UNIVERSITY OF SOUTHAMPTON

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS

School of Mathematics

Optimal and Sequential Design for Bridge Regression with Application in Organic Chemistry

by

Sarah Beth Carnaby

Thesis submitted for the degree of Doctor of Philosophy

September 2010

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING, SCIENCE AND MATHEMATICS

MATHEMATICS

Doctor of Philosophy

OPTIMAL AND SEQUENTIAL DESIGN FOR BRIDGE REGRESSION WITH APPLICATION IN ORGANIC CHEMISTRY

by Sarah Beth Carnaby

This thesis presents and applies methods for the design and analysis of experiments for a family of coefficient shrinkage methods, known collectively as bridge regression, with emphasis on the two special cases of ridge regression and the lasso. The application is the problem of understanding and predicting the melting point of small molecule organic compounds using chemical descriptors.

Experiments typically have a large number of predictors compared to the number of observations, and high correlations between pairs of predictors. In this thesis, bridge regression is used to select linear models which are then compared to models selected by more commonly used methods of variable selection, such as subset selection and stepwise selection. Models including two-way product, or interaction, terms are also considered.

A general method is developed for the selection of an optimal design when accurate estimates of the model coefficients are required. The method exploits a relationship between bridge regression and Bayesian methods which is used to develop a class of D-optimal designs. A necessary approximation to the variance-covariance matrix of coefficient estimators is derived. Designs are found using algorithmic search for ridge regression and the lasso, for experiments with (a) two-level factors and (b) the motivating chemistry problem. Comparisons are made with alternative designs.

A sequential design criterion is developed to enhance an existing design. The criterion selects additional design points, from a finite set of candidate points, that exhibit the highest estimated prediction variance obtained from bootstrapping. The method is applied to the Bayesian D-optimal designs and is shown to be capable of improving design performance through the addition of only a small number of runs.

Contents

List of Figures

List of Tables

DECLARATION OF AUTHORSHIP

I, SARAH BETH CARNABY, declare that the thesis entitled

OPTIMAL AND SEQUENTIAL DESIGN FOR BRIDGE REGRESSION WITH APPLICATION IN ORGANIC CHEMISTRY

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

I would like to thank all those that have contributed to the completion of this thesis, particularly my supervisors Dr Dave Woods and Professor Mike Hursthouse for their invaluable guidance and support, and Professor Sue Lewis for additional guidance and suggestions. I am also grateful to Dr Terry Threlfall, Dr Thomas Gelbrich and Dr Yuanxin Ding in the School of Chemistry at the University of Southampton for the preparation of samples of the organic compounds, crystal structure determination and data collection, and for their advice with respect to scientific interpretation of the results. I would also like to thank Pfizer UK, in particular the members of the non-clinical statistics and computational chemistry groups, for their support through the EPSRC CASE Award scheme.

Chapter 1

Introduction

1.1 Background

During the course of many scientific and industrial experiments, particularly in research science carried out in laboratories, observations are routinely made on the output of processes which are influenced by many possible explanatory variables or factors. It is important to make use of such data sets and to understand what they tell us about the mechanisms that generated them. Where the observed values of the output or response are seen to vary, it is of interest to see if and how this change can be related to one or more of the variables. An important consequence of being able to approximate such a relationship is that it enables us to predict the value of an unobserved response based on the values of the variables, e.g. predicting the yield of a compound based on the known concentrations of reagents used for its synthesis. This knowledge can lead to improvements in the product or process.

Regression analysis is often used to approximate the important underlying trends from a data set. In this method, a data set of values of explanatory variables is related to one or more corresponding responses in order to identify the key variables and build a predictive model. This model can then be used to predict the response of observations corresponding to different values of the explanatory variables, and to estimate the accuracy of the predictions.

The scientific design of experiments can be used to improve the cost-effectiveness of gathering data to investigate relationships. This technique involves obtaining the data set by making observations on a careful choice of the combinations of values of

a set of explanatory variables (known as the design points) that are run in the experiment. These design points are not necessarily distinct. The principles of design also help in choosing an economic size of the experiment, and hence reducing the amount of time needed to run the experiment, and in randomising the experiment to reduce bias in how the experiment is run. These features of designing or planning the collection of data are particularly useful when resources, such as time and materials, are limited. In an industrial setting, when the quality of the end product is of particular importance, experimental design can help to plan the manufacturing process so as to improve the end product or satisfy a standard of quality.

In this thesis, we assume there are f quantitative variables or factors and p predictors that are defined as known functions of one or more variables. We suppose there are N observations taken on a response which are held in a vector $\boldsymbol{Y}$, and described by a linear model

$$\boldsymbol{Y} = \mathbf{1}_N \beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1.1}$$

where $(\mathbf{1}_N \vdots \boldsymbol{X})$ is an $N \times (p+1)$ model matrix, β_0 is the intercept, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^{\mathrm{T}}$ is a p-vector of unknown coefficients and $\boldsymbol{\varepsilon} \sim N(0, \sigma^2)$ is the N-vector of independent and identically distributed random errors. The (i, j)th entry of $\boldsymbol{X}$ is the value in the ith run taken by the jth predictor $(i = 1, \ldots, N; j = 1, \ldots, p)$.

During data analysis, it is frequently necessary to select a subset of variables to include in the fitted model, since the full least squares estimate of (1.1) is often found to be inadequate in describing the data, particularly when a large number of predictors is considered. This inadequacy is due to the least squares estimator of $\boldsymbol{\beta}$ having low bias and high variance, which decreases the prediction accuracy for new observations. The fitted model will also have a large number of terms which can be difficult to interpret scientifically. Variable selection is also needed when there are more predictors than observations and there are not enough available degrees of freedom to fit the full ordinary least squares model (that is, $(p + 1) > N$).

These problems can be solved in two ways:

- discrete variable selection (Miller, 2002) in which a proper subset of the set of all the predictors is chosen,
- coefficient shrinkage (Hastie, Tibshirani and Friedman, 2009, Chapter 3) in

which the estimates of the coefficients are deliberately biased towards zero.

These methods enable a subset of variables that exhibit the strongest relationship with the response to be identified for inclusion in the fitted model. Both these methods are discussed and applied in the thesis.

Generally, in the literature, the methods are applied using model (1.1) by taking the predictors to be only the variables themselves. The work presented here is developed for the general model (1.1) and the applications allow the predictors to include both the variables and products of pairs of variables so that the possibility of interactions between pairs of variables can be investigated.

1.2 Motivating Data Set from Organic Chemistry

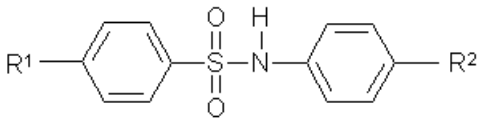
We now describe a motivating experiment and data set that will be used throughout this thesis. In simple terms, the melting point of a compound is the temperature at which the compound melts. It is an important property to chemists who aim to develop novel compounds that exhibit particular thermophysical behaviour. For example, melting point is well known to be closely related to solubility which is important in the oral administration of pharmaceutical solid forms to patients. The melting point may depend on any number of variables related to the molecular and crystal structure of the compound. The number of potential variables, and the possible complexity of their relationship with melting point, presents difficulties in building predictive models for this problem at present, see Section 3.2.

To study the relationship between the properties of organic compounds and their melting points, we investigate a data set based on properties of compounds, which we will refer to as the *Melting Point Data Set* throughout the thesis. Associated with each of these compounds is a large set of explanatory variables, called descriptors. These variables are either physical properties, describing the physical state of the compound such as molecular dipoles, molecular surface area and molecular weight, or chemical properties, describing how a compound will behave during a chemical reaction such as enthalpy of fusion, length of hydrogen bonds and the number of molecules that closely surround a central molecule in the crystal structure. The variables are either obtained by calculations applied directly to the molecular structure (e.g. molecular weight) or collected by measurement (e.g. crystal structure

properties obtained through single crystal X-ray diffraction).

The data set was collected from an experiment performed in the School of Chemistry at the University of Southampton (see Ding, 2007, for the background to a similar experiment). It consists of observations on melting point for a set of sixty 4,4'-disubstituted benzenesulfonamidobenzenes, R^1-C_6H_4-SO_2-NH-C_6H_4-R^2, where R^1, R^2=NO_2, CN, CF_3, I, Br, Cl, F, H, Me or OMe (Gelbrich, Hursthouse and Threlfall, 2007). These compounds are small organic molecules which share the same central molecular structure core, differing in the functional groups, R^1 and R^2, at opposite ends of the structure. Functional groups are groups of atoms within a molecule that contribute to the chemical reactivity of the compound. Figure 1.1 shows the skeletal molecular formula of the compounds, where R^1 and R^2 indicate the positions at which the changes in functional group are made.

Figure 1.1: Generalised skeletal molecular formula of the compounds of the Melting Point Data Set



The compounds used to obtain this data set are closely related in molecular structure, and this enables the dimensionality of the problem to be reduced, since the number of descriptors that vary between compounds will be lower than for a group of less closely related compounds. This restriction on the compounds aimed to enable a better understanding to be gained from the study about why particular relationships between properties are observed. In total there are 21 variables, 8 of which were calculated directly from the molecular structure; the remaining 13 were only available by physical measurement. Further details on the data set are given in Chapter 3. The complete data set is given in Appendix A.

The data displays a high degree of multicollinearity because molecules of similar structure tend to have the same values for many of the variables. For this reason, shrinkage regression methods are applied in the thesis to build predictive models for melting point. A key aspect of the data set is the limitations on the combinations of values of the descriptors (variables) that occur. Studies on compounds using descriptors present problems for the design of experiments because the values of the

descriptors are defined by the compounds available for experimentation and cannot be selected independently of each other. These issues are addressed in the thesis (Chapters 4 and 5).

1.3 Aims and Outline of the Thesis

The work in this thesis is concerned primarily with developing and applying statistical methodology to the problem of predicting the melting point of small molecule organic compounds, through both designing effective experiments and analysing the resulting data.

Inherent to building a descriptive model for this problem is the issue of variable selection. This thesis aims to investigate the application of a family of coefficient shrinkage methods, collectively known as bridge regression, to identify a subset of descriptors that exhibit the strongest relationship with the response.

A further, equally important, aim of the work is the selection of design points at which responses should be observed to obtain efficient (i.e. low variance) estimates of the parameters in model (1.1) using bridge regression. We develop methods for finding designs for two problems:

- the selection of an optimal design for a batch experiment,
- the construction of a point sequential (or adaptive) design to enhance an existing data set.

To obtain experimental data for the prediction of melting points of organic compounds, these compounds must first be synthesised, which is costly in terms of time and materials. Many of the data collection methods themselves, for example X-ray diffraction to obtain crystal structure properties, are also expensive and the processes involved are time consuming. Thus we aim to provide an effective criterion to guide the selection of design points which would be a useful tool to improve the efficiency of the data collection process in terms of time, materials and costs. Typically these kinds of experiments have many descriptors whose values can be computed simply and cheaply. This results in experiments where the number of predictors p is bigger than N, or around the value of N.

The remainder of the thesis is structured as follows. In Chapter 2, bridge regression and the special cases of ridge regression and the lasso are reviewed, including methods of estimating the model parameters and the variance of the resulting estimators. In Chapter 3, we investigate the application of variable selection and regression methods to the Melting Point Data Set.

In Chapter 4, we obtain D-optimal experimental designs for the efficient estimation of model (1.1) using bridge regression in general, and in particular for ridge regression and the lasso using a Bayesian approach. A normal approximation for the posterior distribution for the lasso is derived in order to find designs, and Bayesian inference for general bridge regression is discussed. Also, D-optimal designs are obtained for the chemistry experiment, and designs are obtained for a simulated example. In Chapter 5, the sequential selection of design points is developed as a means of improving an existing data set for the estimation of model (1.1) via bridge regression, and a criterion for selecting and assessing design points is proposed. The proposed criterion is applied to sequentially improve the D-optimal designs of Chapter 4, and the results are critically assessed by comparison with random sampling. Finally, in Chapter 6, some conclusions are drawn and the work is summarised, together with suggestions for possible directions for further work.

Chapter 2

Review of Bridge Regression and its Applications

2.1 Bridge Regression

In this chapter, we review a family of coefficient shrinkage methods called bridge regression, first introduced by Frank and Friedman (1993). We examine four available algorithms for estimation and also methods for selecting values for the tuning parameter. A comparison of methods is made for the lasso using a simulation. Also, the methods are compared in application to a data set on prostate cancer (Stamey et al., 1989). Four methods of approximating the variance-covariance matrix of the coefficient estimators are also reviewed and compared on the same data set.

In situations where there is a large number of predictors (relative to the number of observations), bridge regression methods may provide an estimate of model (1.1) which has lower prediction error than standard regression using ordinary least squares (OLS), in situations where OLS can be applied. Prediction error can be measured using mean squared error (MSE), defined by

$$\text{MSE}(\boldsymbol{\beta}) = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \beta_0 - \boldsymbol{x}_i^{\mathrm{T}}\boldsymbol{\beta})^2, \tag{2.1}$$

where $\boldsymbol{x}_i^{\mathrm{T}}$ is a p-vector of predictors for the ith observation, holding the ith row of matrix $\boldsymbol{X}$.

The reduction in MSE is achieved through a variance-bias trade-off: as the com-

plexity of a regression model increases, for example, through including more terms in the model, the variance increases and the bias simultaneously decreases. Including more terms allows the model to adapt to more complicated relationships in the data and reduces bias by reducing the difference between the average of the predicted and true mean responses. However, a model with too many terms may overfit the data, and have few degrees of freedom remaining to estimate the variance. Such overfitting leads to a model which may not describe future data well. A model selection criterion will try to choose the model complexity to trade off the size of variance of predictions against bias in order to minimise the prediction error.

Bridge regression can also alleviate problems of multicollinearity, that is, a situation where there is large correlation between pairs of predictors leading to estimators of the coefficients of model (1.1) which have high variances. Bridge regression may also produce parsimonious models (with small numbers of terms) that are consequently easier to interpret scientifically.

For simplicity, throughout the thesis, we refer to a fitted model for model (1.1) that is fitted using bridge regression as a 'bridge regression model'.

In bridge regression, the coefficient estimates, $\hat{\boldsymbol{\beta}}$, of the linear model (1.1) are found by minimising the penalised residual sum of squares

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left[(\boldsymbol{Y} - \mathbf{1}_N\beta_0 - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{Y} - \mathbf{1}_N\beta_0 - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j|^{\gamma} \right], \tag{2.2}$$

where $0 < \gamma \leq 2$, and $\lambda \geq 0$ is a tuning parameter called the complexity parameter. Note that $\lambda = 0$ is the special case of OLS. The tuning parameter controls the degree of shrinkage in the coefficient estimates, with shrinkage increasing as λ increases. Equation (2.2) can also be written as a constrained minimisation problem

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(Y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2, \tag{2.3}$$

subject to

$$\sum_{j=1}^{p} |\beta_j|^{\gamma} \leq t,$$

where $t \geq 0$ is a tuning parameter to be chosen by model selection (see Section 2.3).

Fu (1998) stated that the minimisation problems (2.2) and (2.3) are equivalent in the sense that for a given $\lambda \geq 0$, there is a value of $t \geq 0$ that produces the same parameter estimates. In this thesis both formulations of the problem are used.

Hastie et al. (2009, page 63), and other authors, have noted that, for data where a full least squares solution exists, bridge regression coefficient estimates gradually and independently shrink towards zero as $t \to 0$, away from the least squares solution, at different rates. Thus the set of possible predictors can be screened to detect those that have an important effect on the response. Consequently, the bridge regression method can be viewed as a more continuous method of variable selection than, for example, discrete subset selection.

The shrinkage parameter, $0 < \gamma \leq 2$, indicates different types of coefficient shrinkage and below we describe two important special cases of wide application. Alternatively, the value of γ may be selected according to the data via a model selection criterion, see for example Fu (1998).

2.1.1 Ridge Regression

When $\gamma = 2$, bridge regression is identical to ridge regression, which was first introduced by Hoerl and Kennard (1970). Here, the parameters are estimated not by OLS, but by equation (2.2) or (2.3) with $\gamma = 2$, where the second expression gives

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(Y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2, \tag{2.4}$$

subject to

$$\sum_{j=1}^{p} \beta_j^2 \leq t.$$

The ridge estimates are linear functions of the response, and there is a closed form solution for the coefficient estimates $\hat{\boldsymbol{\beta}}$. The ridge regression method does not shrink any of the coefficient estimates to zero. Instead $\hat{\beta}_j \to 0$ as $t \to 0$, see for example Draper and Smith (1998, Chapter 17).

2.1.2 The Lasso

A further approach to coefficient shrinkage is the least absolute shrinkage and selection operator (lasso), first introduced by Tibshirani (1996), which is obtained from equations (2.2) or (2.3) using $\gamma = 1$. Equation (2.2) gives

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left[(\boldsymbol{Y} - \mathbf{1}_N\beta_0 - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{Y} - \mathbf{1}_N\beta_0 - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \sum_{j=1}^{p} |\beta_j| \right], \tag{2.5}$$

where $\lambda \geq 0$, and (2.3) gives

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{i=1}^{N} \left(Y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j \right)^2, \tag{2.6}$$

subject to

$$\sum_{j=1}^{p} |\beta_j| \leq t.$$

Unlike ridge regression, there is no closed form solution for estimating the model parameters, with the lasso estimates being non-linear and non-differentiable functions of the response. Different methods for estimating the lasso parameters, as well as the closed form solution for estimating the ridge regression parameters, are discussed in Section 2.2. Additionally, equation (2.6) allows any number of the lasso coefficient estimates $\hat{\beta}_j$ $(j = 1, \ldots, p)$ to be shrunk equal to 0, provided that the value of parameter t is small enough.

Tibshirani (1996) proposed the lasso as an improvement to subset selection and ridge regression and provided evidence by comparing the three methods on several simulated data sets. Like subset selection, the lasso was shown to produce parsimonious models which were easy to interpret. The lasso shares with ridge regression (and all bridge regression methods) stability in the model selected when small changes are made to the observations, since it is a continuous coefficient shrinkage process. This stability was demonstrated by Tibshirani (1996) through a simulation study, where multiple data sets were generated from a known model. This study also showed that subset selection resulted in very different models being selected for the different simulated data sets. It was found that the lasso performed best on data sets of a small to moderate number of moderate-sized effects.

2.1.3 Other Related Literature

Frank and Friedman (1993) discussed several variable selection methods, including ridge regression, in the context of the analysis of data sets from chemistry where there is likely to be more predictors than observations, as well as a high degree of collinearity between the predictors. They used a simulation study with data generated from a chemistry model to compare ridge regression to four alternative methods. The methods were two commonly used chemometric regression methods (partial least squares and principal components regression), ordinary least squares, and variable subset selection using all subsets regression.

A set of 36 Monte Carlo experiments were run as a full factorial design with four factors. These factors were: the number of predictors, $p = 5, 40, 100$; the degree of collinearity between the columns of $\boldsymbol{X}$; the true values of the coefficients; the size of the error variance. For each of the 36 experiments, a set of 50 observations was simulated and each of the 5 methods used to select a model. For each fitted model, MSE (see (2.1)) was calculated using 100 additional simulated observations. Each experiment was performed 100 times and the MSE values for the 100 fitted models for each method were averaged. It was found that ridge regression had lower MSE averaged over all the 36 experiments than each of the other four methods. The authors also found that ridge regression performed best when average MSE was examined for different subsets of the experiments, e.g. for the 12 experiments in which there were $p = 5$ predictors.

Several authors have adapted the lasso method to suit particular goals. Lu and Zhang (2007) developed a new penalty for use with the lasso, called the adaptive-lasso shrinkage penalty, in an effort to fit a proportional odds model by maximising the likelihood subject to a shrinkage-type penalty. The new penalty was proposed because, from equation (2.5), the lasso applies the same penalty to all the predictor coefficients, and therefore the coefficient estimates for the predictors most significant in predicting the response may suffer from substantial bias. This method places weights on each of the predictors in (2.5) with values depending on their importance. For $j = 1, \ldots, p$, the jth weight was chosen to be equal to the inverse of the absolute value of the maximum likelihood estimate of β_j. This results in smaller weights being placed on the more important predictors so that the coefficients of less important predictors are more likely to be shrunk to zero.

Hsu, Hung and Chang (2008) used the lasso to develop a subset selection strategy for use with vector autoregressive processes. Meier, van der Geer and Buhlmann (2008) presented the group lasso which allowed the selection of groups of variables in linear regression models, and was developed to be applicable to logistic regression models.

Fu (1998) investigated the entire family of bridge regression methods and demonstrated how both λ and γ can be selected by generalised cross-validation. He also compared the performance of bridge regression to that of ordinary least squares, and the special cases of the lasso and ridge regression through a simulation study, where data were generated using different values of $\gamma = 1, 1.5, 2, 3, 4$ and fixed $\lambda = 1$. The study showed that bridge regression and the lasso perform similarly in terms of MSE, and both perform well when the value of γ used to generate the data is small ($\gamma = 1, 1.5$). By contrast, ridge regression performed well for all the values of γ considered, and had better performance than bridge regression and the lasso for the larger values of γ ($\gamma = 2, 3, 4$).

2.2 Estimation

2.2.1 Centring the Data

For the remainder of this chapter and the analysis throughout the thesis, both $\boldsymbol{Y}$ and $\boldsymbol{X}$ are centred, that is, $\boldsymbol{Y}^{\mathrm{T}}\mathbf{1} = 0$ and $\boldsymbol{X}^{\mathrm{T}}\mathbf{1} = \mathbf{0}_p$. This allows an intercept to be included implicitly in every model considered, and adjusts for any lack of balance in the design. The estimate of β_0 is then $\hat{\beta}_0 = \sum_i Y_i/N$, and is not shrunk towards zero.

2.2.2 Ridge Regression

When $\gamma = 2$, (2.2) can be expressed in matrix form as

$$\mathrm{RSS}(\lambda) = (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda\boldsymbol{\beta}^{\mathrm{T}}\boldsymbol{\beta}. \tag{2.7}$$

Differentiating RSS(λ) with respect to $\boldsymbol{\beta}$ gives

$$\frac{\partial \text{RSS}(\lambda)}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta}. \tag{2.8}$$

Setting (2.8) equal to zero leads to the biased estimator for the ridge regression

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}, \tag{2.9}$$

where $\boldsymbol{I}$ is the $p \times p$ identity matrix and λ can be chosen by a model selection criterion. For ridge regression, $\hat{\beta}_j \to 0$ as $\lambda \to \infty$, see for example Draper and Smith (1998, Chapter 17), and therefore it is not possible for any of the coefficient estimates to be exactly equal to zero.

2.2.3 The Lasso

There is no closed form solution for estimating $\boldsymbol{\beta}$ for the lasso (Section 2.1.2), and several algorithms have been developed for solving the equations (2.2) or (2.3) when $\gamma = 1$. These are now briefly described and their strengths and weaknesses are compared.

Tibshirani's Algorithm

Tibshirani (1996) expressed the lasso problem (as stated in equation (2.6)) as a least squares problem with 2^p inequality constraints corresponding to each of the possible different combinations of the signs of the values of the β_j. This approach was motivated by the work of Lawson and Hansen (1974) who solved the linear least squares problem subject to a general linear inequality constraint. Tibshirani (1996) adapted the approach by introducing each of the 2^p constraints sequentially so that the lasso problem is computationally feasible to solve using this method when p is large. The resulting algorithm works as follows.

Step 1: For a particular value of $t \geq 0$, the algorithm begins by finding $\hat{\boldsymbol{\beta}}^k$, with $k = 1$ denoting Step 1, as the solution to the unconstrained least squares minimisation

problem of finding $\boldsymbol{\beta}$ to minimise

$$g(\boldsymbol{\beta}) = \sum_{i=1}^{N} \left(Y_i - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 . \tag{2.10}$$

Step 2: $g(\boldsymbol{\beta})$ is minimised subject to $\boldsymbol{\varphi}_{i_k}^{\mathrm{T}} \boldsymbol{\beta} \leq t$, where $\boldsymbol{\varphi}_{i_k} = \mathrm{sgn}(\hat{\boldsymbol{\beta}}^k)$, the signum function, and

$$\mathrm{sgn}(\hat{\beta}_j^k) = \begin{cases} -1 & \text{if} \quad \hat{\beta}_j^k < 0 \\ 0 & \text{if} \quad \hat{\beta}_j^k = 0 \\ +1 & \text{if} \quad \hat{\beta}_j^k > 0. \end{cases}$$

Step 3: If $\sum_j |\hat{\beta}_j| \leq t$, then the computation is complete; otherwise $\boldsymbol{\varphi}_{i_k}$ becomes the kth row of the matrix $\boldsymbol{G}$.
Step 4: The coefficient estimates, $\hat{\boldsymbol{\beta}}$, that minimise $g(\boldsymbol{\beta})$ subject to $\boldsymbol{G}\boldsymbol{\beta} \leq t\mathbf{1}$ are found, where $\mathbf{1}$ is a unit vector which has length equal to the number of rows in $\boldsymbol{G}$.
Step 5: If $\sum_j |\hat{\beta}_j| > t$ then $\boldsymbol{\varphi}_{i_k} = \mathrm{sgn}(\hat{\boldsymbol{\beta}}^k)$ is entered as the next row of $\boldsymbol{G}$ and the algorithm continues from Step 3; otherwise the algorithm is ended. In this way coefficient estimates continue to be found until $\sum_j |\hat{\beta}_j| \leq t$.

Shooting Algorithm

Fu (1998) proposed the shooting algorithm to compute the lasso estimator of $\boldsymbol{\beta}$ for a given value of $0 \leq \lambda < +\infty$. His method was derived from the Newton-Raphson algorithm using the result that the limit of the bridge estimator (2.2) is the lasso estimator as γ tends to 1 from above.

At each step $k = 2, 3, \ldots$ of the algorithm, there is a simple closed form for the current coefficient estimates $\hat{\boldsymbol{\beta}}^k = (\hat{\beta}_1^k, \hat{\beta}_2^k, \ldots, \hat{\beta}_p^k)^{\mathrm{T}}$ of $\boldsymbol{\beta}$. Each entry $\hat{\beta}_j^k$ is updated using the entries $\hat{\beta}_l^{k-1}$ $(l = 1, \ldots, p; l \neq j)$ of $\hat{\boldsymbol{\beta}}^{k-1}$ which are the coefficient estimates other than $\hat{\beta}_j^{k-1}$ that were estimated at step $k-1$. The algorithm works as follows.
Step 1: For a particular λ, the algorithm finds $\hat{\boldsymbol{\beta}}^k$ (with $k = 1$), as the solution to the unconstrained least squares minimisation problem (see (2.10)).
Step 2: To update $\hat{\boldsymbol{\beta}}^k$ to $\hat{\boldsymbol{\beta}}^{k+1}$, we update each coefficient $\hat{\beta}_j^k$ to $\hat{\beta}_j^{k+1}$ as follows. Calculate, for $j = 1, \ldots, p$,

$$S_j(\hat{\beta}_j^k, \hat{\boldsymbol{\beta}}_{-j}^k, \boldsymbol{X}, \boldsymbol{Y}) = 2\boldsymbol{c}_j^{\mathrm{T}} \boldsymbol{c}_j \hat{\beta}_j^k + \sum_{l \neq j} 2\boldsymbol{c}_j^{\mathrm{T}} \boldsymbol{c}_l \hat{\beta}_l^k - 2\boldsymbol{c}_j^{\mathrm{T}} \boldsymbol{Y}.$$

Set $\hat{\beta}_j^k = 0$, so that

$$S_0(0, \hat{\boldsymbol{\beta}}_{-j}^k, \boldsymbol{X}, \boldsymbol{Y}) = \sum_{l \neq j} 2\boldsymbol{c}_j^{\mathrm{T}} \boldsymbol{c}_l \hat{\beta}_l^k - 2\boldsymbol{c}_j^{\mathrm{T}} \boldsymbol{Y},$$

where $\boldsymbol{\beta}_{-j}^k$ is a $(p-1)$-vector with entries $\hat{\beta}_l^k$ $(1 \leq l \neq j \leq p)$ and $\boldsymbol{c}_j$ is an N-vector holding the jth column of $\boldsymbol{X}$. Here, S_j is the partial derivative of the residual sum of squares with respect to β_j.
Step 3: Calculate the jth entry in $\hat{\boldsymbol{\beta}}^{k+1}$ as

$$\hat{\beta}_j^{k+1} = \begin{cases} \frac{\lambda - S_0}{2\boldsymbol{c}_j^{\mathrm{T}} \boldsymbol{c}_j} & \text{if } S_0 > \lambda \\ \frac{-\lambda - S_0}{2\boldsymbol{c}_j^{\mathrm{T}} \boldsymbol{c}_j} & \text{if } S_0 < -\lambda \\ 0 & \text{if } |S_0| \leq \lambda. \end{cases} \tag{2.11}$$

Enter the values into the vector of updated estimates $\hat{\boldsymbol{\beta}}^{k+1}$.
Step 4: If $\hat{\boldsymbol{\beta}}^{k+1} \neq \hat{\boldsymbol{\beta}}^k$ the algorithm continues from Step 2; otherwise the algorithm is ended.

We now give a brief explanation of this algorithm. The partial derivative of the lasso problem (2.5), when written in matrix form, with respect to β_j is

$$\begin{aligned} \frac{\partial \mathrm{RSS}(\lambda)}{\partial \beta_j} &= S_j + \lambda \mathrm{sgn}(\beta_j) \\ &= \begin{cases} S_j + \lambda & \text{if} \quad \beta_j > 0 \\ S_j - \lambda & \text{if} \quad \beta_j < 0 \end{cases} \\ &= \begin{cases} 2\boldsymbol{c}_j^{\mathrm{T}} \boldsymbol{c}_j \beta_j + S_0 + \lambda & \text{if} \quad \beta_j > 0 \\ 2\boldsymbol{c}_j^{\mathrm{T}} \boldsymbol{c}_j \beta_j + S_0 - \lambda & \text{if} \quad \beta_j < 0. \end{cases} \end{aligned} \tag{2.12}$$

Setting (2.12) equal to zero leads to the coefficient estimates

$$\hat{\beta}_j = \begin{cases} \frac{-\lambda - S_0}{2\boldsymbol{c}_j^{\mathrm{T}} \boldsymbol{c}_j} & \text{if} \quad \beta_j > 0 \\ \frac{\lambda - S_0}{2\boldsymbol{c}_j^{\mathrm{T}} \boldsymbol{c}_j} & \text{if} \quad \beta_j < 0. \end{cases}$$

In reality, β_j is unknown and cannot be used to determine the value that $\hat{\beta}_j$ should take. To determine the value of $\hat{\beta}_j$, S_0 is compared to λ. This works as

follows. When $\beta_j < 0$

$$S_0 = \lambda - 2\boldsymbol{c}_j^{\mathrm{T}}\boldsymbol{c}_j\beta_j.$$

In this equation $\boldsymbol{c}_j^{\mathrm{T}}\boldsymbol{c}_j > 0$. Therefore since $\beta_j < 0$, the term $2\boldsymbol{c}_j^{\mathrm{T}}\boldsymbol{c}_j\beta_j$ is negative and $S_0 > \lambda$. Alternatively, when $\beta_j > 0$

$$-S_0 = \lambda + 2\boldsymbol{c}_j^{\mathrm{T}}\boldsymbol{c}_j\beta_j.$$

Again, $\boldsymbol{c}_j^{\mathrm{T}}\boldsymbol{c}_j > 0$, therefore since $\beta_j > 0$, the term $2\boldsymbol{c}_j^{\mathrm{T}}\boldsymbol{c}_j\beta_j$ is positive and $-S_0 > \lambda$, or equivalently $S_0 < -\lambda$. It is this reasoning that leads to defining the coefficient estimates as in (2.11).

When $|S_0| \leq \lambda$, the derivative is within λ of being equal to zero; as the second derivative of RSS(λ) is positive, a solution for (2.5) is found. In this situation $\hat{\beta}_j$ is set equal to zero.

LARS Algorithm with Lasso Modification

Efron, Hastie, Johnstone and Tibshirani (2004) presented a simple modification to their least angle regression (LARS) algorithm to solve the lasso problem. We begin by describing the LARS algorithm which performs variable selection in a similar way to forward selection.

At the first step, with all coefficients equal to zero, the LARS algorithm finds the predictor most correlated with the response. The predictor enters the model and its coefficient is increased in the direction of the sign of its correlation with $\boldsymbol{Y}$. The residuals, $r_i = Y_i - \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}$ $(i = 1, \ldots, N)$, are calculated as the coefficient is increased. This continues until a new predictor is shown to exhibit at least as much correlation with the residuals as the predictor already included in the model; this predictor then enters the model. The coefficients of the two predictors are then increased in their joint least squares direction until a third predictor exhibits at least as much correlation with the residuals as the first two; the third predictor then enters the model. The process continues until all of the predictors have entered the model.

Following Yuan, Joseph and Lin (2007), the LARS algorithm may be described as follows.

Step 1: The initial estimates held in $\hat{\boldsymbol{\beta}}^k$ (where $k = 1$) have all the coefficients set to 0, i.e. corresponding to a null model. Then the current vector of residuals is equal

to the response, that is $\boldsymbol{r}^k = \boldsymbol{Y}$.

Step 2: The predictor $\boldsymbol{x}_j$ that has the highest correlation with $\boldsymbol{r}^k$ is identified and its index is included in the set $\mathcal{B}_k = \{j\}$, where $\mathcal{B}_k$ keeps track of which predictors are included in the model at stage k ($k = 1, 2, \dots$).

Step 3: The current direction, ψ, of the variable most correlated with the response is computed as

$$\psi_{\mathcal{B}_k} = (\boldsymbol{X}_{\mathcal{B}_k}^{\mathrm{T}} \boldsymbol{X}_{\mathcal{B}_k})^{-1} \boldsymbol{X}_{\mathcal{B}_k}^{\mathrm{T}} \boldsymbol{r}^{k-1}.$$

This value is the estimate of the coefficients of the predictors in the set $\mathcal{B}_k$ when the response is the current residual vector; $\boldsymbol{X}_{\mathcal{B}_k}$ contains the columns in $\boldsymbol{X}$ corresponding to $\mathcal{B}_k$.

Step 4: If $\mathcal{B}_k \neq \{1, \dots, p\}$, i.e. $\mathcal{B}_k$ does not already include all the predictors, then $a = \min_{i \notin \mathcal{B}_k} a_i \equiv a_{i^*}$, where, for every $i \notin \mathcal{B}_k$, a_i is calculated by solving

$$\boldsymbol{x}_i^{\mathrm{T}}(\boldsymbol{r}^{k-1} - a_i \boldsymbol{X}\psi) = \pm \boldsymbol{X}_{\mathcal{B}_k}^{\mathrm{T}}(\boldsymbol{r}^{k-1} - a_i \boldsymbol{X}\psi),$$

for a_i, where $a_i \in [0, 1]$. Additionally, $\mathcal{B}_{k+1} = \mathcal{B}_k \cup \{i^*\}$.

Step 5: After updating $\boldsymbol{\beta}^k = \boldsymbol{\beta}^{k-1} + a\psi$, $k = k+1$ and $\boldsymbol{r}^k = \boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^k$, the algorithm continues from Step 2, and is repeated until $a = 1$.

The **lasso modification to the LARS algorithm** allows any predictor $\boldsymbol{x}_j$ that is in the current active set $\mathcal{B}_k$ at stage k to be removed from the calculation if the sign of the coefficient of the predictor is not equal to the sign of the current correlation $\widehat{\mathrm{corr}}_j = \boldsymbol{c}_j^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}^k)$.

Naïve Iterative Scheme

The problem of finding $\hat{\boldsymbol{\beta}}$ in (2.5), the 'lasso problem', can be solved, for a particular value of λ, by an iterative process using the formula

$$\left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \frac{\lambda}{2}\mathrm{diag}(|\beta_j|^{-1})\right)\boldsymbol{\beta} = \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}, \tag{2.13}$$

which was defined by Fu (1998) for general bridge regression. Equation (2.13) can be rearranged to give

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \frac{\lambda}{2}\mathrm{diag}(|\beta_j|^{-1}) \right)^{-1} \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}. \tag{2.14}$$

The algorithm works as follows.
Step 1: Arbitrary values are chosen for the initial estimates held in $\hat{\boldsymbol{\beta}}^k$ (where $k = 1$).
Step 2: The entries of $\hat{\boldsymbol{\beta}}^k$ are used as the β_j in the right hand side of (2.14) to obtain $\hat{\boldsymbol{\beta}}^{k+1}$.
Step 3: If $\hat{\boldsymbol{\beta}}^{k+1} \neq \hat{\boldsymbol{\beta}}^k$ the algorithm continues from Step 2; otherwise the algorithm is ended.

Discussion of Lasso Algorithms

One of the advantages of using Tibshirani's algorithm is that it finds the solution to the lasso problem (2.6) in terms of the tuning parameter t. This can be standardised to $s = t/\sum_j |\hat{\beta}_j|$, which is in the range $[0, 1]$; it is easy to interpret and make comparisons of values of s. The algorithm also converges to a solution, for each value of the tuning parameter, in a maximum of 2^p steps. However, as p increases the algorithm becomes inefficient and computationally expensive. Tibshirani (1996) noted that in practical examples the algorithm has been found to converge in the range of $0.5p$ to $0.75p$ iterations. The algorithm is not usable when $N < (p+1)$ since it begins with the solution to the full OLS problem, which cannot be solved when $N < (p+1)$. The algorithm would be of little use in solving descriptive and predictive problems in chemistry for instance, where predictors often outnumber observations. Osborne, Presnell and Turlach (2000) argued that the algorithm is inefficient when p is smaller than N, but still large, when the value of the complexity parameter is such that it causes a medium to large amount of shrinkage, since in this situation most of the lasso estimates will be equal to zero. They added that this inefficiency tends to be increased when, for example, selecting the value of the complexity parameter by cross-validation or estimating the standard errors using bootstrapping.

The shooting algorithm also cannot be used when $N < (p+1)$ since it too begins with the solution to the full OLS problem. Also, it does not have a standardised range of complexity parameters from which to choose the optimum lasso model.

The algorithm benefits from having a simple closed form at each step, leading to a simple update of $\hat{\boldsymbol{\beta}}$ at each iteration which makes the algorithm efficient in terms of computation time. Fu (1998) stated that its convergence rate is in the order of $p\log(p)$, although this result has not been proved theoretically. He added that in situations where the model matrix is orthogonal, the algorithm is typically able to converge in p steps.

Efron et al. (2004) discussed the advantages of the LARS algorithm in terms of computational efficiency. The algorithm is able to calculate every possible lasso estimate for the same order of magnitude of computational cost required to solve the ordinary least squares problem for the full set of p predictors. When $p < N$, this cost is typically in the order of $p^3 + Np^2$ computations. The lasso modification to the LARS algorithm requires of the order of p^2 more operations to take into account the added steps due to the occasional removal of predictors during the lasso iterations. Unlike the previous two algorithms, this procedure is feasible in the case when $N < (p + 1)$, since the algorithm begins with the null model and terminates at the saturated least squares fit when $N - 1$ predictors have entered the model. It is possible for more than $N - 1$ separate predictors to have entered the model, since predictors can also be removed from the model. Efron et al. (2004) noted that the model sequence can be unstable 'towards the saturated end' with respect to small changes in $\boldsymbol{Y}$, which is a disadvantage of the method.

In the naïve iterative scheme, the equation used to obtain $\hat{\boldsymbol{\beta}}$ contains the term $\text{diag}(|\beta_j|^{-1})$, see equation (2.14). When solving the lasso problem for a particular value of λ, it is possible for any of the $\hat{\beta}_j$ to be set to zero. This fact appears to make the naïve iterative scheme infeasible, since for any $\beta_j = 0$ it is not possible to calculate $|\beta_j|^{-1}$. Here, a generalised inverse of $\text{diag}(|\beta_j|)$ would need to be applied.

The algorithm chosen to solve the lasso problem throughout the work in this thesis is the LARS algorithm with the lasso modification. This is because this algorithm can be used in situations where there are more predictors than observations, which is a common occurrence when building models for the prediction of properties of organic compounds. The Melting Point Data Set, which motivates the work in this thesis, has fewer variables than observations, with 21 variables and 60 observations. However if all the 210 pairwise or two-factor interactions between the variables were also to be considered as possible predictors, then there would be many more predic-

tors than observations. A further advantage of using the LARS algorithm with lasso modification is that there is a freely available implementation of the algorithm for the statistical package `R` (R Development Core Team, 2010). The package calculates a lasso solution for each value of a user-specified sequence of complexity parameter values.

2.3 Choice of Tuning Parameters

To fit a bridge regression model with a given value of γ requires the selection of an optimum value for a tuning parameter, which we label α in general, which gives the best performance for prediction. This value identifies an optimal model. Four methods for choosing the tuning parameter are discussed below: cross-validation, generalised cross-validation, AIC and BIC.

The last three methods make use of the concept of 'effective degrees of freedom', $d(\alpha)$, for a fitted model with tuning parameter α. For OLS and $N \geq (p+1)$, the number of degrees of freedom for fitting model (1.1) is simply the number of parameters $(p+1)$, an integer. The idea of effective degrees of freedom is a generalisation of degrees of freedom which is used for models fitted using shrinkage methods. When $\hat{\boldsymbol{\beta}}$ is a linear function of $\boldsymbol{Y}$, then $d(\alpha) = \text{trace}(\boldsymbol{H})$, where $\boldsymbol{H}$ is known as the 'hat matrix' and is defined through $\hat{\boldsymbol{Y}} = \boldsymbol{H}\boldsymbol{Y}$. In the special case of OLS with $N \geq (p+1)$, we have $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\mathrm{T}}$ and $\text{trace}(\boldsymbol{H}) = p$.

For each method, a simulation study is used to investigate how well the fitted model approximates a 'true', or known, model from which a sample of data is generated. The study for cross-validation is presented in Section 2.3.1; the three other methods are outlined in Sections 2.3.2 and 2.3.3, with a simulation using these methods in Section 2.3.5.

The known model used in the simulation was obtained from the data from the Melting Point Data Set, with 21 descriptors standardised to the range $[-1, +1]$. The simulation model is a first order model of the form

$$\boldsymbol{Y} = \boldsymbol{1}_N\beta_0 + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{2.15}$$

where $\beta_0 = -2.74$, and

$$\boldsymbol{\beta}^{\mathrm{T}} = (0, 32.80, 0, 0, 0, 28.59, 49.86, -190.73, 166.75, 0, 0, 0, 2.85, 0, 0, 0, 0, 0, 0, 0, 0), \tag{2.16}$$

with random error vector, $\boldsymbol{\varepsilon}$, having entries generated as 32 independent random draws from a $N(0, 1)$ distribution. Here, β_0 and $\boldsymbol{\beta}$ are indicative of the values obtained by fitting (1.1) to the Melting Point Data Set using the lasso.

A response was simulated from (2.15) for 32 of the compounds, that is, combinations of the 21 descriptor values, selected at random from the full set of compounds. These simulated data are used to illustrate each of the methods. The computations in the simulation study were performed using functions within the `lars` package (Hastie and Efron, 2007) in `R`.

2.3.1 Cross-Validation

The method of K-fold cross-validation has been used to select the value of the tuning parameter for the lasso (see for example, Hastie et al., 2009, page 61). During K-fold cross-validation, the data set is randomly divided into K 'folds', or non-overlapping subsets, of data, each containing an equal number of data points (or as close as possible to an equal number). For each fold k ($k = 1, \ldots, K$) and a particular value of the tuning parameter,

- the kth fold is removed from the data set, and a model is fitted using the data from the $(K - 1)$ remaining folds,
- the fitted model is used to calculate the MSE for the kth fold of data (MSE_k) as in formula (2.1).

The overall MSE, using K-folds, is then calculated as the weighted sum

$$\mathrm{MSE}_K(\alpha) = \frac{1}{N} \sum_{k=1}^{K} |\mathcal{F}_k| \mathrm{MSE}_k, \tag{2.17}$$

where $|\mathcal{F}_k|$ denotes the size of the kth fold of data ($k = 1, \ldots, K$). The value of the tuning parameter is chosen by

(i) identifying the value of α that has the smallest value of (2.17), called α^*,

(ii) selecting the final choice of α so that the corresponding model is the most parsimonious model such that $\text{MSE}_K(\alpha)$ is within one standard error of $\text{MSE}_K(\alpha^*)$.

Here, the standard error of (2.17) is calculated as a weighted standard error over the K-folds.

An optimal model was fitted by the lasso to the data set of 32 compounds using K-fold cross-validation to choose $\alpha = t$, with $K = 10$; this value of K was chosen because at this value the cross-validation estimator of the prediction error ($\text{MSE}_K(\alpha)$) has lower variance than, for instance, leave-one-out cross-validation, see Hastie et al. (2009, page 242-243) for the choice of K. The cross-validation procedure was repeated 100 times, in other words for 100 random allocations of data points to folds. For each repeat, the value of t selected by (i) and (ii) was recorded, together with the value of the standardised complexity parameter

$$s = t / \sum_{j=1}^{p} |\hat{\beta}_j|. \tag{2.18}$$

Since $\sum_{j=1}^{p} |\hat{\beta}_j| \leq t$ (see equation (2.6)), it follows that $0 \leq s \leq 1$ and hence s has a simpler interpretation than t (Hastie et al., 2009, page 69).

