-1	0.409

Table 1.3: The design of the wind tunnel experiment. The WP is short for the "whole plot".

Table 1.4: Freeze dried coffee experiment factors

Factor (label)	Type	Low level -1	Center level 0	High level 1
Pressure	Hard-to-change (w)	30	50	70
Temperature	Easy-to-change (s_1)	25	35	45
Solids content	Easy-to-change (s_2)	0.1	0.2	0.3
Slab thickness	Easy-to-change (s_3)	1.0	1.5	2.0
Freezing rate	Easy-to-change (s_4)	Slow*	Medium*	Fast*

Slow*: 18 hours at 18°C and 6 hours at -35°C.
Medium*: 6 hours at 18°C and 18 hours at -35°C.
Fast*: 24 hours at -35°C.

This experiment was explained in more detail by Gilmour and Trinca (2000) and Gilmour and Goos (2009). Five factors were studied, the pressure in the drying chamber (w), the heating temperature (s_1), the initial solids content in the coffee solution (s_2), the thickness of the slab freeze-dried (s_3), and the freezing rate (s_4). Each factor was included at three levels, as shown in Table 1.4. For the freezing rate (s_4), the linear term estimates the difference between fast and slow freezing rate and the quadratic term estimates the difference between the average response for the fast and slow freezing rate and the response from the medium freezing rate. The pressure was controlled manually by a needle valve, which is a very time-consuming process, and the experimenters preferred to run all treatment combinations at each level of the pressure. Having only three main whole plots each with a specific level of the pressure would not allow a clear estimate of the whole plot variance. Therefore, a replication was performed for the three main units yielded in six main units each containing five runs to be able to estimate the variance of the pressure unit random effect. The design of this experiment is shown in Table 1.5. The four responses from this experiment are the drying rate (y_1), benzaldehyde (y_2), 4-ethylbenzaldehyde (y_3), and 2-methoxy-4-methylphenol (y_4) retained after freeze-drying. In this work, we discuss the analysis for drying rate y_1 by our approaches to be compared with the analysis of y_1 in Gilmour and Goos (2009).

1.3 Model and Analysis

The model for the split-plot experiments in Section 1.1.2 includes two types of errors; whole-plot random errors and subplot random errors. Hence, linear mixed models (LMM) are used to analyse responses from the split-plot experiments.

Run	WP	w	s_1	s_2	s_3	s_4	y_1
1	1	1	0	0	0	1	66.000
2	1	1	0	0	1	0	66.094
3	1	1	-1	0	0	0	57.848
4	1	1	0	0	0	0	66.000
5	1	1	0	1	0	0	51.871
6	2	0	0	0	0	0	70.884
7	2	0	-1	1	-1	1	56.763
8	2	0	1	1	1	-1	62.423
9	2	0	1	-1	-1	-1	83.570
10	2	0	-1	-1	1	1	65.191
11	3	-1	0	0	0	0	71.379
12	3	-1	1	1	1	1	97.931
13	3	-1	-1	1	-1	-1	54.947
14	3	-1	-1	-1	1	-1	61.704
15	3	-1	1	-1	-1	1	80.410
16	4	1	0	0	-1	0	66.934
17	4	1	1	0	0	0	79.220
18	4	1	0	0	0	-1	65.203
19	4	1	0	-1	0	0	73.835
20	4	1	0	0	0	0	67.941
21	5	-1	0	0	0	0	69.184
22	5	-1	1	1	-1	1	85.379
23	5	-1	1	-1	1	-1	74.300
24	5	-1	-1	1	1	-1	50.360
25	5	-1	-1	-1	-1	1	60.266
26	6	0	1	-1	1	1	89.160
27	6	0	0	0	0	0	68.500
28	6	0	1	1	-1	-1	75.570
29	6	0	-1	1	1	1	56.470
30	6	0	-1	-1	-1	-1	68.388

Table 1.5: The design of the freeze-dried coffee experiment. The WP is short for the "whole plot".

1.3.1 Linear Mixed Model for Analysing the Split-Plot Experiments

Linear mixed effects models (LMMs) introduce correlations between observations through the use of random effects. This leads to the use of generalised least squares (GLS) estimation, combined with restricted maximum likelihood estimation (REML) of the variance components as will be discussed in Section 1.3.2. This type of analysis is used by most design of experiments textbooks that deal with split-plot designs.

In matrix notation, the model corresponding to a split-plot design is written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{1.1}$$

where $\mathbf{Y}$ is $n \times 1$ vector of observations on the response of interest, $\mathbf{X}$ is the $n \times p$ model design matrix containing the polynomial expansions of the m factor levels at the n experimental runs, $\boldsymbol{\beta}$ is the $p \times 1$ vector of unknown fixed parameters, $\mathbf{Z}$ is an $n \times b$ random design matrix which represents the allocation of the runs to whole plot, and whose (i, j)th element is one where the ith observation belongs to the jth whole plot, and zero otherwise. If the runs of the experiment are grouped per whole-plot, then $\mathbf{Z}$ is of the form

$$\mathbf{Z} = \text{diag}[\mathbf{1}_{k_1}, \mathbf{1}_{k_2}, \ldots, \mathbf{1}_{k_b}],$$

where $\mathbf{1}_k$ is a k vector of ones, and $k_1, k_2, \ldots, k_b$ are the blocks sizes.

The random effects of the b whole-plots are contained within the $b \times 1$ vector $\boldsymbol{\gamma}$, and the random errors are contained within the $n \times 1$ vector $\boldsymbol{\epsilon}$. It is assumed that $\boldsymbol{\gamma}$ and $\boldsymbol{\epsilon}$ are independent and normally distributed, i.e. Cov $(\boldsymbol{\gamma}, \boldsymbol{\epsilon}) = \mathbf{0}_{b \times n}$, where $\mathbf{0}_{b \times n}$ is the $b \times n$ matrix of zeros. Hence, $\boldsymbol{\gamma} \sim \text{N}(\mathbf{0}_b, \sigma^2_\gamma \mathbf{I}_b)$, and $\boldsymbol{\epsilon} \sim \text{N}(\mathbf{0}_n, \sigma^2_\epsilon \mathbf{I}_n)$, where $\mathbf{0}_b$ and $\mathbf{0}_n$ are the b and n column vectors of zeros respectively, and $\mathbf{I}_b$ and $\mathbf{I}_n$ are the $b-$dimensional and $n-$dimensional identity matrices respectively.

Under these assumptions, $\mathbf{Y}$ is a normally distributed random variable with mean $\mathbb{E}(\mathbf{Y}) = \mathbf{X}\boldsymbol{\beta}$, and the variance-covariance matrix of the response $\mathbf{Y}$ can be written as

$$\begin{aligned}
\mathbf{V} = \text{Var}(\mathbf{Y}) &= \text{Var}(\mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}) \\
&= \text{Var}(\mathbf{Z}\boldsymbol{\gamma}) + \text{Var}(\boldsymbol{\epsilon}) \\
&= \mathbf{Z}\text{Var}(\boldsymbol{\gamma})\mathbf{Z}' + \sigma^2_\epsilon \mathbf{I}_n \\
&= \sigma^2_\gamma \mathbf{Z}\mathbf{Z}' + \sigma^2_\epsilon \mathbf{I}_n.
\end{aligned}$$

$\mathbf{V}$ can be given as a block diagonal,

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_1 & 0 & \dots & 0 \\ 0 & \mathbf{V}_2 & \dots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & \dots & 0 & \mathbf{V}_b \end{bmatrix},$$

where

$$\mathbf{V}_i = \sigma_\epsilon^2 \mathbf{I}_{k_i} + \sigma_\gamma^2 \mathbf{1}_{k_i} \mathbf{1}'_{k_i},$$

and

$$\mathbf{V}_i = \begin{bmatrix} \sigma_\epsilon^2 + \sigma_\gamma^2 & \sigma_\gamma^2 & \dots & \sigma_\gamma^2 \\ \sigma_\gamma^2 & \sigma_\epsilon^2 + \sigma_\gamma^2 & \dots & \sigma_\gamma^2 \\ \vdots & \ddots & \ddots & \vdots \\ \sigma_\gamma^2 & \dots & \sigma_\gamma^2 & \sigma_\epsilon^2 + \sigma_\gamma^2 \end{bmatrix}.$$

As a result, the variance-covariance matrix $\mathbf{V}_i$ of all observations within one whole-plot is compound symmetric: the main diagonal of the matrix contains the variances of the observations, while the off-diagonal elements are covariances. However, $\mathbf{V}_i$ can be rewritten as

$$\begin{aligned} \mathbf{V}_i &= \sigma_\epsilon^2 (\mathbf{I}_{k_i \times k_i} + \frac{\sigma_\gamma^2}{\sigma_\epsilon^2} \mathbf{1}_{k_i} \mathbf{1}'_{k_i}), \\ &= \sigma_\epsilon^2 (\mathbf{I}_n + \eta \mathbf{Z}\mathbf{Z}'), \end{aligned}$$

where $\eta = \sigma_\gamma^2 / \sigma_\epsilon^2$ is a measure for the extent to which observations within the same whole-plot are correlated. The larger this variance ratio, the stronger observations within the same whole-plot are correlated.

1.3.2 Likelihood Inference for the Split-Plot Model

When the random error terms as well as the group effects are normally distributed, the maximum likelihood estimate of the unknown model parameter $\boldsymbol{\beta}$ in (1.1) is the generalised least squares (GLS) estimate. Detecting the estimator $\hat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, requires to minimise

$$(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}) = \mathbf{y}'\mathbf{V}^{-1}\mathbf{y} - 2\boldsymbol{\beta}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{y} + \boldsymbol{\beta}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta} \tag{1.2}$$

with respect to $\boldsymbol{\beta}$, which is tantamount to detecting $\hat{\boldsymbol{\beta}}$, so that

$$(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}.$$

Therefore, the generalised least squares (GLS) estimator of $\boldsymbol{\beta}$ is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y}, \tag{1.3}$$

and the variance-covariance matrix of the estimators is given by

$$\begin{aligned}
\mathrm{Var}(\hat{\boldsymbol{\beta}}) &= \mathrm{Var}\Big((\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{Y})\Big) \\
&= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathrm{Var}(\mathbf{Y})\Big((\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\Big)' \\
&= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}\mathbf{X}'\mathbf{V}^{-1}\mathbf{V}\mathbf{V}^{-1}\mathbf{X}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})(\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1} \\
&= (\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})^{-1}.
\end{aligned} \tag{1.4}$$

Often, the variances σ_γ^2 and σ_ϵ^2 are not known and therefore, (1.3) and (1.4) cannot be used directly. Instead, the estimates of the variance components, $\hat{\sigma}_\gamma^2$ and $\hat{\sigma}_\epsilon^2$, are substituted in the GLS estimator (1.3), yielding

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{Y}, \tag{1.5}$$

where

$$\hat{\mathbf{V}} = \hat{\sigma}_{\epsilon}^2 \mathbf{I}_n + \hat{\sigma}_{\gamma}^2 \mathbf{Z}\mathbf{Z}'. \tag{1.6}$$

In that case, the variance-covariance matrix (1.4) can be approximated by

$$\text{Var}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X})^{-1}. \tag{1.7}$$

The generalised least square (GLS) estimator is unbiased, meaning that $\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}$, and is equal to the maximum likelihood estimator (MLE). The likelihood function defined as it is the joint probability density function for the observed data examined as a function of the parameters. Hence, the likelihood function for $\mathbf{Y}$ in (1.1) is

$$L(\boldsymbol{\beta}|\mathbf{Y}) = (2\pi)^{-n/2}|\mathbf{V}|^{-1/2}\text{exp}\big[-\frac{1}{2}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})\big], \tag{1.8}$$

where π is a constant which does not depend on $\boldsymbol{\beta}$. The maximum likelihood estimator (MLE) is the estimator that maximises the likelihood function, which is tantamount to detecting the $\hat{\boldsymbol{\beta}}$ as

$$\frac{\partial}{\partial \boldsymbol{\beta}} L(\hat{\boldsymbol{\beta}}|\mathbf{Y}) = 0, \tag{1.9}$$

which is equal to

$$\frac{\partial}{\partial \boldsymbol{\beta}} \text{ln}L(\hat{\boldsymbol{\beta}}|\mathbf{Y}) = 0, \tag{1.10}$$

where $\text{ln}L(\hat{\boldsymbol{\beta}}|\mathbf{Y})$ is the log likelihood function. As (1.2) is proportionate to log of (1.8), the GLS estimator in (1.3) is the result of (1.9) and (1.10).

Moreover, $\mathbf{V}$ can be estimated when observed data is obtained. In this work, we used the Restricted Maximum Likelihood (REML) estimator to estimate $\mathbf{V}$. According to Matthews (2015),“ 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”. The likelihood in REML includes knowledge about the variance components yet does not include knowledge about the fixed effects (Corbeil and Searle, 1976).

The restricted maximum likelihood (REML) used to estimate σ_ϵ^2 and σ_γ^2 is

$$l_{REML}(\sigma_\epsilon^2, \sigma_\gamma^2; \mathbf{Y}) = -\frac{1}{2}\ln|\mathbf{V}| - \frac{1}{2}\ln|\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}| - \frac{1}{2}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\hat{\boldsymbol{\beta}}),$$

where $\hat{\boldsymbol{\beta}}$ is defined in (1.5). The restricted log-likelihood $l_{REML}(\sigma_\epsilon^2, \sigma_\gamma^2; \mathbf{Y})$ is minimised with respect to the variance components σ_ϵ^2 and σ_γ^2 to obtain an unbiased estimate for the the variance components. In this work, REML is minimised by using the function `'fmincon'` in `Matlab`.

The work by Gilmour and Goos (2009) that explained the weakness of likelihood analysis for non-orthogonal and small designs must be mentioned. They concluded that the likelihood methods are based mainly on asymptotic results which require a large number of whole-plot units. Therefore, for split-plot experiments and for non-orthogonal structures, the likelihood methods require large numbers of whole-plot units to allow enough degrees of freedom to estimate the variance random effect $\hat{\sigma}_\gamma^2$. For orthogonal split-plot structures, the number of whole-plot units does not strongly affect the estimated variance components. In this work, the two different experiments in Section 1.2.1 and Section 1.2.2 are examined, the wind tunnel experiment with $n = 45$ observations and 9 whole-plot units, and the freeze-dried coffee experiment with $n = 30$ observations and 6 whole-plot units.

1.4 Introduction to Frequentist Variable Selection Methods

The traditional model selection such as backward elimination, forward selection and stepwise selection have been utilized for long time to select a subset of linear model variables. However, the traditional model selection methods have some drawbacks and more effective methods have been discussed in literature (Hastie et al., 2009). In variable selection methods, we give attention to the cohesion in variable selection and producing easy to interpret model.

1.4.1 Subset Selection

One of the traditional variable selection methods is the best subset regression. In the subsets regression, we fit all the available models up to a certain size, and the best model will be selected with respect to some model selection criteria. The model selection criteria will be explained in Section 2.3. There are many types of subset selection methods, for example, forward selection, backward elimination, and stepwise selection.

- **Forward selection**

  Forward selection starts with an empty set, and adds one variable at a time. The most significant variable, meaning its corresponding $p-$value is below α significance level, will be added to the model. The main disadvantage of this procedure is that the addition of a new variable at each step may change the state of significance of the old variables which were already included in previous steps.

- **Backward elimination**

  Backward elimination starts with the full model, then the least significant variable, which corresponds to the highest p-value above a significance level α, will be dropped. The reduced models in each step are re-fitted by following the rule of significance level until all remaining variables are statistically significant meaning the corresponding variable has a $p-$value $\leq \alpha$.

- **Stepwise selection**

  Stepwise selection allows movements in either direction, dropping or adding variables at different steps. The process is one of alternations between choosing the least significant variable which has the highest $p-$value and for a given significance level α to drop and then re-considering all variables including those variables dropped previously (except the one that was dropped most recently) for re-introduction into the model.

However, these traditional approaches have some drawbacks as they are unstable in the sense that small changes in the data can result in completely different estimates. Also, they can be computationally cumbersome if there are a large number of predictors, as

these methods require the estimation of a large number of models (Hastie et al., 2009). In this work, backward elimination has been used as an example of the classical subset variable selection as the intention is to compare the models selected by classical subset variable selection methods and by the shrinkage model selection methods described in the next section.

1.4.2 Penalised Generalised Least Squares (PGLS)

To overcome the drawbacks in the classical approaches of variable selection, Hoerl and Kennard (1970) and Breiman et al. (1996) proposed regression modelling by regularisation technique. This technique prevents over fitting by restricting the model, typically to reduce its complexity. The regularisation methods are based on shrinkage penalties, where penalty functions are added to the residual sum of squares or subtracted from the log-likelihood, and minimisation or maximisation of penalised functions with respect to coefficients yields penalised likelihood estimators. The shrinkage penalty method can be explained as there is a penalty for any non-zero estimate of the parameters when we minimise the sum of the squared residuals. Thus, the penalty will shrink the size of the estimated coefficients towards zero. It places a constraint on the size of the regression coefficients (Fan and Li, 2001; Hastie et al., 2009). Shrinkage methods do not explicitly select variables, instead they minimise the sum of the squared residuals by applying a penalty on the size of the estimated coefficients. They have the advantage of selecting variables and estimating the coefficients simultaneously. The advantage of shrinkage methods is that their use often improve the prediction accuracy and helps with the selection of a more parsimonious model, though there is a trade-off between bias and the variance of the final model (see Tibshirani, 1996; Efron et al., 2004).

Based on the size of the estimated coefficients, the penalised estimates might be only shrunk in size while in the case of small estimated coefficients, it is more likely the penalised estimates will be set to zero. Hence, the choice of the shrinkage parameter λ is sensitive and important. Unlike traditional subset selection, penalised regression is a continuous process as it shrinks the size of the coefficients and yields stable models with low prediction errors. However, as some shrinkage penalty functions shrink the size of

the coefficients towards zero and not explicitly zero in the case of large size of estimated coefficients or the case of very small amount of shrinkage parameter λ, the resulting models in such case suffer from complexity and over fitting.

For linear mixed models, the penalised generalised least squares estimates has been discussed by Li and Lin (2003), it can be found by minimising

$$\mathbf{Q}_{PGLS}(\boldsymbol{\beta}) = \frac{1}{2n}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \sum_{j=1}^{d} p_\lambda(\mid \beta_j \mid), \tag{1.11}$$

with respect to $\boldsymbol{\beta}$, where $\hat{\mathbf{V}} = \hat{\sigma}^2_\epsilon \mathbf{I}_n + \hat{\sigma}^2_\gamma \mathbf{Z}\mathbf{Z}'$, and d is the number of the model coefficients $\boldsymbol{\beta}$. The expression $p_\lambda(.)$ is a penalty function and λ is an unknown strictly positive thresholding parameter, which is often selected using information selection criteria as in Section 2.3.

Well-known methods from this family are ridge regression (Hoerl and Kennard, 1970) with L_2 penalty $p_\lambda(\mid \beta \mid) = \lambda \mid \beta \mid^2$, bridge regression (Frank and Friedman, 1993) with L_q penalty $p_\lambda(\mid \beta \mid) = \lambda \mid \beta \mid^q$ for $q > 0$, and LASSO (Tibshirani, 1996, 1997) with L_1 penalty $p_\lambda(\mid \beta \mid) = \lambda \mid \beta \mid$. Note that LASSO and ridge regression are special cases of bridge regression with $q = 1$ and 2 respectively. Fan and Li (2001) developed the smoothly clipped absolute deviation penalty (SCAD). Many other penalty functions in the literature have been proposed, but in this work, some of the most popular functions are used as will be shown in Chapter 2.

1.5 Introduction to Bayesian Variable Selection Methods

An important task in regression building is to determine which variables should be included in the model. Such a method represents the fact that there are small coefficients close to zero on one hand and larger coefficients on the other hand. The priors can be built as a combination of two distributions, a narrow normal distribution centred at zero with a small variance called a "spike", and the other with a flat normal distribution with a large variance to spread over a wide range of parameter values called a "slab". This type of priors are called "Spike-and-Slab" priors (see Figure 1.2). These priors are

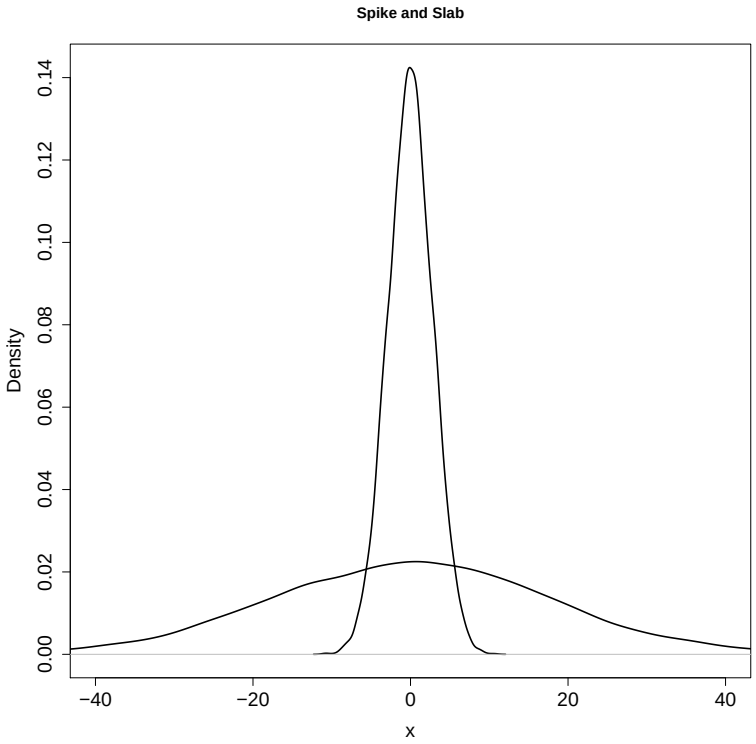


Figure 1.2: Gaussian mixture prior for $\boldsymbol{\beta}$

beneficial for purposes of variable selection because they permit the classification of the regression coefficients into two groups: one group of large, significant effects, and the other group with small, negligible effects.

1.5.1 Stochastic Search Variable Selection (SSVS)

George and McCulloch (1993, 1997) proposed a Bayesian variable selection approach for linear models. Their procedure, called Stochastic Search Variable Selection (SSVS), entails the specification of a hierarchical Bayes mixture prior in Section 4.3 that uses the data to assign larger posterior probability to the more promising models. The SSVS is based on building the entire regression setup in a hierarchical Bayes normal mixture model in which latent variables are used to identify subset choices. With this procedure, the promising subsets of variables can be recognised as those with higher posterior probability. According to George and McCulloch (1997), the SSVS uses the Gibbs sampler to simulate a sample from the posterior distribution. Because high probability models

are more likely to appear quickly, the Gibbs sampler can sometimes identify such models with relatively short runs.

In this work, we apply the SSVS to linear mixed models following Tan and Wu (2013). The difference between our implementation and Tan and Wu (2013) is that we use MCMC to sample from the posterior distributions while their study numerically integrates $\boldsymbol{\beta}$ and σ^2 from the joint posterior distribution of all parameters. They used the normal-inverse gamma prior for those parameters. They were able to do the integration step because they used Gaussian quadrature algorithms to compute the integration.

1.6 Overview

In this work, the analysis of data from split-plot designs has been examined. In particular, a frequentist approach via penalised regression is proposed, so that model selection and parameter estimation can be performed simultaneously when responses from a split-plot experiment are analysed using a linear mixed effects model. The performance of the frequentist approach via penalised regression is compared to the performance of the Bayesian approach via Bayesian variable selection methods. Chapter 2 will describe the frequentist variable selection methods by introducing backward elimination and shrinkage methods. Chapter 3 will present the main results from frequentist analysis. Chapter 4 will discuss Bayesian variable selection for split-plot design. Chapter 5 will present the numerical results from the Bayesian approach, and compares the analysis of the frequentist approach to the analysis of the Bayesian approach. Chapter 6 will summarise the discussion and concludes the thesis with suggestions for future work.

Chapter 2

Frequentist Analysis Methods for Response from Split-Plot Experiments

Experimenters very often would like to know the factors that impact the response variables in their experiment for quality and economic purposes. Model selection is critical for an accurate analysis and model interpretation. In the literature, many model selection methods have been proposed. We have provided a general review of some of these methods. As explained in Section 1.3.1, the linear mixed models (LMMs) are used to analyse responses from split-plot experiments. There are many prior works on mixed penalised regression models. For example, the work by Schelldorfer et al. (2011) proposed an L_1 penalised estimation procedure for high-dimensional linear mixed-effects models. Their aim is to estimate the fixed and random effects parameters as well as the variance components. Their main findings are theoretical results concerning consistency of their proposed estimators. Moreover, the work by Bondell et al. (2010) proposed a methodology to jointly select the fixed and random effects in linear mixed-effects models. Their aim is to identify the significant predictors which correspond to both the fixed and random effects components in a linear mixed-effects models. They apply their method by using a modified Cholesky decomposition.They constrain an Expectation-Maximisation algorithm on a penalised joint log-likelihood with an adaptive penalty to select and to

estimate both fixed and random effects. Another example is the work by Ibrahim et al. (2011). They use a novel re-parametrisation to re-formulate the selection of mixed effects into the problem of grouped variable selection in models with missing data. They use the penalised likelihood methods to select both fixed and random effects. Their aim is to develop a simultaneous fixed and random effects selection procedure based on the SCAD and Adaptive LASSO (ALASSO) penalties for application to longitudinal models and mixed effects models. They apply the Expectation-Maximisation algorithm to simultaneously optimise the penalised likelihood function and estimate the penalty parameters.

Compared to the large body of literature on variable selection procedures using the frequentist analysis, we make a novel contributions in this thesis. Our work is motivated by the split-plot design as we have two different strata. To adapt the issue of having two different strata, we split the penalty function into two parts, and we apply a shrinkage parameter for subplot effects term different than the shrinkage parameter for the whole-plot effects term. Mylona and Goos (2011) introduced the idea of the PGLS-SPD. In this work, we improve the performance of the PGLS-SPD by using different penalty functions and different methods of choosing the shrinkage parameter. The PGLS-SPD reduces the Type I and II error rates as well as it reduces the prediction error of the model for data from split-plot designs compare to the PGLS that applies one penalty function for all factors.

2.1 Penalised Generalised Least Squares for Split-Plot Design (PGLS-SPD)

The penalised regression in (1.11) is a good approach if we want to apply equal sized penalties for all factor effects. According to Jones and Nachtsheim (2009),"A split-plot experiment is a blocked experiment, where the blocks themselves serve as experimental units for a subset of the factors. Thus, there are two levels of experimental units. The blocks are referred to as whole plots, while the experimental units within blocks are called subplots". Therefore, different effect sizes for the two strata are to be expected.

We expect bigger differences between these larger whole plots, naturally leading to larger whole-plot effects. Hence, the PGLS in (1.11) is modified by using two shrinkage penalties, as one will be applied for the subplot effect factors and the other will be applied for the whole-plot effect factors. Therefore, the penalty function $p_\lambda(.)$ in (1.11) will be split into two penalty functions, $p_{\lambda w}(.)$ and $p_{\lambda s}(.)$. The $p_{\lambda w}(.)$ includes the penalty parameter which will be applied to the whole-plot effect factors while the $p_{\lambda s}(.)$ includes the penalty parameter which will be applied to the subplot effect factors. The penalised generalised least squares estimates for split-plot design are obtained by minimising

$$\mathbf{Q}_{PGLS}(\boldsymbol{\beta}) = \frac{1}{2n}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta})'\hat{\mathbf{V}}^{-1}(\mathbf{Y}-\mathbf{X}\boldsymbol{\beta}) + \sum_{j=1}^{d_w} p_{\lambda w}(\mid \beta_j \mid) + \sum_{j=d_w+1}^{d} p_{\lambda s}(\mid \beta_j \mid), \quad (2.1)$$

with respect to $\boldsymbol{\beta}$, where $\hat{\mathbf{V}} = \hat{\sigma}^2_\epsilon \mathbf{I}_n + \hat{\sigma}^2_\gamma \mathbf{Z}\mathbf{Z}'$, and d_w denotes the number of the whole-plot model coefficients, $\beta_1, \ldots, \beta_{d_w}$ which represents these whole-plot factor effects. Furthermore, $\beta_{d_w+1}, \ldots, \beta_d$ represent the subplot factor effects. Moreover, $p_{\lambda w}(.)$ and $p_{\lambda s}(.)$ are the shrinkage penalty functions which will be applied to the whole-plot factor effects and subplot factor effects respectively. Both λ_w and λ_s are unknown positive shrinkage parameters, which can be selected by several selection methods presented in Section 2.3. In this thesis, we improved the performance of the PGLS-SPD by using different shrinkage tuning parameter and different model selection penalty function.

In this chapter, Section 2.2.1 will discuss the computational algorithm for the Penalised Generalised Least Square (PGLS) estimator in which one shrinkage tuning parameter is applied. Also, Section 2.2.2 will discuss the computational algorithm for Penalised Generalised Least Square for Split-Plot Design (PGLS-SPD) estimator in which two shrinkage parameters are applied. One shrinkage tuning parameter is for subplot effect factors and the other is for whole-plot effect factors. The selection of the shrinkage tuning parameters will be explained in Section 2.3. Moreover, Sections 2.4 to 2.9 will discuss the model selection methods that are used in this work. A discussion of this chapter will be provided in Section 2.10.

2.2 Frequentist Variable Selection Algorithms

In this section, we introduce the computational algorithms which we use in the frequentist analysis for variable selection using the PGLS, and the PGLS-SPD. We should standardise the variables when our regression model involves interaction terms and quadratic terms. Although these types of terms may provide useful information about the relationship between the response and the variables, they produce an excessive amount of multicollinearity. Multicollinearity refers to variables that are correlated with other variables in the model. This increases the variance of the coefficient estimates and makes the estimates very sensitive to minor changes in the model. The result is that the coefficient estimates are unstable and difficult to interpret.

Mansfield and Helms (1982) discussed several indications of the multicollinearity problem such as high correlation between pairs of independent variables, statistically non significant regression coefficients on significant variables, and the extreme effects of the changes of sign or magnitude of regression coefficients when an independent variable is included or excluded. Therefore, we standardise our variables before the analysis as this would reduce multicollinearity and the associated problems that are caused by the higher-order terms.

2.2.1 Penalised Generalised Least Square (PGLS) Algorithm

1. Let $\boldsymbol{\beta}^{(0)}$ be the generalised least squares estimator $\hat{\boldsymbol{\beta}}_{GLS}$ as in (1.5), for the full model, a model fitted by all the variables in the experiment, and $\hat{\sigma}^2_{\epsilon}$, $\hat{\sigma}^2_{\gamma}$ be the REML estimates of the variance components for this model.

2. (a) Set a grid for λ with l values $\lambda_1, \lambda_2, \ldots, \lambda_l$ for each grid.

   (b) For $i = 1, 2, \ldots, l$ of the grid, use λ_i to estimate the model parameters of the corresponding tuning.

   (c) For $i = 1, 2, \ldots, l$ of the grid, choose λ_i that minimises cAIC(λ_i) and BIC(λ_i) for λ as defined in Section 2.3.

   (d) Return λ.

3. (a) Set $\boldsymbol{\beta}^{(0)}$ as the GLS estimator and λ given from the previous loop.

   (b) We set the $\text{SE}(\hat{\boldsymbol{\beta}})$ as our proposed threshold to find out if the estimates are statistically significant or not, where $\text{SE}(\hat{\beta}) = \sqrt{\text{Var}(\hat{\beta})}$ in (1.7) is the standard error of $\hat{\boldsymbol{\beta}}$. Thus, we set values of $|\hat{\boldsymbol{\beta}}| \leq \text{SE}(\hat{\boldsymbol{\beta}})$ equal to zero.

   (c) All other (non-zero) coefficients are jointly updated using

$$\hat{\boldsymbol{\beta}}^{(1)} = \left\{\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X} + \mathbf{W}^{(0)}\right\}^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{Y},$$

   where $\hat{\boldsymbol{\beta}}^{(1)}$ is the vector of all non-zero coefficients, and $\mathbf{W}^{(0)}$ is a penalty matrix for the initial values $\hat{\boldsymbol{\beta}}^{(0)}$ which it can be defined as

$$\mathbf{W}^{(0)} = \text{diag}\left\{\frac{p'_\lambda(|\hat{\beta}_1^{(0)}|)}{|\hat{\beta}_1^{(0)}|}, \ldots, \frac{p'_\lambda(|\hat{\beta}_{d^*}^{(0)}|)}{|\hat{\beta}_{d^*}^{(0)}|}\right\},$$

   where the d^* is the total number of non-zero model coefficients.

   (d) Any elements of $\hat{\boldsymbol{\beta}}^{(1)}$ that are $|\hat{\boldsymbol{\beta}}| \leq SE(\hat{\boldsymbol{\beta}})$ are set to zero and the non-zero coefficients are jointly updated along with the matrix $\mathbf{W}^{(0)}$.

   (e) Steps (c) to (d) are repeated until convergence takes place and no more factors can be removed.

4. Denote $\hat{\boldsymbol{\beta}}$ the final estimates of the non-zero model coefficients and $\widehat{\mathbf{W}}$ the corresponding estimated $\mathbf{W}$ penalty matrix.

The covariance of the non-zero parameter estimates can then be obtained from the sandwich formula (Li and Lin, 2003):

$$\widehat{cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X} + \widehat{\mathbf{W}})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X} + \widehat{\mathbf{W}})^{-1}$$

The non-significant coefficients will be replaced with zero indicating that the their associated variables are non-significant.

2.2.2 Penalised Generalised Least Square for Split-plot design (PGLS-SPD) Algorithm

1. Let $\boldsymbol{\beta}^{(0)}$ be the generalised least squares estimator $\hat{\boldsymbol{\beta}}_{GLS}$ as in (1.5), for the full model, a model fitted with all the variables in the experiment, and $\hat{\sigma}^2_\epsilon$, $\hat{\sigma}^2_\gamma$ be the REML estimates of the variance components for this model.

2. (a) Set two similar grids for λ_s and λ_w such that $\lambda_{s_l} = \lambda_{w_l}$ for l values $\lambda_1, \lambda_2, \ldots, \lambda_l$ for each grid.

   (b) For $i = 1, 2, \ldots, l$ from λ_s grid, and for $k = 1, 2, \ldots, l$ from λ_w grid, use λ_i and λ_k to estimate the model parameters of the corresponding tuning.

   (c) Choose λ_i and λ_k that minimises cAIC and BIC for both λ_s and λ_w as defined in Section 2.3.

   (d) Return λ_s and λ_w.

3. (a) Set $\boldsymbol{\beta}^{(0)}$ as the GLS estimator and λ_s and λ_w given from the previous loop.

   (b) We set the SE($\hat{\boldsymbol{\beta}}$) as our proposed threshold to find out if the estimates are statistically significant or not, where $\text{SE}(\hat{\beta}) = \sqrt{\text{Var}(\hat{\beta})}$ in (1.7) is the standard error of $\hat{\boldsymbol{\beta}}$. Thus, we set values of $|\hat{\boldsymbol{\beta}}| \leq \text{SE}(\hat{\boldsymbol{\beta}})$ equal to zero.

   (c) All other (non-zero) coefficients are jointly updated using

$$\hat{\boldsymbol{\beta}}^{(1)} = \left\{ \mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X} + \mathbf{W}^{(0)} \right\}^{-1} \mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{Y},$$

   where $\hat{\boldsymbol{\beta}}^{(1)}$ is the vector of all non-zero coefficients, and $\mathbf{W}^{(0)}$ is a penalty matrix for the initial values $\hat{\boldsymbol{\beta}}^{(0)}$ and it can be defined as $\mathbf{W}^{(0)} = \text{diag}\,(\mathbf{W}_w^{(0)}, \mathbf{W}_s^{(0)})$ where

$$\mathbf{W}_w^{(0)} = \text{diag}\left\{ \frac{p'_{\lambda_w}(|\hat{\beta}_1^{(0)}|)}{|\hat{\beta}_1^{(0)}|}, \ldots, \frac{p'_{\lambda_w}(|\hat{\beta}_{d_w^*}^{(0)}|)}{|\hat{\beta}_{d_w^*}^{(0)}|} \right\},$$

corresponds to the whole-plot effects and d_w^* is the number of non-zero model coefficients that corresponds to the whole-plot effects. And

$$\mathbf{W}_s^{(0)} = \text{diag}\left\{\frac{p'_{\lambda_s}(|\hat{\beta}^{(0)}_{d_{w+1}}|)}{|\hat{\beta}^{(0)}_{dw+1}|}, \ldots, \frac{p'_{\lambda_s}(|\hat{\beta}^{(0)}_{d^*}|)}{|\hat{\beta}^{(0)}_{d^*}|}\right\},$$

corresponds to the sub-plot effects, and the d^* is the total number of non-zero model coefficients.

(d) Any elements of $\hat{\boldsymbol{\beta}}^{(1)}$ that are $|\hat{\boldsymbol{\beta}}| \leq SE(\hat{\boldsymbol{\beta}})$ are set to zero and the non-zero coefficients are jointly updated along with the matrix $\mathbf{W}^{(0)}$.

(e) Steps (c) to (d) are repeated until convergence takes place and no more factors can be removed.

4. Denote $\hat{\boldsymbol{\beta}}$ the final estimates of the non-zero model coefficients and $\widehat{\mathbf{W}}$ the corresponding estimated $\mathbf{W}$ penalty matrix.

The covariance of the non-zero parameter estimates can then be obtained from the sandwich formula (Li and Lin, 2003):

$$\widehat{cov}(\hat{\boldsymbol{\beta}}) = (\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X} + \widehat{\mathbf{W}})^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X}(\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X} + \widehat{\mathbf{W}})^{-1}$$

The non-significant coefficients will be replaced with zero indicating that the their associated variables are non-significant.

2.3 Selection of Shrinkage Tuning Parameter

This section illustrates the selection of the shrinkage parameters, λ, λ_w and λ_s. In the literature, there is no fixed formula to select the tuning parameter, as this depends on the data themselves. In this work, a grid for λ from 0 to 3 was set, and two similar grids were set for λ_w and λ_s from 0 to 3 each. The precise grid of values is (0, 0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2, 2.2, 2.4, 2.6, 2.8, 3). The tuning parameters selected using several information selection criteria.

The maximised log-likelihood for the model fitted to a split-plot design discussed in

Sections 1.3.1 and 1.3.2 is

$$\ln\hat{L} = -\frac{n}{2}\ln(2\pi) - \frac{1}{2}\ln|\hat{\mathbf{V}}| - \frac{1}{2}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})'\hat{\mathbf{V}}^{-1}(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}), \tag{2.2}$$

where $\hat{\boldsymbol{\beta}}$ and $\hat{\mathbf{V}}$ are as in (1.5) and (1.6) respectively.

Let $K = 1 + p_t + v_c$, such that p_t is the number of the active fixed parameters in the fitted penalised least squares model (Wahba, 1980). The p_t can then be computed as $p_t = \text{tr}\big[\mathbf{X}\{\mathbf{X}'\hat{\mathbf{V}}^{-1}\mathbf{X} + \widehat{\mathbf{W}}\}^{-1}\mathbf{X}'\hat{\mathbf{V}}^{-1}\big]$, where $\widehat{\mathbf{W}}$ is a penalty matrix (Li and Lin, 2003) and is computed in Section 2.2. The v_c is the number of variance components that are used in fitting the penalised model. The following information selection criteria can be used in selecting the shrinkage parameters λ, λ_w and λ_s.

1. Corrected Akaike Information Criterion (cAIC) is given by Sugiura (1978) and Hurvich and Tsai (1989),

$$\text{cAIC} = -2\ln\hat{L} + \frac{2Kn}{n - K - 1}.$$

   In the context of regression, several researchers e.g., Sugiura (1978); Hurvich and Tsai (1989); and Anderson and Burnham (1999) have suggested using a corrected version (cAIC) which applies a slightly heavier penalty that depends on K and n.

2. Bayesian Information Criterion (BIC) is given by Schwarz et al. (1978),

$$\text{BIC} = -2\ln\hat{L} + K\ln n.$$

   BIC provides a different balance between lack of fit and complexity. BIC penalises larger models more heavily due to dependency on n and so will tend to prefer smaller models.

2.4 Backward Elimination

We did the generalised least square estimator using the backward elimination to compare with other methods. Backward elimination starts with the full model, and eliminates one variable at a time based on the Wald (**Wa**) test statistic. The Wald test statistic is suitable to compare nested models when the variance covariance matrix changes after each drop. We estimate the variance components σ^2_ϵ and σ^2_γ by REML using the full model, so we deal with the Wald test statistic as the variance components are known. We used the Wald test statistic at 5% significance level, and we compare nested models through the process of backward elimination.

Assuming that $\mathbf{y}$ is normally distributed and the variance components are known, we test the hypothesis:

$$H_0 : \beta_i = 0 \quad \text{vs} \quad H_1 : \beta_i \neq 0$$

The Wald (**Wa**) test statistic will follow an F distribution with 1 and r degrees of freedom.

$$\mathbf{Wa} = \frac{\hat{\beta_i}^2}{\sigma^2_{ii}} \sim F_{(1,r)}$$

where σ^2_{ii} is the (i,i)th element of $\text{Var}(\hat{\boldsymbol{\beta}})$. The degrees of freedom in a statistical calculation represent how many values involved in a calculation that have the freedom to vary. The degrees of freedom can be defined as they are equal to the number of independent observations minus the number of parameters (Toothaker and Miller, 1996). The degrees of freedom could be calculated to guarantee the statistical accuracy of tests statistics such as chi-square tests, t-tests and F-tests. Often, these tests are utilized to make a comparison between the observed data with the data that could be expected to be achieved according to a particular hypothesis.

We want to show that under the above null hypothesis, the distribution of the Wald test statistic is approximately F distribution for our sample sizes. Therefore, for the designs that we have, we will simulate 1000 datasets from the null model and generate the empirical distribution of the test statistic. Then we compare the 1000 realisations of the test statistic to F distribution. We note that the whole-plot factors cannot be approximated well if their associated critical values of F-test statistic have the subplot

degrees of freedom. In contrast, the whole-plot factors have good approximations to F distribution if we used degrees of freedom for the critical value of F-test statistic for whole-plot fixed effects.

We used different degrees of freedom for the critical value of F-test statistic, for the subplot fixed effects than the degrees of freedom for the critical value of F-test statistic, for the whole-plot fixed effects, to have good approximations for both the subplot and the whole-plot factors. We give examples of the advantage of this variation for whole-plot factors in the degrees of freedom as in Figure 2.1 and Figure 2.2.

Also, we keep the degrees of freedom fixed through the process of backward elimination as the variance components are estimated once, using the full model. The degrees of freedom for subplots r_s are, $r_s = n - p_s - p_w - r_w - 1$. Where n is the number of observations, p_s is the number of subplot fixed effects, p_w is the number of whole-plot fixed effects, r_w are the degrees of freedom used for whole-plot variance. Moreover, the degrees of freedom for whole-plots r_w are, $r_w = n_w - p_w - 1$. Where n_w is the number of whole plots or blocks, p_w is the number of whole-plot fixed effects. For example, in this work we studied two examples:

- In the wind tunnel experiment, we have $n = 45$ observations within 9 blocks. We have $p_s = 8$, $p_w = 4$, and $r_w = 4$. Therefore, the Wald (**Wa**) test statistic for subplot effect factors follows $F_{(1,r_s)}$, where $r_s = 28$. Also, we have $n_w = 9$ and $p_w = 4$. Therefore, the Wald (**Wa**) test statistic for whole-plot effect factors follows $F_{(1,r_w)}$, where $r_w = 4$.

- In the freeze-dried coffee experiment, we have $n = 30$ observations within 6 blocks. We have $p_s = 18$, $p_w = 2$, and $r_w = 3$. Therefore, the Wald (**Wa**) test statistic for subplot effect factors follows $F_{(1,r_s)}$, where $r_s = 6$. Also, we have $n_w = 6$ and $p_w = 2$. Therefore, the Wald (**Wa**) test statistic for whole-plot effect factors follows $F_{(1,r_w)}$, where $r_w = 3$.

