Multivariate response predictor selection methods: With applications to telecommunications time series data

Aaron Paul Lowther, B.Sc.(Hons.), M.Res

Submitted for the degree of Doctor of Philosophy at Lancaster University.

September 2019

STOR-i
excellence with impact

Dedicated to my Grandad, who sadly passed away before I was able to complete this thesis.

Abstract

This thesis looks at developing a semi-automated approach to estimate multiple, sparse, linear regression models simultaneously. We are motivated by a telecommunications application and aim to produce interpretable models.

Firstly, we generalise the best-subset problem (Miller, 1996) which is often used to estimate sparse linear regression models. We call our problem the *Simultaneous Best-Subset (SBS) problem* and use it to simultaneously estimate multiple linear regression models. The so-called SBS approach produces models that perform more favourably in comparison to models estimated individually using the best-subset approach. We solve the SBS problem by formulating a Mixed Integer Quadratic Optimisation (MIQO) program which can often be solved quickly using an optimisation solver. The MIQO framework allows us to have some control over the regression models estimated which is desirable in an automated setting.

Secondly, we propose a simultaneous shrinkage operator. This operator shrinks coefficients between models towards a common value. We show that this operator can further improve parameter estimation when simultaneously estimating multiple linear regression models. This operator was found to be particularly useful when noisy predictors entered the models.

Thirdly, we show how the SBS approach can be integrated into a two-step *semi-automated* procedure for fitting REGression Seasonal AutoRegressive Integrated Moving Average (Reg-SARIMA) models. We apply this automated approach to estimate models for a telecommunications dataset and compare it to the current approach employed by our industrial collaborator. We show how the Reg-SARIMA models provide a better fit to the data, are more interpretable, and perform more favourably for future short-term predictions. In addition to this, the two-step procedure requires much less human intervention into the modelling procedure than procedures currently used by industry.

Finally, we propose fast approaches to simultaneously estimate multiple sparse linear regression models. Using a simulation study we show that these approaches often produce models that perform as favourably as the SBS approach, despite producing models in far less time.

Acknowledgements

Firstly, I would like to express huge gratitude to my supervisors, Dr Matt Nunes, Professor Paul Fearnhead, and Dr Kjeld Jensen, for their patience and guidance over the past years. I am very grateful for their commitment towards my research project, the time that they have given me and for all that they have taught me.

I would also like to thank all STOR-i staff and students, past and present, for making the CDT a stimulating, warm, and friendly environment to study. In particular, I would like to thank Professors Jonathan Tawn and Idris Eckley for the time and effort they invest into making the CDT such a success. Cyrus Gaviri and Dave Sole from the Mathematics & Statistics department have helped me overcome countless computing difficulties, and for this they deserve a special thanks for their exceptional IT support. I am very grateful for the funding provided by the EPSRC and BT, and for the helpful suggestions provided by my examiners, Dr Ines Wilms, Dr Claire Gormley, Dr Marco Battiston and (Chair) Professor Jonathan Tawn, that improved the initial version of this thesis.

My final thanks is to my family. My parents, grandparents, and sister, have always encouraged me throughout my academic pursuits. Their support has given me the inspiration, commitment and confidence to pursue my goals. Finally I wish to thank my wife, Dr Jamie-Leigh Chapman for her love and support especially during the past few months.

Declaration

I declare that the work in this thesis has been done by myself and has not been submitted elsewhere for the award of any other degree.

Aaron Paul Lowther

A condensed version of Chapter 3 has been submitted to *Statistics & Computing* as: Lowther, A. P., Fearnhead, P., Nunes, M. A. and Jensen, K. (2019). Semi-automated simultaneous predictor selection for Regression-SARIMA models.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The work in this thesis considers estimating statistical models that are suitable for a range of industrial applications. Suitable applications include scenario's where multiple linear models can be estimated simultaneously and particularly when similarity may be expected across the models. In addition to this, the data can be time ordered and our approach is able to select suitable predictors that can be used to explain the variation observed in a response variable. We apply our methodology to an industrial dataset provided by our industrial collaborator, BT, to better understand how weather variables affect the rate of telecommunication events. Our methodology has also been applied by BT to understand how electricity consumption for different types of telecommunication buildings can be affected by weather variables. Suitable applications from the statistical literature include understanding *commodity dynamics* (Barbaglia et al., 2016) and modelling sales data (Wilms et al., 2018), where the effects of multiple predictor variables are expected to affect sales across multiple stores similarly.