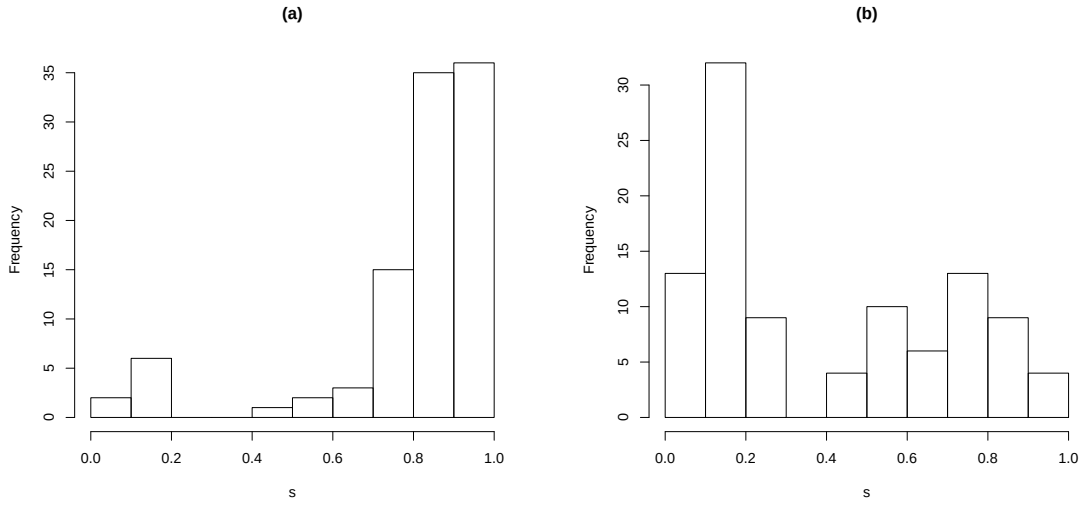
The distribution of s for the 100 repetitions, shown in Figure 2.1(a), has a mean of 0.80, a median of 0.87 and a variance of 0.05. This standard deviation of 0.224 is a cause for concern, because it is large relative to the size of s. Also, Figure 2.1(a) shows that, although in the majority of cases the selected value of s is within the range of 0.8 to 1.0, approximately 10% of the repetitions select a value of s between 0.0 and 0.2 which would give an entirely different model. This can be due only to the different allocations of data points to folds.

The lasso fitted model generated using the mean value of s calculated over the 100 repetitions was able to identify the 6 important variables in the true model and these had coefficients of the same order of magnitude as they had in the true model. In addition, the lasso model included an additional 12 predictors which made it much more complicated to interpret.

In order to assess how the choice of K affects the distribution of s, the analysis was also carried out using five-fold cross-validation, see Figure 2.1(b). The mean of this distribution of s is now 0.40, the median is 0.25 and the variance is 0.09. The fitted lasso model was obtained using the mean value of s calculated over the

repetitions. It again identified all of the important coefficients in the model (see (2.16)), and their estimates had the same order of magnitude as in the true model with two exceptions: the large negative term (-190.73) which was estimated to be zero; the largest positive term (166.75) which was an order of magnitude smaller than its true value. In addition, the fitted model was again more complicated and included 13 further variables. These variables were the same as those selected when choosing s by ten-fold cross-validation, apart from one variable.

Figure 2.1: Distribution of s from 100 repetitions of cross-validation: (a) 10-fold; (b) 5-fold



For comparison, an analysis using $K = N$, that is leave-one-out cross-validation, was performed. The value of s selected was 0.97. This value is obviously much larger than the mean values of the complexity parameter obtained through either ten- or five-fold cross-validation, and results in a much more complicated model with a non-zero coefficient estimate for every term.

The high variability in the chosen standardised complexity parameters in the above study obtained for $K = 5$ and 10 could occur for a number of reasons. As the data set studied is relatively small and is split into ten or five folds, removing a fold of data in cross-validation reduces the data used for model fitting to only about 29 or 26 observations respectively. This reduces the accuracy of the fitted models and

hence may lead to more variability in the chosen s.

In general, if there are outliers or influential points present in the data set, allocating these points to the $(K-1)$ folds used for model fitting during cross-validation would potentially cause a poorly fitting model. As a consequence, the data points in the kth fold used for prediction would have large residuals and the resulting value of MSE_k would be high. Since the data are divided into folds randomly, the outliers or influential points will not necessarily be located in the same folds every time cross-validation is carried out. This would result in a different value of MSE_k and ultimately a different value of the complexity parameter being chosen each time cross-validation is performed. This would also explain the observed increase in the variability in the value of the complexity parameter chosen as the number of folds, K, decreases, as the effect of any outliers or influential points would be magnified in a smaller subset of data.

For these reasons it is important to investigate alternative means of choosing the complexity parameter.

2.3.2 Generalised Cross-Validation

Fu (1998) used **generalised cross-validation** (GCV), defined by Craven and Wahba (1979), to select the parameter α for the lasso. This method could, more generally, be used to select γ, in addition to α, for other regression methods in the bridge family.

The value of GCV for each model in a set indexed by the value of a tuning parameter α can be defined as

$$\text{GCV}(\alpha) = \frac{\text{RSS}}{N(1-d(\alpha)/N)^2}, \tag{2.19}$$

where $d(\alpha)$ is the effective degrees of freedom, $d(\alpha) < N$, and

$$\text{RSS} = \sum_{i=1}^{N}(Y_i - \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})^2. \tag{2.20}$$

A model is chosen out of a set of competing models that has the minimum value of GCV of all models in the set.

2.3.3 Akaike Information Criterion and the Bayesian Information Criterion

We now define two model selection criteria, both of which penalise the selection of larger models. Akaike (1974) defined **Akaike's Information Criterion** (AIC) for a model with tuning parameter α as

$$\text{AIC}(\alpha) = -2l(\hat{\boldsymbol{\theta}}) + 2d(\alpha), \tag{2.21}$$

where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate of $\boldsymbol{\theta}$ which holds the unknown model parameters, $l(\hat{\boldsymbol{\theta}})$ is the log-likelihood and $d(\alpha)$ is the number of estimable parameters, or degrees of freedom of the fitted model. Burnham and Anderson (2002, page 60-64) described the development of the AIC criterion.

The AIC criterion was motivated by the Kullback-Leibler information which measures the discrepancy between the true model and a candidate model, and can be described as the information lost when a candidate model is used to approximate the true model. For a given model, there is a unique value of $\boldsymbol{\theta}$ that minimises the Kullback-Leibler information. In data analysis, $\boldsymbol{\theta}$ is unknown and must be estimated, with associated uncertainty, and the Kullback-Leibler information cannot then be calculated. Instead, an estimate of the expected value of the Kullback-Leibler information is minimised in order to find the best model from a set of candidate models. AIC(α) defined in (2.21) is an approximately unbiased estimator for the expected value of the Kullback-Leibler information, where the degrees of freedom penalty, $2d(\alpha)$, is a bias-correction term.

We first consider the situation where the error variance σ^2 is unknown and must be estimated. For the linear model (1.1), the log-likelihood is defined as

$$l(\hat{\boldsymbol{\theta}}) = -\frac{N}{2}\log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{N}(Y_i - \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})^2. \tag{2.22}$$

where

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{i=1}^{N}(Y_i - \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})^2 = \frac{\text{RSS}}{N}, \tag{2.23}$$

is an estimate of σ^2 if the error variable of each model considered is assumed to be normally distributed with constant variance, and RSS is defined in equation (2.20).

The log-likelihood (2.22) can now be expressed in terms of $\hat{\sigma}^2$ as

$$l(\hat{\boldsymbol{\theta}}) = -\frac{N}{2}\log(2\pi\hat{\sigma}^2) - \frac{N}{2}. \tag{2.24}$$

On substitution of (2.24) into (2.21) we obtain

$$\begin{aligned} \text{AIC}(\alpha) &= N\log(2\pi\hat{\sigma}^2) + N + 2d(\alpha) \\ &= N\log(\hat{\sigma}^2) + 2d(\alpha) + N\log(2\pi). \end{aligned}$$

The term $N\log(2\pi)$ does not depend on the model. Therefore the AIC can be expressed as

$$\text{AIC}(\alpha) = N\log\left(\frac{\text{RSS}}{N}\right) + 2d(\alpha). \tag{2.25}$$

For situations where a model-independent unbiased estimator for the error variance, σ^2, is available, Hastie et al. (2009, page 231) defined the AIC criterion, for a given set of models indexed by the parameter α, by

$$\text{AIC}(\alpha) = \frac{\text{RSS}}{N} + 2\frac{d(\alpha)}{N}\hat{\sigma}_{\varepsilon}^2, \tag{2.26}$$

where $d(\alpha)$ is the effective degrees of freedom. Here $\hat{\sigma}_{\varepsilon}^2$ denotes an unbiased estimator of the error variance which is independent of the value of α. This might be pure error obtained from replicated design points. Alternatively, $\hat{\sigma}_{\varepsilon}^2$ may be obtained by fitting a model with low bias, for example, the full ordinary least squares model. Therefore in situations where $N < (p+1)$, it may be more appropriate to use (2.25) for model selection. When using either definition of AIC, a model is chosen, out of the set of models, that has the lowest value of AIC.

Zou, Hastie and Tibshirani (2007) applied the **Bayesian Information Criterion** (BIC) to select the optimum value of the complexity parameter for the lasso. For each value of α the value of BIC is

$$\text{BIC}(\alpha) = \frac{\text{RSS}}{N} + \frac{\log(N)}{N}d(\alpha)\hat{\sigma}_{\varepsilon}^2, \tag{2.27}$$

and α is chosen to achieve the minimum value of BIC.

The BIC is similar to AIC (see (2.26)), with the factor $\log(N)$ replacing the factor 2 in the penalty term. Hastie et al. (2009, page 233) stated that the BIC

modification to the penalty term results in the BIC penalising complex models more heavily than AIC, leading to simpler models being selected.

2.3.4 Effective Degrees of Freedom for the Lasso

We begin by briefly discussing effective degrees of freedom for ridge regression, and then concentrate on the lasso. In line with equations (2.5) and (2.6), we now have either $\alpha = \lambda$ or $\alpha = t$ (or equivalently s).

For ridge regression, where $\hat{\boldsymbol{\beta}}$ is a linear function of the response, the effective degrees of freedom has a closed form expression, given by

$$d(\lambda) = \text{trace}(\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}^{\mathrm{T}}), \tag{2.28}$$

see Hastie et al. (2009, page 68). This is the trace of the hat matrix for ridge regression.

For the lasso, since the coefficient estimators for (1.1) are non-linear and non-differentiable functions of the response, there is no closed form for the effective degrees of freedom. Therefore several approximations or estimates to the effective degrees of freedom have been developed.

The simplest estimate is given by Zou et al. (2007), who defined an unbiased estimator as

$$\hat{d}(\alpha) = |T|, \tag{2.29}$$

where $T = \left\{ j; \hat{\beta}_j \neq 0 \right\}$, i.e. $\hat{d}(\alpha)$ is equal to the number of non-zero coefficients in $\hat{\boldsymbol{\beta}}(\lambda)$, see (2.5). The authors used this estimate in both AIC and BIC, and also applied (2.28) in fitting a ridge regression model.

Several other methods for estimating the effective degrees of freedom for the lasso are discussed below.

Ridge Approximation to the Lasso

The ridge approximation to estimate the effective degrees of freedom for the lasso model was defined by Fu (1998) who applied it to GCV. For each value of λ, a lasso estimate $\hat{\boldsymbol{\beta}}(\lambda)$ is obtained. From this, an estimate of the effective number of

parameters for this lasso model can be calculated using the ridge approximation

$$\hat{d}_R(\lambda) = \text{trace}(\boldsymbol{X}(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda \boldsymbol{W}^{-})^{-1}\boldsymbol{X}^{\mathrm{T}}) - n_0, \tag{2.30}$$

where $\boldsymbol{W}^{-}$ is a generalised inverse of $\boldsymbol{W} = \text{diag}(2|\hat{\beta}_j(\alpha)|)$, and n_0 is the number of $\hat{\beta}_j(\lambda)$ such that $\hat{\beta}_j(\lambda) = 0$, which compensates for the fact that ridge regression shrinks none of the parameter estimates to zero.

Approximation via Sum of Covariances

Efron et al. (2004) proposed approximating $d(\alpha)$ in the LARS approach (Section 2.2.3), which includes the lasso as a special case, by using an approach developed for the case of a linear fitting method (i.e. linear estimators). This is described by Hastie et al. (2009, page 231). We assume for the purposes of model fitting ((2.26) for AIC, (2.27) for BIC), that the error variance σ^2 is known. Then

$$\sum_{i=1}^{N} \text{Cov}(\hat{Y}_i, Y_i) = \sigma^2 \text{trace}(\boldsymbol{H}),$$

and the effective degrees of freedom is

$$d(\lambda) = \frac{1}{\sigma^2} \sum_{i=1}^{N} \text{Cov}(\hat{Y}_i, Y_i). \tag{2.31}$$

Note that if this estimate for the effective degrees of freedom is used with GCV, then it must be divided through by $\hat{\sigma}^2_{\varepsilon}$, which is taken to be the full OLS estimate of the variance.

Efron et al. (2004) suggested using bootstrapping to estimate (2.31), using the full OLS model to generate a bootstrap sample of observations. To avoid the use of the full ordinary least squares model, which cannot be fitted when $N < (p+1)$, we suggest the following algorithm for generating the bootstrap data and calculating the sum of the covariances for each value of the tuning parameter λ in the set being considered.

1. `Fit a lasso model to the data, for a given` λ`, and obtain estimated coefficients` $\hat{\boldsymbol{\beta}}$ `and residuals` $\hat{\boldsymbol{r}}$

2. Randomly resample the residuals with replacement to obtain $\boldsymbol{r}_i^*$, where $i = 1, \ldots, N$

3. Set the value of the ith resampled observation to $Y_i^* = \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}} + \boldsymbol{r}_i^*$ $(i = 1, \ldots, N)$

4. Fit the lasso, with λ fixed, to the resampled data $(\boldsymbol{x}_1, Y_1^*), \ldots, (\boldsymbol{x}_N, Y_N^*)$

5. Obtain the estimated coefficients $\hat{\boldsymbol{\beta}}^*$ and calculate the fitted values $\hat{\mu}_i^* = \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}}^*$

6. Repeat steps 2-5 B times

7. For each i calculate the covariance

$$\mathrm{Cov}(Y_i^*, \hat{\mu}_i^*) = \frac{1}{B-1}\sum_{h=1}^{B}(Y_{ih}^* - \bar{Y}_i^*)(\hat{\mu}_{ih}^* - \bar{\mu}_i^*),$$

where

$$\bar{Y}_i^* = \frac{1}{B}\sum_{h=1}^{B} Y_{ih}^*,$$

and

$$\bar{\mu}_i^* = \frac{1}{B}\sum_{h=1}^{B} \hat{\mu}_{ih}^*.$$

8. Calculate $\hat{d}(\lambda)\sigma^2 = \sum_{i=1}^{N}\mathrm{Cov}(Y_i^*, \hat{\mu}_i^*)$.

To give some insight into this approximation, we now investigate how the correlation between $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$ is related to the model complexity. Figure 2.2 shows the sum of the correlations, rather than covariances, plotted against the standardised complexity parameter, s, for a lasso model including linear and two-way product terms (so that $N < (p+1)$). Here, bootstrap samples are generated according to steps 1-6 of the algorithm defined above. Instead of calculating the covariances between the bootstrap re-sampled response and the fitted values obtained from the model fitted to the bootstrap data, the correlation between these two quantities is

calculated. For every $i = 1, \ldots, N$ the correlation is calculated by the formula

$$\text{Corr}(Y_i^*, \hat{\mu}_i^*) = \frac{\sum_{h=1}^{B}(Y_{ih}^* - \bar{Y}_i^*)(\hat{\mu}_{ih}^* - \bar{\mu}_i^*)}{(B-1)s_{Y_i^*}s_{\hat{\mu}_i^*}},$$

where

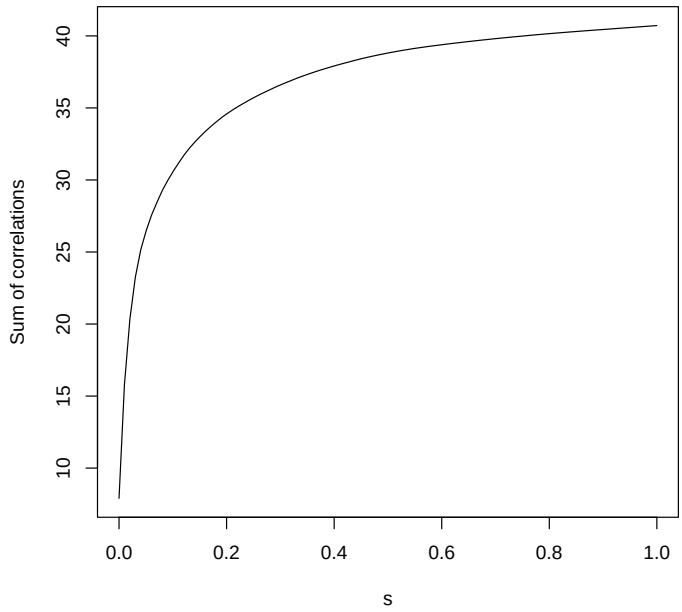
$$s_{Y_i^*} = \sqrt{\frac{1}{B-1}\sum_{h=1}^{B}(Y_{ih}^* - \bar{Y}_i^*)^2},$$

and

$$s_{\hat{\mu}_i^*} = \sqrt{\frac{1}{B-1}\sum_{h=1}^{B}(\hat{\mu}_{ih}^* - \bar{\mu}_i^*)^2}. \tag{2.32}$$

The correlations of each observation are then summed over i for each of the 100 values of $0 \leq s \leq 1$. Figure 2.2 shows that the sum of the correlations increases as the complexity parameter increases, as the fit of the model to the training data improves.

Figure 2.2: Sum of correlations between $\boldsymbol{Y}$ and $\hat{\boldsymbol{Y}}$ vs. s for lasso model with product terms



Modified Degrees of Freedom

More generally, Fu (2005) proposed an estimate for the effective degrees of freedom to take account of the nonlinearity of bridge regression coefficient estimators when $\gamma \neq 2$, and applied it in GCV (equation (2.19)) for model selection. It is defined for the lasso as

$$\hat{d}_M(\alpha) = p \, \frac{\sum_{j=1}^{p} |\hat{\beta}_j(\alpha)|}{\sum_{j=1}^{p} |\hat{\beta}_j^0|}, \tag{2.33}$$

where $\hat{\boldsymbol{\beta}}^0$ is the full ordinary least squares estimator obtained from OLS for model (1.1). This will be called the modified df.

2.3.5 Comparison of Tuning Parameter Selection Methods for the Lasso

We now compare the values selected for the tuning parameter in the lasso by

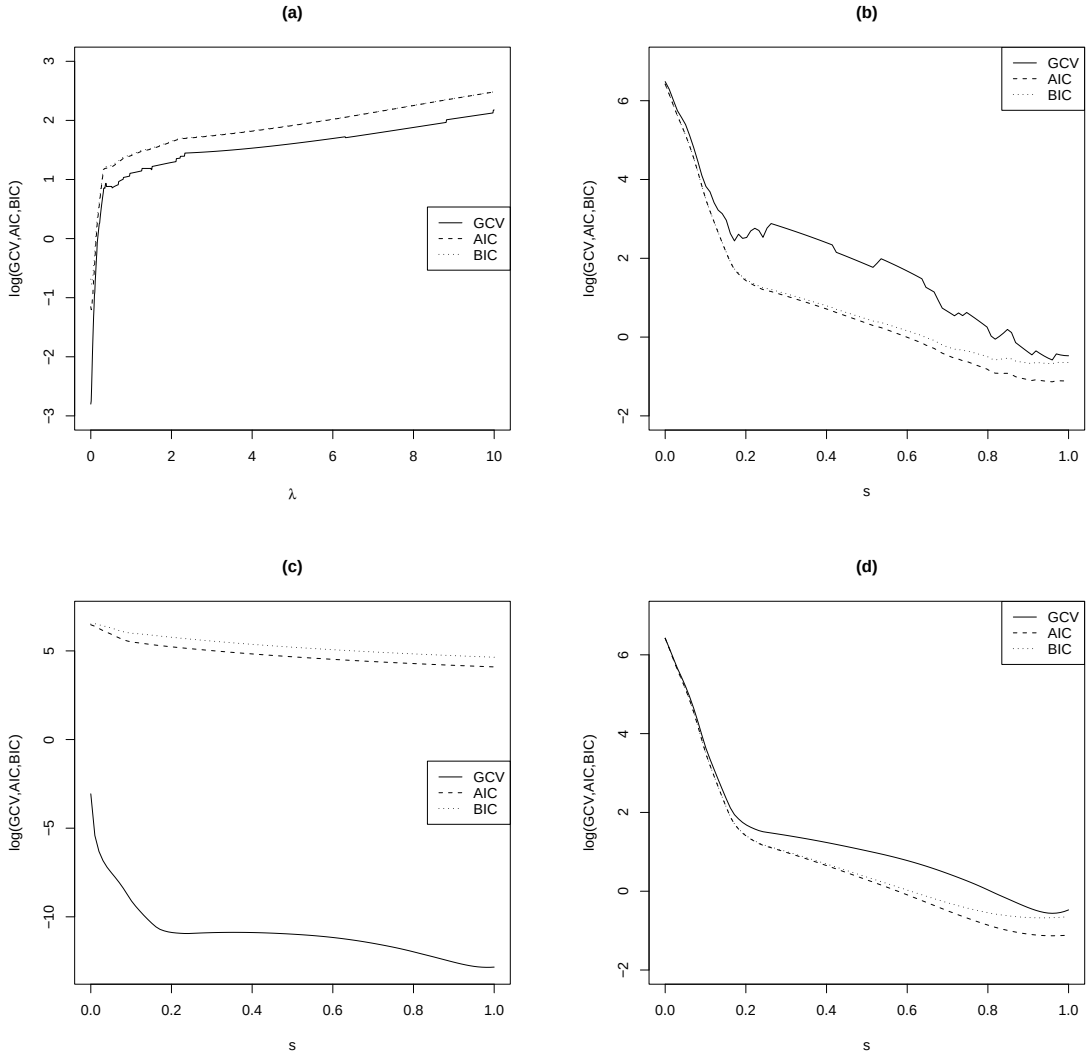
- three model selection criteria: GCV (2.19), AIC (2.26) and BIC (2.27), each implemented with
- four methods of estimating the effective degrees of freedom: simple and ridge approximations, sum of the covariances and the modified df.

The comparison is made by analysing the simulated data set of $N = 32$ observations generated as described in Section 2.3, with $p = 21$.

Figure 2.3 shows the value of the objective function for each model selection criterion plotted against the tuning parameter for each estimate of effective degrees of freedom. Table 2.1 gives the value of the tuning parameter chosen by each method. For the ridge approximation method, the tuning parameter is the complexity parameter λ; s is used in the other methods.

The use of two different tuning parameters, s and λ, in this study can make it difficult to compare the value of the complexity parameter chosen using the ridge approximation and the other methods of estimating the effective degrees of freedom. However, for this data set, models with coefficients that exhibit very little shrinkage were found to be optimum under each method, which corresponds to $\lambda \approx 0$ or $s \approx 1$. The twelve methods are broadly consistent in the choice of optimum model. The

Figure 2.3: Tuning parameter, s or λ, vs. GCV, AIC and BIC calculated using different estimated degrees of freedom: (a) ridge approximation; (b) simple approximation; (c) sum of the covariances; (d) modified df. In plots (a), (b) and (d) the plots for AIC and BIC almost coincide



corresponding models all included the important variables with a coefficient of the same order of magnitude as the true model.

The smallest value of s was obtained using BIC with the modified df estimate of the effective degrees of freedom, resulting in a model that exhibits the most shrinkage and is slightly more parsimonious (having the same number of terms but

Table 2.1: Value of tuning parameter, s or λ, chosen by three model selection criteria using four estimates of effective degrees of freedom

Criterion	$\hat{d}(\alpha)$	Value of tuning parameter
GCV	Ridge approximation	$\lambda = 0.0$
	Simple approximation	$s = 0.96$
	Sum of the covariances	$s = 0.98$
	Modified df	$s = 0.96$
AIC	Ridge approximation	$\lambda = 0.01$
	Simple approximation	$s = 0.96$
	Sum of the covariances	$s = 1.0$
	Modified df	$s = 0.96$
BIC	Ridge approximation	$\lambda = 0.02$
	Simple approximation	$s = 0.96$
	Sum of the covariances	$s = 1.0$
	Modified df	$s = 0.94$

with smaller coefficient estimates) than the models found using either AIC or GCV.

We next compare the results with those from cross-validation in Section 2.3.1 for the same data set. All values of s are slightly higher than the mean value of s found by multiple repetitions of the ten-fold cross-validation procedure, which indicates less coefficient shrinkage and therefore a more complicated model. However, the s values lie within the modal group of s values chosen by ten-fold cross-validation and shown in Figure 2.1(a).

All three values of s are very similar to the value, $s = 0.97$, chosen by N-fold cross-validation. This latter result is expected for GCV in particular, as this method is an approximation to leave-one-out cross-validation (see, for example Hastie et al., 2009, page 244).

When performing model selection using an estimate of the effective degrees of freedom obtained from the ridge approximation, simple approximation or modified df, there is no variability in the value of the complexity parameter selected for a particular data set. This is an advantage over cross-validation, which was shown to exhibit high variability in the value of the complexity parameter selected for these data (Section 2.3.1). My experience of applying, to this example, the sum of the covariances estimate (2.31), using the bootstrapping algorithm, for different values of B, indicated that there was little variation in the model chosen for $B > 200$.

In this study we have compared twelve ways of selecting a model for the lasso using a simulated data set for a known model with $N > (p+1)$. Methods for data sets in which $N < (p+1)$ will be discussed in Chapter 3.

2.4 Estimation of the Variance-Covariance Matrix for Bridge Estimators

In this section, we outline different methods that have been developed for estimating the variance-covariance matrix of the coefficient estimators in bridge regression with emphasis on the lasso. These methods will be compared in the next section.

2.4.1 Inference for Ridge Regression

The ridge estimator of $\boldsymbol{\beta}$ is the only shrinkage estimator in the bridge family to be a linear function of the response. The closed form solution for the variance-covariance matrix of the ridge regression estimator is obtained as

$$\begin{aligned} \text{Var}(\hat{\boldsymbol{\beta}}) &= \text{Var}\left[(\boldsymbol{X}^{\text{T}}\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^{\text{T}}\boldsymbol{Y}\right] \\ &= (\boldsymbol{X}^{\text{T}}\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^{\text{T}}\text{Var}(\boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X}^{\text{T}}\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}\hat{\sigma}^2 \\ &= (\boldsymbol{X}^{\text{T}}\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}\boldsymbol{X}^{\text{T}}\boldsymbol{X}(\boldsymbol{X}^{\text{T}}\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}\hat{\sigma}^2, \end{aligned} \tag{2.34}$$

where $\hat{\sigma}^2$ is an estimator of the error variance.

Fu (1998) considered the variance of the bridge estimator when $\gamma > 1$, which includes ridge regression ($\gamma = 2$). The variance was derived using the delta method (see Davison and Hinkley, 1997, Chapter 2). The variance has the form

$$\text{Var}(\hat{\boldsymbol{\beta}}) = \left(\boldsymbol{X}^{\text{T}}\boldsymbol{X}+D(\hat{\boldsymbol{\beta}})\right)^{-1}\boldsymbol{X}^{\text{T}}\text{Var}(\boldsymbol{Y})\boldsymbol{X}\left(\boldsymbol{X}^{\text{T}}\boldsymbol{X}+D(\hat{\boldsymbol{\beta}})\right)^{-1}, \tag{2.35}$$

where $D(\hat{\boldsymbol{\beta}}) = \text{diag}(\lambda\gamma(\gamma-1)|\hat{\beta}_j|^{\gamma-2}/2)$. The variance, $\text{Var}(\boldsymbol{Y})$, is replaced by a variance estimator $\hat{\sigma}^2$. For the special case $\gamma = 2$, the function $D(\hat{\boldsymbol{\beta}}) = \lambda\boldsymbol{I}$ and (2.35) is equivalent to (2.34). Attention was also drawn to a second special case, that of ordinary least squares regression when $\lambda = 0$. Then the function $D(\hat{\boldsymbol{\beta}})$ is a zero matrix, so that $\text{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^{\text{T}}\boldsymbol{X})^{-1}\boldsymbol{X}^{\text{T}}\text{Var}(\boldsymbol{Y})\boldsymbol{X}(\boldsymbol{X}^{\text{T}}\boldsymbol{X})^{-1}$, the usual definition of the variance of the OLS estimator, see for example Hastie et al. (2009, page 46-47).

2.4.2 Methods for the Lasso

Two methods have been proposed in the literature for obtaining approximate standard errors for the lasso estimator; analytic approximation and bootstrapping.

2.4.3 Analytic Approximation

Since the lasso estimator of $\boldsymbol{\beta}$ is a non-linear and non-differentiable function of the response, it is difficult to obtain a satisfactory estimate of the standard errors of the coefficient estimators. Fu (1998) stated that the form of the variance in (2.35) is not applicable to the lasso, $\gamma = 1$, since the lasso sets some $\hat{\beta}_j = 0$ and the delta method used to derive this variance does not apply in such cases. Tibshirani (1996) proposed a formula for the variance-covariance matrix of estimates of the form

$$\text{Var}(\hat{\boldsymbol{\beta}}) = (\boldsymbol{X}^{\text{T}}\boldsymbol{X} + \lambda \boldsymbol{W}^{-})^{-1}\boldsymbol{X}^{\text{T}}\boldsymbol{X}(\boldsymbol{X}^{\text{T}}\boldsymbol{X} + \lambda \boldsymbol{W}^{-})^{-1}\hat{\sigma}^2, \tag{2.36}$$

where $\boldsymbol{W}^{-}$ is a diagonal matrix with entries

$$\boldsymbol{W}^{-}_{jj} = \begin{cases} \frac{1}{|\hat{\beta}_j|} & \text{if } \hat{\beta}_j \neq 0 \\ \frac{1}{10^{-11}} & \text{if } \hat{\beta}_j = 0, \end{cases} \tag{2.37}$$

and λ is chosen such that $\sum |\beta_j^{\dagger}| = t$ with $\boldsymbol{\beta}^{\dagger} = (\boldsymbol{X}^{\text{T}}\boldsymbol{X} + \lambda \boldsymbol{W}^{-})^{-1}\boldsymbol{X}^{\text{T}}\boldsymbol{Y}$, an approximation based on ridge regression. Hence, by taking the square root of the diagonal entries of the covariance matrix of estimates, the standard errors of the coefficient estimates will be obtained.

As for the form of the variance in (2.35), this approximation is also only appropriate for non-zero $\hat{\beta}_j$. This is because calculating the standard errors via this method will result in a standard error of zero for any $\hat{\beta}_j = 0$, which is inaccurate because there is some uncertainty in the estimation of the coefficient, even if the estimated value is zero. In fact, Fan and Li (2001) define the asymptotic covariance matrix of the estimators, equivalent to the form shown in (2.36), only for coefficients that are not equal to zero.

Osborne et al. (2000) treated the lasso as a convex programming problem, derived its dual and obtained the lasso estimator by considering the primal and dual problems together. In linear programming, the principle of duality states that for any

standard minimising problem, called the primal problem, there is a corresponding maximising problem, known as the dual of the primal problem. The primal and dual problems are related by the coefficients that bound the inequalities and the variables. The variables in the primal problem become the inequalities to satisfy in the dual problem and the inequalities in the primal problem become the variables in the dual problem. The coefficients that bound the inequalities in the primal problem become the function to optimise in the dual problem, and vice versa. Osborne et al. (2000) defined the dual objective function of the lasso primal problem (2.6) as

$$\mathcal{L}_*(\lambda) = \inf_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}, \lambda),$$

where $\inf_{\boldsymbol{\beta}}$ is the infimum, or greatest lower bound, of $\boldsymbol{\beta}$,

$$\mathcal{L}(\boldsymbol{\beta}, \lambda) = f(\boldsymbol{\beta}) - \lambda h(\boldsymbol{\beta}),$$

and

$$f(\boldsymbol{\beta}) = \frac{1}{2}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}), \qquad h(\boldsymbol{\beta}) = t - \sum_{j=1}^{p} |\beta_j| \geq 0.$$

The dual problem for the lasso is

$$\max_{\lambda \geq 0} \mathcal{L}_*(\lambda).$$

Osborne et al. (2000) proposed an improved method for the estimation of the covariance matrix of the lasso estimator to try to overcome the above criticism of Tibshirani's covariance matrix method (that (2.36) does not yield an appropriate estimate of the covariance matrix of $\hat{\boldsymbol{\beta}}$). They illustrated this by applying the formula to the prostate cancer study of Stamey et al. (1989). They therefore proposed three analytic approximations by reformulating λ and $\boldsymbol{W}^{-}$ and using these in (2.36).

Approximation 1. Choose

$$\lambda = ||\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r}||_{\infty}, \tag{2.38}$$

where $\boldsymbol{r}$ are the residuals of the fitted lasso model and $||\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r}||_{\infty}$ is the largest (absolute) entry of the vector $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r}$, which ensures the constraint $||\boldsymbol{\beta}^{\dagger}||_1 = t$, subject to $||\boldsymbol{\beta}^{\dagger}||_1 = \sum |\beta_j^{\dagger}|$. The matrix $\boldsymbol{W}^{-}$ is as defined in (2.37).

Approximation 2. Choose λ as in (2.38) and define the matrix $\boldsymbol{W}$ as

$$\boldsymbol{W}_{jj} = \begin{cases} |\hat{\beta}_j| & \text{if } \hat{\beta}_j \neq 0 \\ 0 & \text{if } \hat{\beta}_j = 0. \end{cases} \tag{2.39}$$

The matrix $\boldsymbol{W}^-$ is then generated by taking the Moore-Penrose generalised inverse of $\boldsymbol{W}$,

$$\boldsymbol{W}^-_{jj} = \begin{cases} \frac{1}{|\hat{\beta}_j|} & \text{if } \hat{\beta}_j \neq 0 \\ 0 & \text{if } \hat{\beta}_j = 0. \end{cases} \tag{2.40}$$

Approximation 3. Choose

$$\lambda = \frac{1}{||\hat{\boldsymbol{\beta}}||_1 ||\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r}||_\infty}. \tag{2.41}$$

A $\boldsymbol{W}^-$ matrix of the form

$$\boldsymbol{W}^- = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r})(\boldsymbol{X}^{\mathrm{T}}\boldsymbol{r})^{\mathrm{T}}, \tag{2.42}$$

is used provided that $N \geq (p+1)$.

Following the evaluation by Osborne et al. (2000) of each of the methods on the prostate cancer data, the authors concluded that Approximation 3 was the most appropriate for estimating the standard errors of the lasso estimates. Unlike Approximation 1, this approximation does not give an estimated standard error of zero for any $\hat{\beta}_j = 0$. Osborne et al. (2000) went on to show that the variance-covariance matrix (2.36) can be obtained by approximating the lasso problem (2.6) with a series of smooth functions. These smooth approximations break down in the context of Approximations 1 and 2 as (2.6) is approached, specifically when any $\hat{\beta}_j = 0$. Therefore it was concluded that Approximation 3 should be used to estimate the standard errors of the lasso estimates.

We next give an outline of bootstrapping which is used for the study in the final section of this chapter.

2.4.4 Bootstrapping

Tibshirani (1996) and Fu (1998) suggested estimating the standard errors of the coefficients through a bootstrapping procedure (see, for example, Davison and Hinkley,

1997). Bootstrapping is a method of assessing the uncertainty of an estimator by sampling from either the original data or a fitted model to form replicate data sets. Refitting the model to these replicate data sets allows the desired measure of model uncertainty to be calculated. Bakker and Heskes (2003) stated that the collection of models fitted to the bootstrap samples reflect the variability of the original data set, and for a data set that exhibits only global similarities this collection of models will be centred around an average model, with the bootstrapping process providing an unbiased version of the average model.

Typically, a bootstrap data set is generated by randomly selecting rows of data, with replacement, from the original data set to create a new data set of the same dimension as the original data set. This is repeated B times; the higher the value of B the more accurate the measure of model uncertainty will be. Efron and Tibshirani (1993, page 50-53) stated that a value of $B = 200$ would usually be sufficient for estimating a standard error. The model fitted to the original data set is then refitted to each of the bootstrap data sets. From these models a measure of uncertainty of the estimator can be obtained, for instance the variance of each of the model coefficients can be calculated over the B bootstrap replicates.

For this section of the thesis, a slightly modified version of this procedure is used, which matches that used by Tibshirani (1996) to obtain the standard errors of the coefficients of a lasso model fitted to the prostate cancer data by Stamey et al. (1989). Instead of resampling entire observations with replacement from the original data set, the residuals of the full least squares model are resampled and combined with the fitted values obtained from the model fitted to the original data set in order to create the bootstrap data sets. This is done to preserve the structure of the data set. Hastie et al. (2009, page 264) refer to this type of bootstrapping as the parametric bootstrap since the method uses a specific parametric model to simulate new responses for the bootstrap data sets. The model refitting is then performed as before. This method is defined by the following algorithm.

`Bootstrapping Algorithm:`

1. `Fit the full least squares model to the data, and obtain estimated coefficients` $\hat{\boldsymbol{\beta}}$ `and residuals` $\hat{\boldsymbol{r}}$

2. `Evaluate` $\mathrm{Var}(\hat{\beta}_j)$ `using bootstrapping`

a. Randomly resample the residuals with replacement, to obtain $\boldsymbol{r}_i^*$, where $i = 1, \ldots, N$

b. Set $Y_i^* = \boldsymbol{x}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}} + \boldsymbol{r}_i^*$

c. Fit the lasso, with λ fixed at the value estimated for the original data, to the resampled data $(\boldsymbol{x}_1, Y_1^*), \ldots, (\boldsymbol{x}_N, Y_N^*)$

d. Obtain the estimated coefficients $\hat{\boldsymbol{\beta}}^*$

e. Repeat steps a-d B times, and calculate the sample variance of the estimated coefficients.

2.5 Comparison of Methods of Lasso Model Selection and Coefficient Standard Error Estimation on Prostate Cancer Data

We now apply the methods described in Sections 2.3 and 2.4.2 to analyse the prostate cancer study from Stamey et al. (1989) by fitting a first order linear model (1.1). In this study the level of prostate-specific antigen (a continuous response variable) was measured on 97 men about to undergo a radical prostatectomy. For each man, the values of 8 explanatory variables were recorded with labels listed in Table 2.2.

First, the values of each of the variables in the data set were standardised to mean = 0 and standard deviation = 1 and the response was mean centred. Table 2.2 gives the values calculated for the standardised complexity parameter, s, and each estimated coefficient obtained using each of the 3 model selection criteria and 4 methods of estimating the effective degrees of freedom (Section 2.3). The value of the complexity parameter λ is reported when the ridge approximation for estimating the effective degrees of freedom was used during model selection.

This table enables us to compare with the method used by Tibshirani (1996). He used generalised cross-validation (2.19), with an estimate of the effective degrees of freedom similar to the ridge approximation (2.30) but without the n_0 compensation term, to select the optimum value of s, which was $s = 0.44$. The corresponding model has five coefficients, for age, lbph, lcp, gleason and pgg45, set to exactly zero. The coefficient estimates for this model are shown in the column headed 'Estimated coefficients' in Table 2.3.

The method in the study with results that most closely replicate the coefficients obtained by Tibshirani (1996) is BIC with the effective degrees of freedom estimated by the sum of the covariances method. The AIC and GCV model selection criteria using the same effective degrees of freedom choose a slightly larger value for s (0.59 for AIC; 0.58 for GCV). As a consequence the resulting lasso models both include two additional non-zero coefficients.

The remaining model selection methods choose values of the complexity parameter that result in models that exhibit less shrinkage, with a greater number of non-zero coefficients.

The inference methods for estimating standard errors for the model coefficients discussed in Section 2.4.2 were compared on the prostate cancer data; the findings were compared to comparisons presented in the papers where the methods were proposed. In order to test the implementation of the methods, each method was applied to the data set. To ensure consistency with the published results, estimates of the coefficients given in the papers were used in $\boldsymbol{W}$ and $\boldsymbol{W}^{-}$ (see (2.37) and (2.39)). The results of this analysis can be found in Table 2.3. Results generated with $\lambda = 2$ are included, and are held in the column headed '$\lambda = 2$'. This is the value of λ chosen using the data by Tibshirani (1996) to generate the estimates of the coefficient standard errors using (2.36) and (2.37). Osborne et al. (2000) argued that, under this value of λ, $||\boldsymbol{\beta}^{\dagger}||_1 \neq t$, a result which led the authors to define λ as in (2.38).

The standard errors produced under each of the three analytic approximations and the bootstrapping method agree with the results of Tibshirani (1996) and Osborne et al. (2000), and this provides a check on the methods having been correctly implemented in my work.

The results in Table 2.3 show that standard errors of zero are obtained only for $\hat{\beta}_j$ with value 0 when using (2.36) with (2.37). Standard errors of zero are not obtained for any $\hat{\beta}_j = 0$ when the standard errors are estimated using either bootstrapping, Approximation 2 using (2.38) and (2.39), or Approximation 3 using (2.41) and (2.42).

Table 2.2: Coefficient estimates for the prostate cancer example using three model selection criteria and four estimates of the effective degrees of freedom

	GCV				AIC				BIC			
Predictor	Ridge approx.	Simple approx.	Sum of cov.	Modified df	Ridge approx.	Simple approx.	Sum of cov.	Modified df	Ridge approx.	Simple approx.	Sum of cov.	Modified df
lcavol	0.6289	0.6223	0.5995	0.6276	0.6239	0.6223	0.6019	0.6276	0.6016	0.6066	0.5701	0.6088
lweight	0.2050	0.1923	0.1589	0.2027	0.1947	0.1923	0.1623	0.2027	0.1619	0.1691	0.1163	0.1723
age	-0.0826	-0.0522	0.0000	-0.0770	-0.0581	-0.0522	0.0000	-0.0770	0.0000	0.0000	0.0000	-0.0017
lbph	0.1238	0.1058	0.0538	0.1205	0.1094	0.1058	0.0592	0.1205	0.0585	0.0700	0.0000	0.0749
svi	0.2553	0.2472	0.2194	0.2537	0.2486	0.2472	0.2236	0.2537	0.2230	0.2318	0.1705	0.2353
lcp	-0.0005	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
gleason	0.0091	0.0000	0.0000	0.0071	0.0000	0.0000	0.0000	0.0071	0.0000	0.0000	0.0000	0.0000
pgg45	0.0722	0.0654	0.0301	0.0711	0.0680	0.0654	0.0334	0.0711	0.0330	0.0401	0.0000	0.0432
Tuning parameter	$\lambda = 2.10$	$s = 0.70$	$s = 0.58$	$s = 0.74$	$\lambda = 3.14$	$s = 0.70$	$s = 0.59$	$s = 0.74$	$\lambda = 7.56$	$s = 0.61$	$s = 0.46$	$s = 0.62$

Table 2.3: Estimated coefficients (Tibshirani, 1996) and standard error estimates calculated in the study of the prostate cancer example

		Estimated standard errors				
				Using (2.36) with		
	Estimated	Using	(2.41)	$\boldsymbol{W}^{-}$=(2.37)		$\boldsymbol{W}$ = (2.39)
Predictor	coefficients	bootstrapping	& (2.42)	$\lambda = 2$	λ=(2.38)	λ=(2.38)
lcavol	0.5588	0.0767	0.0986	0.0789	0.0537	0.0610
lweight	0.0970	0.0680	0.0805	0.0602	0.0246	0.0233
age	0.0000	0.0111	0.0791	0.0000	0.0000	0.0812
lbph	0.0000	0.0391	0.0810	0.0000	0.0000	0.0779
svi	0.1556	0.0831	0.0936	0.0714	0.0312	0.0302
lcp	0.0000	0.0124	0.1146	0.0000	0.0000	0.1044
gleason	0.0000	0.0141	0.1039	0.0000	0.0000	0.1112
pgg45	0.0000	0.0392	0.1141	0.0000	0.0000	0.1233

Larger standard errors for the zero coefficients were obtained when the standard errors were estimated under Approximation 2 or Approximation 3. This suggests that these two approximations are overestimating the variance of the zero coefficients compared to the bootstrapping method.

2.6 Conclusions

In this chapter the bridge regression family of coefficient shrinkage methods was introduced and studied as a continuous method of variable selection. Two special cases of bridge regression, ridge regression and the lasso, which are investigated in the thesis were discussed in detail, with emphasis on the lasso. Details were also given of methods of selecting the tuning parameter, estimating the model coefficients and estimating the standard errors of the coefficient estimators of the resulting models.

Four algorithms, from the literature, for finding the solutions to the lasso problem have been outlined and the advantages and disadvantages of each discussed. The LARS algorithm with lasso modification holds a key advantage over the other algorithms considered as it is the only method that can be used when there are more predictors than observations, $N < (p + 1)$, which is common in many chemistry experiments. Tibshirani's algorithm and the shooting algorithm both begin by obtaining the solution to the full ordinary least squares problem, which cannot be fitted when there are fewer observations than predictors.

The optimum bridge regression model is identified by the value for a tuning

parameter which gives the best prediction performance. Four methods of selecting the tuning parameter for a lasso model were considered: cross-validation, GCV, AIC and BIC. For the three most common forms of cross-validation, 5-fold, 10-fold and leave-one-out, we found that the first two methods produced high variability in the value of the tuning parameter chosen for a particular small data set, and the third method produced large, overfitted models.