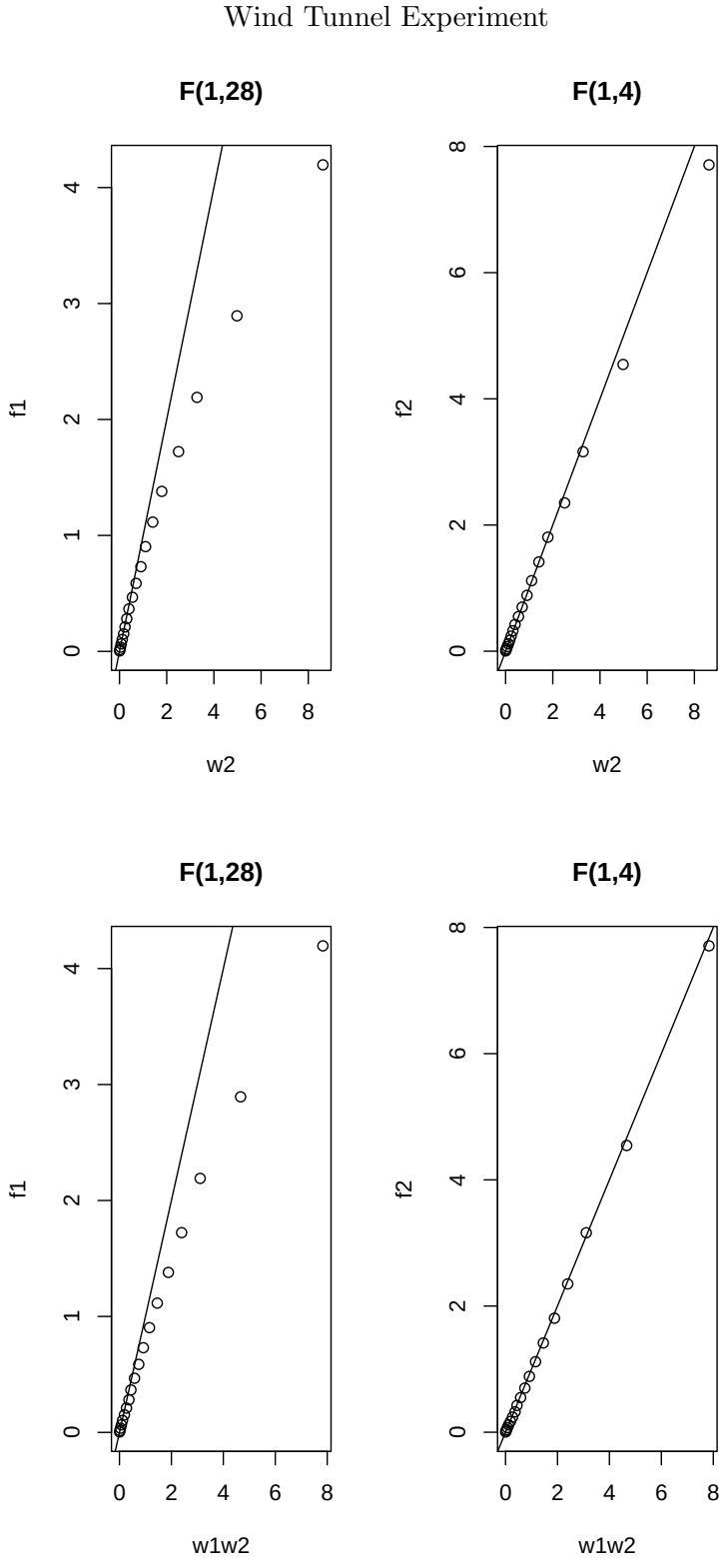


Figure 2.1: 19 quantile for w_2 and w_1w_2 against F(1,28) and F(1,4) in the wind tunnel experiment.

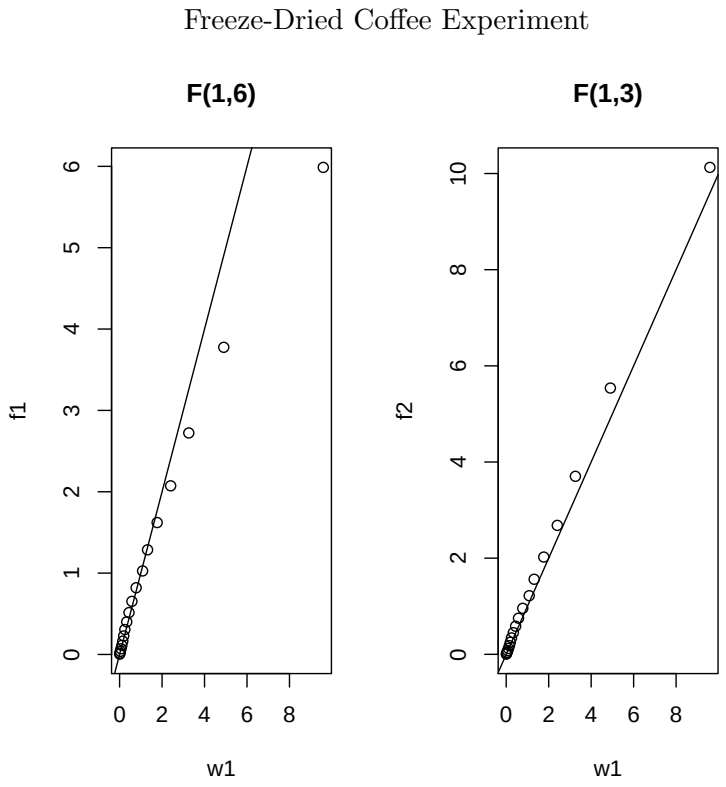


Figure 2.2: 19 quantile for w_1 against F(1,6) and F(1,3) in the freeze-dried coffee experiment.

2.5 Least Absolute Shrinkage and Selection Operator Penalty (LASSO)

According to Tibshirani (1996, 1997), LASSO minimises the residual sum of squares subject to the sum of the absolute value of the coefficients being less than a constant. Thus, this constraint tends to produce some coefficients are exactly zero and shrinks the other coefficients toward zero. Recall the PGLS in (1.11), the penalty term by LASSO is,

$$p_\lambda(\beta) = \lambda \sum_{j=1}^{d} (\mid \beta_j \mid),$$

and recall the PGLS-SPD in (2.1), the penalty terms by LASSO are,

$$p_{\lambda w}(\beta) = \lambda_w \sum_{j=1}^{d_w} (\mid \beta_j \mid) \quad \text{and} \quad p_{\lambda s}(\beta) = \lambda_s \sum_{j=d_w+1}^{d} (\mid \beta_j \mid).$$

Here λ, λ_w, and $\lambda_s \geq 0$ are complexity parameters that control the amount of shrinkage: the larger the value of λ, λ_w, and λ_s, the greater the amount of shrinkage. The LASSO penalty shrinks small coefficients to zero and hence results in a sparse representation of the solution. However, estimation of large coefficients may suffer from substantial bias if the chosen shrinkage parameter is too big, whereas the model may not be sufficiently sparse if the chosen shrinkage parameter is too small (Tibshirani, 1996, 1997). The LASSO shrinkage causes the estimates of the non-zero coefficients to be biased towards zero, and not statistically consistent (as the sample size grows, the estimates do not converge to the true values). The LASSO continuously shrinks the coefficients toward 0 as the shrinkage parameter increases, and some coefficients are shrunk exactly to 0 if the shrinkage parameter is sufficiently large.

2.6 Adaptive LASSO Penalty (ALASSO)

Despite the advantage of LASSO, it has some limitations. In the situation of high dimensional data in which N $<<$ p, LASSO can at most select N variables. Also, in the situations where we have correlated variables, LASSO will select only one variable from a group of correlated variables. In addition, Tibshirani (1996, 1997) discussed the case of N $>>$ p as it is found that for correlated variables, ridge regression proposed by Hoerl and Kennard (1970) in Section 1.4.2 has better prediction accuracy than LASSO. Though the ridge regression does not enforce non significant effects to be zero, but rather minimises their effects on the response variable. Moreover, LASSO is not an oracle procedure as indicated by Zou (2006).

The oracle properties of an estimator are discussed by Fan and Li (2001). They involve the unbiasedness (to avoid unnecessary bias for non-zero parameters), the sparsity (non-informative variable coefficients are estimated as zeros to reduce the model complexity), and the continuity. Zou (2006) proposed the Adaptive LASSO (ALASSO) as an alternative approach to improve the LASSO with respect to obtaining consistency in variable selection and prediction accuracy. ALASSO is based on using a weighted L_1, the LASSO penalty with weight determined by an initial estimator.

Recall the PGLS in (1.11), the penalty term by ALASSO is,

$$p_\lambda(\beta) = \lambda \sum_{j=1}^{d} \widehat{wg}_j |\beta_j|,$$

and recall the PGLS-SPD in (2.1), the penalty terms by ALASSO are,

$$p_{\lambda w}(\beta) = \lambda_w \sum_{j=1}^{d_w} \widehat{wg}_j |\beta_j| \quad \text{and} \quad p_{\lambda s}(\beta) = \lambda_s \sum_{j=d_w+1}^{d} \widehat{wg}_j |\beta_j|.$$

The vector of the weights $\widehat{wg} = (\widehat{wg_1}, \widehat{wg_2}, \ldots, \widehat{wg_d})'$ are the adaptive data-driven weights, where $\widehat{wg}_j$, $j = 1, \ldots, d$ can be constructed by $\widehat{wg}_j = (\hat{\beta}_j^{(0)})^{-\psi}$, where ψ is a positive constant and $\hat{\boldsymbol{\beta}}^{(0)}$ is $\hat{\boldsymbol{\beta}}_{GLS}$ (Zou, 2006).

To construct the adaptive weights $\widehat{wg}$, Zou (2006) suggested to pick a $\psi > 0$. In this work, we set $\psi = 0.5$ as this choice was according our simulation study. We found that this choice provided the lowest Type I and II error rates among the choices that have been investigated during our research process.

2.7 Smoothly Clipped Absolute Deviation Penalty (SCAD)

Fan (1997) and Fan and Li (2001) developed a new penalty function name the Smoothly Clipped Absolute Deviation (SCAD) penalty for variable selection in linear models, which cleverly avoids excessive penalties on large coefficients and enjoys the oracle properties. The SCAD penalty is symmetric, continuous on $(0, \infty)$ and singular at the origin. Fan and Li (2001) showed that under certain regular conditions, the SCAD penalised estimators perform as well as the oracle procedures; in other word, zero coefficients are estimated as zero with probability tending to 1, and non-zero coefficients are estimated as well as if the correct model was known. Recall the PGLS in (1.11), the penalty term by SCAD is,

$$p_\lambda(\beta) = \begin{cases} \lambda|\beta| & \text{if } 0 \le |\beta| < \lambda, \\ \frac{(\alpha^2-1)\lambda^2 - (|\beta| - \alpha\lambda)^2}{2(\alpha-1)} & \text{if } \lambda \le |\beta| < \alpha\lambda, \\ \frac{(\alpha+1)\lambda^2}{2} & \text{if } |\beta| \ge \alpha\lambda, \end{cases}$$

and recall the PGLS-SPD in (2.1), the penalty terms by SCAD are,

$$p_{\lambda w}(\beta) = \begin{cases} \lambda_w|\beta| & \text{if } 0 \leq |\beta| < \lambda_w, \\ \frac{(\alpha^2-1)\lambda_w^2-(|\beta|-\alpha\lambda_w)^2}{2(\alpha-1)} & \text{if } \lambda_w \leq |\beta| < \alpha\lambda_w, \\ \frac{(\alpha+1)\lambda_w^2}{2} & \text{if } |\beta| \geq \alpha\lambda_w, \end{cases}$$

and

$$p_{\lambda s}(\beta) = \begin{cases} \lambda_s|\beta| & \text{if } 0 \leq |\beta| < \lambda_s, \\ \frac{(\alpha^2-1)\lambda_s^2-(|\beta|-\alpha\lambda_s)^2}{2(\alpha-1)} & \text{if } \lambda_s \leq |\beta| < \alpha\lambda_s, \\ \frac{(\alpha+1)\lambda_s^2}{2} & \text{if } |\beta| \geq \alpha\lambda_s, \end{cases}$$

where λ, λ_w, λ_s and α are tuning parameters to be determined. For the tuning parameter α, Fan (1997) and Fan and Li (2001) demonstrated that $\alpha = 3.7$ works well in practice for most models because this value gave a satisfactory performance in variety of variable selection problems. The LASSO shrinkage causes the estimates of the non-zero coefficients to be biased towards zero, and in general they are not consistent. The SCAD applies the LASSO for small coefficients in the first branch. The second branch of the SCAD function would reduce the amount of shrinkage for larger values of coefficients. The last branch of the SCAD function would deal with very large coefficients where not excessive penalisation will be applied for this branch. Unlike the LASSO, the SCAD penalty function gives the best performance in selecting significant variables without creating excessive biases in linear models (Fan and Li, 2001).

2.8 Elastic Net Penalty (EN)

Zou and Hastie (2005) proposed a new regularisation technique that used when there are unknown groups of multicollinear predictors. Recall the PGLS in (1.11), the penalty term by EN is,

$$p_\lambda(\beta) = \lambda \left[(1-\alpha)|\beta|_1 + \alpha\,|\beta|^2\right],$$

and recall the PGLS-SPD in (2.1), the penalty terms by EN are,

$$p_{\lambda w}(\beta) = \lambda_w \left[(1-\alpha)|\beta|_1 + \alpha\,|\beta|^2\right],$$

and

$$p_{\lambda s}(\beta) = \lambda_s \left[(1-\alpha)|\beta|_1 + \alpha\,|\beta|^2\right],$$

where $|\beta|_1 = \sum_{j=1}^{d} |\beta_j|$ represents the LASSO penalty, $|\beta|^2 = \sum_{j=1}^{d} {\beta_j}^2$ represents the ridge regression penalty, and $\alpha \in [0, 1)$. The elastic net estimator can be interpreted as a stabilised version of the LASSO. The EN penalty function is a singular (without first derivative) at 0 and strictly convex for all $\alpha > 0$, thus having the characteristics of both the LASSO and ridge regression (Zou and Hastie, 2005; Zou and Zhang, 2009). We set three values of $\alpha = 0.1, 0.01$, and 0.001 as in Zou and Hastie (2005). The information criteria which will be used to select the shrinkage tuning parameter will also be used to select one of the values of α which minimises the selection criteria. According to Zou and Zhang (2009), the LASSO part performs an automatic variable selection while the ridge regression stabilizes the solution paths and hence improves the prediction. They state that in orthogonal designs where the LASSO is shown to be optimal, the EN automatically reduces to the LASSO. However, when the correlations among the predictors become high, the EN can significantly improve the prediction accuracy of the LASSO. Also, Zou and Hastie (2005) state that the EN simultaneously does automatic variable selection and continuous shrinkage, and it can select groups of correlated variables.

2.9 Least Angle Regression Selection (LARS)

The LARS algorithm has been described in Efron et al. (2004). Let $\mathbf{X}$ is the $n \times p$ design matrix for p factors. LARS builds up estimates $\hat{\boldsymbol{\mu}} = \mathbf{X}\hat{\boldsymbol{\beta}}$ of the response where $\hat{\boldsymbol{\beta}}$ is the generalised least square estimate of the vector of coefficients. Yue (2010) summarised the LARS algorithm as follows

1. The algorithm starts at $\hat{\mu}_0 = 0$ and locate all coefficients to zero.

2. Find the variable, x_1, which is most correlated with the response.

3. Fit the model using the generalised least square estimator in the direction of x_1 (or u_1, the unit vector along x_1) until another variable, say x_2, has similar correlation with the recent residual as x_1 does.

4. At this step, the LARS estimate is updated to $\hat{\mu}_1 = \hat{\mu}_0 + \hat{\gamma}_1 u_1$, where $\hat{\gamma}_1$ can be chosen such that the recent residual $y - \hat{\mu}_1$ halves the angle between the two variables x_1 and x_2.

5. Rather than continuing along x_1, LARS progresses in the direction of u_2, the unit bisector of the two variables x_1 and x_2, until a third variable x_3 gains its way into the most correlated set.

6. The LARS estimate then will be upgraded to $\hat{\mu}_2 = \hat{\mu}_1 + \hat{\gamma}_2 u_2$, where $\hat{\gamma}_2$ is selected so that the recent residual $y - \hat{\mu}_2$ has similar angles with x_1, x_2, and x_3.

7. After that, LARS progresses along u_3, the equiangular unit vector, i.e. along the least angle direction, until a fourth variable enters, etc. LARS builds up estimates in each step by adding one variable to the model, therefore, only p steps are required for the full set, where p is the number of variables.

As we deal with mixed effect models, we can weight our data as $\mathbf{X}^* = \sqrt{\mathbf{C}}\mathbf{X}$ and $\mathbf{Y}^* = \sqrt{\mathbf{C}}\mathbf{Y}$, where $\mathbf{C} = \mathbf{V}^{-1}$, and apply the `lars` function in the `lars` package in `R`. However, in this work, we modify the Gram matrix $\mathbf{G} = \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$ to calculate the generalised least square estimator (GLS) in the `lars` function in `Matlab`.

2.10 Discussion

This chapter discussed the theoretical framework of the frequentist analysis using the PGLS and PGLS-SPD. It provided the steps of the computational algorithm for the variable selection by PGLS in Section 2.2.1 and by PGLS-SPD in Section 2.2.2. The PGLS approach utilises one shrinkage parameter for all factor effects whereas the PGLS-SPD approach utilises two shrinkage parameters as one for the subplot effect factors and the

other for whole-plot effect factors. The two shrinkage parameters in PGLS-SPD may have the same grid; however, different values for the two shrinkage parameters would be selected by the information criteria. This extension in PGLS-SPD was applied to adapt the issue of the restricted randomisation as it yields two different strata for split-plot design. The backward elimination was presented as an example of the classical model selection methods. Also, some shrinkage methods involving LASSO, ALASSO, SCAD, EN, and LARS were described.

In Chapter 3, we will summarise the results of the implementation for both the PGLS in (1.11) and the PGLS-SPD in (2.1) on the data for the wind tunnel experiment and the freeze-dried coffee experiment. Moreover, we will show how the performance of the PGLS-SPD overcomes the performance of PGLS with respect to variable selection and parameter estimation.

Chapter 3

Application of the Frequentist Methods for Response from Split-Plot Experiments

This chapter is a practical demonstration of the analyses described in Chapter 2. It provides a comparison between penalised generalised least squares (PGLS) in (1.11) and penalised generalised least squares for split-plot experiments (PGLS-SPD) in (2.1). The numerical results will be summarised for both motivating examples explained in Sections 1.2.1 and 1.2.2. The focus is on two different designs of split-plot experiments. In the first example, the wind tunnel experiment, the used design is an orthogonal (see Figure 3.1 and Figure 3.2). In the second example, the freeze-dried coffee experiment, the used design is clearly non-orthogonal (see Figure 3.3 and Figure 3.4). The graphs show the correlation coefficients for each pair of variables. The correlation that we have is the scaled inner-product between columns in the design matrix. The performance of PGLS and PGLS-SPD using these two different types of designs have been studied. Section 3.1 will represent the results from the real dataset from both experiments, while the simulation results will be summarised in Section 3.2. This chapter concludes with a short discussion and comparison of PGLS, PGLS-SPD, backward elimination, and LARS selection methods.

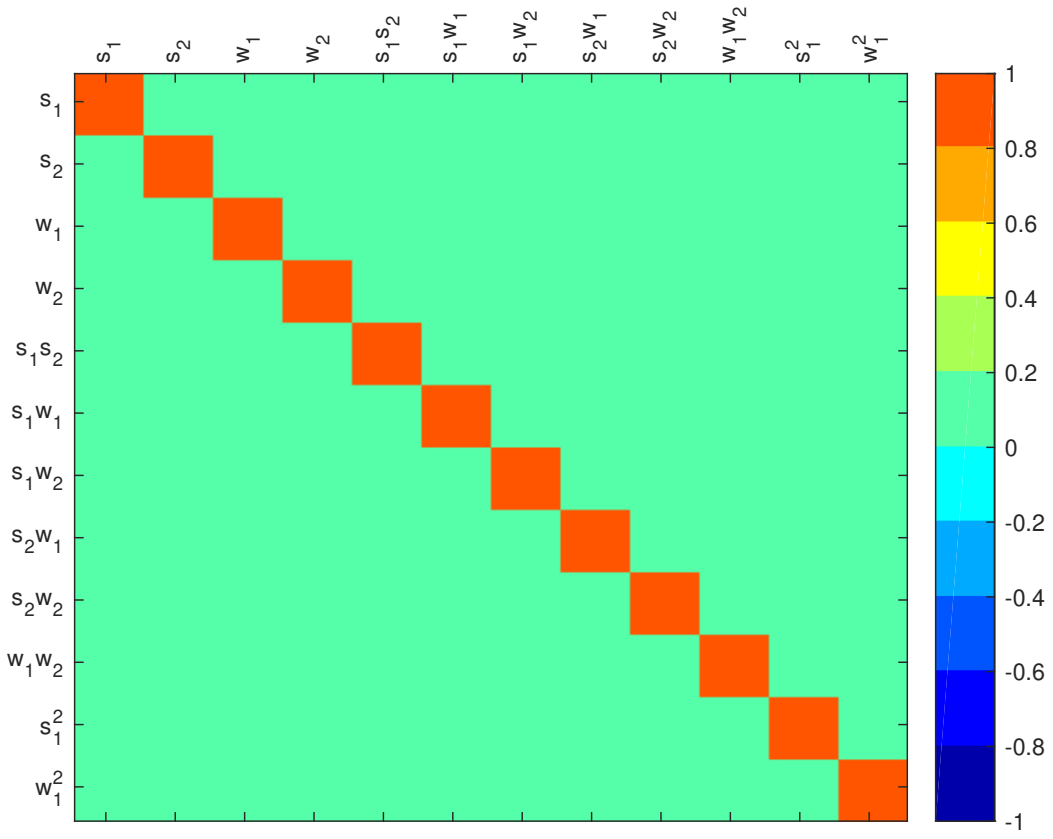


Figure 3.1: Heat map of column correlation matrix for the quadratic model in Section 3.1.1 of the wind tunnel experiment.

3.1 Practical Examples

The real datasets for the wind tunnel experiment and the freeze-dried coffee experiment have been used to apply the PGLS, PGLS-SPD, backward elimination, and LARS approaches for each experiment. The estimated coefficients were obtained as well as the standard errors. The selected tuning parameters in each experiment were also displayed. The estimated variance components using REML have been reported as well.

Using the quadratic model for the wind tunnel experiment and the full quadratic model for the freeze-dried coffee experiment in Sections 3.1.1 and 3.1.2, Figure 3.1 to Figure 3.4 have been produced. The quadratic model for the wind tunnel experiment involves four main effects (w_1, w_2, s_1, s_2), six two-factor interaction effects $(s_1s_2,\ s_1w_1,\ s_1w_2,\ s_2w_1,\ s_2w_2,\ w_1w_2)$, and two quadratic effects (w_1^2, s_1^2). Also, the full quadratic model for the freeze-dried coffee experiment involves five main effects $(w$,$s_1,\ s_2,\ s_3,\ s_4)$, 10 two-factor

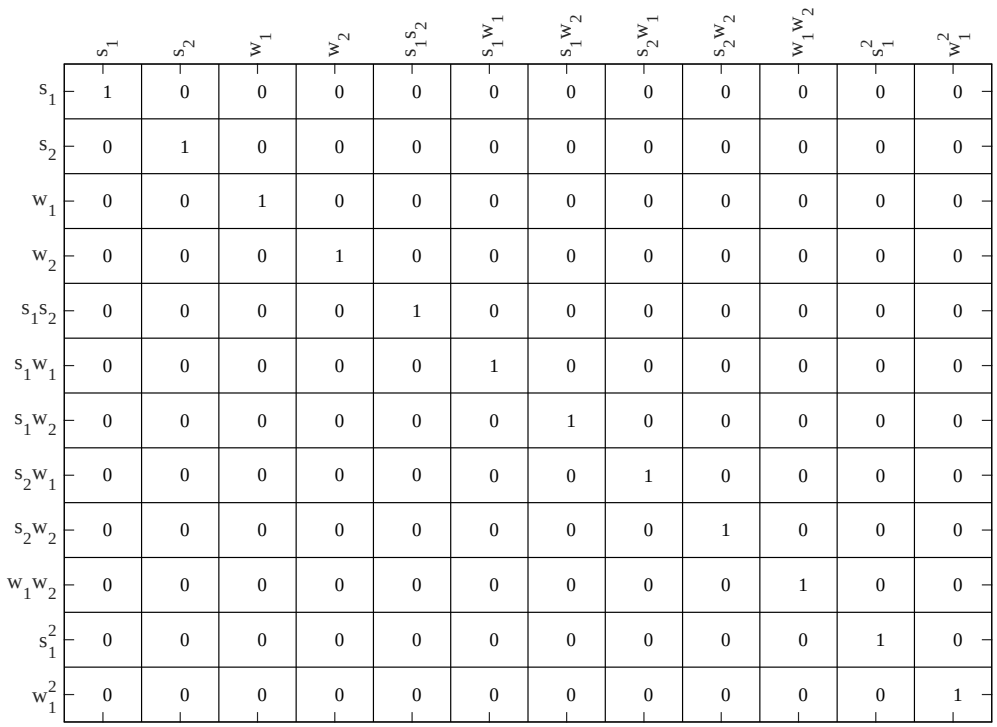


Figure 3.2: Column correlation matrix for the quadratic model in Section 3.1.1 of the wind tunnel experiment displayed by the values of the correlation coefficients.

interaction effects (ws_1, ws_2, ws_3, ws_4, s_1s_2, s_1s_3, s_1s_4, s_2s_3, s_2s_4, s_3s_4), and five quadratic effects (w^2, s_1^2, s_2^2, s_3^2, s_4^2).

3.1.1 Analysis of the Wind Tunnel Experiment

The wind tunnel experiment has been discussed in Section 1.2.1. In this section, the results of applying the PGLS, PGLS-SPD, backward elimination, and LARS for this data has been discussed. The maximal model is the quadratic model as in Section 3.1. In this design, w_1^2 and w_2^2 are fully correlated with each other. The s_1^2 and s_2^2 are also fully correlated with each other. Therefore, only one of the quadratic whole-plot effects and only one of the quadratic subplot effects could be estimated as in Section 3.1. The estimates of the 12 parameters of the quadratic model have been obtained using four different penalty functions for the PGLS and PGLS-SPD, as well as using the LARS and backward elimination as described in Sections 2.4 to 2.9.

Table 3.1 and Table 3.2 present the estimated coefficients and standard errors by two

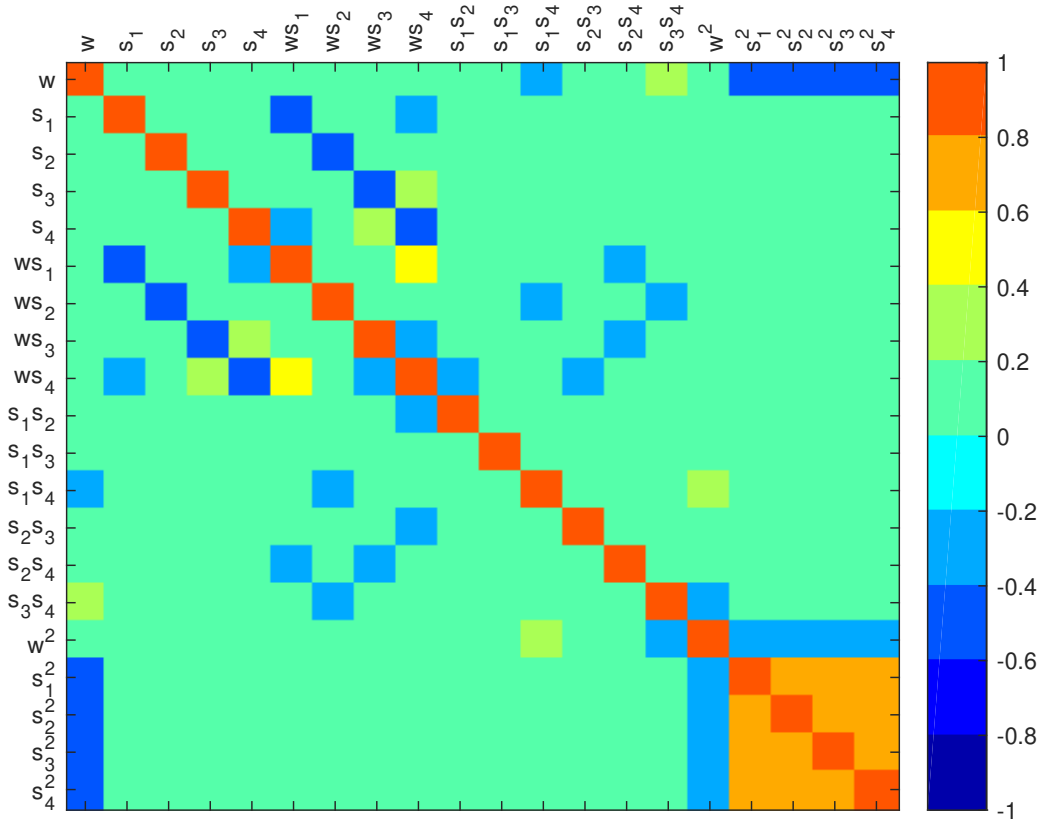


Figure 3.3: Heat map of column correlation matrix for the full quadratic model in Section 3.1.2 of the freeze-dried coffee experiment.

different shrinkage tuning parameter selection criteria, i.e. cAIC and BIC, as in Section 2.3, for y_3 displayed in Table 1.3. In Table 3.1, the model subset selection by backward elimination and LARS, are displayed for comparison. Moreover, the estimated coefficients with their $p-$values from Simpson et al. (2004) are presented. In their work, they compared the analysis for y_3 with the quadratic model using the completely randomised design and the split-plot design. We report their results from the split-plot design to compare it with our results. They estimated the parameters of the quadratic model using the generalised least squares.

Table 3.3 then presents the selected tuning parameters for the wind tunnel experiment by both PGLS and PGLS-SPD. The REML has been used to estimate the variance components from the full quadratic model. We found that both $\hat{\sigma}^2_\epsilon$ and $\hat{\sigma}^2_\gamma$ equal to 1.0000×10^{-5}. The value 1.0000×10^{-5} is the lower bound that we assumed in the

	w	s_1	s_2	s_3	s_4	ws_1	ws_2	ws_3	ws_4	s_1s_2	s_1s_3	s_1s_4	s_2s_3	s_2s_4	s_3s_4	w^2	s_1^2	s_2^2	s_3^2	s_4^2
w	1	0	0	0	0	0	0	0	0	0	0	-0.22	0	0	0.22	0	-0.50	-0.50	-0.50	-0.50
s_1	0	1	0	0	0	-0.45	0	0	-0.30	0	0	0	0	0	0	0	0	0	0	0
s_2	0	0	1	0	0	0	-0.45	0	0	0	0	0	0	0	0	0	0	0	0	0
s_3	0	0	0	1	0	0	0	-0.45	0.30	0	0	0	0	0	0	0	0	0	0	0
s_4	0	0	0	0	1	-0.30	0	0.30	-0.45	0	0	0	0	0	0	0	0	0	0	0
ws_1	0	-0.45	0	0	-0.30	1	0	0	0.40	0	0	0	0	-0.32	0	0	0	0	0	0
ws_2	0	0	-0.45	0	0	0	1	0	0	0	0	-0.32	0	0	-0.32	0	0	0	0	0
ws_3	0	0	0	-0.45	0.30	0	0	1	-0.40	0	0	0	0	-0.32	0	0	0	0	0	0
ws_4	0	-0.30	0	0.30	-0.45	0.40	0	-0.40	1	-0.32	0	0	-0.32	0	0	0	0	0	0	0
s_1s_2	0	0	0	0	0	0	0	0	-0.32	1	0	0	0	0	0	0	0	0	0	0
s_1s_3	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
s_1s_4	-0.22	0	0	0	0	0	-0.32	0	0	0	0	1	0	0	0	0.39	0	0	0	0
s_2s_3	0	0	0	0	0	0	0	0	-0.32	0	0	0	1	0	0	0	0	0	0	0
s_2s_4	0	0	0	0	0	-0.32	0	-0.32	0	0	0	0	0	1	0	0	0	0	0	0
s_3s_4	0.22	0	0	0	0	0	-0.32	0	0	0	0	0	0	0	1	-0.39	0	0	0	0
w^2	0	0	0	0	0	0	0	0	0	0	0	0.39	0	0	-0.39	1	-0.29	-0.29	-0.29	-0.29
s_1^2	-0.50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.29	1	0.72	0.72	0.72
s_2^2	-0.50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.29	0.72	1	0.72	0.72
s_3^2	-0.50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.29	0.72	0.72	1	0.72
s_4^2	-0.50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	-0.29	0.72	0.72	0.72	1

Figure 3.4: Column correlation matrix for the full quadratic model in Section 3.1.2 of the freeze-dried coffee experiment displayed by the values of the correlation coefficients.

function `fmincon` of Matlab. Table 3.1, Table 3.2 and Table 3.3, indicate that the selected tuning parameters by cAIC and BIC in both PGLS and PGLS-SPD have the same values. This resulted in providing the same final model with nine variables considered to be significant for both the PGLS and the PGLS-SPD and by all penalty functions in this experiment due to the similarity in the amount of shrinkage. Furthermore, the estimated coefficients as well as the standard errors are identical for all methods. This is because of the small variance of the response and of the experimental design properties which was used for the wind tunnel experiment. The design is orthogonal, so the model parameters can be estimated independently.

A comparison between the penalised methods, backward elimination and LARS showed that backward yielded four main variables while the LARS yielded seven variables considered to be significant. The extra two variables which were included by PGLS and PGLS-SPD are s_1w_1 and s_1w_2. We note that these two variables have a little effect with a standard error equal to a half of their estimated size, which makes the standard errors

for these estimates quite large compared to their size effects. The PGLS and PGLS-SPD prefer to keep these two variables in the final model. By increasing the values of the tuning parameters, extra shrinkage amount has been applied to other variables (i.e. s_1, s_2, w_1, and s_1^2). However, the results are reported using grids from 0 to 3 as (0, 0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2, 2.2, 2.4, 2.6, 2.8, 3) for λ, λ_s, and λ_w.

	Method	β_{s_1}	β_{s_2}	β_{w_1}	β_{w_2}	$\beta_{s_1s_2}$	$\beta_{s_1w_1}$	$\beta_{s_1w_2}$	$\beta_{s_2w_1}$	$\beta_{s_2w_2}$	$\beta_{w_1w_2}$	$\beta_{s_1^2}$	$\beta_{w_1^2}$
	LASSO	-0.0117	-0.0049	0.0086	0.0090	-0.0005	0.0004	0.0004	0	0.0007	0	0.0012	0
		(0.0002)	(0.0002)	(0.0005)	(0.0005)	(0.0002)	(0.0002)	(0.0002))	(-)	(0.0002)	(-)	(0.0005)	(-)
	$\text{ALASSO}_{0.5}$	-0.0117	-0.0049	0.0086	0.0090	-0.0005	0.0004	0.0004	0	0.0007	0	0.0012	0
		(0.0002)	(0.0002)	(0.0005)	(0.0005)	(0.0002)	(0.0002)	(0.0002)	(-)	(0.0002)	(-)	(0.0005)	(-)
	SCAD	-0.0117	-0.0049	0.0086	0.0090	-0.0005	0.0004	0.0004	0	0.0007	0	0.0012	0
		(0.0002)	(0.0002)	(0.0005)	(0.0005)	(0.0002)	(0.0002)	(0.0002)	(-)	(0.0002)	(-)	(0.0005)	(-)
cAIC	EN	-0.0117	-0.0049	0.0086	0.0090	-0.0005	0.0004	0.0004	0	0.0007	0	0.0012	0
		(0.0002)	(0.0002)	(0.0005)	(0.0005)	(0.0002)	(0.0002)	(0.0002)	(-)	(0.0002)	(-)	(0.0005)	(-)
	LASSO	-0.0117	-0.0049	0.0086	0.0090	-0.0005	0.0004	0.0004	0	0.0007	0	0.0012	0
		(0.0002)	(0.0002)	(0.0005)	(0.0005)	(0.0002)	(0.0002)	(0.0002)	(-)	(0.0002)	(-)	(0.0005)	(-)
	$\text{ALASSO}_{0.5}$	-0.0117	-0.0049	0.0086	0.0090	-0.0005	0.0004	0.0004	0	0.0007	0	0.0012	0
		(0.0002)	(0.0002)	(0.0005)	(0.0005)	(0.0002)	(0.0002)	(0.0002)	(-)	(0.0002)	(-)	(0.0005)	(-)
	SCAD	-0.0117	-0.0049	0.0086	0.0090	-0.0005	0.0004	0.0004	0	0.0007	0	0.0012	0
		(0.0002)	(0.0002)	(0.0005)	(0.0005)	(0.0002)	(0.0002)	(0.0002)	(-)	(0.0002)	(-)	(0.0005)	(-)
BIC	EN	-0.0117	-0.0049	0.0086	0.0090	-0.0005	0.0004	0.0004	0	0.0007	0	0.0012	0
		(0.0002)	(0.0002)	(0.0005)	(0.0005)	(0.0002)	(0.0002)	(0.0002)	(-)	(0.0002)	(-)	(0.0005)	(-)
	Backward	-0.0117	-0.0049	0.0085	0.0089	0	0	0	0	0	0	0	0
		(0.0002)	(0.0002)	(0.0006)	(0.0006)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
	LARS	-0.0117	-0.0049	0.0086	0.0090	-0.0005	0	0	0	0.0007	0	0.0012	0
		(0.0002)	(0.0002)	(0.0005)	(0.0005)	(0.0002)	(-)	(-)	(-)	(0.0002)	(-)	(0.0005)	(-)
	Original paper	-0.0117	-0.0049	0.0086	0.0089	-0.0006	0	0	0	0.0008	0	0.0006	0
		(0)	(0)	(0)	(0)	(0.0360)	(-)	(-)	(-)	(0.0070)	(-)	(0.0480)	(-)

Table 3.1: Estimated coefficients and standard errors (in parentheses) for the wind tunnel experiment for y_3 by PGLS. $\text{ALASSO}_{0.5}$ is the ALASSO with $\psi = 0.5$. The last row is the estimated coefficients and $p-$values (in parentheses) from Simpson et al. (2004).

Method		β_{s_1}	β_{s_2}	β_{w_1}	β_{w_2}	$\beta_{s_1s_2}$	$\beta_{s_1w_1}$	$\beta_{s_1w_2}$	$\beta_{s_2w_1}$	$\beta_{s_2w_2}$	$\beta_{w_1w_2}$	$\beta_{s_1^2}$	$\beta_{w_1^2}$
cAIC	LASSO	-0.0117	-0.0049	0.0086	0.0090	-0.0005	0.0004	0.0004	0	0.0007	0	0.0012	0
		(0.0002)	(0.0002)	(0.0005)	(0.0005)	(0.0002)	(0.0002)	(0.0002)	(-)	(0.0002)	(-)	(0.0005)	(-)
	$ALASSO_{0.5}$	-0.0117	-0.0049	0.0086	0.0090	-0.0005	0.0004	0.0004	0	0.0007	0	0.0012	0
		(0.0002)	(0.0002)	(0.0005)	(0.0005)	(0.0002)	(0.0002)	(0.0002)	(-)	(0.0002)	(-)	(0.0005)	(-)
	SCAD	-0.0117	-0.0049	0.0086	0.0090	-0.0005	0.0004	0.0004	0	0.0007	0	0.0012	0
		(0.0002)	(0.0002)	(0.0005)	(0.0005)	(0.0002)	(0.0002)	(0.0002)	(-)	(0.0002)	(-)	(0.0005)	(-)
	EN	-0.0117	-0.0049	0.0086	0.0090	-0.0005	0.0004	0.0004	0	0.0007	0	0.0012	0
		(0.0002)	(0.0002)	(0.0005)	(0.0005)	(0.0002)	(0.0002)	(0.0002)	(-)	(0.0002)	(-)	(0.0005)	(-)
BIC	LASSO	-0.0117	-0.0049	0.0086	0.0090	-0.0005	0.0004	0.0004	0	0.0007	0	0.0012	0
		(0.0002)	(0.0002)	(0.0005)	(0.0005)	(0.0002)	(0.0002)	(0.0002)	(-)	(0.0002)	(-)	(0.0005)	(-)
	$ALASSO_{0.5}$	-0.0117	-0.0049	0.0086	0.0090	-0.0005	0.0004	0.0004	0	0.0007	0	0.0012	0
		(0.0002)	(0.0002)	(0.0005)	(0.0005)	(0.0002)	(0.0002)	(0.0002)	(-)	(0.0002)	(-)	(0.0005)	(-)
	SCAD	-0.0117	-0.0049	0.0086	0.0090	-0.0005	0.0004	0.0004	0	0.0007	0	0.0012	0
		(0.0002)	(0.0002)	(0.0005)	(0.0005)	(0.0002)	(0.0002)	(0.0002)	(-)	(0.0002)	(-)	(0.0005)	(-)
	EN	-0.0117	-0.0049	0.0086	0.0090	-0.0005	0.0004	0.0004	0	0.0007	0	0.0012	0
		(0.0002)	(0.0002)	(0.0005)	(0.0005)	(0.0002)	(0.0002)	(0.0002)	(-)	(0.0002)	(-)	(0.0005)	(-)

Table 3.2: Estimated coefficients and standard errors (in parentheses) for the wind tunnel experiment for y_3 by PGLS-SPD. $ALASSO_{0.5}$ is the ALASSO with $\psi = 0.5$.

Information criteria	cAIC			BIC		
Method	λ	λ_s	λ_w	λ	λ_s	λ_w
LASSO	3	3	3	3	3	3
$\text{ALASSO}_{0.5}$	0	0	0	0	0	0
SCAD	2.8	2.8	2.8	2.8	2.8	2.8
EN	3	3	3	3	3	3

Table 3.3: The selected tuning parameters λ from PGLS and $(\lambda_s,\ \lambda_w)$ from PGLS-SPD for the wind tunnel experiment.

3.1.2 Analysis of the Freeze-Dried Coffee experiment

The freeze-dried coffee experiment was introduced in Section 1.2.2. In this section, the results of applying the PGLS, PGLS-SPD, backward elimination, and LARS for this data are summarised. The maximal model is the full quadratic model as in Section 3.1. The estimates of the 20 parameters of the full quadratic model have been obtained using four different penalty functions for the PGLS and PGLS-SPD, as well as using LARS and the backward elimination as described in Sections 2.4 to 2.9.

Table 3.4 and Table 3.5 present the estimated coefficients and their standard errors using two different shrinkage tuning parameter selection criteria, i.e. cAIC and BIC as in Section 2.3, for y_1 displayed in Table 1.5. The model subset selection by backward elimination and the LARS, are also displayed for comparison. Moreover, the estimated coefficients with their $p-$values from Gilmour and Goos (2009) are presented to compare with our results. In their work, they carried out REML-GLS analysis for y_1 with the full model using the `R lme` function.