One important aspect of producing statistical models for industry is model *interpretability*. Models that are interpretable are often simple and can support or provide an explanation relating the external (**predictor**) variables to the (**response**) variable of interest. We achieve model interpretability in the following ways. Firstly, we estimate multiple models simultaneously for *related* response variables. We encourage the models to include effects from *similar* predictor variables, which is expected in practise. This is not always achieved using current procedures due to the challenging modelling conditions primarily caused by highly correlated predictor variables. As well as improving model interpretability we show that simultaneous model estimation can greatly improve model selection and estimation accuracy. Secondly, we fit *sparse* models that include the effects of only the most important predictor variables. Finally, we exploit expert knowledge at the model estimation stage to ensure the effects of external variables are in agreement with this expert knowledge.

This greatly reduces the human intervention required to ensure models are interpretable.

By reducing the amount of intervention needed to produce interpretable models we have developed a *semi-automated* approach that can select important predictors to include into models and estimate their effects. Traditional *manual approaches* to statistical modelling, whereby an analyst produces models *by hand*, are becoming infeasible due to the amount of related data that is routinely recorded. We apply our approach to an example where the number of predictors is in the thirties and the automated nature allows it to be applied to many groups of *related* response variables. The maximum number of models estimated simultaneously in our application is seven, although this and the number of predictors considered can be larger, but subject to increased computational time.

The main contribution of this thesis is the development of multiple simultaneous predictor selection methods. We investigate how a shrinkage operator, only available when jointly fitting models, can improve parameter estimation. By implementing our methods using mixed integer quadratic optimisation techniques, we can estimate the models easily and ensure they demonstrate desirable properties. We consider a two-step procedure that can be used to select predictors for our models and account for the serial correlation often observed in the response variables. Finally, we produced a statistical software package during this project which has had significant impact in industry. The package has allowed BT to seamlessly integrate our work into their systems and produce large numbers of sensible models with minimal effort. These models can be used as excellent baseline models and compared to models produced by hand, requiring significantly more effort.

1.1 Telecommunications event dataset

Our industrial collaborator, BT own and are responsible for maintaining much of the UK's telecommunications network. Statistical models are used by BT to better understand how the network behaves. Specifically, statistical models can be used to quantify the impact of external influences on the network. In addition to this, predictions from these models can also be used to plan effectively by efficiently allocating resources. In Chapter 3 and Chapter 4 we apply our methodology to the *telecommunications event dataset*, provided to us by BT. This dataset provides a good example of the modelling challenges often encountered by researchers within the organisation. In this section we provide details of the telecommunications event dataset to motivate much of the methodology developed in this thesis.

The telecommunication event dataset records the daily *event rates* by *type* at a given *location* in the network. Figure 1.1.1 shows the daily event rate for a particular event type and location for approximately 3 years and 8 months. A telecommunication event may correspond to a problem with