To apply the GCV, AIC and BIC criteria the effective degrees of freedom of the fitted model is required. For the lasso, no closed form exists for the effective degrees of freedom. Therefore four methods of estimating this quantity were investigated. In total, twelve approaches were applied to a simulated data set with $N > (p + 1)$ to select the value of the complexity parameter for fitting a lasso model. The twelve methods were broadly consistent in the value of the complexity parameter chosen as optimum. Each method also holds an advantage over cross-validation in that there is limited variability in the value of the complexity parameter chosen for a particular data set.

For data sets in which $N < (p + 1)$ the modified df estimate of the effective degrees of freedom cannot be used as it makes direct use of the full OLS estimator. For such data sets, a model independent estimator of σ^2 will not usually be available and hence AIC (2.26) and BIC (2.27) used in the study of Section 2.3.5 cannot be applied. Instead equation (2.25) and its equivalent for BIC should be applied with the ridge or the simple approximation to the effective degrees of freedom. Methods for $N < (p + 1)$ will be discussed further in Chapter 3.

The lack of a closed form of the lasso parameters makes it difficult to obtain a satisfactory estimate for the standard errors of the coefficient estimators. Four methods of calculating estimates of the standard errors for the lasso, including a bootstrapping procedure and three approximations to the variance-covariance matrix of the coefficient estimators, have been compared on data from a prostate cancer study by Stamey et al. (1989), which has previously been studied by both Tibshirani (1996) and Osborne et al. (2000) for the same purpose. Three of the twelve model selection methods were able to approximately replicate the coefficient estimates obtained by Tibshirani (1996) for this data. The four methods of estimating the standard errors of the coefficient estimators produced standard errors that agreed with those given by Tibshirani (1996) and Osborne et al. (2000).

The approximations to the variance-covariance matrix for the lasso estimators have several limitations. Fan and Li (2001) stated that an asymptotic covariance matrix of the estimators equivalent to that of (2.36) can be used only for coefficients that are not equal to zero. This variance-covariance matrix is used in Approximation 1 and Approximation 2 defined in Section 2.4.3. The third approximation defined by Osborne et al. (2000), Approximation 3, is limited to cases where $N \geq (p+1)$. Additionally, Approximation 1 particularly gives unsatisfactory estimates for the standard errors of the lasso estimators which for any $\hat{\beta}_j = 0$ are equal to zero.

Chapter 3

Prediction of Melting Point via Regression Methods

3.1 Introduction

In this chapter we describe an investigation of the combined and separate influences of properties of organic compounds on melting point. We build a predictive model for melting point by applying regression methods (including ridge regression and the lasso) to the data set of 60 compounds introduced in Section 1.2, and discuss the interpretation of the results.

3.2 Chemistry Background

Melting occurs when the vibrations between the individual molecules increase (as temperature is increased) to the point where they overcome the forces that hold the molecules together in the solid phase. Melting point (°C) can be measured fairly easily and accurately, and this is done routinely at the conclusion of a reaction in order to determine the purity of the product synthesised. It has proved a difficult quantity to predict using other common molecular properties or descriptors which can be measured or calculated directly from the molecular structure. These descriptors are insufficient for describing the diverse and subtle factors that influence melting point. In particular they are lacking in being able to describe crystal packing information and intra- and intermolecular forces present in the compounds' solid forms. Despite

these problems, there are still hundreds of molecular descriptors that are available for use in modelling. Attempts at including them all would lead to a complicated model that is difficult to interpret and which overfits the data, and hence would not generalise well to unseen data.

To overcome these limitations, we consider a data set with the following features

(i) chemistry knowledge is used to select a set of interpretable and meaningful descriptors thought most likely to be related to melting point. The features explored in this chapter are given by the chemical descriptors in Table 3.1,

(ii) the data set includes variables that characterise molecular properties as well as variables that describe the arrangement of molecules in the crystal structure, such as descriptors B, F, G, H, I, M, N, O (see Table 3.1).

The existence of polymorphs, compounds with the same molecular structure but different crystal structures, that have widely different melting points show that it is important to consider descriptors that represent the arrangement of the molecules in the solid form and not to rely on molecular structure descriptors alone. In fact, the compounds used in our study all have similar molecular structure but their melting points have a range of over 160°C, which therefore can not be influenced by molecular structure properties alone.

There are, however, several features on which the melting point of a compound is known to depend. For organic compounds (carbon-based compounds) the dominant feature is regarded, by most researchers, to be intermolecular forces, particularly hydrogen bonding if present (see, for example, Karthikeyan, Glen and Bender, 2005).

Hydrogen bonding depends on the arrangement of the individual molecules in the solid form, known as crystal packing. Higher melting points are often observed when the crystal packing is dense and symmetric. In contrast, lower melting points result from defective crystals, which have weaker intermolecular bonds because they have a disturbed repeating pattern to the arrangement of the individual molecules. Further intermolecular forces that are inherently weaker than hydrogen bonds but nevertheless affect the melting point are van der Waals forces and intermolecular attraction and repulsion. These forces are induced by the polarity of the molecule and descriptors G, M, N (Table 3.1) characterise these features. All of the above features are forces which hold molecules together in the solid form and require energy to be overcome. These are called enthalpic forces.

Table 3.1: Variables (descriptors) for the organic chemistry example

Variable	Description
A	Approximate average width of melting peak
B	Molecular weight (Da)
C	Enthalpy of fusion (J g^{-1})
D	Enthalpy of fusion (KJ mol^{-1})
E	Unit cell density (g cm^{-3})
F	Partition coefficient (ml g^{-1})
G	Polar surface area (Å^2)
H	Molecular volume (Å^3)
I	Molecular volume from Spartan (Å^3)
J	z, number of molecules in the unit cell
K	Unit cell volume (Å^3)
L	Molecular volume/unit cell volume (%)
M	Molecular dipoles from Hartree (debye)
N	Molecular dipoles from Semi (debye)
O	Molecular surface area (Å^2)
P	IR frequency of H-bonding (cm^{-1})
Q	Angle of H-bonding (°)
R	Length of H-bonding (Å)
S	Torsion angle of C^1-S^1-N^1-C^7 bond (°)
T	Number of molecules around one molecule
U	Number of short contacts of one molecule excluding hydrogen bonding

Entropic forces are also known to affect the melting point. Entropy is a measure of disorder within a system. Compounds with densely packed and symmetrical crystal structures typically have a low entropy of fusion and will exhibit a high melting point. A compound may be described as having high rotational entropy, for instance, and will therefore have a low melting point. To address hydrogen bonding, weaker intermolecular forces and some entropic forces we included descriptors $E, J, K, P, Q, R, S, T, U$ in the list of descriptors considered for the experiment (see Table 3.1).

All the remaining descriptors in Table 3.1 were chosen because they were thought likely to be important, e.g. C is the energy required to melt the compound, and were simple to obtain.

3.3 Literature on the Statistical Modelling of Melting Points

In this section, we describe the literature on modelling melting points for organic compounds. After giving brief details of the methods used by other authors, we focus on describing the quality of the generated models.

The problem of modelling the relationship between melting point and various descriptors has, in recent publications, mainly been approached using group contribution methods or quantitative structure-property relationship (QSPR) methods and on data sets consisting of compounds more structurally diverse than those of our motivating data set. Key entries in this literature are summarised in Table 3.2.

Group contribution methods are used within the thermodynamic melting point expression

$$T_m = \frac{\Delta H_m}{\Delta S_m}. \tag{3.1}$$

Here T_m is the melting point and ΔS_m is the entropy of melting, the increase in the degree of disorder when a compound changes from a solid to a liquid. Group contribution methods are used to calculate ΔH_m, the enthalpy of melting, the total energy required to change a compound from a solid to a liquid. The method assumes that the effects of individual functional groups on the enthalpy of melting, and hence the melting point, are additive. The enthalpy of melting is calculated by multiplying the number of times each functional group appears in the molecular structure by the contribution of the functional group towards the enthalpy of melting and summing over all functional groups. The contributions of each functional group are often obtained by multiple linear regression of experimental enthalpy values against the individual group counts. The functional groups chosen ideally cover a wide variety of organic compounds.

The total entropy of melting can be calculated using the semiempirical equation proposed by Dannenfelser and Yalkowsky (1996)

$$\Delta S_m = 50 - 8.31 \ln \omega + 8.31 \ln \eta, \tag{3.2}$$

where ω is the rotational symmetry number of the compound and η is the molecular flexibility number of the compound.

Wang, Ma and Neng (2009) constructed a predictive melting point model using a positional group contribution method. This model predicted melting point using group contributions combined with a distribution function that summarised the position of each functional group in the compound. This method was applied to a data set of 730 organic compounds.

Godavarthy, Robinson and Gasem (2006) argued that group contribution methods in most cases can only be applied to future, unseen compounds that contain the functional groups that are present in the compounds of the data set used to develop the model. Jain and Yalkowsky (2006) noted that they also do not take into account geometric properties that are known to influence the melting point, and are therefore generally unsuitable for estimating melting points.

QSPR methods build regression models to describe a molecular property of interest in terms of chemical structure. This allows models to be built that estimate the property using descriptors that characterise many different aspects of the chemical structure. Such models are developed using variable selection and a variety of regression techniques. This empirical approach is flexible and allows a variety of potential descriptors to be investigated; it is this approach we apply in this thesis.

O'Boyle, Palmer, Nigsch and Mitchell (2008) developed a QSPR model using a winnowing artificial ant colony algorithm for variable selection and model optimisation. This is an extension to the ant colony optimisation algorithm used for variable selection. The 'winnowing' takes place after the optimisation stage and proceeds by discarding the descriptors that are present in less than 20% of the best models. Two types of regression analysis were used; partial least squares and support vector machines. For comparison, a genetic algorithm was also used to separately carry out variable selection, and random forest and k-nearest neighbour models were separately fitted. These methods were applied to a data set of 4119 diverse organic compounds, with melting points with a range of between 14°C and 392.5°C, and 203 2D and 3D descriptors generated from the Molecular Operating Environment software (Chemical Computing Group Inc.).

Hughes, Palmer, Nigsch and Mitchell (2008) investigated building models for the prediction of several properties, including melting point. They built QSPR models using several regression analysis methods, including partial least squares, random forest, k-nearest neighbour and support vector machine modelling procedures. These

methods were applied to a set of 287 drug and drug-like organic compounds. The Molecular Operating Environment was used to generate 168 2D descriptors and 53 3D descriptors. An additional 346 descriptors, including functional group counts, atom-centred fragments and geometrical descriptors were generated by the computer package Dragon (Tetko et al., 2005). Models were developed using each of the three descriptor types separately and in combination. Dimensionality reduction was performed using an ant colony optimisation algorithm before model fitting was carried out.

The results from these publications, as well as several other recent publications, are given in Table 3.2. For simplicity, only the best performing model reported in each publication is included in Table 3.2. These results will be used to help assess the performance of the prediction models developed in this chapter of the thesis. Definitions of the abbreviations used in the table are given below.

SVM Support vector machine

PLS Partial least squares

PC/PCA Principal components/principal components analysis

GA Genetic algorithm

ANN Artificial neural network

UPPER Unified physical property estimating relationships

The models summarised in Table 3.2 have mostly been developed using data sets made up of compounds with more diverse molecular structures than those of our data set. This means that the compounds can range from small unsubstituted hydrocarbons to heavily functionalised, heterocyclic structures with many branches. Models developed using data sets of diverse compounds are likely to exhibit good coverage and may potentially achieve accurate predictions for the melting points of other structurally dissimilar compounds. A model developed using compounds with similar molecular structures, such as those developed in this chapter, should, however, provide more accurate predictions for future compounds from the same series of compounds.

Table 3.2: Summary of published melting point models for organic compounds

Author	Method	Compound types	Training set	Test set	RMSE training set	R^2 training set	RMSE test set	R^2 test set
O'Boyle et al. (2008)	WAAC-SVM	Diverse organic compounds	1831	1373	30.7°C	0.77	45.1°C	0.54
Hughes et al. (2008)	SVM	Drug and drug-like compounds	150	87	27.5°C	0.84	52.8°C	0.46
Bhat et al. (2008)	Extreme learning machine	Non-drug molecules	3173	1000	39.7°C	0.62	45.4°C	0.5
Zhou et al. (2008)	Kernel based PLS	Organic compounds	3000	804	43.4°C	0.54	48.3°C	0.47
Habibi-Yangjeh et al.(2008)	PC-GA-ANN	Drug like compounds	195	64			12.77°C	0.9843
Azencott et al. (2007)	2D kernel based SVM	Non-drug compounds	4173		42.71°C	0.56		
Jain et al. (2007)	UPPER group contribution	Organic compounds	2230		39.7°C	0.830		
Godavarthy et al. (2006)	Nonlinear QSPR back propagation neural network	Organic compounds	770	360			12.6°C	0.95
Modaressi et al. (2006)	Stepwise selection & best subset selection	Drug like compounds	278	45	40.4°C	0.673	42.3°C	0.766
Nigsch et al. (2006)	k-nearest neighbours	Diverse organic compounds	3119	1000			47.3°C	0.48
		Diverse drug compounds	197	80			46.3°C	0.3
Keshavarz (2006)	Group contribution	Nitramines, nitrates & nitroaliphatics	33			0.951		
Karthikeyan et al. (2005)	PCA-ANN	Diverse organic compounds	2087	1043	48.0°C	0.661	49.3°C	0.658
Jain et al. (2004)	Group contribution	Organic compounds	1215			0.977		
Dyekjær et al. (2004)	Forward stepwise	Carbohydrates	11	3		0.951	5.91°C	

The models summarised in Table 3.2 have also, in most cases, been developed using larger data sets, with many more compounds than possible predictors. This is possible because of the use of compounds that are diverse in their molecular structures, more of which are available than compounds restricted to be structurally similar. A larger data set with many more observations than predictors makes the task of identifying the trends between the variables and the melting point, and the subsequent regression analysis, slightly easier. The development of models where there are a large number of predictors relative to the number of observations, and also where there are more predictors than observations, will be investigated in subsequent sections of this chapter.

Most of the models summarised in Table 3.2 have moderate predictive ability when assessed using R^2 and subsequently exhibit poorer performance when assessed over an independent test set of compounds. Among some of the best performing

models are those developed using group contribution methods (see, for example, Jain, Yang and Yalkowsky, 2004). However, as was mentioned in the earlier discussion, if the molecular structures of future compounds do not include the same functional groups as those used to develop the model, then the model is unlikely to perform well when applied to the future compounds. This would apply to the group contribution model reported by Keshavarz (2006) which was developed using only nitramines, nitrates and nitroaliphatics.

Other models that perform well are the principal components-genetic algorithm-artificial neural network model developed by Habibi-Yangjeh, Pourbasheer and Danandeh-Jenagharad (2008) and the nonlinear QSPR model developed using a back propogation neural network by Godavarthy, Robinson Jr. and Gasem (2006). These models were assessed over an independent test set of compounds, and the high R^2 shows that QSPR methods produce models that can be applied, with high predictive accuracy, to future compounds. The model developed using forward stepwise selection by Dyekjær and Jónsdóttir (2004) also performed well, on a limited test set, even though it was developed using a very small set of compounds, consisting of only 11 carbohydrate compounds. This thesis considers a similar, but more ambitious, data set of structurally related compounds.

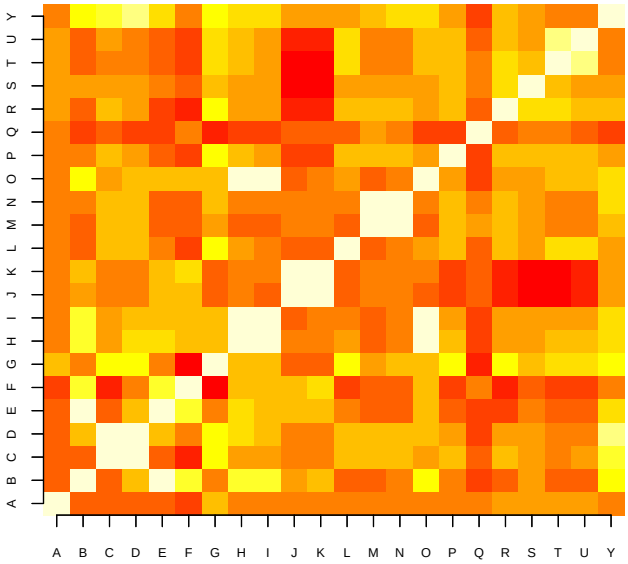
3.4 Exploratory Data Analysis

The organic chemistry data set consisting of observations made on a series of 60 compounds closely related in molecular structure was described in detail in Section 1.2. The 21 descriptors (variables) that were observed for each compound are defined in Table 3.1. Before any analysis was carried out, the explanatory variables were standardised to have mean = 0 and standard deviation = 1 and the response was mean centred, see Section 2.2.1.

After simple plots of each variable were examined, the correlation between each pair of variables, including melting point (Y), was then investigated. Figure 3.1 is a 'hotspot' map for the degree of correlation between each pair of variables: the lighter the shade, the higher the strength of the correlation. It was decided that one variable from each of the six pairs of variables that exhibited greater than 0.9 correlation should be removed from the analysis, since the variables of such a pair

will exert a similar effect on the response. Some of these highly correlated variables are the same variable measured differently, e.g. variables M and N both measure molecular dipoles. By initially considering pairs of such variables, we obtain a check on the quality of the data set.

Figure 3.1: Hotspot map describing the correlation between variables



To decide which variable from each pair should be removed, the correlations between each variable and the response was calculated (i.e. a simple linear regression of melting point on a single variable). The variable to be removed from a pair was the one that was least correlated with the response. This analysis resulted in the removal of enthalpy of fusion in J g^{-1} (C), unit cell density (E), molecular volume from Spartan (I), number of molecules in the unit cell (J), molecular dipoles from Hartree (M) and molecular surface area (O). Therefore, in this chapter of the thesis, regression models for the prediction of melting point were built using a training data set of 60 compounds and 15 explanatory variables.

The investigation of correlations between each descriptor and the response also showed that only two of these 15 explanatory variables had greater than 0.5 corre-

lation with the response; enthalpy of fusion in KJ mol^{-1} (D) and polar surface area (G).

3.5 Variable Selection and Regression Modelling

In this section, variable selection and regression techniques are described which are subsequently applied to the data. Four methods that have not yet been discussed are introduced: subset selection, forward and backward stepwise selection and the Dantzig selector. The application of ridge regression, the lasso and LARS is also described. These methods will be applied in the next section to fit model (1.1) with linear terms (main effects) to the Melting Point Data Set.

3.5.1 Measures of Model Fit

We first define three statistics that each measure the proportion of variation in the data that is explained by the fitted model: C_P, R^2 and adjusted R^2. These statistics are used as criteria to select the best subset size during subset selection. R^2 and adjusted R^2 are also used to measure how well each fitted model describes the data, and to determine which is the best performing model.

We calculate C_P for a model fitted using P predictors, selected from a total of p predictors, as

$$C_P = \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_{Pi})^2}{S^2} - N + 2P,$$

where $\hat{Y}_{Pi}$ is the predicted value of the ith observation obtained from the model fitted using P predictors and S^2 is calculated for the model fitted using all p predictors as

$$S^2 = \frac{\text{RSS}}{N - (p+1)}.$$

The R^2 statistic, the square of the correlation coefficient, is calculated as

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})^2}{\sum_{i=1}^{N}(Y_i - \bar{Y})^2},$$

where $\bar{Y} = \sum_{i=1}^{N} Y_i/N$. The R^2 statistic will always increase as more terms are included in the model, with $R^2 = 1$ when there are $p = N - 1$ predictors in the

model; $R^2 = 1$ denotes a perfect fit, with all the variation in the data explained and fitted values $\hat{\boldsymbol{Y}} = \boldsymbol{Y}$.

Adjusted R^2 penalises the value of R^2 for more complex models. For a model with complexity parameter α, it is defined as

$$\text{Adj.}R^2 = 1 - (1 - R^2)\left(\frac{N-1}{N - d(\alpha) - 1}\right),$$

where $d(\alpha)$ is the degrees of freedom (see Section 2.3.4). The adjusted R^2 does not necessarily increase with the number of terms in the model.

The performance measures C_P and adjusted R^2 are similar statistics that both penalise the prediction error by the number of predictors included in the model; therefore they both tend to select similar models from a set of candidate models. The R^2 statistic will always choose the model with the most predictors, as R^2 always increases with the number of terms in the model.

3.5.2 Subset Selection

In this chapter, subset selection was performed using a branch-and-bound algorithm available within the `leaps` package (Lumley and Miller, 2009) of the statistical program `R`. This algorithm searches exhaustively for the best subset, of size P, of the predictors to include in model (1.1). The algorithm was described by Miller (2002, page 52-54) for OLS. Here, best is defined as having minimum residual sum of squares.

The algorithm, as described by Miller (2002) for model selection using RSS or R^2, begins by dividing all the possible subsets of size P into two branches; those that contain the first predictor to be considered (e.g. predictor A) and those that do not. These branches can be further divided into two sub-branches; those that contain the second predictor to be considered (e.g. predictor B) and those that do not, and so on for other variables. Now assume that, in one of the branches that contains A or B or both, a subset has been identified that has RSS=q_1. If we then consider the branch of subsets that do not contain A or B, a lower-bound on the smallest RSS, q_2, in this branch will be given by the model containing all remaining $p - 2$ predictors. If $q_2 > q_1$, then there can be no subset of P predictors in this branch with RSS$< q_1$, and hence all subsets in this branch can be discarded.

Miller (2002, page 54) noted that the branch-and-bound algorithm is infeasible when there are more predictors than observations since the lower bound on the RSS of each branch is almost always zero. The algorithm becomes computationally infeasible when the number of subsets to be considered is greater than an order of $10^7 \approx 1.192 \times 2^{23}$, or in other words having around 23 possible predictors.

In the work in this chapter, the 10 best subsets of size P, for $P = 0, \ldots, p$, were found. At any stage of the algorithm, we compare the lower-bound on RSS for a given subset to the 10th best subset of size P found up to that stage.

Three different model selection criteria were separately used to select the optimum model from each of the $10 \times (p+1)$ subsets identified via the branch-and-bound algorithm. These criteria are minimum C_P, maximum R^2 and maximum adjusted R^2.

3.5.3 Forward and Backward Stepwise Selection

The methods of forward and backward stepwise selection do not search exhaustively through every possible subset of predictors to identify the best, but instead attempt to find the best by sequentially adding or deleting predictors from a starting model. Forward stepwise selection begins with the null model, i.e. the model where only the intercept is estimated. At the kth step, an F statistic is calculated for each predictor not yet included in the model to test the null hypothesis $H_0 : \beta = 0$. The F statistic is calculated using the formula

$$\begin{aligned} F &= \frac{[\text{Total sum of squares} - \text{RSS}]\,/p_k}{\text{RSS}/(N - p_k - 1)} \\ &= \frac{\text{Regression sum of squares}/p_k}{\text{RSS}/(N - p_k - 1)} \\ &= \frac{\left[\sum_{i=1}^{N}(Y_i - \bar{Y})^2 - \sum_{i=1}^{N}(Y_i - \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})^2\right]/p_k}{\sum_{i=1}^{N}(Y_i - \boldsymbol{x}_i^{\mathrm{T}}\hat{\boldsymbol{\beta}})^2/(N - p_k - 1)} \\ &= \frac{\left[\hat{\boldsymbol{\beta}}^{\mathrm{T}}\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\hat{\boldsymbol{\beta}} - N\bar{\boldsymbol{Y}}\right]/p_k}{(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})/(N - p_k - 1)}, \end{aligned} \tag{3.3}$$

where p_k is the number of predictors included in the model at step k, and $\hat{\boldsymbol{\beta}}$ contains the estimates of the coefficients of the predictors included in the model at step

$k-1$ and the predictor being tested. The predictor with the largest value for the F statistic (3.3) is added to the model if the statistic is greater than the 95th percentile of the $F_{p_k-1,N-p_k}$ distribution. This process is continued until no further predictors have a value for (3.3) greater than this 95th percentile. At each step the coefficients are estimated by ordinary least squares.

Backward stepwise selection begins with the full model, i.e. the model with every predictor included. At the kth step, the F statistic (3.3) is calculated for each submodel obtained by removing a single predictor. The predictor whose removal produced the submodel with the smallest value of (3.3) is removed from the model, provided the value of (3.3) is less than the 95th percentile of the $F_{p_k-1,N-p_k}$ distribution. The process is continued until the removal of any of the predictors remaining in the model produces a value of (3.3) greater than the 95th percentile of the $F_{p_k-1,N-p_k}$ distribution. Unlike forward stepwise selection, backward stepwise selection can only be used when $N \geq (p+1)$.

3.5.4 Ridge Regression and the Lasso

The coefficient shrinkage methods of ridge regression and the lasso were described in detail in Chapter 2. In this chapter of the thesis they have both been applied with the aim of understanding the relationship between the melting point and variables of Table 3.1 for the compounds of the motivating data set. The lasso will also enable variable selection, and the identification of a parsimonious and interpretable model. Coefficient estimates are obtained for ridge regression for a sequence of 1000 values of the complexity parameter, evenly spaced in the range $0 \leq \lambda \leq 10$. Draper and Smith (1998, Chapter 17) consider $0 \leq \lambda \leq 1$; however, in this example the minimum AIC (Section 2.3.3) lies outside this range. For the lasso, coefficient estimates are obtained using the LARS algorithm with the lasso modification for a sequence of 100 values of the standardised complexity parameter, evenly spaced in the range $0 \leq s \leq 1$. For the fitting of first-order main effect models, the value of the complexity parameter that leads to the optimum ridge regression and lasso models is selected using the AIC criterion (2.26). This is because $N > (p+1)$ and therefore an estimator of the error variance, independent of the complexity parameter, is available from fitting model (1.1) using OLS. For each model fitted via the lasso, the effective degrees of freedom, $\hat{d}(\alpha)$, is approximated using the sum

of the covariances between $\hat{\boldsymbol{Y}}$ and $\boldsymbol{Y}$, see (2.31) in Section 2.3.4 and Efron et al. (2004). For ridge regression, $d(\alpha)$ is calculated using (2.28).

3.5.5 LARS

The technical details of the LARS algorithm were given in Section 2.2.3 in the context of a modification to obtain the lasso estimates. The main difference between the lasso and LARS is that in the lasso predictors may leave the model at any stage, whereas at each stage of the LARS algorithm one predictor enters the model, so at step k there are k non-zero coefficients. The LARS algorithm requires p steps to obtain the full set of solutions for a particular value of the complexity parameter.

In this chapter, coefficient estimates are obtained using LARS for a sequence of 100 values of the standardised complexity parameter, evenly spaced in the range $0 \leq s \leq 1$. For the fitting of first order models with linear terms, the value of the complexity parameter that leads to the optimum LARS model is chosen using the AIC criterion (2.26). The effective degrees of freedom is estimated using (2.31), the sum of covariances method, which was originally defined by Efron et al. (2004) for use with the LARS algorithm.

3.5.6 Dantzig Selector

The Dantzig selector is another coefficient shrinkage method, introduced by Candes and Tao (2007), but is not part of the bridge regression family. The parameters, $\boldsymbol{\beta}$, in model (1.1) are estimated by solving the convex programming problem

$$\min_{\hat{\boldsymbol{\beta}} \in \mathbb{R}^p} ||\hat{\boldsymbol{\beta}}||_1, \tag{3.4}$$

subject to

$$||\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})||_\infty \leq \rho,$$

for some $\rho \geq 0$ where $||\hat{\boldsymbol{\beta}}||_1 = \sum_j |\hat{\beta}_j|$ and $||\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})||_\infty = \max_i |(\boldsymbol{X}^{\mathrm{T}}(\boldsymbol{Y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}))_i|$. The convex programming problem (3.4) can be easily reformulated as a linear program by replacing the inequality constraint with a pair of inequality

constraints

$$\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\hat{\boldsymbol{\beta}} \geq \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} - \mathbf{1}_p\rho, \qquad -\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\hat{\boldsymbol{\beta}} \geq -\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} - \mathbf{1}_p\rho. \tag{3.5}$$

In the work in this chapter, coefficient estimates are obtained using the Dantzig selector for a sequence of 100 values of ρ evenly spaced in $0 \leq \rho \leq ||\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}||_\infty$. When $\rho = 0$ we have the full ordinary least squares estimator, and when $\rho = ||\boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y}||_\infty$, all $\hat{\beta}_j = 0$ for $j = 1, \ldots, p$ (see, for example, James, Radchenko and Lv, 2009).

The value of ρ that leads to the optimum Dantzig selector model is chosen using the AIC criterion (2.26). Following Phao, Pan and Xu (2009), we use $|T|$ as the effective degrees of freedom for the Dantzig selector, where $T = \left\{j; \hat{\beta}_j \neq 0\right\}$.

3.6 Models with Linear Terms

The variable selection and regression analysis methods described in Section 3.5 were each applied to the Melting Point Data Set in an effort to build a descriptive model that identifies variables which have an important effect on the melting point of each compound. In this section the only predictors considered for inclusion in the model are linear terms in each variable, often referred to in this thesis as main effects. A summary of the results of each modelling method is given in Table 3.3. The R^2 and adjusted R^2 statistics have been calculated over the training data that was used to fit each of the models.

Table 3.3: Summary of results of models with linear terms

Method	Variables included the model	R^2	Adjusted R^2
Subset selection, min. C_P	B, D, G, K, N, P, Q, R	0.844	0.816
Subset selection, max. R^2	$A, B, D, F, G, H, K, L, N, P, Q, R, S, T, U$	0.867	0.822
Subset selection, max. adj. R^2	$A, B, D, F, G, H, K, N, P, Q, R$	0.860	0.828
Forward stepwise	A, B, D, N	0.788	0.772
Backward stepwise	B, D, G, N, P	0.823	0.807
Ridge regression	$A, B, D, F, G, H, K, L, N, P, Q, R, S, T, U$	0.863	0.827
Lasso	$A, B, D, G, K, N, P, Q, R, T$	0.840	0.807
LARS	$A, B, D, G, K, N, P, Q, R, T$	0.840	0.807
Dantzig	A, B, D, G, K, N, P, R	0.833	0.806

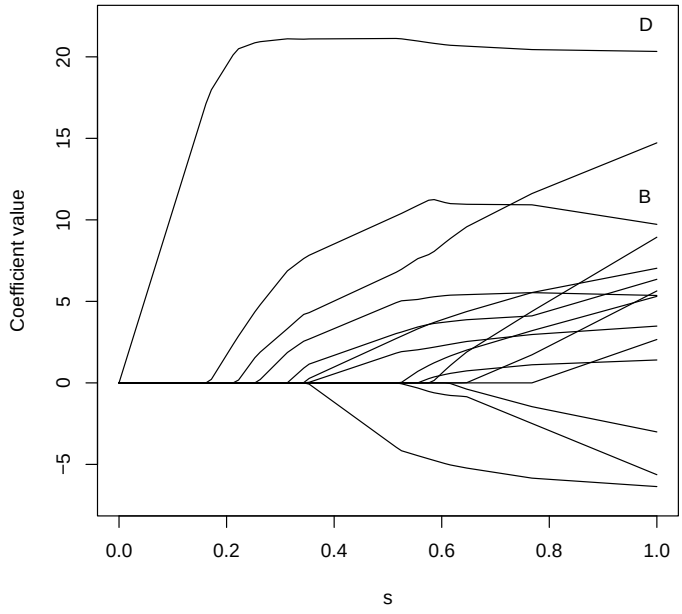
The models summarised in Table 3.3 all exhibit similar performance in terms of prediction. They also have several predictors in common, such as B, D and N. As expected, the model found by subset selection maximising R^2 includes all of the 15 possible explanatory variables and is, in fact, the full ordinary least squares model. Even though it is the most complex model, it is still the third best performing model in terms of adjusted R^2. The ridge regression model also includes all of the 15 possible explanatory variables. This was also to be expected since ridge regression does not shrink any of the coefficients to exactly zero. Some shrinkage has, however, been applied to the coefficients of the ridge regression model. In shrinking the coefficients, the model is less complex than the full ordinary least squares model, as measured by the effective degrees of freedom (2.28), which is equal to 12.4. The performance of the ridge regression model is improved in terms of adjusted R^2 when compared to the subset selection model chosen by maximising the R^2.

The lasso and LARS methods have chosen the same model with the same estimates for each of the coefficients, resulting in the same values for both R^2 and adjusted R^2. As the algorithm used to find the lasso solutions is a modification to the LARS algorithm, it is not unusual for the optimum models to be similar. The variables entered the model in the same order with increasing s for both the lasso and LARS; Efron et al. (2004) obtained a similar result when they applied the lasso and LARS algorithms to data obtained from a study of diabetes patients. They found that the lasso took two more steps than LARS on the way to completing the algorithm, with the additional steps taken due to a variable being removed from the model and then re-entering a few steps later. The variables had joined the model in almost the same order. For our organic chemistry example, no variables were removed from the lasso model once they had entered. As a consequence, both the lasso and LARS models have the same effective degrees of freedom at each value of s and the same optimum model was chosen for both the lasso and LARS using AIC.

Figure 3.2 shows how the coefficient estimates of the lasso model change as s is increased from 0 to 1. Note that only the lines of the two largest coefficients have been labelled with variable names for clarity. Plots such as these are called 'trace plots'. It is clear to see from Figure 3.2 that as s increases more predictors enter the model, and at $s = 1$ all predictors have entered the model. At the value of s chosen for the lasso and LARS models, $s = 0.55$, 10 predictors had entered the

model. Figure 3.2 also shows that the models obtained using forward and backward stepwise selection can be considered 'submodels' of the lasso and LARS models, obtained by the selection of a smaller value of s. For example, B and D are selected for inclusion in the models obtained by forward and backward stepwise selection, and these predictors are the first to enter the lasso model at small values of s, shown in Figure 3.2.

Figure 3.2: Trace plot of the estimated coefficients for a lasso model with linear terms



The model obtained from using the Dantzig selector is a more parsimonious model than those obtained by the lasso and LARS, with eight predictors selected for inclusion in the model. The model has two fewer terms than the lasso and LARS models, but the performance of these models in terms of adjusted R^2 is very similar.

Three of the variables are common to all of the fitted models. These are variable B (molecular weight) variable D (enthalpy of fusion) and variable N (molecular dipole). The inclusion of these particular variables is scientifically sound and we now comment on their chemistry interpretation.

Molecular weight is the combined weight of all the atoms that make up one

molecule of a compound. Since the compounds in this study all have the same central molecular structure, the differences in molecular weight must be due to the functional groups that are substituted at each end of the molecules. A higher molecular weight implies that these terminal functional groups are larger, meaning the molecules in the solid form are arranged further apart from one another. This reduces the strength of the intermolecular forces that hold the molecules together in the solid form, meaning less energy would be required to overcome these forces and the compound would melt at a lower temperature.

Enthalpy of fusion is the amount of thermal energy which must be absorbed for one mol of a substance to change states from a solid to a liquid. The temperature at which this occurs is the melting point, therefore if a higher energy is required then the melting point for that compound will be greater.

Molecular dipoles describe the distribution of charge through a molecule. A large molecular dipole indicates that one end of the molecule is more negatively charged than the other. Molecules with a large dipole will be held together by strong intermolecular forces in the solid form, since the negatively charged end of one molecule will be attracted to the positively charged end of the next. More energy, and therefore a higher melting point, will be required to overcome these forces than for a molecule with a smaller molecular dipole.

The inclusion of any of the other variables in the models is not unexpected as the complete set of variables were initially chosen as they were thought likely to contribute to the observed melting point. One exception is variable A, the approximate average width of the melting peak. As an explanatory variable this feature has not been considered in any analysis found in the literature. The melting peak itself provides the measurement of the melting point. Through discussion with chemists, it has been proposed that the width of the melting peak can give an indication as to the degree of defects in the crystal structure of the compound. A crystal structure with more defects would be expected to have a wider melting peak and a lower melting point. It would be interesting to conduct further investigation into the potential causes of the relationship between the width of the melting peak and the melting point.

The best performing model, in terms of adjusted R^2, was obtained through subset selection using the maximum adjusted R^2 as the criterion for choosing the

optimum subset size. A slightly more parsimonious model was obtained using the lasso with only a slight decrease in R^2 and adjusted R^2. The lasso model included 10 predictors with non-zero coefficient estimates compared to 11 non-zero coefficients in the subset selection model. The lasso model had a root mean squared error of 14.43°C, where root mean squared error (RMSE) is

$$\text{RMSE} = \sqrt{\text{MSE}},$$

and MSE is defined in (2.1); this is around 10% of our mean melting point of $\bar{Y} = 142.05$. This RMSE is smaller than many of the RMSE values for models reported in published literature (see Table 3.2).

The lasso model including linear terms was of the form

$$\begin{aligned} \text{E}(Y) &= 1.99A + 10.72B + 21.02D + 7.41G + 3.33K + 5.09N \\ &- 4.36P + 0.48Q + 3.17R - 0.23T. \end{aligned} \tag{3.6}$$

Recall, from Section 3.4, that this model was developed using a mean-centred response and mean-centred and scaled variables.

Cross-validation was used to obtain out-of-sample estimates of the R^2, adjusted R^2 and root mean squared error for the models listed in Table 3.3. Cross-validation was used because the data set did not contain enough observations to split into adequately sized training and test sets. The results of carrying out 5-fold, 10-fold and N-fold cross-validation (see Section 2.3.1) are shown in Table 3.4. The performance of model (3.6) is consistent with the models obtained using all other model fitting methods investigated, including the other coefficient shrinkage methods. The performance of all models is lower when performance is assessed out-of-sample than when it is assessed within-sample. This is to be expected as during cross-validation the models are fitted using fewer observations, and we are predicting for 'unseen' compounds.

The model obtained through subset selection using maximum R^2 as the criterion for choosing the optimum subset size performs poorly when assessed using cross-validation, this is because it is overfitting to the data. In contrast, the models obtained using forward and backward stepwise selection, which were the most parsimonious of all the models, have good out-of-sample performance. However, these

methods have not selected some scientifically relevant predictors that were selected by the lasso, such as Q (angle of hydrogen bonding) and R (length of hydrogen bonding).

Model (3.6) also compares favourably to the models reported in the recent literature (Table 3.2). It has one of the highest values for R^2 and the lowest value of root mean squared error, calculated over the data used to build the model (training data). It also performs comparatively well out-of-sample, as approximated by cross-validation. This has been achieved with a relatively small set of compounds and a regression method that has not before been applied in the published literature to building predictive melting point models. A true comparison between the models is difficult to make since the types of compounds present in the data sets are so different, with the models found in the literature generally developed for data sets made up of compounds with diverse molecular structures. We would not recommend extrapolation to compounds that do not have the same central molecular structure as those of the motivating data set; model (3.6) may perform poorly in such cases.

Model (3.6) was used to predict the melting point of two compounds from the same series of structurally related compounds as those in the Melting Point Data Set that were not used to develop the model. The first of these compounds has R^1=Br and R^2=I. Its melting point is 153.97°C. Model (3.6) predicted the melting point of this compound to be 161.21°C. The second of these compounds has R^1=I and R^2=Br. Its melting point is 170.35°C. Model (3.6) predicted the melting point of this compound to be 167.48°C. The RMSE for these predictions is 5.51°C. This error is relatively large for the purpose of predicting melting points, where the melting points of structurally different compounds can be within a few degrees of each other.

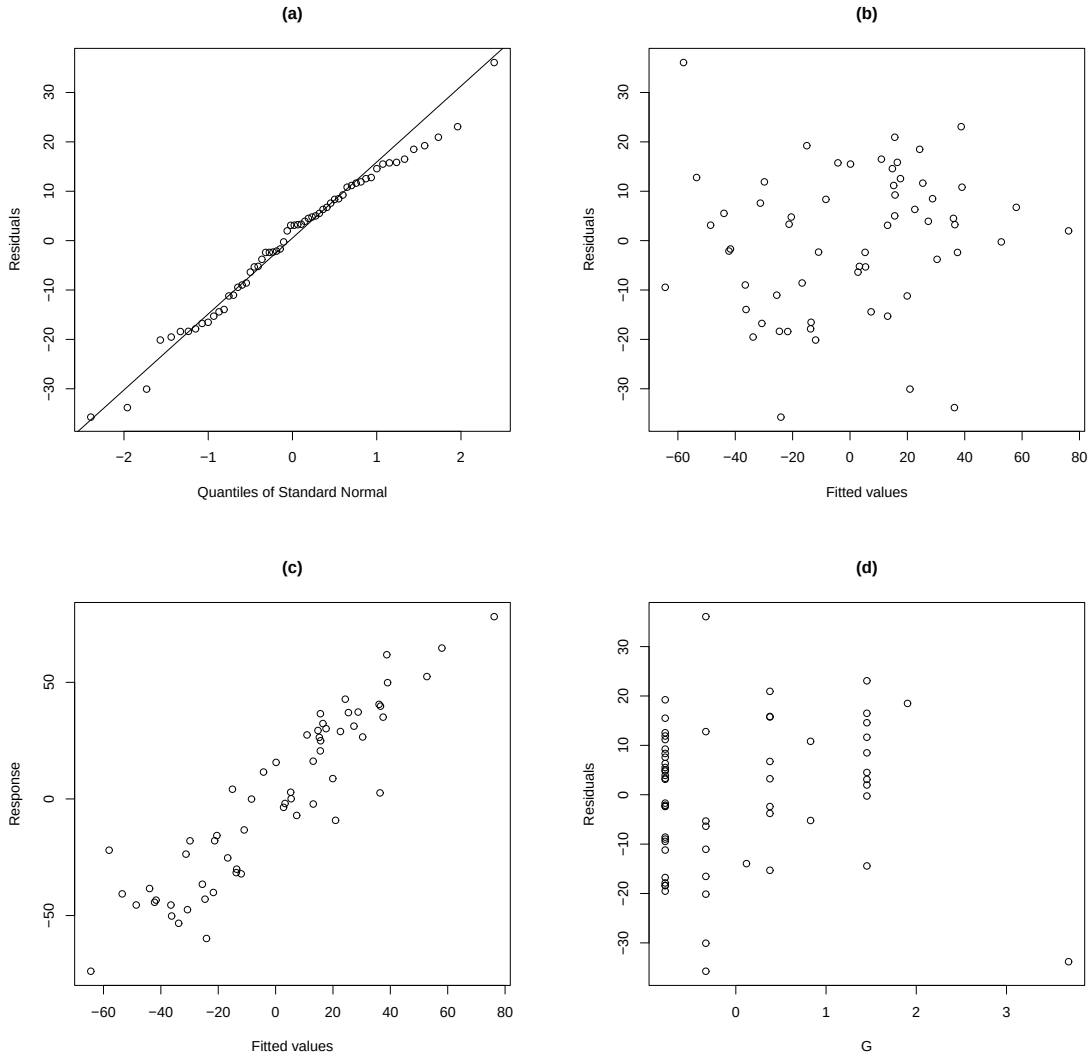
Diagnostic residual plots for model (3.6) are shown in Figure 3.3, where each residual has been calculated as $Y_i - \hat{Y}_i$. These types of plots are commonly used to check the model assumptions and for departures from the fitted model. The normal Quantile-Quantile (Q-Q) plot of the residuals, Figure 3.3(a), shows a line that deviates only slightly from a straight line and passes through the origin. This supports the assumption that the normal distribution is appropriate for the error term in (1.1).

The plot of the residuals against the fitted values obtained from the model, Figure 3.3(b), displays a random scatter of points with no systematic pattern or curvature.

Table 3.4: Model validation statistics for models with linear terms

	5-fold c-v			10-fold c-v			N-fold c-v		
Method	R^2	Adj.R^2	RMSE (°C)	R^2	Adj.R^2	RMSE (°C)	R^2	Adj.R^2	RMSE (°C)
Subset selection, min. C_P	0.73	0.69	18.9	0.76	0.72	17.6	0.76	0.72	17.7
Subset selection, max. R^2	0.64	0.51	22.9	0.75	0.66	18.4	0.73	0.64	18.8
Subset selection, max. adj. R^2	0.75	0.69	18.2	0.76	0.71	17.7	0.76	0.71	17.6
Forward stepwise	0.76	0.74	17.6	0.73	0.71	18.7	0.74	0.72	18.2
Backward stepwise	0.78	0.76	17.0	0.76	0.73	17.8	0.76	0.74	17.6
Ridge regression	0.73	0.63	18.8	0.74	0.65	18.4	0.74	0.66	18.1
Lasso	0.71	0.65	19.4	0.71	0.65	19.3	0.73	0.67	18.7
LARS	0.73	0.70	18.6	0.74	0.67	18.2	0.73	0.66	18.7
Dantzig	0.69	0.65	19.8	0.71	0.67	19.1	0.75	0.71	17.9

Figure 3.3: Diagnostic residual plots for Model (3.6): (a) Normal Q-Q plot of residuals; (b) Fitted values vs. residuals; (c) Fitted values vs. response; (d) Variable G vs. residuals



This is ideal and supports the model assumption that the error is constant across all observations. There are also no obvious outliers in this plot.

The plot of the true response values against the fitted values, Figure 3.3(c), displays a clear positive linear relationship with close to unit gradient between the response and the fitted values. This reinforces the result that was suggested by

the high value of both R^2 and adjusted R^2, that the model performs well when predicting the melting point, across the data to which the model was fitted.