Table 3.6 presents the selected tuning parameters for the freeze-dried coffee experiment by both PGLS and PGLS-SPD. REML has been used to estimate the variance components for the response, y_1, as $\hat{\sigma}^2_\epsilon$ is 5.8457 and $\hat{\sigma}^2_\gamma$ is 1.0000×10^{-5}. The experimental design used in the freeze-dried coffee experiment is non-orthogonal as shown in Figure 3.3 and Figure 3.4, and has four variables varied within the whole plots. Table 3.6 shows that the selected tuning parameters by cAIC is larger than the selected tuning parameters by BIC. This resulted in final models with an extra one variable by BIC (e.g. the subplot factor ws_4 in all penalty functions) more than the final models by cAIC in both PGLS and PGLS-SPD by $\text{ALASSO}_{0.5}$ and SCAD. Moreover, the subplot factor ws_4 has been included in the final model by LASSO and EN in which the cAIC

	cAIC				BIC						
	LASSO	ALASSO$_{0.5}$	SCAD	EN	LASSO	ALASSO$_{0.5}$	SCAD	EN	Backward	LARS	Original paper
β_w	-2.1874	-2.3123	-2.21561	-2.3060	-2.3823	-2.4468	-2.3735	-2.3735	-2.8639	-2.5771	-2.5021
	(0.4951)	(0.5096)	(0.5417)	(0.5291)	(0.5291)	(0.5096)	(0.5728)	(0.5290)	(0.6483)	(0.6261)	(0.0439)
β_{s_1}	10.1782	10.7517	10.8903	10.2342	10.2516	10.5002	10.6284	10.2096	10.8903	10.8903	10.4748
	(0.5875)	(0.6202)	(0.6228)	(0.6014)	(0.5898)	(0.6140)	(0.6175)	(0.5898)	(0.6834)	(0.6261)	(0)
β_{s_2}	-4.4176	-4.7988	-5.0763	-4.6146	-4.47749	-4.9667	-5.1815	-4.7427	-5.4188	-5.1321	-5.1321
	(0.5804)	(0.6496)	(0.6793)	(0.6139)	(0.6119)	(0.6496)	(0.6761)	(0.6115)	(0.6834)	(0.7229)	(0.0003)
β_{s_3}	0	0	0	0	0	0	0	0	0	0	0
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
β_{s_4}	2.5214	3.1381	3.0495	2.6054	2.7556	3.0018	3.0018	2.7410	3.3947	3.3947	3.0201
	(0.5234)	(0.5633)	(0.5865)	(0.5653)	(0.5832)	(0.6152)	(0.6398)	(0.5831)	(0.6834)	(0.6261)	(0.0070)
β_{ws_1}	0	0	0	0	0	0	0	0	0	0	0
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
β_{ws_2}	-3.5703	-4.1675	-4.6797	-3.8356	-4.0574	-4.3543	-4.6928	-4.0105	-5.4049	-4.5446	-4.5446
	(0.7205)	(0.8508)	(0.9037)	(0.8113)	(0.8412)	(0.8508)	(0.9745)	(0.8400)	(0.9169)	(0.0844)	(0.0047)
β_{ws_3}	0	0	0	0	-0.9967	0	0	-0.9983	0	0	0
	(-)	(-)	(-)	(-)	(0.5936)	(-)	(-)	(0.5940)	(-)	(-)	(-)
β_{ws_4}	-1.4509	0	0	-1.5558	-1.9974	-1.4401	-1.1788	-2.0161	0	0	0
	(0.5260)	(-)	(-)	(0.6416)	(0.7478)	(0.5835)	(0.6284)	(0.7493)	(-)	(-)	(-)
$\beta_{s_1s_2}$	1.1283	1.5138	1.3449	1.2439	1.2109	1.3493	1.4274	1.2032	1.9295	1.9295	0
	(0.4441)	(0.4315)	(0.4777)	(0.5118)	(0.5307)	(0.4315)	(0.5602)	(0.5301)	(0.7249)	(0.6640)	(-)
$\beta_{s_1s_3}$	0	0	0	0	0	0	0	0	0	0	0
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
$\beta_{s_1s_4}$	1.4638	1.2106	1.0156	1.5242	1.5287	1.4105	1.3778	1.5214	0	1.5775	0
	(0.4880)	(0.3578)	(0.3928)	(0.5472)	(0.5646)	(0.5188)	(0.5811)	(0.5647)	(-)	(0.7341)	(-)
$\beta_{s_2s_3}$	0	0	0	0	0	0	0	0	0	0	0
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
$\beta_{s_2s_4}$	2.4477	2.5669	2.4114	2.5817	2.4177	2.7314	2.8746	2.4090	2.8861	2.8861	2.5554
	(0.5329)	(0.5592)	(0.5922)	(0.5479)	(0.5737)	(0.5592)	(0.6155)	(0.5735)	(0.7249)	(0.6640)	(0.0132)
$\beta_{s_3s_4}$	2.9522	2.9656	2.6535	3.0369	3.0884	3.104	3.1984	3.1033	3.0911	3.2345	3.0856
	(0.5677)	(0.6106)	(0.6432)	(0.6085)	(0.6109)	(0.6106)	(0.6107)	(0.6117)	(0.7249)	(0.7341)	(0.0075)
β_{w^2}	0	0	0	0	0	0	0	0	0	0	0
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
$\beta_{s_1^2}$	0	0	0	0	0	0	0	0	0	0	0
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
$\beta_{s_2^2}$	0	0	0	0	0	0	0	0	0	0	0
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
$\beta_{s_3^2}$	0	0	0	0	0	0	0	0	0	0	0
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
$\beta_{s_4^2}$	0	0	0	0	0	0	0	0	0	0	0
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)

Table 3.4: Estimated coefficients and standard errors (in parentheses) for the freeze-dried coffee experiment for y_1 by PGLS. ALASSO$_{0.5}$ is the ALASSO with $\psi = 0.5$. The last column is the estimated coefficients and $p-$values (in parentheses) from Gilmour and Goos (2009).

was used to select the tuning parameter. This was the same for both the PGLS and PGLS-SPD as the selected tuning parameters λ and λ_s were the same as in Table 3.6. The difference between PGLS and PGLS-SPD appears in the subplot factor ws_3 as it was included in the PGLS to the final model by LASSO and EN in which the BIC used to select the tuning parameters whereas PGLS-SPD removed this factor for all functions and all selection criteria. A comparison between the penalised methods, backward elimination and LARS showed that the ALASSO$_{0.5}$ and the SCAD in which the cAIC was used to select the tuning parameters have almost similar performance to the backward elimination and LARS as they resulted in a final model with eight and nine variables respectively considered to be significant by both PGLS and PGLS-SPD.

	cAIC				BIC			
	LASSO	$\text{ALASSO}_{0.5}$	SCAD	EN	LASSO	$\text{ALASSO}_{0.5}$	SCAD	EN
β_w	-2.2524	-2.5541	-2.23606	-2.3060	-2.3823	-2.5726	-2.3765	-2.4407
	(0.5110)	(0.5902)	(0.5430)	(0.5291)	(0.5291)	(0.5888)	(0.5366)	(0.5438)
β_{s_1}	10.1782	10.7517	10.8903	10.2342	10.2516	10.5002	10.6284	10.2096
	(0.5875)	(0.6202)	(0.6261)	(0.6014)	(0.5898)	(0.6140)	(0.6198)	(0.5898)
β_{s_2}	-4.4176	-4.8016	-5.0726	-4.6146	-4.47749	-4.9671	-5.1796	-4.7427
	(0.5804)	(0.6496)	(0.6747)	(0.6139)	(0.6119)	(0.6608)	(0.6928)	(0.6115)
β_{s_3}	0	0	0	0	0	0	0	0
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
β_{s_4}	2.5214	3.1381	3.0495	2.6054	2.7556	2.7783	3.0018	2.7410
	(0.5234)	(0.5633)	(0.5508)	(0.5653)	(0.5832)	(0.5885)	(0.6398)	(0.5831)
β_{ws_1}	0	0	0	0	0	0	0	0
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
β_{ws_2}	-3.5703	-4.1755	-4.6704	-3.8356	-4.0574	-4.3556	-4.6873	-4.0105
	(0.7204)	(0.8504)	(0.9037)	(0.8113)	(0.8412)	(0.9261)	(1.0219)	(0.8399)
β_{ws_3}	0	0	0	0	0	0	0	0
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
β_{ws_4}	-1.4509	0	0	-1.5558	-1.9974	-1.4401	-1.1788	-2.0161
	(0.5260)	(-)	(-)	(0.6416)	(0.7478)	(0.5835)	(0.6390)	(0.7493)
$\beta_{s_1s_2}$	1.1283	1.5138	1.3450	1.2439	1.2109	1.3493	1.4274	1.2032
	(0.4441)	(0.4315)	(0.4356)	(0.5118)	(0.5307)	(0.4886)	(0.5596)	(0.5301)
$\beta_{s_1s_3}$	0	0	0	0	0	0	0	0
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
$\beta_{s_1s_4}$	1.4638	1.2106	1.0156	1.5242	1.5287	1.4105	1.3778	1.5214
	(0.4880)	(0.3578)	(0.3928)	(0.5472)	(0.5646)	(0.5188)	(0.5811)	(0.5647)
$\beta_{s_2s_3}$	0	0	0	0	0	0	0	0
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
$\beta_{s_2s_4}$	2.4477	2.5669	2.4114	2.5817	2.4177	2.7314	2.8746	2.4090
	(0.5329)	(0.5592)	(0.5365)	(0.5679)	(0.5737)	(0.5942)	(0.6370)	(0.5735)
$\beta_{s_3s_4}$	2.9585	3.0271	2.7153	3.0369	3.0884	3.1362	3.1487	3.1033
	(0.5688)	(0.6174)	(0.5841)	(0.6085)	(0.6109)	(0.6528)	(0.7032)	(0.6117)
β_{w^2}	0	0	0	0	0	0	0	0
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
$\beta_{s_1^2}$	0	0	0	0	0	0	0	0
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
$\beta_{s_2^2}$	0	0	0	0	0	0	0	0
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
$\beta_{s_3^2}$	0	0	0	0	0	0	0	0
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)
$\beta_{s_4^2}$	0	0	0	0	0	0	0	0
	(-)	(-)	(-)	(-)	(-)	(-)	(-)	(-)

Table 3.5: Estimated coefficients and standard errors (in parentheses) for the freeze-dried coffee experiment for y_1 by PGLS-SPD. $\text{ALASSO}_{0.5}$ is the ALASSO with $\psi = 0.5$.

Information criteria	cAIC			BIC		
Method	λ	λ_s	λ_w	λ	λ_s	λ_w
LASSO	1.2	1.2	1.0	0.6	0.6	0.6
$\text{ALASSO}_{0.5}$	2.8	2.8	0.6	1.2	1.4	0.2
SCAD	1.6	1.6	1.2	0.8	0.8	1.2
EN	0.6	0.8	0.8	0.6	0.6	0.4

Table 3.6: The selected tuning parameters λ from PGLS and (λ_s, λ_w) from PGLS-SPD for the freeze-dried coffee experiment.

3.2 Simulation Study

To examine our methods, we need to run simulation studies in order to find out how the resulting model will be compared to the true model which used in the simulation. In the simulation, we set $\sigma_\epsilon^2 + \sigma_\gamma^2 = 10$, grids for λ, λ_s, and λ_w from 0 to 3 as (0, 0.2, 0.4, 0.6, 0.8, 1, 1.2, 1.4, 1.6, 1.8, 2, 2.2, 2.4, 2.6, 2.8, 3), and the variance components ratio to two different levels, $\eta = 1$ and 10. Similar values for η have been used for the analysis of data from many blocked and split-plot experiments (see, for instance, Letsinger et al. (1996) and Gilmour and Trinca (2000)). We generate 1000 datasets using the design structure from both motivating experiments in Table 1.3 for the wind tunnel design and Table 1.5 for the freeze-dried coffee design, and given the assumed true model as given in Sections 3.2.1 and 3.2.2 respectively. We compare the performance of the LASSO, $\text{ALASSO}_{0.5}$, SCAD, EN, backward elimination, and LARS by PGLS and PGLS-SPD. We focus on the properties of the estimated models by investigating the following properties:

1. consistency in variable selection (frequency in selecting the active/ non-active variable), and

2. prediction performance.

For point 1, at 5% significant level, we report Type I error rate (an effect that is truly not significant but the corresponding procedure estimate indicates that it is significant). We also report Type II error rate (an effect that is truly present but the corresponding procedure estimate indicates that it is not significant).

For point 2, following Fan and Li (2001) and Ibrahim et al. (2011), prediction accuracy

is measured by computing the mean-squared error for each penalised estimate $\hat{\boldsymbol{\beta}}_\lambda$ as,

$$\mathbf{ME}(\hat{\boldsymbol{\beta}}_\lambda) = (\mathbf{X}\hat{\boldsymbol{\beta}}_\lambda - \mathbf{X}\boldsymbol{\beta})'(\mathbf{X}\hat{\boldsymbol{\beta}}_\lambda - \mathbf{X}\boldsymbol{\beta}).$$

The relative model error (**RME**) is the ratio of the model error of the penalised estimates to the model error for the GLS estimates of the fixed effects,

$$\mathbf{RME} = \frac{\mathbf{ME}(\hat{\boldsymbol{\beta}}_\lambda)}{\mathbf{ME}(\hat{\boldsymbol{\beta}}_{GLS})},$$

where $\hat{\boldsymbol{\beta}}_{GLS}$ in (1.5) is the generalised least squares estimator of $\boldsymbol{\beta}$. The median of the relative model error (**MRME**) over 1000 simulated data sets were reported for each example. **MRME** values greater than one indicate that the PGLS and PGLS-SPD estimates perform worse than the GLS estimates, values near to one indicate that the PGLS and PGLS-SPD estimates performs in a similar way to the GLS estimates, values less than one indicate that the PGLS and PGLS-SPD estimates performs better than the GLS estimates.

3.2.1 Simulation Study Using the Design of the Wind Tunnel Experiment

A simulation study was performed to examine the performance of the PGLS and PGLS-SPD estimates and compare them to the traditional model selection methods. Using the design of the wind tunnel experiment from Table 1.3, the response variable was generated given the true model

$$\mathbb{E}(\mathbf{Y}) = 4w_1 + 2s_2 - 4w_1w_2 + 2w_1s_2 + 6w_1^2 + 4s_1^2.$$

In this experiment, six active variables ($w_1, s_2, w_1w_2, w_1s_2, w_1^2$ and s_1^2) and six non-active variables ($s_1, w_2, s_1s_2, s_1w_1, s_1w_2$ and s_2w_2) were assumed. We assumed this model as we would like to check a model with variety of factor types. Figure 3.5, Figure 3.6 and Figure 3.7 present the performance of PGLS and PGLS-SPD at $\eta = 1$ whereas

Figure 3.8, Figure 3.9 and Figure 3.10 present the performance of PGLS and PGLS-SPD at $\eta = 10$, by cAIC and BIC for all functions in Sections 2.4 to 2.9. The left side of the figures displays the Type II error rate while the right side presents the Type I error rate. For details about Type I and II error rates, see Table A.1 to Table A.4 in Appendix A.

With respect to the Type I error rate at both $\eta = 1$ and $\eta = 10$, Figure 3.5 to Figure 3.10 along with Table A.1 and Table A.2 in Appendix A show that backward elimination and LARS achieved the lowest Type I error rate compared to PGLS and PGLS-SPD though backward elimination at $\eta = 1$ had the best performance. Comparing PGLS to PGLS-SPD revealed that PGLS-SPD yielded a lower Type I error rate than PGLS with respect to the subplot factors. The function $\text{ALASSO}_{0.5}$ had the lowest Type I error rate in both PGLS and PGLS-SPD.

We can explain the advantages of the PGLS-SPD in reducing the Type I error rate for the subplot factors over the PGLS from Figure 3.11 and Figure 3.12. Figure 3.11 presents the frequency of selected the shrinkage parameters λ for the PGLS, and λ_s and λ_w for PGLS-SPD at $\eta = 1$ by $\text{ALASSO}_{0.5}$. The first row shows the frequency of selected shrinkage parameters by the selection criterion cAIC. Also, the second row shows the frequency of selected shrinkage parameters by the selection criterion BIC. The λ_s which was chosen by BIC (Figure 3.11 (e)) selects a variety of values of λ (small and large) values, and rarely selects zero. This indicates that the non significant subplot factors are more likely to be penalised more than the whole-plot factors. This results in a lower Type I error rate for subplot factors by PGLS-SPD. The cAIC in Figure 3.11 (b) selected small values of λ_s and at some models, selected zero. This makes the BIC penalised the subplot factors more than cAIC and resulted in Type I error rates by BIC which are often lower than those by cAIC.

With respect to λ_w, it must be noted that both cAIC and BIC chose zero or very close to zero values for λ_w resulting in not penalising the whole-plot effects. This increased the Type I error rate for the non significant whole-plot factor w_2. Having one shrinkage parameter λ which was chosen by cAIC (Figure 3.11 (a)) and by BIC (Figure 3.11 (d)) resulted in having zero values of λ on some models which increased the Type I error rate for both subplot and whole-plot factors. We conclude that by introducing two shrinkage

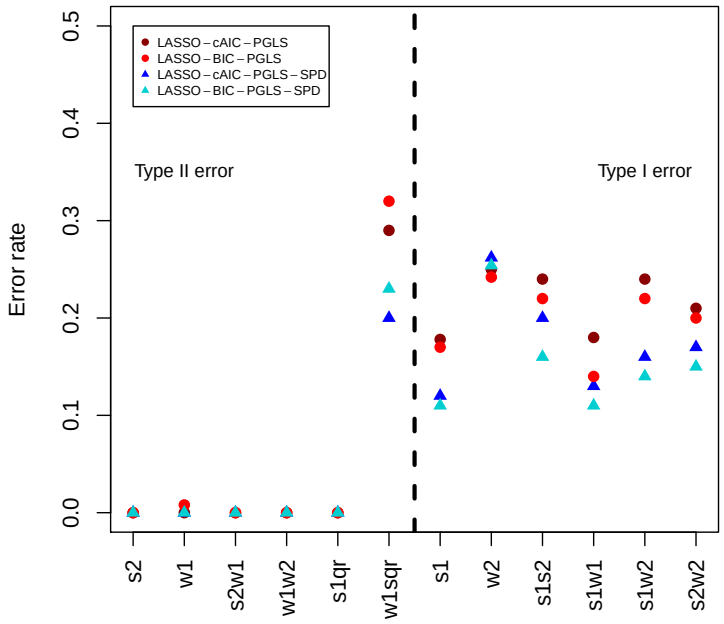
parameters λ_s and λ_w in the PGLS-SPD, the PGLS-SPD resulted in a considerable reduction in the Type I error rate compared to the Type I error rate by PGLS.

A similar analysis can be provided from Figure 3.12 in which $\eta = 10$. The figure confirms the advantage of the PGLS-SPD at $\eta = 10$ as discussed above. For example, λ_s which was chosen by cAIC (Figure 3.12 (b)) and by BIC (Figure 3.12 (e)) tends to be assigned larger values than the λ which was chosen by cAIC (Figure 3.12 (a)) and by BIC (Figure 3.12 (d)), this also resulted in a reduction of the Type I error rate for subplot factors. Furthermore, we note the λ_w which was chosen by cAIC (Figure 3.12 (c)) and by BIC (Figure 3.12 (f)) does not encourage to the penalisation of the whole-plot factors as this results in a high Type I error rate for w_2.

Regarding the Type II error rate, at $\eta = 1$ and $\eta = 10$ in Figure 3.5 to Figure 3.10 along with Table A.3 and Table A.4 in Appendix A, it can be seen that apart from w_1^2, all factors have been identified as active terms by all methods. Focusing on w_1^2, PGLS and PGLS-SPD resulted in a lower Type II error rate than with backward elimination and LARS. In a screening experiments, detecting active variables is very important for investigator. Apart from the quadratic whole-plot effect factor, the LARS and PGLS-SPD provide good alternatives to detect the active variables.

With respect to the Median Relative Model Error (MRME) discussed in Section 3.2, it can be seen from Figure 3.13 that the ALASSO$_{0.5}$ at Figure 3.13 (a) $\eta = 1$ followed by the ALASSO$_{0.5}$ at Figure 3.13 (b) $\eta = 10$ achieved the lowest MRME among the penalised methods. This indicates that the estimated coefficients with ALASSO$_{0.5}$ are better at $\eta = 1$ than the estimated coefficients with GLS, and are similar at $\eta = 10$ to the estimates with GLS. Similar to the ALASSO$_{0.5}$, we find the backward and LARS in Figure 3.13 (c) at $\eta = 10$ has similar estimates to the GLS estimator. The LASSO and EN at Figure 3.13 (a) $\eta = 1$ and Figure 3.13 (b) $\eta = 10$ as well as the SCAD at Figure 3.13 (a) $\eta = 1$ had close estimates to GLS estimator though the lowest MRME for them was at Figure 3.13 (a) $\eta = 1$. With respect to SCAD, PGLS-SPD using BIC had the lowest MRME for this function.

(a) LASSO



(b) $\text{ALASSO}_{0.5}$

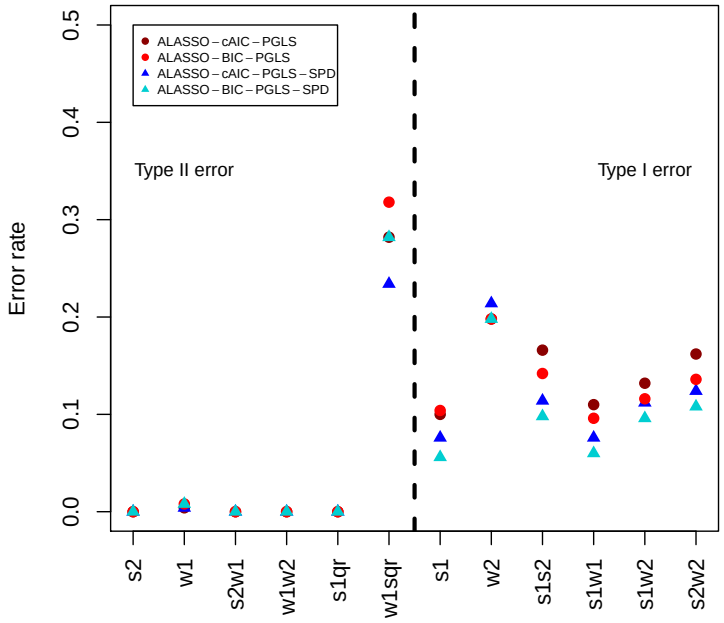
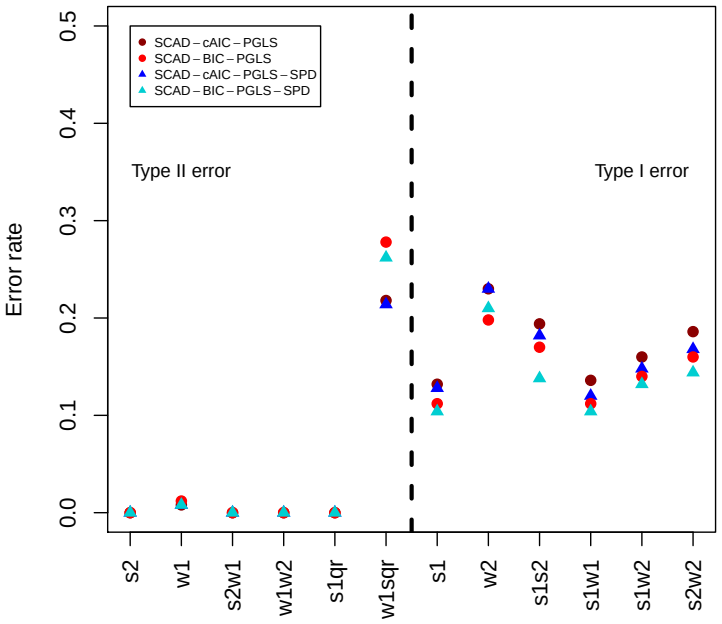


Figure 3.5: Type I and II error rates for the wind tunnel design by LASSO and $\text{ALASSO}_{0.5}$ at $\eta = 1$.

(c) SCAD



(d) EN

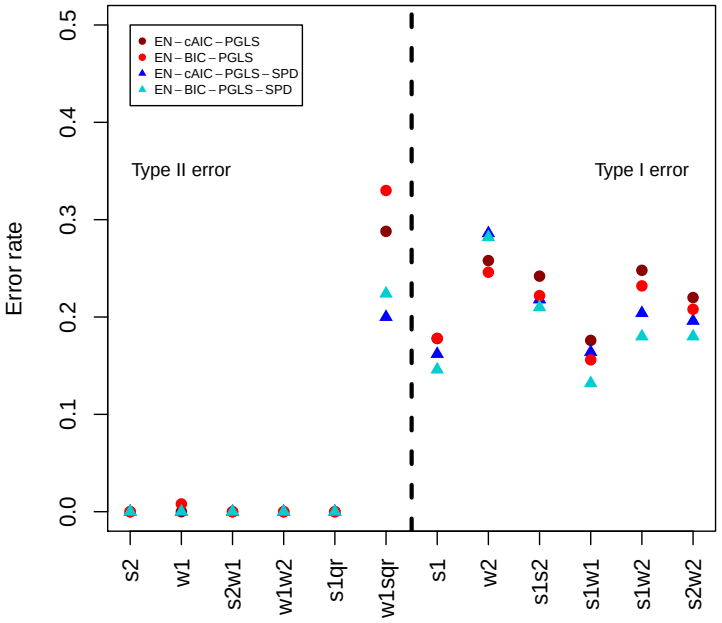


Figure 3.6: Type I and II error rates for the wind tunnel design by SCAD and EN at $\eta = 1$.

(e) Backward and LARS

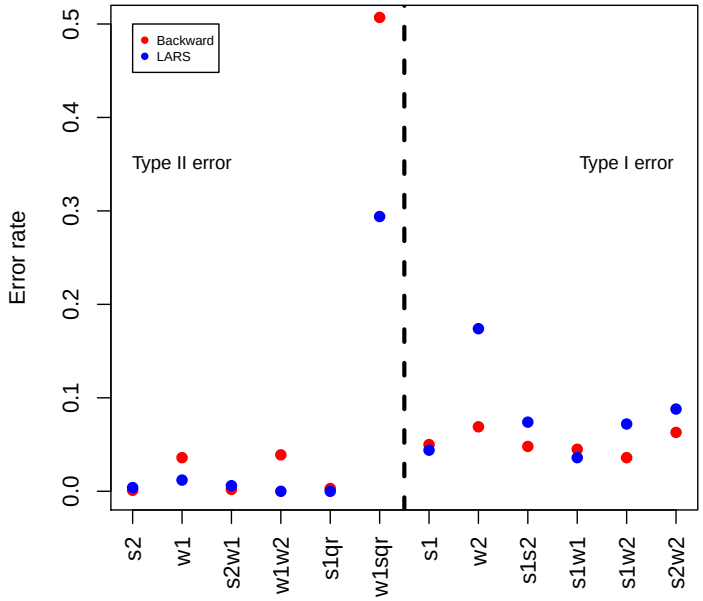
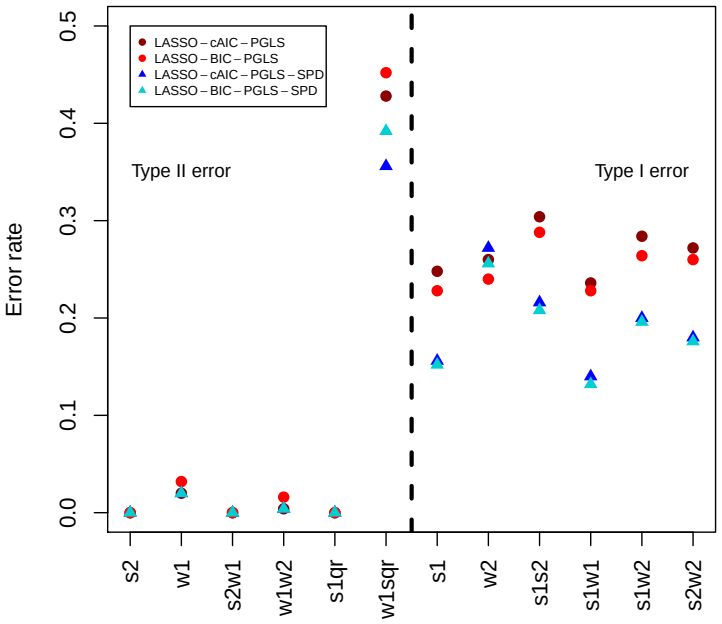


Figure 3.7: Type I and II error rates for the wind tunnel design by Backward and LARS at $\eta = 1$.

(a) LASSO



(b) $\text{ALASSO}_{0.5}$

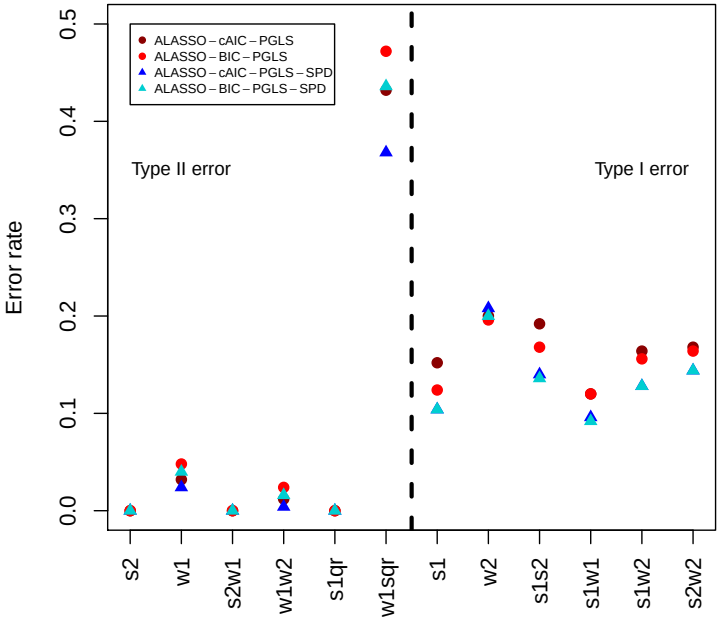
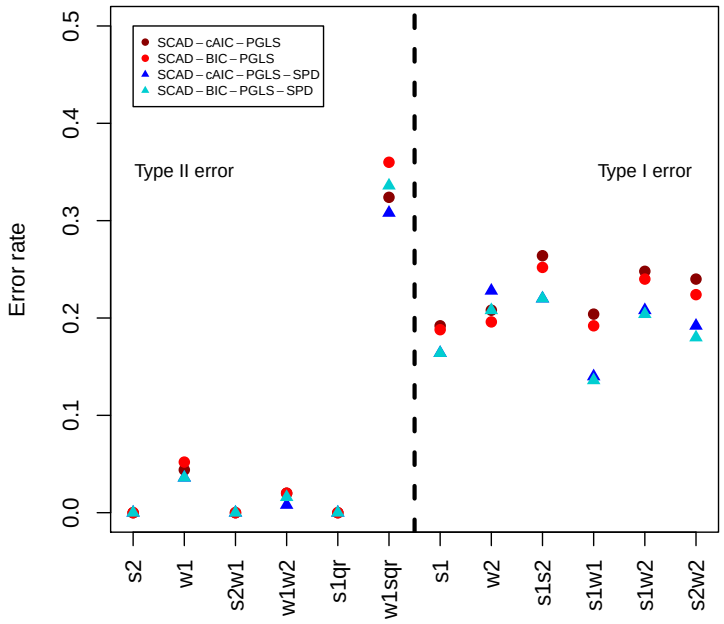


Figure 3.8: Type I and II error rates for the wind tunnel design by LASSO and $\text{ALASSO}_{0.5}$ at $\eta = 10$.

(c) SCAD



(d) EN

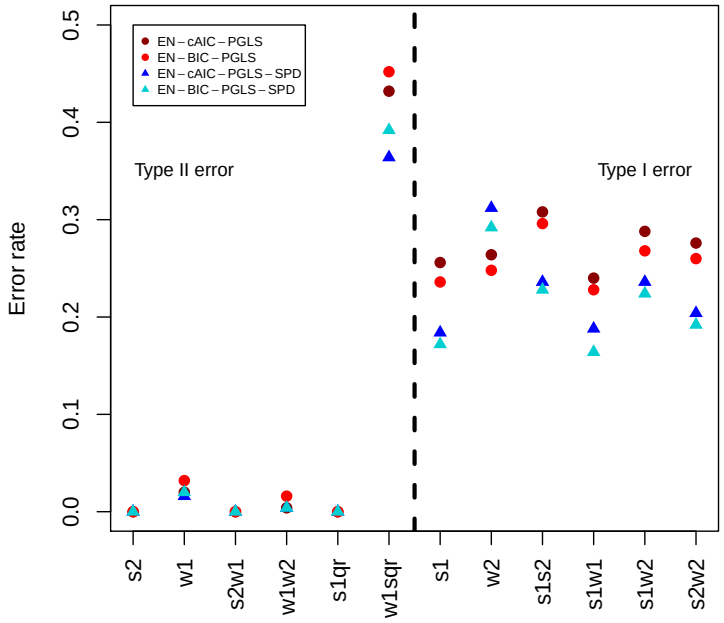


Figure 3.9: Type I and II error rates for the wind tunnel design by SCAD and EN at $\eta = 10$.

(e) Backward and LARS

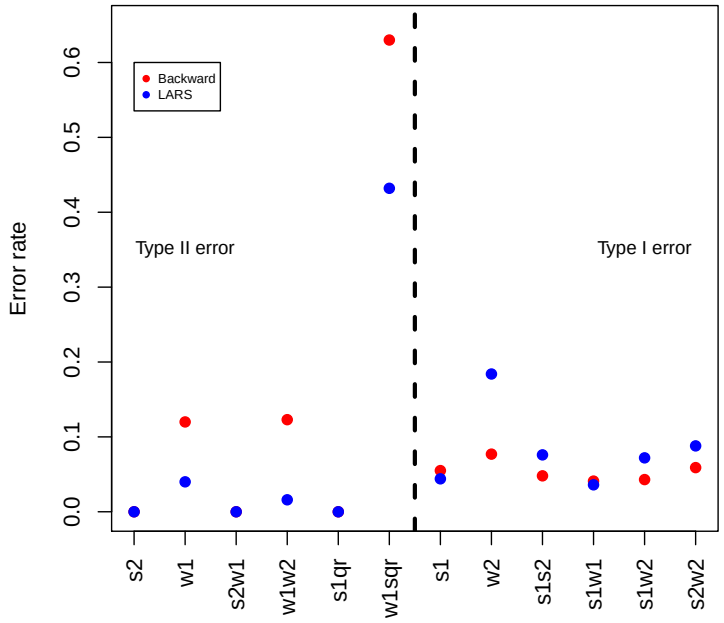


Figure 3.10: Type I and II error rates for the wind tunnel design by Backward and LARS at $\eta = 10$.

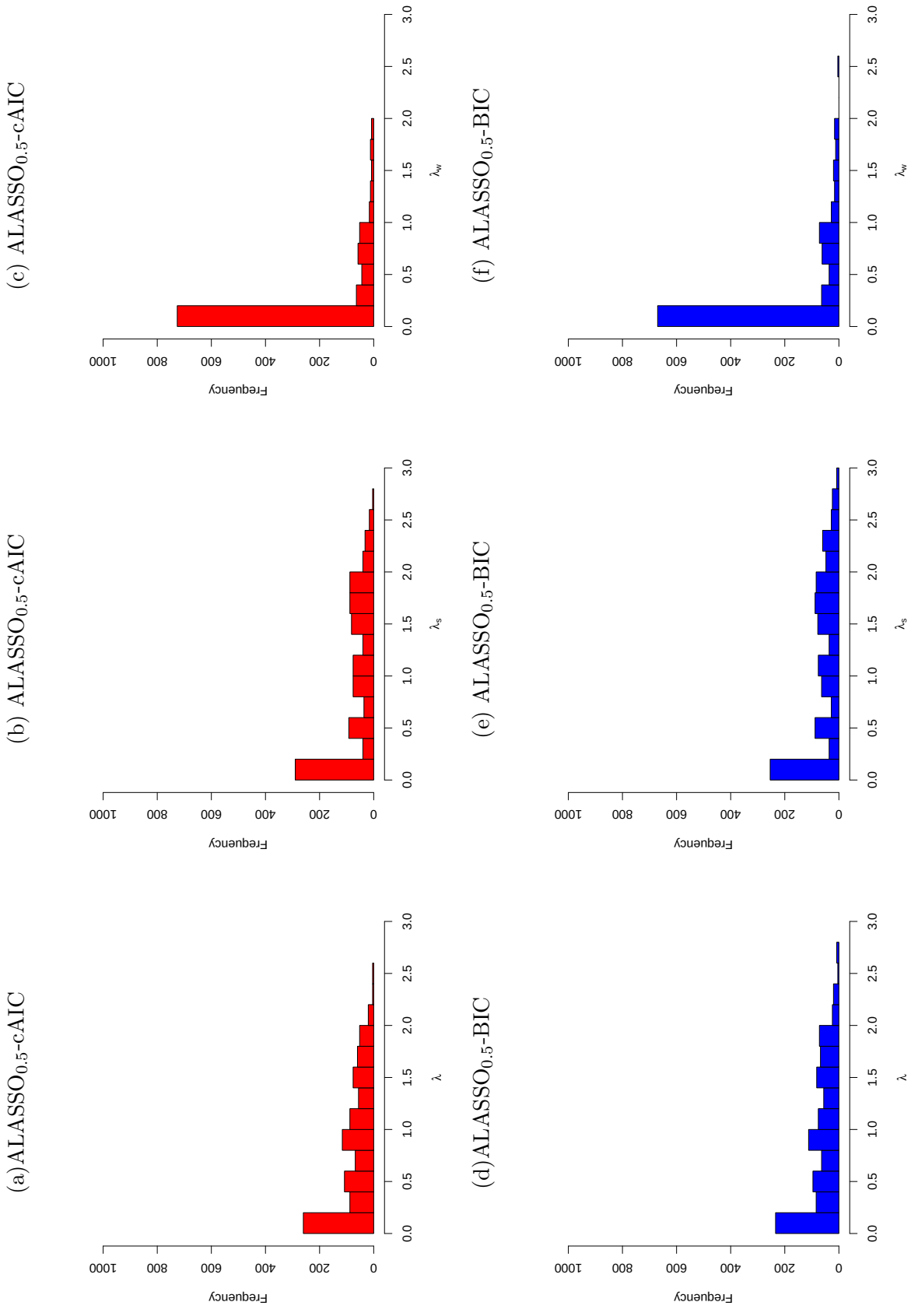


Figure 3.11: The frequency of selected λ for $\text{ALASSO}_{0.5}$ using the wind tunnel design at $\eta = 1$.

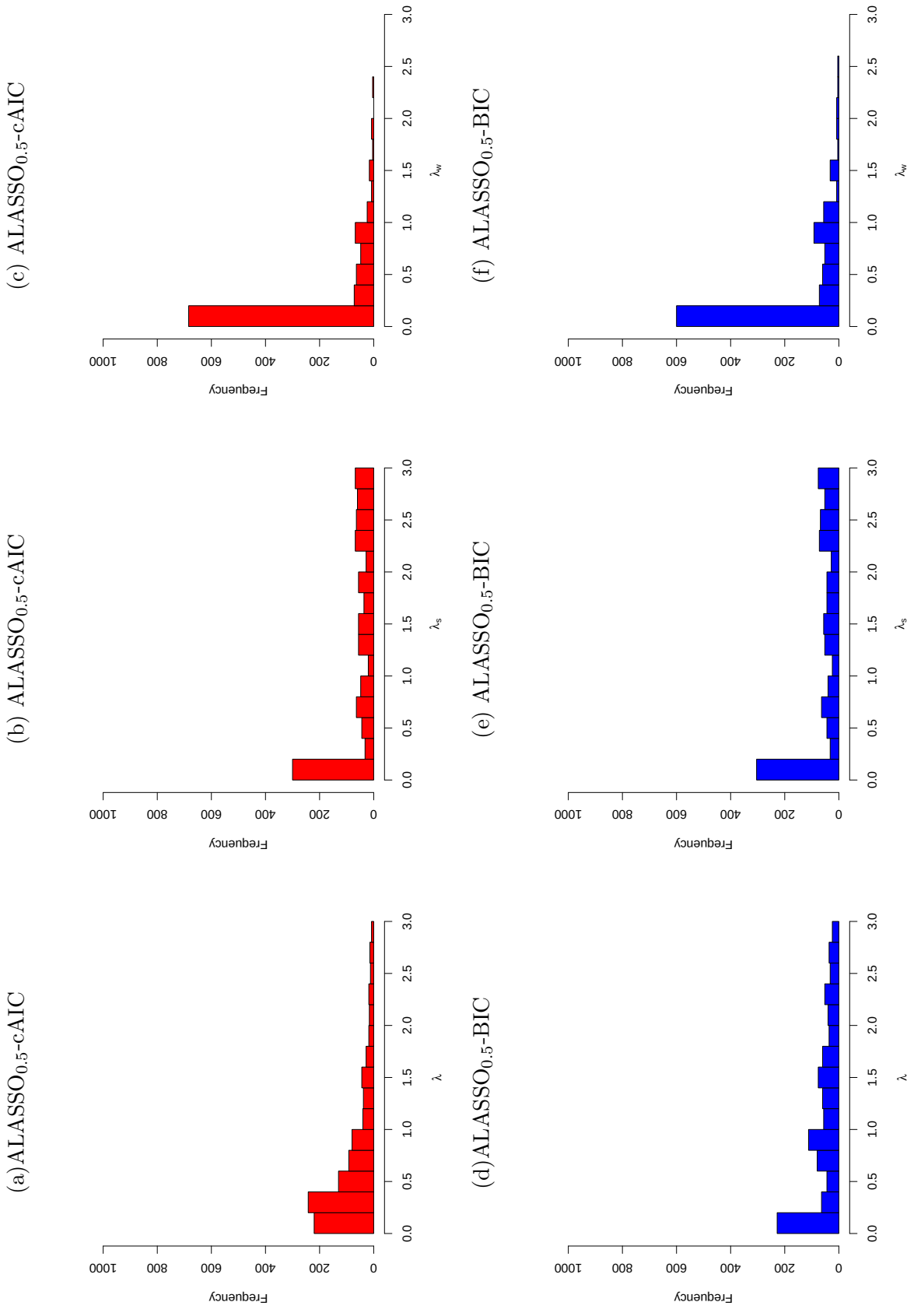


Figure 3.12: The frequency of selected λ for ALASSO$_{0.5}$ using the wind tunnel design at $\eta = 10$.

3.2.2 Simulation Study Using the Design of the Freeze-Dried Coffee Experiment

A simulation study was performed to examine the performance of the PGLS and PGLS-SPD estimates and compare them to the traditional model selection methods. Using the design of the freeze-dried coffee experiment from Table 1.5, the response variable was generated given the true model

$$\mathbb{E}(\mathbf{Y}) = 4w + 4s_1 - 3s_2 + 2s_3 - 4ws_1 + 3ws_2 - ws_3 + 4s_1s_2 + 3s_1s_3 + 2s_1s_4 + s_2s_4 + 4w^2 + 2s_1^2 - s_2^2 + 2s_4^2$$

In this experiment, 15 active variables (w, s_1, s_2, s_3, ws_1, ws_2, ws_3, s_1s_2, s_1s_3, s_1s_4, s_2s_4, w^2, s_1^2, s_2^2, and s_4^2) and five non-active variables ($s_4, ws_4, s_2s_3, s_3s_4,$ and s_3^2) were assumed. We assumed this model as we would like to study a challenging model for this design. This simulation represents the most challenge case as the design is non-orthogonal with six whole plots. The aim was to detect 15 active variable in a model with 20 variables in total. The quadratic terms in this design were very hard to detect by all methods. There were considerable correlations between the variables in this model (see Figure 3.3 and Figure 3.4).

Figure 3.14, Figure 3.15, and Figure 3.16 present the performance of PGLS and PGLS-SPD at $\eta = 1$ whereas Figure 3.17, Figure 3.18, and Figure 3.19 present the performance of PGLS and PGLS-SPD at $\eta = 10$, by cAIC and BIC for all functions in Sections 2.4 to 2.9. The left side of the figures displays the Type II error rate while the right side presents the Type I error rate. For details about Type I and II error rates, see Table B.1 and Table B.4 in Appendix B.