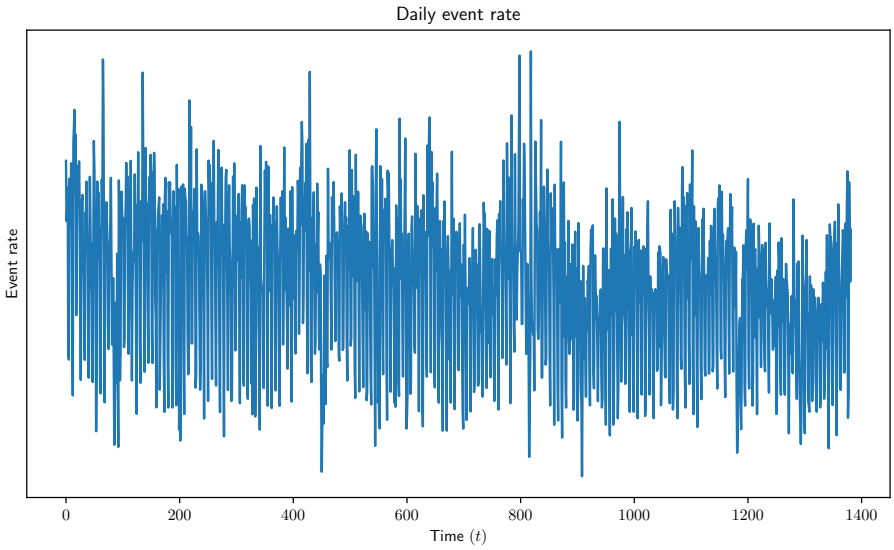


Figure 1.1.1: A daily time series plot of telecommunications event rate data. We do not show the vertical scale or the time window in the interest of anonymising the data.

a service provided by the network such as an interrupted service. The measure is an *event rate* as the daily number of events is scaled by the daily number of *active services*. This is to ensure that we take into account the number of active services, which changes on a daily basis, when considering the number of events. In **this dataset** the *location* corresponds to a specific region in the United Kingdom. Other telecommunication datasets could identify a location specific to a component in the network, in contrast to a geographical location. In the interest of developing a method which is widely applicable, we focus on estimating models that do not explicitly use *spatial information*. The event *type* identifies the particular problem with a service. Given that BT have a national network it's obvious that the task of producing excellent models for all types of events across many locations poses a significant challenge to the industry.

It is known by experts that events of the same type are influenced by the same weather variables, no matter the location in the network. However, it is thought that the effect of the weather variables on events may differ between locations. For example, consider an event type that is known to be influenced by precipitation. The effects of precipitation on this event type may be more similar between regions in Scotland than compared to one region from Scotland and one region from London. This difference could be explained by the geography of the two locations. Figure 1.1.2 shows that there may exist strong correlation for suitably grouped events. The similarity between events within a group combined with expert opinion suggests that it may be advantageous to jointly model groups of events. We do not consider the problem of determining suitable groups and leave this to the expert judgement of our industrial collaborator.

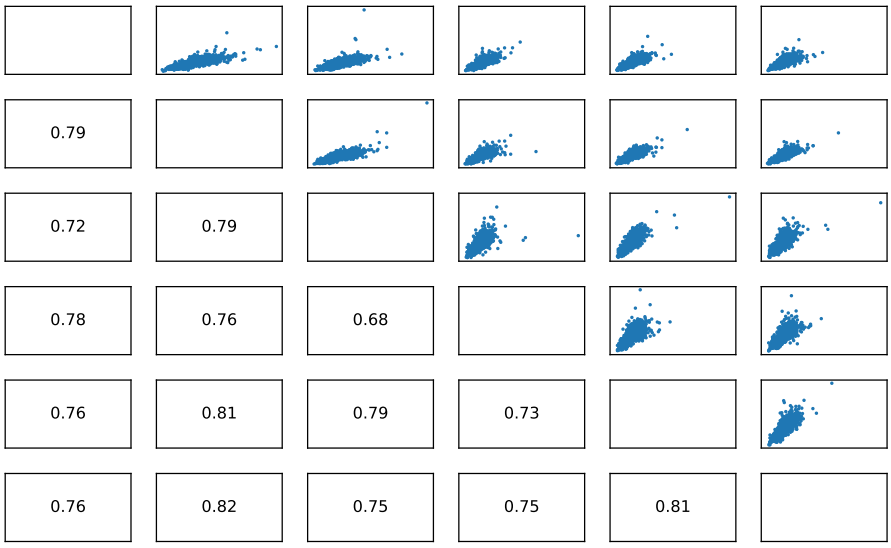


Figure 1.1.2: Correlation between suitably grouped events. The upper-right grids show the pairwise scatter-plots of the events between the six locations within the group for a fixed event type. The lower-left grid shows the Pearson correlation coefficient between the events shown in the scatter-plots reflected across the main diagonal.

Complex relationships may exist between response and predictor variables. Expert knowledge can often provide appropriate non-linear transformations that reveal these relationships, but the parameters of these transformations are typically not known. Reasonable estimates of the transformation parameters may be obtained by selecting the *best* predictor among a *group of predictors*. We make the following distinction between a group of predictor variables and a group of response variables,

- **Group of predictors:** A set of predictors that are produced by applying a transformation to an observed *base predictor* for multiple values of a transformation parameter.
- **Group of response variables:** A set of response variables that are deemed suitable for joint analysis due to the similarity in their behaviour or physical properties.