Each of the variables included in the model were plotted against the residuals and in these plots very few outliers were observed. Figure 3.3(d) shows one of these plots, where the variable G, polar surface area, has been plotted against the residuals. There is one outlier, in the 'x' direction, which was identified as the compound with a melting point of 144.64°C and a polar surface area of 137.817Å^2, the highest value for this variable. This compound has the functional group NO_2 substituted at either end of the molecular structure and is the only compound with this combination of functional groups. Therefore, it is also the only compound of the data set that exhibits this value of polar surface area, which is much greater than the next largest value of polar surface area, 101.227Å^2. Compounds with NO_2 substituted at one end of the molecular structure exhibit the next highest polar surface area and so it is not unexpected that with two of these functional groups, this compound has the highest value of polar surface area. A further investigation may provide a chemical reason as to why the polar surface area of this compound is so much greater than for other compounds, and whether or not this compound would affect the predictive accuracy of any fitted models. Such an investigation would, potentially, take a great deal of time. Although this data point was not identified by a preliminary data analysis to be either an influential or leverage point, it may be possible to obtain more accurate predictions if, in the future, the models are built with this observation removed from the data set. Removing this data point would, however, reduce the coverage of the model.

3.7 Models Including Product Terms

Further models are now sought that include predictors constructed from products of two variables ('two-factor interactions') as well as linear terms. Product terms describe the joint action of two variables on the response. In the remainder of the thesis, these terms will be called two-factor interactions.

Adding all two-factor interactions to the set of predictors to be considered during variable selection means there is now a total of 120 potential predictors; there are now 61 more predictors than observations. In this section, we use only shrinkage

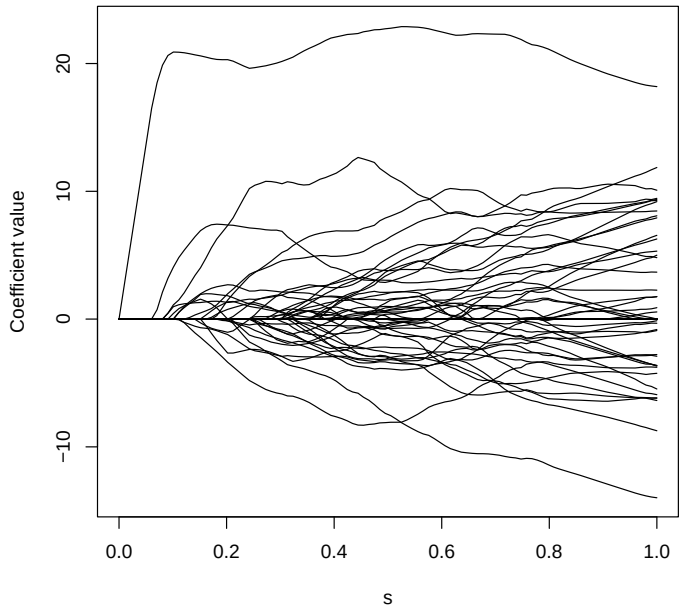
methods (ridge regression, the lasso, LARS and the Dantzig selector) which allow models containing all predictors to be considered.

Yuan et al. (2007) commented on the difficulties in selecting the optimum model out of those generated from the LARS algorithm when $N < (p + 1)$. For the example they presented, based on the analysis of a simulated experiment where the true model was known, they investigated the use of leave-one-out cross-validation to select the optimum model, minimising the overall mean squared error of (2.17). They found that either the null model or the full model was chosen as optimum using this method; these models were not the same as the model used to simulate the data. The authors suggested an alternative method of model selection. This heuristic approach involved plotting the values of each coefficient against s, and then selecting s to be the value at which the observed initial, rapid increase in a small number of the coefficients slowed significantly.

A trace plot for a lasso model including both linear terms and interaction terms for the organic chemistry example is shown in Figure 3.4. Note that the lines have not been labelled with variable names for clarity. It was previously shown, in Figure 3.2, that for a lasso model including linear terms there are 15 variables to consider, so it is easy to see on the plot which of the variables are increasing quickly, although it could be difficult to justify the choice of one value of s over another. For the lasso model including interaction terms there are many more variables. It then becomes difficult to see what is happening to the values of their coefficients as s is increased and the lines become entwined, see Figure 3.4. This makes it even more difficult to justify the choice of the complexity parameter. A value of $s = 0.15$ could be suggested based on the definition (Yuan et al., 2007) of the model selection method. The model resulting from this value of s has 12 non-zero coefficients, including 4 linear or main effect terms and 8 interaction terms. This model has an $R^2 = 0.841$ and an adjusted $R^2 = 0.796$. Based on this latter value, the performance of this model is poorer than model (3.6) which included only linear terms.

When using any of the shrinkage methods to fit a model including both main effects and interactions for the Melting Point Data Set, there are complicating issues caused by having fewer observations than predictors, particularly the choice of the tuning parameter. Firstly, when $N < (p + 1)$, there is no model-independent or reliable, low bias estimator of σ^2. Therefore σ^2 must be viewed as an additional

Figure 3.4: Trace plot of the estimated coefficients for a lasso model with linear and interaction terms



model parameter to be estimated; the form of AIC defined in equation (2.25), or the equivalent formulation for BIC, should be used.

Secondly, any model containing $p = N-1$ predictors will have RSS ≈ 0 which will always be selected using either the AIC or BIC model selection criteria. Therefore modifications to the degrees of freedom penalty must be applied so that complicated models are subject to an even heavier penalty. Finally, the effective degrees of freedom, for the lasso and LARS, can only be estimated using the simple approximation, $\hat{d}(\lambda) = |T|$. The sum of covariances estimate (2.31) cannot be used because a model-independent estimator of σ^2 is not available. The modified df estimate (2.33) cannot be used because this method makes direct use of the full ordinary least squares estimate.

In this section, two modifications to the degrees of freedom penalty of the AIC criterion (2.25) are discussed. The original penalty and the two modifications are then applied within the AIC criterion to select optimum lasso models containing only linear terms. These models, and their performances, are compared to model

(3.6) which was selected using the alternative AIC criterion (2.26) and (2.31) to approximate the effective degrees of freedom. The original penalty and the two modifications are then applied with AIC to select an optimum lasso model that includes linear and interaction terms. The degrees of freedom penalty that selects the best performing lasso model is then used within AIC to select optimum ridge regression, LARS and Dantzig selector models that also include interactions. These models are presented and their performances discussed.

3.7.1 Modifying the Penalty in the AIC Criterion

In this section the two modifications to the degrees of freedom penalty of the AIC criterion defined in (2.25) are discussed. This definition of the AIC is appropriate when σ^2 is estimated as another model parameter, as is necessary when there is not a model independent, low bias, estimate of the error variance σ^2 available. It is common to use the full ordinary least squares model to provide an estimate of the error variance, however this model can not be fitted when there are fewer observations than variables. This is the case when linear and interaction terms are considered with the organic chemistry example.

The problems that are often encountered when selecting the complexity parameter for the lasso when $N < (p+1)$ can be attributed to the fact that the degrees of freedom penalty term in the AIC criterion (2.25) does not increase quickly enough to be able to compensate for the rapid decrease in $\log(\text{RSS}/N)$ as the complexity and number of terms in the model increases. This results in a saturated model, i.e. a model with the same number of predictors as observations, being consistently chosen as optimum with RSS ≈ 0. Therefore, in this section, two modifications to the penalty are considered in order to improve model selection. The degrees of freedom penalties considered are

(i) $2d(\alpha)$,

(ii) $2d(\alpha)^2$,

(iii) $2d(\alpha)\frac{N}{N-d(\alpha)-1}$.

In each case the effective degrees of freedom $d(\alpha)$ is estimated using the simple approximation, i.e. the number of non-zero coefficients in the model, see (2.29).

Penalty (i) was used in the original definition of AIC by Akaike (1974) and is linear in the number of predictors. Phao et al. (2009) noted that using AIC with penalty (i) can tend towards overfitting the model when the number of observations is relatively small; as commented earlier, in the extreme case of $N < (p + 1)$, a saturated model will be selected.

Penalty (ii) was suggested by Phoa et al. (2009) for model selection for supersaturated designs, which are designs with not enough runs to enable estimation of all the main effects. Penalty (ii) is quadratic in the number of terms in the model and will penalise complex models more heavily than penalty (i), leading to the selection of a more parsimonious model.

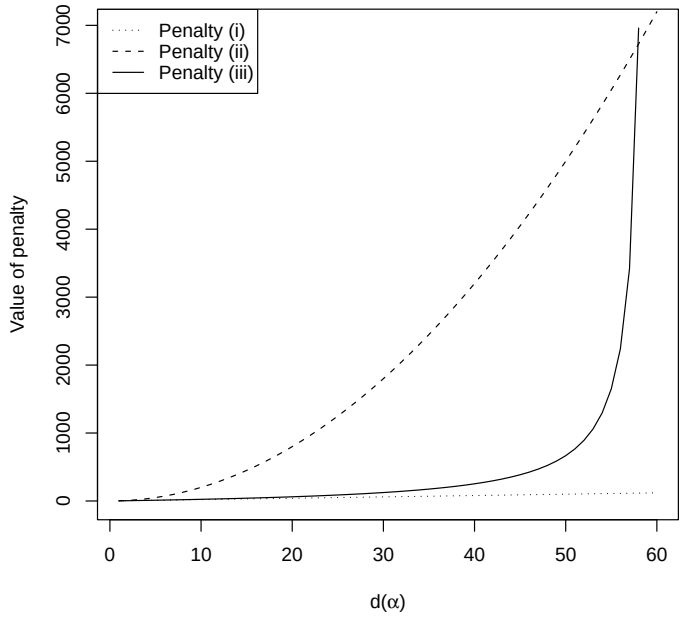
Hurvich and Tsai (1989) noted that using AIC with penalty (i) will overfit severely unless a restriction was placed on the largest model considered. They commented that this overfitting is caused by AIC becoming a biased estimate of the Kullback-Leibler information as $d(\alpha)$ increases compared to N. They therefore developed penalty (iii) as an additional bias-correction term. As $d(\alpha) \to N$, penalty (iii) becomes close to quadratic and is expected to behave similarly to penalty (ii), heavily penalising complex models; for smaller $d(\alpha)$, penalty (iii) behaves more like penalty (i). Burnham and Anderson (2002, page 66), suggested the use of penalty (iii) particularly when the ratio $N/d(\alpha) < 40$ for the model of highest dimension being considered. A saturated model for the Melting Point Data Set has a ratio of $N/d(\alpha) = 1$ therefore the use of penalty (iii) for this example is justified.

When using penalty (iii) the effective degrees of freedom must be restricted so that $d(\alpha) < N - 1$. This prevents the denominator of the penalty equalling zero. For the lasso and LARS and $d(\alpha) = |T|$ this is not a problem. The LARS algorithm used to select the variables to be included in the model terminates at a saturated least squares fit, when $N - 1$ variables have entered the model. For ridge regression and the Dantzig selector the effective degrees of freedom increases as the value of the tuning parameter, λ and ρ respectively, decreases. It is possible for the effective degrees of freedom to be greater than $N - 1$ if the value of the tuning parameter is small enough. Therefore, for these two methods, any value of the tuning parameter that results in a model where $d(\alpha) \geq N - 1$ should be excluded from the search for the optimum model.

Figure 3.5 shows how each of the degrees of freedom penalties change for $N = 60$

as the number of terms in the model is increased from that of the null model, $d(\alpha) = 0$, to a saturated model, $d(\alpha) = 59$. The increase in the value of penalty (i) as the number of terms in the model increases is so small compared to the other penalties that it appears to be constant. Penalty (ii) increases rapidly as the number of terms in the model increases. Penalty (iii) increases at a similar rate as penalty (i) when there are few terms in the model up until $d(\alpha) \approx 20$ after which it begins to increase more rapidly. As $d(\alpha)$ increases past 50, penalty (iii) increases very quickly. As $d(\alpha) \to 60$, penalty (iii) tends to infinity. Thus penalty (iii) would be able to compensate for the very small RSS of the saturated model, leading to a more parsimonious model being chosen as optimum, while not overly penalising smaller models.

Figure 3.5: Value of AIC penalties (i)-(iii) against effective degrees of freedom



3.7.2 Lasso Models with Linear Terms

Each of the degrees of freedom penalties (i)-(iii) were used with AIC (2.25) to select the optimum value of the standardised complexity parameter, s, for a lasso model

including only linear terms. The resulting models are summarised in Table 3.5 along with their R^2 and adjusted R^2 statistics.

Penalty (i) selects the largest complexity parameter of $s = 1$ which corresponds to a model including all of the predictors with coefficients equal to the ordinary least squares estimates. Penalty (iii) selects a model that exhibits slightly more coefficient shrinkage and fewer terms. Penalty (ii) selects the smallest value of the complexity parameter corresponding to a model including only two terms. The models selected using penalties (i) and (iii) perform similarly when assessed using the R^2 and adjusted R^2 statistics. The values of these statistics are also similar to those exhibited by the linear term only models summarised in Table 3.3, and the models have several terms in common, including variables B and D which have been selected by all variable selection methods. Variable N, selected in Section 3.6, is selected using penalty (iii) but not (ii).

Table 3.5: Summary of lasso models with linear terms chosen by AIC with penalties (i)-(iii)

Penalty	s	Terms in the model	R^2	Adjusted R^2
(i)	1	$A, B, D, F, G, H, K, L, N, P, Q, R, S, T, U$	0.867	0.822
(ii)	0.21	B, D	0.709	0.699
(iii)	0.52	A, B, D, G, K, N, P, R	0.836	0.810

The model selected using penalty (ii) exhibits much smaller values of R^2 and adjusted R^2. This indicates that penalty (ii) penalises the model according to its complexity too heavily and results in a model that is too simple and underfits the data.

These results are supported by the trends in the values of the three penalties shown in Figure 3.5. The term $N \log(\text{RSS}/N)$ in the AIC criterion is the same irrespective of the penalty being applied, so any differences in the selected value of the complexity parameter are caused only by the different penalties. Within the range $0 \leq d(\alpha) \leq 15$, which corresponds to the range of models between the null model and the model including every linear term, the values of penalties (i) and (iii) are very similar. Therefore it is not unexpected for similar models to be chosen when using these penalties. Penalty (iii) increases slightly faster than penalty (i) as $d(\alpha)$ is increased towards $d(\alpha) = 15$, therefore penalty (iii) will penalise complicated

models more strongly than penalty (i), resulting in the slightly smaller model chosen (Table 3.5). Penalty (ii) increases much more quickly than either penalty (i) or (iii) as $d(\alpha)$ increases towards $d(\alpha) = 15$, therefore it will penalise larger models much more heavily than the other two penalties. It is for these reasons that the simplest model is chosen when using penalty (ii).

3.7.3 Lasso Models with Interactions

Each of the degrees of freedom penalties (i)-(iii) were used with AIC to select the optimum value of the standardised complexity parameter, s, for a lasso model including linear and interaction terms. The resulting models are summarised in Table 3.6 along with their R^2 and adjusted R^2 statistics.

Penalty (i) again, as expected, selects $s = 1$ which corresponds to a saturated model with coefficients equal to the least squares estimate. This model has $R^2 = 1$ but it is not possible to calculate the adjusted R^2 since there are $p = 59$ predictors in the model. It is likely that this model is overfitted to the training data set and is unlikely to generalise well to unseen data.

Table 3.6: Summary of lasso models with linear and interaction terms chosen by AIC with penalties (i)-(iii)

Penalty	s	Terms in the model	R^2	Adjusted R^2
(i)	1	$A, B, D, F, G, N, Q, R, T, U, AB, AD, AG,$ $AH, AK, AN, AR, AS, AT, BF, BH, BN,$ $BT, BU, DH, DK, DL, DN, DQ, DU, FP,$ $FR, GH, GP, GU, HL, HS, KP, KS, KT,$ $LN, LP, LR, LU, NP, NQ, NR, NS, NU,$ $PS, PU, QR, QS, QT, RS, RT, ST, SU,$ TU	1	NA
(ii)	0.08	B, D	0.706	0.696
(iii)	0.23	$B, D, G, N, T, AB, AH, BF, BL, DN, DR,$ $GP, GR, GU, HL, KT, LR, QR, RS$	0.914	0.873

Penalty (iii) selects a smaller value of s than penalty (i), which leads to a much simpler model. The difference in the value of s selected by penalties (i) and (iii) is much larger for the models including interactions than for the models including only linear terms. This is supported by the evidence provided in Figure 3.5; the

values of penalties (i) and (iii) were similar when $0 \leq d(\alpha) \leq 15$ but the value of penalty (iii) increases much more rapidly than penalty (i) as $d(\alpha)$ increases towards $d(\alpha) = 60$ and the saturated model is reached. Therefore penalty (iii) penalises complex models more heavily than penalty (i) resulting in the selection of a simpler model.

Penalty (ii) selects the smallest value of the complexity parameter, leading to a model with only two terms. These are the same two terms that were included in the model selected by penalty (ii) when only linear terms were considered. Figure 3.5 shows that the value of penalty (ii) is greater than the values of penalties (i) and (iii) for all values of $d(\alpha)$, therefore it penalises the models more heavily and tends to select much simpler models. The model chosen by penalty (ii) has smaller values of R^2 and adjusted R^2 than the model chosen by penalty (iii) which indicates that the model is too simple and underfits the data.

The model chosen by penalty (iii) exhibits the best values of R^2 and adjusted R^2 of the three models; this model also outperforms the models containing only linear terms summarised in Table 3.3.

3.7.4 Applying Penalty (iii) with other Coefficient Shrinkage Methods

Penalty (iii) is now applied with AIC to select the value of the complexity parameter for ridge regression, LARS and the Dantzig selector. The optimal models selected are summarised in Table 3.7.

For this scenario, the lasso and LARS choose different models, with 19 and 17 terms respectively. The models include both linear and interaction terms and are not saturated. The ridge regression model includes every linear and interaction term, but a large amount of shrinkage has been applied to the coefficients of the variables, with $\lambda = 102.95$ chosen as optimum. The Dantzig selector selects a model including 19 terms as optimum.

The interaction terms chosen by the lasso, LARS and Dantzig selector are all scientifically appropriate. For example, all three models include the interaction KT, which describes the joint action of unit cell volume and the number of molecules that surround one molecule in the crystal structure on the melting point. A compound with a unit cell that has low volume and with many molecules surrounding

Table 3.7: Summary of results of models including interaction terms chosen by AIC using penalty (iii)

Method	Terms in the model	R^2	Adjusted R^2
Lasso	$B, D, G, N, T, AB, AH, BF, BL, DN, DR, GP, GR,$ GU, HL, KT, LR, QR, RS	0.914	0.873
LARS	$B, D, G, N, AB, BF, BL, DN, DR, GP, GR, GU, HL,$ KT, LR, QR, TU	0.896	0.854
Ridge	All main effects and two-factor interactions	0.890	0.829
Dantzig	$B, D, G, N, T, AB, AH, BH, DN, DR, FU, GP, GR,$ GT, HL, KT, LR, QR, TU	0.906	0.862

a single molecule will cause the molecules to be closer together. As a consequence the intermolecular forces between the molecules will be stronger, requiring a higher temperature to melt the compound. All three models also include the interaction QR, which describes the joint action of the angle of hydrogen bonds and the length of hydrogen bonds on the melting point. Shorter hydrogen bonds that are close to 180° will be strong, requiring a higher temperature to break the bond in order to separate the molecules and melt the compound.

In terms of R^2 and adjusted R^2 the model fitted via the lasso is the best performing model. In the absence of an external test set, cross-validation was carried out on each of these models in 5-, 10- and N-folds. Out-of-sample R^2, adjusted R^2 and RMSE were obtained for each model and for each fold size. The results of the cross-validation are shown in Table 3.8.

Under cross-validation the largest model, the ridge regression model, performs poorly. This is because including every term makes the model complex and overfitted to the data. For this model, a negative adjusted R^2 is obtained from cross-validation. This tells us that a model including only the mean would be a better predictive model. The lasso outperforms the other methods in out-of-sample performance. When compared to the models including only linear terms (Table 3.3) the lasso model with interactions also has better summary measures. This model also equals the performance of the models including only linear terms when evaluated using cross-validation.

The model including interaction terms fitted using the lasso was used to predict the melting points of the two compounds that were not included in the set of

Table 3.8: Model validation statistics for models including product terms chosen by AIC using degrees of freedom penalty (iii)

	5-fold c-v			10-fold c-v			N-fold c-v		
Method	R^2	Adj.R^2	RMSE	R^2	Adj.R^2	RMSE	R^2	Adj.R^2	RMSE
Lasso	0.72	0.64	19.2	0.74	0.64	18.3	0.71	0.57	19.4
LARS	0.63	0.51	21.8	0.61	0.39	22.5	0.67	0.55	20.7
Ridge	0.30	-0.09	30.1	0.22	-0.21	31.9	0.24	-0.19	31.4
Dantzig	0.71	0.62	19.3	0.74	0.59	18.4	0.70	0.56	19.8

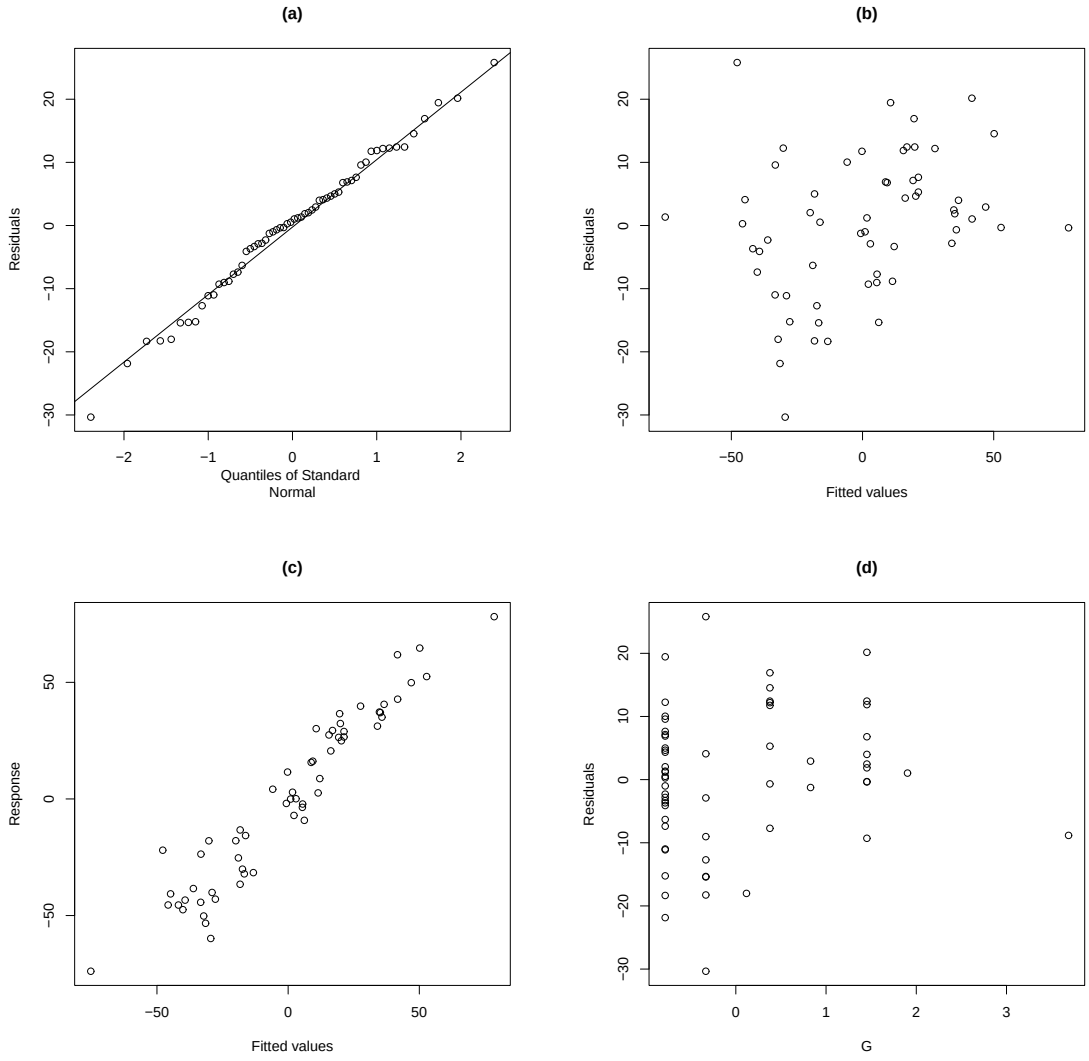
compounds used to develop the model. For the compound with R^1=Br and R^2=I, the model gave a predicted value of 170.09°C. For the compound with R^1=I and R^2=Br, the model gave a predicted value of 167.54°C. This second predicted value is very similar to the predicted melting point for the same compound from model (3.6). The RMSE for these predictions is 11.57°C. This is approximately twice that of the RMSE for the same two compounds obtained from model (3.6). Of course, few conclusions can be drawn from such a small test set; a better indication of out-of-sample performance is probably the cross-validation statistics.

Figure 3.6 shows the diagnostic residual plots for the lasso model. The distribution of points in each of these plots is indicative of a well fitting model that satisfies all model assumptions.

3.8 Conclusions

In this chapter variable selection and regression methods were applied to the motivating data set, the Melting Point Data Set, in order to build a descriptive model and identify the variables that have an important effect on melting point. Several methods of variable selection and regression modelling were investigated when building models involving only linear terms (main effects). The best performing model, in terms of R^2 and adjusted R^2, was found when variable selection was carried out using best subset selection, choosing the final model by maximising the adjusted R^2. The models fitted using the coefficient shrinkage methods (ridge regression, the lasso, LARS and the Dantzig selector) also performed well under the same criteria. The performance of the model fitted using the lasso compared favourably with predictive melting point models reported in recent literature.

Figure 3.6: Diagnostic residual plots for lasso model including product terms: (a) Normal Q-Q plot of residuals; (b) Fitted values vs. residuals; (c) Fitted values vs. response; (d) Variable G vs. residuals



Models involving both linear terms and interaction terms were also considered. Including interactions led to more potential predictors than observations. Models were fitted using the four coefficient shrinkage methods. Since a model independent or low bias estimator for the error variance is not available when $N < (p + 1)$, the tuning parameter was selected using the AIC criterion defined in (2.25), with

$d(\alpha) = |T|$ for the lasso, LARS and Dantzig selector. For ridge regression, $d(\alpha)$ was calculated using (2.28).

When $N < (p + 1)$ it is possible to fit a saturated model for which RSS $\approx$ 0; hence with the standard AIC penalty, $2d(\alpha)$, a saturated model is consistently chosen as optimum. This model is unlikely to generalise well to future unseen data as it interpolates between every observation. Therefore we have considered two modifications to the degrees of freedom penalty that are able to penalise complex models more heavily.

The best performing lasso models, in terms of R^2 and adjusted R^2, were obtained when the degrees of freedom penalty was equal to

$$2d(\alpha)\frac{N}{N - d(\alpha) - 1}. \tag{3.7}$$

When only linear terms are considered, and $N \geq (p+1)$, this penalty selects a similar model, using the lasso, to that selected by the alternative AIC criterion (2.26) using the sum of the covariances (2.31) to estimate the effective degrees of freedom. This adjusted penalty was also used with AIC to select optimal models including linear and interaction terms under the lasso, LARS, ridge regression and Dantzig selector. It performed particularly well when using the lasso, LARS and the Dantzig selector.

The model including interaction terms fitted using the lasso was the best performing in terms of within- and out-of-sample R^2 and adjusted R^2, and performed better than any of the models including only linear terms. Therefore, in situations where $N < (p + 1)$ it is suggested that the AIC criterion (2.25) with the degrees of freedom penalty (3.7) is used to select the optimal model.

The predictors, including main effects and interactions terms, chosen for inclusion in the best performing models all have good scientific reasons for their inclusion. Variables previously thought to have a significant affect on melting point, such as molecular weight, enthalpy of fusion and molecular dipoles, were frequently chosen for inclusion and the resulting models performed well when assessed within-sample and out-of-sample. Many of the variables considered apply directly to the functional groups substituted at either end of the molecular structure, such as the angle of hydrogen bonds, which would be applicable to other compounds that include the same functional groups but are more structurally diverse than those of our motivating data set. Other descriptors are more general and describe the molecule as a whole,

such as molecular weight. It would be important to consider both types of variables when developing models for sets of more structurally diverse compounds.

The descriptive melting point models obtained in this chapter were developed using small data sets of small organic compounds with similar molecular structures. These models may not generalise well to a set of larger, more complex compounds with more diverse structures (see, for example, Hughes et al., 2008). There is no reason to suggest, however, that the variable selection and regression methods employed in this chapter may not be applied, with similar results, to data sets of more complex molecules. This would also raise the question of whether the variables considered in this chapter are adequate in describing the characteristics that affect the melting points of these more structurally diverse compounds. This would be the natural next direction in which to take the investigation.

Chapter 4

Optimal Design of Experiments for Bridge Regression

4.1 Introduction

In this chapter we investigate the problem of how to select design points for an experiment to enable efficient estimation of a bridge regression model from the data obtained. The selection of a design is especially needed in situations where there are limited resources or when data collection is slow. In this chapter, designs are found by a Bayesian approach. Chaloner and Verdinelli (1995) presented a review of Bayesian experimental design, including Bayesian D-optimality as well as other Bayesian 'alphabetical' optimality criteria.

We use a Bayesian D-optimality criterion which is particularly useful when we want to estimate the parameters in the model accurately, for example, when they provide knowledge and understanding about the science, as in the chemistry example in Section 1.2. A Bayesian D-optimality criterion for finding designs for bridge regression has not, as far as we are aware, been developed yet in the literature. We develop a Bayesian D-optimality methodology to find designs for bridge regression, and find designs for two cases: ridge regression and the lasso. For the lasso, the criterion requires the derivation of a normal approximation to the posterior distribution since the posterior distribution is not available in closed form.

We begin by establishing in Section 4.2 a link between bridge regression and Bayesian estimation for the linear model. For bridge regression with $0 < \gamma \leq 2$, we

derive a normal approximation for the posterior distribution of $\boldsymbol{\beta}$. The objective functions used to select ridge regression and lasso Bayesian D-optimal designs are defined in Section 4.3. In Section 4.4, designs are given for experiments where the factors each have two levels and both main effects and two-factor interactions are included in the model. These designs are compared to a catalogue of main effect orthogonal designs provided by Sun, Li and Ye (2002). In Section 4.5 designs are found for an example based on the motivating data set introduced in Section 1.2. For this example, the values for a variable cannot be chosen independently of the values of all other variables, since it is highly unlikely that a compound exists for every possible combination of variable values. Instead, design points from a candidate set of allowable design points are considered for inclusion in the designs. In Section 4.6 some conclusions are drawn.

4.2 Connections Between Bayesian Inference and Bridge Regression

Bridge regression can be considered from a Bayesian perspective with $\boldsymbol{\beta}$ in linear model (1.1) having prior density

$$f(\boldsymbol{\beta}) \propto \prod_{j=1}^{p} e^{-\lambda|\beta_j|^\gamma/2\sigma^2}, \tag{4.1}$$

where we assume σ^2 is known, $\lambda \geq 0$ and $0 < \gamma \leq 2$. Parameter β_0 in (1.1) is assumed to have a noninformative prior distribution defined by $f(\beta_0) \propto 1$. As in Chapters 2 and 3, $\boldsymbol{Y}$ and each column of $\boldsymbol{X}$ are assumed to have been centred by subtraction of their respective means.

The log posterior density for $\boldsymbol{\beta}$ is

$$\begin{aligned} \log f(\boldsymbol{\beta}|\boldsymbol{Y}) &\propto \log f(\boldsymbol{Y}|\boldsymbol{\beta}) + \log f(\boldsymbol{\beta}) \\ &\propto -(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta})^{\mathrm{T}}(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}) - \lambda\sum_{j=1}^{p}|\beta_j|^\gamma. \end{aligned} \tag{4.2}$$

It follows that the posterior mode for $\boldsymbol{\beta}$ is the solution $\hat{\boldsymbol{\beta}}$ of the bridge problem (2.2). This was observed for the special case of the lasso by Park and Casella (2008). This

relationship between the bridge penalty function and a Bayesian prior distribution was also considered by Fu (1998).

When $\gamma = 2$ (ridge regression), from (4.1) $\boldsymbol{\beta}$ has a multivariate normal prior distribution with mean $\mathbf{0}$ and variance-covariance matrix $\frac{\sigma^2}{\lambda}\boldsymbol{I}$ and, from (4.2), the posterior distribution of $\boldsymbol{\beta}$ is also normal, provided that the assumptions of model (1.1) hold for $\boldsymbol{Y}$. Hence the log posterior density (4.2) has the variance-covariance matrix

$$\text{Var}(\boldsymbol{\beta}|\boldsymbol{Y}) \propto (\boldsymbol{X}^{\text{T}}\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}, \tag{4.3}$$

where the constant of proportionality does not depend on $\boldsymbol{X}$.

When $0 < \gamma < 2$, the posterior distribution of $\boldsymbol{\beta}$ is not available in closed form. However, an approximation to the variance-covariance matrix of the posterior distribution of $\boldsymbol{\beta}$ is the inverse of the observed information

$$\left[-\frac{\partial^2}{\partial\boldsymbol{\beta}^2}\log f(\boldsymbol{\beta}|\boldsymbol{Y})\right]^{-1}, \tag{4.4}$$

see, for example, Gelman, Carlin, Stern and Rubin (2004, Chapter 4).

The first and second derivatives of the log likelihood are well known and have the form

$$\frac{\partial\log f(\boldsymbol{Y}|\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} \propto \boldsymbol{X}^{\text{T}}\boldsymbol{Y} - \boldsymbol{X}^{\text{T}}\boldsymbol{X}\boldsymbol{\beta}, \qquad \frac{\partial^2\log f(\boldsymbol{Y}|\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^2} \propto -\boldsymbol{X}^{\text{T}}\boldsymbol{X}.$$

The log prior density can be seen from (4.1) to contain the term $|\beta_j|^\gamma$. For $0 < \gamma < 2$, $|\beta_j|^\gamma$ is not differentiable at $\beta_j = 0$, because $|\beta_j|$ is not differentiable there. However, $|\beta_j|$ is a convex function and hence subderivatives, described below, can be used to find the first and second derivatives of the log posterior distribution of $\boldsymbol{\beta}$.

The absolute value function $|x|$ is subdifferentiable with respect to x at every $x \in \mathbb{R}$ but differentiable only when $x \neq 0$ (see Rockafellar, 1970, page 215). A *subderivative* of a function $f(x)$ at a point x_0 is a real number w such that

$$f(x) - f(x_0) \geq w(x - x_0). \tag{4.5}$$

The *subdifferential* is defined as the interval $[u, v]$ such that for all $w \in [u, v]$, equation (4.5) holds. If $u = v$, then the function $f(x)$ is differentiable at x_0.

Intuitively, a subderivative can be thought of as the gradient of a tangent to $f(x)$ at x_0 which passes through the point $(x_0, f(x_0))$ and everywhere is either touching or below $f(x)$. Therefore, for the convex function $f(x) = |x|$, when $x \neq 0$ there is only one tangent that satisfies this condition and hence $|x|$ is differentiable at $x \neq 0$. Conversely, at $x = 0$, there is an infinite number of tangents to the function that satisfy this condition, with slope in $[-1, 1]$. The slope of one of these tangents is a subderivative of $|x|$ at $x = 0$, and the set of all subderivatives is the subdifferential of $|x|$ at $x = 0$. One subderivative of $|x|$, which summarises the gradient of the absolute value function at every x, is $\text{sgn}(x)$, defined by

$$\text{sgn}(x) = \begin{cases} 1 & x > 0 \\ 0 & x = 0 \\ -1 & x < 0. \end{cases}$$

It follows that the first subderivative of the log prior density

$$\log f(\boldsymbol{\beta}) \propto -\lambda \sum_{j=1}^{p} |\beta_j|^{\gamma},$$

has the form

$$\frac{\partial \log f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \propto -\lambda\gamma \boldsymbol{b}, \tag{4.6}$$

where

$$\boldsymbol{b} = \begin{bmatrix} |\beta_1|^{\gamma-1}\text{sgn}(\beta_1) \\ \vdots \\ |\beta_p|^{\gamma-1}\text{sgn}(\beta_p) \end{bmatrix}. \tag{4.7}$$

We next find the second subderivative of the log prior distribution of $\boldsymbol{\beta}$. To do this we use the Dirac delta function for β_j which is the distributional derivative of the signum function, $\text{sgn}(\beta_j)$. This function takes the value zero everywhere except at $\beta_j = 0$, where its value is infinitely large in such a way that its total integral over the real line $\mathbb{R}$ is 1, i.e.

$$\delta(\beta_j) = \begin{cases} +\infty & \beta_j = 0 \\ 0 & \beta_j \neq 0, \end{cases} \tag{4.8}$$

thus representing the rate of change of $\text{sgn}(\beta_j)$, and reflecting the lack of continuity at $\beta_j = 0$. It is technically not a function, but a distribution that generalises the

derivative of $\text{sgn}(\beta_j)$ making it possible to differentiate the signum function, which is not differentiable in a classical sense at $\beta_j = 0$.

Hence, we derive the second subderivative of the log prior as

$$\frac{\partial^2 \log f(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} \propto -\lambda\gamma \boldsymbol{A},$$

where $\boldsymbol{A}$ is a diagonal matrix with $\boldsymbol{A}_{jj} = 2\delta(\beta_j)|\beta_j|^{\gamma-1} + \text{sgn}^2(\beta_j)(\gamma-1)|\beta_j|^{\gamma-2}$. The form of the first and second derivatives of the log posterior density for general γ can now be obtained from (4.2) as

$$\frac{\partial \log f(\boldsymbol{\beta}|\boldsymbol{Y})}{\partial \boldsymbol{\beta}} \propto \boldsymbol{X}^{\mathrm{T}}\boldsymbol{Y} - \boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}\boldsymbol{\beta} - \lambda\gamma \boldsymbol{b}, \qquad \frac{\partial^2 \log f(\boldsymbol{\beta}|\boldsymbol{Y})}{\partial \boldsymbol{\beta}^2} \propto -\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} - \lambda\gamma \boldsymbol{A}.$$

Hence, an approximate posterior variance-covariance matrix is given by

$$\text{Var}(\boldsymbol{\beta}|\boldsymbol{Y}) \approx \left[-\frac{\partial^2 \log f(\boldsymbol{\beta}|\boldsymbol{Y})}{\partial \boldsymbol{\beta}^2} \right]^{-1} = (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\gamma \tilde{\boldsymbol{A}})^{-1}, \tag{4.9}$$

where $\tilde{\boldsymbol{A}}_{jj} = 2\delta(\tilde{\beta}_j)|\tilde{\beta}_j|^{\gamma-1} + \text{sgn}^2(\tilde{\beta}_j)(\gamma-1)|\tilde{\beta}_j|^{\gamma-2}$ and the $\tilde{\beta}_j$ are treated as hyper-parameters quantifying prior knowledge about $\boldsymbol{\beta}$.

As an example, consider the lasso ($\gamma = 1$). In this case the prior density (4.1) is the double exponential or Laplace distribution. In (4.9), $\tilde{\boldsymbol{A}}_{jj} = 2\delta(\tilde{\beta}_j)$, therefore the variance approximation is equivalent to approximating (4.1) using improper uniform prior distributions for those β_j thought likely to be non-zero ($\delta(\beta_j) = 0$), and point mass prior distributions at 0 for those β_j thought likely to be equal to zero ($\delta(\beta_j) = \infty$). This approximation will be used in Sections 4.4 and 4.5 to find Bayesian D-optimal designs for the lasso.

4.3 Bayesian D-optimal Design Criteria

The selection of a design for an experiment under a particular model may be viewed as a decision problem. When prior information is available on unknown parameters, Bayesian methods can be used (Chaloner and Verdinelli, 1995), and a design selected that maximises a specified utility function which is relevant to the purpose of the experiment. For example, the purpose of the experiment might be prediction or

model estimation, and the optimal designs for each of these goals may be different.

The Bayesian D-optimality criterion aims to choose an optimal design for a regression model, using prior information about the value of $\boldsymbol{\beta}$, by selecting design points that maximise the determinant of the inverse of the posterior variance-covariance matrix. To find designs for bridge regression, a Bayesian D-optimality criterion can be applied by minimising the determinant of (4.9).

A Bayesian design criterion is sometimes thought of as an extension of the corresponding non-Bayesian criterion which, for D-optimality under linear model (1.1), selects a design which maximises the determinant of the information matrix, $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$. This criterion is equivalent to minimising the determinant of the variance-covariance matrix of the least squares parameter estimators, under the assumption that the distributions for the errors are independent and identical normal distributions. An equivalent interpretation arises from the fact that the volume of the joint confidence region for the $(p+1)$ parameters is inversely proportional to the square root of $|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}|$ (see Atkinson, Donev and Tobias, 2007, Chapter 6). Hence a D-optimal design minimises the volume of the joint confidence region. Chaloner and Verdinelli (1995) stated that a non-Bayesian criterion is the limiting case of a Bayesian criterion when little prior information is available. In this case there would be no practical advantage in using a Bayesian criterion over a non-Bayesian criterion.

The Bayesian D-optimality criterion is appropriate when inference about the model parameters is important, for example, for scientific interpretation. The design is chosen to maximise the expected gain in the Shannon information between the prior and the posterior distributions, where the Shannon information is defined as a measure of information about the values of a finite number of parameters provided by an experiment. This measure is obtained by a comparison of the knowledege of the parameters before and after the experiment has been carried out, using prior and posterior distributions. Since the prior distribution does not depend on the design, this criterion is the same as choosing the design to maximise the expected Shannon information of the posterior distribution. In practice, this selection is achieved by maximising the determinant of the inverse of the posterior variance-covariance matrix.

For the remainder of this chapter, we will assume that $\sigma^2 = 1$. For linear model (1.1) and prior distribution $N(\mathbf{0}, \boldsymbol{K}^{-1})$ for the model parameters $\boldsymbol{\beta}$, the criterion for

Bayesian D-optimality is

$$\arg\max_{\xi\in\Xi} |\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{K}|, \tag{4.10}$$

where ξ is a design from the set of all possible designs Ξ and $\boldsymbol{K}^{-1}$ is the prior variance-covariance matrix (see Chaloner and Verdinelli, 1995). A special case of this criterion is when the prior distributions of β_j $(j = 1, \ldots, p)$ are independent and identical normal distributions, i.e. $\boldsymbol{\beta} \sim N(\boldsymbol{0}, \tau^2\boldsymbol{I})$, where $\tau^2 > 0$ is the prior variance of β_j $(j = 1, \ldots, p)$. A Bayesian D-optimal design is then

$$\xi^* = \arg\max_{\xi\in\Xi} |\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \frac{1}{\tau^2}\boldsymbol{I}|. \tag{4.11}$$

DuMouchel and Jones (1994) found designs for the linear model (1.1) using a Bayesian D-optimality criterion that took account of the presence of two kinds of terms:

- c *primary* terms which are likely to have large effects, and
- d *potential* terms which are thought unlikely to have large effects.

This partition of terms led to the authors to choose matrix $\boldsymbol{K}$ in (4.10) as a $p \times p$ diagonal matrix, $\boldsymbol{Q}$, with the first c diagonal elements equal to 0, and last d elements equal to $1/\tau^2$, where $c + d = p$. Here, τ^2 is the prior variance of the coefficients of the potential terms. Hence, their criterion was

$$\max_{\xi\in\Xi} |\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \boldsymbol{Q}|. \tag{4.12}$$

Bayesian *D*-optimal Designs for Ridge Regression

We define a ridge regression Bayesian D-optimality design criterion as: select a design, ξ_R^*, such that

$$\xi_R^* = \arg\max_{\xi\in\Xi} |\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda\boldsymbol{I}|. \tag{4.13}$$

This criterion may be obtained from (4.11) by substituting $\lambda = 1/\tau^2$.

Bayesian *D*-optimal Designs for the Lasso

For the lasso, prior density (4.1) is approximated by a marginal point mass prior distribution for each coefficient thought to be equal to zero a priori, and each re-

maining coefficient has a diffuse, improper prior distribution. This classification of coefficients is analogous to that of DuMouchel and Jones (1994) into primary and potential terms.

In order to ensure that the approximate posterior variance-covariance matrix, (4.9), is positive definite, this prior density is adjusted to obtain

$$\mathrm{Var}(\boldsymbol{\beta}|\boldsymbol{Y}) \approx (\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + 2\lambda\Delta_0(\boldsymbol{\beta}))^{-1},$$

where $\Delta_0(\boldsymbol{\beta})$ is a diagonal matrix with jth entry $\delta_0 > 0$ if $\beta_j = 0$, and 0 otherwise. Here the value of δ_0 encapsulates prior knowledge about the size of the potential terms in the model: a large value of δ_0 indicates that the potential terms are small. The diagonal elements of $\Delta_0(\boldsymbol{\beta})$ that are assigned value zero correspond to the primary terms (i.e. the large terms). In other words, we relax the implicit assumption in (4.9) that those β_j thought likely to be zero have point mass prior distributions at 0, and instead give them prior distributions with non-zero, but small, variance.

For the lasso, a Bayesian D-optimal design is defined as

$$\xi_L^* = \arg\max_{\xi\in\Xi} |\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + 2\lambda\Delta_0(\boldsymbol{\beta})|. \tag{4.14}$$

A comparison of (4.14) and (4.12) shows that the $2\lambda\Delta_0(\boldsymbol{\beta})$ matrix is equivalent to the matrix $\boldsymbol{Q}$ of the designs found for the linear model by DuMouchel and Jones (1994) with $\tau^2 = 1/(2\lambda\delta_0)$. DuMouchel and Jones (1994) suggested $\tau = 1$ as an appropriate value for the prior standard deviation of the coefficients of the potential terms. However, Jones, Lin and Nachtsheim (2008) and Marley and Woods (2010) have reported that the design selected is not greatly affected by the choice of τ. The sensitivity of the choice of design under (4.14) to the value of δ_0 is studied in later sections.