With respect to the Type I error rate, from Table B.1 and Table B.2 as well as Figure 3.14 to Figure 3.19, it can be seen that the Type I error has a rate larger than 0.05 by all methods. However, s_2s_3 achieved a Type I error rate of 0.05 using ALASSO$_{0.5}$ for PGLS and PGLS-SPD at $\eta = 1$. This was the lowest value among all methods. Moreover, backward elimination and LARS provided lower Type I error rates than the PGLS and PGLS-SPD. This is similar to the analysis of the simulation of the wind tunnel design. The frequency of selected the shrinkage parameters in Figure 3.20 and Figure 3.21 plays

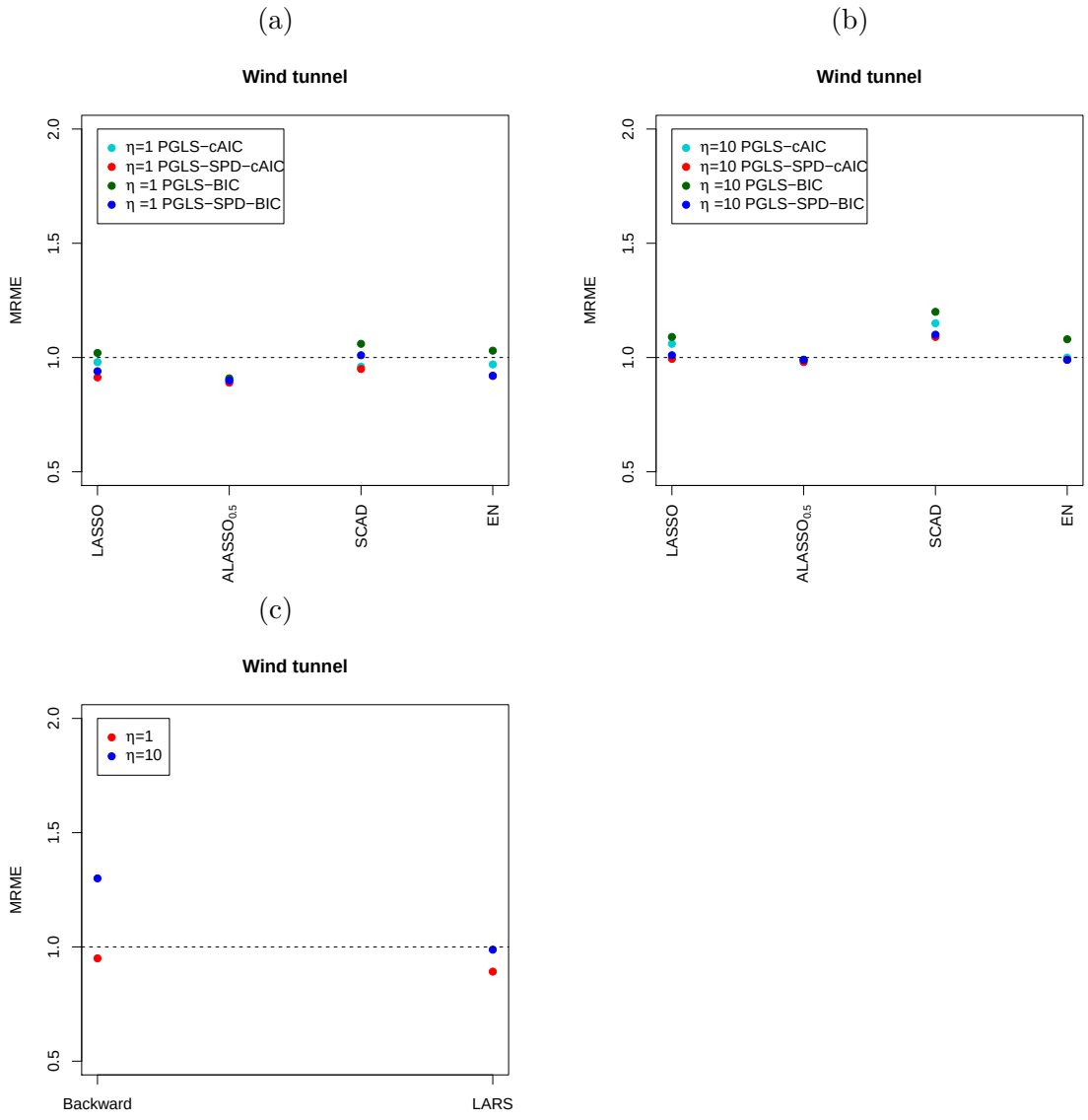


Figure 3.13: Median relative model error (MRME) for the wind tunnel design.

a role in increasing the Type I error rate for the PGLS and PGLS-SPD. For example, in Figure 3.20 at $\eta = 1$, we note that the selected λ by cAIC (Figure 3.20 (a)) and the selected λ_s by cAIC (Figure 3.20 (b)) tends to be assigned small values and, for some models, to be zero. This results in the inclusion of subplot factors by PGLS and PGLS-SPD. The assumed non active terms are all subplot factors; thus, the PGLS and PGLS-SPD recorded a high Type I error rate when they were included in the model. Moreover, in Figure 3.21 at $\eta = 10$, we present the selected λ_s by cAIC (Figure 3.21 (b)) and by BIC (Figure 3.21 (e)). The cAIC tends to select large values more frequently than the BIC. This resulted in the Type I error rate for EN by cAIC being lower than

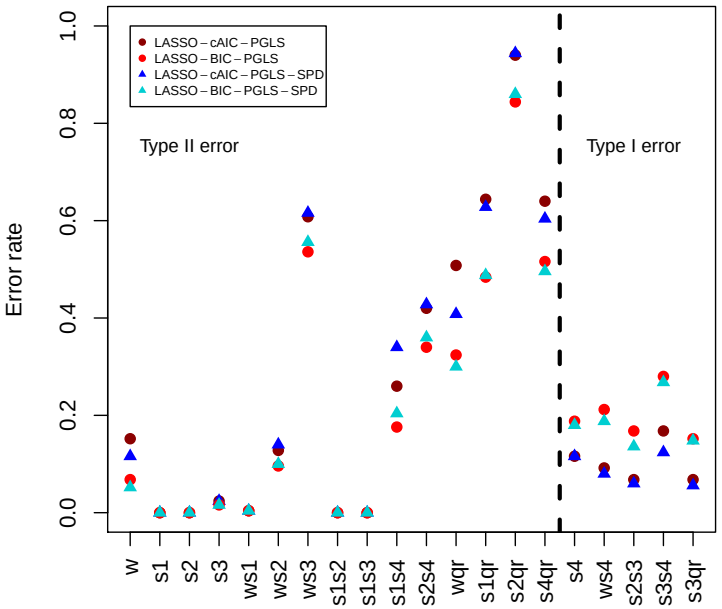
the Type I error rate for EN by BIC at $\eta = 10$. Overall, in a comparison between PGLS and the PGLS-SPD, the PGLS-SPD has a lower Type I error rate than the PGLS.

With respect to the Type II error rate at $\eta = 1$, Table B.3 in Appendix B and Figure 3.14 to Figure 3.16 show that it is hard to detect the quadratic factors and some of the interaction factors because this design is non-orthogonal. It can be noted that the LARS overcome the PGLS-SPD with respect to the main effect factors. However, the EN recorded the lowest Type II error rate for subplot factors. This can be explained using Figure 3.20 and Figure 3.21. For example, the selected λ_s by cAIC (Figure 3.20 (b)) and the selected λ_s by BIC (Figure 3.20 (e)) are more frequently leaving the subplot factors without penalisation in some models. This results in a better detection of the active subplot factors by PGLS-SPD using EN than by PGLS using EN for some factors (e.g. s_3, ws_2 and ws_3). This performance of selecting the λ_s encourages the methods to include the subplot factors which results in a high Type I error rate and a low Type II error rate. The main whole-plot factor w recorded a Type II error rate of 0.05 for both EN-BIC and LASSO-BIC in which the PGLS-SPD approach was applied. Due to the lack of the number of whole plots, it is difficult to detect w and w^2.

Table B.4 in Appendix B and Figure 3.17 to Figure 3.19 show that increasing η to 10, causes the σ^2_γ to become 10 times larger than σ^2_ϵ. Therefore, the subplot factors are more likely to be included to the model. This is true for most of the (main and interaction) subplot factors as the Type II error rate is zero for all of the subplot factors except for ws_3 and s_2s_4. However, s_2s_4 recorded a Type II error rate of 0.02 which is still under control. Although it is hard to detect w and w^2 as explained above and the high Type II error rate by both PGLS and PGLS-SPD, it can be observed that the PGLS-SPD reduced the Type II error rate compared to the PGLS.

Figure 3.22 shows the MRME values of all methods at different settings compared to the the GLS estimator as in Section 3.2. From Figure 3.22 (a) at $\eta = 1$ and Figure 3.22 (b) at $\eta = 10$, the cAIC for all methods provided a higher MRME than BIC. This indicates the estimates by BIC are closer to the GLS estimator than the cAIC. The EN, ALASSO$_{0.5}$, LASSO at Figure 3.22 (a) $\eta = 1$ and at Figure 3.22 (b) $\eta = 10$ by PGLS-SPD using the BIC has similar estimates to the GLS. The backward elimination and LARS in Figure 3.22 (c) have MRME values worse than the GLS estimator.

(a) LASSO



(b) $\text{ALASSO}_{0.5}$

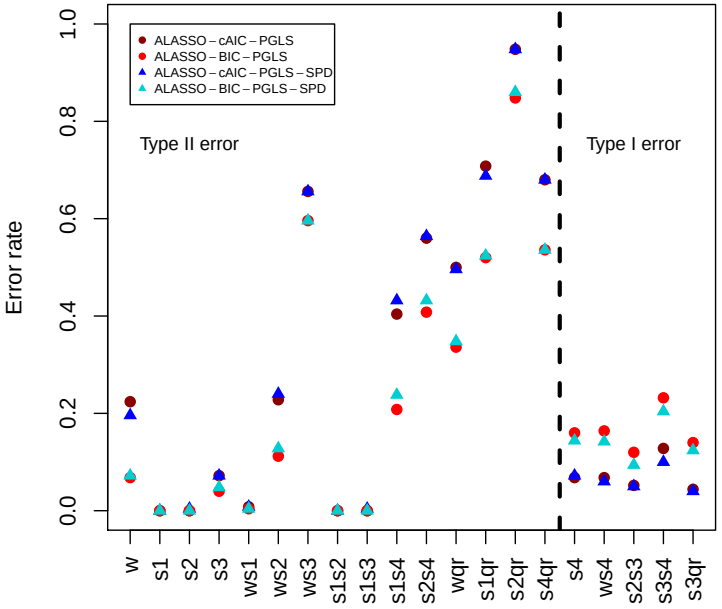
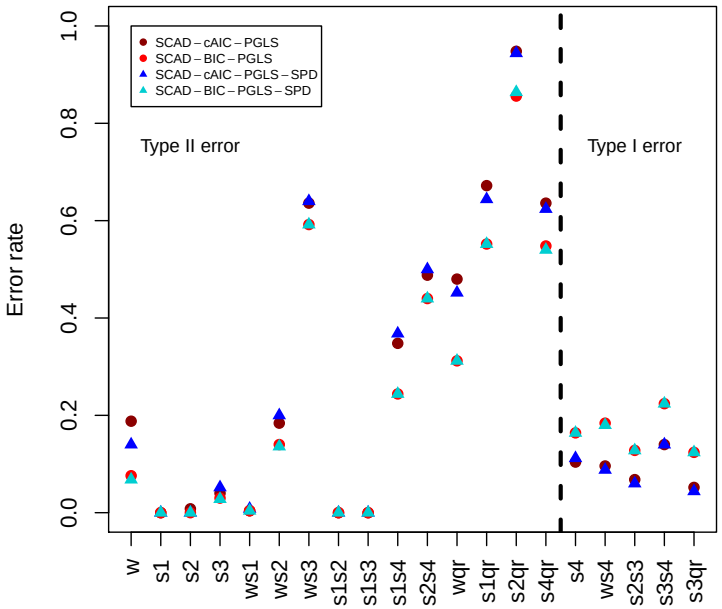


Figure 3.14: Type I and II error rates for the freeze-dried coffee design by LASSO and $\text{ALASSO}_{0.5}$ at $\eta = 1$.

(c) SCAD



(d) EN

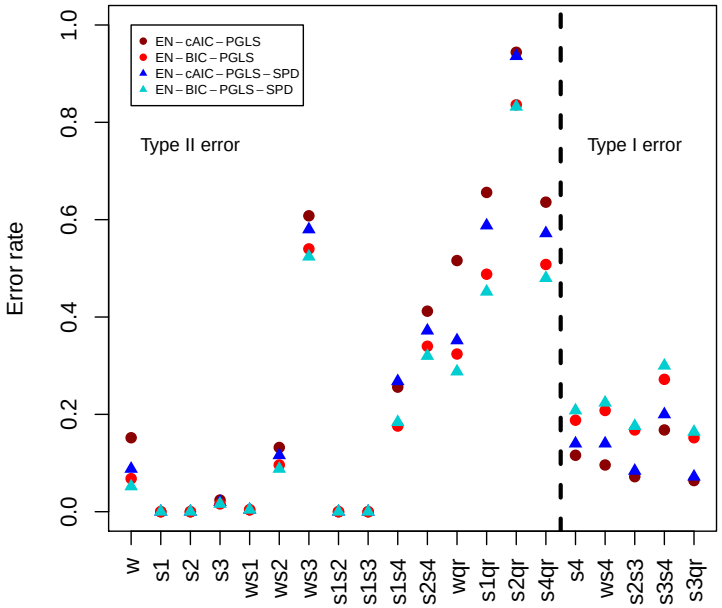


Figure 3.15: Type I and II error rates for the freeze-dried coffeee design by SCAD and EN at $\eta = 1$.

(e) Backward and LARS

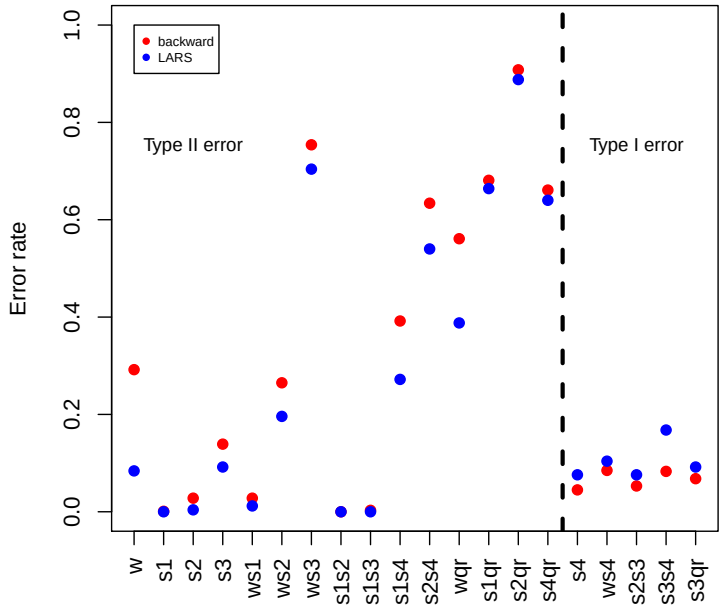
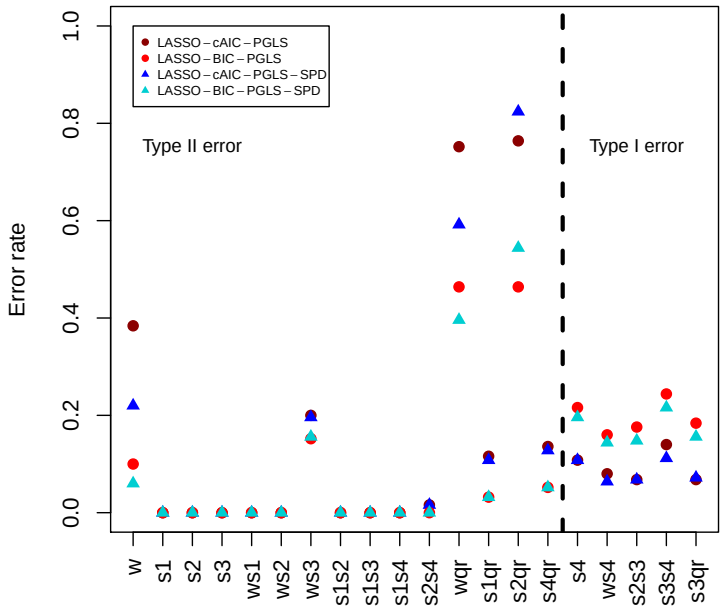


Figure 3.16: Type I and II error rates for the freeze-dried coffee design by Backward and LARS at $\eta = 1$.

(a) LASSO



(b) $\text{ALASSO}_{0.5}$

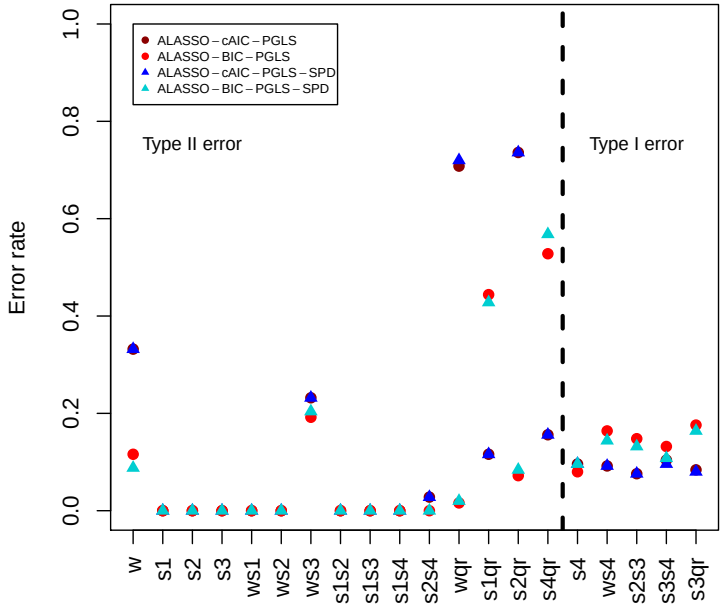
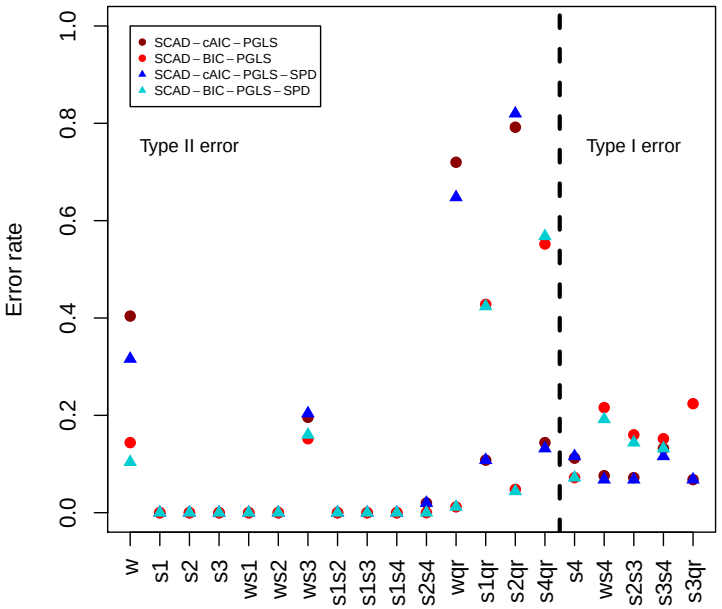


Figure 3.17: Type I and II error rates for the freeze-dried coffee design by LASSO and $\text{ALASSO}_{0.5}$ at $\eta = 10$.

(c) SCAD



(d) EN

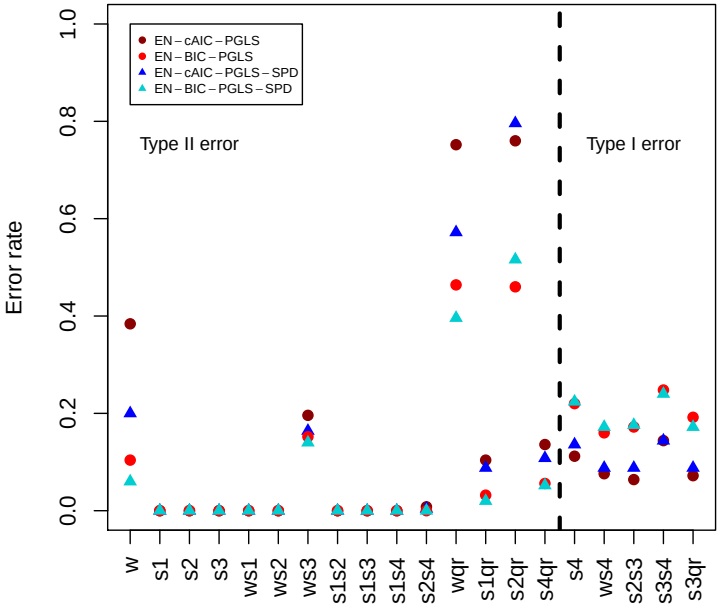


Figure 3.18: Type I and II error rates for the freeze-dried coffee design by SCAD and EN at $\eta = 10$.

(e) Backward and LARS

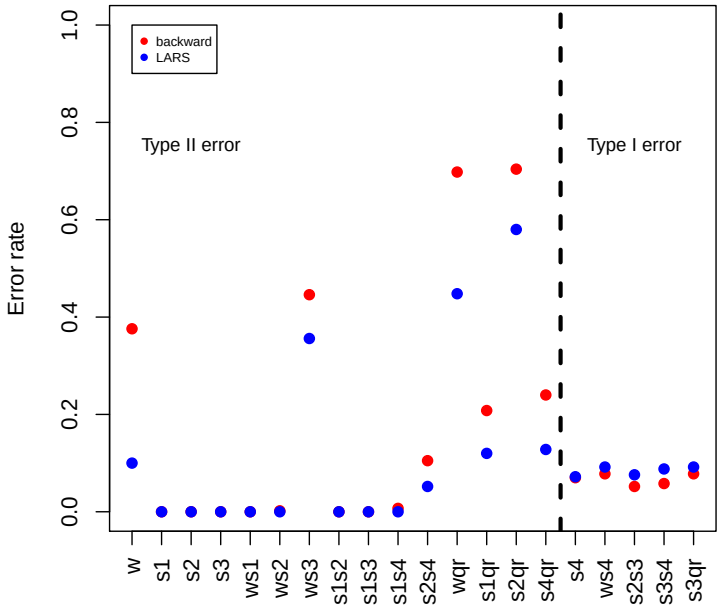


Figure 3.19: Type I and II error rates for the freeze-dried coffee design by Backward and LARS at $\eta = 10$.

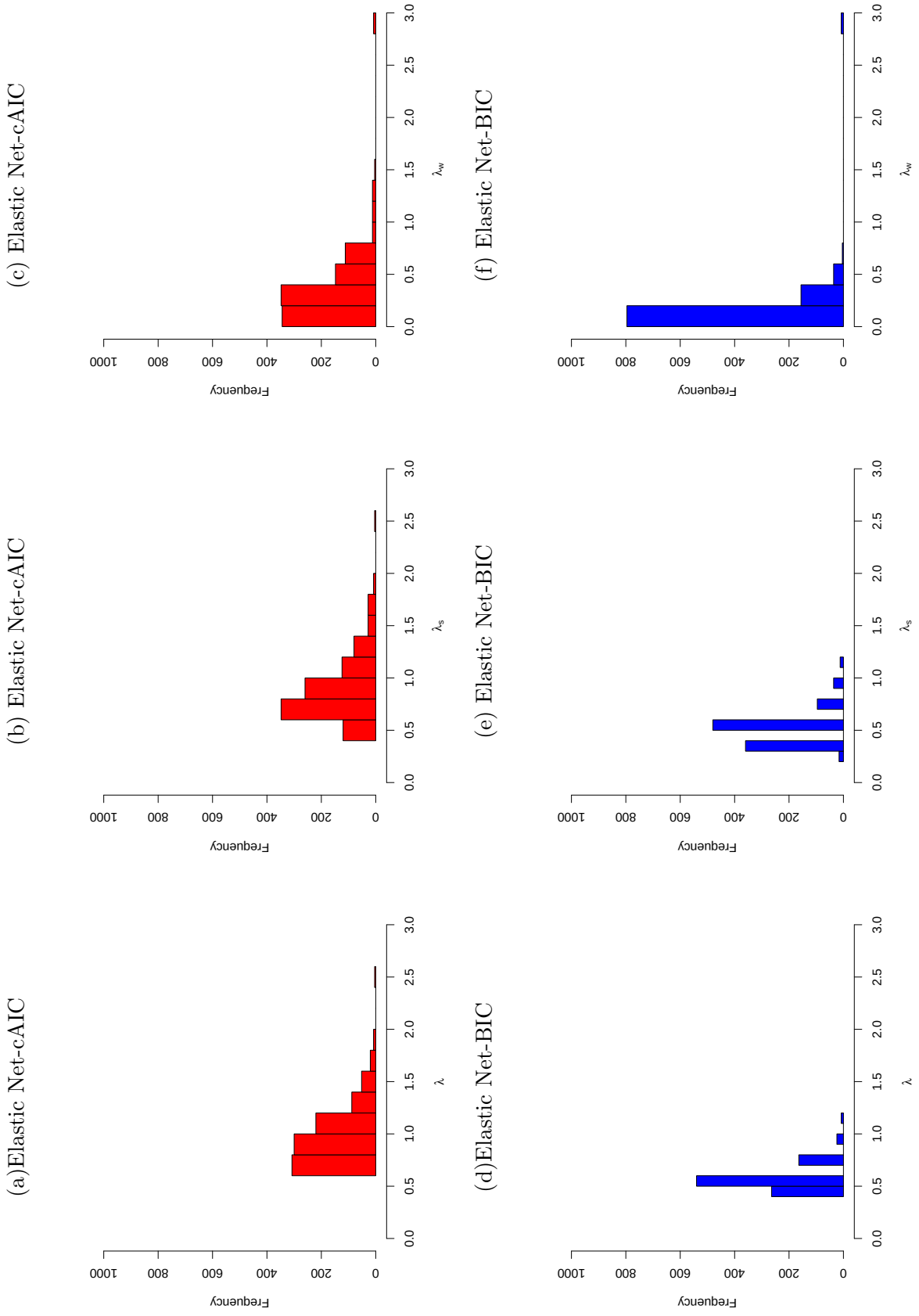


Figure 3.20: The frequency of selected λ for Elastic Net using the freeze-dried coffee design at $\eta = 1$.

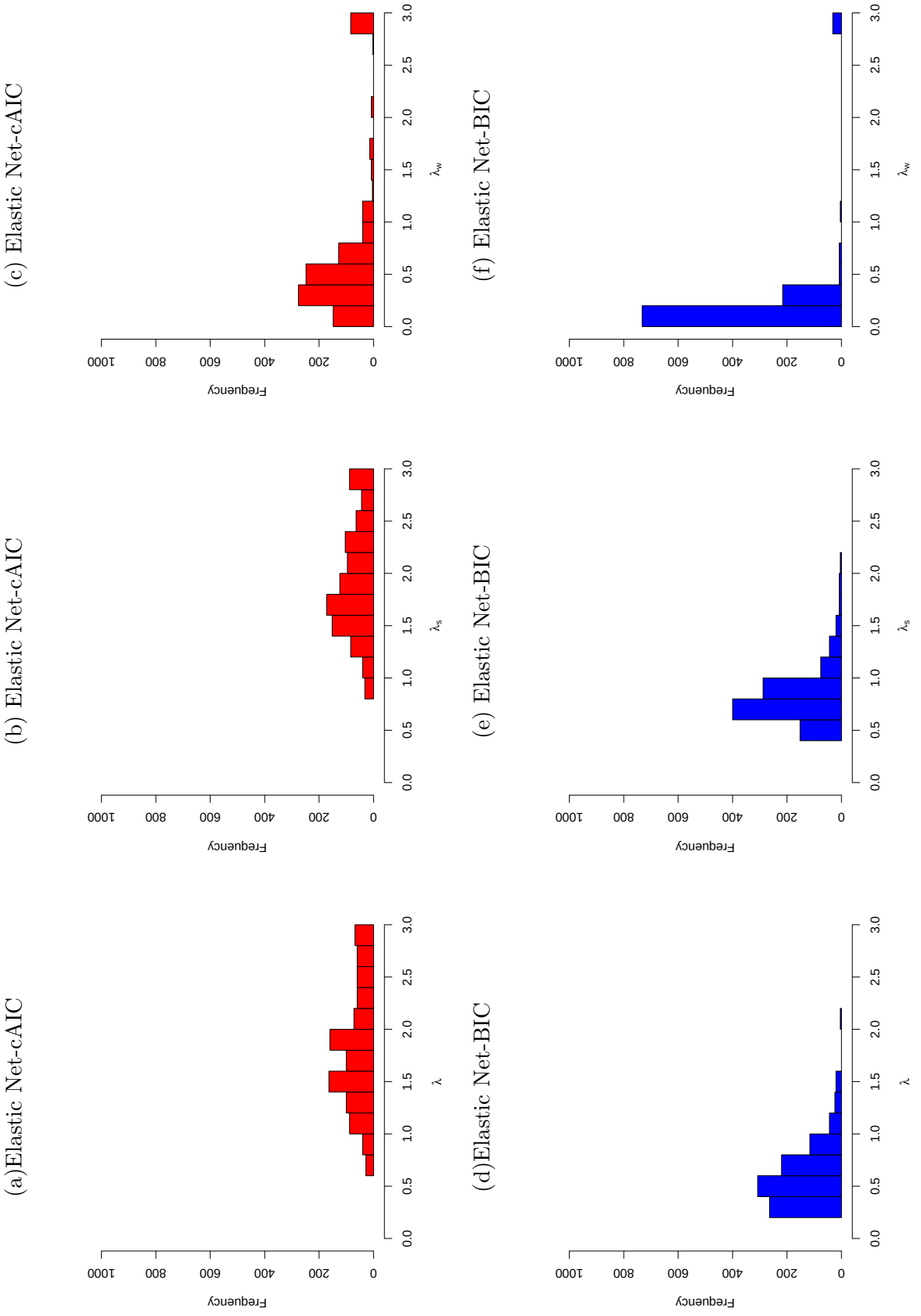


Figure 3.21: The frequency of selected λ for Elastic Net using the freeze-dried coffee design at $\eta = 10$.

3.3 Discussion

In this chapter, the theoretical approach described in Section 2.2 using the PGLS estimator in Section 2.2.1 and PGLS-SPD estimator in Section 2.2.2 was applied to two industrial experiments. Four different sorts of penalty functions in relation to the PGLS and PGLS-SPD approaches were investigated. The features of the design were found to affect the performance of the PGLS and PGLS-SPD as well as the trade off between Type I and II error rate. For example, in non-orthogonal designs and with few main stratum as in the design of the freeze-dried coffee experiment, REML showed its weakness in estimating the variance of the random effects as discussed in Section 1.3. This indeed affected the point estimates of the fixed effects by the PGLS and PGLS-SPD. The analysis from the real-life experiments showed that for an orthogonal design such as in the wind tunnel experiment, the backward elimination and LARS tend to end up with simpler models than the PGLS and PGLS-SPD. On the other hand, for non-orthogonal design such as in the freeze-dried coffee experiment, for both PGLS and PGLS-SPD and by applying the ALASSO$_{0.5}$, SCAD, and EN in which the cAIC was used to select the tuning parameters had almost similar final models to the backward and LARS.

The analysis from the simulation study which used an orthogonal design from the wind tunnel experiment showed that backward elimination followed by LARS could control the Type I error rate at both $\eta = 1$ and 10 better than PGLS and PGLS-SPD. With regard to the Type II error rate, apart from the quadratic whole-plot factor w_1^2, all methods succeeded in controlling the Type II error rate to be less than 0.05 for all factors. However, PGLS-SPD using the EN at $\eta = 1$ and using the SCAD at $\eta = 10$ controlled the quadratic whole-plot factor w_1^2 better than other methods (see Table A.3 and Table A.4 in Appendix A for more details).

The analysis from the simulation study which used the non-orthogonal design from the freeze-dried coffee experiment showed that all methods failed in controlling Type I error rate under 0.05. Moreover, the Backward yields the lowest Type I error rate. After that, the LARS which overcomes the PGLS-SPD method. However, the ALASSO$_{0.5}$ by PGLS at $\eta = 1$ had the lowest Type I error rate. Although the backward elimination

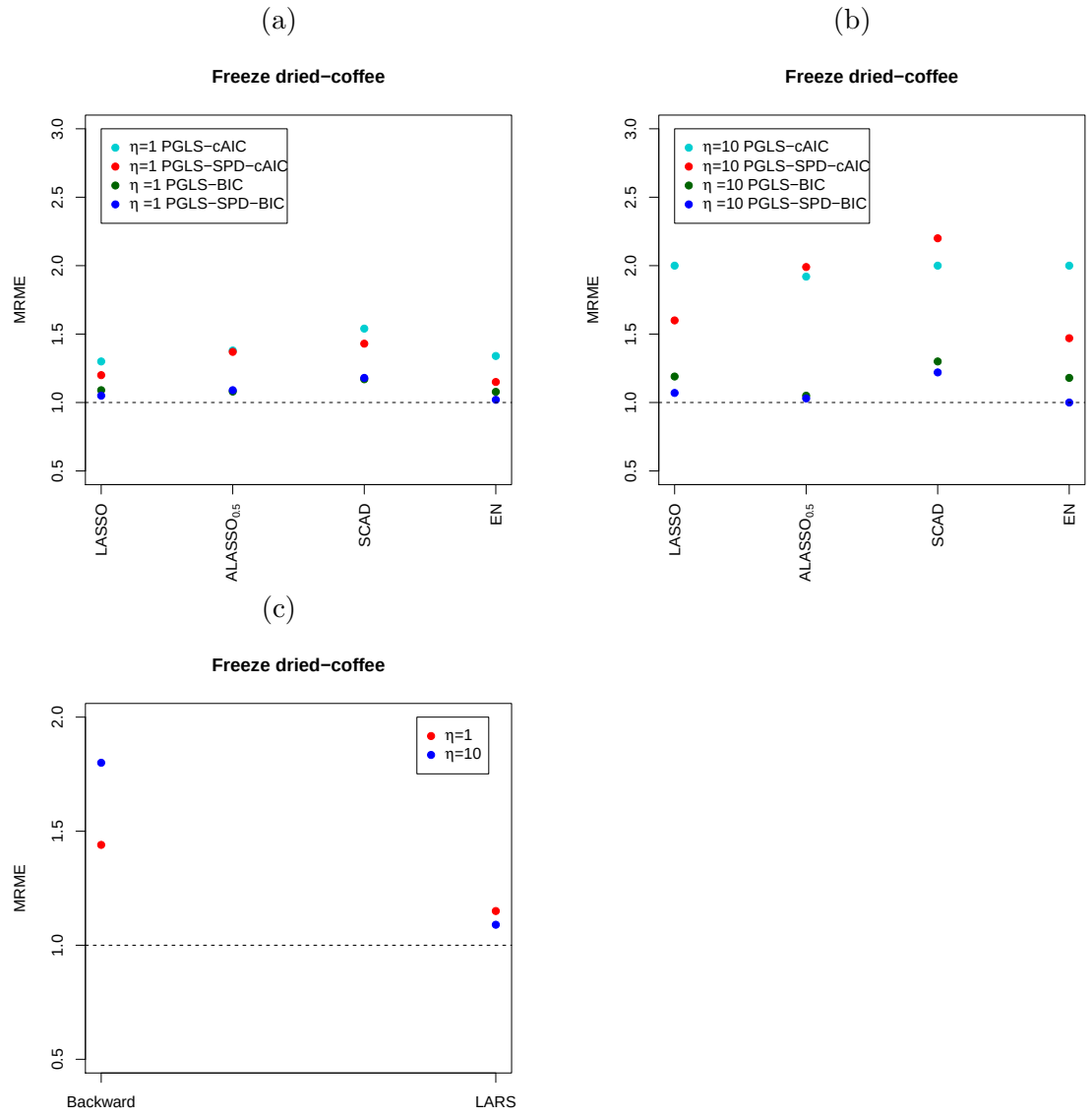


Figure 3.22: Median relative model error (MRME) for the freeze-dried coffee design.

seemed to have the lowest Type I error rate at $\eta = 10$, the SCAD overcomes the backward elimination at some factors. For example, s_3^2 has the lowest Type II error rate using SCAD by PGLS and PGLS-SPD in which the cAIC was used to select the tuning parameter (see Table B.1 and Table B.2 in Appendix B for more details). On the other hand, the Type II error rate was hard to control by both PGLS and PGLS-SPD mainly for the quadratic effect factors and some of the interaction effect factors. Apart from these effect factors, LARS overcomes the PGLS-SPD with respect to the main effect factors in some cases. Also, we note the PGLS-SPD using the EN had the lowest Type

II error rate among other methods at $\eta = 1$ and 10.

Also, in a small and a non-orthogonal split-plot design such as the freeze-dried coffee design, both the PGLS and PGLS-SPD yielded a higher prediction error for most of the methods than the orthogonal design in the simulation by the wind tunnel design as can be seen in Figure 3.13 and Figure 3.22. Furthermore, the PGLS-SPD at both simulations had lower MRME than the PGLS at both $\eta = 1$ and $\eta = 10$. To sum up, the simulation studies recommend the use of the LARS and the PGLS-SPD to detect the active effect factors for both an orthogonal and non-orthogonal split-plot designs.

Chapter 4

Bayesian Analysis Methods for Responses from Split-Plot Experiments

This chapter discusses Bayesian variable selection methods for split-plot designs. Bayesian methods are important approaches due to their ability to quantify uncertainty. In such an approach, prior distributions that represent subjective beliefs about parameters are assigned to the regression coefficients. By applying Bayes' rule, prior beliefs are updated by the data and transformed into posterior distributions, on which all inference is based. As reviewed by Tang et al. (2016), Mitchell and Beauchamp (1988) introduced Bayesian variable selection via spike-and-slab prior distributions. The spike prior that they used was a probability mass at zero to remove the non-significant variables. Their slab is the uniform distribution with a large symmetric range in order to keep the significant variables. Following their work, many priors were proposed to implement the spike-and-slab property. George and McCulloch (1993) proposed the Stochastic Search Variable Selection (SSVS) in which the coefficients are sampled from a mixture of two normal distributions with different variances. The spike part is the distribution with a small variance while the slab part is the distribution with a much larger variance. Also, Geweke (1996) proposed positive mass at zero for the spike part and a normal distribution for the slab part. In addition, Tan and Wu (2013) proposed a Bayesian approach for model

selection in fractionated split-plot experiments with application to robust-parameter design. In their work, they extend the SSVS algorithm of George and McCulloch (1997) to account for the split-plot error structure. They derive an expression for the posterior probability of a model that requires computation of, at most, two unidimensional integrals, and employ this quantity for model selection. They were able to integrate out the coefficients and the variance components from the joint posterior distribution of all parameters because they use the conjugate normal-inverse gamma prior for these parameters. The integrals are computed with Gaussian quadrature, and Global and Local search algorithms to find models with high posterior probabilities. More recently, the work by Matthews (2015) applied the SSVS algorithm to select variables from a multivariate linear mixed model in which the response is a $n \times m$ matrix for n observations and m response vectors.

The novel contribution of this thesis to Bayesian variable selection is motivated by a very specific experimental design of data from experiments subject to restricted randomisation. In this work, we use the split-plot design as an example for the restricted randomisation experiments. We have two different levels of the experimental units one for the whole plots and the other for the subplots in the split-plot design; see, for example, Jones and Nachtsheim (2009). To address this issue, we adapt the SSVS algorithm in which we sample the subplot coefficients using a mixture of normal posterior distributions with a slab variance different from the slab variance which will be used in the mixture normal posterior distributions for the whole-plot coefficients. This method reduces Type I and II error rates as well as reducing the prediction error for split-plot design rather than applying the SSVS algorithm in which all coefficients will be sampled from a mixture of normal posterior distributions with one slab variance.

4.1 Motivation and Aim of Work

In the previous chapters, we found that one major disadvantage of the frequentist methods analysis of the split-plot experiments is the high rates of Type I error. The frequentist analysis is dependent on the estimates of the variance components, yet these estimates cannot be precisely calculated because of the deficiency of the degrees of freedom for the

random effects in the split-plot design. This issue was discussed by Gilmour and Goos (2009). Introducing a prior distribution for the variance components in the linear mixed model provides additional some information to overcome the problem of the variance estimation. In this chapter, the analysis of two split-plot experiments will be presented in order to compare the analysis of the Bayesian approach with the frequentist approach. Section 4.2 will introduce the Bayesian methodology that we used in this work. Section 4.3 will explain the hierarchical mixture model for variable selection. In Section 4.4, we will adapt the SSVS method to fit data from split-plot designs by introducing the Stochastic Search Variable Selection for Split-Plot Design (SSVS-SPD). In Section 4.5, we will explain the computational algorithm used in the Bayesian variable selection. Finally, Section 4.6 will provide a discussion of this chapter.

4.2 The Bayesian Methodology

In Section 4.2.1, we will describe the Bayesian framework. In Section 4.2.2, we will discuss the Monte Carlo Markov Chain (MCMC) methods.

4.2.1 Bayesian Framework

The essential philosophy behind Bayesian inference is to update a prior distribution for an unidentified parameter to a posterior distribution by Bayes' theorem. Bayes' theorem can be used to estimate the conditional distributions. While the frequentist approach treats the parameters as unknown and fixed, the Bayesian approach regards them as random variables. We can define the prior distribution $p(\boldsymbol{\theta})$ as the probability density (or mass) function which reflects our beliefs about $\boldsymbol{\theta}$ in the parameter space $\boldsymbol{\Theta}$. For given data $\mathbf{y} = (y_1, y_2, ..., y_n)'$, the likelihood function $f(\mathbf{y}|\boldsymbol{\theta})$ can then be defined given the parameter $\boldsymbol{\theta}$ for the data $\mathbf{y}$. Also, we can define the posterior density (or mass) function $p(\boldsymbol{\theta}|y_1, y_2, ..., y_n)$ which represents our updated belief about $\boldsymbol{\theta}$ given the observed data $\mathbf{y}$.

Using Bayes theorem, the posterior density of $\boldsymbol{\theta}$ given $\mathbf{y}$ is:

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

Bayesian inference continues from this distribution. The denominator of equation (4.1) is the marginal likelihood of $\mathbf{y}$, and it often does not need to be calculated because it is independent of $\boldsymbol{\theta}$. Bayes' rule can then be written as:

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

Equation (4.2) defines the unnormalised posterior density. The posterior then is proportional to the likelihood $\times$ the prior. For more details on Bayesian inference, see Gelman et al. (2014) and O'Hagan and Forster (2004).

A prior distribution can be selected based on past information or experimental practice. It can be informative or uninformative. The informative distribution is given numerical information to estimate the parameter of concern. The uninformative reflects equilibrium among outcomes when weak information about the parameter is presented. There are two types of uninformative priors: proper prior and improper prior. The density for proper prior distribution integrates to 1 whereas the integral of the density for an improper distribution is not finite. If the prior integrates to any positive finite value, it is called an unnormalised density and can be renormalised- multiplied by a constant- to integrate to 1 (Gelman et al., 2014; O'Hagan and Forster, 2004).

We say that the prior $p(\boldsymbol{\theta})$ is a conjugate prior for the likelihood function if its conditional posterior $p(\boldsymbol{\theta}|\mathbf{y})$ belongs to the same parametric family as $p(\boldsymbol{\theta})$. In practice, conjugate priors are broadly used due to their computational properties as they make the calculation of the conditional posterior easier, and yield a standard form of the posterior. However, in several implementations, we are unable to use a conjugate prior (e.g., if one wants to use a reference prior which inserts the smallest amount of individual belief into the analysis). But most frequently, a Bayesian will have a personal belief about the problem that cannot be expressed in terms of a convenient conjugate prior. Consequently, we cannot compute the posterior distribution in a standard form (Gelman et al., 2014; O'Hagan and Forster, 2004).

4.2.2 Markov Chain Monte Carlo (MCMC) Methods

Markov Chain Monte Carlo simulation is a general method based on drawing values of the $\boldsymbol{\theta}$ from approximate distributions, and then correcting those draws to better approximate the target posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ (Gelman et al., 2014; O'Hagan and Forster, 2004; Gilks et al., 1995) . A Markov chain can be defined as a sequence of random variables $\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, ...$ for which for any iteration t, the distribution of $\boldsymbol{\theta}^t$ depends only on the most recent value $\boldsymbol{\theta}^{t-1}$ (Gelman et al., 2014; O'Hagan and Forster, 2004; Gilks et al., 1995). A Markov chain is generated by sampling $\boldsymbol{\theta}^t \sim p(\boldsymbol{\theta}|\boldsymbol{\theta}^{t-1})$. This p is called the transition kernel of the Markov chain. Therefore, $\boldsymbol{\theta}^t$ depends only on $\boldsymbol{\theta}^{t-1}$, not on $\boldsymbol{\theta}^0, \boldsymbol{\theta}^1, ..., \boldsymbol{\theta}^{t-2}$.

As $t \to \infty$, the sampling from Markov chain converges to the posterior for the right choice of transition kernel $p(\boldsymbol{\theta}|\mathbf{y})$. Thus, we should run the simulation long enough so that the distribution of the current draws is close enough to $p(\boldsymbol{\theta}|\mathbf{y})$.

4.2.2.1 Metropolis-Hastings Sampling

Metropolis-Hastings sampling was proposed by Metropolis et al. (1953) and Hastings (1970). The theory of this sampling is based on rejection sampling. The acceptance-rejection method is a technique of getting samples from a distribution with an unknown form. The Metropolis-Hastings algorithm is a common expression for a family of Markov chain simulation methods. We will illustrate how the Gibbs sampler is also a special case of Metropolis-Hastings sampling.

It is worth describing the Metropolis algorithm first, then broadening it to discuss the Metropolis-Hastings algorithm. Let $p(\boldsymbol{\theta}|\mathbf{y})$ be the conditional posterior distribution where we want to sample from. Let $\boldsymbol{\theta}^t$ be the current parameter value, and let $\pi(\boldsymbol{\theta})$ be the proposal density. The proposal density is much like a conventional transition operator for a Markov chain, the proposal distribution depends only on the previous state in the chain. However, the transition operator for the Metropolis algorithm has

a additional step that assesses whether or not the target distribution has a sufficiently large density near the proposed state to warrant accepting the chain. The Metropolis algorithm for sampling $\boldsymbol{\theta}^t$ for $t = 1, 2, ...$ is given as follows:

1. Sample a proposal $\boldsymbol{\theta}_*^{t+1}$ from proposal distribution $\pi(\boldsymbol{\theta})$ at time t, $\pi_t(\boldsymbol{\theta}_*^{t+1}|\boldsymbol{\theta}^t)$. In the Metropolis (but not the Metropolis-Hastings algorithm), the proposal distribution must be symmetric, fulfilling the condition $\pi_t(\boldsymbol{\theta}_a|\boldsymbol{\theta}_b) = \pi_t(\boldsymbol{\theta}_b|\boldsymbol{\theta}_a)$ for all $\boldsymbol{\theta}_a, \boldsymbol{\theta}_b$ and t. An example of choices of symmetric proposals includes Gaussian distribution or Uniform distribution centred at the current state of the chain. The algorithms using Gaussian distribution as a proposal are called "Random-walk Metropolis algorithm".