In the following two sections, we find optimal designs for two types of experiment

- all factors (variables) at two values (levels), and model (1.1) with all main effect and two-factor interaction terms,
- the melting point experiment where the values of the variables in each combination cannot be set independently, and model (1.1) is first order, with main effects only.

For each type of experiment, we describe an appropriate search algorithm to find designs, and investigate the sensitivity of the selected designs to changes in the prior distribution for $\boldsymbol{\beta}$.

4.4 Two-level Designs for Ridge Regression and the Lasso

4.4.1 Generation of Designs

Bayesian D-optimal designs are found for experiments having f factors at two levels, labelled +1 and -1, where model (1.1) consists of an intercept, the f main effect and $f(f-1)/2$ interaction terms. The search algorithm applied is coordinate-exchange (Meyer and Nachtsheim, 1995) which is computationally efficient and appropriate when the possible levels for each factor form a finite set.

In this algorithm, a coordinate is the value of a particular factor in a particular run of a design, and each coordinate is changed independently of the other coordinates to optimise the objective function, which is shown in (4.13) for ridge regression and (4.14) for the lasso. A coordinate-exchange algorithm for two-level Bayesian D-optimal design is described below.

Algorithm:

1. Randomly generate M design matrices, $\boldsymbol{D}^1, \ldots, \boldsymbol{D}^M$, of size $N \times f$, with $\boldsymbol{D}^l = (d_{ij}^l)$ and $d_{ij}^l \in \{-1, +1\}$ for $i = 1, \ldots, N$; $j = 1, \ldots, f$; $l = 1, \ldots, M$.

2. For every $\boldsymbol{D}^l$, generate a corresponding model matrix, $\boldsymbol{X}^l$, that includes the interactions between the columns of $\boldsymbol{D}^l$; mean-centre $\boldsymbol{X}^l$.

3. Select, as the starting point of the algorithm, $\boldsymbol{D} = \arg\max_{\boldsymbol{D}^l \in \chi} \boldsymbol{\Phi}$, where χ is the set of M randomly generated matrices and $\boldsymbol{\Phi}$ is the objective function calculated using $\boldsymbol{X}^l$.

4. Replace d_{ij}, the ijth coordinate of $\boldsymbol{D}$, by $-d_{ij}$.

5. Generate for the new $\boldsymbol{D}$ the corresponding model matrix, $\boldsymbol{X}$, including interactions; mean-centre $\boldsymbol{X}$.

6. If $\Delta_{ij}^{\Phi} > 0$, where Δ_{ij}^{Φ} is the change in the value of Φ when d_{ij} is replaced by $-d_{ij}$, keep the exchange; otherwise set d_{ij} to $-d_{ij}$ in $\boldsymbol{D}$ to reject the exchange.

7. Repeat steps 4-6 for $i = 1, \ldots, N$, $j = 1, \ldots, f$.

8. When all Nf coordinates have been considered: if any d_{ij} have been changed repeat steps 4-7; otherwise the algorithm ends.

4.4.2 Comparison with Main Effects Orthogonal Designs

For $f = 4, 5, 6$ and $N = 16$ runs, Bayesian D-optimal designs for models composed of all main effect and two-factor interaction terms were found using coordinate-exchange for each of ridge regression and the lasso. For ridge regression, values of $\lambda = \{0, 10^{-7}, 0.0001, 0.001, 0.02, 0.5, 1\}$ were investigated; this range of λ was suggested by Draper and Smith (1998, page 388) as appropriate for many studies. For the lasso λ was set to 0.5 and values of $\delta_0 = \{0, 0.001, 0.1, 0.25, 0.5, 0.75, 1\}$ were investigated; keeping $\delta_0 < 1$ prevents the prior variance of the coefficients of the potential terms becoming too small, and hence producing designs that have no ability to estimate them (see page 95). In the first instance, the main effects were the primary terms and the interactions were considered to be the potential terms. The impact of this classification on the designs selected is discussed in Section 4.4.3.

In each case, the designs were compared to non-isomorphic main effects orthogonal (MEO) designs catalogued by Sun et al. (2002) (see Appendix C). These designs consist of f columns from one of the five non-isomorphic 16×16 Hadamard matrices of Hall (1961) (see Appendix B). Each row defines a combination of factor levels in the design. Note that an $N \times N$ Hadamard matrix $\boldsymbol{C}$ has entries +1 and -1 and $\boldsymbol{C}^{\mathrm{T}}\boldsymbol{C} = N\boldsymbol{I}_N$. It follows that, for a 'main effects only' linear model, the estimators of the main effects are uncorrelated with each other. The designs are balanced, that is, each factor is set to +1 and -1 the same number of times. These designs are non-isomorphic in the sense that no design can be obtained from another by relabelling the factors, changing the run order, or relabelling the level labels

in the design. The numbers of designs for $f = 4, 5$ and 6 are 5, 11 and 29 respectively.

Investigation of designs for ridge regression: In order to compare Bayesian D-optimal designs for different choices of tuning parameter and $0 \leq \lambda \leq 1$ with the MEO designs, D-optimal designs were found via (4.13) for $\lambda = 0, 10^{-7}, 10^{-4}, 10^{-3}, 0.02, 0.5, 1$. This range of λ was chosen because as λ increases, the prior variance decreases which means that less information is contributed by the design relative to the prior. So for large λ, the value of (4.13) for all designs will be equal. For each MEO design, the value of the objective function

$$\boldsymbol{\Phi}_R(\xi) = \log\left\{|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + \lambda \boldsymbol{I}|^{1/p}\right\}, \tag{4.15}$$

was calculated. This is a rescaling of the objective function in (4.13).

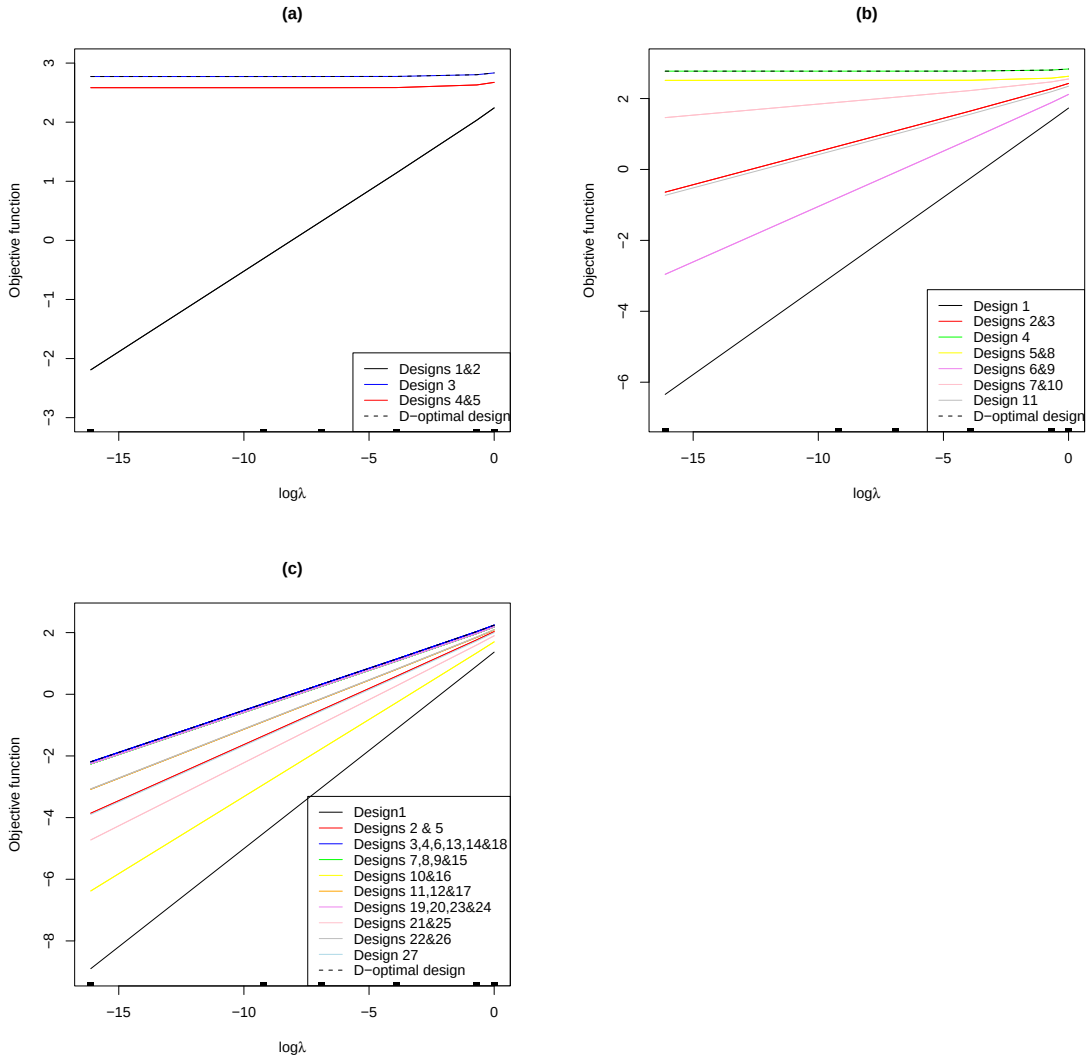
The performance of the designs for each value of f, and varying λ, are shown in Figure 4.1. It was found that, for each value of f, the Bayesian D-optimal design was the same for each value of λ and it was an MEO design. Hence, in Figure 4.1 the performance of the D-optimal design coincides with that of the best MEO design. Note that in Figure 4.1, a design label (e.g. Design 1) indicates a different design for each value of f. The MEO designs have the same labels as given by Sun et al. (2002). Figure 4.1 shows that the performance of the MEO designs improves, and converges, as λ increases, as anticipated. Note that in Figures 4.1(a) and (b), (4.15) does increase with λ for the optimal and other good designs, but only very slightly.

For a regular design, i.e. where all effects are either not aliased or fully aliased, the designs can be classified as follows (Wu and Hamada, 2009, page 217)

- resolution III designs: some main effects aliased with two-factor interactions, but not with other main effects,
- resolution IV designs: no main effects aliased; some two-factor interactions aliased with other two-factor interactions,
- resolution V and VI designs: no aliasing of any main effects or two-factor interactions.

To rationalise the ranking of the designs, by the size of (4.15), the form of the information matrix, $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$, of each design can be examined and the resolution of the

Figure 4.1: Evaluation of 16-run designs (Bayesian D-optimal and main effects orthogonal) for ridge regression: (a) $f = 4$, (b) $f = 5$, (c) $f = 6$. The objective function is (4.15). The values of logλ used to obtain the designs are indicated on the horizontal axis



design can be determined. When $f = 4$ or 5, the best designs for ridge regression were resolution V designs, that is, there is no aliasing of any main effects or two-factor interactions. The value of each diagonal element of the information matrix of these best designs is 16 and all other elements of the matrix are zero, i.e. all column correlations are zero. For ridge regression, the prior distribution is identical for each

β_j; there are no primary or potential terms. Hence it is of little consequence to the performance of the designs which effects, main effects or interactions, are aliased (column correlation 1) or partially aliased (column correlation between 0 and 1). Therefore after the resolution V designs, the next best performing designs are those with partial aliasing. They have off-diagonal elements of $\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X}$, corresponding to partially aliased terms, equal to 8 or -8 (terms of ± 16 indicate full aliasing). The worst performing designs are those with full aliasing. The MEO designs included resolution III and resolution IV designs which were among the worst performing designs; these designs provide no information on some of the terms in model (1.1).

Investigation of designs for the lasso: We now compare Bayesian D-optimal designs for the lasso for $\delta_0 = 0, 0.001, 0.1, 0.25, 0.5, 0.75, 1$ and $\lambda = 0.5$, with the MEO designs. The performance of each design, ξ, is measured by the following objective function, a rescaling of (4.14):

$$\boldsymbol{\Phi}_L(\xi) = \log\left\{|\boldsymbol{X}^{\mathrm{T}}\boldsymbol{X} + 2\lambda\Delta_0(\boldsymbol{\beta})|^{1/p}\right\}. \tag{4.16}$$

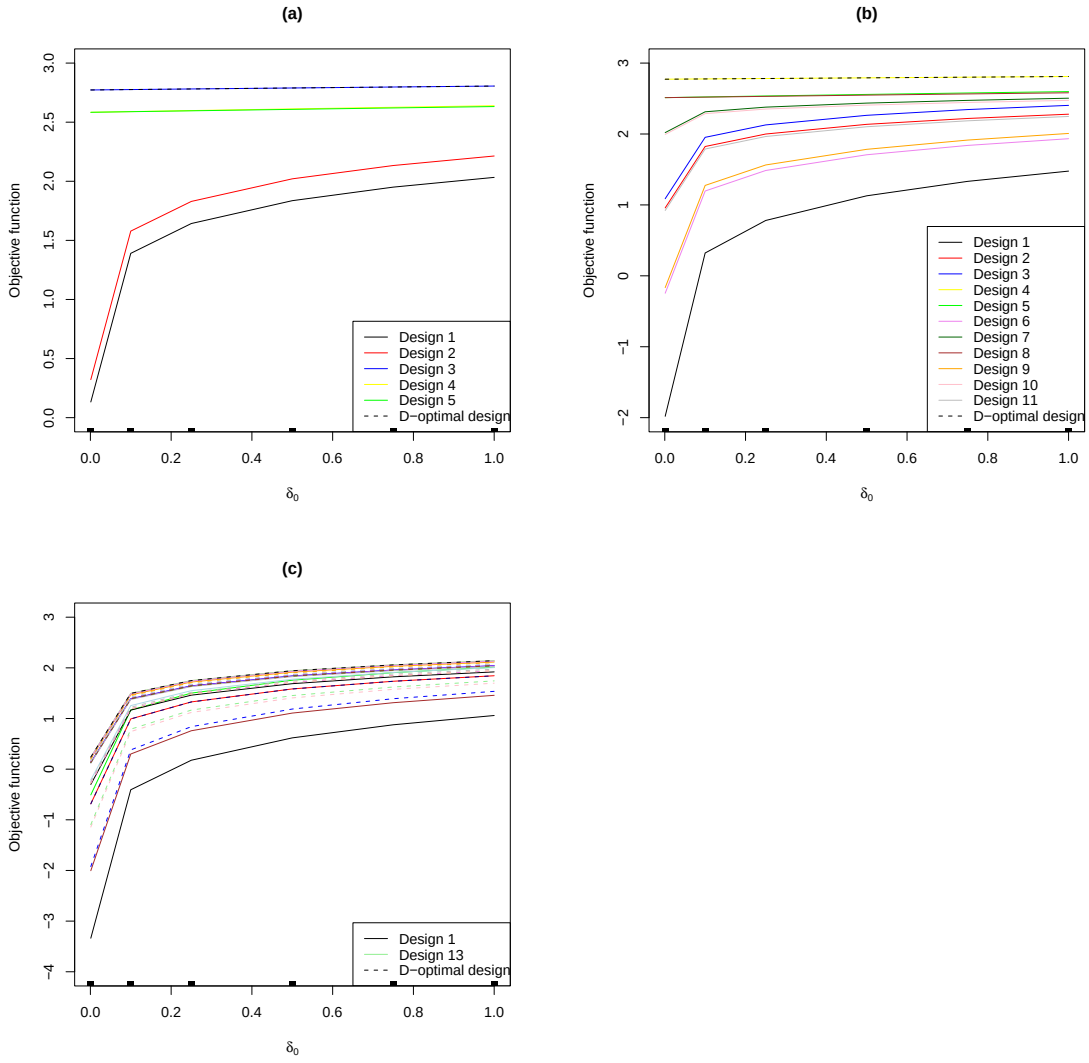
The value $\lambda = 0.5$ makes the prior variance of each potential term (interaction) equal to the prior variance of each term (main effect or interaction) used to find the ridge regression Bayesian D-optimal designs.

The value of (4.16) for each design was plotted against δ_0, see Figure 4.2. For simplicity, the legend of Figure 4.2(c) only labels the D-optimal design and the best and worst performing MEO designs. Note that in Figures 4.2(a) and (b), once again the increase in (4.16) is very small for the optimal and other good designs. As before, for each $f = 4, 5$ and 6, the Bayesian D-optimal design was an MEO design. For $f = 4$ and 5, a resolution V design was again Bayesian D-optimal.

Figure 4.2 shows that the performance of designs converge as δ_0 increases. This trend continues as δ_0 is increased past 1. The explanation is that δ_0 is inversely proportional to the prior variance for the potential terms (the interactions). So large δ_0 means smaller variance, see (4.9). As δ_0 increases, the prior information on the interactions increases and so the design is only required to provide information on the main effects. As the MEO designs all provide the same information on the main effects, the design performances converge as δ_0 increases.

For the lasso, the prior distribution is not the same for all effects. A more

Figure 4.2: Evaluation of 16-run designs (Bayesian D-optimal and main effects orthogonal) for the lasso: (a) $f = 4$, (b) $f = 5$, (c) $f = 6$. The objective function is (4.16). The values of δ_0 used to obtain the designs are indicated on the horizontal axis



informative prior is placed on the potential terms, thus penalising a design that aliases potential terms with primary terms. The ranking of the lasso designs, from best performing to worst performing, can be summarised as follows:

(i) Designs with no aliasing (resolution V designs)

(ii) Designs with partial aliasing between potential terms

(iii) Designs with partial aliasing between primary terms and potential terms

(iv) Designs with full aliasing between potential terms

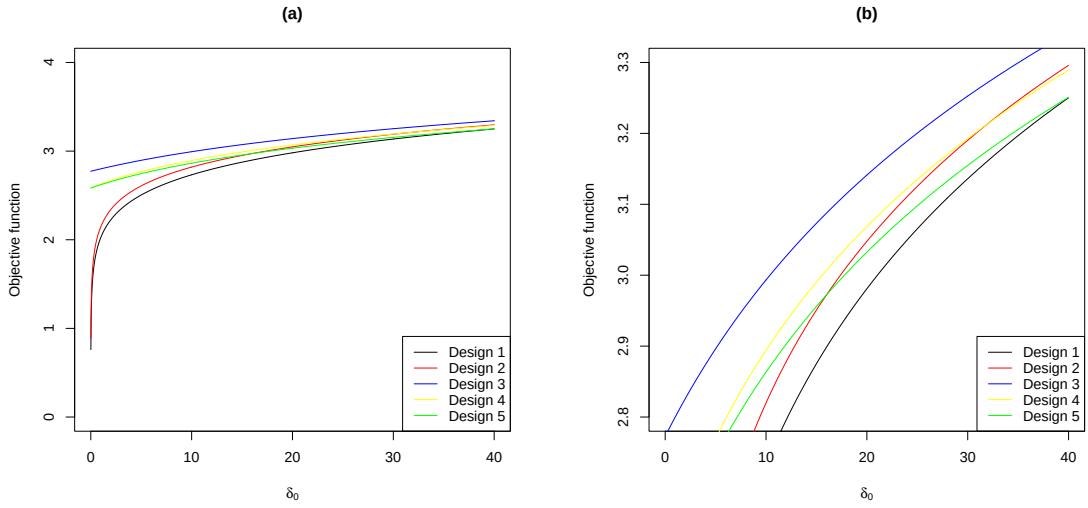
(v) Designs with full aliasing between primary terms and potential terms

Note that, as the designs are all main effect orthogonal, no design has full or partial aliasing between primary terms.

This trend in the performance of the main effects orthogonal designs is exemplified by the $f = 6$ lasso example where the best design exhibited partial aliasing between potential terms (type (ii)) and partial aliasing between primary and potential terms (type (iii)). There was a resolution IV design in the set. However this design exhibited full aliasing between potential terms and hence was not the best performing design (see Appendix D for the aliasing structures of the MEO designs).

Figure 4.3(a) shows the performance of the $f = 4$ main effects orthogonal designs when evaluated under the lasso objective function for values of $0 < \delta_0 \leq 40$. Figure 4.3(b) is the same plot enlarged for a region of interest. This plot shows that the

Figure 4.3: Evaluation of 16-run $f = 4$ main effects orthogonal designs for the lasso over a wider range of δ_0: (a) Shows the full plot, (b) Shows the plot enlarged for a region of interest. The objective function is (4.16)



performance of Design 2, the resolution IV design, which was initially the second worst performing, improves as δ_0 is increased, and becomes better performing than Designs 4 and 5 (which have partial aliasing between primary and potential terms) within the range of δ_0 plotted. It can also be seen that, within this range of δ_0, the performance of Design 1, the resolution III design which was initially the worst performing, increases to approximately the same performance as Design 5. If δ_0 was increased further, the performance of Design 1 would exceed that of Design 5. It should be noted that Design 3, the resolution V design, always performs best, no matter the value of δ_0. This provides more evidence for the choice of $\delta_0 \leq 1$, as an experimenter would need to be very certain that the potential terms were unimportant to be comfortable using the resolution IV or, particularly, resolution III designs.

4.4.3 Effect of Choice of Primary Terms on the Performance of Bayesian *D*-optimal Lasso Designs

The effect of setting different variables as the primary terms in model (1.1) was investigated for the $f = 4$ lasso example, where a pair of main effects and their interaction were set as the primary terms, with the remaining predictors set as the potential terms. The six possible combinations of pairs of factors and their interactions can be summarised as follows:

(i) $\boldsymbol{\nu} = (1, 1, 0, 0, 1, 0, 0, 0, 0, 0)$

(ii) $\boldsymbol{\nu} = (1, 0, 1, 0, 0, 1, 0, 0, 0, 0)$

(iii) $\boldsymbol{\nu} = (1, 0, 0, 1, 0, 0, 1, 0, 0, 0)$

(iv) $\boldsymbol{\nu} = (0, 1, 1, 0, 0, 0, 0, 1, 0, 0)$

(v) $\boldsymbol{\nu} = (0, 1, 0, 1, 0, 0, 0, 0, 1, 0)$

(vi) $\boldsymbol{\nu} = (0, 0, 1, 1, 0, 0, 0, 0, 0, 1)$

Here, $\nu_j = 1$ if the jth predictor is a primary term and $\nu_j = 0$ if the jth predictor is a potential term. The ordering of the predictors is lexicographical: $x_1, x_2, x_3, x_4, x_1x_2, x_1x_3, x_1x_4, x_2x_3, x_2x_4, x_3x_4$.

Bayesian D-optimal designs for the lasso were found for each of (i)-(vi) with, as before, $\lambda = 0.5$ and $\delta_0 = \{0, 0.001, 0.1, 0.25, 0.5, 0.75, 1\}$. For $N = 16$ runs, a full factorial design is available and, for each value of δ_0, this design was optimal for each of (i)-(vi).

Designs in $N = 8$ runs were generated using coordinate-exchange for each of (i)-(vi). For $\delta_0 = 0$, it was not possible to generate designs for $N = 8$ since there are essentially no potential terms in the assumed model. As every term is primary and there are fewer runs than parameters to be estimated, all designs are incapable of estimating the model. Hence, six designs were found for each of six values of δ_0.

Each generated design had its performance evaluated under each of (i)-(vi) using (4.16). As expected, the best performance of a design was for the choice of primary terms for which it was generated. Similarly, for each set of primary terms, the design with best performance was the one generated for that set of primary terms.

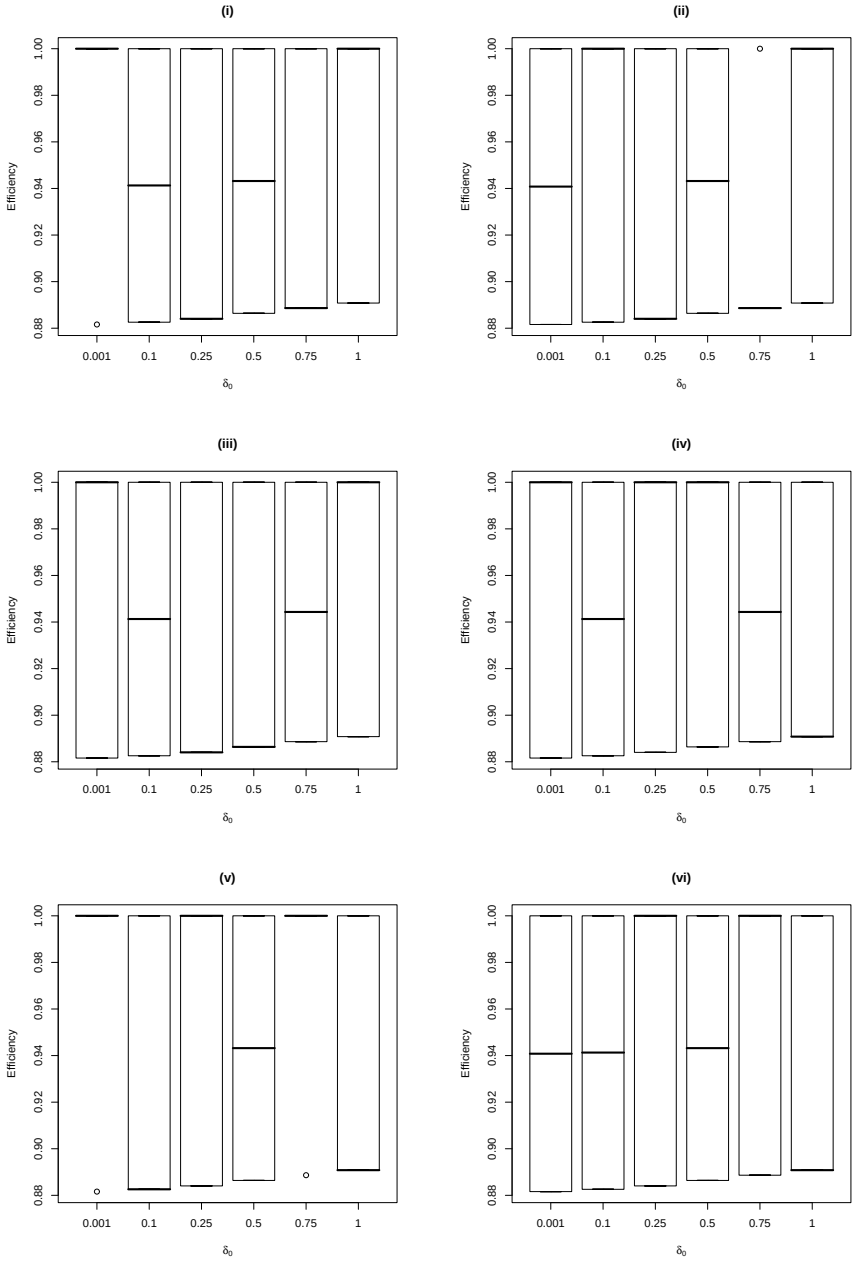
Figure 4.4 shows boxplots for each of (i)-(vi) of the efficiency of each design, ξ, found for each value of δ_0. For each value of δ_0 and set of primary terms, the efficiency of a design ξ, relative to the optimal design ξ^*, is

$$\exp\left\{\mathbf{\Phi}_L(\xi) - \mathbf{\Phi}_L(\xi^*)\right\}.$$

These boxplots have been constructed for the above designs as follows.

1. Select a set of primary terms from (i)−(vi)
2. Select a value of δ_0 from $0.001, 0.1, 0.25, 0.5, 0.75, 1$
3. Calculate the value of the objective function, using the selected primary terms, for each of the six Bayesian D-optimal designs found using the value of δ_0 selected in step 2
4. Select the design which has the largest value of the objective function
5. Calculate the efficiency of each design compared to the best performing design found in step 4
6. Construct a boxplot of the efficiency values

Figure 4.4: Efficiencies of D-optimal lasso designs evaluated for each set of primary terms (i)-(vi)



7. Repeat steps 2-6 for each remaining value of δ_0 and combine the boxplots into a single figure

8. Repeat steps 1-7 for each set of primary terms

For each value of δ_0 and set of primary terms, the efficiencies differ by only a small amount (Figure 4.4), and have a range of approximately 0.88 to 1. This finding indicates that the choice of primary terms does not make a substantial difference to the performance of an optimal design. This robustness is consistent over the values of δ_0. The range of efficiency for each value of δ_0 decreases slightly as δ_0 increases for each set of primary terms (i)-(vi). The smallest range of efficiency values for each set of primary terms is observed when $\delta_0 = 1$ which indicates that the designs are most robust at this value of δ_0. The greatest range of efficiency (0.12) is observed when $\delta_0 = 0.001$ for (i)-(vi) indicating that the designs are least robust at this value of δ_0; for primary terms in (i) and (v), this range in efficiency is caused by a single design which has a much lower efficiency than the rest. Under primary terms (i) and (v), the poorest designs are those that exhibit full aliasing between more than one of the primary terms and the potential terms; in each case, there is only one design with this property.

4.4.4 Further *D*-optimal Designs for Two-level Factors

Bayesian D-optimal designs for factors at two levels, labelled +1 and -1, with $N = 18$ or $N = 33$ runs, including interaction terms as well as main effect terms, were generated using the coordinate-exchange algorithm. We first consider the easier case for 33 runs and designs for ridge regression and the lasso.

Two $N = 33$ Bayesian D-optimal designs were generated for ridge regression using (4.13) with $\lambda = 0.5$: a design with $f = 4$ variables; a design with $f = 6$ variables. These designs are given in Tables 4.1 and 4.2 and are referred to as d_1 and d_2, respectively.

For the lasso, an $N = 33$ design with $f = 6$ variables was generated, using (4.14) with $\lambda = 0.5$ and $\delta_0 = 1$. The primary terms in model (1.1) were the main effects of four factors and their two-factor interactions defined in $\boldsymbol{\nu} = (1, 1, 1, 1, 0, 0, 1, 1, 1, 0, 0, 1, 1, 0, 0, 0, 1, 0, 0, 0, 0)$. The design is given in Table 4.3 and is called d_3. Note that the observations Y are simulated from a known model and will be used in the next chapter (see Section 5.3).

Table 4.4 gives summaries related to the balance and correlation properties for the primary (main effect) terms. The row labelled 'Column sum' is the sum of the levels of each factor and indicates the degree of balance in the design. When the

column sum is equal to +1, then that factor has one more +1 setting than −1 setting in the design. The opposite is true if the column sum is equal to −1. Each design is as close to balanced as possible for a design with an odd number of runs. A fully balanced design would have an equal number of +1 and −1 settings for each factor in the design.

For only 32 runs, and $f = 6$, a resolution VI design can be constructed, that is, having no aliasing between any pairs of main effects and interactions. The 33 run designs in Tables 4.2 and 4.3 are, in fact, the runs of a resolution VI design plus one repeated point, where the extra point is different for the two designs. The design found for $N = 33$, $f = 4$ for ridge regression, shown in Table 4.1, is two replicates of a 2^4 factorial design with the addition of the point (1 1 -1 1).

The row labelled 'Maximum correlation' contains the absolute value of the highest correlation between each column of $\boldsymbol{X}$ corresponding to a main effect and the columns corresponding to the remaining predictors. The three designs, $d_1 - d_3$, all have each correlation between predictors equal to 0.0294 and, in fact, every main effect and two-factor interaction has an equal correlation with all other main effects and two-factor interactions in the design. This small value for the correlation between the columns of the designs reflects the small amount of partial aliasing between predictors. These findings reinforce the result that designs which are close to regular fractional factorial designs of resolution V or higher will, where available, perform well under these models.

We now consider Bayesian D-optimal designs with $N = 18$ runs for $f = 4, 6$. The values of λ and δ_0, and the specification of primary terms, were as for the previous $N = 33$ designs. The $f = 4$ and $f = 6$ ridge regression designs are shown in Tables 4.5 and 4.6, and are referred to as d_4 and d_5, respectively. The $f = 6$ design for the lasso is shown in Table 4.7 and is referred to as d_6.

Table 4.8 summarises the balance and correlation for each of the three $N = 18$ designs found. Each design has at least one main effect whose column sum is either +2 or -2 indicating that the $N = 18$ designs are less balanced than the $N = 33$ designs, even though it would be possible for them to be fully balanced since they have an even number of runs. Each design is non-regular, with every predictor partially aliased with other predictors.

Table 4.1: A Bayesian D-optimal design (d_1) for ridge regression for $N = 33$ runs and $f = 4$ factors, together with the simulated observations. The bold combinations of factor levels are repeated points

Run	A	B	C	D	Y
1	1	-1	1	1	31.19
2	1	1	1	-1	-99.73
3	-1	1	-1	1	-104.48
4	1	-1	1	-1	-111.08
5	-1	-1	-1	1	-97.05
6	-1	-1	1	1	-19.98
7	-1	1	-1	1	-104.72
8	**1**	**1**	**-1**	**1**	**-74.32**
9	-1	1	1	-1	-111.13
10	-1	1	1	1	171.95
11	-1	-1	-1	-1	242.64
12	1	1	1	1	262.55
13	1	-1	1	-1	-111.05
14	1	-1	-1	1	-106.34
15	-1	-1	1	-1	-80.35
16	-1	-1	-1	1	-95.77
17	-1	1	1	1	172.10
18	-1	-1	-1	-1	240.46
19	1	1	1	-1	-100.44
20	1	1	-1	-1	-35.83
21	-1	1	-1	-1	13.10
22	1	-1	-1	-1	152.77
23	**1**	**1**	**-1**	**1**	**-74.15**
24	1	-1	-1	-1	152.53
25	-1	-1	1	1	-20.46
26	-1	1	1	-1	-111.07
27	1	-1	-1	1	-105.50
28	-1	1	-1	-1	13.56
29	1	1	-1	-1	-37.25
30	-1	-1	1	-1	-82.82
31	1	-1	1	1	32.71
32	**1**	**1**	**-1**	**1**	**-73.75**
33	1	1	1	1	262.62

Table 4.8 also gives the absolute value of the highest correlation that each main effect has with any other predictor in the design. It is obvious that the correlations between the predictors do not follow the same pattern that was observed for the

Table 4.2: A Bayesian D-optimal design (d_2) for ridge regression for $N = 33$ runs and $f = 6$ factors, together with the simulated observations. The bold combinations of factor levels are repeated points

Run	A	B	C	D	E	F	Y
1	-1	-1	-1	-1	1	1	241.82
2	-1	1	1	-1	1	1	-112.28
3	**-1**	**1**	**1**	**1**	**1**	**-1**	**169.93**
4	1	-1	-1	1	-1	-1	-106.88
5	-1	1	-1	1	1	1	-105.59
6	-1	-1	-1	1	1	-1	-97.77
7	**-1**	**1**	**1**	**1**	**1**	**-1**	**171.33**
8	1	1	1	1	-1	-1	261.14
9	1	-1	1	-1	1	1	-112.10
10	1	-1	-1	1	1	1	-105.89
11	1	1	-1	1	-1	1	-73.70
12	1	1	1	1	1	1	263.13
13	1	-1	1	-1	-1	-1	-109.76
14	-1	1	-1	-1	-1	1	12.31
15	-1	-1	1	-1	1	-1	-79.91
16	-1	1	-1	-1	1	-1	13.22
17	-1	-1	1	-1	-1	1	-83.56
18	1	-1	1	1	1	-1	31.80
19	1	1	1	-1	-1	1	-98.71
20	-1	-1	-1	-1	-1	-1	242.07
21	-1	-1	1	1	1	1	-20.06
22	1	-1	-1	-1	-1	1	153.07
23	1	-1	1	1	-1	1	28.42
24	-1	-1	-1	1	-1	1	-96.17
25	-1	-1	1	1	-1	-1	-20.78
26	1	1	1	-1	1	-1	-99.39
27	-1	1	1	-1	-1	-1	-110.76
28	1	1	-1	1	1	-1	-75.14
29	1	-1	-1	-1	1	-1	150.38
30	1	1	-1	-1	-1	-1	-36.33
31	-1	1	-1	1	-1	-1	-103.74
32	-1	1	1	1	-1	1	170.43
33	1	1	-1	-1	1	1	-36.36

33 run designs, in that the correlations are not equal. The maximum correlations are also all greater than those for the 33 run designs, indicating a higher degree of partial aliasing between the main effects and the other predictors. Note that for

Table 4.3: A Bayesian D-optimal design (d_3) for the lasso for $N = 33$ runs and $f = 6$ factors, together with the simulated observations. The bold combinations of factor levels are repeated points

Run	A	B	C	D	E	F	Y
1	1	1	1	1	1	-1	262.28
2	-1	-1	1	-1	-1	-1	-82.81
3	-1	-1	1	1	-1	1	-21.91
4	-1	-1	-1	1	-1	-1	-96.16
5	-1	1	1	-1	-1	1	-111.41
6	1	-1	-1	-1	1	1	152.39
7	-1	1	1	-1	1	-1	-109.35
8	-1	1	-1	-1	1	1	11.82
9	-1	1	1	1	-1	-1	171.72
10	1	-1	-1	1	-1	1	-105.68
11	-1	1	-1	-1	-1	-1	11.80
12	1	-1	1	1	-1	-1	31.11
13	-1	-1	-1	1	1	1	-96.29
14	-1	1	1	1	1	1	170.49
15	**-1**	**-1**	**1**	**1**	**1**	**-1**	**-20.68**
16	1	-1	1	-1	-1	1	-112.33
17	1	1	1	1	-1	1	262.97
18	1	-1	1	1	1	1	31.72
19	1	1	1	-1	-1	-1	-100.42
20	-1	-1	-1	-1	1	-1	241.80
21	**-1**	**-1**	**1**	**1**	**1**	**-1**	**-20.78**
22	-1	1	-1	1	-1	1	-104.52
23	1	-1	-1	1	1	-1	-104.99
24	1	1	-1	-1	1	-1	-35.33
25	-1	1	-1	1	1	-1	-105.87
26	1	1	-1	1	1	1	-74.04
27	1	1	-1	-1	-1	1	-34.03
28	1	1	-1	1	-1	-1	-74.59
29	-1	-1	1	-1	1	1	-83.66
30	-1	-1	-1	-1	-1	1	242.53
31	1	1	1	-1	1	1	-98.43
32	1	-1	1	-1	1	-1	-112.01
33	1	-1	-1	-1	-1	-1	151.53

the lasso design, the maximum correlations for main effect columns are much more unbalanced than for the ridge regression designs, with the two potential main effects (E and F) having much higher maximum correlation. Choosing two main effects as

Table 4.4: Column sum and maximum correlation for the $N = 33$ D-optimal designs

		Factor					
Design	Summary statistic	A	B	C	D	E	F
Ridge	Column sum	1	1	-1	1	-	-
$f = 4$	Maximum correlation	0.0294	0.0294	0.0294	0.0294	-	-
Ridge	Column sum	-1	1	1	1	1	-1
$f = 6$	Maximum correlation	0.0294	0.0294	0.0294	0.0294	0.0294	0.0294
Lasso	Column sum	-1	-1	1	1	1	-1
$f = 6$	Maximum correlation	0.0294	0.0294	0.0294	0.0294	0.0294	0.0294

potential terms when generating the lasso designs has a much greater effect on the balance and correlation structure of the 18 run designs than was found for the 33 run designs.

The designs presented in this section will be used to explore sequential design in Chapter 5.

Table 4.5: A Bayesian D-optimal design (d_4) for ridge regression for $N = 18$ runs and $f = 4$ factors, together with the simulated observations

Run	A	B	C	D	Y
1	-1	1	-1	-1	13.58
2	-1	-1	1	1	-20.39
3	1	-1	1	1	30.28
4	1	-1	-1	-1	152.33
5	-1	1	1	1	170.68
6	-1	1	-1	-1	12.62
7	-1	1	-1	1	-104.01
8	1	1	1	-1	-99.24
9	-1	-1	1	-1	-83.88
10	-1	-1	-1	1	-96.76
11	1	1	-1	-1	-36.53
12	1	1	1	-1	-97.88
13	-1	-1	-1	-1	243.12
14	1	1	1	1	261.42
15	1	1	-1	1	-74.08
16	1	-1	-1	1	-103.80
17	-1	1	1	-1	-109.51
18	1	-1	1	-1	-112.05

Table 4.6: A Bayesian D-optimal design (d_5) for ridge regression for $N = 18$ runs and $f = 6$ factors, together with the simulated observations

Run	A	B	C	D	E	F	Y
1	1	-1	1	-1	-1	-1	-110.82
2	1	-1	-1	-1	-1	1	152.20
3	-1	1	1	-1	-1	-1	-111.03
4	1	1	1	-1	-1	1	-97.92
5	-1	-1	1	1	-1	1	-19.51
6	-1	1	1	1	1	1	171.25
7	1	-1	-1	1	-1	-1	-105.54
8	-1	1	-1	-1	1	-1	13.12
9	-1	-1	-1	-1	-1	-1	242.61
10	1	1	1	1	-1	-1	260.87
11	-1	-1	1	1	1	-1	-18.92
12	1	1	1	-1	1	-1	-98.76
13	1	1	-1	1	-1	1	-74.70
14	-1	-1	-1	1	1	1	-95.55
15	-1	1	-1	1	-1	-1	-104.92
16	1	-1	1	1	1	1	31.76
17	-1	-1	1	-1	1	1	-81.81
18	1	1	-1	-1	1	1	-36.33

Table 4.7: A Bayesian D-optimal design (d_6) for the lasso for $N = 18$ runs and $f = 6$ factors, together with the simulated observations

Run	A	B	C	D	E	F	Y
1	-1	1	-1	1	-1	-1	-106.31
2	1	1	-1	1	1	1	-73.83
3	-1	1	1	1	1	1	171.47
4	-1	-1	1	1	-1	-1	-21.69
5	-1	1	1	-1	-1	-1	-111.60
6	1	-1	-1	-1	1	1	151.29
7	1	-1	-1	1	1	-1	-105.90
8	-1	-1	1	-1	1	1	-82.16
9	-1	1	-1	-1	1	1	12.46
10	1	-1	1	-1	1	-1	-112.57
11	1	-1	1	1	1	1	28.06
12	-1	-1	-1	1	1	1	-96.06
13	1	-1	-1	1	-1	1	-104.33
14	1	1	-1	-1	-1	-1	-35.14
15	-1	-1	-1	-1	-1	-1	242.63
16	1	1	1	-1	1	1	-101.62
17	1	-1	1	-1	-1	1	-110.91
18	1	1	1	1	-1	-1	261.50

Table 4.8: Column sum and maximum correlation for the $N = 18$ D-optimal designs

		Factor					
Design	Summary statistic	A	B	C	D	E	F
Ridge	Column sum	0	2	0	-2	-	-
$f = 4$	Maximum correlation	0.1111	0.1	0.1111	0.1	-	-
Ridge	Column sum	0	0	2	0	-2	0
$f = 6$	Maximum correlation	0.3333	0.3333	0.35	0.3333	0.35	0.3333
Lasso	Column sum	2	-2	0	0	2	2
$f = 6$	Maximum correlation	0.4781	0.4781	0.1111	0.1111	0.55	0.55

4.5 Designs for Restricted Factor Level Combinations: Application to the Melting Point Experiment

In this section, we find designs for ridge regression and the lasso for the melting point experiment, introduced in Section 1.2. As the factor levels cannot be freely combined, a row-exchange algorithm is used which is described below. We also investigate the impact of prior information on the number of distinct design points. We assume a model with only main effect predictors, see (2.15). In this section, all 21 variables are considered, including those from highly correlated pairs of variables, in order to study design selection in the presence of a high degree of multicollinearity.

4.5.1 Generation of Bayesian *D*-optimal Designs

Bayesian D-optimal designs for ridge regression and the lasso were generated for model (1.1) with predictors that are all the main effect terms. The row-exchange algorithm used for this work is a modified Federov exchange algorithm, which is described by Cook and Nachtsheim (1980). In this algorithm, the objective function is optimised by repeatedly exchanging a design point (i.e. an entire row of the design matrix) for a point selected from a candidate list of possible points. This algorithm is better suited to the chemistry example than the coordinate-exchange algorithm because the values of the variables for a particular compound cannot be independently chosen; the only choice is how many times, if at all, a compound is included in the design.

For this example, the candidate list consists of all 55 points in the data set. (Note that 55 compounds, not 60, were candidates because there were queries about descriptor values for 5 of the compounds at the time this work was carried out. These queries were later resolved.) Any of the candidate points may be included in the design more than once. The design selection via a row-exchange algorithm is as follows.

`Algorithm:`

1. `Generate` M `designs of` N `runs by randomly selecting the design points, with replacement, from a candidate list of` $N_{\mathcal{C}}$ `points`

2. Calculate the value of the objective function ((4.15) or (4.16)) for each of the M random designs

3. Select the random design having the largest value of the objective function as the starting design for the algorithm

4. Set $k = 1$ and $l = 1$

5. Exchange kth row of this design for the lth point in the candidate list

6. Calculate the value of the objective function for the new design

7. If the value of the objective function is increased, retain the exchange in the design

8. Set $l = l + 1$

9. If $l < N_{\mathcal{C}}$, repeat steps 5-8; otherwise continue

10. Set $k = k + 1$

11. If $k < N$, set $l = 1$ and repeat steps 5-10; otherwise continue

12. If one or more of the exchanges have been retained go to step 4; otherwise end the algorithm

4.5.2 Support Points and Prior Information for Bayesian *D*-optimal Designs

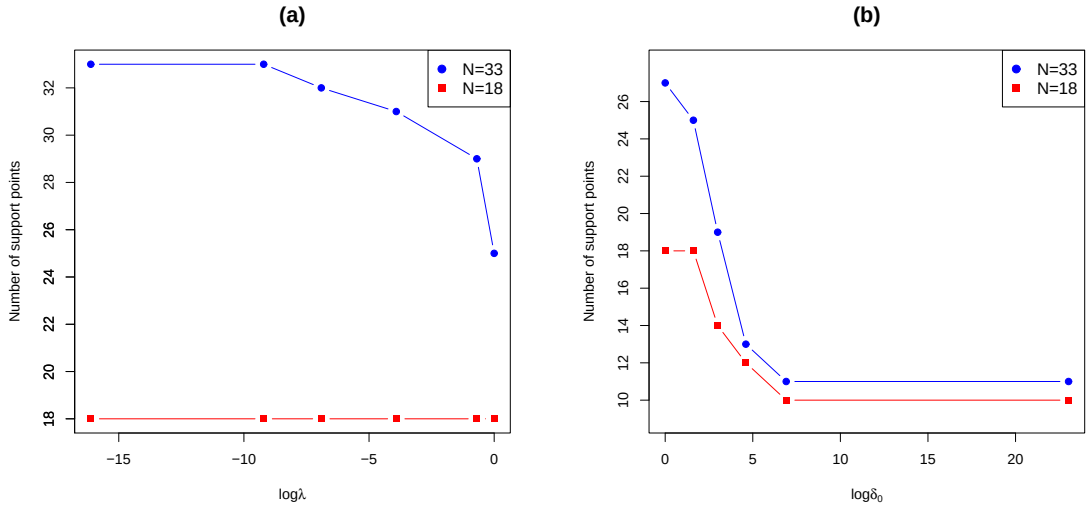
We investigate how the number of distinct or support points in a Bayesian D-optimal design is affected by the choice of values of the parameters λ and δ_0 assumed when selecting designs for 33 runs and 18 runs under ridge regression (4.13) and the lasso (4.14). For the lasso designs, the primary terms are the main effects of the factors B, F, G, H, I, M, i.e. the terms with non-zero coefficients in the model described in Section 2.3 (see Table 3.1 for the factor names). All other terms (the remaining main effects) are potential.

Altering the values of the parameters λ and δ_0 changes the degree of prior information for ridge regression and lasso designs respectively, with prior information increasing (prior variance decreasing) as these parameters are increased, and this will have an effect on the number of support points. This is because the prior variance is lower, and less information is needed to be gained from the design to estimate $\boldsymbol{\beta}$ accurately. For ridge regression, a high value of λ corresponds to model fitting with a high degree of shrinkage and low effective degrees of freedom; therefore there are essentially fewer terms in the model to learn about. For the lasso, a high value of δ_0 (and fixed λ) means the effective degrees of freedom are low, since estimating the potential terms is less important.

For ridge regression, designs for $\lambda = \{0, 10^{-7}, 0.0001, 0.001, 0.002, 0.5, 1\}$ are found ($\lambda = 0$ was not used when generating $N = 18$ designs). Figure 4.5(a) shows the number of support points as a function of $\log \lambda$ for the ridge regression designs. As λ is increased, the number of support points in the $N = 33$ designs decreases, as expected. The $N = 18$ ridge regression designs do not follow this trend. Each of these designs has 18 support points, the maximum possible number, irrespective of the value of λ. As equal prior information is assumed for all the parameters and there are fewer observations in the designs than there are parameters to be estimated, it follows that the maximum possible number of distinct design points will be selected.

For the lasso, λ is fixed at 1.0 and designs are found for $\delta_0 = \{0, 1, 5, 20, 100, 1000, 10^{10}\}$ ($\delta_0 = 0$ was not used when generating $N = 18$ designs). Figure 4.5(b) shows the number of support points as a function $\log \delta_0$ for the lasso designs. As δ_0 is increased, the number of support points in both the $N = 33$ and $N = 18$ designs decreases. When generating the lasso designs, a different prior distribution is assumed for the primary terms to that assumed for the potential terms. In this example there are 7 primary terms; fewer than there are observations in the designs. Therefore the number of support points in the lasso $N = 18$ designs decreases as δ_0 is increased; the same trend is observed for the $N = 33$ designs.

Figure 4.5: Effect of prior information on the number of support points in Bayesian D-optimal designs: (a) Ridge regression; (b) Lasso



4.5.3 Investigation of the Properties of Bayesian D-optimal Designs for the Melting Point Experiment

We now consider four of the designs from the previous section: $N = 18$ and 33 for ridge regression with $\lambda = 1$; $N = 18$ and 33 for the lasso with $\lambda = 1$, $\delta_0 = 5$. As in Section 4.4.4, we investigate correlations between the columns of the design that hold the values of the primary terms. We also compare models fitted by ridge regression and the lasso to simulated observations for the full candidate set of compounds and for each of the four designs.

Bayesian D-optimal designs for $N = 33$ runs and $N = 18$ runs have been generated for ridge regression and the lasso. The values $\lambda = 1$ and $\delta_0 = 5$ were used to generate the lasso designs because we found for $N = 33$ that these values generated the design that had the lowest maximum prediction variance, which was calculated using a bootstrapping procedure (see Section 5.2). A value of $\lambda = 1$ was also used to generate the ridge regression designs to enable a fair comparison of their performance with the lasso designs.

The designs for ridge regression are given in Tables 4.12 and 4.13, and for the lasso in Tables 4.14 and 4.15. For each type, there are 25 support points for the

$N = 33$ designs and 18 support points for the $N = 18$ designs. The two larger designs both have 8 design points that are repeated twice (see Tables 4.12 and 4.14), with six of these points occuring in both designs. From considering all four designs ($N = 33$ and $N = 18$), we see that approximately half of the points in the candidate set occur in at least one of the four designs, 25 of the points do not appear in any of the designs, and four points appear in only one design. This means that 47% of the candidate points appear in two or more of the designs. Other than this there is little pattern as to which points the designs have in common.

Data was simulated from the 'true' model defined in Section 2.3 for the four designs. A model was fitted to each data set using ridge regression and the lasso. Values of the within-sample R^2 and adjusted R^2, and out-of-sample R^2 and adjusted R^2 (from N-fold cross-validation) were calculated and are shown in Table 4.9. The same statistics were calculated for models fitted by ridge regression and the lasso to the observations simulated for each candidate design point. The values of these statistics for each design are comparable to those for the full candidate list. Ridge regression does not shrink any of the coefficients to exactly zero, therefore for the $N = 18$ designs the models are overfitted to the data and as a result the out-of-sample R^2 and adjusted R^2 are lower than for the larger designs.

Table 4.9: Summary statistics for models fitted to D-optimal designs and a design formed from the candidate list

		Design				
Model	Summary statistic	Full candidate list, $N = 55$	Ridge, $N = 33$	Ridge, $N = 18$	Lasso, $N = 33$	Lasso, $N = 18$
Ridge model	R^2	0.999	0.999	0.965	0.999	0.980
	Adjusted R^2	0.999	0.999	0.940	0.999	0.965
	C-V R^2	0.998	0.998	0.667	0.996	0.842
	C-V adjusted R^2	0.998	0.994	0.430	0.988	0.722
Lasso model	R^2	0.999	0.999	0.950	0.999	0.962
	Adjusted R^2	0.999	0.998	0.929	0.999	0.947
	C-V R^2	0.999	0.993	0.874	0.996	0.913
	C-V adjusted R^2	0.998	0.983	0.847	0.989	0.886

Table 4.10 lists, for each design, the maximum of the absolute values of the correlations between each pair of predictors (factor main effects) in the design. Maximum column correlations are only given for factors B, F, G, H, I and M, which are the factors set as primary terms for finding the lasso designs. These maximum correlations are also given between columns of the candidate list, which correspond to the same factors.

Table 4.10: Maximum column correlations for D-optimal designs and a design formed from the candidate list for the melting point experiment

		Design			
Factor	Candidate list, $N = 55$	Ridge, Bayesian, $N = 33$	Ridge, Bayesian, $N = 18$	Lasso, Bayesian, $N = 33$	Lasso, Bayesian, $N = 18$
B	0.924	0.940	0.926	0.923	0.924
F	0.710	0.772	0.770	0.765	0.752
G	0.640	0.586	0.595	0.606	0.576
H	0.997	0.996	0.996	0.997	0.997
I	0.998	0.997	0.997	0.997	0.997
M	0.961	0.972	0.978	0.975	0.964

As designs for ridge regression and the lasso emphasise the estimation of different sets of effects, we might expect the correlation structure to differ for designs found under the two models. For ridge regression designs, all predictors are treated equally; for lasso designs, the primary terms (main effects of B, F, G, H, I, M) are more important than the other predictors. Hence, for this latter case, we might think that there will be lower maximum column correlations for the primary terms. In fact, all designs, both for ridge regression and the lasso, have the same maximum column correlations (to a single decimal place), determined by the correlations present in the full candidate list.

The Bayesian D-optimal designs generated for the organic chemistry example were assessed under (4.15) and (4.16), the ridge regression and lasso D-optimality objective functions. When calculating the values of these objective functions, the terms set as primary in the assumed model and the values of λ and δ_0 were the same as those used to generate the designs. The values of both objective functions for all four of these designs are given in Table 4.11. Keeping in mind that the designs have been generated with the aim of maximising the value of the objective function, it

can be seen that when two designs with equal run sizes are evaluated under a certain criterion the design generated using that criterion performs better. For every design, the value of the lasso objective function (4.16) is consistently larger than that of the ridge regression objective function (4.15). This is due to the relative magnitudes of diagonal entries of the matrices $\lambda \boldsymbol{I}$, of the ridge regression objective function, and $2\lambda\Delta_0(\boldsymbol{\beta})$, of the lasso objective function. Under the values of λ and δ_0 used to generate the designs, the non-zero diagonal entries of $2\lambda\Delta_0(\boldsymbol{\beta})$ are greater than the diagonal entries of $\lambda \boldsymbol{I}$. Therefore, for the same design, the value of the lasso objective function will be larger than the value of the ridge regression objective function. Essentially, the lasso objective function is focussed on the estimation of fewer parameters.

The Bayesian D-optimal designs will be used to explore sequential design point selection in the next chapter.

Table 4.11: Values of the objective functions for the four Bayesian D-optimal designs

Design	Ridge regression objective function	Lasso objective function
Ridge, $N = 33$	5.03	8.93
Lasso, $N = 33$	4.94	9.06
Ridge, $N = 18$	3.52	6.79
Lasso, $N = 18$	3.34	6.89

4.6 Conclusions

In this chapter, the connection between Bayesian inference and bridge regression has been described, including the form of the posterior distribution and a new approximation to the prior distribution of the coefficients for the lasso. This approximation has enabled designs to be found for the lasso using Bayesian D-optimality. The design method was demonstrated for designs with two-level factors (Section 4.4) and designs where levels of factors cannot be freely combined (Section 4.5) using the organic chemistry example. In the first case, a comparison with a catalogue of main effect orthogonal designs for 4, 5 and 6 variables showed that

Table 4.12: Design and simulated observations for $N = 33$ and ridge regression

Run	Cand. point	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	Y
1	1	-0.5	-0.8	-0.1	-0.2	-0.3	-0.5	-1.0	-0.8	-0.8	1.0	0.6	0.4	-0.6	-0.7	-0.8	-0.9	0.1	-0.7	-0.5	-0.1	0.0	-78.9
2	5	-0.8	-0.4	0.0	0.1	-0.2	0.4	-1.0	-0.1	-0.2	1.0	0.8	0.1	-0.6	-0.5	-0.1	-0.4	0.5	-0.9	-0.8	-0.6	-0.6	-58.4
3	9	-0.5	-0.5	0.5	0.6	-0.2	-0.7	0.7	-0.1	-0.2	-0.8	-0.8	0.3	-0.8	-0.9	-0.1	-0.7	0.9	-0.2	-0.4	1.0	0.8	-23.2
4	11	-0.3	0.2	0.2	0.6	0.5	-0.1	0.7	0.4	0.3	-0.8	-0.8	0.5	-0.8	-0.8	0.4	-0.2	-0.2	-0.1	0.0	0.8	0.8	2.2
5	12	0.5	-0.5	-0.2	-0.2	-0.3	-0.7	0.7	-0.1	-0.2	-0.8	-0.8	0.2	0.1	0.3	-0.1	-0.7	0.7	-0.1	-0.1	0.8	1.0	-20.8
6	12	0.5	-0.5	-0.2	-0.2	-0.3	-0.7	0.7	-0.1	-0.2	-0.8	-0.8	0.2	0.1	0.3	-0.1	-0.7	0.7	-0.1	-0.1	0.8	1.0	-22.2
7	14	0.3	0.4	-0.3	0.0	0.9	0.3	-1.0	0.0	-0.1	-0.8	-0.8	-0.1	-0.5	-0.4	-0.1	-0.5	0.9	-0.7	0.2	0.1	-0.2	-31.6
8	16	-0.5	0.6	-0.2	0.2	0.9	0.8	-1.0	0.3	0.3	1.0	0.9	-0.1	-0.6	-0.5	0.2	-0.8	0.6	-0.9	-1.0	-1.0	-1.0	-22.9
9	17	0.0	-0.7	0.1	0.0	-1.0	-0.3	-0.7	0.5	0.6	-1.0	-1.0	-0.7	-0.7	-0.7	0.5	-0.4	0.9	0.2	0.3	-0.3	-0.8	-67.8
10	20	1.0	-0.5	0.1	0.1	-0.8	-0.9	0.2	0.5	0.5	-0.8	-0.7	-0.5	-0.5	-0.7	0.5	-0.5	0.4	-0.3	0.1	-0.3	-0.2	-42.2
11	21	-0.3	-0.8	-0.7	-0.8	-0.7	-0.7	-0.7	-0.2	-0.1	-0.8	-0.8	0.3	-0.4	-0.4	-0.2	-0.4	0.9	-0.3	0.4	0.1	-0.6	-73.7
12	22	0.0	-0.5	-0.6	-0.6	-0.5	-0.7	-0.3	0.8	0.9	-0.8	-0.8	0.8	-0.9	-0.8	0.9	0.0	0.7	0.6	-0.4	0.8	0.7	-67.4
13	22	0.0	-0.5	-0.6	-0.6	-0.5	-0.7	-0.3	0.8	0.9	-0.8	-0.8	0.8	-0.9	-0.8	0.9	0.0	0.7	0.6	-0.4	0.8	0.7	-69.2
14	23	0.0	-0.9	0.7	0.4	-0.7	-1.0	-0.1	-0.5	-0.5	-0.8	-0.8	-0.2	0.6	0.4	-0.5	-0.5	0.6	-0.7	0.5	-0.3	-0.6	-43.6
15	26	-0.5	0.5	0.1	0.5	0.7	0.0	-0.1	0.4	0.4	0.3	0.2	-0.2	0.1	0.1	0.4	-1.0	0.0	0.4	1.0	-0.6	-0.4	0.2
16	26	-0.5	0.5	0.1	0.5	0.7	0.0	-0.1	0.4	0.4	0.3	0.2	-0.2	0.1	0.1	0.4	-1.0	0.0	0.4	1.0	-0.6	-0.4	-1.5
17	27	-0.5	-0.7	0.2	0.2	-0.4	-0.8	0.7	-0.3	-0.4	-0.8	-0.8	0.3	0.8	0.8	-0.3	1.0	0.6	0.1	0.4	0.3	0.0	-20.8
18	28	-0.3	-0.5	1.0	1.0	-0.5	-0.4	0.7	0.4	0.3	-0.8	-0.8	0.3	1.0	1.0	0.4	1.0	0.6	0.0	0.0	-0.8	-0.6	-14.5
19	28	-0.3	-0.5	1.0	1.0	-0.5	-0.4	0.7	0.4	0.3	-0.8	-0.8	0.3	1.0	1.0	0.4	1.0	0.6	0.0	0.0	-0.8	-0.6	-13.5
20	29	-0.8	0.7	0.2	0.8	1.0	0.2	0.7	0.7	0.6	-0.8	-0.8	0.4	0.3	0.6	0.6	-0.5	0.1	-0.2	0.0	-0.1	0.0	25.32
21	30	-0.5	-0.7	0.4	0.3	-0.4	-0.8	0.7	-0.3	-0.4	-0.8	-0.8	0.1	-0.3	-0.4	-0.3	-0.1	0.3	1.0	-0.2	-0.3	-0.2	-24.7
22	31	0.3	-0.5	0.3	0.3	-0.6	-0.4	0.7	0.4	0.3	1.0	0.9	-0.2	-0.1	-0.2	0.4	-0.5	-1.0	0.6	-0.5	-0.3	-0.5	-18.4
23	31	0.3	-0.5	0.3	0.3	-0.6	-0.4	0.7	0.4	0.3	1.0	0.9	-0.2	-0.1	-0.2	0.4	-0.5	-1.0	0.6	-0.5	-0.3	-0.5	-15.7
24	32	-0.5	-0.5	0.7	0.7	-0.7	-0.9	0.2	0.5	0.5	-0.8	-0.7	-0.2	0.8	0.6	0.5	-0.1	0.7	0.2	-0.4	0.6	0.8	-37.7
25	33	-0.5	1.0	-0.5	-0.1	0.7	1.0	-1.0	1.0	1.0	-0.8	-0.7	-1.0	-0.2	-0.1	1.0	-0.3	-0.2	-0.3	0.5	0.1	-0.2	-16.6
26	33	-0.5	1.0	-0.5	-0.1	0.7	1.0	-1.0	1.0	1.0	-0.8	-0.7	-1.0	-0.2	-0.1	1.0	-0.3	-0.2	-0.3	0.5	0.1	-0.2	-14.7
27	36	-0.5	0.6	-0.9	-0.7	0.5	0.2	-0.7	0.8	0.8	1.0	1.0	-0.4	-0.1	-0.2	0.8	-0.8	0.3	-0.9	-0.9	-0.1	-0.4	-26.8
28	36	-0.5	0.6	-0.9	-0.7	0.5	0.2	-0.7	0.8	0.8	1.0	1.0	-0.4	-0.1	-0.2	0.8	-0.8	0.3	-0.9	-0.9	-0.1	-0.4	-28.1
29	37	-0.3	-0.4	-0.7	-0.7	-0.3	0.0	-1.0	0.0	0.1	-0.3	-0.3	-0.2	0.3	0.2	0.1	0.1	-0.4	0.4	-0.5	-0.3	-0.1	-62.1
30	43	-0.5	-0.8	-0.1	-0.2	-0.6	-0.2	-1.0	-0.7	-0.7	-0.8	-0.8	-0.5	-0.6	-0.6	-0.7	-0.4	0.8	0.3	-0.7	0.3	0.1	-68.3
31	53	-0.8	-0.8	0.3	0.1	-0.8	0.0	-1.0	0.1	0.2	-1.0	-1.0	0.4	-0.2	-0.3	0.2	-0.8	0.5	-0.7	-0.2	0.1	0.4	-66.5
32	53	-0.8	-0.8	0.3	0.1	-0.8	0.0	-1.0	0.1	0.2	-1.0	-1.0	0.4	-0.2	-0.3	0.2	-0.8	0.5	-0.7	-0.2	0.1	0.4	-67.0
33	54	-0.8	-0.7	0.4	0.3	-0.6	-0.6	-0.1	0.1	0.2	-0.8	-0.8	0.6	0.8	0.6	0.2	-0.5	0.6	0.7	0.6	1.0	0.2	-38.7

Table 4.13: Design and simulated observations for $N = 18$ and ridge regression

Run	Cand. point	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	Y
1	1	-0.5	-0.8	-0.1	-0.2	-0.3	-0.5	-1.0	-0.8	-0.8	1.0	0.6	0.4	-0.6	-0.7	-0.8	-0.9	0.1	-0.7	-0.5	-0.1	0.0	-79.9
2	11	-0.3	0.2	0.2	0.6	0.5	-0.1	0.7	0.4	0.3	-0.8	-0.8	0.5	-0.8	-0.8	0.4	-0.2	-0.2	-0.1	0.0	0.8	0.8	0.7
3	12	0.5	-0.5	-0.2	-0.2	-0.3	-0.7	0.7	-0.1	-0.2	-0.8	-0.8	0.2	0.1	0.3	-0.1	-0.7	0.7	-0.1	-0.1	0.8	1.0	-21.1
4	16	-0.5	0.6	-0.2	0.2	0.9	0.8	-1.0	0.3	0.3	1.0	0.9	-0.1	-0.6	-0.5	0.2	-0.8	0.6	-0.9	-1.0	-1.0	-1.0	-22.0
5	17	0.0	-0.7	0.1	0.0	-1.0	-0.3	-0.7	0.5	0.6	-1.0	-1.0	-0.7	-0.7	-0.7	0.5	-0.4	0.9	0.2	0.3	-0.3	-0.8	-68.0
6	20	1.0	-0.5	0.1	0.1	-0.8	-0.9	0.2	0.5	0.5	-0.8	-0.7	-0.5	-0.5	-0.7	0.5	-0.5	0.4	-0.3	0.1	-0.3	-0.2	-40.0
7	22	0.0	-0.5	-0.6	-0.6	-0.5	-0.7	-0.3	0.8	0.9	-0.8	-0.8	0.8	-0.9	-0.8	0.9	0.0	0.7	0.6	-0.4	0.8	0.7	-68.1
8	26	-0.5	0.5	0.1	0.5	0.7	0.0	-0.1	0.4	0.4	0.3	0.2	-0.2	0.1	0.1	0.4	-1.0	0.0	0.4	1.0	-0.6	-0.4	0.3
9	27	-0.5	-0.7	0.2	0.2	-0.4	-0.8	0.7	-0.3	-0.4	-0.8	-0.8	0.3	0.8	0.8	-0.3	1.0	0.6	0.1	0.4	0.3	0.0	-20.5
10	28	-0.3	-0.5	1.0	1.0	-0.5	-0.4	0.7	0.4	0.3	-0.8	-0.8	0.3	1.0	1.0	0.4	1.0	0.6	0.0	0.0	-0.8	-0.6	-15.4
11	31	0.3	-0.5	0.3	0.3	-0.6	-0.4	0.7	0.4	0.3	1.0	0.9	-0.2	-0.1	-0.2	0.4	-0.5	-1.0	0.6	-0.5	-0.3	-0.5	-16.2
12	32	-0.5	-0.5	0.7	0.7	-0.7	-0.9	0.2	0.5	0.5	-0.8	-0.7	-0.2	0.8	0.6	0.5	-0.1	0.7	0.2	-0.4	0.6	0.8	-39.2
13	33	-0.5	1.0	-0.5	-0.1	0.7	1.0	-1.0	1.0	1.0	-0.8	-0.7	-1.0	-0.2	-0.1	1.0	-0.3	-0.2	-0.3	0.5	0.1	-0.2	-15.7
14	36	-0.5	0.6	-0.9	-0.7	0.5	0.2	-0.7	0.8	0.8	1.0	1.0	-0.4	-0.1	-0.2	0.8	-0.8	0.3	-0.9	-0.9	-0.1	-0.4	-26.4
15	37	-0.3	-0.4	-0.7	-0.7	-0.3	0.0	-1.0	0.0	0.1	-0.3	-0.3	-0.2	0.3	0.2	0.1	0.1	-0.4	0.4	-0.5	-0.3	-0.1	-59.8
16	43	-0.5	-0.8	-0.1	-0.2	-0.6	-0.2	-1.0	-0.7	-0.7	-0.8	-0.8	-0.5	-0.6	-0.6	-0.7	-0.4	0.8	0.3	-0.7	0.3	0.1	-69.5
17	53	-0.8	-0.8	0.3	0.1	-0.8	0.0	-1.0	0.1	0.2	-1.0	-1.0	0.4	-0.2	-0.3	0.2	-0.8	0.5	-0.7	-0.2	0.1	0.4	-67.7
18	54	-0.8	-0.7	0.4	0.3	-0.6	-0.6	-0.1	0.1	0.2	-0.8	-0.8	0.6	0.8	0.6	0.2	-0.5	0.6	0.7	0.6	1.0	0.2	-39.0

Table 4.14: Design and simulated observations for $N = 33$ and the lasso

Run	Cand. point	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	Y
1	1	-0.5	-0.8	-0.1	-0.2	-0.3	-0.5	-1.0	-0.8	-0.8	1.0	0.6	0.4	-0.6	-0.7	-0.8	-0.9	0.1	-0.7	-0.5	-0.1	0.0	-80.2
2	1	-0.5	-0.8	-0.1	-0.2	-0.3	-0.5	-1.0	-0.8	-0.8	1.0	0.6	0.4	-0.6	-0.7	-0.8	-0.9	0.1	-0.7	-0.5	-0.1	0.0	-78.3
3	9	-0.5	-0.5	0.5	0.6	-0.2	-0.7	0.7	-0.1	-0.2	-0.8	-0.8	0.3	-0.8	-0.9	-0.1	-0.7	0.9	-0.2	-0.4	1.0	0.8	-23.4
4	10	0.0	-0.3	0.4	0.5	-0.1	-0.2	0.7	0.3	0.1	-0.8	-0.8	0.6	-0.9	-0.7	0.2	-0.2	-0.3	0.1	-0.1	0.8	0.6	-11.6
5	12	0.5	-0.5	-0.2	-0.2	-0.3	-0.7	0.7	-0.1	-0.2	-0.8	-0.8	0.2	0.1	0.3	-0.1	-0.7	0.7	-0.1	-0.1	0.8	1.0	-20.0
6	16	-0.5	0.6	-0.2	0.2	0.9	0.8	-1.0	0.3	0.3	1.0	0.9	-0.1	-0.6	-0.5	0.2	-0.8	0.6	-0.9	-1.0	-1.0	-1.0	-24.1
7	17	0.0	-0.7	0.1	0.0	-1.0	-0.3	-0.7	0.5	0.6	-1.0	-1.0	-0.7	-0.7	-0.7	0.5	-0.4	0.9	0.2	0.3	-0.3	-0.8	-65.5
8	20	1.0	-0.5	0.1	0.1	-0.8	-0.9	0.2	0.5	0.5	-0.8	-0.7	-0.5	-0.5	-0.7	0.5	-0.5	0.4	-0.3	0.1	-0.3	-0.2	-42.2
9	22	0.0	-0.5	-0.6	-0.6	-0.5	-0.7	-0.3	0.8	0.9	-0.8	-0.8	0.8	-0.9	-0.8	0.9	0.0	0.7	0.6	-0.4	0.8	0.7	-67.9
10	22	0.0	-0.5	-0.6	-0.6	-0.5	-0.7	-0.3	0.8	0.9	-0.8	-0.8	0.8	-0.9	-0.8	0.9	0.0	0.7	0.6	-0.4	0.8	0.7	-67.0
11	23	0.0	-0.9	0.7	0.4	-0.7	-1.0	-0.1	-0.5	-0.5	-0.8	-0.8	-0.2	0.6	0.4	-0.5	-0.5	0.6	-0.7	0.5	-0.3	-0.6	-44.5
12	26	-0.5	0.5	0.1	0.5	0.7	0.0	-0.1	0.4	0.4	0.3	0.2	-0.2	0.1	0.1	0.4	-1.0	0.0	0.4	1.0	-0.6	-0.4	-0.1
13	26	-0.5	0.5	0.1	0.5	0.7	0.0	-0.1	0.4	0.4	0.3	0.2	-0.2	0.1	0.1	0.4	-1.0	0.0	0.4	1.0	-0.6	-0.4	-0.9
14	27	-0.5	-0.7	0.2	0.2	-0.4	-0.8	0.7	-0.3	-0.4	-0.8	-0.8	0.3	0.8	0.8	-0.3	1.0	0.6	0.1	0.4	0.3	0.0	-20.6
15	28	-0.3	-0.5	1.0	1.0	-0.5	-0.4	0.7	0.4	0.3	-0.8	-0.8	0.3	1.0	1.0	0.4	1.0	0.6	0.0	0.0	-0.8	-0.6	-13.9
16	28	-0.3	-0.5	1.0	1.0	-0.5	-0.4	0.7	0.4	0.3	-0.8	-0.8	0.3	1.0	1.0	0.4	1.0	0.6	0.0	0.0	-0.8	-0.6	-12.4
17	29	-0.8	0.7	0.2	0.8	1.0	0.2	0.7	0.7	0.6	-0.8	-0.8	0.4	0.3	0.6	0.6	-0.5	0.1	-0.2	0.0	-0.1	0.0	22.8
18	31	0.3	-0.5	0.3	0.3	-0.6	-0.4	0.7	0.4	0.3	1.0	0.9	-0.2	-0.1	-0.2	0.4	-0.5	-1.0	0.6	-0.5	-0.3	-0.5	-16.7
19	31	0.3	-0.5	0.3	0.3	-0.6	-0.4	0.7	0.4	0.3	1.0	0.9	-0.2	-0.1	-0.2	0.4	-0.5	-1.0	0.6	-0.5	-0.3	-0.5	-15.1
20	32	-0.5	-0.5	0.7	0.7	-0.7	-0.9	0.2	0.5	0.5	-0.8	-0.7	-0.2	0.8	0.6	0.5	-0.1	0.7	0.2	-0.4	0.6	0.8	-67.9
21	32	-0.5	-0.5	0.7	0.7	-0.7	-0.9	0.2	0.5	0.5	-0.8	-0.7	-0.2	0.8	0.6	0.5	-0.1	0.7	0.2	-0.4	0.6	0.8	-38.9
22	33	-0.5	1.0	-0.5	-0.1	0.7	1.0	-1.0	1.0	1.0	-0.8	-0.7	-1.0	-0.2	-0.1	1.0	-0.3	-0.2	-0.3	0.5	0.1	-0.2	-14.7
23	33	-0.5	1.0	-0.5	-0.1	0.7	1.0	-1.0	1.0	1.0	-0.8	-0.7	-1.0	-0.2	-0.1	1.0	-0.3	-0.2	-0.3	0.5	0.1	-0.2	-15.0
24	36	-0.5	0.6	-0.9	-0.7	0.5	0.2	-0.7	0.8	0.8	1.0	1.0	-0.4	-0.1	-0.2	0.8	-0.8	0.3	-0.9	-0.9	-0.1	-0.4	-24.9
25	37	-0.3	-0.4	-0.7	-0.7	-0.3	0.0	-1.0	0.0	0.1	-0.3	-0.3	-0.2	0.3	0.2	0.1	0.1	-0.4	0.4	-0.5	-0.3	-0.1	-61.0
26	39	-0.3	-0.6	-0.1	-0.1	-0.6	0.2	-1.0	0.0	0.0	0.0	1.0	0.8	-0.3	0.1	-0.2	0.0	-0.8	0.5	-0.8	-0.8	-0.1	-0.2	-61.7
27	41	0.0	-1.0	-0.4	-0.5	-0.5	-0.6	-1.0	-1.0	-1.0	-0.3	-0.4	0.3	-0.1	-0.2	-1.0	-0.5	1.0	-0.6	-0.3	-0.3	-0.4	-77.2
28	43	-0.5	-0.8	-0.1	-0.2	-0.6	-0.2	-1.0	-0.7	-0.7	-0.8	-0.8	-0.5	-0.6	-0.6	-0.7	-0.4	0.8	0.3	-0.7	0.3	0.1	-68.8
29	45	0.3	-0.7	0.1	0.0	-0.3	-0.9	-0.1	-0.3	-0.3	-0.3	-0.4	0.8	-0.9	-1.0	-0.3	-0.4	0.5	0.6	0.0	0.1	0.3	-49.5
30	50	-1.0	0.5	0.0	0.4	0.5	0.8	-1.0	0.8	0.8	-1.0	-1.0	0.1	-0.9	-0.9	0.8	-0.3	0.4	-0.3	0.1	-0.1	-0.5	-36.2
31	53	-0.8	-0.8	0.3	0.1	-0.8	0.0	-1.0	0.1	0.2	-1.0	-1.0	0.4	-0.2	-0.3	0.2	-0.8	0.5	-0.7	-0.2	0.1	0.4	-67.9
32	53	-0.8	-0.8	0.3	0.1	-0.8	0.0	-1.0	0.1	0.2	-1.0	-1.0	0.4	-0.2	-0.3	0.2	-0.8	0.5	-0.7	-0.2	0.1	0.4	-67.2
33	54	-0.8	-0.7	0.4	0.3	-0.6	-0.6	-0.1	0.1	0.2	-0.8	-0.8	0.6	0.8	0.6	0.2	-0.5	0.6	0.7	0.6	1.0	0.2	-37.6

Table 4.15: Design and simulated observations for $N = 18$ and the lasso

Run	Cand. point	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	Y
1	1	-0.5	-0.8	-0.1	-0.2	-0.3	-0.5	-1.0	-0.8	-0.8	1.0	0.6	0.4	-0.6	-0.7	-0.8	-0.9	0.1	-0.7	-0.5	-0.1	0.0	-79.5
2	10	0.0	-0.3	0.4	0.5	-0.1	-0.2	0.7	0.3	0.1	-0.8	-0.8	0.6	-0.9	-0.7	0.2	-0.2	-0.3	0.1	-0.1	0.8	0.6	-14.4
3	12	0.5	-0.5	-0.2	-0.2	-0.3	-0.7	0.7	-0.1	-0.2	-0.8	-0.8	0.2	0.1	0.3	-0.1	-0.7	0.7	-0.1	-0.1	0.8	1.0	-20.2
4	20	1.0	-0.5	0.1	0.1	-0.8	-0.9	0.2	0.5	0.5	-0.8	-0.7	-0.5	-0.5	-0.7	0.5	-0.5	0.4	-0.3	0.1	-0.3	-0.2	-41.9
5	22	0.0	-0.5	-0.6	-0.6	-0.5	-0.7	-0.3	0.8	0.9	-0.8	-0.8	0.8	-0.9	-0.8	0.9	0.0	0.7	0.6	-0.4	0.8	0.7	-67.8
6	26	-0.5	0.5	0.1	0.5	0.7	0.0	-0.1	0.4	0.4	0.3	0.2	-0.2	0.1	0.1	0.4	-1.0	0.0	0.4	1.0	-0.6	-0.4	0.4
7	28	-0.3	-0.5	1.0	1.0	-0.5	-0.4	0.7	0.4	0.3	-0.8	-0.8	0.3	1.0	1.0	0.4	1.0	0.6	0.0	0.0	-0.8	-0.6	-12.3
8	29	-0.8	0.7	0.2	0.8	1.0	0.2	0.7	0.7	0.6	-0.8	-0.8	0.4	0.3	0.6	0.6	-0.5	0.1	-0.2	0.0	-0.1	0.0	24.6
9	31	0.3	-0.5	0.3	0.3	-0.6	-0.4	0.7	0.4	0.3	1.0	0.9	-0.2	-0.1	-0.2	0.4	-0.5	-1.0	0.6	-0.5	-0.3	-0.5	-16.5
10	32	-0.5	-0.5	0.7	0.7	-0.7	-0.9	0.2	0.5	0.5	-0.8	-0.7	-0.2	0.8	0.6	0.5	-0.1	0.7	0.2	-0.4	0.6	0.8	-36.3
11	33	-0.5	1.0	-0.5	-0.1	0.7	1.0	-1.0	1.0	1.0	-0.8	-0.7	-1.0	-0.2	-0.1	1.0	-0.3	-0.2	-0.3	0.5	0.1	-0.2	-17.2
12	36	-0.5	0.6	-0.9	-0.7	0.5	0.2	-0.7	0.8	0.8	1.0	1.0	-0.4	-0.1	-0.2	0.8	-0.8	0.3	-0.9	-0.9	-0.1	-0.4	-26.7
13	39	-0.3	-0.6	-0.1	-0.1	-0.6	0.2	-1.0	0.0	0.0	1.0	0.8	-0.3	0.1	-0.2	0.0	-0.8	0.5	-0.8	-0.8	-0.1	-0.2	-63.5
14	41	0.0	-1.0	-0.4	-0.5	-0.5	-0.6	-1.0	-1.0	-1.0	-0.3	-0.4	0.3	-0.1	-0.2	-1.0	-0.5	1.0	-0.6	-0.3	-0.3	-0.4	-78.1
15	45	0.3	-0.7	0.1	0.0	-0.3	-0.9	-0.1	-0.3	-0.3	-0.3	-0.4	0.8	-0.9	-1.0	-0.3	-0.4	0.5	0.6	0.0	0.1	0.3	-47.7
16	50	-1.0	0.5	0.0	0.4	0.5	0.8	-1.0	0.8	0.8	-1.0	-1.0	0.1	-0.9	-0.9	0.8	-0.3	0.4	-0.3	0.1	-0.1	-0.5	-37.0
17	53	-0.8	-0.8	0.3	0.1	-0.8	0.0	-1.0	0.1	0.2	-1.0	-1.0	0.4	-0.2	-0.3	0.2	-0.8	0.5	-0.7	-0.2	0.1	0.4	-66.8
18	54	-0.8	-0.7	0.4	0.3	-0.6	-0.6	-0.1	0.1	0.2	-0.8	-0.8	0.6	0.8	0.6	0.2	-0.5	0.6	0.7	0.6	1.0	0.2	-38.3

(a) the catalogued designs had different performances under the Bayesian objective function, and

(b) the Bayesian D-optimal design was always in the catalogue.

The objective function for generating Bayesian D-optimal designs for ridge regression places equal prior information on all of the coefficients to be estimated. For the lasso, there is more prior information for the potential terms (the coefficients that are thought to be zero a priori) in the assumed model through the term δ_0. It has been observed that by increasing the strength of the prior information on the coefficients of the potential terms, the structure and properties of the designs change, for example, the estimation of the primary terms is favoured.

There is no such mechanism in place for adjusting ridge regression designs, where the strength of the prior information is controlled by the choice of λ which affects all parameters equally. Thus the criterion produces designs which provide equal information for the estimation of all coefficients.

There are known limitations to each of the methods introduced. The form of the posterior variance for ridge regression is closed form whereas, for the lasso, the posterior variance is only an approximation. It would be interesting, in the future, to investigate the performance of the designs under a different approximation, for example, using Markov Chain Monte Carlo techniques.

Chapter 5

Sequential Design of Experiments for Bridge Regression

5.1 Introduction

In this chapter, we consider the problem of sequential selection of additional design points to enhance an existing data set for the efficient estimation of a bridge regression model. Point sequential design is used to choose the next design point to run to aim to give the greatest improvement in predictions made from the fitted model. Sequential design can be particularly effective if data collection is slow, such as in chemistry experiments where compounds must be synthesised and their properties measured. This means that time is available for the next design point to be selected in an optimal manner. We approach this problem by choosing the next design point, from a finite list of possible candidate points, as the point where the predicted response has highest variance (MacKay, 1992). For general bridge regression, including the lasso, the prediction variance of each candidate point is estimated using a bootstrapping procedure (Section 5.2). The use of bootstrapping to estimate prediction variance for sequential design is a novel approach.

The methodology developed in this chapter is applied to the Melting Point Data Set in Section 5.4, and its effectiveness investigated. The sequential selection of compounds, from a predefined list of compounds, that are most likely to enhance an existing data set and improve a predictive model for the melting points of compounds not included in the data set is a useful tool for a synthetic chemist. It provides

guidance on the best compounds to synthesise. This will save materials, and also time in the lab and on analytical equipment used to measure physical and chemical properties, where data collection can take several hours and use of the machine is in high demand. Costs would be reduced as a positive consequence. As in the previous chapter, the data set used in this chapter is made up of 55 compounds described by 21 quantitative descriptors. The simulated responses are generated from the known model with 7 non-zero coefficients (including the intercept) defined in Section 2.3.

5.1.1 Literature Review

In an important paper on sequential design using prediction variance, MacKay (1992) considered three different criteria for the selection of a new design point. These criteria can be applied to linear and non-linear models, and correspond to different aims of the experiment. For each criterion, a function of the next design point, $\tilde{\boldsymbol{x}}$, was defined that predicted the information gain from an observation at $\tilde{\boldsymbol{x}}$. These criteria were

1. select $\tilde{\boldsymbol{x}}$ to be the point that has highest prediction variance, as an approximation to improving the overall predictive accuracy of the fitted model. This is the criterion applied in this chapter (Section 5.2),

2. select $\tilde{\boldsymbol{x}}$ from a region of particular interest within the design space to maximise the accuracy of predictions within that region,

3. select $\tilde{\boldsymbol{x}}$ to determine the most appropriate model from two or more competing models.

There has been a variety of recent work on sequential design for regression-type models, focussed on choosing design points to satisfy different performance criteria. This includes work by Mandal, Wu and Johnson (2006) who introduced a Sequential Elimination of Level Combinations algorithm, in which the initial experiment is carried out using an orthogonal array design. The design points that are worst at achieving the goal of the experiment, e.g. those where the response is low when the aim is to maximise the response, are placed in a 'forbidden array' from which future runs may not be chosen. A genetic algorithm is then used to generate a specified number of better designs using the current best design points. The best design points

to use to create new experimental runs during the crossover and mutation processes of the genetic algorithm are determined according to probabilities proportional to a fitness measure, such as the value of the response. In this way the experimental design is guided by the information gained from the previously collected data. The authors applied the method to a combinatorial chemistry problem with the aim of selecting combinations of reagents that produce compounds that exhibit a maximum of desired efficacy.

Deng, Joseph, Sudjianto and Wu (2009) used a combination of stochastic approximation and optimal design methods to develop an 'active learning through sequential design' approach to estimate an optimal threshold hyperplane. This hyperplace is a decision boundary which may be linear, non-linear or nonparametric. In this paper, the hyperplane was used to decide whether or not a bank account had been used for money laundering according to a threshold 'probability of suspiciousness'. The procedure is:

- estimate the threshold hyperplane from the currently available data, e.g. by fitting a logistic regression model,
- identify those points in a candidate list of data which are closest, in terms of Euclidean distance, to this estimated threshold hyperplane,
- from these points, select for inclusion in the design the point that maximises the determinant of the Fisher information matrix of the fitted model for the threshold hyperplane.

These recent articles show that the issue of sequential design is of current interest within the experimental design literature, and that there are several methods of approaching the problem with applications in various different situations.

5.2 Sequential Design Improvement Algorithm for Bridge Regression

Point sequential design can be employed to improve an initial experiment from which model (1.1) has been estimated using bridge regression. At each step, an additional design point, $\tilde{\boldsymbol{x}}$, is selected from a fixed set of possible candidate design points,

$\mathcal{C} = \{\tilde{\boldsymbol{x}}_1, \ldots, \tilde{\boldsymbol{x}}_{N_\mathcal{C}}\}$, $\tilde{\boldsymbol{x}}_i \in \chi$, and the corresponding experimental run performed. The experiment continues one-step-at-a-time. At the kth step, the next point is chosen as

$$\tilde{\boldsymbol{x}}_k = \arg\max_{\tilde{\boldsymbol{x}} \in \mathcal{C}} \mathrm{Var}_{k-1}(\hat{Y}(\tilde{\boldsymbol{x}})), \qquad k = 1, \ldots, N_k, \tag{5.1}$$

see also MacKay (1992). Here, $\mathrm{Var}_{k-1}(\hat{Y}(\tilde{\boldsymbol{x}}))$ is the prediction variance based on the data collected up to step $k-1$, including from the initial design. We denote the size of the initial experiment by N_0, and the size of the experiment after the kth step by $N_k = N_0 + k$.

For ridge regression, there is a closed form for the variance-covariance matrix of the coefficients and hence the variance of $\hat{Y}(\boldsymbol{x})$ for $\boldsymbol{x} \in \chi$. However, this is not the case for the lasso (see Section 2.4). The approximations to the variance-covariance matrix for the lasso, detailed in Section 2.4.3, are not used to approximate the prediction variance of the candidate points for point sequential design. Instead, a bootstrap procedure (see Section 2.4.4) is used which makes fewer assumptions, for example, it does not assume a linear relationship between $\hat{\boldsymbol{Y}}$ and $\boldsymbol{Y}$.

For $0 < \gamma \leq 2$, the prediction variance can be approximated by using the bootstrap model fitted to each bootstrap sample to predict the response at each point in $\mathcal{C}$. Then the variance of each candidate point is estimated over its B bootstrap fitted values, $\hat{Y}_h^*(\tilde{\boldsymbol{x}})$ $(h = 1, \ldots, B)$, as

$$\mathrm{Var}_{k-1}(\hat{Y}(\tilde{\boldsymbol{x}})) = \frac{1}{B-1} \sum_{h=1}^{B} (\hat{Y}_h^*(\tilde{\boldsymbol{x}}) - \bar{Y}(\tilde{\boldsymbol{x}}))^2, \tag{5.2}$$

where

$$\bar{Y}(\tilde{\boldsymbol{x}}) = \frac{1}{B} \sum_{h=1}^{B} \hat{Y}_h^*(\tilde{\boldsymbol{x}}),$$

for all $\tilde{\boldsymbol{x}} \in \mathcal{C}$. The design point chosen to be next included is the candidate point that exhibits the highest prediction variance estimated from the B bootstrap samples.

The design improvement procedure at step k is then given by the following algorithm.