2. Calculate the ratio of the densities,

$$r = \frac{p(\boldsymbol{\theta}_*^{t+1}|\mathbf{y})}{p(\boldsymbol{\theta}^t|\mathbf{y})}, \tag{4.3}$$

   where we do not need the normalising constant to calculate this ratio.

3. Calculate the acceptance probability $\alpha(\boldsymbol{\theta}^t, \boldsymbol{\theta}_*^{t+1}) = \min(r, 1)$.

4. Sample u from $U(0, 1)$.

5. Set

$$\boldsymbol{\theta}^{t+1} = \begin{cases} \boldsymbol{\theta}_*^{t+1} & \text{if } u < \alpha \\ \boldsymbol{\theta}^t & \text{otherwise} \end{cases}.$$

The Metropolis-Hastings algorithm generalises the basic Metropolis algorithm in two approaches. First, the proposal distribution need no longer be symmetric. Second, the ratio r in step (2) becomes

$$r = \frac{p(\boldsymbol{\theta}_*^{t+1}|\mathbf{y})\pi(\boldsymbol{\theta}^t|\theta_*^{t+1})}{p(\boldsymbol{\theta}^t|\mathbf{y})\pi(\boldsymbol{\theta}_*^{t+1}|\boldsymbol{\theta}^t)}. \tag{4.4}$$

When we use the prior distribution $\pi(\boldsymbol{\theta})$ as a proposal distribution, the proposal distribution is not conditional on the current value of $\boldsymbol{\theta}$. Then the acceptance probability

equals

$$\alpha(\boldsymbol{\theta}^t, \boldsymbol{\theta}_*^{t+1}) = \min\left\{1, \frac{p(\mathbf{y}|\boldsymbol{\theta}_*^{t+1})}{p(\mathbf{y}|\boldsymbol{\theta}^t)}\right\},$$

where $p(\mathbf{y}|\boldsymbol{\theta})$ is the likelihood function of the parameter $\boldsymbol{\theta}$. The acceptance rate in this case is very high, so this may affect the efficiency of sampling. However, it works for some examples, and the experimenter always needs to check the efficiency of MCMC sampling.

The choice of the proposal distribution is not unique. A good proposal distribution should maintain the acceptance rate of the proposal in a reasonable range. Gelman et al. (2014, p. 605) suggested an acceptance rate value of 0.44 (in one dimension sampling) and 0.23 if more than one parameter is being updated. The choice of the proposal distribution is crucial in Metropolis-Hastings. The most common Metropolis-Hastings algorithms are Random-walk Metropolis-Hastings algorithms.

4.2.2.2 Gibbs Sampling

A Gibbs sampler is the simplest of the Markov chain simulation algorithms, and it is used to sample from the conditional conjugate models, where we can directly sample from each conditional posterior (Gelman et al., 2014; O'Hagan and Forster, 2004). It is rare to find all the conditional posteriors in a model in known forms. One may find some conditional posterior distributions that are possible to directly sample from. Furthermore, one may find some of the conditional posteriors that cannot be straightforwardly sampled from. Therefore, the procedure for this issue is to update the parameters one at a time with the Gibbs sampler used where possible, and one-dimensional Metropolis updating where necessary. This process is called the Metropolis-Hastings within Gibbs sampling and will be used in this work.

The Gibbs sampling involves a proposal from the full conditional distribution which always has a Metropolis-Hastings ratio of 1, i.e., the proposal is always accepted. According to Gelman et al. (2014), at step j of the iteration t, we can update θ_j conditional on all other elements of $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)'$, $j = 1, 2, ..., p$ where p is the number of unknown parameters. The proposal density in Gibbs sampling is the conditional distribution of the current parameter, $\boldsymbol{\theta}_j^t$, hence $\pi(\boldsymbol{\theta}_j^t) = p(\theta_j^t|\boldsymbol{\theta}_{-j}^t, \mathbf{y})$, where $\boldsymbol{\theta}_{-j}^t$ is the current value

for the other elements of $\boldsymbol{\theta} = (\theta_1, \theta_2, ..., \theta_p)'$, and $j = 1, 2, ..., p$.

The acceptance probability in step 3 of Metropolis-Hastings sampling when the proposal density is the conditional distribution is

$$\alpha\left(\theta_j^t, \boldsymbol{\theta}_{j*}^{t+1}\right) = \min\left\{1, \frac{p(\boldsymbol{\theta}_{j*}^{t+1}|\boldsymbol{\theta}_{-j}^t, \mathbf{y})\ p(\boldsymbol{\theta}_j^t|\boldsymbol{\theta}_{-j}^t, \mathbf{y})}{p(\boldsymbol{\theta}_j^t|\boldsymbol{\theta}_{-j}^t, \mathbf{y})\ p(\boldsymbol{\theta}_{j*}^{t+1}|\boldsymbol{\theta}_{-j}^t, \mathbf{y})}\right\} = 1.$$

Thus, every proposed sample from the conditional distribution is accepted. The Gibbs sampling algorithm at each iteration can be processed then to sample from $p(\theta_j^t|\boldsymbol{\theta}_{-j}^t, \mathbf{y})$, so

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

4.2.2.3 Inference and Assessing Convergence

There are good reviews of MCMC convergence diagnostics; see, for example, Cowles and Carlin (1996), Brooks and Roberts (1998), and Sinharay (2003). The fundamental role of the MCMC algorithm is to generate Markov chains whose stationary distribution is the same as the target distribution. We need to evaluate whether the Markov chains in MCMC algorithms converge to their stationary distribution or not. Thus, if the iterations have not proceeded for long enough, the simulations might not symbolize the target distribution. Even when simulations have reached approximate convergence, early iterations still reflect the starting approximation rather than the target distribution (Gelman et al., 2014; Sinharay, 2003).

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. The trace plot demonstrates the history of a parameter value across iteration t of the chain. If the chain is stationary, it should not be revealing any long-term trends. The average value of the chain should be approximately flat. If the trace plot shows a clear pattern (e.g.,

always increasing or always decreasing) such a plot indicates that the MCMC algorithm may not have converged.

Also, we can determine the level of autocorrelation in each chain by the AutoCorrelation Function (ACF) plot. The ACF plots show the diagnostic calculation of the autocorrelation between iterative samples against different lags where lag is a function of time. It tells us how much information is offered in the Markov chain. For example, sampling 1000 iterations from a highly-correlated Markov chain yields less information about the stationary distribution than we would obtain from 1000 samples independently drawn from the stationary distribution. Also, highly autocorrelated parameter values require a large number of iterations to be able to traverse the whole sample space of the parameter. It is worth comparing the ACF plots to the trace plots because it helps to interpret the trace plots. For example, a chain traversing the sample space very slowly could be a result of high autocorrelation. We also throw away some iterations at the beginning of an MCMC run, so the chain has a chance to burn in when enter a high probability region.

4.3 A Hierarchical Mixture Model for Variable Selection

The linear mixed model fitted to data from a split-plot experiment with n responses is

$$\mathbf{y} \sim N(\beta_0 \mathbf{1}_n + \mathbf{X}\boldsymbol{\beta}, \mathbf{V}), \tag{4.5}$$

where $\mathbf{y}$ is $n \times 1$ vector of random responses, β_0 is the intercept, $\mathbf{1}_n$ is a $n \times 1$ vector of ones, $\mathbf{X}$ is the $n \times p$ model matrix without the column of the intercept, $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed effect parameters and $\mathbf{V}$ is

$$\mathbf{V} = \sigma_\epsilon^2(\mathbf{I}_n + \frac{\sigma_\gamma^2}{\sigma_\epsilon^2}\mathbf{Z}\mathbf{Z}'),$$

where $\mathbf{Z}$ is the random effect design matrix. As $\rho = \frac{\sigma_\gamma^2}{\sigma^2}$, and $\sigma^2 = \sigma_\epsilon^2 + \sigma_\gamma^2$, then $\mathbf{V}$ can be written as

$$\mathbf{V} = \sigma^2(1-\rho)\Big(\mathbf{I}_n + \frac{\rho}{1-\rho}\mathbf{Z}\mathbf{Z}'\Big). \tag{4.6}$$

We need to find the highest posterior probability of an indicator vector $\boldsymbol{\nu} = (\nu_1, \nu_2, ...\nu_p)$ such that

$$\nu_j = \begin{cases} 0 & \text{if} \quad \beta_j = 0 \\ 1 & \text{if} \quad \beta_j \neq 0 \end{cases},$$

for $j = 1, 2, ...p$. When $\nu_j = 1$ the term is assumed to be active and will be included in the model, and when $\nu_j = 0$ the term is assumed to be non-active and will not be included in the model.

Following George and McCulloch (1997), and Tan and Wu (2013), we assume that $\boldsymbol{\beta}|\sigma^2, \boldsymbol{\nu}, c \sim \mathbf{N}(\mathbf{0}_p, \sigma^2\mathbf{D}_{\boldsymbol{\nu},c})$, where $\boldsymbol{\nu}$ is the indicator vector, c is the prior variance of the slab distribution, and $\mathbf{D}_{\boldsymbol{\nu},c}$ is a diagonal matrix with the jth diagonal element $cI(\nu_j = 1) + dI(\nu_j = 0), j = 1, ..., p$. The parameters $\sigma^2, \boldsymbol{\nu}$ and c will be given prior distributions, and the parameter d is assumed to be a small fixed non-negative number because we want the spike distribution to have a smaller variance than the slab distribution. Formally the prior construction of $\boldsymbol{\beta}$ is the following:

$$\beta_j|\sigma^2, \nu_j, c \sim (1-\nu_j)\mathbf{N}(0, d\sigma^2) + \nu_j\mathbf{N}(0, c\sigma^2).$$

For every coefficient β_j, a Bernoulli variable ν_j is defined taking values 1 and 0 with probability of inclusion ω, as $p(\nu_j = 1) = \omega$ and $p(\nu_j = 0) = (1-\omega)$. Often, ν_j's are taken as independent Bernoulli (ω) random variables, where $0 < \omega < 1$. It is common to fix ω in the normal mixture, however, we shall deal with ω as a parameter to investigate different values of ω, and sample it from the Beta distribution as it will be explained in Section 4.3.2.5.

4.3.1 Prior Distributions

Following the prior distributions used by Tan and Wu (2013), we assume that the prior distribution for the fixed effects is $\boldsymbol{\beta} \sim \mathbf{N}(0, \sigma^2 \mathbf{D}_{\boldsymbol{\nu},c})$

$$p(\boldsymbol{\beta}|\sigma^2, \boldsymbol{\nu}, c) \propto |\sigma^2 \mathbf{D}_{\boldsymbol{\nu},c}|^{-1/2} \exp\Big(-\frac{1}{2} \boldsymbol{\beta}' (\sigma^2 \mathbf{D}_{\boldsymbol{\nu},c})^{-1} \boldsymbol{\beta} \Big).$$

The prior distribution for the total variance is $\sigma^2 \sim \mathbf{IG}(a, b)$,

$$p(\sigma^2) \propto (\sigma^2)^{-a-1} \exp\Big(-\frac{b}{\sigma^2} \Big).$$

For this work, we used $a = 0$ and $b = 0$ following Tan and Wu (2013) as this yields the common non-informative prior for σ^2. This prior is improper, however we will sample from the posterior distribution, which should be a proper gamma distribution.

The prior distribution for the correlation parameter is $\rho \sim \mathbf{Beta}(a', b')$ with shape parameters a', $b' > 0$. We consider $a' = b' = 2.5$ following Gilmour and Goos (2009). According to Gilmour and Goos (2009), " A $\mathbf{Beta}(a', b')$ prior distribution for a correlation parameter can be interpreted as indicating a prior point estimate of $a'/(a'+b')$, this prior information being worth $a' + b'$ observations". Our prior was selected to be centred at $2.5/(2.5 + 2.5) = 0.5$ and to be worth five observations. For an experiment with $a' + b'$ observations, the posterior distribution would give equal weight to the prior and the likelihood (Gilmour and Goos, 2009). In relation to our two examples, in the wind tunnel experiment, we have 45 observations, meaning our prior is worth 1/7th of the weight of the likelihood. In the freeze dried-coffee experiment, we have 30 observations, meaning our prior worth 1/6th of the likelihood. The prior density for ρ is

$$p(\rho) \propto \rho^{(a'-1)} (1 - \rho)^{(b'-1)}.$$

The prior distribution for the elements of the indicator vector is $\nu_j \sim \mathbf{Bernoulli}(\omega)$,

$$p(\nu_j) = \begin{cases} \omega & \text{if} \quad \nu_j = 1 \\ 1 - \omega & \text{if} \quad \nu_j = 0 \end{cases},$$

where ω is the prior probability that β_j is active following Tan and Wu (2013). The prior distribution for the elements of the probability of inclusion is $\omega \sim \mathbf{Beta}(c_0, d_0)$,

$$p(\omega) \propto \omega^{(c_0-1)} \ (1-\omega)^{(d_0-1)}.$$

Tan and Wu (2013) set $\omega = 0.25$. However, we select c_0 and d_0 such that the prior of ω has a mode $= 0.25$. The choice of $c_0 = 2$ and $d_0 = 4$ results in a prior with a mode $= 0.25$ and the upper cumulative percentile at 5% equals 0.66. Meaning a 5% chance the observations have a pdf ≥ 0.66.

In our simulation study, this choice of c_0 and d_0 is suitable for the wind tunnel experiment because this experiment uses an orthogonal design in which we estimate truly 6 active variables from a true model of 12 variables as the assumed model in Section 5.2.1. However, in the freeze-dried coffee experiment, this option of c_0 and d_0 does not allow us to detect the active effects. This is because this experiment uses a non orthogonal design in which we estimate truly 15 active variables from a true model of 20 variables as the assumed model in Section 5.2.2. Thus, we select c_0 and d_0 such that the prior of ω has a mode $= 0.50$ following Gilmour and Goos (2009) in which they analysed this dataset. For the freeze-dried coffee experiment, we assumed $c_0 = 2$ and $d_0 = 2$ as this yields in a prior of ω with a mode $= 0.50$. This is because this choice of c_0 and d_0 yields in upper cumulative percentile at 5% equals to 0.86, meaning a 5% chance the observations have a pdf ≥ 0.86.

The prior distribution for the slab variance c is a discrete uniform prior distribution with support points T = {1/4, 9/16,1, 4, 9, 16, 25} as given by Tan and Wu (2013). They found that large values of c tend to favor sparse models with large effects and in this case small effects will be missed. On the other hand, small values of c tend to favor less sparse models. Moreover, very small values of c tend to favor sparse models again. They select the support points in T such that it covers small and large values of c. The prior distribution for c is

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

4.3.2 Full Conditional Distributions

We use the prior distributions presented in Section 4.3.1 to derive the full conditional distributions.

4.3.2.1 The Conditional Distribution for $\boldsymbol{\beta}$

The likelihood of the data depends on $\boldsymbol{\beta}$, so we can derive the conditional distribution for $\boldsymbol{\beta}$ using the prior distribution $\boldsymbol{\beta}|\sigma^2, \boldsymbol{\nu}, c \sim \mathbf{N}(\mathbf{0}_p, \sigma^2\mathbf{D}_{\boldsymbol{\nu},c})$ and the likelihood for the model (1.1)

$$L(\mathbf{y}|\boldsymbol{\beta}) \propto |\mathbf{V}|^{-1/2}\exp\big[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\big].$$

Note that we standardise both $\mathbf{X}$ and $\mathbf{y}$ so the fixed effect vector $\boldsymbol{\beta}$ does not include the intercept. The conditional distribution for $\boldsymbol{\beta}$ can be expressed as:

$$\begin{aligned}
p(\boldsymbol{\beta}|\boldsymbol{\nu},\sigma^2,c,\omega,\rho,\mathbf{y}) \propto & p(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\nu},\sigma^2,c,\omega,\rho)\ \ p(\boldsymbol{\beta}|\sigma^2,\boldsymbol{\nu},c) \\
& \propto |\mathbf{V}|^{-1/2}\exp\big[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\big] \times |\sigma^2\mathbf{D}_{\boldsymbol{\nu},c}|^{-1/2} \\
& \exp\Big(-\frac{1}{2}\boldsymbol{\beta}'(\sigma^2\mathbf{D}_{\boldsymbol{\nu},c})^{-1}\boldsymbol{\beta}\Big) \\
& \propto |\mathbf{V}|^{-1/2}|\sigma^2\mathbf{D}_{\boldsymbol{\nu},c}|^{-1/2}\exp\Big[-\frac{1}{2}\Big(\boldsymbol{\beta}'(\sigma^2\mathbf{D}_{\boldsymbol{\nu},c})^{-1}\boldsymbol{\beta}\Big) \\
& -\frac{1}{2}\Big(\mathbf{y}'\mathbf{V}^{-1}\mathbf{y}-\mathbf{y}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}-\boldsymbol{\beta}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{y}+\boldsymbol{\beta}'\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\boldsymbol{\beta}\Big)\Big] \\
& \propto |\mathbf{V}|^{-1/2}|\sigma^2\mathbf{D}_{\boldsymbol{\nu},c}|^{-1/2}\exp\Big[-\frac{1}{2}\boldsymbol{\beta}'\Big((\sigma^2\mathbf{D}_{\boldsymbol{\nu},c})^{-1}+\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\Big)\boldsymbol{\beta}\Big] \\
& \exp\Big[-\frac{1}{2}(-\mathbf{y}'\mathbf{V}^{-1}\mathbf{X})\boldsymbol{\beta}\Big]\exp\Big[-\frac{1}{2}\boldsymbol{\beta}'(-\mathbf{X}'\mathbf{V}^{-1}\mathbf{y})\Big] \\
& \propto |\mathbf{V}|^{-1/2}|\sigma^2\mathbf{D}_{\boldsymbol{\nu},c}|^{-1/2}\exp\Big[-\frac{1}{2}\boldsymbol{\beta}'\Big((\sigma^2\mathbf{D}_{\boldsymbol{\nu},c})^{-1}+\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\Big)\boldsymbol{\beta}\Big] \\
& \exp\Big[\frac{1}{2}(\mathbf{y}'\mathbf{V}^{-1}\mathbf{X})\boldsymbol{\beta}+\frac{1}{2}\boldsymbol{\beta}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y})\Big] \\
& \propto |\mathbf{V}|^{-1/2}|\sigma^2\mathbf{D}_{\boldsymbol{\nu},c}|^{-1/2}\exp\Big[-\frac{1}{2}\boldsymbol{\beta}'\Big((\sigma^2\mathbf{D}_{\boldsymbol{\nu},c})^{-1}+\mathbf{X}'\mathbf{V}^{-1}\mathbf{X})\Big)\boldsymbol{\beta} \\
& +\boldsymbol{\beta}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y})\Big].
\end{aligned}$$

The key to deriving the joint posterior distribution is to rewrite the expression in the exponential part in a more convenient form. This can happen by using the multivariate

completion of squares:

$$\mathrm{U}'\mathrm{AU}-2\mathrm{U}'\boldsymbol{\alpha} = (\mathrm{U}-\mathrm{A}^{-1}\boldsymbol{\alpha})'\mathrm{A}(\mathrm{U}-\mathrm{A}^{-1}\boldsymbol{\alpha}) - \boldsymbol{\alpha}'\mathrm{A}^{-1}\boldsymbol{\alpha},$$

where A is a symmetric positive definite (hence invertible) matrix. We assume $\mathrm{U} = \boldsymbol{\beta}$, $\mathrm{A} = (\sigma^2\mathbf{D}_{\boldsymbol{\nu},c})^{-1} + \mathbf{X}'\mathbf{V}^{-1}\mathbf{X}$, and $\alpha = \mathbf{X}'\mathbf{V}^{-1}\mathbf{y}$.
The conditional distribution for $\boldsymbol{\beta}$ can be written as:

$$\begin{aligned} p(\boldsymbol{\beta}|\boldsymbol{\nu},\sigma^2,c,\omega,\rho,\mathbf{y}) &\propto |\mathbf{V}|^{-1/2}|\sigma^2\mathbf{D}_{\boldsymbol{\nu},c}|^{-1/2}\exp\Big[-\frac{1}{2}\boldsymbol{\beta}'\Big((\sigma^2\mathbf{D}_{\boldsymbol{\nu},c})^{-1}+\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\Big)\boldsymbol{\beta} - 2\boldsymbol{\beta}'(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y})\Big] \\ &\propto |\mathbf{V}|^{-1/2}|\sigma^2\mathbf{D}_{\boldsymbol{\nu},c}|^{-1/2}\exp\Big[\big[\boldsymbol{\beta}-\Big((\sigma^2\mathbf{D}_{\boldsymbol{\nu},c})^{-1}+\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\Big)^{-1}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y})\big]' \\ &\quad \big[(\sigma^2\mathbf{D}_{\boldsymbol{\nu},c})^{-1}+\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\big]\big[\boldsymbol{\beta}-\Big((\sigma^2\mathbf{D}_{\boldsymbol{\nu},c})^{-1}+\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\Big)^{-1}(\mathbf{X}'\mathbf{V}^{-1}\mathbf{y})\big]\Big] \\ &\propto \exp\Big[-\frac{1}{2}(\boldsymbol{\beta}-\boldsymbol{\beta}_*)'\mathrm{D}_*^{-1}(\boldsymbol{\beta}-\boldsymbol{\beta}_*)\Big]. \end{aligned}$$

Thus, we can sample $\boldsymbol{\beta}$ from the conditional posterior $\mathrm{N}(\boldsymbol{\beta}_*, \mathrm{D}_*)$, where

$$\boldsymbol{\beta}_* = \Big((\sigma^2\mathbf{D}_{\boldsymbol{\nu},c})^{-1}+\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\Big)^{-1}(\mathbf{X}\mathbf{V}^{-1}\mathbf{y}), \text{ and } \mathrm{D}_* = \Big((\sigma^2\mathbf{D}_{\boldsymbol{\nu},c})^{-1}+\mathbf{X}'\mathbf{V}^{-1}\mathbf{X}\Big)^{-1}. \tag{4.7}$$

4.3.2.2 The Conditional Distribution for ρ

The likelihood of the data depends on ρ, so the conditional distribution for ρ can be derived by

$$\begin{aligned} p(\rho|\boldsymbol{\beta},\boldsymbol{\nu},\sigma^2,c,\omega,\mathbf{y}) &\propto p(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\nu},\sigma^2,c,\omega,\rho)\ \ p(\rho) \\ &\propto |\mathbf{V}|^{-1/2}\exp\big[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\big] \times \rho^{(a'-1)}(1-\rho)^{(b'-1)}. \end{aligned}$$

We note here that the likelihood depends on ρ through $\mathbf{V}$ as in (4.6), so we can express $\mathbf{V}$ as a function of ρ,

$$\mathrm{V} = \sigma^2\ (1-\rho)\ (\mathrm{I}_n + \frac{\rho}{1-\rho}\mathbf{Z}\mathbf{Z}').$$

The conditional distribution for ρ is a non-standard distribution that cannot be sampled directly. Therefore, we use the Metropolis-Hastings (M-H) rejection sampling. As explained in Section 4.2.2.1, our correlation parameter is $\rho \in (0,1]$, and has a prior $\beta(a',b')$.

We apply the Random-Walk Metropolis-Hastings algorithm, and select a proposal distribution of log-normal distribution for the variance ratio η where $\eta = f(\rho) = \frac{\rho}{1-\rho}$ with a mean equal to the current value of η^t at iteration t and variance s^2. The choice of s^2 affects the jumping rule in the random walk proposal distribution. As we have one parameter to be updated in the random walk algorithm which is ρ, we follow Gelman et al. (2014) and Gelman et al. (1996) to set $s^2 = g^2\Sigma$. The most efficient jump has a scale $g \approx 2.4/\sqrt{h}$ where h is the number of parameters which will be updated. In this work, we set $g = 2.4$ and $h = 1$ following Gelman et al. (1996), and we set $\Sigma = 100$ as this yields an appropriate acceptance rate associated with the independent sampler of the ACF plot. Thus, $\eta \in (0, \infty)$ and

$$g(\eta) = \frac{1}{\eta\sqrt{2\pi s^2}}\exp\big[-\frac{1}{2s^2}(\ln\eta - \eta^t)^2\big]$$

We can use $\eta = \frac{\rho}{1-\rho}$ as a transformation function between η and ρ as $\rho = \frac{\eta}{1+\eta}$ and the Jacobian function of ρ is $J(\rho) = \frac{d\eta}{d\rho} = \frac{1}{(1-\rho)^2}$.

We draw a proposal value η_* from a log-normal(η^t, s^2) distribution, and the probability of accepting or rejecting η_* is the minimum of 1 and the ratio r where r is

$$r = \frac{p(\rho_*|\text{all})}{p(\rho^t|\text{all})} \times \frac{q(\eta^t|\eta_*)}{q(\eta_*|\eta^t)},$$

which is equivalent to

$$r = \frac{p(\rho_*|\text{all})}{p(\rho^t|\text{all})} \times \frac{q(\rho^t|\rho_*)J(\rho^t)}{q(\rho_*|\rho^t)J(\rho_*)}.$$

Our proposal ratio is

$$\frac{q(\eta^t|\eta_*)}{q(\eta_*|\eta^t)} = \frac{\eta_*\exp\big[-\frac{1}{2s^2}(\ln\eta^t - \eta_*)^2\big]}{\eta^t\exp\big[-\frac{1}{2s^2}(\ln\eta_* - \eta^t)^2\big]},$$

which is equivalent to

$$\frac{q(\rho^t|\rho_*)J(\rho^t)}{q(\rho_*|\rho^t)J(\rho_*)} = \frac{(\frac{\rho_*}{1-\rho_*})\exp\big[-\frac{1}{s^2}\big(\ln(\frac{\rho^t}{1-\rho^t}) - (\frac{\rho_*}{1-\rho_*})\big)^2\big] \times |\frac{1}{(1-\rho^t)^2}|}{(\frac{\rho^t}{1-\rho^t})\exp\big[-\frac{1}{s^2}\big(\ln(\frac{\rho_*}{1-\rho_*}) - (\frac{\rho^t}{1-\rho^t})\big)^2\big] \times |\frac{1}{(1-\rho_*)^2}|}.$$

The ratio r can be expressed as

$$\begin{aligned} r =& \frac{|\mathbf{V}(\rho_*)|^{-1/2}\exp\big[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}(\rho_*)^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\big]\times(\rho_*)^{(a'-1)}(1-\rho_*)^{(b'-1)}}{|\mathbf{V}(\rho^t)|^{-1/2}\exp\big[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}(\rho^t)^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\big]\times(\rho^t)^{(a'-1)}(1-\rho^t)^{(b'-1)}} \\ &\times\frac{\rho_*(1-\rho^t)\exp\big[-\frac{1}{s^2}\big(\ln(\frac{\rho^t}{1-\rho^t})-(\frac{\rho_*}{1-\rho_*})\big)^2\big]\ |(1-\rho^t)^{-2}|}{\rho^t(1-\rho_*)\exp\big[-\frac{1}{s^2}\big(\ln(\frac{\rho_*}{1-\rho_*})-(\frac{\rho^t}{1-\rho^t})\big)^2\big]\ |(1-\rho_*)^{-2}|}. \end{aligned} \tag{4.8}$$

Where $\mathrm{V}(\rho_*)=\sigma^2\ (1-\rho_*)\ (\mathrm{I}_n+\frac{\rho_*}{1-\rho_*}\mathbf{ZZ}')$, and $\mathrm{V}(\rho^t)=\sigma^2\ (1-\rho^t)\ (\mathrm{I}_n+\frac{\rho^t}{1-\rho^t}\mathbf{ZZ}')$.

4.3.2.3 The Conditional Distribution for σ^2

The likelihood of the data depends on σ^2, so we can express the conditional distribution of σ^2 as

$$\begin{aligned} p(\sigma^2|\boldsymbol{\beta},\rho,\boldsymbol{\nu},c,\omega,\mathbf{y}) &\propto p(\mathbf{y}|\boldsymbol{\beta},\boldsymbol{\nu},\sigma^2,c,\omega,\rho)\ \ p(\sigma^2) \\ &\propto |\mathbf{V}|^{-1/2}\exp\big[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\mathbf{V}^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\big]\times(\sigma^2)^{-a-1}\exp\big(-\frac{b}{\sigma^2}\big). \end{aligned}$$

We know that $\mathrm{V}=\sigma^2\ (1-\rho)\ (\mathrm{I}_n+\frac{\rho}{1-\rho}\mathbf{ZZ}')$, so the conditional posterior for σ^2 can be written as

$$\begin{aligned} p(\sigma^2|\mathbf{y},...) \propto& |(1-\rho)\ \sigma^2(\mathrm{I}_n+\frac{\rho}{1-\rho}\mathbf{ZZ}')|^{-1/2}(\sigma^2)^{-a-1}\ \times\ \ \exp\ \big(-\frac{b}{\sigma^2}\big) \\ &\exp\big[-\frac{1}{2}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\Big((1-\rho)\ \sigma^2\ (\mathrm{I}_n+\frac{\rho}{1-\rho}\mathbf{ZZ}')\Big)^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})\big] \\ &\propto (\sigma^2)^{-(a+\frac{n}{2})-1}\exp\Big(-\frac{1}{\sigma^2}\Big[\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\Big((1-\rho)(\mathrm{I}_n+\frac{\rho}{1-\rho}\mathbf{ZZ}')\Big)^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2}+b\Big]\Big). \end{aligned}$$

This is the inverse gamma distribution with a shape parameter a^* and a scale parameter b^* such that

$$a^*=a+\frac{n}{2},\quad\text{and}\quad b^*=\frac{(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})'\Big((1-\rho)(\mathrm{I}_n+\frac{\rho}{1-\rho}\mathbf{ZZ}')\Big)^{-1}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta})}{2}+b. \tag{4.9}$$

4.3.2.4 The Conditional Distribution for $\boldsymbol{\nu}$

The indicator vector can be drawn conditionally on the regressor coefficient and computation of the marginal likelihood is not required. The prior probabilities for ν_j are

$$p(\nu_j) = \begin{cases} \omega & \text{if} \quad \nu_j = 1 \\ 1-\omega & \text{if} \quad \nu_j = 0 \end{cases},$$

where ω is the prior probability that β_j is active. The joint conditional posterior distribution for $\boldsymbol{\nu}$ has mass function

$$\begin{aligned} p(\boldsymbol{\nu}|\boldsymbol{\beta}, \mathbf{y}) &\propto p(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\nu})\; p(\boldsymbol{\beta}, \boldsymbol{\nu}) \\ &= p(\mathbf{y}|\boldsymbol{\beta})\; p(\boldsymbol{\beta}, \boldsymbol{\nu}) \\ &\propto p(\boldsymbol{\beta}, \boldsymbol{\nu}) \\ &= p(\boldsymbol{\beta}|\boldsymbol{\nu})\; p(\boldsymbol{\nu}). \end{aligned}$$

The conditional density for $\boldsymbol{\beta}$ given $\boldsymbol{\nu}$ is

$$\begin{aligned} p(\boldsymbol{\beta}|\boldsymbol{\nu}) &\propto |\sigma^2 \mathrm{diag}\big[c\mathrm{I}(\nu_j = 1) + d\mathrm{I}(\nu_j = 0)\big]|^{-1/2} \\ &\times \exp\Big[-\frac{1}{2}\boldsymbol{\beta}'\Big(\sigma^2 \mathrm{diag}\big[c\mathrm{I}(\nu_j = 1) + d\mathrm{I}(\nu_j = 0)\big]\Big)^{-1}\boldsymbol{\beta}\Big]. \end{aligned}$$

The conditional distribution for the jth component given ν_j is

$$\begin{aligned} p(\beta_j|\nu_j) &\propto |\sigma^2\big[c\mathrm{I}(\nu_j = 1) + d\mathrm{I}(\nu_j = 0)\big]|^{-1/2} \\ &\times \exp\Big[-\frac{\beta_j^2}{2\sigma^2\big[c\mathrm{I}(\nu_j = 1) + d\mathrm{I}(\nu_j = 0)\big]}\Big]. \end{aligned}$$

The conditional posterior probabilities for ν_j are therefore

$$\begin{aligned} p(\nu_j = 1|\beta_j, \mathbf{y}) &= p(\nu_j = 1)\; p(\beta_j|\nu_j = 1) \\ &\propto \omega\; |c\;\sigma^2|^{-1/2} \exp\Big[-\frac{\beta_j^2}{2\;c\;\sigma^2}\Big], \end{aligned} \tag{4.10}$$

and

$$\begin{aligned} p(\nu_j = 0|\beta_j, \mathbf{y}) &= p(\nu_j = 0)\ p(\beta_j|\nu_j = 0) \\ &\propto (1-\omega)\ |d\ \sigma^2|^{-1/2}\exp\Big[-\frac{\beta_j^2}{2\ d\ \sigma^2}\Big]. \end{aligned} \tag{4.10}$$

4.3.2.5 The Conditional Distribution for ω

The probability of inclusion ω can be drawn conditionally on the indicator and computation of the marginal likelihood is not required. Hence the conditional distribution for ω is

$$\begin{aligned} p(\omega|\boldsymbol{\nu}, \sigma^2, c, \boldsymbol{\beta}, \rho, \mathbf{y}) &\propto p(\mathbf{y}|\omega, \boldsymbol{\nu}, \sigma^2, c, \boldsymbol{\beta}, \rho)\ p(\omega, \boldsymbol{\nu}, \sigma^2, c, \boldsymbol{\beta}, \rho) \\ &= p(\mathbf{y}|\boldsymbol{\nu}, \sigma^2, c, \boldsymbol{\beta}, \rho)\ p(\omega, \boldsymbol{\nu}) \\ &\propto p(\omega, \boldsymbol{\nu}) \\ &= p(\boldsymbol{\nu}|\omega)\ p(\omega) \\ &\propto \omega^{\sum_{j=1}^{p}\nu_j}(1-\omega)^{p-\sum_{j=1}^{p}\nu_j} \times \omega^{(c_0-1)}(1-\omega)^{(d_0-1)} \\ &\propto \omega^{c_0+\sum_{j=1}^{p}\nu_j-1}(1-\omega)^{p-\sum_{j=1}^{p}\nu_j+d_0-1}. \end{aligned}$$

Hence,

$$\omega|\boldsymbol{\nu} \sim Beta\Big(c_0 + \sum_{j=1}^{p}\nu_j,\ p - \sum_{j=1}^{p}\nu_j + d_0\Big),\ \text{ where }\ p - \sum_{j=1}^{p}\nu_j = \sum_{j=1}^{p}\mathrm{I}(\nu_j = 0). \tag{4.11}$$

4.3.2.6 The Conditional Distribution for c

The prior distribution for c is a discrete uniform distribution with support points T = {1/4, 9/16,1, 4, 9, 16, 25}, and it can be drawn conditionally on the regressor coefficient. The computation of the marginal likelihood is not required. Hence, the conditional

distribution for c

$$
\begin{aligned}
p(c|\boldsymbol{\beta},\sigma^2,\boldsymbol{\nu},\omega,\rho,\mathbf{y}) &\propto p(\mathbf{y}|\boldsymbol{\beta},\sigma^2,\boldsymbol{\nu},c)\; p(\boldsymbol{\beta},\boldsymbol{\nu},\sigma^2,c)\\
&= p(\mathbf{y}|\boldsymbol{\beta},\sigma^2)\; p(\boldsymbol{\beta},\boldsymbol{\nu},\sigma^2,c)\\
&\propto p(\boldsymbol{\beta},\sigma^2,\boldsymbol{\nu},c)\\
&= p(\boldsymbol{\beta}|\sigma^2,\boldsymbol{\nu},c)\; p(c)\\
&\propto \frac{1}{7}|\sigma^2\mathrm{diag}\big[c\mathrm{I}(\nu_j=1)+d\mathrm{I}(\nu_j=0)\big]|^{-1/2}\\
&\times \exp\Big[-\frac{1}{2}\boldsymbol{\beta}'\Big(\sigma^2\mathrm{diag}\big[c\mathrm{I}(\nu_j=1)+d\mathrm{I}(\nu_j=0)\big]\Big)^{-1}\boldsymbol{\beta}\Big]\\
&\propto \frac{1}{7}\Big[\prod_{j=1}^{p}\big[c\mathrm{I}(\nu_j=1)+d\mathrm{I}(\nu_j=0)\big]\Big]^{-1/2}\\
&\times \exp\Big[-\frac{1}{2c}(\boldsymbol{\beta}'\mathrm{I}_{\sum_j \nu_j=1}\boldsymbol{\beta})\Big]\times\exp\Big[-\frac{1}{2d}(\boldsymbol{\beta}'\mathrm{I}_{\sum_j \nu_j=0}\boldsymbol{\beta})\Big]\\
&\propto \frac{1}{7}c^{\frac{-\sum_{j=1}^{p}\nu_j}{2}}+\frac{1}{7}d^{\frac{-\sum_{j=1}^{p}(1-\nu_j)}{2}}\times\exp\Big[-\frac{1}{2c}(\boldsymbol{\beta}'\mathrm{I}_{\sum_j \nu_j=1}\boldsymbol{\beta})\Big]\\
&\times \exp\Big[-\frac{1}{2d}(\boldsymbol{\beta}'\mathrm{I}_{\sum_j \nu_j=0}\boldsymbol{\beta})\Big]\\
&\propto \frac{1}{7}c^{\frac{-\sum_{j=1}^{p}\nu_j}{2}}\times\exp\Big[-\frac{1}{2c}(\boldsymbol{\beta}'\mathrm{I}_{\sum_j \nu_j=1}\boldsymbol{\beta})\Big]\\
&\propto \frac{1}{7}c^{\frac{-\sum_{j=1}^{p}\nu_j}{2}}\times\exp\Big[-\frac{1}{2c}(\boldsymbol{\beta}'\boldsymbol{\beta})\Big].
\end{aligned}
$$

Therefore,

$$
p(c|\mathbf{y},...)\propto\begin{cases}\frac{1}{7}c^{\frac{-\sum_{j=1}^{p}\nu_j}{2}}\times\exp\Big[-\frac{1}{2c}(\boldsymbol{\beta}'\boldsymbol{\beta})\Big] & \text{if}\quad c\in\mathrm{T}\\ 0 & \text{otherwise}\end{cases}.
$$

Then, the posterior probabilities of the conditional distribution $p(c|\mathbf{y},...)$ are

$$
p(c=\frac{1}{4}|\mathbf{y},...)\propto\frac{1}{7}\big(\frac{1}{4}\big)^{\frac{-\sum_{j=1}^{p}\nu_j}{2}}\times\exp\Big[-\frac{1}{2\big(\frac{1}{4}\big)}(\boldsymbol{\beta}'\boldsymbol{\beta})\Big],
$$

$$
p(c=\frac{9}{16}|\mathbf{y},...)\propto\frac{1}{7}\big(\frac{9}{16}\big)^{\frac{-\sum_{j=1}^{p}\nu_j}{2}}\times\exp\Big[-\frac{1}{2\big(\frac{9}{16}\big)}(\boldsymbol{\beta}'\boldsymbol{\beta})\Big],
$$

$$
p(c=1|\mathbf{y},...)\propto\frac{1}{7}\big(1\big)^{\frac{-\sum_{j=1}^{p}\nu_j}{2}}\times\exp\Big[-\frac{1}{2\big(1\big)}(\boldsymbol{\beta}'\boldsymbol{\beta})\Big],
$$

$$p(c=4|\mathbf{y},...) \propto \frac{1}{7}(4)^{\frac{-\sum_{j=1}^{p}\nu_j}{2}} \times \exp\Big[-\frac{1}{2(4)}(\boldsymbol{\beta}'\boldsymbol{\beta})\Big],$$

$$p(c=9|\mathbf{y},...) \propto \frac{1}{7}(9)^{\frac{-\sum_{j=1}^{p}\nu_j}{2}} \times \exp\Big[-\frac{1}{2(9)}(\boldsymbol{\beta}'\boldsymbol{\beta})\Big],$$

$$p(c=16|\mathbf{y},...) \propto \frac{1}{7}(16)^{\frac{-\sum_{j=1}^{p}\nu_j}{2}} \times \exp\Big[-\frac{1}{2(16)}(\boldsymbol{\beta}'\boldsymbol{\beta})\Big],$$

and

$$p(c=25|\mathbf{y},...) \propto \frac{1}{7}(25)^{\frac{-\sum_{j=1}^{p}\nu_j}{2}} \times \exp\Big[-\frac{1}{2(25)}(\boldsymbol{\beta}'\boldsymbol{\beta})\Big].$$

The conditional posterior $p(c|\mathbf{y},...)$ can be written as

$$p(c|\mathbf{y},...) = \begin{cases} \dfrac{c^{-\frac{\sum_{j=1}^{p}\nu_j}{2}}\exp[-\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{2c}]}{\sum_{c\in T} c^{-\frac{\sum_{j=1}^{p}\nu_j}{2}}\exp[-\frac{\boldsymbol{\beta}'\boldsymbol{\beta}}{2c}]} & \text{if } c \in \text{T} \\ 0 \quad \text{otherwise} \end{cases} . \tag{4.12}$$

4.4 Stochastic Search Variable Selection for Split-Plot Design (SSVS-SPD)

We adapt the SSVS for the analysis of data from split-plot designs by taking into account the two types of factors, i.e. the whole-plot factors and the subplot factors which expected to have different effect sizes for the two strata in split-plot design (Jones and Nachtsheim, 2009) . This approach can be reported as the Stochastic Search Variable Selection for Split-Plot Design (SSVS-SPD). While the SSVS samples all parameters from one slab variance posterior distribution, the SSVS-SPD samples the whole-plot parameters and the subplot parameters from two different slab variance posterior distributions given that the whole-plot and the subplot effects might have different sizes. The computational algorithms for both SSVS and SSVS-SPD will be given in Sections 4.5.1 and 4.5.2 respectively.

4.5 Bayesian Variable Selection Algorithms

In this section, we introduce the computational algorithms which we use in the Bayesian analysis for variable selection using the SSVS, and the SSVS-SPD. In this work, we choose an asymmetric proposal distribution, the log-normal density. We apply the Metropolis-Hastings algorithms to sample the variance ratio η where $\eta = f(\rho) = \frac{\rho}{1-\rho}$, and ρ is the correlation parameter. This is because of the fact that in our experiments, observations from different subplots within the same wholeplot are positively correlated as $\rho = \frac{\sigma_\gamma^2}{\sigma_\epsilon^2+\sigma_\gamma^2}$; also observations from different wholeplots are independent (Tan and Wu, 2013).

4.5.1 The Stochastic Search Variable Selection (SSVS) Algorithm

We process the MCMC estimation of the parameters $\boldsymbol{\beta}, \rho, \sigma^2, \boldsymbol{\nu}, \omega,$ and c. We use the priors of all these parameters as in Section 4.3.1. The following Metropolis-Hastings within Gibbs sampling algorithm can be implemented. Let $\mathbf{y}$ be the $n \times 1$ vector of random responses, $\mathbf{X}$ is the $n \times p$ model matrix without the column of the intercept, $\boldsymbol{\beta}$ is the $p \times 1$ vector of fixed effect parameters, where p is the number of fixed effect parameters that need to be estimated. We set initial values for the parameters as $\boldsymbol{\beta}^{(0)} = \mathbf{1}_p$, $\boldsymbol{\nu}^{(0)} = \mathbf{1}_p$, $\rho^{(0)} = 0.5$, $\sigma^{2(0)} = 20$, $c^{(0)} = 1$, $\omega^{(0)} = 0.5$, $d = 0.001$. Starting at the tth iteration such that $t = 1, 2, ..., its$ where $its = 10000$, and setting $j = 1, 2, ..., p$, the sampling algorithm is:

1. For $j = 1, 2, ..., p$, sample $\nu_j^{(t)}$ of the indicator vector $\boldsymbol{\nu}^{(t)}$ using (4.10) for $\beta_j^{(t-1)}$, $c^{(t-1)}$, $\sigma^{2(t-1)}$, and $\omega^{(t-1)}$.

2. Sample the mixture weight $\omega^{(t)}$ using (4.11) for $\boldsymbol{\nu}^{(t)}$.

3. Sample the regressor coefficients $\boldsymbol{\beta}^{(t)}$ using (4.7) for $\mathbf{X}$, $\mathbf{y}$, $\mathbf{V}^{(t-1)}$, $\mathbf{D}^{(t-1)}$, $c^{(t-1)}$, $\boldsymbol{\nu}^{(t)}$, and $\sigma^{2(t-1)}$.

4. Sample the total variance $\sigma^{2(t)}$ using (4.9) for $\mathbf{X}$, $\mathbf{y}$, $\mathbf{Z}$, $\boldsymbol{\beta}^{(t)}$, and $\rho^{(t-1)}$.

5. (a) Sample $\rho_*^{(t)}$ from $\beta(a', b')$.

   (b) Calculate $\alpha^{(t)}$ using (4.8) for $\mathbf{X}$, $\mathbf{y}$, $\mathbf{V}(\rho_*^{(t)})$, $\mathbf{V}(\rho^{(t-1)})$, and $\boldsymbol{\beta}^{(t)}$.

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

(d) If $\alpha^{(t)} > u^{(t)}$, then set $\rho^{(t)} = \rho_*^{(t)}$, otherwise set $\rho^{(t)} = \rho^{(t-1)}$.

6. Sample $c^{(t)}$ from the set T with probability mass function given in (4.12) for $\boldsymbol{\beta}^{(t)}$, and $\boldsymbol{\nu}^{(t)}$.

4.5.2 The Stochastic Search Variable Selection for Split-Plot Design (SSVS-SPD) Algorithm

While the SSVS samples all parameters from one slab variance posterior distribution, the SSVS-SPD samples the whole-plot parameters and the subplot parameters from two different slab variance posterior distributions. We use the same priors as in the SSVS for all the parameters of interest as in Section 4.3.1. Basically, the SSVS-SPD can be seen as running the SSVS twice in one process, one for subplot factors and the other one for whole-plot factors. The algorithm can be explained as follows:

We process the MCMC estimation of the parameters $\boldsymbol{\beta}, \rho, \sigma^2, \boldsymbol{\nu}, \omega,$ and c. The following Metropolis-Hastings within Gibbs sampling algorithm can be implemented. Let $\mathbf{y}$ be the $n \times 1$ vector of random responses, $\mathbf{X}$ is the $n \times p$ model matrix without the column of the intercept, $\mathbf{X.S}$ is the $n \times p_s$ model matrix for subplot factors where p_s is the number of subplot fixed effect parameters.

Also, $\mathbf{X.W}$ is the $n \times p_w$ model matrix for whole-plot factors where p_w is the number of whole-plot fixed effect parameters. The $\boldsymbol{\beta} = (\boldsymbol{\beta}_s, \boldsymbol{\beta}_w)$ is the $p \times 1$ vector of fixed effect parameters, where p is the number of fixed effect parameters that need to be estimated, $\boldsymbol{\beta}_s$ is the $p_s \times 1$ subplot effect parameters, and $\boldsymbol{\beta}_w$ is the $p_w \times 1$ whole-plot effect parameters.

We set initial values for the parameters as $\boldsymbol{\beta}_s^{(0)} = \mathbf{1}_{p_s}$, $\boldsymbol{\beta}_w^{(0)} = \mathbf{1}_{p_w}$. The initial values for the indicator vectors for the subplot factor $\boldsymbol{\nu_s}$ and the whole-plot factor $\boldsymbol{\nu_w}$ are $\boldsymbol{\nu_s}^{(0)} = \mathbf{1}_{ps}$ and $\boldsymbol{\nu_w}^{(0)} = \mathbf{1}_{pw}$. Also, $\rho^{(0)} = 0.5$, $\sigma^{2(0)} = 20$, and $d = 0.001$. The initial values for the slab variance for the subplot factors c_s and for the slab variance for the whole-plot factors c_w are $c_s^{(0)} = c_w^{(0)} = 1$. Finally, the initial weights for the subplot factors ω_s and for the whole-plot factors ω_w are $\omega_s^{(0)} = \omega_w^{(0)} = 0.5$.

Starting at the tth iteration such that $t = 1, 2, ..., its$ where $its = 10000$, and setting $j = 1, 2, ..., p_s$ and $k = 1, 2, ..., p_w$, the sampling algorithm is:

1. For $j = 1, 2, ..., p_s$, and $k = 1, 2, ..., p_w$ sample $\nu_{sj}^{(t)}$ and $\nu_{wk}^{(t)}$ of the indicator vectors $\boldsymbol{\nu}_s^{(t)}$ and $\boldsymbol{\nu}_w^{(t)}$ using (4.10) for $\beta_{sj}^{(t-1)}$, $\beta_{wk}^{(t-1)}$, $c_s^{(t-1)}$, $c_w^{(t-1)}$, $\sigma^{2(t-1)}$, $\omega_s^{(t-1)}$, and $\omega_w^{(t-1)}$.
2. Allocate $\boldsymbol{\nu}^{(t)} = (\boldsymbol{\nu}_s^{(t)}, \boldsymbol{\nu}_w^{(t)})$.
3. Sample the mixture weights $\omega_s^{(t)}$ and $\omega_w^{(t)}$ using (4.11) for $\boldsymbol{\nu}_s^{(t)}$ and $\boldsymbol{\nu}_w^{(t)}$.
4. Allocate $\omega^{(t)} = (\omega_s^{(t)}, \omega_w^{(t)})$.
5. Sample the regressor coefficients $\boldsymbol{\beta}_s^{(t)}$ and $\boldsymbol{\beta}_w^{(t)}$ using (4.7) for $\mathbf{X}$, $\mathbf{y}$, $\mathbf{V}^{(t-1)}$, $\mathbf{D}_s^{(t-1)}$, $\mathbf{D}_w^{(t-1)}$, $c_s^{(t-1)}$, $c_w^{(t-1)}$, $\boldsymbol{\nu}_s^{(t)}$, $\boldsymbol{\nu}_w^{(t)}$, and $\sigma^{2(t-1)}$. Where the $\mathbf{D}_s$ is a diagonal matrix with the jth diagonal element $c_s^{(t-1)} I(\nu_{sj} = 1) + dI(\nu_{sj} = 0)$, and $\mathbf{D}_w$ is a diagonal matrix with the kth diagonal element $c_w^{(t-1)} I(\nu_{wk} = 1) + dI(\nu_{wk} = 0)$.
6. Allocate $\boldsymbol{\beta}^{(t)} = (\boldsymbol{\beta}_s^{(t)}, \boldsymbol{\beta}_w^{(t)})$ and $\mathbf{D}^{(t)} = \text{diag}(\mathbf{D}_s^{(t)}, \mathbf{D}_w^{(t)})$.
7. Sample the total variance $\sigma^{2(t)}$ using (4.9) for $\mathbf{X}$, $\mathbf{y}$, $\mathbf{Z}$, $\boldsymbol{\beta}^{(t)}$, and $\rho^{(t-1)}$.
8. (a) Sample $\rho_*^{(t)}$ from $\beta(a, b)$.
   (b) Calculate $\alpha^{(t)}$ using (4.8) for $\mathbf{X}$, $\mathbf{y}$, $\mathbf{V}(\rho_*^{(t)})$, $\mathbf{V}(\rho^{(t-1)})$, and $\boldsymbol{\beta}^{(t)}$.
   (c) Sample $u^{(t)}$ from $U(0, 1)$.
   (d) If $\alpha^{(t)} > u^{(t)}$, then set $\rho^{(t)} = \rho_*^{(t)}$, otherwise set $\rho^{(t)} = \rho^{(t-1)}$.
9. Sample $c_s^{(t)}$ and $c_w^{(t)}$ from the set T with probability mass function given in (4.12) for $\boldsymbol{\beta}_s^{(t)}$, $\boldsymbol{\beta}_w^{(t)}$, $\boldsymbol{\nu}_s^{(t)}$ and $\boldsymbol{\nu}_w^{(t)}$.
10. Allocate $c^{(t)} = (c_s^{(t)}, c_w^{(t)})$.

4.6 Discussion

This chapter discussed Bayesian variable selection methods for models from split-plot designs using samples from Metropolis-Hastings within Gibbs sampling algorithm. Bayesian variable selection is easy to implement due to the improvement in computing via MCMC sampling. In Section 4.2, we described the Bayesian methodology by introducing the

Bayesian framework, and explaining Markov Chain Monte Carlo (MCMC) sampling. The Metropolis-Hastings within Gibbs sampling in Section 4.2.2 was used to draw dependent samples from the full conditional distributions which were explained in Section 4.3.2. We assumed a mixture normal distribution as a prior for the fixed effect parameters. We jointly sampled the indicator vector and the fixed effect parameters within our sampler. In Section 4.5.1, we discussed the computational algorithm for the Stochastic Search Variable Selection (SSVS) in linear mixed models. We extended the computational algorithm of SSVS to fit models from split-plot design by introducing the algorithm of the Stochastic Search Variable Selection for Split-plot Design (SSVS-SPD) in Section 4.5.2. The key difference between the SSVS and the SSVS-SPD is that, instead of introducing one slab variance prior distribution (c) in SSVS, we introduced two slab variance prior distributions, one for subplot factors (c_s) and the other for whole-plot factors (c_w). The motivation of this extension is that we have two different levels of the experimental units one for the whole plots and the other for subplots in the split-plot design; see, for example, Jones and Nachtsheim (2009). The slab prior distributions c_s and c_w are assumed to have the same prior in this work; however, they might have different posterior distributions. In the next chapter, we will show how this extension in SSVS-SPD improved the variable selection for models in split-plot design.

Chapter 5

Application of the Bayesian Methods for Variable Selection from Split-Plot Experiments

This chapter is a practical demonstration of the methods described in Chapter 4. It demonstrates the usefulness of Bayesian variable selection for split-plot experiments. We will summarise the numerical results for both motivating examples explained in Section 1.2.1 and Section 1.2.2. We focus on two different split-plot experiments. In the first example, the wind tunnel experiment, the design used is an orthogonal (see Figure 3.1 and Figure 3.2). In the second example, the freeze-dried coffee experiment, the design used is clearly non-orthogonal (see Figure 3.3 and Figure 3.4). We will study the performance of our approach using these two different types of designs. In Section 5.1, we will present the results from the real dataset from both experiments, while the simulation results will be summarised in Section 5.2. In Section 5.3, we will provide a comparison of the frequentist and Bayesian methods. Finally, Section 5.4 will provide a discussion of this chapter.

5.1 Practical Examples

The real datasets for the wind tunnel experiment and the freeze-dried coffee experiment have been used to apply the Bayesian variable selection approach. The quadratic model for the wind tunnel experiment involves four main variables (w_1, w_2, s_1, s_2), six of two-factor interaction variables $(s_1s_2, s_1w_1, s_1w_2, s_2w_1, s_2w_2, w_1w_2)$, and two quadratic variables (w_1^2, s_1^2). Also, the full quadratic model for the freeze-dried coffee experiment involves five main variables (w,s_1, s_2, s_3, s_4), 10 of two-factor interaction variables (ws_1, ws_2, ws_3, ws_4, s_1s_2, s_1s_3, s_1s_4, s_2s_3, s_2s_4, s_3s_4), and five quadratic variables (w^2, s_1^2, s_2^2, s_3^2, s_4^2).

We used the prior distributions in Section 4.3.1. Following Barbieri et al. (2004), in a Bayesian framework the final model could be the median probability model consisting of those variables whose posterior inclusion probability $p\ (\nu_j = 1|y) \geq 0.5$. The posterior probability of parameter $\beta_j, j = 1, \ldots, p$, being active is approximated by

$$\sum_{q=1}^{its} \frac{\nu_j^{(q)}}{its}, \tag{5.1}$$

where ${\nu_j}^{(q)}$ is ν_j sampled at iteration $q = 1, \ldots, its$ of the Metropolis-Hastings within Gibbs sampling algorithm.

5.1.1 Analysis of the Wind Tunnel Experiment

The wind tunnel experiment has been discussed in Section 1.2.1. In this section, we summarise the results of applying the Bayesian variable selection for these data. Our maximal model is the quadratic model. In this design w_1^2 and w_2^2 are fully correlated to each other. The s_1^2 and s_2^2 are also fully correlated to each other. Therefore, only one of the quadratic whole-plot effects and only one of the quadratic subplot effects could be estimated. We use this model as we would like to study a challenging case where we include second-order factors. In Table 5.1, the estimates of the 12 parameters of the quadratic model have been reported using the Stochastic Search Variable Selection (SSVS) and Stochastic Search Variable Selection for Split-plot Design (SSVS-SPD) explained in Sections 4.5.1 and 4.5.2 for response y_3 which is displayed in Table 1.3.

Figure 5.1 shows a comparison between the (a) SSVS and (b) SSVS-SPD with respect to the resulting approximate posterior probability for the wind tunnel experiment. The four main parameters $\beta_{s_1}, \beta_{s_2}, \beta_{w_1}$, and β_{w_2} have the highest posterior probability of being active by both SSVS and SSVS-SPD. This indicates that the four associated variables to these terms play a significant role in this experiment. Followed by these terms, we find that the parameter $\beta_{s_2w_2}$ has an approximate posterior probability of about 0.70 by both SSVS and SSVS-SPD. Moreover, the SSVS tends to consider the $\beta_{s_1^2}$ to be significant at an approximate posterior probability of 0.60 while the SSVS-SPD yielded an approximate posterior probability of about 0.50. The last term which has an approximate posterior probability of 0.50 by both SSVS and SSVS-SPD is $\beta_{s_1s_2}$. The Bayesian analysis for the real data of the wind tunnel experiment yielded seven significant variables which is similar to the resulting final model by Simpson et al. (2004). The rest of the terms seem to have a low approximate posterior probability of being active by both SSVS and SSVS-SPD which is a similar analysis of the frequentist methods that were used in Chapter 3. In summary, with respect to the final model from the Bayesian methods, it seems that the variables s_1, s_2, w_1, w_2 followed by s_2w_2 would have the most significant role in this experiment. In addition, compared to the frequentist analysis in Chapter 3, we found that there is agreement between the frequentist and the Bayesian analysis with regard to including those five variables to the final model.

Table 5.1 represents the posterior means of the coefficients and standard deviations by both the SSVS and SSVS-SPD methods for the wind tunnel experiment. Table 5.2 shows the posterior mean of the correlation $\hat{\rho}$ and the posterior mean of the total variance $\hat{\sigma}^2$ for both SSVS and SSVS-SPD methods in the wind tuunel experiment. Figure E.1 and Figure E.2 in Appendix E show (a) the trace and (b) the ACF plots for the Markov chain using the Metropolis-Hastings within Gibbs sampling algorithm used by SSVS and SSVS-SPD to sample $\boldsymbol{\beta}$. The trace plots show that 10000 iterations are enough to achieve a reasonable approximation to the posterior model probability as the sampling of the chains converge to the posterior distribution. Furthermore, the ACF plots indicate an independent sampling for $\beta_j, j = 1, \ldots, 12$. Also, Figure E.3 and Figure E.4 represent (a) the trace and (b) the ACF plots for the Markov chain formed by sampling the correlation ρ and the total variance σ^2 by both SSVS and SSVS-SPD. The trace plots in

	SSVS	SSVS-SPD
β_{s_1}	-0.0112 (0.0003)	-0.0112 (0.0003)
β_{s_2}	-0.0047 (0.0003)	-0.0047 (0.0003)
β_{w_1}	0.0063 (0.0009)	0.0068 (0.0011)
β_{w_2}	0.0066 (0.0009)	0.0071 (0.0009)
$\beta_{s_1 s_2}$	-0.0002 (0.0002)	-0.0002 (0.0002)
$\beta_{s_1 w_1}$	0.0002 (0.0002)	0.0002 (0.0002)
$\beta_{s_1 w_2}$	-0.0001 (0.0002)	-0.0001 (0.0002)
$\beta_{s_2 w_1}$	0.0001 (0.0001)	0.0001 (0.0001)
$\beta_{s_2 w_2}$	0.0005 (0.0004)	0.0005 (0.0004)
$\beta_{w_1 w_2}$	0.0001 (0.0005)	0.0001 (0.0004)
$\beta_{s_1^2}$	0.0006 (0.0004)	0.0005 (0.0004)
$\beta_{w_1^2}$	0.0001 (0.0010)	0.0001 (0.0010)

Table 5.1: Posterior means of the coefficients and standard deviations (in parenthesis) for the wind tunnel experiment.

Method	$\hat{\sigma}^2$	$\hat{\rho}$
SSVS	0.000010	0.65
SSVS-SPD	0.000009	0.59

Table 5.2: Posterior means of $\hat{\sigma}^2$ and $\hat{\rho}$ by the SSVS and SSVS-SPD for the wind tunnel experiment.

Figure E.3 and Figure E.4 for both ρ and σ^2 shows that sampling of the chains converge to the posterior distribution. The ACF plot for the σ^2 represents independently-drawn sampling for both SSVS and SSVS-SPD. Finally, the ACF plot for the correlation in Figure E.3 and Figure E.4 does display some high correlation at small lags, however, the autocorrelation decays quickly for larger lags.

(a) SSVS (b) SSVS-SPD

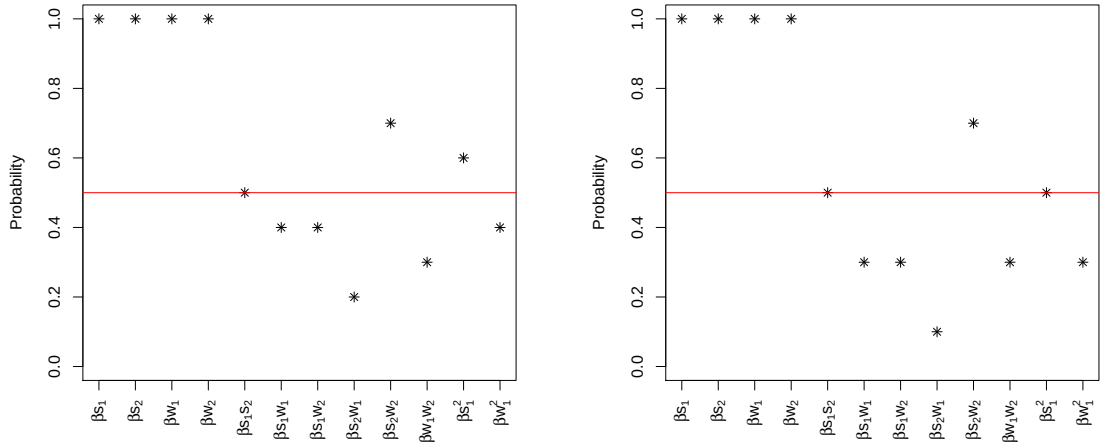


Figure 5.1: Approximate posterior probability for the wind tunnel design.

5.1.2 Analysis of the Freeze-Dried Coffee Experiment

The freeze-dried coffee experiment has been discussed in Section 1.2.2. In this section, we summarise the results of applying the Bayesian variable selection for this data. Our maximal model is the full quadratic model. In Table 5.3, the estimates of the 20 parameters of the quadratic model have been reported using the Stochastic Search Variable Selection (SSVS) and Stochastic Search Variable Selection for Split-plot Design (SSVS-SPD) explained in Sections 4.5.1 and 4.5.2 for response y_1 which is displayed in Table 1.5. Figure 5.2 shows a comparison between the (a) SSVS and (b) SSVS-SPD with respect to the resulting approximate posterior probability for the freeze-dried coffee experiment. The four parameters $\beta_{s_1}, \beta_{s_2}, \beta_{s_4}$, and β_{ws_2} have the highest posterior probability of being active by both SSVS and SSVS-SPD. This indicates that the four associated variables to these terms play a significant role in this experiment. Followed by these terms, we find that the parameters $\beta_{s_2s_4}$ and $\beta_{s_3s_4}$ have a lower, but still high, approximate posterior probability of around 0.90 by SSVS-SPD for both terms, yet the term $\beta_{s_3s_4}$ has an approximate posterior probability of around 0.80 by SSVS. Furthermore, for both SSVS and SSVS-SPD, the terms $\beta_w, \beta_{ws_4}, \beta_{s_1s_2}$, and $\beta_{s_1s_4}$ have an approximate posterior probability of between $(0.50 - 0.60)$. The SSVS-SPD preferred to include an extra term, which is β_{ws_3}, with a probability of 0.60 being active.

The Bayesian analysis for the real data of the freeze-dried coffee experiment yielded 10 significant variables by SSVS and 11 significant variables by SSVS-SPD which is similar to the resulting final model by Gilmour and Goos (2009). The rest of the terms seem to have a low approximate posterior probability of being active by both SSVS and SSVS-SPD which is similar to the analysis of the frequentist methods used in Chapter 3 as nine significant variables have been included in the final model (see Table 3.4 and Table 3.5). In summary, with respect to the final model from a Bayesian methods, it seems that the variables s_1, s_2, s_4, ws_2 followed by s_2s_4 would have the most significant role in this experiment. In addition, compared to the frequentist analysis in Chapter 3, we found that there is agreement between the frequentist and the Bayesian analysis with regard to including those five variables in the final model.

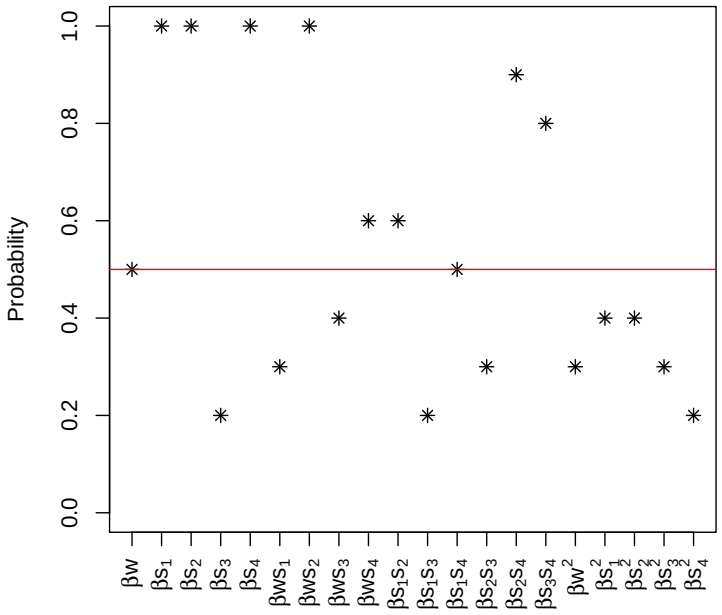
Table 5.3 represents the posterior means of the coefficients and standard deviations by both the SSVS and SSVS-SPD methods for the freeze-dried coffee experiment. Also, Table 5.4 shows the posterior means of the correlation $\hat{\rho}$ and the posterior means of the total variance $\hat{\sigma}^2$ for both SSVS and SSVS-SPD methods in the freeze-dried coffee experiment. Figure E.5, Figure E.6, Figure E.7 and Figure E.8 in Appendix E show (a) the trace and (b) the ACF plots for the Markov chain using Metropolis-Hastings within Gibbs sampling algorithm used by SSVS and SSVS-SPD to sample $\boldsymbol{\beta}$. Similar to the wind tunnel experiment, the trace plots showed that 10000 iterations are enough to achieve a reasonable approximation to the posterior model probability as the sampling from the chains converge to the posterior distribution. The ACF plots show some independent sampling (e.g. β_{w^2}) and some dependence at small lags but decays at large lags (e.g. $\beta_{s_2^2}$). Also, Figure E.9 and Figure E.10 in Appendix E represent (a) the trace and (b) the ACF plots for the Markov chain formed by sampling the correlation ρ and the total variance σ^2 by both SSVS and SSVS-SPD. The trace plots in Figure E.9 and Figure E.10 for both ρ and σ^2 are saturated and summarised good converges. The ACF plot for σ^2 represents a decay for large lags for both SSVS and SSVS-SPD. Finally, the ACF plot for the correlation in Figure E.9 and Figure E.10 does display some correlated sampling at small lags, however, the autocorrelation decays early at around lag 10.

From the previous discussions in Section 5.1.1 and Section 5.1.2 , we note that there is not a big difference between the analysis by frequentist methods and Bayesian methods

	SSVS	SSVS-SPD
β_w	-1.2247	-1.1074
	(1.5384)	(1.4115)
β_{s_1}	10.1046	10.0486
	(0.9915)	(0.9542)
β_{s_2}	-5.0055	-4.9870
	(1.0319)	(0.9761)
β_{s_3}	-0.0816	-0.1133
	(0.4042)	(0.4502)
β_{s_4}	2.8448	2.8581
	(1.0636)	(0.9252)
β_{ws_1}	-0.0361	-0.0139
	(0.8407)	(0.8979)
β_{ws_2}	-4.8085	-4.6960
	(1.7409)	(1.6248)
β_{ws_3}	-0.7308	-0.9598
	(1.1883)	(1.2380)
β_{ws_4}	-1.7546	-2.0428
	(1.888)	(1.7676)
$\beta_{s_1s_2}$	0.8963	1.0214
	(0.9923)	(0.9661)
$\beta_{s_1s_3}$	0.0717	0.0662
	(0.3759)	(0.3562)
$\beta_{s_1s_4}$	0.8160	1.0397
	(1.0819)	(1.1101)
$\beta_{s_2s_3}$	-0.2744	-0.3590
	(0.6353)	(0.6668)
$\beta_{s_2s_4}$	2.4827	2.4893
	(1.0126)	(0.9760)
$\beta_{s_3s_4}$	2.1795	2.2888
	(1.3949)	(1.3142)
β_{w^2}	-0.1510	-0.1727
	(1.4810)	(1.4487)
$\beta_{s_1^2}$	0.9158	1.0733
	(1.4810)	(1.5771)
$\beta_{s_2^2}$	-0.8341	-0.9861
	(0.9464)	(1.5265)
$\beta_{s_3^2}$	0.2128	0.2864
	(0.9464)	(1.1276)
$\beta_{s_4^2}$	0.0142	0.0271
	(0.8999)	(0.9634)

Table 5.3: Posterior means of the coefficients and standard deviations (in parentheses) for the freeze-dried coffee experiment.

(a) SSVS



(b) SSVS-SPD

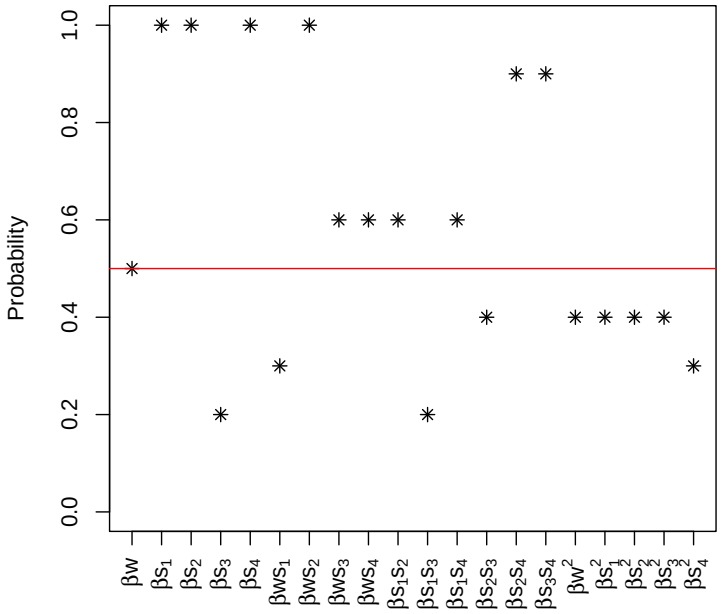


Figure 5.2: Approximate posterior probability for the freeze-dried coffee design.

Method	$\hat{\sigma}^2$	$\hat{\rho}$
SSVS	16.6	0.41
SSVS-SPD	16.7	0.40

Table 5.4: Posterior means of $\hat{\sigma}^2$ and $\hat{\rho}$ by the SSVS and SSVS-SPD for the freeze-dried coffee experiment.

with respect to the variable selection in the real data examples. We note the differences on the estimated effect sizes and their standard-deviations between the frequentist and the Bayesian methods. This is because in the Bayesian methods the results are sensitive to the choice of the prior distributions. A comparison between the frequentist and the Bayesian methods will be addressed in more details in Section 5.3.

Table 5.1 and Table 5.3 show the posterior means of the parameters. Our assumption of the prior for the probability of inclusion affects the sampling of the parameters. For example, the effect sizes and standard-deviations for the parameters who have high approximated posterior probability are slightly lower than the effect sizes and standard-deviations of the active parameters in the frequentist analysis. The choice of the prior of the probability of inclusions play a significant role in this variation between the two analysis. For example, in the wind tunnel experiment, the prior of ω has a mode $=$ 0.25 yields 7 variables that have approximate posterior probability above 0.50. While, in the freeze dried-coffee experiment, the prior of ω has a mode $=$ 0.50 which yields 10 variables by SSVS and 11 variables by SSVS-SPD that have approximate posterior probability above 0.50. Moreover, if the approximated probability of the parameter in Figure 5.1 and Figure 5.2 is low, then there is a spike in the marginal posterior densities for the associated parameter. We do not expect to have zero effect size for the non active parameters because the estimated posterior density for the non active parameters are concentrated around zero as the spike is a normal distribution with a small variance, see Figure 5.6 and Figure 5.9.

5.2 Simulation Study

We performed a simulation study by generating 1000 datasets, where each dataset would run for 10000 iterations using MCMC, and the SSVS and the SSVS-SPD would be

applied at two levels of $\eta = 1$ and 10. We assume that $\sigma_\epsilon^2 + \sigma_\gamma^2 = 10$. Thus, the true value of σ^2 is 10. Also, the true value of ρ is 0.5 at $\eta = 1$ while the true value at $\eta = 10$ of ρ is 0.9. We used the same model as in Section 3.2.1 for the simulation that used the design of the wind tunnel. Also, we used the same model as in Section 3.2.2 for the simulation that used the design of the freeze-dried coffee.

As we are keen on variable selection, we shall calculate the Type I error rate which tells us how the methods identify the non-active effects, and the Type II error rate which tells us how the methods identify the active effects. In this simulation, we follow Barbieri et al. (2004) to define the Type I and II error rates using the indicator vector $\boldsymbol{\nu}$ and the approximate posterior probability from (5.1). If the true variable is active but the algorithm yielded a corresponding approximation posterior probability of less than 0.5, this variable would then have a Type II error rate. Also, if the true variable is non-active but the algorithm yielded a corresponding approximation posterior probability larger than or equal to 0.5, this variable would then have Type I error rate. We are also keen to find out the precision of the point estimates by SSVS and SSVS-SPD. This can be measured by calculating the median relative model error (MRME) for the estimates of the SSVS and the SSVS-SPD as explained in Section 3.2.

5.2.1 Simulation Study Using the Design of the Wind Tunnel Experiment

We performed a simulation study to examine the performance of the SSVS and SSVS-SPD methods. Using the design of the wind tunnel experiment from Table 1.3, we generated the response given the true model as

$$\mathbb{E}(\mathbf{Y}) = 4w_1 + 2s_2 - 4w_1w_2 + 2w_1s_2 + 6w_1^2 + 4s_1^2.$$

Type I and II error rates are displayed in Figure 5.3 , for two settings of (a) $\eta = 1$ and (b) $\eta = 10$. The left side of the figures display the Type II error rate while the right side represents the Type I error rate. Also, the slab posterior distributions for SSVS and SSVS-SPD are displayed in Figure 5.4. The details of the Type I and II error rates are presented in Table C.1 and Table C.2 in Appendix C.

In this experiment, we assume six active variables ($s_2, w_1, s_2w_1, w_1w_2, s_1^2$ and w_1^2) and six non-active variables ($s_1, w_2, s_1s_2, s_1w_1, s_1w_2$ and s_2w_2). For both $\eta = 1$ and $\eta = 10$, by comparing the SSVS to the SSVS-SPD, we note that five active variables (s_2, w_1, s_2w_1, w_1w_2, s_1^2) have been identified with low Type II error rates (see Table C.2 in Appendix C). However, the quadratic whole-plot variable w_1^2 was more likely to be estimated as non-active by an error rate of 0.2 using both SSVS and SSVS-SPD. The SSVS-SPD recorded slightly lower Type II error rate of w_1^2 than the SSVS.

In addition, the advantage of the SSVS-SPD is clear as it represents a lower Type I error rate than the SSVS for both subplot and whole-plot variables. This can be explained by looking at the slab posterior distribution in Figure 5.4. At $\eta = 1$ in the first row, we note the slab posterior of the whole-plot parameters at (c) c_w have a similar slab posterior of the SSVS shown in (a) c which are a variety of small and large values. This gives the opportunity of the slab posterior of subplot parameters at (b) c_s to pick the small values more frequently. It indicates that the subplot's factors by SSVS-SPD are more likely to be sampled from a distribution of small variance. This allows the subplot factors to have a lower Type I error rate by SSVS-SPD than the SSVS. Tan and Wu (2013) explained how the slab posterior would affect the detection of small and large effects. They state that large values of c tend to favor sparse models with large effects. Also, small values of c tend to favor less sparse models. However, very small values of c would again favor sparse models because the Bayesian model does not support the hypothesis of a true model with larger effects than $\sqrt{c\sigma}$, where σ is the standard deviation of the total variance σ^2.

Apart from w_2, the non-active terms have been detected with lower Type I error rate at $\eta = 10$ than at $\eta = 1$ for both SSVS and SSVS-SPD. Similar performance of the slab posterior distribution of SSVS and SSVS-SPD for $\eta = 10$ is shown in the second row of Figure 5.4.

From the Median Relative Model Error (MRME) in Figure 5.5 (a), we note that there is not a considerable difference between the MRME of the models by SSVS and the MRME of the models by SSVS-SPD. At both $\eta = 1$ and $\eta = 10$, both SSVS and SSVS-SPD show that the estimated models have a slightly higher prediction error than the estimated model by the Generalised Least Square (GLS) estimator. We found that although

Method	η	$\hat{\sigma}^2$	$\hat{\rho}$
SSVS	1	11.7	0.51
	10	7.40	0.77
SSVS-SPD	1	11.8	0.50
	10	7.30	0.76

Table 5.5: Posterior means of $\hat{\sigma}^2$ and $\hat{\rho}$ by the SSVS and SSVS-SPD from the simulation by using the design of the wind tunnel experiment.

(a) $\eta = 1$ (b) $\eta = 10$

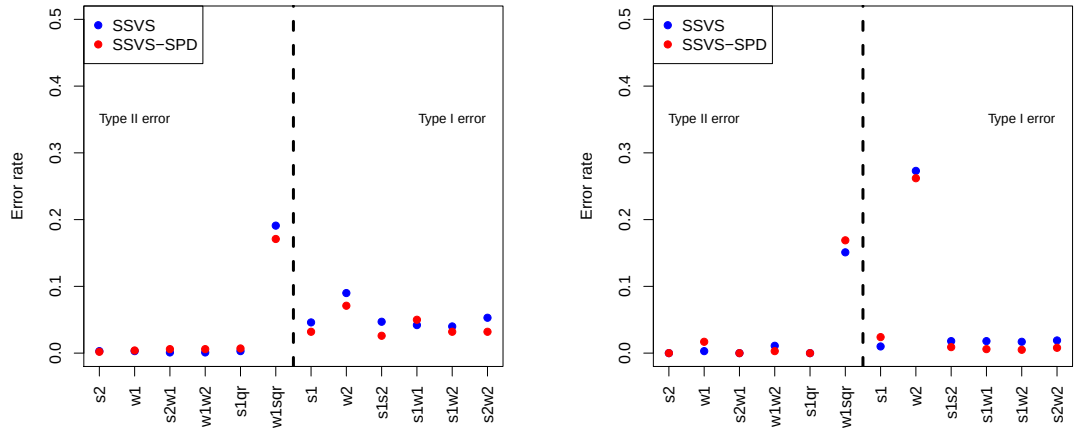


Figure 5.3: Type I and II error rates for the wind tunnel design.

the inclusion probability of large effects is high, the estimated coefficients are slightly smaller than the true coefficient value. Also, the SSVS and SSVS-SPD tend to sample the non-significant terms from a distribution with a small variance. This implies that true zero effects, although they have lower and non-considerable probability of inclusion, have small estimates and are not explicitly zero. Figure 5.6 displays the Boxplots for the coefficient estimates as we see the non-active variables have small estimates and not exactly zero by both the SSVS and SSVS-SPD. This indeed makes the MRME of the SSVS and the SSVS-SPD slightly higher than the MRME of the GLS estimator.

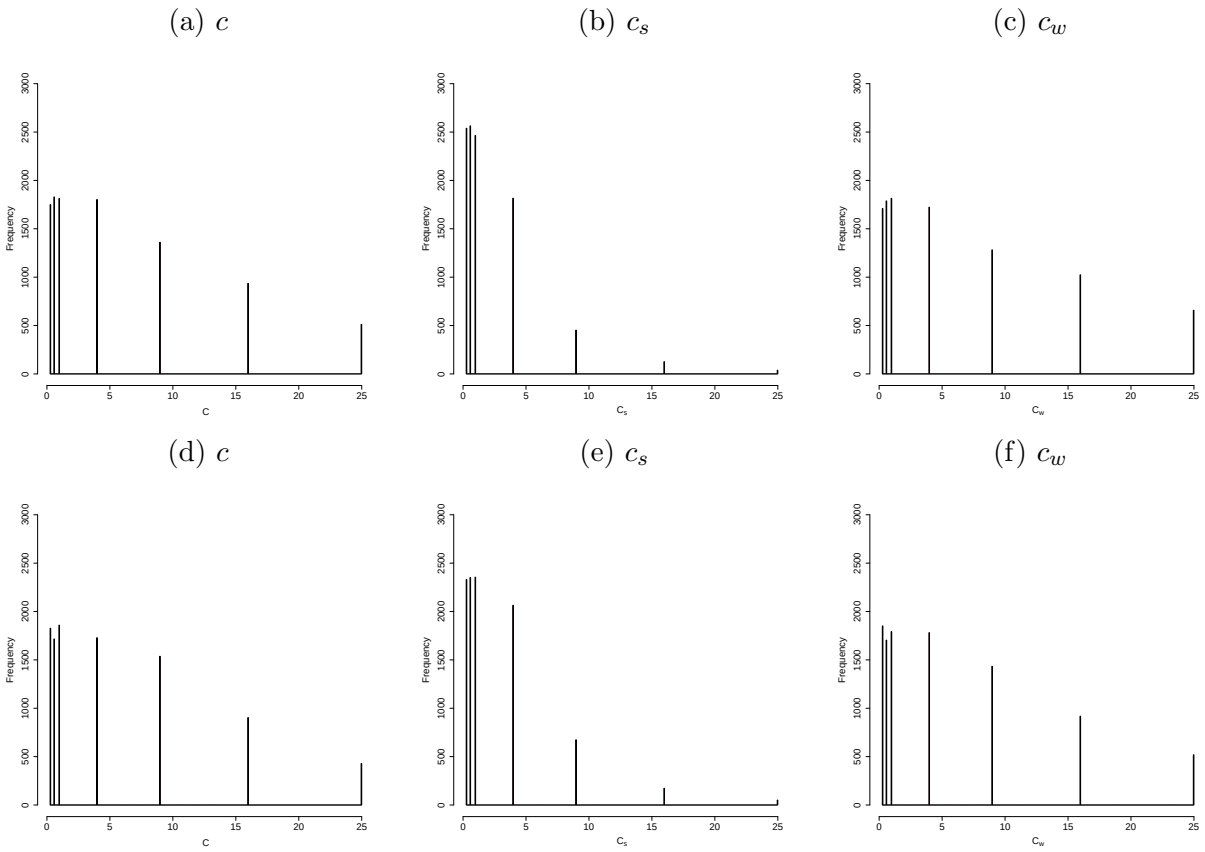


Figure 5.4: The slab posterior distributions for the wind tunnel design at $\eta = 1$ in the first row and $\eta = 10$ in the second row.

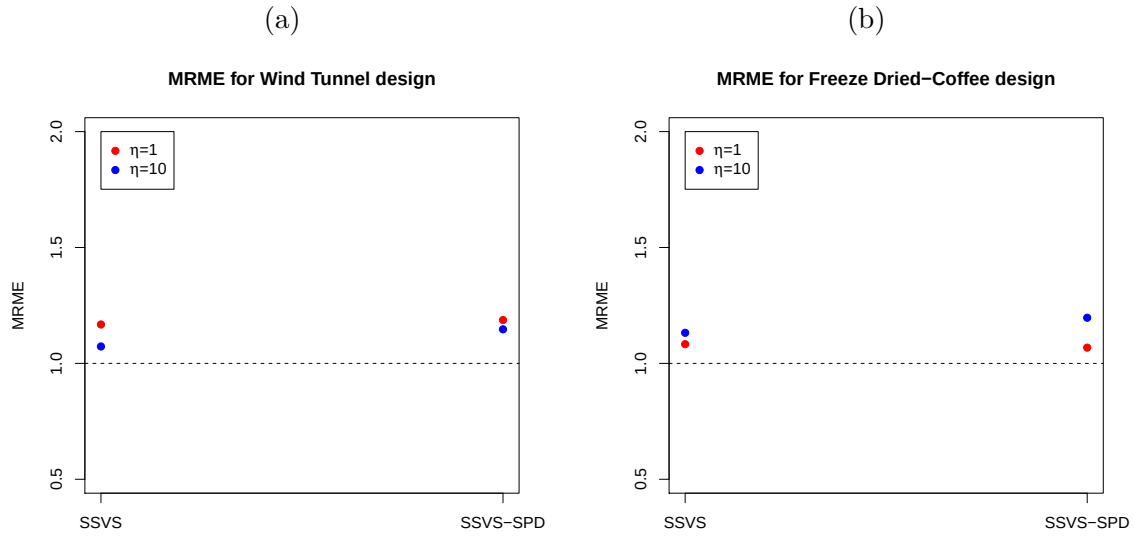


Figure 5.5: Median relative model error (MRME).

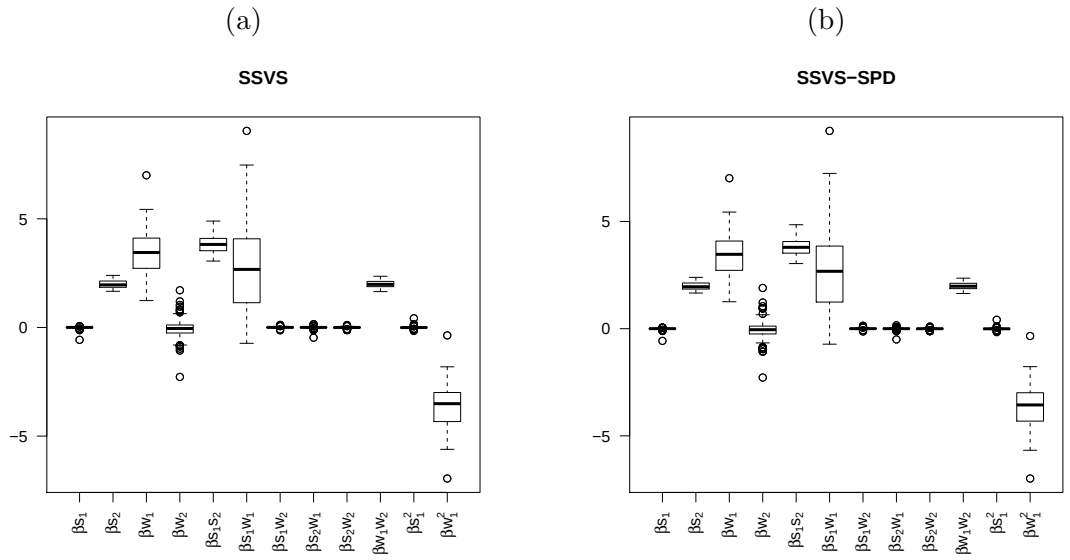


Figure 5.6: Box plots of the posterior means of the coefficients for the wind tunnel design at $\eta = 10$.

5.2.2 Simulation Study Using the Design of the Freeze-Dried Coffee Experiment

We performed a simulation study to examine the performance of the SSVS and SSVS-SPD methods. Using the design of the freeze-dried coffee experiment from Table 1.5, we generate the response giving the true model as

$$\mathbb{E}(\mathbf{Y}) = 4w + 4s_1 - 3s_2 + 2s_3 - 4ws_1 + 3ws_2 - ws_3 + 4s_1s_2 + 3s_1s_3 + 2s_1s_4 + s_2s_4 + 4w^2 + 2s_1^2 - s_2^2 + 2s_4^2$$

Type I and II error rates are displayed Figure 5.7, for two settings of (a) $\eta = 1$ and (b) $\eta = 10$. Type II error rates are displayed in the left side of the figure while the Type I error rates are shown in the right side. Also, the slab posterior distributions for SSVS and SSVS-SPD are displayed in Figure 5.8. The details of the Type I and II error rates are presented in Table D.1 and Table D.2 in Appendix D. In this experiment, we assume 15 active variables (w, s_1, s_2, s_3, ws_1, ws_2, ws_3, s_1s_2, s_1s_3, s_1s_4, s_2s_4, w^2, s_1^2, s_2^2, and s_4^2) and five non-active variables ($s_4, ws_4, s_2s_3, s_3s_4,$ and s_3^2).