`Algorithm:`

1. `Use bridge regression to fit model (1.1) to the current data,` $(\boldsymbol{x}_1, Y_1), \ldots, (\boldsymbol{x}_{N_{k-1}}, Y_{N_{k-1}})$`, and obtain estimated coefficients` $\hat{\boldsymbol{\beta}}$ `and`

residuals $\hat{\boldsymbol{r}}$

2. For each $\tilde{\boldsymbol{x}} \in \mathcal{C}$, evaluate $\mathrm{Var}_{k-1}(\hat{Y}(\tilde{\boldsymbol{x}}))$ using bootstrapping

   a. Randomly resample the residuals with replacement, to obtain $\boldsymbol{r}_i^*$, where $i = 1, \ldots, N_{k-1}$

   b. Set $Y_i^* = \boldsymbol{x}_i^{\mathrm{T}} \hat{\boldsymbol{\beta}} + \boldsymbol{r}_i^*$

   c. Fit a bridge regression to the resampled data $(\boldsymbol{x}_1, Y_1^*), \ldots, (\boldsymbol{x}_{N_{k-1}}, Y_{N_{k-1}}^*)$, with λ re-estimated

   d. Calculate $\hat{Y}^\star(\tilde{\boldsymbol{x}})$ for all $\tilde{\boldsymbol{x}} \in \mathcal{C}$

   e. Repeat steps a-d B times, and calculate the sample variance of the predictions for each $\tilde{\boldsymbol{x}} \in \mathcal{C}$, using (5.2)

3. Choose $\tilde{\boldsymbol{x}}_k$ such that (5.1) is satisfied, and perform a run at this new point

5.2.1 Comments on the Bootstrapping Procedure

The bootstrap procedure used for the sequential selection of design points differs in two ways from that outlined in Section 2.4.4. These differences are

- how the resampled residuals are obtained,
- the re-estimation of the tuning parameter for the models fitted to each bootstrap sample.

The reasons for these differences are discussed below for the lasso; as there is no closed form for the variance of $\hat{Y}(\boldsymbol{x})$, the lasso is the motivation for developing the bootstrapping procedure. The comments can be generalised to methods for the bridge regression family in general.

1. Residual resampling: as $N_k < (p+1)$ for many sequentially designed studies, especially in the early steps, we resample residuals obtained from fitting a lasso regression model rather than a linear model fitted via OLS. We use the corrected AIC (2.25) to avoid overfitting the model and under-estimating σ^2.

2. Re-estimating the tuning parameter for each bootstrap sample: the tuning parameter s is re-estimated for each bootstrap sample to avoid under-estimating the prediction variance. In extreme cases s may be estimated as $s = 0$ for the original data and then, if s is not re-estimated, the predictions will be equal for all $\tilde{\boldsymbol{x}} \in \mathcal{C}$ within each bootstrap sample, and (5.2) will be the same for each $\tilde{\boldsymbol{x}}$. Hence, the criterion will be unable to choose between the candidate points.

5.3 Sequentially Developed Designs for Two-level Factors

In this section, we apply the sequential design process of Section 5.2 to the Bayesian D-optimal designs from Section 4.4.4, for model (1.1) having all main effects and two-factor interaction terms, and for f factors each having two levels, +1 and -1. Designs for ridge regression were found using (4.13) with identical prior distributions assumed for all predictors and $\lambda = 0.5$. Designs for the lasso were found using (4.14) with $\lambda = 0.5$ and $\delta_0 = 1$ assumed for the potential terms. The potential terms are identified as those with $\beta_j = 0$ in (5.5) below.

For each size of design, $N = 18$ and $N = 33$, we consider two ridge regression D-optimal designs, one with $f = 4$ variables and the other with $f = 6$ variables, and one lasso D-optimal design with $f = 6$ variables. The design and observations for the $N = 33$ designs, d_1 (ridge, $f = 4$), d_2 (ridge, $f = 6$) and d_3 (lasso, $f = 6$), are shown in Tables 4.1, 4.2 and 4.3 respectively. The design and observations for the $N = 18$ designs, d_4 (ridge, $f = 4$), d_5 (ridge, $f = 6$) and d_6 (lasso, $f = 6$), are shown in Tables 4.5, 4.6 and 4.7 respectively. The bootstrapping procedure for obtaining the prediction variances of the candidate points was applied to the ridge regression designs as well as the lasso designs.

The candidate list from which the design points are sequentially selected and then included in the initial designs is made up of all possible combinations of +1 and −1 variable values, therefore for the $f = 4$ example there are 16 candidate points and for the $f = 6$ examples there are 64 candidate points. The known model used to simulate the response for the design points in the D-optimal initial designs

and for those points subsequently chosen for inclusion is

$$Y_i = \beta_0 + \sum_{j=1}^{f} x_{ij}\beta_j + \sum_{l=1}^{f}\sum_{j>l}^{f} x_{il}x_{ij}\beta_{lj} + \varepsilon_i, \tag{5.3}$$

where $\beta_0 = 2$. For the $f = 4$ ridge regression designs

$$\boldsymbol{\beta}^{\mathrm{T}} = (0.5, 0.7, 3, 6, 10, 15, 20, 50, 55, 100), \tag{5.4}$$

where the first $f = 4$ entries of $\boldsymbol{\beta}$ are the 'true' coefficients of the main effect terms and the remaining $p - f = 6$ entries are the 'true' coefficients of the interaction terms.

For the $f = 6$ ridge regression and lasso designs

$$\boldsymbol{\beta}^{\mathrm{T}} = (0.5, 0.7, 3, 6, 0, 0, 10, 15, 20, 0, 0, 50, 55, 0, 0, 100, 0, 0, 0, 0, 0), \tag{5.5}$$

where the first $f = 6$ entries of $\boldsymbol{\beta}$ are the coefficients of the main effect terms and the remaining $p - f = 15$ entries are the coefficients of the interaction terms. For all designs, the random error, ε_i, was generated as a random sample from a $N(0, 1)$ distribution.

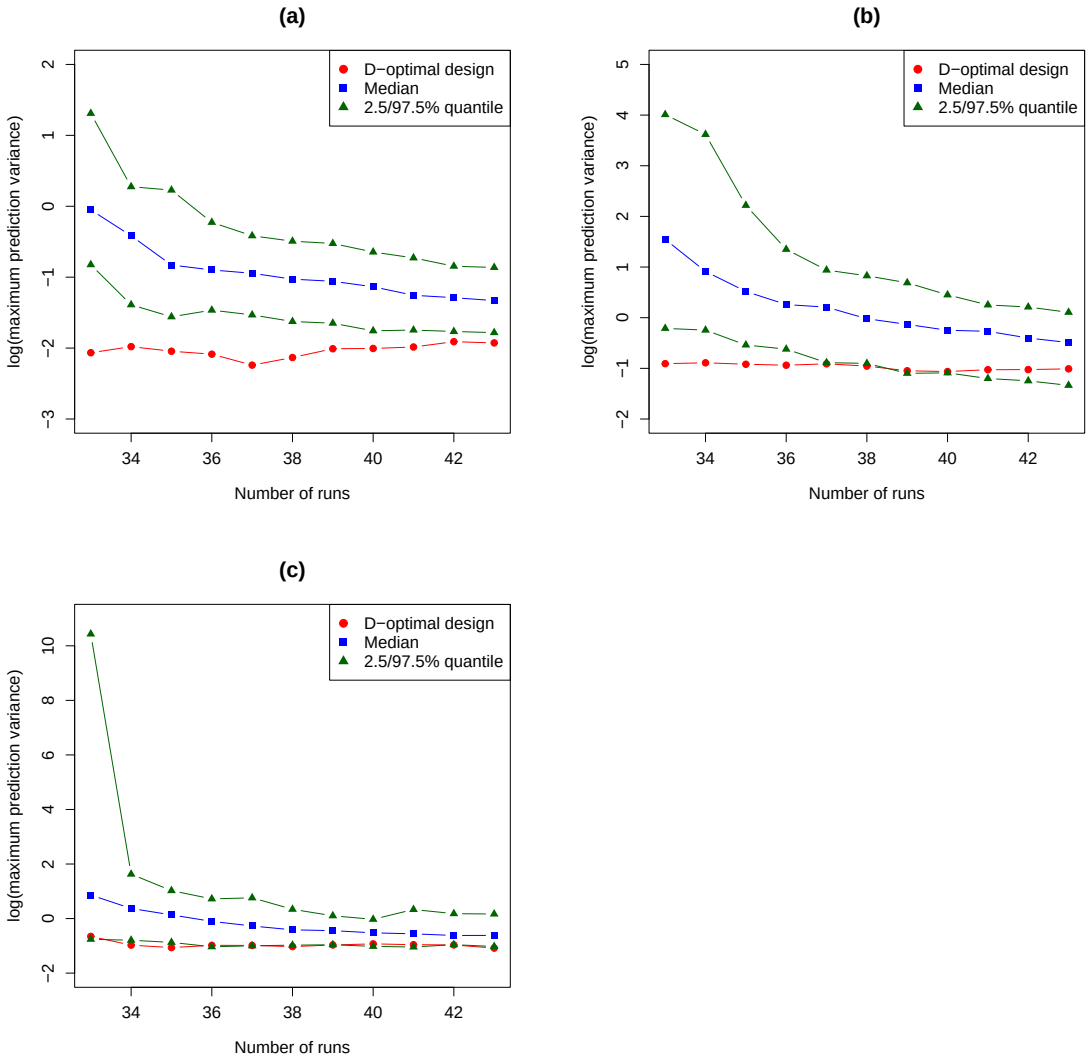
5.3.1 Results

For the $N_0 = 33$ Bayesian D-optimal initial designs, ten candidate points were added sequentially according to criterion (5.1) and using the algorithm in Section 5.2. The candidate list of points included the points already in the initial design, and each candidate point could be selected more than once. It should be noted that the Bayesian D-optimal initial designs are found under a estimation-based criterion rather than through a function of the prediction variance. Subsequently the maximum prediction variance is calculated without assuming the prior information used to generate the D-optimal designs and therefore it is not necessary to update the prior information after the addition of each design point.

Properties of the designs: The results of this process for the ridge regression designs and the lasso design are shown in Figure 5.1. For comparison, summaries of

the results obtained from 100 random initial designs are also included in the plots. For each of these designs, 33 design points were chosen at random from $\{-1, +1\}^f$. The plot shows, at each step, the median and the 2.5% and 97.5% percentiles of the maximum prediction variance obtained across all designs.

Figure 5.1: Maximum prediction variance of a two-level sequentially developed design from an initial 33-run Bayesian D-optimal design and from random initial designs for: (a) Ridge regression, $f = 4$; (b) Ridge regression, $f = 6$; (c) Lasso, $f = 6$. Also shown are the 50%, 2.5% and 97.5% quantiles for 100 designs from random initial designs



The performance of the sequential designs developed from all three D-optimal initial designs remains approximately constant as design points are added. The performance of the $f = 6$ ridge regression and lasso designs, d_2 and d_3, remain approximately at the same level as the 2.5% quantile of the 100 random initial designs throughout the sequential design procedure. The sequential design developed from the $f = 4$ ridge regression initial design, d_1, has maximum prediction variance lower than the 2.5% quantile for the designs from the 100 random initial designs at every step.

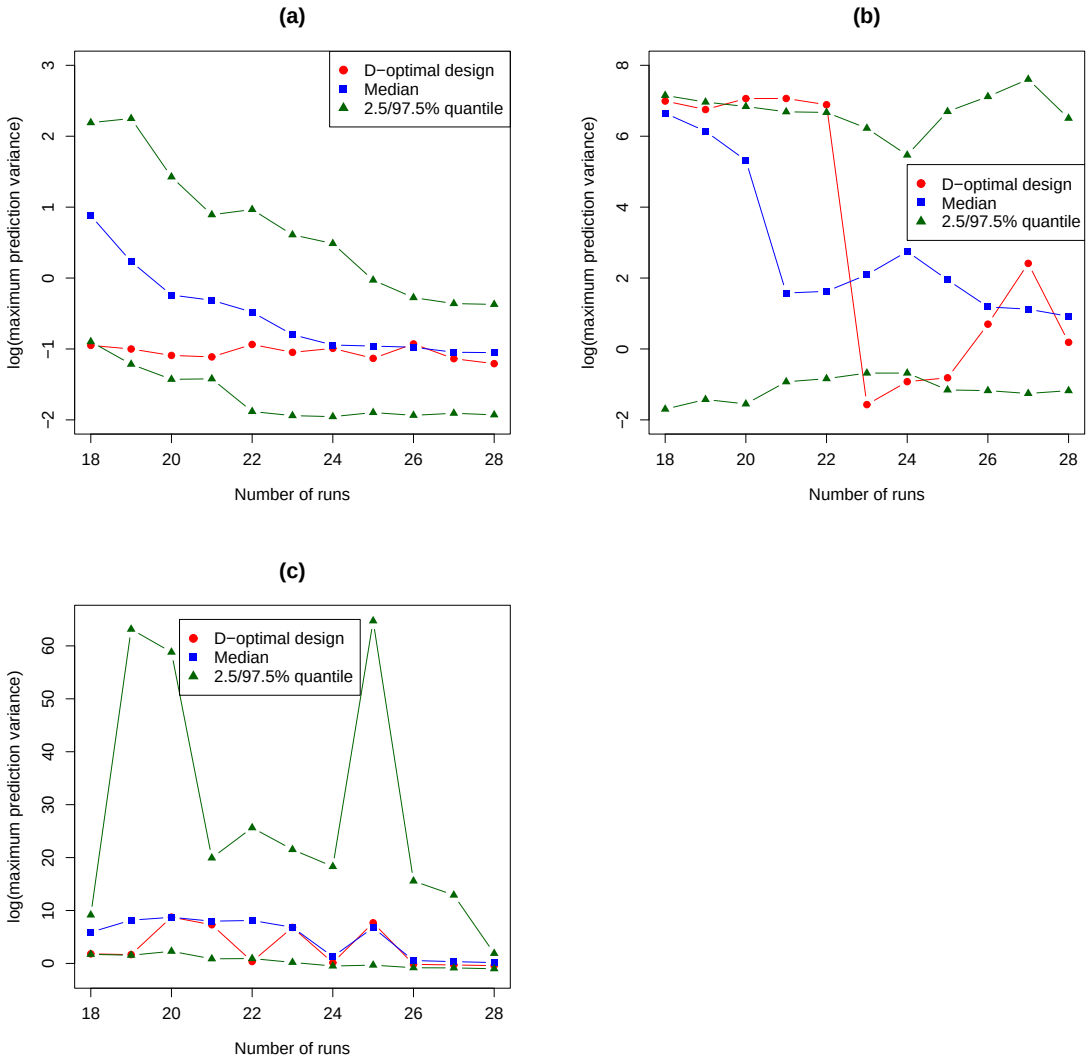
The three 33-run D-optimal initial designs were discussed in Section 4.4.4, where it was noted that these designs were as balanced as possible. These designs would therefore be expected to perform well under most criteria and Figure 5.1 shows that this is the case for maximum prediction variance. The models fitted to the simulated data from the initial D-optimal designs were previously found to be very similar to the known models from which the data were generated, with each coefficient having an estimated value within approximately 0.5 of the true value. We repeated the model-fitting study as each new point was added to each design. The estimated coefficient values from the fitted models were again very similar to those of the true model. Therefore, there is little scope for the addition of a few runs to substantially improve the designs.

The worst performance of the designs developed from the random initial designs compares poorly to the D-optimal initial designs. However, the sequential design procedure is able to produce a substantial improvement over the worst performing of these designs after only a few new design points have been selected. This is especially apparent for $f = 6$ and the lasso (Figure 5.1(c)) and the 97.5% percentile of the 100 random initial designs, where the performance is dramatically improved after the inclusion of only one new design point. Note that there are also substantial improvements for the ridge regression designs but the log scale in Figure 5.1 somewhat flattens the curves.

Ten new points were added sequentially to the $N_0 = 18$ Bayesian D-optimal designs using criterion (5.1), and these results are shown in Figure 5.2. As before, 100 random initial designs were compared and summaries of their results are also included on the same plots.

The maximum prediction variance, at each step, of the sequential design de-

Figure 5.2: Maximum prediction variance of a two-level sequentially developed design from an initial 18-run Bayesian D-optimal design and from random initial designs for: (a) Ridge regression, $f = 4$; (b) Ridge regression, $f = 6$; (c) Lasso, $f = 6$



veloped from the $f = 4$ ridge regression design, d_4, remains roughly constant as design points are added sequentially to the design. The maximum prediction variance for this design is consistently lower than that of the median of the sequential designs from the 100 random initial designs, and is only slightly higher than the 2.5% quantile. Again, substantial improvements in the designs obtained from the

random initial designs can be observed.

For $f = 6$ variables, there are $p = 21$ main effects and interactions in the model. Hence, for $N_0 = 18$, $N_0 < (p + 1)$. Although the $f = 6$ ridge regression design, d_5, initially performs worse than the 97.5% quantile obtained from the 100 random initial designs, subsequent performance improves dramatically once there are more distinct design points in the design than there are predictors (this occurs at $N_5 = 23$, see later). The design performs better than the median from the 100 random initial designs by $N_{10} = 28$. The scale of the vertical axis in Figure 5.2(b) makes the increase in maximum prediction variance of the $f = 6$ ridge regression design at $N_9 = 27$ appear deceptively large. In fact, between $N_8 = 26$ and $N_9 = 27$ the maximum prediction variance is only increased from 2.0 to 11.2. A maximum prediction variance of 11.2 is a substantial improvement on the performance of the initial design, which had a maximum prediction variance of 1088.2. On average, the performance of the random designs is also substantially improved by the addition of well chosen extra points.

The performance of the designs obtained from the $f = 6$ lasso initial design, d_6, neither monotonically increases or decreases as design points are added sequentially (see discussion in next paragraph). However, the performance remains better than that of the median of the 100 random initial designs and at certain points equals that of the 2.5% quantile of the 100 random initial designs.

When $N_0 < (p + 1)$, as for this example with $N_0 = 18$ and $f = 6$, the results of applying the sequential design method to an initial ridge regression or lasso Bayesian D-optimal design are somewhat erratic. The models fitted to these initial D-optimal designs are not very similar to the true simulation model, with the ridge regression model in particular severely underestimating the coefficients that were non-zero in the true model. Each additional point is more influential and causes larger changes to the estimated coefficients of the fitted model than for designs with $N_0 \geq p$.

Tables 5.1, 5.2 and 5.3 show the candidate points selected by criterion (5.1) for inclusion in the $N_0 = 18$ initial designs for $f = 4$ ridge regression, $f = 6$ ridge regression and $f = 6$ lasso respectively. For the $f = 4$ ridge regression design, the criterion selects ten different candidate points for inclusion in the design, all of which were already present in the initial design. This is not unexpected because there are only 16 candidate points to choose from for the $f = 4$ example and there are only 10

predictors in the model. For the $f = 6$ ridge regression design, the criterion selects 8 distinct points for inclusion in the design, with one of the points repeated three times. For the $f = 6$ lasso design, the criterion selects 7 distinct points for inclusion in the design, with one of the points repeated four times. The repeated additional point included in the $f = 6$ ridge regression design is not the same candidate point repeated in the sequentially developed lasso design. The only design point of the ten additional points included in the $f = 6$ lasso design that was present in the initial design is the repeated point.

Table 5.1: The ten design points selected for addition to the $f = 4$, $N_0 = 18$ ridge regression initial design

Run	A	B	C	D	Y
19	1	-1	1	1	30.52
20	1	-1	-1	-1	153.53
21	-1	-1	-1	-1	243.03
22	-1	-1	1	-1	-81.52
23	-1	-1	1	1	-19.91
24	-1	1	-1	1	-103.74
25	1	-1	-1	1	-103.36
26	-1	-1	-1	1	-95.11
27	1	1	-1	-1	-36.39
28	1	1	1	1	262.60

Figures 5.3, 5.4 and 5.5 show projections of the design points included in the $N_0 = 18$ initial designs for $f = 4$ ridge regression, $f = 6$ ridge regression and $f = 6$ lasso respectively. In these plots, the design points included in the initial designs are coloured blue and those subsequently chosen by criterion (5.1) are coloured red. For the designs developed from the $f = 4$ ridge regression initial design (Figure 5.3) the plot shows that for factors which have an uneven balance of design points across the factor levels in the initial design, the sequential design procedure is capable of improving this balance. This is shown by an even number of design points at each corner of every subplot of the figure. This contributes to the observed stability in maximum prediction variance during the sequential design procedure for this design. For the designs developed from the $f = 6$ ridge regression and lasso designs (Figures 5.4 and 5.5), where $N_0 < (p + 1)$, we do not see the same tendency to 'balance' the

Table 5.2: The ten design points selected for addition to the $f = 6$, $N_0 = 18$ ridge regression initial design. The bold combinations of factor levels are repeated points

Run	A	B	C	D	E	F	Y
19	**1**	**-1**	**-1**	**-1**	**-1**	**1**	**152.36**
20	**1**	**-1**	**-1**	**-1**	**-1**	**1**	**152.71**
21	-1	1	1	-1	-1	-1	-108.59
22	**1**	**-1**	**-1**	**-1**	**-1**	**1**	**151.77**
23	-1	-1	1	1	-1	1	-18.73
24	1	1	1	-1	-1	1	-100.87
25	-1	1	-1	-1	1	-1	13.44
26	-1	1	-1	1	1	1	-104.21
27	1	1	1	1	1	-1	263.29
28	1	1	-1	1	1	-1	-72.97

Table 5.3: The ten design points selected for addition to the $f = 6$, $N_0 = 18$ lasso initial design. The bold combinations of factor levels are repeated points

Run	A	B	C	D	E	F	Y
19	**1**	**1**	**1**	**1**	**-1**	**-1**	**260.38**
20	**1**	**1**	**1**	**1**	**-1**	**-1**	**261.84**
21	-1	1	1	1	-1	1	170.20
22	1	1	1	1	1	-1	261.07
23	**1**	**1**	**1**	**1**	**-1**	**-1**	**263.89**
24	**1**	**1**	**1**	**1**	**-1**	**-1**	**260.66**
25	1	1	-1	-1	1	-1	-36.39
26	1	1	1	1	-1	1	262.18
27	-1	-1	1	-1	-1	1	-80.43
28	-1	1	-1	-1	-1	-1	12.61

design points.

Figure 5.3: Projection of design points of the $f = 4$, $N_0 = 18$ ridge regression initial design and the ten additional design points. The blue dots are the design points of the initial design and the red dots are the additional design points. A small amount of random noise has been added to the values in order to separate them

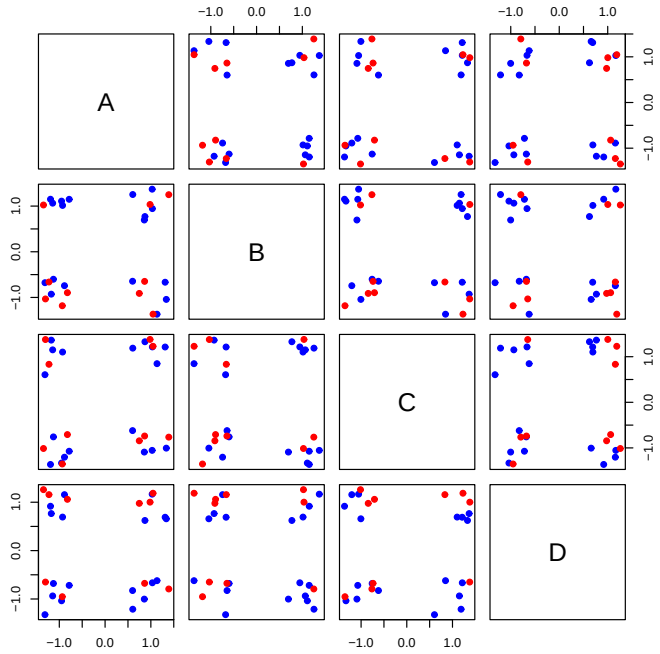


Figure 5.4: Projection of design points of the $f = 6$, $N_0 = 18$ ridge regression initial design and the ten additional design points. The blue dots are the design points of the initial design and the red dots are the additional design points. A small amount of random noise has been added to the values in order to separate them

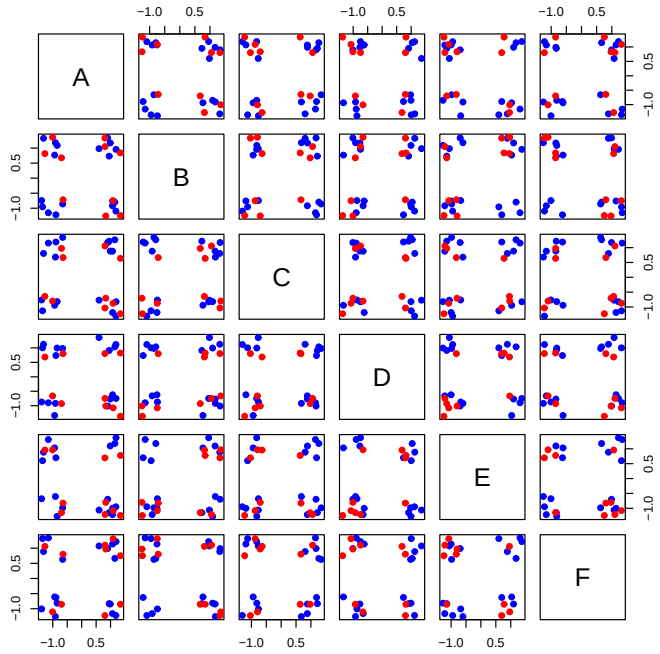
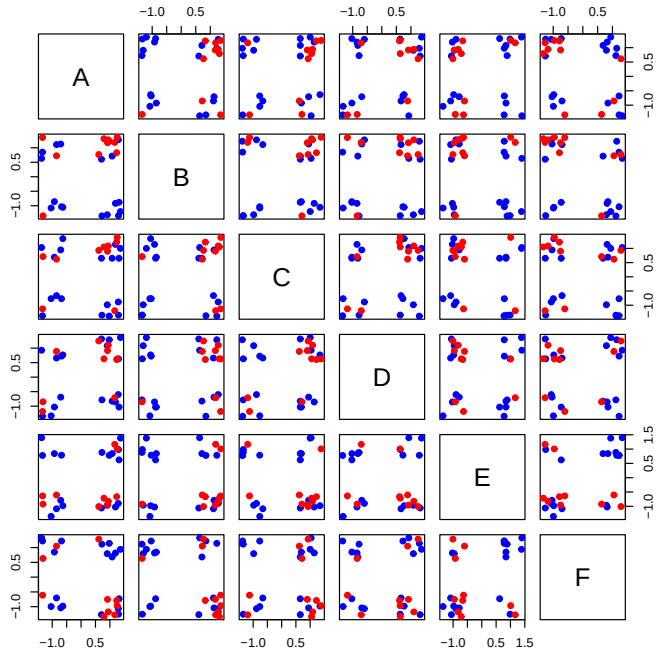


Figure 5.5: Projection of design points of the $f = 6$, $N_0 = 18$ lasso initial design and the ten additional design points. The blue dots are the design points of the initial design and the red dots are the additional design points. A small amount of random noise has been added to the values in order to separate them



5.4 Sequentially Developed Designs when Factor Levels cannot be Freely Combined

In this section, we investigate a sequential design procedure for experiments in which not all combinations of factor levels can be investigated. We use the melting point experiment (Section 1.2) and the assumed known model (2.15) as the focus of the work, and use the candidate list of 55 compounds (the possible design points) to find designs.

In Section 5.4.1, we describe a study to explore how the candidate point chosen by criterion (5.1) improves the prediction variances of the remaining candidate points, and compare this improvement to that obtained by all other choices of the next design point. In Section 5.4.2, we apply the sequential design procedure to the Bayesian D-optimal initial designs from Section 4.5.3, which were generated for the melting point experiment.

5.4.1 Investigation of $N_0 + 1$ Run Designs

In this section, we compare the improvement gained in prediction variance from the addition of a single point, chosen according to (5.1), to all possible alternative N_0+1 point designs. The 55 points of the candidate list were split randomly to form a $N_0 = 45$ initial design and a candidate list containing the remaining 10 points. Using the algorithm outlined below, all ten possible $N_1 = 46$ designs were compared by ranking the additional point according to the improvement in prediction variance of the remaining 9 candidate points. The whole procedure was repeated 100 times.

To be more specific, the $N_1 = 46$ point designs are thus investigated by

- selecting, at random, an initial design of $N_0 = 45$ points chosen from the 55-point candidate list,
- taking each of the 10 remaining candidate points in turn and adding them to the 45 points of the initial design to create a design with $N_1 = 46$ points,
- fitting the lasso model to the $N_1 = 46$ run design,
- calculating the average prediction variance of the remaining 9 candidate points from this fitted model,

- comparing the average prediction variance of the set of 9 candidate points from each $N_1 = 46$ design,
- repeating the entire process 100 times.

The steps actually performed are

`Algorithm:`

1. `For each point` $\tilde{\boldsymbol{x}} \in \mathcal{C}$ `calculate the variance of its predicted response according to the bootstrap procedure of Section 5.2`
2. `Rank the points` $\tilde{\boldsymbol{x}} \in \mathcal{C}$ `from 1 to` $N_{\mathcal{C}}$ `in order of decreasing variance,` $\tilde{\boldsymbol{x}}_{(1)}, \ldots, \tilde{\boldsymbol{x}}_{(N_{\mathcal{C}})}$
3. `Include` $\tilde{\boldsymbol{x}}_{(1)} \in \mathcal{C}$ `in the training set and refit the bridge regression`
4. `For each remaining point` $\tilde{\boldsymbol{x}}_{(2)}, \ldots, \tilde{\boldsymbol{x}}_{(N_{\mathcal{C}})} \in \mathcal{C}$ `calculate the variance of its predicted response according to the bootstrapping procedure of Section 5.2`
5. `Calculate the average of the variances calculated in step 4`
6. `Repeat steps 3-5, including in the design each point` $\tilde{\boldsymbol{x}}_{(2)}, \ldots, \tilde{\boldsymbol{x}}_{(N_{\mathcal{C}})} \in \mathcal{C}$ `in turn`
7. `Rank the points` $\tilde{\boldsymbol{x}}_{(j)} \in \mathcal{C}$ $(j = 1, \ldots, \mathcal{C})$ `in order of increasing average of the variances from step 5`

The results of this study are shown in Figure 5.6, as histograms of the ranking of the candidate point with the highest prediction variance for each of the 100 repetitions for both ridge regression and the lasso. If the sequential design criterion performs well across different training sets and candidate lists, the same point $\tilde{\boldsymbol{x}}^*$ will

1. have the highest prediction variance (step 1 of the above algorithm)

$$\tilde{\boldsymbol{x}}^* = \arg\max_{\tilde{\boldsymbol{x}} \in \mathcal{C}} \mathrm{Var}_{45}(\hat{Y}(\tilde{\boldsymbol{x}})),$$

2. have lowest average variance across $\mathcal{C} \setminus \{\tilde{\boldsymbol{x}}^*\}$ (step 5)

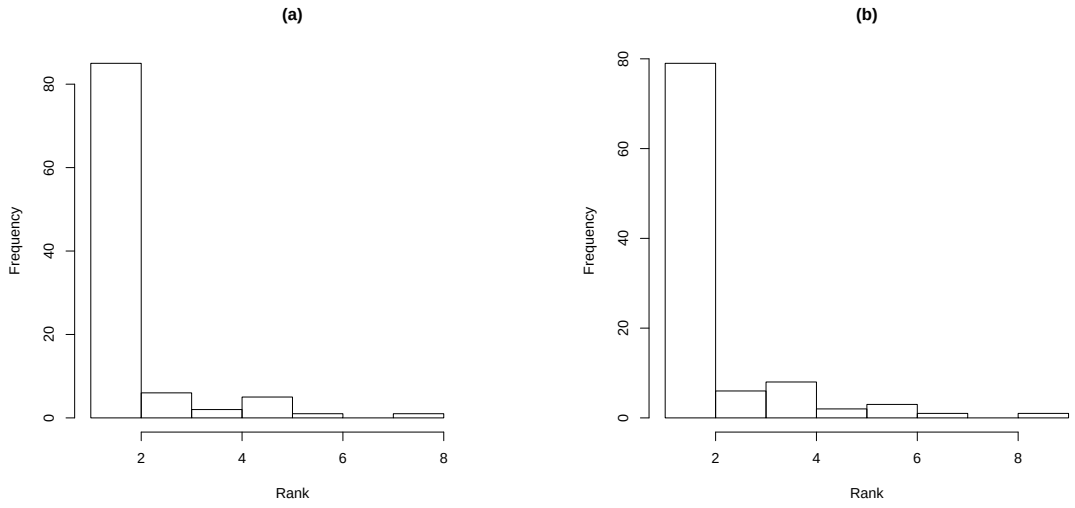
$$\min_{\tilde{\boldsymbol{x}} \in \mathcal{C}} \sum_{\mathcal{C} \setminus \{\tilde{\boldsymbol{x}}^*\}} \text{Var}_{46}(\hat{Y}(\tilde{\boldsymbol{x}}))/9,$$

and hence

3. have rank 1 when $\tilde{\boldsymbol{x}} \in \mathcal{C}$ are ranked by increasing average prediction variance (step 7).

For the majority of 100 repetitions of the procedure, we can see that this is the case. For ridge regression, the mean ranking of the candidate point $\tilde{\boldsymbol{x}}^*$ was 1.72, the median ranking was 1 and the variance was 1.68. For the lasso the mean ranking was 1.92, the median ranking was 1 and the variance was 2.40.

Figure 5.6: Histograms of ranks of improvement to the initial design of candidate point (5.1): (a) Ridge regression; (b) Lasso



In the vast majority of the 100 repetitions of the procedure, the sequential design criterion (5.1) is able to select the candidate point that improves the performance of the design the most; over 80% of the time for ridge regression and just under 80% of the time for the lasso. This study has shown that choosing the next point via (5.1) is able to improve the average of the bootstrap prediction variances of the remaining candidate points after a point is added to the design.

5.4.2 Application to *D*-optimal Initial Designs

The sequential design process of Section 5.2 has been applied to each of the D-optimal designs found for the organic chemistry example, with main effect terms only, that were presented in Section 4.5.3. Ten points have been added, one at a time, to each of these designs according to criterion (5.1) and using the algorithm described in Section 5.2. These points were selected from a candidate list made up of 55 points, the compounds in the data set. This candidate list included the points in the initial design and each of the candidate points could be selected more than once. The response values for the candidate data points, including the points of the initial designs, were simulated from the explanatory data and the known model described in Section 2.3.

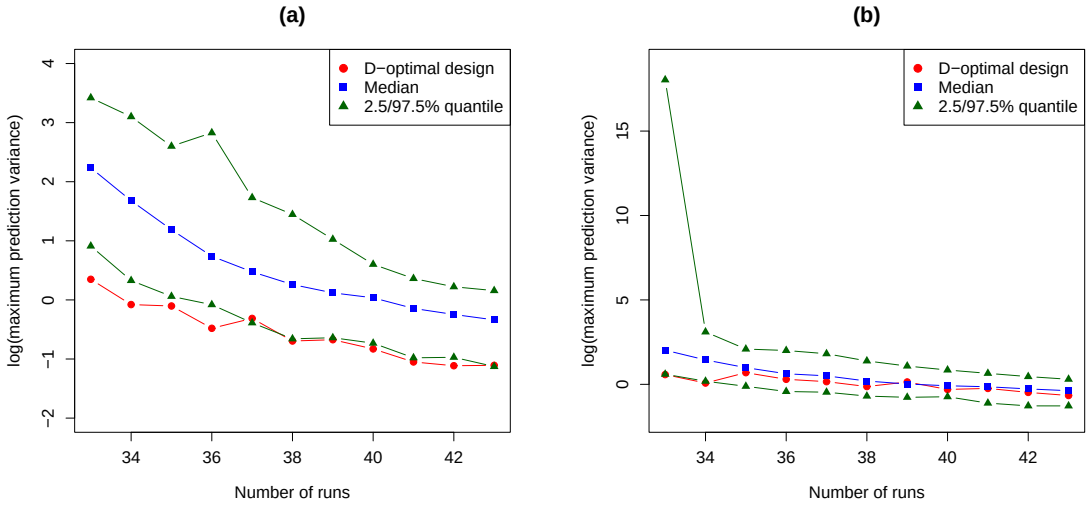
For comparison, the sequential design procedure was also carried out on 100 random initial designs generated from the Melting Point Data Set. To generate these designs, points were selected at random, with replacement, from the candidate list of 55 points. Summaries of the results are presented in Figure 5.7, along with the results of the corresponding D-optimal design. At each step, we plot the median, 2.5% and 97.5% quantiles of the maximum prediction variances obtained from all designs.

The sequential design criterion was applied to develop designs for model (1.1) for ridge regression and the lasso with an initial Bayesian D-optimal design of $N_0 = 33$ points. The results for each of these designs are shown in Figure 5.7, starting from the $N_0 = 33$ point initial design; summaries of the results obtained for 100 random initial designs are also included. The ridge regression initial design is shown in Table 4.12, and the lasso initial design in Table 4.14.

It is clear, from Figure 5.7, that all designs developed, including from the random initial designs, are improved in terms of maximum prediction variance after only a few points have been added sequentially. The ridge regression design has better performance than the lasso design, in terms of maximum prediction variance, throughout the procedure. At its best ($N_2 = 35$), the maximum prediction variance of the ridge regression design is approximately half that of the lasso design. The performances of sequential designs from the ridge regression and lasso designs remain better than the median for designs from the 100 random initial designs. The maximum prediction variance for the ridge regression design is lower than that of

the 2.5% quantile of the 100 random initial designs at every stage of the procedure.

Figure 5.7: Maximum prediction variance for $N_0 = 33$ Bayesian D-optimal designs, Melting Point Data Set: (a) Ridge regression; (b) Lasso



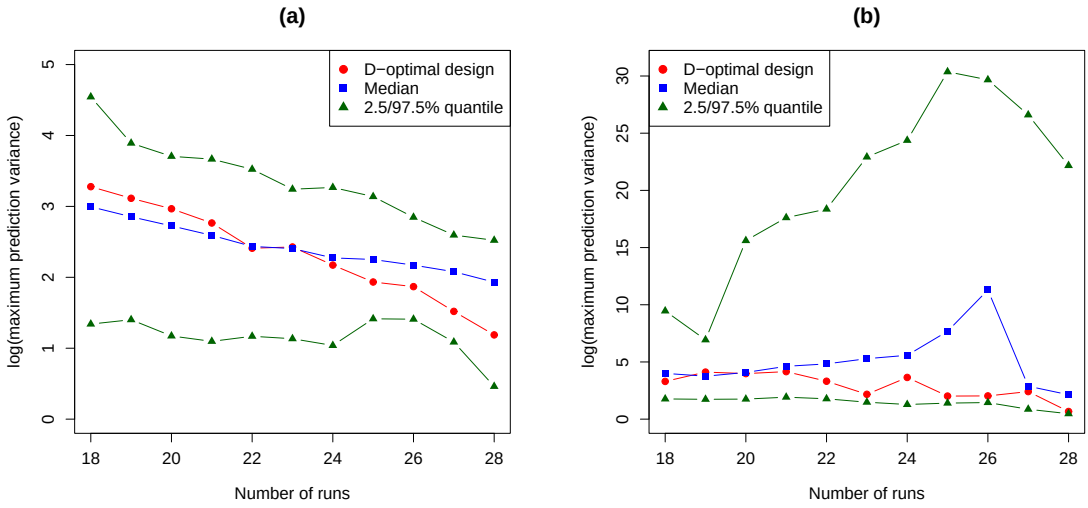
The sequential design criterion was also applied to the $N_0 = 18$ point Bayesian D-optimal designs presented in Section 4.5.3; the ridge regression initial design is shown in Table 4.13 and the lasso initial design is shown in Table 4.15. Again, summaries of the results obtained for 100 random initial designs are included for comparison. The results are shown in Figure 5.8.

The maximum prediction variance of the two D-optimal initial designs is improved once the additional ten design points have been included. The performance of these sequential designs is only slightly worse than the 2.5% quantile of the 100 random initial designs by the point at which all ten of the additional design points have been included in the designs. Again, the D-optimal designs have been found under an estimation-based criterion, and the maximum prediction variance is not calculated using the same prior information used to generate the D-optimal initial designs.

When using ridge regression, the performance of all the initial designs, including the random initial designs, are improved by the sequential addition of extra design points. When using the lasso, the maximum prediction variance of the worst

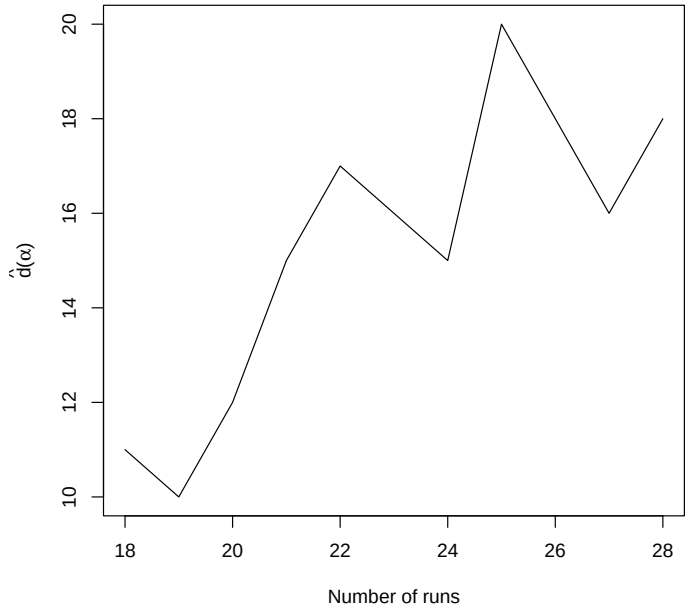
performing designs, represented by the 97.5% quantile of the 100 random initial designs, is unusual as it increases as design points are added sequentially. Figure 5.9 demonstrates how the complexity of the fitted lasso model that generates the 97.5% quantile changes as the size of the design increases. The effective degrees of freedom, in this case estimated using the simple approximation (i.e. by the number of non-zero terms in the model, see Section 2.3.4) is plotted against the number of runs in the design. The effective degrees of freedom increases as the number of runs increases; hence the complexity of the model is increasing. For the worst designs, additional (and perhaps spurious) terms are being added to the model. This leads to a less stable estimated model under resampling, and higher prediction variance. This behaviour is also reflected, to a lesser extent, in the median performance of the random designs.

Figure 5.8: Maximum prediction variance for $N_0 = 18$ Bayesian D-optimal designs, Melting Point Data Set: (a) Ridge regression; (b) Lasso



These results show that, in most cases, the sequential criterion can improve the maximum prediction variance of even the poorest performing designs after only a few steps and can also improve designs that already perform well under a particular measure of performance. Also, the Bayesian D-optimal designs are amongst the best initial designs when assessed under maximum prediction variance.

Figure 5.9: Estimated effective degrees of freedom of lasso models fitted at each run size for the worst performing random initial designs of Figure 5.8(b)



5.5 Conclusions

In this chapter, we have defined the criterion for the sequential selection of design points for the improvement of an existing design and the algorithm used to calculate the value of the criterion and select the design points. The method was applied to the D-optimal designs generated in Chapter 4 for both the two-level example and the example based on the Melting Point Data Set. The bootstrapping procedure allows the assessment of the performance of the initial designs via the maximum prediction variance of the points in the candidate set. The sequential selection of design points was able to improve the performance of poorly performing designs significantly, and also improve the performance of designs that already perform reasonably well, after only a few candidate points had been selected and included in the design.

Although there is a closed form for the prediction variance for ridge regression, the bootstrapping method provides a consistent method for estimating the prediction variance for both ridge regression and the lasso.

The results of applying the sequential design process to the D-optimal designs have shown that the criterion is capable of improving the performance of poor designs in very few steps. Hence, such methods could be used to 'repair' experiments that had been poorly designed initially. Together with intelligent choice of initial designs, the sequential design process would provide an advantage to synthetic chemists in their ability to select the most significant compounds to next synthesise, saving time, materials and money.

In the next chapter, we discuss two possible future directions for the work on sequential design for bridge regression. The first is an alternative criterion for the sequential selection of design points, which selects points that give the best expected improvement in overall prediction variance. The second is the selection of batches of design points which would be more time and cost efficient, especially for a synthetic chemist, than performing the experiments one-step-at-a-time.

Chapter 6

Discussion and Future Work

6.1 Conclusions

In this thesis, we have presented and applied methods for the design and analysis of experiments for bridge regression, with emphasis on the two commonly applied cases of ridge regression and the lasso. Novel methods for the optimal and sequential design of experiments have been developed and demonstrated when a response is approximated by a linear model and ridge regression or the lasso is to be used to analyse the experimental data.

Statistical models have been developed to predict the melting point of small molecule organic compounds from the values of their chemical descriptors and to gain understanding on which chemical features influence melting point. Ridge regression and the lasso were used to estimate models composed only of linear terms and also models containing two-factor interactions. These regression methods have not before been applied to this chemistry problem.

In many experiments, such as the one that motivated the work in this thesis, there are often more predictors than observations and there may be a high degree of multicollinearity between the predictor values run in an experiment. We have shown that both ridge regression and the lasso were able to screen the set of possible predictors for those that have an important effect on the response, through shrinking the coefficient estimates towards zero. The resulting models performed well in terms of R^2 and adjusted R^2 calculated over the data set on which the models were fitted, and in terms of out-of-sample R^2 and adjusted R^2 calculated by cross-validation. Their

performance also matched that of models obtained through more commonly used methods of variable selection. For simplicity and scientific interpretation, the lasso-fitted model is preferred because the lasso achieves a parsimonious model through variable selection.

The inclusion of interaction terms in the set of possible predictors led to a lasso model that outperformed the model fitted using only main effect (or linear) terms, based on the above model-fitting statistics. The model performance statistics obtained by cross-validation indicated that the ridge regression model including interactions was overfitted to the data. This is because ridge regression, unlike the lasso, does not shrink any of the coefficients to exactly zero. However, the performance of the ridge regression model including only main effects terms was similar to that of the lasso model including only main effects terms.

Although the predictive melting point models developed in this thesis may not be directly applicable to compounds from outside our set of structurally similar compounds, it would be likely that the same variables identified as having a significant influence on the melting point would also exert a similar effect on the melting point of more structurally diverse compounds. The identification of variable A, the width of the melting peak, as having an influence on melting point is an unexpected result. After discussion with the chemists, this variable is thought to be related to defects in the crystal structure which are known to influence melting point, but are difficult to quantify. It would be interesting to conduct further investigations into the significance of this variable.

The relationship between bridge regression and Bayesian methods was explored in order to develop a class of D-optimal designs using appropriate prior distributions. This provides a method for the selection of an optimal design for an experiment, with the aim of obtaining accurate estimates of the coefficients in the model. It was necessary to derive an approximation to the posterior variance for general bridge regression; for the lasso, for example, the corresponding prior distribution is Laplace and the posterior distribution is not available in closed form. Bayesian D-optimal designs could then be generated for ridge regression and lasso models by criteria which aim to maximise the determinant of the inverse of the posterior variance-covariance matrix. Such an objective function has not previously been defined or applied to bridge regression. The D-optimality criteria were applied to find and critically assess

designs for both two-level experiments and for an experiment when factors cannot be freely combined based on the motivating organic chemistry example.

The Bayesian D-optimality criterion for ridge regression places equal prior information on all of the coefficients to be estimated. For the lasso, stronger prior information is placed on terms thought to be zero a priori than those thought to be non-zero; hence the designs favour the estimation of model terms thought to be non-zero.

A point-sequential design criterion was developed to enhance an existing design. The criterion selects design points, from a finite set of candidate points, that exhibit the highest estimated prediction variance. This criterion is an approximation to improving the predictive ability of the model fitted to the augmented design. The lasso is a non-linear and non-differentiable function of the response, and so there is no closed form for the variance-covariance matrix of the coefficient estimators. Therefore a bootstrapping procedure was developed to provide estimates of the prediction variance of the candidate points for the lasso. This procedure may also be applied to ridge regression.

The sequential design criterion was applied to improve the Bayesian D-optimal designs found for two-level experiments, and also for the melting point example. The performance of the designs at each stage of the sequential design procedure was assessed in terms of the improvement to the maximum prediction variance of the candidate points evaluated over the candidate list. The method was critically assessed by comparison with randomly chosen initial designs. The criterion was able to improve the performance of all designs, including the worst performing of the random designs, after the selection and inclusion of a small number of design points. This assessment also showed that the D-optimal initial designs were some of the best designs found.

6.2 Future Work

Finally, we outline some additional research which could extend that presented in this thesis.

6.2.1 Modelling via Bridge Regression

The current work on using regression methods to predict melting points is limited by the nature of the available data set to compounds that are closely related in molecular structure. The types of compounds considered could be widened to include compounds that are more diverse and complex in their molecular structures, which would allow a more thorough comparison with the published literature (see Section 3.3). These further compounds could be used to develop new predictive models for different sets of compounds using the same regression methods. Additional data could be used as an independent test set to assess the performance of the predictive models developed in the thesis.

The set of variables considered in this thesis may not be adequate in characterising the features that affect the melting point across a wide range of compounds. One way to determine whether this is true would be to use compounds that are more diverse in molecular structure. Different classes of descriptors could be developed for this purpose, particularly 3D descriptors that are determined by molecular conformation, defined by the arrangement of the molecules in the crystal structure.

The models in the thesis only considered the inclusion of main effect predictors and two-way products of the linear predictors, representing two-factor interactions. These types of models could be extended to also include quadratic terms and three-way product terms, although the latter are often found to be less important in practice. The models currently considered are also linear in the unknown parameters. It may be beneficial to also consider more complex approximations to the response, for example, derived from any available physical theory.

When models with $N < (p + 1)$ were considered, reasonably good prediction accuracy was obtained for models fitted using the bridge regression methods, when measured both over the data used in developing the models, and also by cross-validation. Other modelling methods, such as regression trees, can be useful for performing variable selection in situations where $N < (p + 1)$ and we could perform a wider comparison of variable selection methods, including ridge regression and the lasso, by using simulation studies to investigate how the methods would fare in extreme cases, for example, where there are very highly correlated variables.

To obtain a critical comparison between the predictive models developed in this thesis and those found in the literature, the ridge regression and the lasso could

be applied to the same sets of compounds used to develop the models found in the literature, where these are available. The performance of the fitted ridge regression and lasso models would then be compared to the published model for the same data set.

6.2.2 Sequential Design

Ideally, the design methods presented here could be applied to real chemistry experiments. This would allow

(i) a practical assessment of the methods,

(ii) increased efficiency in the chemistry laboratory, through smaller and more informative experiments.

There are a number of possible extensions to the sequential design methodology. Design points could be selected using a maximum squared error criterion rather than a maximum prediction variance criterion. The squared error for the ith candidate point at step $k-1$ of the sequential procedure is defined as

$$\mathrm{Err}^2(\hat{Y}(\tilde{\boldsymbol{x}}))_{k-1} = (\bar{Y}(\tilde{\boldsymbol{x}}) - \hat{Y}_i)^2, \qquad i = 1, \ldots, N_{\mathcal{C}}, \tag{6.1}$$

where $\hat{Y}_i$ is the predicted value of the ith candidate point from the bridge regression model fitted to the design at step $k-1$ of the procedure, and

$$\bar{Y}(\tilde{\boldsymbol{x}}) = \frac{1}{B}\sum_{h=1}^{B} \hat{Y}_h^*(\tilde{\boldsymbol{x}}),$$

is the average of the predicted values of the ith candidate point from the B bootstrap samples. The candidate point chosen for inclusion in the design would be the one that exhibits the maximum value of (6.1). This criterion has been applied to the $N = 33$ Bayesian D-optimal ridge regression and lasso designs generated for the motivating Melting Point Data Set. Similar results were observed to those obtained using the sequential design criterion based on maximum prediction variance. It is expected that this criterion would have similar results to those of the maximum prediction variance criterion for all initial designs, with potentially larger differences for smaller designs.

The sequential design problem could be approached using an alternative criterion, where the data point would be chosen so as to give the best expected improvement in overall prediction (Cohn, 1996)

$$\tilde{\boldsymbol{x}} = \arg\min_{\tilde{\boldsymbol{x}} \in \mathcal{C}} \int_{\boldsymbol{x}} \mathrm{Var}(\hat{Y}(\tilde{\boldsymbol{x}}))d\boldsymbol{x},$$

where $\mathrm{Var}(\hat{Y}(\tilde{\boldsymbol{x}}))$ is approximated for $N+1$ observations. For nonlinear estimators, such as the lasso, this criterion would include an extra level of computational complexity, for example, using a nested bootstrapping algorithm. This sequential design criterion would then be compared to the one developed in this thesis by assessing their efficiencies across multiple stages via a simulation study.

In practice it is more efficient, in terms of time and materials, to synthesise a batch of compounds at any one time. Therefore it is important that we consider how to choose sequentially these multiple additional compounds. Construction of the new batches here cannot be carried out one point at a time, since once one compound has been selected for inclusion in the design and a model has been fitted to that design the whole prediction variance landscape will vary, potentially by a large amount. Hence, batches must be assessed as a whole, creating a considerable combinatorial problem.

Appendix A

Melting Point Data Set

Table A.1: Melting Point Data Set, variables Y and A-H

R^1	R^2	Y	A	B	C	D	E	F	G	H
F	F	101.94	20	269.29	94.37	25.41	1.553	3.030	46.169	209.148
F	Cl	146.22	20	285.74	96.95	27.70	1.569	3.545	46.169	217.753
F	Br	157.72	15	330.19	93.40	30.84	1.759	3.676	46.169	222.103
Cl	F	96.58	35	285.74	43.89	12.54	1.553	3.545	46.169	217.753
Cl	Cl	142.00	15	302.19	100.36	30.33	1.584	4.059	46.169	226.358
Cl	Br	162.65	15	346.64	98.98	34.31	1.786	4.190	46.169	230.707
Br	F	124.09	30	330.19	51.10	16.87	1.765	3.676	46.169	222.103
Br	Cl	126.35	20	346.64	64.10	22.22	1.765	4.190	46.169	230.707
NO_2	F	179.07	20	296.30	138.48	41.03	1.565	2.826	91.993	227.551
NO_2	Cl	179.31	30	312.75	130.00	40.66	1.621	3.340	91.993	236.156
NO_2	Br	203.89	25	357.20	115.09	41.11	1.809	3.471	91.993	240.505
F	NO_2	169.49	40	296.30	85.74	25.40	1.554	2.826	91.993	227.551
Cl	NO_2	158.24	25	312.75	105.25	32.92	1.613	3.340	91.993	236.156
NO_2	NO_2	144.64	20	323.31	134.33	43.43	1.644	2.621	137.817	245.954
I	F	144.93	35	377.19	74.17	27.98	1.941	3.950	46.169	228.207
I	Cl	170.97	20	393.64	92.65	36.47	1.927	4.464	46.169	236.812
Cl	I	173.27	20	393.64	85.76	33.76	1.950	4.464	46.169	236.812
Me	OMe	111.94	30	277.37	105.02	29.13	1.302	3.208	55.403	241.393
Cl	OMe	120.07	35	297.78	23.83	7.10	1.478	3.438	55.403	238.367
I	OMe	132.88	40	389.23	87.18	33.93	1.862	3.843	55.403	248.822

Table A.2: Melting Point Data Set, variables Y and A-H, continued

R^1	R^2	Y	A	B	C	D	E	F	G	H
NO_2	OMe	184.83	30	308.34	108.22	33.37	1.551	2.719	101.227	248.166
CN	OMe	140.12	50	288.35	105.08	30.30	1.370	2.515	79.195	241.691
OMe	H	101.34	25	263.34	46.41	12.22	1.389	2.760	55.403	224.832
OMe	OMe	91.85	30	293.37	55.55	16.30	1.458	2.816	64.637	250.377
H	CN	178.60	30	258.32	150.27	38.82	1.400	2.458	69.961	216.145
F	CN	139.87	20	276.31	123.87	34.23	1.458	2.622	69.961	221.077
Cl	CN	181.84	30	292.76	139.84	40.94	1.488	3.136	69.961	229.681
I	CN	206.75	20	384.21	103.31	39.69	1.872	3.541	69.961	240.135
H	NO_2	134.98	20	278.31	116.32	32.37	1.500	2.662	91.993	222.620
Me	NO_2	194.53	25	292.34	174.08	50.89	1.467	3.111	91.993	239.181
I	NO_2	220.22	15	404.20	115.91	46.85	1.981	3.745	91.993	246.610
NO_2	H	171.40	20	278.31	125.32	34.88	1.489	2.662	91.993	222.620
NO_2	Me	182.62	35	292.34	124.61	36.43	1.428	3.111	91.993	239.181
CF_3	OMe	138.45	15	331.34	105.69	35.02	1.546	3.655	55.403	256.129
OMe	CN	191.91	20	288.35	155.18	44.75	1.395	2.515	79.195	241.691
I	CF_3	168.45	20	427.20	63.89	27.29	1.870	4.682	46.169	254.573
F	OMe	105.47	15	281.33	97.40	27.40	1.447	2.924	55.403	229.763
OMe	F	82.21	30	281.33	74.84	21.05	1.468	2.924	55.403	229.763
OMe	I	109.93	20	389.23	32.60	12.69	1.810	3.843	55.403	248.822
H	CF_3	96.56	25	301.31	46.58	14.04	1.527	3.598	46.169	230.583
H	Cl	118.39	20	267.75	82.38	22.06	1.527	3.381	46.169	212.822
Me	Cl	128.76	25	281.78	92.90	26.18	1.427	3.830	46.169	229.383
Me	Br	150.78	25	326.23	108.02	35.24	1.611	3.961	46.169	233.732
Me	I	172.19	30	373.23	74.01	27.62	1.787	4.235	46.169	239.837
H	F	103.63	30	251.30	68.21	17.14	1.478	2.867	46.169	204.217
Me	H	99.07	20	247.34	112.29	27.77	1.320	3.152	46.169	215.847
Cl	H	94.56	20	267.75	88.05	23.58	1.444	3.381	46.169	212.822
F	Me	97.73	20	265.33	81.37	21.59	1.427	3.315	46.169	220.778
CN	F	153.58	35	276.31	107.88	29.81	1.540	2.622	69.961	221.077
Br	Me	98.67	15	326.23	44.21	14.42	1.636	3.961	46.169	233.732
Br	CN	177.13	15	337.21	116.27	39.21	1.734	3.267	69.961	234.031
Br	H	110.45	20	312.20	90.80	28.35	1.721	3.512	46.169	217.171
Br	OMe	142.14	15	342.23	102.44	35.06	1.691	3.569	55.403	242.717
CF_3	Br	166.99	10	380.20	101.13	38.45	1.816	4.408	46.169	248.469
CF_3	Me	124.13	20	315.34	77.90	24.56	1.530	4.047	46.169	247.144
CF_3	CN	168.62	25	326.32	116.64	38.06	1.613	3.353	69.961	247.443
Me	Me	116.77	15	261.37	122.24	31.95	1.359	3.600	46.169	232.408
Me	CN	174.39	15	272.35	128.51	35.00	1.429	2.907	69.961	232.706
Me	F	68.17	20	265.33	23.33	6.19	1.352	3.315	46.169	220.778
Cl	Me	88.73	20	281.78	78.21	22.04	1.483	3.830	46.169	229.383

Table A.3: Melting Point Data Set, variables I-Q

R^1	R^2	I	J	K	L	M	N	O	P	Q
F	F	236.95	18	5182.31	72.64	3.86	3.33	260.12	3222	163.60
F	Cl	245.92	18	5443.81	72.00	3.90	3.23	269.95	3230	169.85
F	Br	250.46	18	5610.05	71.26	3.84	3.24	274.62	3237	168.60
Cl	F	245.91	4	1222.13	71.27	3.80	3.81	269.94	3244	165.84
Cl	Cl	254.88	18	5701.59	71.46	3.84	3.65	279.77	3256	170.81
Cl	Br	259.42	18	5799.51	71.60	3.79	3.67	284.45	3232	174.12
Br	F	250.46	4	1242.19	71.52	4.04	3.74	274.65	3248	174.50
Br	Cl	259.42	2	652.03	70.77	4.10	3.57	284.44	3256	168.91
NO_2	F	254.04	4	1257.16	72.40	3.23	2.64	280.32	3238	176.81
NO_2	Cl	263.00	4	1281.13	73.73	3.00	3.10	290.13	3267	156.65
NO_2	Br	267.55	4	1311.34	73.36	3.25	3.03	294.81	3262	157.61
F	NO_2	254.04	4	1266.56	71.86	5.91	5.95	280.34	3236	174.57
Cl	NO_2	263.00	4	1287.73	73.36	5.84	6.61	290.15	3330	165.61
NO_2	NO_2	271.12	4	1306.41	75.31	2.43	2.81	300.49	3327	148.64
I	F	257.07	4	1290.87	70.71	4.11	3.90	280.93	3248	176.68
I	Cl	266.05	4	1356.59	69.83	4.19	3.73	290.77	3255	166.01
Cl	I	266.03	18	6032.58	70.66	3.78	3.64	290.74	3230	171.16
Me	OMe	273.69	2	707.58	68.23	3.55	3.17	299.48	3251	176.87
Cl	OMe	268.76	4	1338.34	71.24	2.82	2.50	294.69	3245	176.88
I	OMe	279.92	4	1388.42	71.68	2.80	2.54	305.67	3246	161.11
NO_2	OMe	276.87	4	1320.00	75.20	4.81	4.36	305.03	3275	158.53
CN	OMe	273.54	4	1397.62	69.17	4.22	3.14	298.67	3249	168.59
OMe	H	255.24	4	1244.58	72.26	4.39	3.94	279.18	3255	177.57
OMe	OMe	282.61	4	1345.95	74.41	2.98	2.87	309.53	3277	173.24
H	CN	246.22	4	1225.85	70.53	7.33	6.28	268.45	3248	172.12
F	CN	250.70	4	1258.83	70.25	5.41	4.72	273.96	3228	171.14
Cl	CN	259.66	4	1306.54	70.32	5.37	5.39	283.78	3228	169.67
I	CN	270.82	12	4089.69	70.46	5.88	5.51	294.77	3219	160.91
H	NO_2	249.55	4	1232.17	72.27	7.80	7.48	274.77	3331	171.80
Me	NO_2	267.94	4	1323.37	72.29	8.35	7.93	294.96	3333	171.35

Table A.4: Melting Point Data Set, variables I-Q, continued

R^1	R^2	I	J	K	L	M	N	O	P	Q
I	NO_2	274.17	4	1355.07	72.80	6.36	6.75	301.16	3250	163.39
NO_2	H	249.55	4	1241.36	71.73	4.88	4.16	274.79	3273	166.50
NO_2	Me	267.92	18	6118.62	70.36	5.29	4.47	294.91	3245	143.42
CF_3	OMe	288.35	4	1423.23	71.99	3.59	2.79	316.49	3249	172.21
OMe	CN	273.48	4	1372.58	70.43	7.72	6.77	298.54	3273	173.70
I	CF_3	285.62	4	1517.44	67.11	5.01	4.81	312.57	3258	157.28
F	OMe	259.79	4	1291.32	71.17	2.81	2.38	284.85	3240	170.74
OMe	F	259.75	4	1272.82	72.21	5.26	4.79	284.74	3245	173.91
OMe	I	279.87	18	6425.57	69.70	5.26	4.47	305.55	3230	165.81
H	CF_3	261.02	8	2621.35	70.37	6.43	5.57	286.21	3279	154.97
H	Cl	241.44	4	1164.50	73.10	5.41	4.36	264.42	3225	175.79
Me	Cl	259.82	18	5901.50	69.96	5.86	4.70	284.59	3230	169.71
Me	Br	264.36	18	6051.08	69.53	5.70	4.74	289.24	3225	170.08
Me	I	270.97	18	6242.16	69.16	5.70	4.68	295.55	3219	170.04
H	F	232.48	8	2258.20	72.35	5.29	4.59	254.61	3248	179.09
Me	H	246.36	4	1244.18	69.39	5.09	4.31	269.22	3233	172.88
Cl	H	241.42	4	1231.06	69.15	4.03	3.54	264.40	3252	175.42
F	Me	250.86	8	2469.00	71.54	4.22	3.39	274.78	3274	161.02
CN	F	250.71	8	2383.29	74.21	3.07	2.36	273.96	3254	170.93
Br	Me	264.36	4	1324.40	70.59	4.25	3.55	289.25	3269	167.90
Br	CN	264.19	4	1291.41	72.49	5.75	5.28	288.44	3225	161.82
Br	H	245.95	4	1205.18	72.08	4.10	3.49	269.05	3246	169.93
Br	OMe	273.31	4	1343.88	72.24	2.80	2.48	299.37	3247	167.16
CF_3	Br	279.02	2	695.33	71.47	2.98	2.53	306.26	3257	168.30
CF_3	Me	279.41	2	684.52	72.21	4.46	3.40	306.37	3252	167.69
CF_3	CN	279.25	8	2687.57	73.66	3.48	3.16	305.58	3227	170.52
Me	Me	264.75	2	638.68	72.78	5.07	4.27	289.36	3230	170.89
Me	CN	264.59	4	1265.69	73.54	7.87	6.71	288.58	3248	171.33
Me	F	250.85	2	651.51	67.77	5.70	4.94	274.76	3256	171.04
Cl	Me	259.82	2	630.97	72.71	4.21	3.59	284.58	3227	175.23

Table A.5: Melting Point Data Set, variables *R*-*U*

R^1	R^2	*R*	*S*	*T*	*U*
F	F	2.882	55.03	7	12
F	Cl	2.844	51.89	3	6
F	Br	2.843	49.07	3	6
Cl	F	2.911	69.57	7	9
Cl	Cl	2.860	50.68	5	6
Cl	Br	2.866	48.67	5	4
Br	F	2.916	67.59	6	10
Br	Cl	2.921	65.21	8	10
NO_2	F	2.943	57.50	12	20
NO_2	Cl	2.984	62.71	11	18
NO_2	Br	2.959	65.12	11	20
F	NO_2	2.957	63.03	11	22
Cl	NO_2	2.950	66.40	10	17
NO_2	NO_2	2.914	69.88	10	17
I	F	2.879	69.20	8	10
I	Cl	2.911	66.64	7	11
Cl	I	2.856	46.43	3	2
Me	OMe	2.986	70.56	6	4
Cl	OMe	2.917	67.87	10	16
I	OMe	2.905	68.72	8	10
NO_2	OMe	3.090	58.57	11	20
CN	OMe	2.925	67.53	6	10
OMe	H	2.933	72.20	8	6
OMe	OMe	3.035	58.30	11	19
H	CN	2.883	73.61	6	6
F	CN	3.052	70.52	6	6
Cl	CN	3.053	71.40	5	5
I	CN	3.010	83.22	5	8
H	NO_2	2.982	71.38	9	12
Me	NO_2	2.971	65.60	4	6

Table A.6: Melting Point Data Set, variables *R*-*U*, continued

R^1	R^2	*R*	*S*	*T*	*U*
I	NO_2	2.946	64.69	7	12
NO_2	H	3.090	61.90	6	10
NO_2	Me	3.041	56.47	6	7
CF_3	OMe	2.881	66.55	9	10
OMe	CN	2.991	56.76	10	20
I	CF_3	2.926	74.74	8	10
F	OMe	2.889	68.40	9	14
OMe	F	2.949	68.24	12	18
OMe	I	2.855	47.74	7	8
H	CF_3	3.015	55.30	6	11
H	Cl	2.904	65.36	8	8
Me	Cl	2.865	50.41	7	10
Me	Br	2.855	50.10	6	10
Me	I	2.852	47.78	5	9
H	F	2.898	59.98	6	8
Me	H	2.925	50.31	7	12
Cl	H	3.008	51.79	9	13
F	Me	2.930	54.82	9	13
CN	F	3.045	65.25	8	15
Br	Me	2.981	72.59	10	18
Br	CN	3.021	71.99	9	11
Br	H	2.915	74.20	8	7
Br	OMe	2.905	68.76	8	16
CF_3	Br	2.925	66.60	7	7
CF_3	Me	2.959	65.75	7	10
CF_3	CN	3.031	76.16	7	8
Me	Me	2.881	61.41	8	16
Me	CN	3.059	75.46	12	14
Me	F	3.014	73.89	8	8
Cl	Me	2.867	59.47	7	12

Appendix B

$N = 16$ Hadamard Matrices

Table B.1: $N = 16$ Hadamard matrix $\boldsymbol{C}_{16.\mathrm{I}}$ (Hall, 1961)

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1
1	1	1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1
1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1
1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1
1	1	-1	-1	1	1	-1	-1	-1	-1	1	1	-1	-1	1	1
1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	1	1
1	1	-1	-1	-1	-1	1	1	-1	-1	1	1	1	1	-1	-1
1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1
1	-1	1	-1	1	-1	1	-1	-1	1	-1	1	-1	1	-1	1
1	-1	1	-1	-1	1	-1	1	1	-1	1	-1	-1	1	-1	1
1	-1	1	-1	-1	1	-1	1	-1	1	-1	1	1	-1	1	-1
1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1
1	-1	-1	1	1	-1	-1	1	-1	1	1	-1	-1	1	1	-1
1	-1	-1	1	-1	1	1	-1	1	-1	-1	1	-1	1	1	-1
1	-1	-1	1	-1	1	1	-1	-1	1	1	-1	1	-1	-1	1

Table B.2: $N = 16$ Hadamard matrix $\boldsymbol{C}_{16.\mathrm{II}}$ (Hall, 1961)

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1
1	1	1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1
1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1
1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1
1	1	-1	-1	1	1	-1	-1	-1	-1	1	1	-1	-1	1	1
1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	1	1
1	1	-1	-1	-1	-1	1	1	-1	-1	1	1	1	1	-1	-1
1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1
1	-1	1	-1	1	-1	1	-1	-1	1	-1	1	-1	1	-1	1
1	-1	1	-1	-1	1	-1	1	1	-1	1	-1	-1	1	-1	1
1	-1	1	-1	-1	1	-1	1	-1	1	-1	1	1	-1	1	-1
1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	-1	1	1	-1
1	-1	-1	1	1	-1	-1	1	-1	1	1	-1	1	-1	-1	1
1	-1	-1	1	-1	1	1	-1	1	-1	-1	1	1	-1	-1	1
1	-1	-1	1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1

Table B.3: $N = 16$ Hadamard matrix $\boldsymbol{C}_{16.\mathrm{III}}$ (Hall, 1961)

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1
1	1	1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1
1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1
1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1
1	1	-1	-1	1	1	-1	-1	-1	-1	1	1	-1	-1	1	1
1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	1	1
1	1	-1	-1	-1	-1	1	1	-1	-1	1	1	1	1	-1	-1
1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1
1	-1	1	-1	1	-1	1	-1	-1	1	-1	1	-1	1	-1	1
1	-1	1	-1	-1	1	-1	1	1	-1	-1	1	1	-1	-1	1
1	-1	1	-1	-1	1	-1	1	-1	1	1	-1	-1	1	1	-1
1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	-1	1	1	-1
1	-1	-1	1	1	-1	-1	1	-1	1	1	-1	1	-1	-1	1
1	-1	-1	1	-1	1	1	-1	1	-1	1	-1	-1	1	-1	1
1	-1	-1	1	-1	1	1	-1	-1	1	-1	1	1	-1	1	-1

Table B.4: $N = 16$ Hadamard matrix $\boldsymbol{C}_{16.\mathrm{IV}}$ (Hall, 1961)

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1
1	1	1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1
1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1
1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1
1	1	-1	-1	1	1	-1	-1	-1	-1	1	1	-1	-1	1	1
1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1	1	1
1	1	-1	-1	-1	-1	1	1	-1	-1	1	1	1	1	-1	-1
1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1	1	-1
1	-1	1	-1	1	-1	-1	1	1	-1	-1	1	-1	1	-1	1
1	-1	1	-1	-1	1	1	-1	-1	1	-1	1	1	-1	-1	1
1	-1	1	-1	-1	1	-1	1	-1	1	1	-1	-1	1	1	-1
1	-1	-1	1	1	-1	1	-1	-1	1	-1	1	-1	1	1	-1
1	-1	-1	1	1	-1	-1	1	-1	1	1	-1	1	-1	-1	1
1	-1	-1	1	-1	1	1	-1	1	-1	1	-1	-1	1	-1	1
1	-1	-1	1	-1	1	-1	1	1	-1	-1	1	1	-1	1	-1

Table B.5: $N = 16$ Hadamard matrix $\boldsymbol{C}_{16.\mathrm{V}}$ (Hall, 1961)

0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
1	1	1	1	1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1
1	1	1	1	-1	-1	-1	-1	1	1	1	1	-1	-1	-1	-1
1	1	1	1	-1	-1	-1	-1	-1	-1	-1	-1	1	1	1	1
1	1	-1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1	-1
1	1	-1	-1	1	1	-1	-1	-1	-1	1	1	-1	-1	1	1
1	1	-1	-1	-1	-1	1	1	1	-1	1	-1	1	-1	1	-1
1	1	-1	-1	-1	-1	1	1	-1	1	-1	1	-1	1	-1	1
1	-1	1	-1	1	-1	1	-1	1	1	1	-1	-1	-1	1	1
1	-1	1	-1	1	-1	1	-1	-1	-1	1	1	1	1	-1	-1
1	-1	1	-1	-1	1	-1	1	1	-1	-1	1	-1	1	1	-1
1	-1	1	-1	-1	1	-1	1	-1	1	1	-1	1	-1	-1	1
1	-1	-1	1	1	-1	-1	1	1	-1	1	-1	-1	1	-1	1
1	-1	-1	1	1	-1	-1	1	-1	1	-1	1	1	-1	1	-1
1	-1	-1	1	-1	1	1	-1	1	-1	-1	1	1	-1	-1	1
1	-1	-1	1	-1	1	1	-1	-1	1	1	-1	-1	1	1	-1

Appendix C

Main Effects Orthogonal Designs

The $N = 16$ non-isomorphic main effects orthogonal designs for $f = 4, 5$ and 6 factors catalogued by Sun et al. (2002) are given below. Each design is made up of columns from one of the five $N = 16$ Hadamard matrices, defined in Appendix B. The roman numerals $\mathrm{I}, \mathrm{II}, \mathrm{III}, \mathrm{IV}, \mathrm{V}$ are used to indicate, for each MEO design, which Hadamard matrix the columns are taken from, e.g. I is Hadamard matrix $\boldsymbol{C}_{16.\mathrm{I}}$. The numbers in brackets indicate the columns of the Hadamard matrix that are used, e.g. $\mathrm{I}(1, 2, 3, 4)$ are the 1st, 2nd, 3rd and 4th columns of Hadamard matrix $\boldsymbol{C}_{16.\mathrm{I}}$.

Designs with $f = 4$ factors

Design 1: $\mathrm{I}(1, 2, 3, 4)$
Design 2: $\mathrm{I}(1, 2, 4, 7)$
Design 3: $\mathrm{I}(1, 2, 4, 8)$
Design 4: $\mathrm{II}(1, 4, 8, 12)$
Design 5: $\mathrm{II}(4, 5, 8, 12)$

Designs with $f = 5$ factors

Design 1: $\mathrm{I}(1, 2, 3, 4, 5)$
Design 2: $\mathrm{I}(1, 2, 3, 4, 8)$
Design 3: $\mathrm{I}(1, 2, 4, 7, 8)$
Design 4: $\mathrm{I}(1, 2, 4, 8, 15)$
Design 5: $\mathrm{II}(1, 2, 4, 8, 12)$
Design 6: $\mathrm{II}(1, 4, 5, 8, 12)$
Design 7: $\mathrm{II}(1, 4, 6, 8, 12)$
Design 8: $\mathrm{II}(4, 5, 6, 8, 12)$
Design 9: $\mathrm{II}(4, 5, 8, 9, 12)$
Design 10: $\mathrm{II}(4, 5, 8, 10, 12)$
Design 11: $\mathrm{III}(2, 4, 8, 10, 12)$

Designs with $f = 6$ factors

Design 1: I$(1, 2, 3, 4, 5, 6)$
Design 2: I$(1, 2, 3, 4, 5, 8)$
Design 3: I$(1, 2, 3, 4, 8, 12)$
Design 4: I$(1, 2, 3, 4, 8, 13)$
Design 5: I$(1, 2, 4, 7, 8, 11)$
Design 6: II$(1, 2, 3, 4, 8, 12)$
Design 7: II$(1, 2, 4, 5, 8, 12)$
Design 8: II$(1, 2, 4, 7, 8, 12)$
Design 9: II$(1, 4, 5, 6, 8, 12)$
Design 10: II$(1, 4, 5, 8, 9, 12)$
Design 11: II$(1, 4, 5, 8, 10, 12)$
Design 12: II$(1, 4, 6, 8, 10, 12)$
Design 13: II$(1, 4, 6, 8, 11, 12)$
Design 14: II$(4, 5, 6, 7, 8, 12)$
Design 15: II$(4, 5, 6, 8, 9, 12)$
Design 16: II$(4, 5, 8, 9, 12, 13)$
Design 17: II$(4, 5, 8, 9, 12, 14)$
Design 18: II$(4, 5, 8, 10, 12, 15)$
Design 19: III$(1, 2, 4, 8, 10, 12)$
Design 20: III$(2, 3, 4, 8, 10, 12)$
Design 21: III$(2, 4, 6, 8, 10, 12)$
Design 22: III$(2, 4, 7, 8, 10, 12)$
Design 23: III$(2, 4, 8, 9, 10, 12)$
Design 24: III$(2, 4, 8, 9, 10, 14)$
Design 25: III$(2, 4, 8, 10, 12, 14)$
Design 26: III$(2, 4, 8, 10, 12, 15)$
Design 27: IV$(2, 4, 6, 8, 10, 12)$

Appendix D

Aliasing Structure of Main Effects Orthogonal Designs

The aliasing structure of the $N = 16$, $f = 4, 5, 6$ main effect orthogonal designs, defined in Appendix C, are given in Tables D.1, D.2 and D.3 respectively. Columns headed 'M.E.-Int.' indicate whether the design has aliasing between main effect and two-factor interaction terms; columns headed 'Int.-Int.' indicate whether the design has aliasing between interaction terms. For regular designs, the resolution of the design is also given.

Each design was evaluated under the ridge regression Bayesian D-optimality objective function (4.15) and the lasso Bayesian D-optimality objective function (4.16). For each of $f = 4, 5, 6$, and for each of the two objective functions, the designs are ranked in order of decreasing performance, i.e. the design ranked 1 exhibits the highest value of the objective function. Recall that for the lasso, the main effects are the primary terms, and the two-factor interactions are the potential terms.

Table D.1: Aliasing structure and performance ranking of MEO designs with $f = 4$ factors

	Rank under	Rank under	Full aliasing		Partial aliasing		
Design	ridge	lasso	M.E.-Int.	Int.-Int.	M.E.-Int.	Int.-Int.	Resolution
1	4	5	×				III
2	4	4		×			IV
3	1	1					V
4	2	2			×		
5	2	3			×		

Table D.2: Aliasing structure and performance ranking of MEO designs with $f = 5$ factors

	Rank under	Rank under	Full aliasing		Partial aliasing		
Design	ridge	lasso	M.E.-Int.	Int.-Int.	M.E.-Int.	Int.-Int.	Resolution
1	11	11	×	×			III
2	6	7	×				III
3	6	6		×			IV
4	1	1					V
5	2	2			×		
6	9	10	×		×	×	
7	4	4			×		
8	2	3			×		
9	9	9		×	×		
10	4	5			×		
11	8	8			×	×	

Table D.3: Aliasing structure and performance ranking of MEO designs with $f = 6$ factors

	Rank under	Rank under	Full aliasing		Partial aliasing		
Design	ridge	lasso	M.E.-Int.	Int.-Int.	M.E.-Int.	Int.-Int.	Resolution
1	27	27	×	×			III
2	20	21	×	×			III
3	1	11	×				III
4	1	4	×	×			III
5	20	20		×			IV
6	1	5	×		×	×	
7	11	13	×		×	×	
8	11	2		×	×	×	
9	11	14	×		×	×	
10	25	26	×	×	×	×	
11	17	19	×		×	×	
12	17	15		×	×	×	
13	1	1			×	×	
14	1	3		×	×		
15	11	9		×	×		
16	25	25		×	×		
17	17	17		×	×		
18	1	10			×		
19	7	6			×	×	
20	7	8			×	×	
21	23	24	×		×	×	
22	15	16			×	×	
23	7	12			×	×	
24	7	7			×	×	
25	23	23		×	×	×	
26	15	18			×	×	
27	22	22			×	×	

References

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

Atkinson, A.C., Donev, A.N., Tobias, R.D. (2007). *Optimum Experimental Designs, with SAS, 2nd Edition*. Oxford: Oxford University Press.

Azencott, C.A., Ksikes, A., Swamidass, S.J., Chen, J.H., Ralaivola, L., Baldi, P. (2007). One- to four-dimensional kernels for virtual screening and the prediction of physical, chemical and biological properties. *Journal of Chemical Information and Modeling*, **47**, 965-974.

Bakker, B., Heskes, T. (2003). Clustering ensembles of neural network models. *Neural Networks*, **16**, 261-269.

Bhat, A.U., Merchant, S.S., Bhagwat, S.S. (2008). Prediction of melting points of organic compounds using extreme learning machines. *Industrial and Engineering Chemistry Research*, **47**, 920-925.

Burnham, K.P., Anderson, D.R. (2002). *Model Selection and Multimodel Inference, 2nd Edition*. New York: Springer.

Candes, E., Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than N. *The Annals of Statistics*, **35**, 2313-2351.

Chaloner, K., Verdinelli, I. (1995). Bayesian experimental design: a review. *Statistical Science*, **10**, 273-304.

Cohn, D.A. (1996). Neural network exploration using optimal experiment design. *Neural Networks*, **9**, 1071-1083.

Cook, R.D., Nachtsheim, C.J. (1980). A comparison of algorithms for constructing exact D-optimal designs. *Technometrics*, **22**, 315-324.

Craven, P., Wahba, G. (1979). Smoothing noisy data with spline functions. *Numerische Mathematik*, **31**, 377-403.

Dannenfelser, R.M., Yalkowsky, S.H. (1996). Estimation of entropy of melting from molecular structure: a non-group contribution method. *Industrial and Engineering Chemistry Research*, **35**, 1483-1486.

Davison, A.C., Hinkley, D.V. (1997). *Bootstrap Methods and their Application*. Cambridge: Cambridge University Press.

Deng, X., Joseph, V.R., Sudjianto, A., Wu, C.F.J. (2009). Active learning through sequential design, with applications to detection of money laundering. *Journal of the American Statistical Association*, **104**, 969-981.

Ding, Y. (2007). *Structure-property relationships in organic solid forms: structures and melting points*. PhD Thesis. University of Southampton, School of Chemistry.

Draper, N.R., Smith, H. (1998). *Applied Regression Analysis, 3rd Edition*. New York: Wiley.

DuMouchel, W., Jones, B. (1994). A simple Bayesian modification of D-optimal designs to reduce dependence on an assumed model. *Technometrics*, **36**, 37-47.

Dyekær, J.D., Jónsdóttir, S.Ó. (2004). QSPR models for various physical properties of carbohydrates based on molecular mechanics and quantum chemical calculations. *Carbohydrate Research*, **339**, 269-280.

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

Efron, B., Tibshirani, R.J. (1993). *An Introduction to the Bootstrap*. London: Chapman and Hall.

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

Frank, I.E., Friedman, J.H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, **35**, 109-135.

Fu, W.J. (1998). Penalized regressions: the bridge versus the lasso. *Journal of Computational and Graphical Statistics*, **7**, 397-416.

Fu, W.J. (2005). Nonlinear GCV and quasi-GCV for shrinkage models. *Journal of Statistical Planning and Inference*, **131**, 333-347.

Gelbrich, T., Hursthouse, M.B., Threlfall, T.L. (2007). Structural systematics of 4,4'-disubstituted benzenesulfonamidobenzenes. 1. Overview and dimer-based isostructures. *Acta Crystallographica Section B*, **63**, 621-632.

Gelman, A., Carlin, J.B., Stern, H.S., Rubin, D.B. (2004). *Bayesian Data Analysis, 2nd Edition*. Boca Raton, Florida: Chapman and Hall/CRC.

Godavarthy, S.S., Robinson Jr., R.L., Gasem, K.A.M. (2006). An improved structure-property model for predicting melting-point temperatures. *Industrial and Engineering Chemistry Research*, **45**, 5117-5126.

Habibi-Yangjeh, A., Pourbasheer, E., Danandeh-Jenagharad, M. (2008). Prediction of melting point for drug-like compounds using principal component-genetic algorithm-artificial neural network. *Bulletin of the Korean Chemical Society*, **29**, 833-841.

Hall, M.J. (1961). Hadamard matrices of order 16. *Research Summary No. 36-10, Jet Propulsion Laboratory, Pasadena, California*, **1**, 21-26.

Hastie, T., Efron, B. (2007). *lars: least angle regression, lasso and forward stagewise*. R package version 0.9-7.

Hastie, T., Tibshirani, R., Friedman, J. (2009). *The Elements of Statistical Learning, 2nd Edition*. New York: Springer.

Hoerl, A.E., Kennard, R.W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55-67.

Hsu, N.J., Hung, H.L., Chang, Y.M. (2008). Subset selection for vector autoregressive processes using lasso. *Computational Statistics and Data Analysis*, **52**, 3645-3657.

Hughes, L.D., Palmer, D.S., Nigsch, F., Mitchell, J.B.O. (2008). Why are some properties more difficult to predict than others? A study of QSPR models of solubility, melting point and log p. *Journal of Chemical Information and Modeling*, **48**, 220-232.

Hurvich, C.M., Tsai, C.L. (1989). Regression and time series model selection in

small samples. *Biometrika*, **76**, 297-307.

Jain, A., Yalkowsky, S.H. (2006). Estimation of melting points of organic compounds-II. *Journal of Pharmaceutical Sciences*, **95**, 2562-2618.

Jain, A., Yalkowsky, S.H. (2007). Comparison of two methods for estimation of melting points of organic compounds. *Industrial and Engineering Chemistry Research*, **46**, 2589-2592.

Jain, A., Yang, G., Yalkowsky, S.H. (2004). Estimation of melting points of organic compounds. *Industrial and Engineering Chemistry Research*, **43**, 7618-7621.

James, G.M., Radchenko, P., Lv, J. (2009). DASSO: connections between the Dantzig selector and lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **71**, 127-142.

Jones, B., Lin, D.K.J., Nachtsheim, C.J. (2008). Bayesian D-optimal supersaturated designs. *Journal of Statistical Planning and Inference*, **138**, 86-92.

Karthikeyan, M., Glen, R.C., Bender, A. (2005). General melting point prediction based on a diverse compound data set and artificial neural networks. *Journal of Chemical Information and Modeling*, **45**, 581-590.

Keshavarz, M.H. (2006). Approximate prediction of melting point of nitramines, nitrate esters, nitrate salts and nitroaliphatics energetic compounds. *Journal of Hazardous Materials*, **138**, 448-451.

Lawson, C., Hansen, R. (1974). *Solving Least Squares Problems*. Englewood Cliffs: Prentice Hall.

Lu, W., Zhang, H.H. (2007). Variable selection for proportional odds model. *Statistics in Medicine*, **26**, 3771-3781.

Lumley, T., Miller, A. (2009). *leaps: regression subset selection*. R package version 2.9.

MacKay, D.J.C. (1992). Information-based objective functions for active data selection. *Neural Computation*, **4**, 589-603.

Mandal, A., Ranjan, P., Wu, C.F.J. (2009). G-SELC: optimization by sequential elimination of level combinations using genetic algorithms and Gaussian processes. *Annals of Applied Statistics*, **3**, 398-421.

Mandal, A., Wu, C.F.J., Johnson, K. (2006). SELC: sequential elimination of level combinations by means of modified genetic algorithms. *Technometrics*, **48**, 273-283.

Marley, C.J., Woods, D.C. (2010). A comparison of design and model selection methods for supersaturated experiments. *Computational Statistics and Data Analysis*, **54**, 3158-3167.

Meier, L., van der Geer, S.A., Buhlmann, P. (2008). The group lasso for logistic regression. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, **70**, 53-71.

Meyer, R.K., Nachtsheim, C.J. (1995). The coordinate-exchange algorithm for constructing exact optimal experimental designs. *Technometrics*, **37**, 60-69.

Miller, A.J. (2002). *Subset Selection in Regression, 2nd Edition.* Boca Raton, Florida: Chapman and Hall/CRC.

Modarresi, H., Dearden, J.C., Modarress, H. (2006). QSPR correlation of melting point for drug compounds based on different sources of molecular descriptors. *Journal of Chemical Information and Modeling*, **46**, 930-936.

Molecular Operating Environment. Chemical Computing Group Inc., Montreal, Quebec, Canada. URL http://www.chemcomp.com.

Nigsch, F., Bender, A., van Buuren, B., Tissen, J., Nigsch, E., Mitchell, J.B.O. (2006). Melting point prediction employing k-nearest neighbor algorithms and genetic parameter optimization. *Journal of Chemical Information and Modeling*, **46**, 2412-2422.

O'Boyle, N.M., Palmer, D.S., Nigsch, F., Mitchell, J.B.O. (2008). Simultaneous feature selection and parameter optimisation using an artificial ant colony: case study of melting point prediction. *Chemistry Central Journal*, **2**:21.

Osborne, M.R., Presnell, B., Turlach, B.A. (2000). On the lasso and its dual. *Journal of Computational and Graphical Statistics*, **9**, 319-337.

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

Phao, F.K.H., Pan, Y.H., Xu, H. (2009). Analysis of supersaturated designs via the Dantzig selector. *Journal of Statistical Planning and Inference*, **139**, 2362-2372.

R Development Core Team (2007). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Rockafellar, R.T. (1970). *Convex Analysis*. Princeton, New Jersey: Princeton University Press.

Stamey, T., Kabalin, J., McNeal, J., Johnstone, J., Freiha, F., Redwine, E., Yang, N. (1989). Prostate specific antigen in the diagnosis and treatment of adenocarcinoma of the prostate, ii: radical prostatectomy treated patients. *The Journal of Urology*, **16**, 1076-1083.

Sun, D.X., Li, W., Ye, K.Q. (2002). *An algorithm for sequentially constructing non-isomorphic orthogonal designs and its applications*. Technical Report. SUNYSB-02-13, State University of New York at Stony Brook.

Tetko, I.V., Gasteiger, J., Todeschini, R., Mauri, A., Livingstone, D., Ertl, P., Palyulin, V.A., Radchenko, E.V., Zefirov, N.S., Makarenko, A.S., Tanchuk, V.Y., Prokopenko, V.V. (2005). Virtual computational chemistry laboratory - design and description. *Journal of Computer-Aided Molecular Design*, **19**, 453-63.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, **58**, 267-288.

Wang, Q., Ma, P., Neng, S. (2009). Position group contribution method for estimation of melting point of organic compounds. *Chinese Journal of Chemical Engineering*, **17**, 468-472.

Wu, C.F.J., Hamada, M. (2009). *Experiments: Planning, Analysis, and Parameter Design Optimization, 2nd Edition*. New York: Wiley.

Yuan, M., Joseph, V.R., Lin, Y. (2007). An efficient variable selection approach for analyzing designed experiments. *Technometrics*, **49**, 430-439.

Zhou, D., Alelyunas, Y., Liu, R. (2008). Scores of extended connectivity fingerprint as descriptors in QSPR study of melting point and aqueous solubility. *Journal of Chemical Information and Modeling*, **48**, 981-987.

Zou, H., Hastie, T., Tibshirani, R. (2007). On the 'degrees of freedom' of the lasso. *The Annals of Statistics*, **35**, 2173-2192.