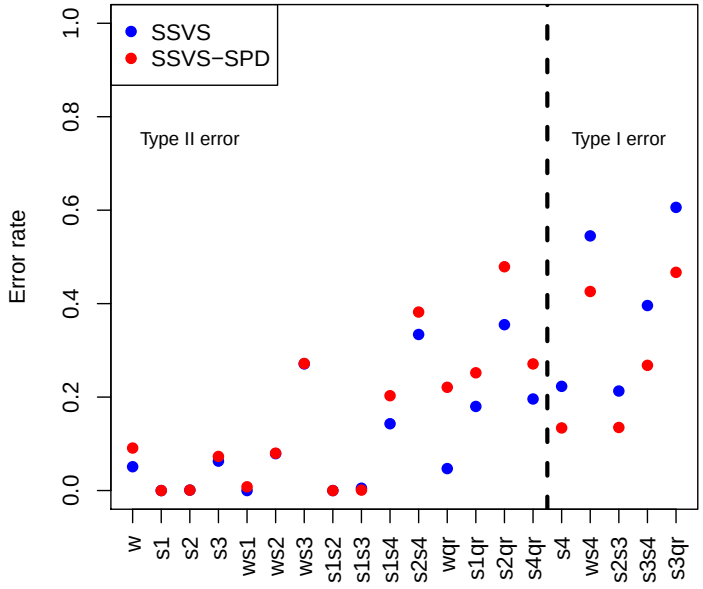
At $\eta = 1$, both methods found difficulties in detecting the quadratic effect terms and some interaction effect terms (e.g. s_2s_4 and s_2^2). This design is non-orthogonal and

considerable correlations between variables exist (see Figure 3.3 and Figure 3.3). The non-orthogonality of this design does not help the SSVS and the SSVS-SPD to detect the active terms precisely at $\eta = 1$ in which the $\sigma_\epsilon^2 = \sigma_\gamma^2$. On the other hand, at $\eta = 10$ in which the $\sigma_\epsilon^2 < \sigma_\gamma^2$, the subplot's factors are better included in the model recorded as small or zero Type II error rates by both SSVS and SSVS-SPD. Tan and Wu (2013) stated the same result for correlated data from the split-plot design. For example, from Figure 5.7 (b) and Table D.2 in Appendix D, most of the subplot (main and interaction) effects have a zero Type II error rate.

With respect to the Type I error rate, at both $\eta = 1$ and $\eta = 10$, both SSVS and SSVS-SPD fail in controlling it at a significance level of 5%. However, we note the benefit of SSVS-SPD as it decreases the Type I error rate for all non-active terms in contrast to the SSVS, similar to the analysis of the simulation from the wind tunnel design. We conclude that having two different slab posterior distributions in the SSVS-SPD allowed a reduction of the Type I error rates for both subplot and whole-plot effect terms. Figure 5.8 (b) shows that the subplots are more likely to be sampled from a distribution with a small variance as this causes the reduction in Type I error rates for the subplot terms. We note that due to the non-orthogonality of the design and the assumed true model which is very challenging, both the Type I and II error rate is higher than the rates of the simulation study by the wind tunnel design.

The analysis of the Median Relative Model Error (MRME) in Figure 5.5 (b) for both SSVS and SSVS-SPD using the freeze-dried coffee design is similar to the analysis of the (MRME) in Figure 5.5 (a) for the wind tunnel design. Figure 5.9 displays the Boxplots for the coefficients estimates for the freeze-dried coffee design as this explains the high MRME since the true zero effects have been estimated as small values by SSVS and SSVS-SPD and not exactly zero.

(a) $\eta = 1$



(b) $\eta = 10$

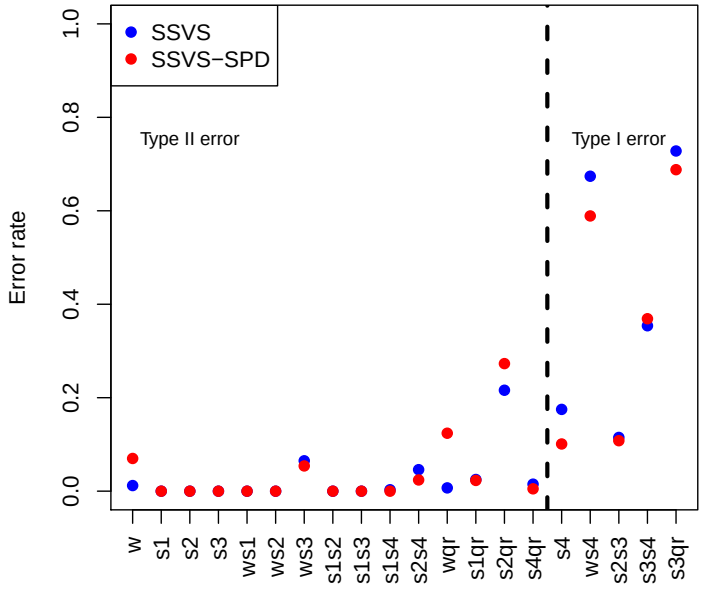


Figure 5.7: Type I and II error rates for the freeze-dried coffee design.

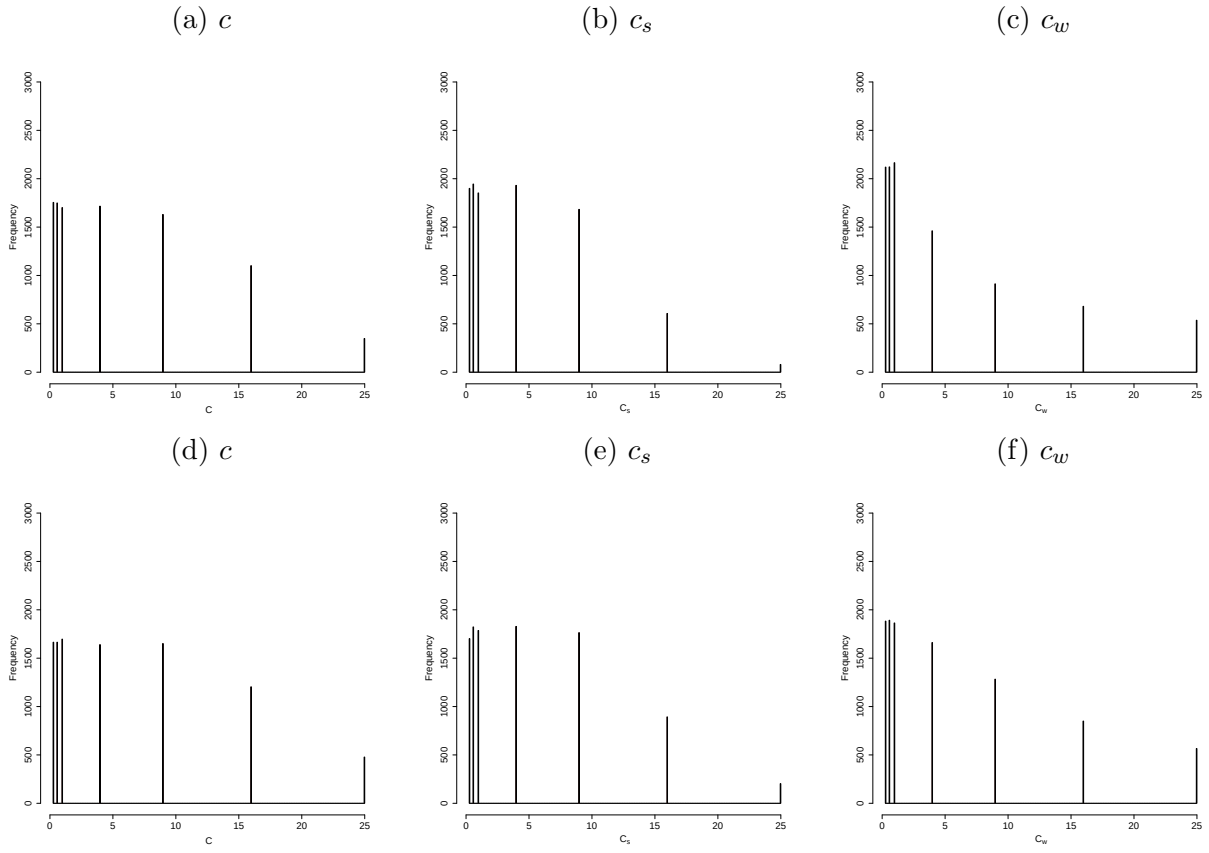


Figure 5.8: The slab posterior distributions for the freeze-dried coffee design at $\eta = 1$ in the first row and $\eta = 10$ in the second row.

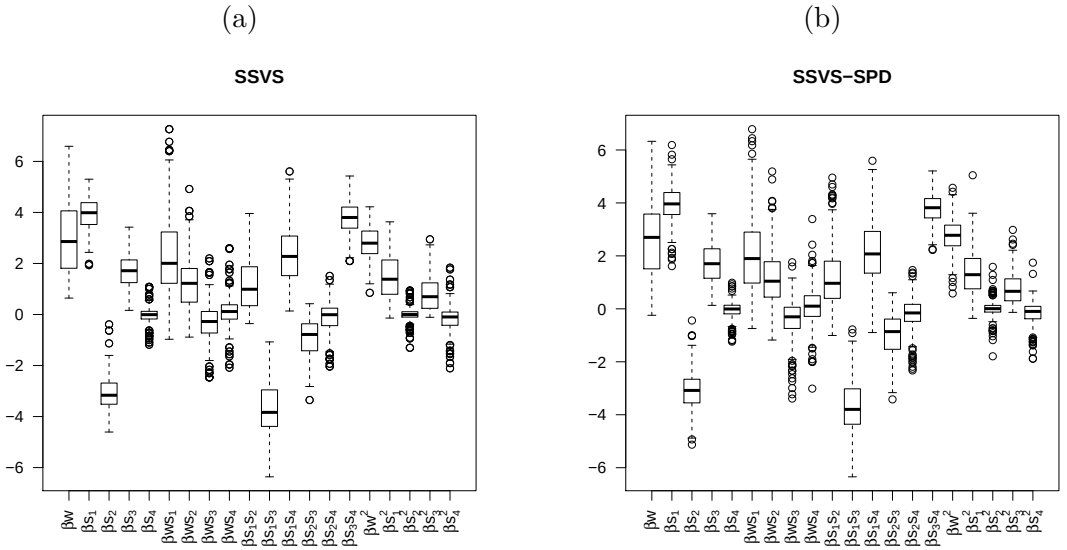


Figure 5.9: Box plots of posterior means of the coefficients for the freeze-dried coffee design at $\eta = 1$.

Method	η	$\hat{\sigma}^2$	$\hat{\rho}$
SSVS	1	11.2	0.50
	10	10.6	0.70
SSVS-SPD	1	11.0	0.49
	10	11.0	0.70

Table 5.6: Posterior means of $\hat{\sigma}^2$ and $\hat{\rho}$ by the SSVS and SSVS-SPD from the simulation by using the design of the freeze-dried coffee experiment.

5.3 Comparison of the Frequentist and the Bayesian methods

In this work, we used the frequentist analysis by applying the Penalised Generalised Least Square (PGLS) in (1.11) and the Penalised Generalised Least Square for Split-Plot Design (PGLS-SPD) in (2.1) to two different examples from industrial settings in which a split-plot design is used. We selected the wind tunnel experiment in Section 1.2.1 to present the case of an orthogonal design whereas the freeze-dried coffee experiment in Section 1.2.2 was chosen to present the case of a non-orthogonal design. In Chapter 3, in a comparison between the PGLS and the PGLS-SPD, we reported the improvement of variable selection and parameter estimation for both examples by using the PGLS-SPD as this method reduces both Type I and II error rates compared to the PGLS method. In addition, the LARS performs a very good performance by having lower Type I error rate than the PGLS and PGLS-SPD. LARS also yields almost similar Type II error rate to the PGLS and the PGLS-SPD. We note that the backward yields the lowest Type I error rate among all methods, however has higher Type II error rates than all methods. There is a trade off between Type I and Type II error rates for all methods.

Also, the PGLS-SPD provides lower MRME than the PGLS. However, in general, detecting the non active effect factors by both PGLS and PGLS-SPD was not achieved, for both the design from the wind tunnel experiments (see Table A.1 and Table A.2 in Appendix A) and for the design from the freeze-dried coffee experiment (see Table B.1 and Table B.2 in Appendix B).

To solve the issue of the high Type I error rate, we applied the Bayesian approach using Bayesian variable selection methods as in Chapters 4 and 5. We used the Stochastic Search Variable Selection (SSVS) applied by Tan and Wu (2013), yet we used the Markov

Chain Monte Carlo (MCMC) sampler from the full conditional distribution instead of the numerical integration of β and σ^2 that was used in Tan and Wu (2013). Furthermore, we adapted the SSVS to fit data from split-plot designs and we introduced the Stochastic Search Variable Selection for Split-Plot Design (SSVS-SPD). In Chapter 5, we presented the results of the SSVS and the SSVS-SPD with respect to variable selection and model prediction error. In the example of an orthogonal design, we note that the high rate of Type I error which we found in the frequentist approach has been reduced by the Bayesian approach. This is due to the advantage of mixture normal distribution and the choice of the spike variance which allows us to identify the non active effect factors. Comparing Table A.1 and Table A.2 in Appendix A to Table C.1 in Appendix C, we found a remarkable reduction in Type I error rate for the wind tunnel design by the Bayesian approach. Although Bayesian methods recorded lower Type I error rate than the frequentist methods, LARS detect w_2 better than SSVS and SSVS-SPD at $\eta = 10$. In addition to the reduction of Type I error rates, by comparing Table A.3 and Table A.4 in Appendix A to Table C.2 in Appendix C, we found that Bayesian methods, particularly the SSVS-SPD, overcome the problem of frequentist methods in having high Type II error rates for w_1^2. More specifically, in a comparison between the SSVS and the SSVS-SPD, we note that our extended Bayesian approach SSVS-SPD method provides a lower Type I error rate at $\eta = 1$ and $\eta = 10$ than the SSVS method (see Table C.1 in Appendix C).

On the other hand, in the example of non-orthogonal design, by comparing Table B.1 and Table B.2 in Appendix B to Table D.1 in Appendix D, we found that the frequentist approach identified non-active effect factors better than the Bayesian approach. In particular, the LARS method detects the non active effect factors better than the SSVS and SSVS-SPD methods. However, the Bayesian approach provided lower Type II error rates than the frequentist approach. Also, apart from the quadratic effect factors, LARS overcomes all methods with respect to Type II error rate. This is noticeable when comparing Table B.3 and Table B.4 in Appendix B to Table D.2 in Appendix D.

We note the design structure affects the performance of the frequentist and the Bayesian approaches. In orthogonal designs, we recommend a Bayesian analysis to identify active effect factors. Apart from the whole-plot non active effect factors which were detected

better by LARS, Bayesian analysis is recommended for identifying non active effect factors as well. Furthermore, for non-orthogonal designs, if we are keen to detect the active effect factors as in screening experiments, we recommend the use of Bayesian analysis and mainly the SSVS-SPD if the design used is a split-plot design. However, LARS overcomes the SSVS-SPD in detecting the non active effect factors.

5.4 Discussion

In this chapter, we discussed the implementation of Stochastic Search Variable Selection (SSVS) and the Stochastic Search Variable Selection for Split-plot Design (SSVS-SPD) to two different examples from split-plot experiments. We introduced the results of the real datasets for the wind tunnel experiment in Section 5.1.1, and the results of the real datasets for the freeze-dried coffee experiment in Section 5.1.2. Both the SSVS and the SSVS-SPD agreed to consider the four main variables s_1, s_2, w_1, and w_2 as the most significant variables in the wind tunnel experiment. Moreover, the SSVS and SSVS-SPD agreed to consider four variables s_1, s_2, s_4, and ws_4 to be the most significant variables in the freeze-dried coffee experiment. It is worth mentioning that these selected variables in both experiments have also been selected by the frequentist analysis in Chapter 3 which also agreed with analysis by Simpson et al. (2004) for the wind tunnel experiment and by Gilmour and Goos (2009) for the freeze-dried coffee experiment. In more detail, with respect to the wind tunnel experiment, Bayesian methods select the same set of variables as in the original paper though both s_1s_2 and s_1^2 have a probability of inclusion that equals 0.50. Furthermore, with respect to the freeze-dried coffee experiment, Bayesian methods select the same set of variables as in the original paper though Bayesian methods added three extra variables. They are s_1s_4 with a probability of inclusion that equals 0.50 and both s_1s_2 and ws_4 with a probability of inclusion that equals 0.60. The assessment of MCMC sampling shows good convergences for β, σ^2, and ρ for both experiments.

Two simulation studies generating 1000 datasets have been approached using the wind tunnel design in Section 5.2.1, and by using the freeze-dried coffee design in Section 5.2.2. Both SSVS and SSVS-SPD were successful in identifying five active variables

from the assumed six true active variables in the simulation by using the design of the wind tunnel experiment. The most challenging analysis was in the simulation using the design of the freeze-dried coffee experiment. This is because we tried to detect 15 active variables in a small and non-orthogonal design using a model of 20 variables. However, the case of $\eta = 10$ resulted in including most of the active subplot factors because in that case σ_ϵ^2 is less than the σ_γ^2 which yields a remarkable low Type II error rate in this case for the subplot factors.

The frequentist approach using the PGLS-SPD by $\text{ALASSO}_{0.5}$ provided a lower prediction error than the Bayesian approach for the wind tunnel example. The frequentist approach by BIC for the freeze-dried coffee example has a similar prediction error to the Bayesian approach. However, we do not expect the Bayesian approach to have a better prediction error than the frequentist approach. This is because in the Bayesian approach, we sample the non-active effects from a normal distribution with a small variance. This results in non-significant effects having very small estimates but not exactly zero. This increases the differences between the resulting model and the true model. On the other hand, we find in the frequentist approach, with a proper choice of shrinkage parameter, non-active variables are estimated to be zero exactly (see for example the $\text{ALASSO}_{0.5}$ in the wind tunnel example).

In Section 5.3, we provided a comparison of the frequentist analysis and the Bayesian analysis methods, and we showed how the LARS and the SSVS-SPD method surpasses the other methods with respect to variable selection. In summary, from this work, we note that both SSVS-SPD and LARS have advantages and disadvantages. However, the LARS overcomes Bayesian methods with respect to Type I error rate. Also, apart from the quadratic effect factors, LARS overcomes the Bayesian methods with respect to Type II error rate.

Chapter 6

Conclusion and Future Work

6.1 Conclusion

The focus of this thesis is the analysis of data from split-plot experiments using motivating examples from industrial environment. In Chapter 1, we presented the details of the two motivating examples that used in this work. We also provided an introduction of the frequentist variable selection and the Bayesian variable selection.

In Chapter 2, we introduced the computational algorithm of the Penalised Generalised Least Squares (PGLS) in which we apply one shrinkage penalty to all factors. Also, we discussed the computational algorithm of the Penalised Generalised Least Squares for Split-Plot Design (PGLS-SPD) in which we apply two shrinkage penalties for subplot factors and whole-plot factors. We explained that the motivation of the extension in PGLS-SPD is that we have two different strata in split-plot design. Another reason is that we found that even when the two shrinkage parameters are being selected the same grid of values, the information criteria tend to select different values of the shrinkage parameter for subplot factors and whole-plot factors.

In Chapter 3, we analysed data from two real-life experiments, and we have done two simulation studies in which we have compared different variable selection methods. The main finding from our analysis for both examples, is that Backward yields the lowest Type I error rate. After that, the LARS as it overcomes the PGLS-SPD method. Moreover, the active quadratic whole-plot effect factor was hardly to detect by all methods.

Apart from that, the LARS yields similar Type II error rate to the PGLS-SPD for the wind tunnel example. For the freeze-dried coffee example, the LARS overcomes the PGLS-SPD with respect to the main effect factors. However, in screening experiments where the main aim is to identify the active effect factors, the LARS and the PGLS-SPD offer good alternatives to the traditional methods. In addition, we recommend to use the $\text{ALASSO}_{0.5}$ for analysing data from orthogonal split-plot designs whereas the EN can be utilised for non-orthogonal split-plot designs.

In Chapter 4, we described the a Metropolis-Hastings within Gibbs sampling algorithm, which generates dependent samples from the posterior distribution of the parameters in linear mixed models. We explained the hierarchical mixture model for variable selection and discussed the full conditional distributions. Also, we discussed the Stochastic Search Variable Selection (SSVS) and its adapted version Stochastic Search Variable Selection for Split-Plot Design (SSVS-SPD). We also explained that the motivation of this extension in SSVS-SPD is that we have two different strata in split-plot design. Also, because we found that although we use the same prior distribution for the slab variance of the subplot factors and whole-plot factors, different posterior distributions of the slab variance will be used to sample the subplot and whole-plot effects.

In Chapter 5, with respect to Bayesian methods, we found that the SSVS-SPD detects the active and non-active effect factors with lower error rate than the SSVS. We provided a comparison between the performance of the frequentist analysis and the Bayesian analysis with regard the variable selection and the model prediction error. The LARS and the SSVS-SPD overcome the other methods with respect to variable selection. In conclusion, we note that both SSVS-SPD and LARS have pros and cons. However, the LARS overcomes SSVS-SPD with respect to Type I error rate. Also, apart from the quadratic effect factors, LARS overcomes SSVS-SPD with respect to Type II error rate.

6.2 Future Work

Possible extensions to the work in this thesis are discussed as in Sections 6.2.1 and 6.2.2.

6.2.1 Extension of the Frequentist Analysis Approach

- The usefulness of other choices for the shrinkage parameters based on the Akaike Information Criterion (AIC) (Akaike, 1973) and the Generalised Cross Validation (GCV) (Wahba, 1980) could be investigated in selecting the shrinkage parameters.

- Extend the Penalised Generalised Least Square for Split-Plot Design (PGLS-SPD) to more restricted randomised designs. For example, the Split-Split-Plot Design which has three different types of factors. One can investigate the application of Penalised Generalised Least Square for the Split-Split-Plot Design (PGLS-SSPD) by minimising

$$\mathbf{Q}_{PGLS}(\boldsymbol{\beta}) = \frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta})'\hat{\mathbf{V}}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}) + \sum_{j=1}^{d_w} p_{\lambda w}(\mid \beta_j \mid) + \sum_{j=d_w+1}^{d_s} p_{\lambda s}(\mid \beta_j \mid) + \sum_{j=d_s+1}^{d} p_{\lambda ss}(\mid \beta_j \mid),$$

  with respect to $\boldsymbol{\beta}$. In the above formula, d_w denotes the number of the whole-plot model coefficients, $\beta_1, \ldots, \beta_{d_w}$ represent these whole-plot coefficients, the $\beta_{d_w+1}, \ldots, \beta_s$ are the subplot coefficients, and the $\beta_{d_s+1}, \ldots, \beta_d$ are the sub-subplot coefficients. Also, $p_{\lambda w}(.)$, $p_{\lambda s}(.)$, and $p_{\lambda ss}(.)$ are the shrinkage penalty functions which will be applied to the whole-plot factor effects, subplot factor effects, sub-subplot factor effects respectively. All λ_w, λ_s, and λ_{ss} are unknown positive shrinkage parameters, which can be chosen by several selection methods as in Section 2.3. In Split-Split-Plot design, there are more hierarchical stages, the small units in Split-Plot design are splitted further into split-split-plot observational units. In such case the experimental design contains the whole-plot error, split-plot error and the split-split-plot error in which 3 variance components need to be estimated.

6.2.2 Extension of the Bayesian Analysis Approach

- Further investigation can be drawn using another prior distribution which provide variable selection such as Bayesian LASSO. Tibshirani (1996) suggested that

LASSO estimates can be interpreted as posterior mode estimates when the regression parameters have independent and identical Laplace (i.e., double-exponential) priors. Bayesian LASSO was introduced by Park and Casella (2008). Bayesian LASSO was proposed for linear models, hence it needs to be adopted to linear mixed model to fit data from a split-plot design. For linear model, the conditional Laplacian prior for $\boldsymbol{\beta}$ is

$$\pi(\beta|\sigma^2) = \prod_{j=1}^{d} \frac{\lambda}{2\sigma} \exp\Big\{ -\lambda|\beta_j|/\sigma \Big\}.$$

for $\lambda \geq 0$ and σ^2 is the response variance and needs an appropriate prior to estimate it.

- In the Stochastic Search Variable Selection, we introduced mixture of normal density for both spike and slab distributions in which the spike has a small variance while the slab has larger variance. One can replace the spike prior with a point mass at zero (Dirac spike) and can keep the slab to be normal distribution. The probability of Dirac spike is a discrete random variable which is exactly equal to zero. By considering the Dirac spike, the indicator vector ν_j, total variance σ^2, and slab variance c, one can use a prior for $\boldsymbol{\beta}$ from

$$p(\beta_j|\nu_j) = (1-\nu_j)p_{spike}(\beta_j) + \nu_j p_{slab}(\beta_j).$$

where $p_{slab}(\beta_j) = \mathrm{N}(0, c\sigma^2)$, and $p_{spike}(\beta) = I_0(\beta_j)$. The non significant coefficients hence will be set to zero.

Appendix A

Type I and II error rates by PGLS and PGLS-SPD for the wind tunnel design.

	True non-active variable		s_1	w_2	s_1s_2	s_1w_1	s_1w_2	s_2w_2
	Method	Criteria	0	0	0	0	0	0
PGLS	LASSO	cAIC	0.178	0.250	0.242	0.180	0.248	0.216
		BIC	0.170	0.242	0.222	0.148	0.224	0.208
	$ALASSO_{0.5}$	cAIC	0.100	0.198	0.166	0.110	0.132	0.162
		BIC	0.104	0.198	0.142	0.096	0.116	0.136
	SCAD	cAIC	0.132	0.230	0.194	0.136	0.160	0.186
		BIC	0.112	0.198	0.170	0.112	0.140	0.160
	EN	cAIC	0.178	0.258	0.242	0.176	0.248	0.220
		BIC	0.178	0.246	0.222	0.156	0.232	0.208
PGLS-SPD	LASSO	cAIC	0.128	0.262	0.202	0.132	0.168	0.176
		BIC	0.108	0.254	0.162	0.112	0.144	0.156
	$ALASSO_{0.5}$	cAIC	0.076	0.214	0.114	0.076	0.112	0.124
		BIC	0.056	0.198	0.098	0.060	0.096	0.108
	SCAD	cAIC	0.128	0.230	0.182	0.120	0.148	0.168
		BIC	0.104	0.210	0.138	0.104	0.132	0.144
	EN	cAIC	0.162	0.286	0.218	0.164	0.204	0.196
		BIC	0.146	0.282	0.210	0.132	0.180	0.180
	Backward		0.050	0.069	0.048	0.045	0.036	0.063
	LARS		0.044	0.174	0.074	0.036	0.072	0.088

Table A.1: Type I error rate for the wind tunnel design at $\eta = 1$ for the PGLS and PGLS-SPD. $ALASSO_{0.5}$ is the ALASSO with $\psi = 0.5$.

	True non-active variable		s_1	w_2	s_1s_2	s_1w_1	s_1w_2	s_2w_2
	Method	Criteria	0	0	0	0	0	0
PGLS	LASSO	cAIC	0.248	0.260	0.304	0.236	0.284	0.272
		BIC	0.228	0.240	0.288	0.228	0.264	0.260
	$ALASSO_{0.5}$	cAIC	0.152	0.200	0.192	0.120	0.164	0.168
		BIC	0.124	0.196	0.168	0.120	0.156	0.164
	SCAD	cAIC	0.192	0.208	0.264	0.204	0.248	0.240
		BIC	0.188	0.196	0.252	0.192	0.240	0.224
	EN	cAIC	0.256	0.264	0.308	0.240	0.288	0.276
		BIC	0.236	0.248	0.296	0.228	0.268	0.260
PGLS-SPD	LASSO	cAIC	0.156	0.272	0.216	0.140	0.200	0.180
		BIC	0.152	0.256	0.208	0.132	0.196	0.176
	$ALASSO_{0.5}$	cAIC	0.104	0.208	0.140	0.096	0.128	0.144
		BIC	0.104	0.200	0.136	0.092	0.128	0.144
	SCAD	cAIC	0.164	0.228	0.220	0.140	0.208	0.192
		BIC	0.164	0.208	0.220	0.136	0.204	0.180
	EN	cAIC	0.184	0.312	0.236	0.188	0.236	0.204
		BIC	0.172	0.292	0.228	0.164	0.224	0.192
	Backward		0.055	0.077	0.048	0.041	0.043	0.059
	LARS		0.044	0.184	0.076	0.036	0.072	0.088

Table A.2: Type I error rate for the wind tunnel design at $\eta = 10$ for the PGLS and PGLS-SPD. $ALASSO_{0.5}$ is the ALASSO with $\psi = 0.5$.

	True active variable		s_2	w_1	s_2w_1	w_1w_2	s_1^2	w_1^2
	Method	Criteria	2	4	2	-4	4	6
PGLS	LASSO	cAIC	0	0	0	0	0	0.290
		BIC	0	0.008	0	0	0	0.326
	$ALASSO_{0.5}$	cAIC	0	0.004	0	0	0	0.282
		BIC	0	0.008	0	0	0	0.318
	SCAD	cAIC	0	0.008	0	0	0	0.218
		BIC	0	0.012	0	0	0	0.278
	EN	cAIC	0	0	0	0	0	0.288
		BIC	0	0.008	0	0	0	0.330
PGLS-SPD	LASSO	cAIC	0	0	0	0	0	0.208
		BIC	0	0	0	0	0	0.232
	$ALASSO_{0.5}$	cAIC	0	0.004	0	0	0	0.234
		BIC	0	0.008	0	0	0	0.282
	SCAD	cAIC	0	0.008	0	0	0	0.214
		BIC	0	0.008	0	0	0	0.262
	EN	cAIC	0	0	0	0	0	0.200
		BIC	0	0	0	0	0	0.224
	Backward		0.001	0.036	0.002	0.039	0.003	0.507
	LARS		0.004	0.012	0.006	0	0	0.294

Table A.3: Type II error rate for the wind tunnel design at $\eta = 1$ for the PGLS and PGLS-SPD. $ALASSO_{0.5}$ is the ALASSO with $\psi = 0.5$.

	True active variable		s_2	w_1	s_2w_1	w_1w_2	s_1^2	w_1^2
	Method	Criteria	2	4	2	-4	4	6
PGLS	LASSO	cAIC	0	0.020	0	0.004	0	0.428
		BIC	0	0.032	0	0.016	0	0.452
	$ALASSO_{0.5}$	cAIC	0	0.032	0	0.012	0	0.432
		BIC	0	0.048	0	0.024	0	0.472
	SCAD	cAIC	0	0.044	0	0.020	0	0.324
		BIC	0	0.052	0	0.020	0	0.360
	EN	cAIC	0	0.020	0	0.004	0	0.432
		BIC	0	0.032	0	0.016	0	0.452
PGLS-SPD	LASSO	cAIC	0	0.020	0	0.004	0	0.356
		BIC	0	0.020	0	0.004	0	0.392
	$ALASSO_{0.5}$	cAIC	0	0.024	0	0.004	0	0.368
		BIC	0	0.040	0	0.016	0	0.436
	SCAD	cAIC	0	0.036	0	0.008	0	0.308
		BIC	0	0.036	0	0.016	0	0.336
	EN	cAIC	0	0.016	0	0.004	0	0.364
		BIC	0	0.020	0	0.004	0	0.392
	Backward		0	0.120	0	0.123	0	0.630
	LARS		0	0.048	0	0.016	0	0.432

Table A.4: Type II error rate for the wind tunnel design at $\eta = 10$ for the PGLS and PGLS-SPD. $ALASSO_{0.5}$ is the ALASSO with $\psi = 0.5$.

Appendix B

Type I and II error rates by PGLS and PGLS-SPD for the freeze dried-coffee design.

	True non-active variables	Criteria	s_4	ws_4	s_2s_3	s_3s_4	s_3^2
	Method	Criteria	0	0	0	0	0
PGLS	LASSO	cAIC	0.116	0.092	0.068	0.168	0.068
		BIC	0.188	0.212	0.168	0.280	0.152
	$ALASSO_{0.5}$	cAIC	0.068	0.068	0.052	0.128	0.044
		BIC	0.160	0.164	0.120	0.232	0.140
	SCAD	cAIC	0.104	0.096	0.068	0.140	0.052
		BIC	0.164	0.184	0.128	0.224	0.124
	EN	cAIC	0.116	0.096	0.072	0.168	0.064
		BIC	0.188	0.208	0.168	0.272	0.152
PGLS-SPD	LASSO	cAIC	0.116	0.080	0.060	0.124	0.056
		BIC	0.180	0.188	0.136	0.268	0.148
	$ALASSO_{0.5}$	cAIC	0.072	0.068	0.052	0.104	0.044
		BIC	0.144	0.142	0.094	0.204	0.124
	SCAD	cAIC	0.112	0.088	0.060	0.140	0.044
		BIC	0.164	0.180	0.128	0.224	0.124
	EN	cAIC	0.140	0.140	0.084	0.200	0.072
		BIC	0.208	0.224	0.176	0.300	0.164
	Backward		0.049	0.085	0.053	0.083	0.068
	LARS		0.076	0.104	0.076	0.168	0.092

Table B.1: Type I error rate the freeze dried-coffee design design at $\eta = 1$ for the PGLS and PGLS-SPD. $ALASSO_{0.5}$ is the ALASSO with $\psi = 0.5$.

	True non-active variables	Criteria	s_4	ws_4	s_2s_3	s_3s_4	s_3^2
	Method	Criteria	0	0	0	0	0
PGLS	LASSO	cAIC	0.108	0.080	0.068	0.140	0.068
		BIC	0.216	0.160	0.176	0.244	0.184
	$ALASSO_{0.5}$	cAIC	0.096	0.092	0.076	0.104	0.084
		BIC	0.164	0.148	0.132	0.176	0.148
	SCAD	cAIC	0.112	0.076	0.072	0.132	0.068
		BIC	0.216	0.160	0.152	0.224	0.156
	EN	cAIC	0.112	0.076	0.064	0.144	0.072
		BIC	0.220	0.160	0.172	0.248	0.192
PGLS-SPD	LASSO	cAIC	0.108	0.064	0.068	0.112	0.072
		BIC	0.196	0.144	0.148	0.216	0.156
	$ALASSO_{0.5}$	cAIC	0.096	0.092	0.076	0.096	0.080
		BIC	0.144	0.132	0.108	0.164	0.140
	SCAD	cAIC	0.116	0.068	0.068	0.116	0.068
		BIC	0.192	0.144	0.132	0.212	0.144
	EN	cAIC	0.136	0.088	0.088	0.144	0.088
		BIC	0.224	0.172	0.176	0.240	0.172
	Backward		0.070	0.078	0.052	0.058	0.078
	LARS		0.072	0.092	0.076	0.088	0.092

Table B.2: Type I error rate the freeze dried-coffee design design at $\eta = 10$ for the PGLS and PGLS-SPD. $ALASSO_{0.5}$ is the ALASSO with $\psi = 0.5$.

	True active var.	Criteria	w	s_1	s_2	s_3	ws_1	ws_2	ws_3	s_1s_2	s_1s_3	s_1s_4	s_2s_4	w^2	s_1^2	s_2^2	s_4^2
	Method		4	4	-3	2	-4	3	-1	4	3	2	1	4	2	-1	2
PGLS	LASSO	cAIC	0.152	0	0	0.024	0.004	0.128	0.608	0	0	0.260	0.420	0.508	0.644	0.940	0.640
		BIC	0.068	0	0	0.016	0.004	0.096	0.536	0	0	0.176	0.340	0.324	0.484	0.844	0.516
	$ALASSO_{0.5}$	cAIC	0.224	0	0	0.072	0.008	0.228	0.656	0	0	0.404	0.560	0.500	0.708	0.948	0.680
		BIC	0.068	0	0	0.040	0.004	0.112	0.596	0	0	0.208	0.408	0.336	0.520	0.848	0.536
	SCAD	cAIC	0.188	0	0.008	0.040	0.004	0.184	0.636	0	0	0.348	0.488	0.480	0.672	0.948	0.636
		BIC	0.076	0	0	0.028	0.004	0.140	0.592	0	0	0.244	0.440	0.312	0.552	0.856	0.548
	EN	cAIC	0.152	0	0	0.024	0.004	0.132	0.608	0	0	0.256	0.412	0.516	0.656	0.944	0.636
		BIC	0.068	0	0	0.016	0.004	0.096	0.540	0	0	0.176	0.340	0.324	0.488	0.836	0.508
PGLS-SPD	LASSO	cAIC	0.116	0	0	0.024	0.004	0.140	0.616	0	0	0.340	0.428	0.408	0.628	0.949	0.604
		BIC	0.052	0	0	0.016	0.004	0.100	0.556	0	0	0.204	0.360	0.300	0.488	0.860	0.496
	$ALASSO_{0.5}$	cAIC	0.196	0	0.004	0.072	0.008	0.240	0.656	0	0.004	0.432	0.564	0.496	0.688	0.948	0.680
		BIC	0.072	0	0	0.048	0.004	0.128	0.596	0	0	0.238	0.432	0.348	0.524	0.860	0.536
	SCAD	cAIC	0.140	0	0	0.052	0.008	0.200	0.640	0	0	0.368	0.500	0.452	0.644	0.944	0.624
		BIC	0.068	0	0	0.028	0.004	0.136	0.592	0	0	0.244	0.440	0.312	0.552	0.864	0.540
	EN	cAIC	0.088	0	0	0.020	0.004	0.116	0.580	0	0	0.268	0.372	0.352	0.588	0.936	0.572
		BIC	0.052	0	0	0.016	0.004	0.088	0.524	0	0	0.184	0.320	0.288	0.452	0.832	0.480
	Backward		0.292	0.001	0.028	0.139	0.028	0.265	0.754	0	0.003	0.392	0.634	0.561	0.681	0.908	0.661
	LARS		0.084	0	0.004	0.092	0.012	0.196	0.704	0	0	0.272	0.540	0.388	0.664	0.888	0.640

Table B.3: Type II error rate for the freeze dried-coffee design at $\eta = 1$ for PGLS and PGLS-SPD. $ALASSO_{0.5}$ is the ALASSO with $\psi = 0.5$.

	True active variables	Criteria	w	s_1	s_2	s_3	ws_1	ws_2	ws_3	s_1s_2	s_1s_3	s_1s_4	s_2s_4	w^2	s_1^2	s_2^2	s_4^2
	Method		4	4	-3	2	-4	3	-1	4	3	2	1	4	2	-1	2
PGLS	LASSO	cAIC	0.384	0	0	0	0	0	0.200	0	0	0	0.016	0.752	0.116	0.764	0.136
		BIC	0.100	0	0	0	0	0	0.152	0	0	0	0	0.464	0.032	0.464	0.052
	$ALASSO_{0.5}$	cAIC	0.332	0	0	0	0	0	0.232	0	0	0	0.028	0.708	0.116	0.736	0.156
		BIC	0.116	0	0	0	0	0	0.192	0	0	0	0.016	0.444	0.072	0.528	0.080
	SCAD	cAIC	0.404	0	0	0	0	0	0.196	0	0	0	0.020	0.720	0.108	0.792	0.144
		BIC	0.144	0	0	0	0	0	0.152	0	0	0	0.012	0.428	0.048	0.552	0.072
	EN	cAIC	0.384	0	0	0	0	0	0.196	0	0	0	0.008	0.752	0.104	0.760	0.136
		BIC	0.104	0	0	0	0	0	0.152	0	0	0	0	0.464	0.032	0.460	0.056
PGLS-SPD	LASSO	cAIC	0.220	0	0	0	0	0	0.196	0	0	0	0.016	0.592	0.108	0.824	0.128
		BIC	0.060	0	0	0	0	0	0.156	0	0	0	0	0.396	0.032	0.544	0.052
	$ALASSO_{0.5}$	cAIC	0.332	0	0	0	0	0	0.232	0	0	0	0.028	0.720	0.116	0.736	0.156
		BIC	0.088	0	0	0	0	0	0.204	0	0	0	0.020	0.428	0.084	0.568	0.096
	SCAD	cAIC	0.316	0	0	0	0	0	0.204	0	0	0	0.020	0.648	0.108	0.820	0.132
		BIC	0.104	0	0	0	0	0	0.160	0	0	0	0.012	0.424	0.044	0.568	0.072
	EN	cAIC	0.120	0	0	0	0	0	0.164	0	0	0	0.004	0.572	0.088	0.796	0.108
		BIC	0.060	0	0	0	0	0	0.140	0	0	0	0	0.392	0.020	0.516	0.052
	Backward		0.376	0	0	0	0	0.002	0.446	0	0	0.007	0.105	0.698	0.208	0.704	0.240
	LARS		0.100	0	0	0	0	0	0.356	0	0	0	0.052	0.448	0.120	0.580	0.128

Table B.4: Type II error rate for the freeze dried-coffee design at $\eta = 10$ for PGLS and PGLS-SPD. $ALASSO_{0.5}$ is the ALASSO with $\psi = 0.5$.

Appendix C

Type I and II error rates by SSVS and SSVS-SPD for the wind tunnel design.

	True non-active variable	s_1	w_2	s_1s_2	s_1w_1	s_1w_2	s_2w_2
	Method	0	0	0	0	0	0
$\eta = 1$	SSVS	0.046	0.090	0.047	0.042	0.040	0.053
	SSVS-SPD	0.032	0.071	0.026	0.050	0.032	0.038
$\eta = 10$	SSVS	0.010	0.273	0.018	0.018	0.017	0.019
	SSVS-SPD	0.024	0.262	0.009	0.006	0.005	0.008

Table C.1: Type I error rates for the wind tunnel design.

	True active variable	s_2	w_1	s_2w_1	w_1w_2	s_1^2	w_1^2
	Method	2	4	2	-4	4	6
$\eta = 1$	SSVS	0.003	0.003	0.001	0.001	0.003	0.191
	SSVS-SPD	0.002	0.004	0.006	0.006	0.007	0.171
$\eta = 10$	SSVS	0	0.003	0	0.011	0	0.151
	SSVS-SPD	0	0.017	0	0.003	0	0.169

Table C.2: Type II error rates for the wind tunnel design.

Appendix D

Type I and II error rates by SSVS and SSVS-SPD for the freeze dried-coffee design.

	True non-active variables	s_4	ws_4	s_2s_3	s_3s_4	s_3^2
	Method	0	0	0	0	0
$\eta = 1$	SSVS	0.223	0.545	0.213	0.396	0.606
	SSVS-SPD	0.134	0.426	0.135	0.268	0.467
$\eta = 10$	SSVS	0.175	0.674	0.115	0.354	0.728
	SSVS-SPD	0.101	0.589	0.108	0.369	0.688

Table D.1: Type I error rates for the freeze dried-coffee design.

	True active variables Method	w 4	s_1 4	s_2 -3	s_3 2	ws_1 -4	ws_2 3	ws_3 -1	s_1s_2 4	s_1s_3 3	s_1s_4 2	s_2s_4 1	w^2 4	s_1^2 2	s_2^2 -1	s_4^2 2
$\eta = 1$	SSVS	0.051	0	0.001	0.063	0	0.079	0.271	0	0.005	0.143	0.334	0.047	0.180	0.355	0.196
	SSVS-SPD	0.091	0	0.001	0.073	0.008	0.080	0.272	0	0.001	0.203	0.382	0.221	0.252	0.479	0.271
$\eta = 10$	SSVS	0.012	0	0	0	0	0	0.065	0	0	0.003	0.046	0.007	0.025	0.216	0.015
	SSVS-SPD	0.070	0	0	0	0	0	0.054	0	0	0	0.024	0.124	0.023	0.273	0.005

Table D.2: Type II error rates for the freeze dried-coffee design.

Appendix E

Assessment of MCMC samples from Chapter 5

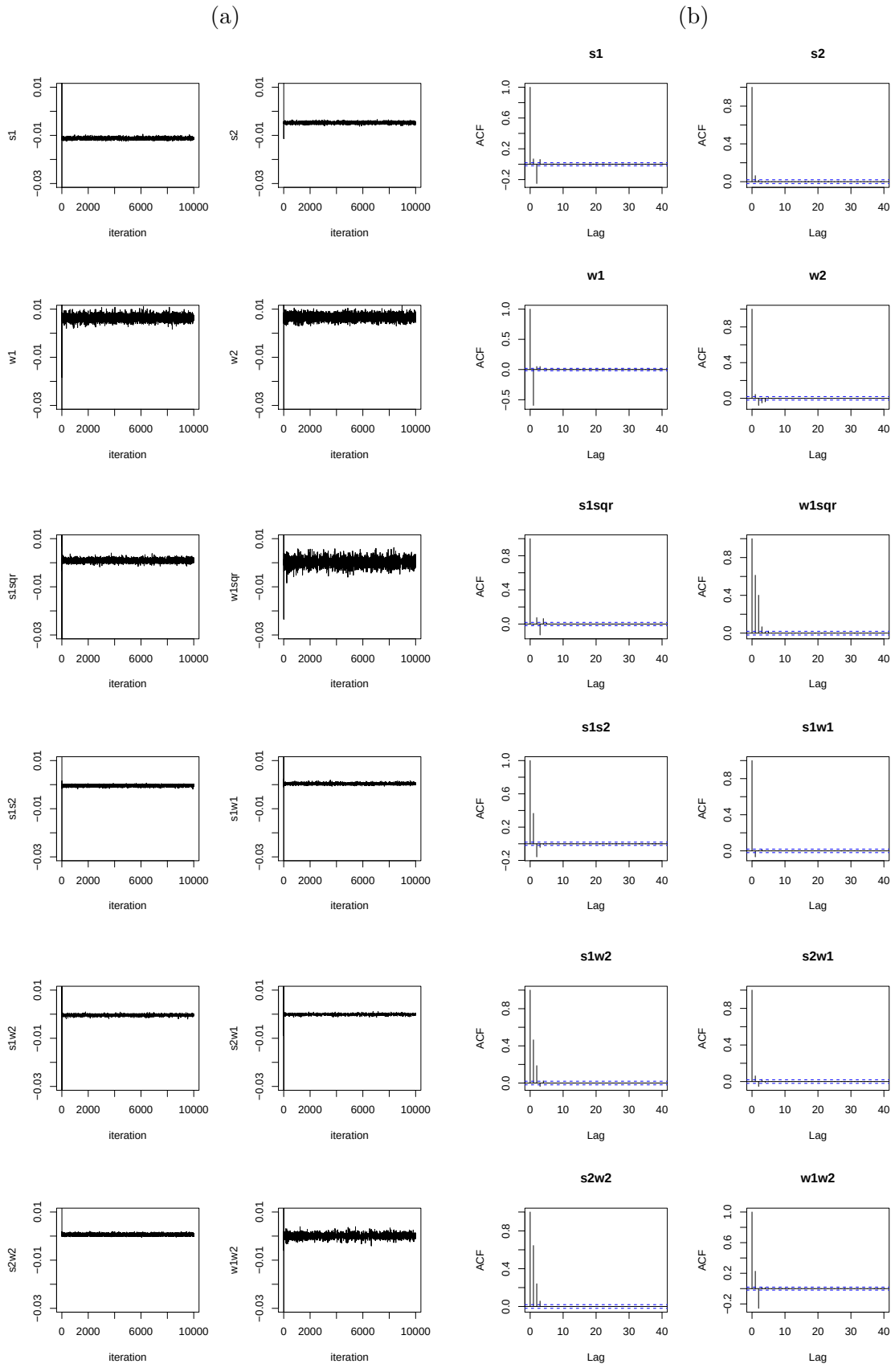


Figure E.1: Column (a) trace and column (b) ACF plot for the Markov chain formed by sampling β by SSVS for the wind tunnel design.

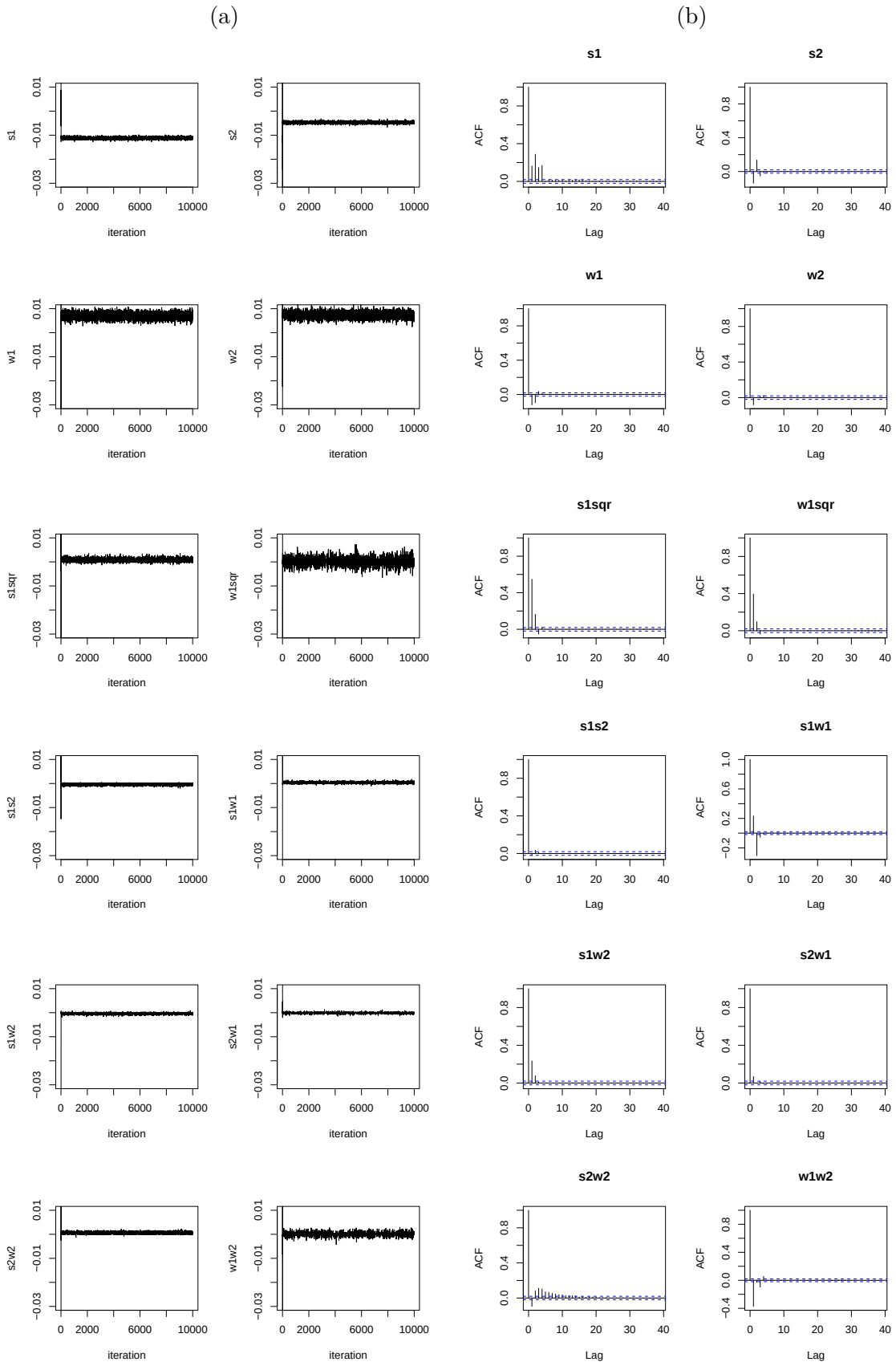


Figure E.2: Column (a) trace and column (b) ACF plot for the Markov chain formed by sampling β by SSVS-SPD for the wind tunnel design.

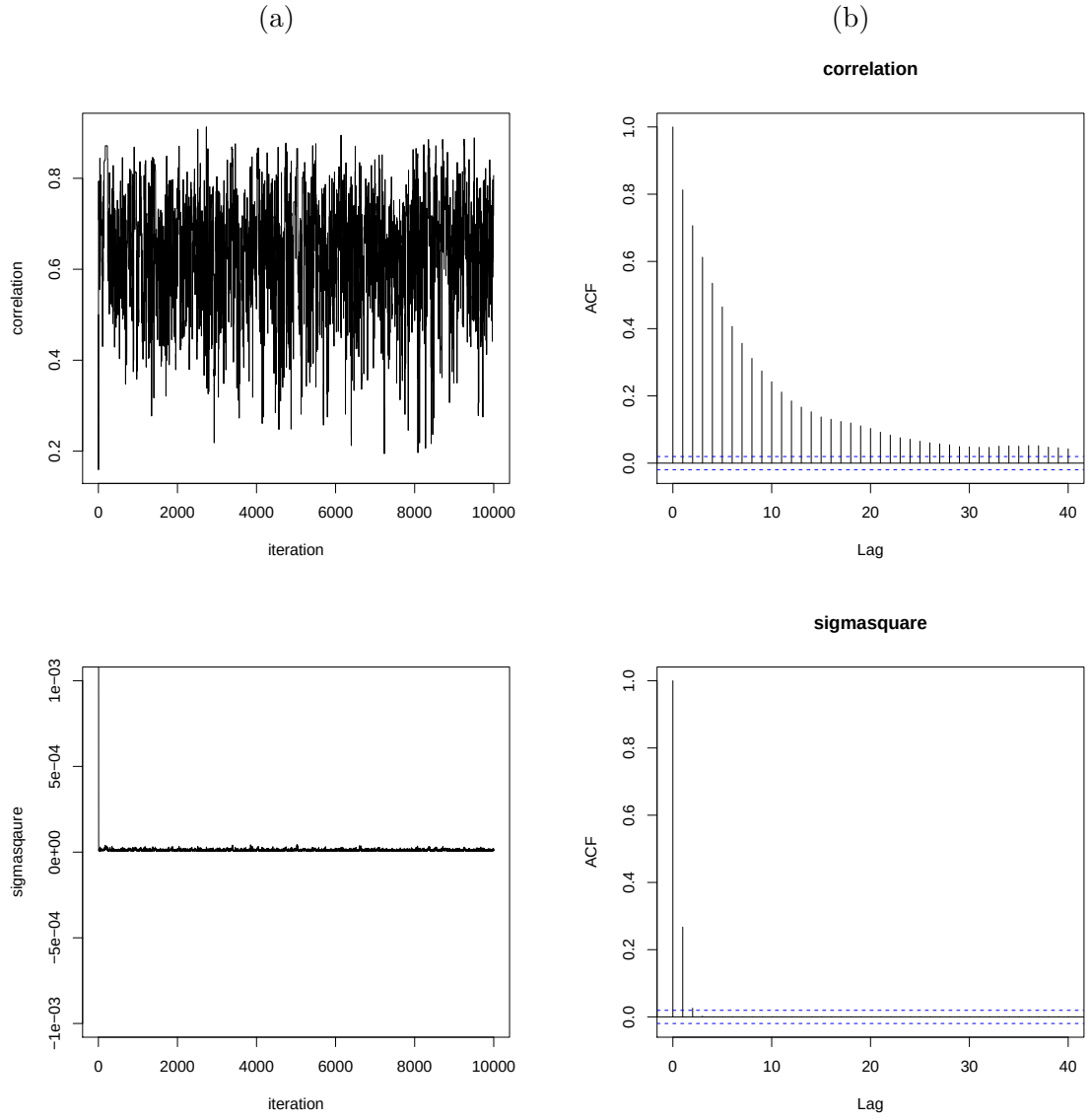


Figure E.3: Column (a) trace and column (b) ACF plot for the Markov chain formed by sampling the total variance σ^2 and the correlation ρ by SSVS for the wind tunnel design.

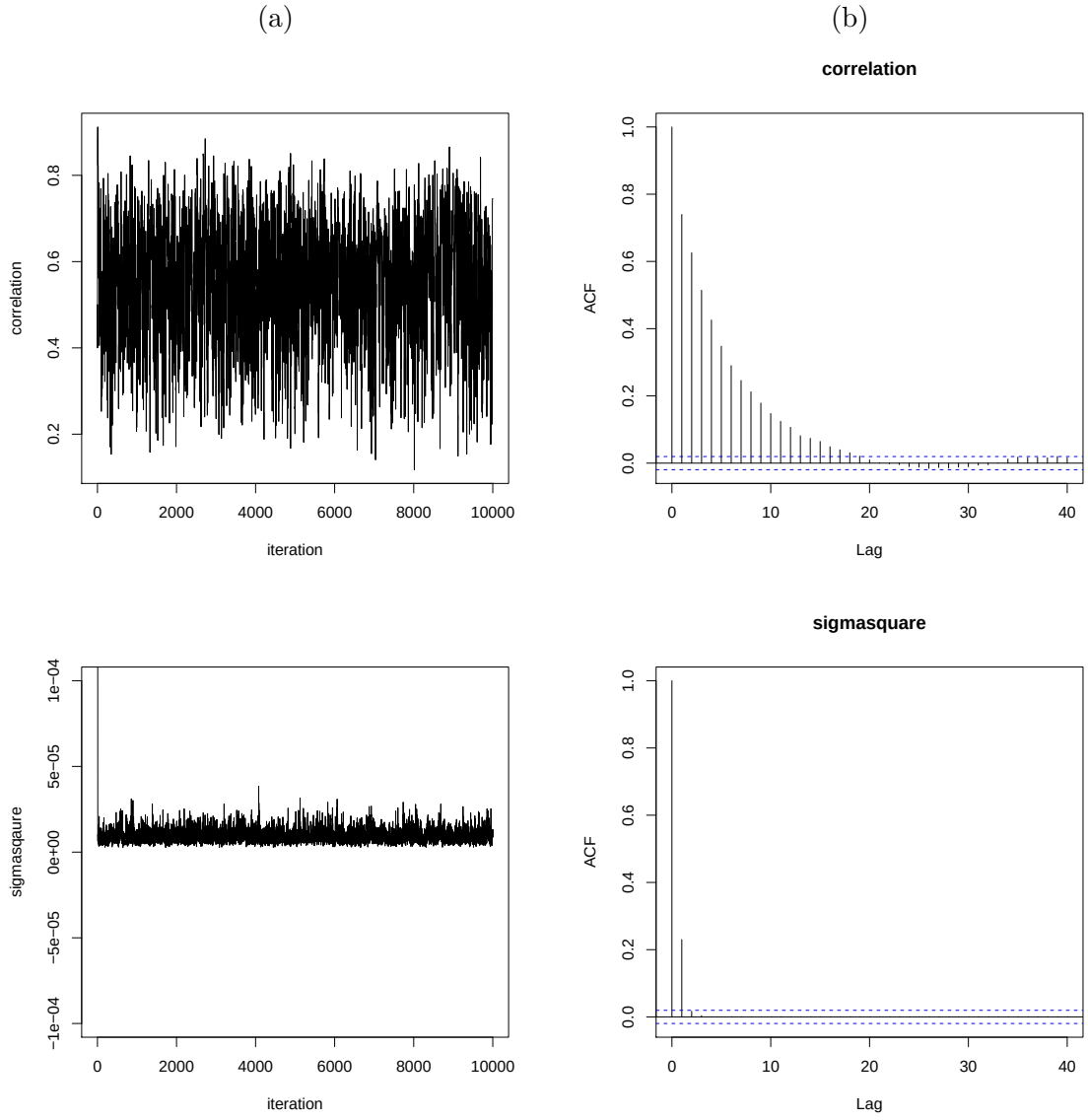


Figure E.4: Column (a) trace and column (b) ACF plot for the Markov chain formed by sampling the total variance σ^2 and the correlation ρ by SSVS-SPD for the wind tunnel design.

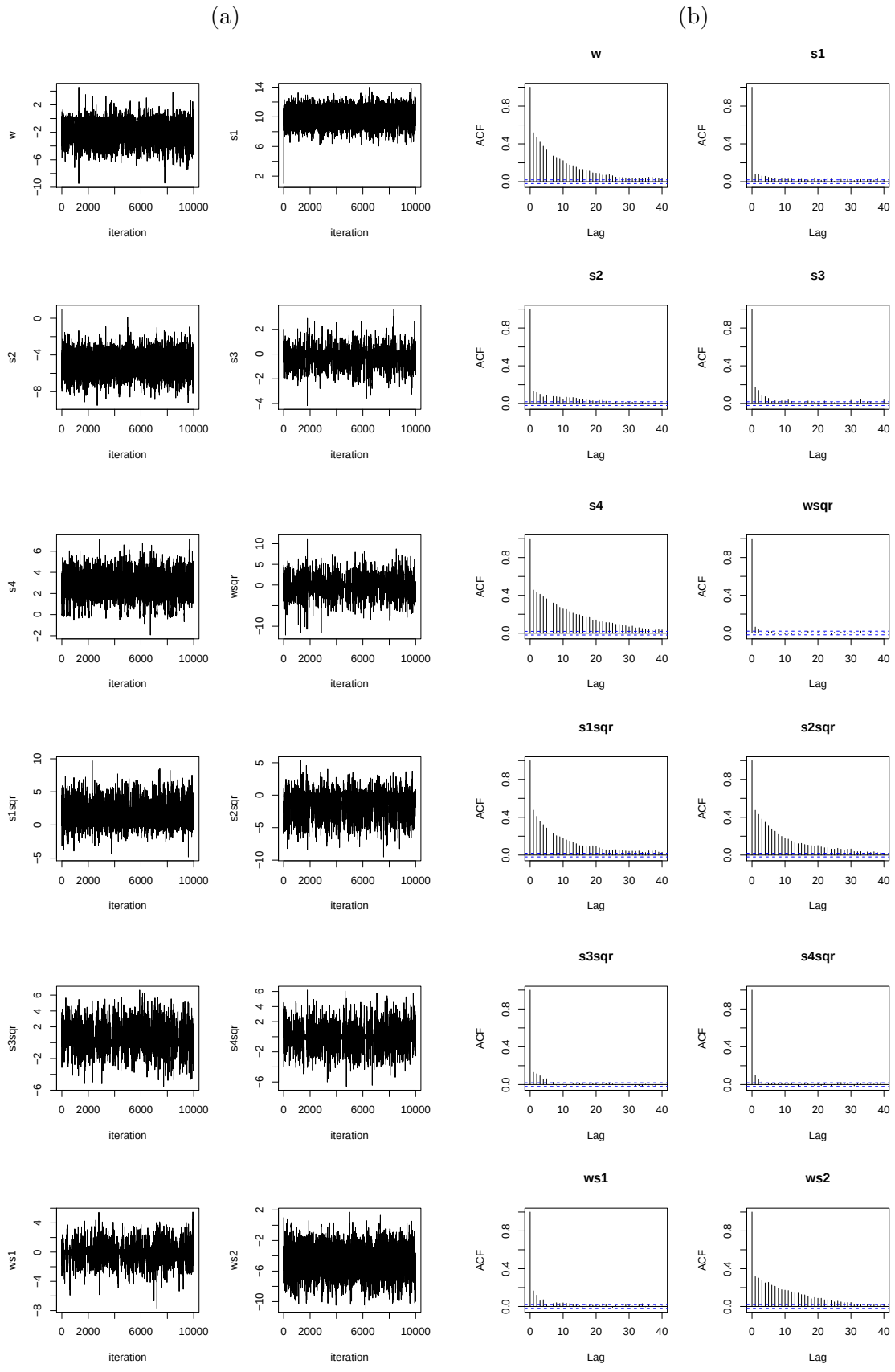


Figure E.5: Column (a) trace and column (b) ACF plot for the Markov chain formed by sampling β by SSVS for the freeze dried-coffee design.

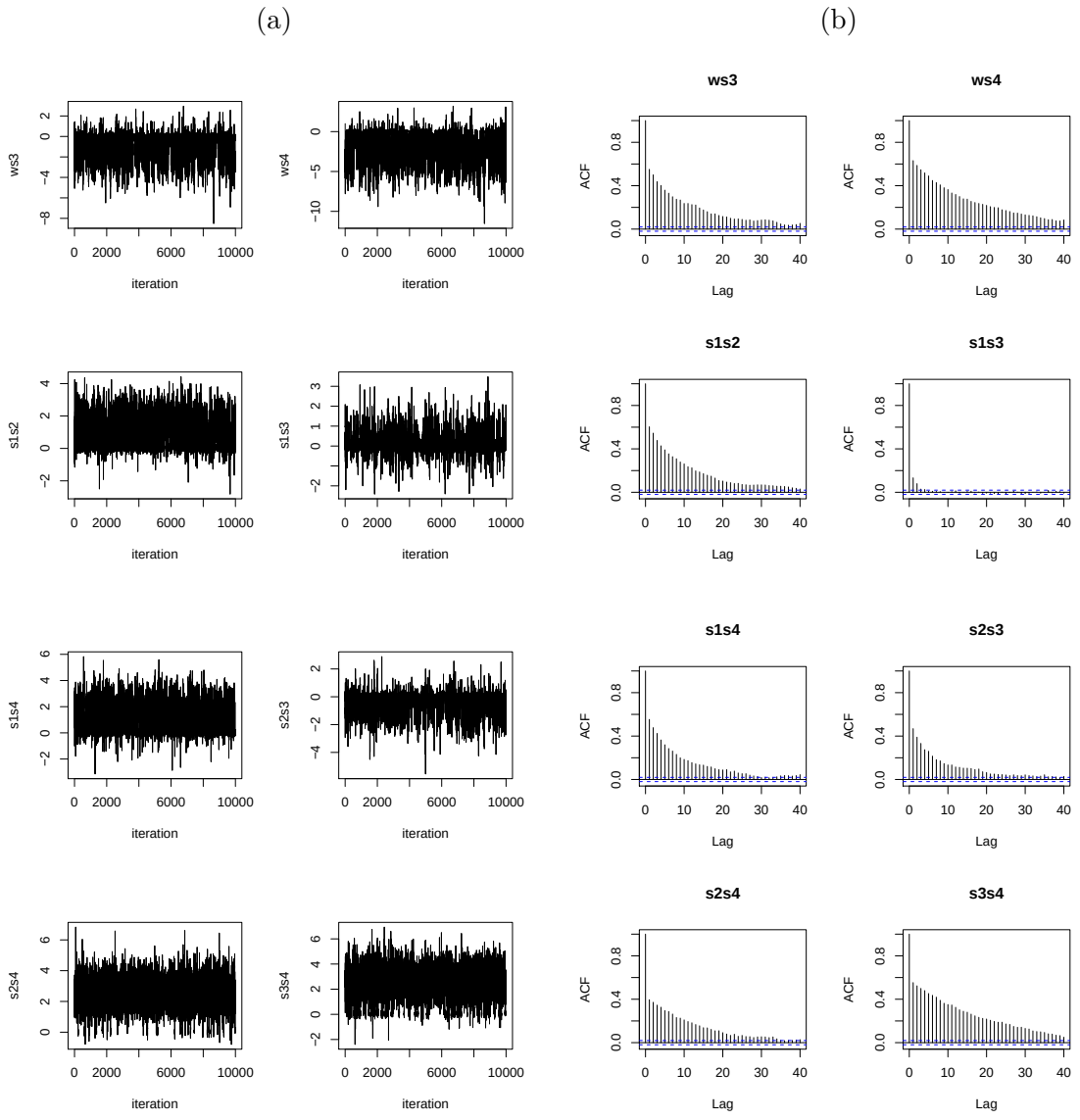


Figure E.6: Column (a) trace and column (b) ACF plot for the Markov chain formed by sampling β by SSVS for the freeze dried-coffee design.

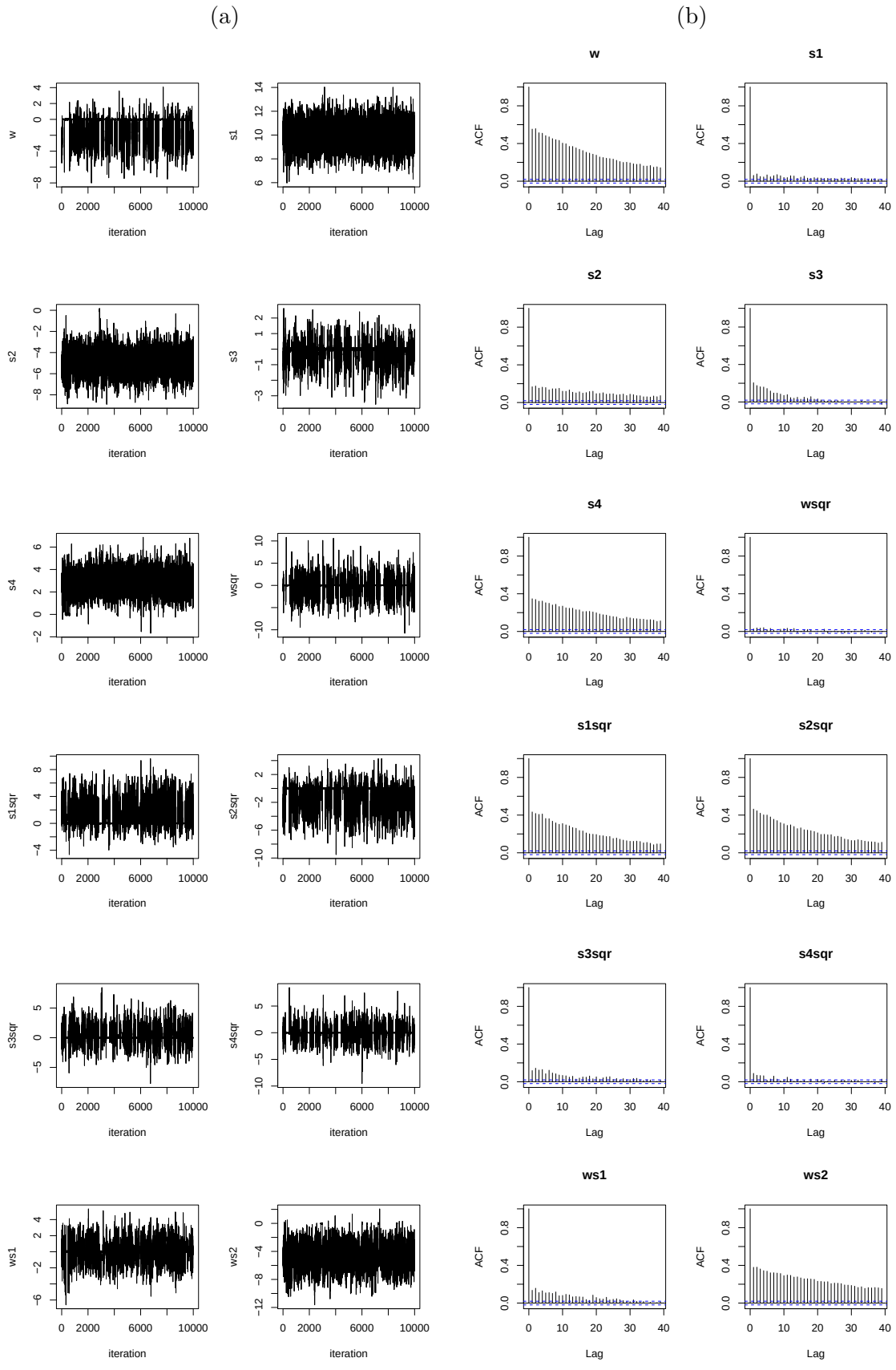


Figure E.7: Column (a) trace and column (b) ACF plot for the Markov chain formed by sampling β by SSVS-SPD for the freeze dried-coffee design.

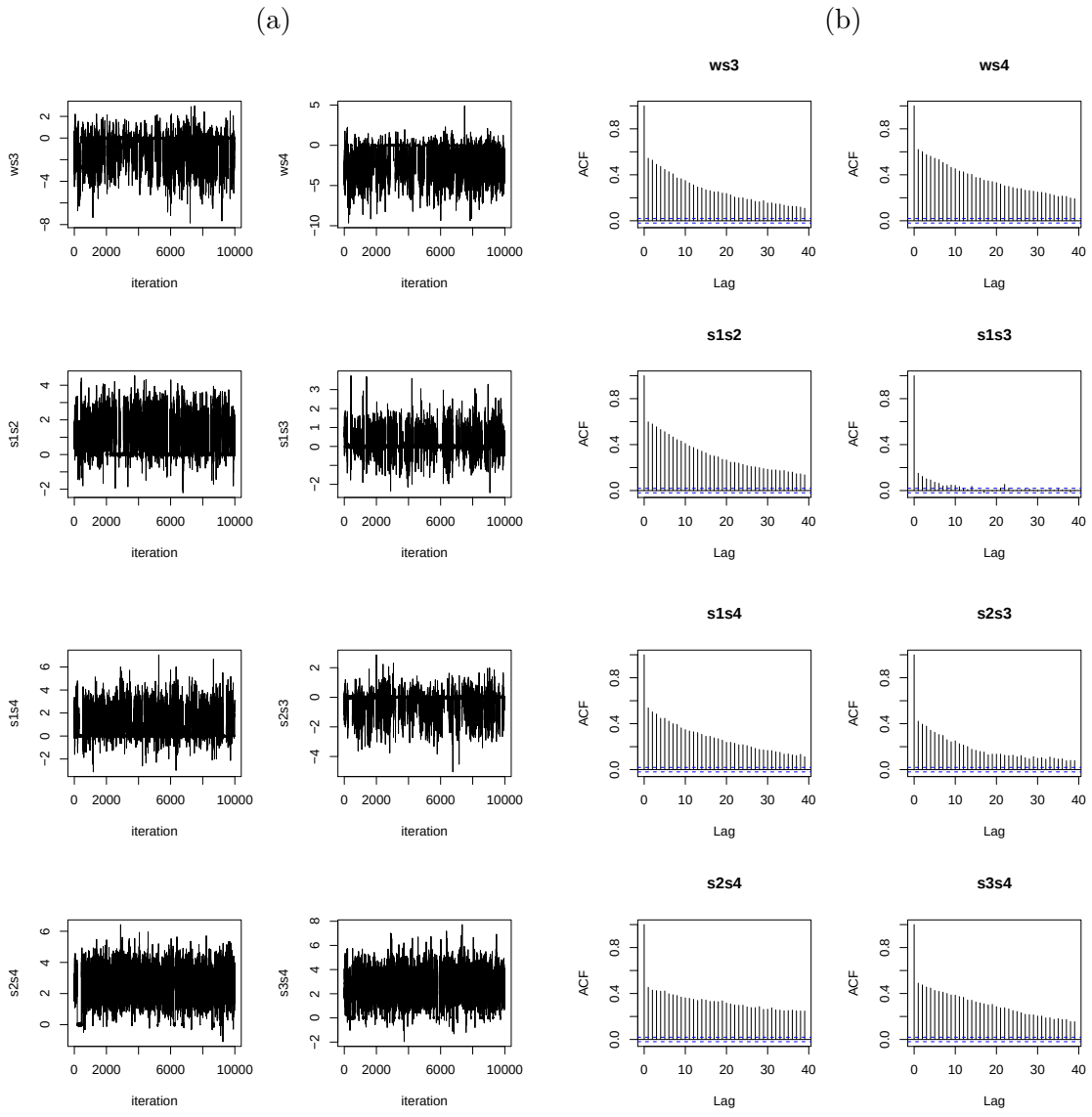


Figure E.8: Column (a) trace and column (b) ACF plot for the Markov chain formed by sampling β by SSVS-SPD for the freeze dried-coffee design.

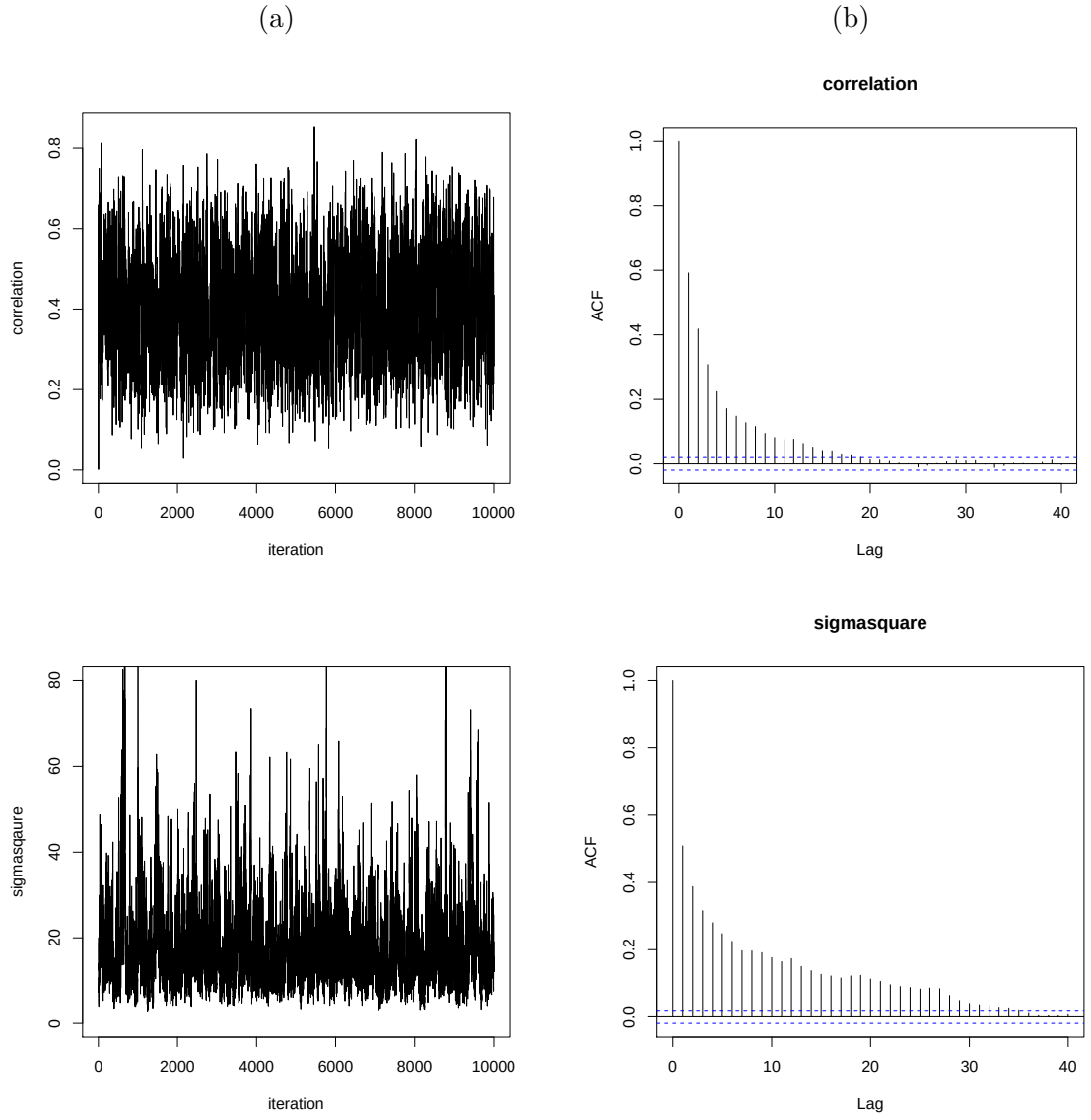


Figure E.9: Column (a) trace and column (b) ACF plot for the Markov chain formed by sampling the total variance σ^2 and the correlation ρ by SSVS for the freeze dried-coffee design.

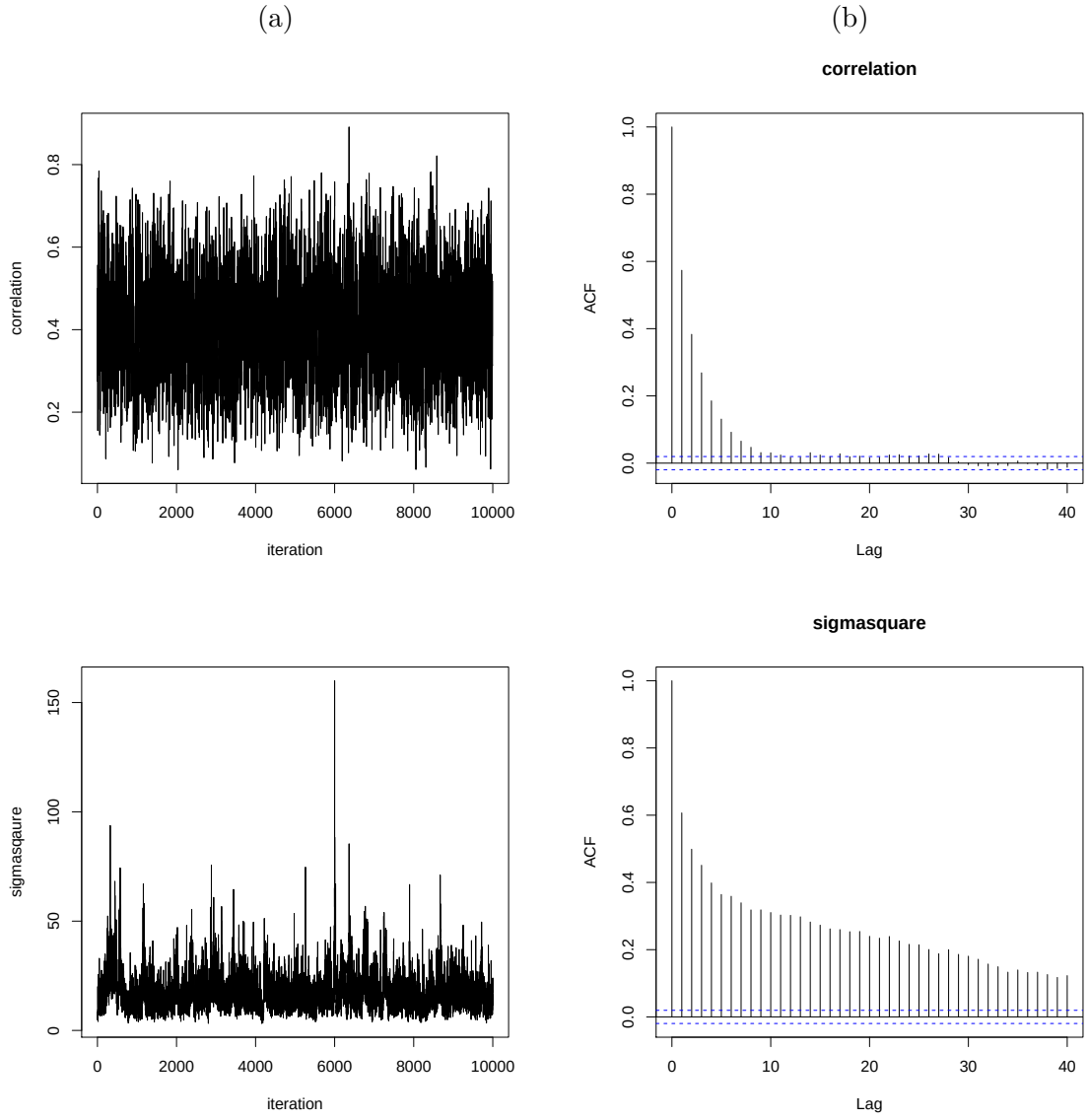


Figure E.10: Column (a) trace and column (b) ACF plot for the Markov chain formed by sampling the total variance σ^2 and the correlation ρ by SSVS-SPD for the freeze dried-coffee design.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory.*, pages 267–281. Akademiai Kiado.

Anbari, F. T. and Lucas, J. M. (1994). Super-efficient designs: how to run your experiment for higher efficiency and lower cost. In *Annual Quality Congress Proceedings-American Society for Quality Control*, pages 853–853.

Anderson, D. R. and Burnham, K. P. (1999). Understanding information criteria for selection among capture-recapture or ring recovery models. *Bird Study*, 46(sup1):S14–S21.

Barbieri, M. M., Berger, J. O., et al. (2004). Optimal predictive model selection. *The Annals of Statistics*, 32(3):870–897.

Bondell, H. D., Krishna, A., and Ghosh, S. K. (2010). Joint variable selection for fixed and random effects in linear mixed-effects models. *Biometrics*, 66(4):1069–1077.

Breiman, L. et al. (1996). Heuristics of instability and stabilization in model selection. *The Annals of Statistics*, 24(6):2350–2383.

Brooks, S. P. and Roberts, G. O. (1998). Convergence assessment techniques for markov chain monte carlo. *Statistics and Computing*, 8(4):319–335.

Corbeil, R. R. and Searle, S. R. (1976). Restricted maximum likelihood (reml) estimation of variance components in the mixed model. *Technometrics*, 18(1):31–38.

Cowles, M. K. and Carlin, B. P. (1996). Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–499.

Fan, J. (1997). Comments on wavelets in statistics: A review by a. antoniadis. *Journal of the Italian Statistical Society*, 6(2):131.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Fisher, R. A. (1925). *Statistical Methods for Research Workers*. Genesis Publishing Pvt Ltd.

Frank, L. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2):109–135.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2014). *Bayesian Data Analysis*, volume 2. CRC Press Boca Raton, FL.

Gelman, A., Roberts, G. O., Gilks, W. R., et al. (1996). Efficient metropolis jumping rules. *Bayesian Statistics*, 5(599-608):42.

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

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian variable selection. *Statistica Sinica*, pages 339–373.

Geweke, J. (1996). Variable selection and model comparison in regression. *In Bayesian Statistics 5*.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. (1995). *Markov Chain Monte Carlo in Practice*. CRC press.

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 (Applied Statistics)*, 58(4):467–484.

Gilmour, S. G. and Trinca, L. A. (2000). Some practical advice on polynomial regression analysis from blocked response surface designs. *Communications in Statistics-Theory and Methods*, 29(9-10):2157–2180.

Goos, P. (2012). *The Optimal Design of Blocked and Split-Plot Experiments*, volume 164. Springer Science & Business Media.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). Unsupervised learning. In *The Elements of Statistical Learning*, pages 485–585. Springer.

Hastings, W. K. (1970). Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika*, 76(2):297–307.

Ibrahim, J. G., Zhu, H., Garcia, R. I., and Guo, R. (2011). Fixed and random effects selection in mixed effects models. *Biometrics*, 67(2):495–503.

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

Kowalski, S. M. and Potcner, K. J. (2003). How to recognize a split-plot experiment. *Quality Progress*, 36(11):60.

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

Li, R. and Lin, D. K. (2003). Analysis methods for supersaturated design: some comparisons. *Journal of Data Science*, 1(3):249–260.

Mansfield, E. R. and Helms, B. P. (1982). Detecting multicollinearity. *The American Statistician*, 36(3a):158–160.

Matthews, E. S. (2015). *Design of factorial experiments in blocks and stages.* PhD thesis, University of Southampton.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Miller, A. (1997). Strip-plot configurations of fractional factorials. *Technometrics*, 39(2):153–161.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83(404):1023–1032.

Mylona, K. and Goos, P. (2011). Penalized generalized least squares for model selection under restricted randomization. Manuscript submitted for Isaac Newton Institute for Mathematical Sciences; NI 11032.

O'Hagan, M. G. and Forster, J. (2004). *Kendall's Advanced Theory of Statistics: Bayesian Statistics. Vol. 2B.* Arnold.

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(482):681–686.

Peck, R., Olsen, C., and Devore, J. L. (2015). *Introduction to Statistics and Data Analysis.* Cengage Learning.

Schelldorfer, J., Bühlmann, P., DE, G., and VAN, S. (2011). Estimation for high-dimensional linear mixed-effects models using 1-penalization. *Scandinavian Journal of Statistics*, 38(2):197–214.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464.

Simpson, J. R., Kowalski, S. M., and Landman, D. (2004). Experimentation with randomization restrictions: Targeting practical implementation. *Quality and Reliability Engineering International*, 20(5):481–495.

Sinharay, S. (2003). Practical applications of posterior predictive model checking for assessing fit of common item response theory models. *ETS Research Report Series*, 2003(2).

Sugiura, N. (1978). Further analysts of the data by akaike's information criterion and the finite corrections: Further analysts of the data by akaike's. *Communications in Statistics-Theory and Methods*, 7(1):13–26.

Tan, M. H. and Wu, C. (2013). A Bayesian approach for model selection in fractionated split plot experiments with applications in robust parameter design. *Technometrics*, 55(3):359–372.

Tang, X., Xu, X., Ghosh, M., and Ghosh, P. (2016). Bayesian variable selection and estimation based on global-local shrinkage priors. *Sankhya A*, pages 1–32.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):385–395.

Toothaker, L. E. and Miller, L. (1996). *Introductory Statistics for the Behavioral Sciences*. Thomson Brooks/Cole Publishing Co.

Wahba, G. (1980). Spline bases, regularization, and generalized cross-validation for solving approximation problems with large quantities of noisy data. *Approximation Theory III*, 2.

Yue, L. H. (2010). Cost-efficient variable selection using branching lars.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, 37(4):1733.