By using a fine grid of parameters it may be possible to obtain an accurate estimate of the transformation and it's associated parameter. A fine grid of parameter values can lead to highly correlated predictors, as shown in Figure 1.1.3. Here, we observe the correlation between pairs of predictors from a group of predictors. These predictors were obtained by smoothing precipitation across a grid of smoothing parameter values. Clearly, including all smoothed precipitation predictors in a model for telecommunication events will tell us very little about the relationship between events and precipitation. However, if we can identify a single smoothed precipitation predictor (among other

predictors) that can adequately explain the behaviour of telecommunication events we are likely to gain from a much better understanding of a relationship.

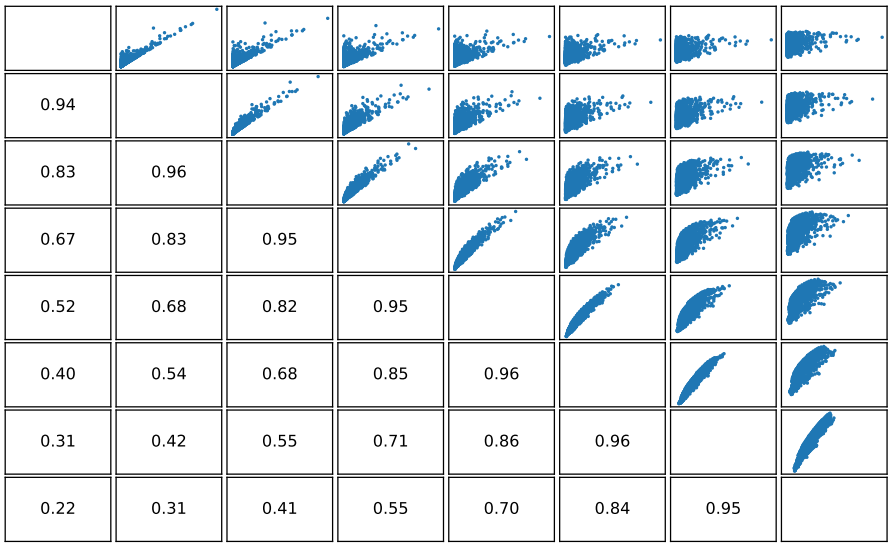


Figure 1.1.3: Correlation amongst predictors within a group of predictors. The upper-right grids show the pairwise scatter-plots between the eight predictors within the group produced by smoothing the precipitation observations. The lower-left grid shows the Pearson correlation coefficient between the predictors shown in the scatter-plots reflected across the main diagonal.

Building statistical models that adequately explain the physical relationship between a response variable and predictor variables can be challenging. It is important to obtain simple models that can be interpreted easily and these models should describe a relationship that aligns with expert judgement and opinion. Currently, great effort is required to obtain interpretable models. Highly correlated predictors present challenging conditions to select the *best* predictors and estimate their effects. Often with the current procedure, pairs of highly correlated predictors are selected for a model. The problem here is that one predictor appears to have a large positive effect on the response, and the other a large negative effect. The effect of these predictors may effectively cancel, which makes it hard to interpret the resulting model.

The event data described in this section is recorded daily. This means that the event dataset is time series data. Daily time series typically exhibit seasonality (Hyndman and Athanasopoulos, 2019). This can be identified in Figure 1.1.4 which shows an estimate of the auto-correlation function for the event data presented in Figure 1.1.1. A large peak at a lag l in the auto-correlation function indicates that the values separated by l days are often highly correlated. The repeated pattern in the auto-correlation function every 7 lags indicates the presence of weekly seasonality. In practice, large peaks at lag seven suggests that the events on each day of the week are similar to that of the

same day on the previous week. Seasonality may be a characteristic of a response variable that is not induced by a predictor of interest. For example, it is unlikely that the weekly seasonality present in the event data is caused by weather variables, and thus, the seasonality could be explicitly included into the models. In Chapter 2 we discuss how BT currently estimate and remove the seasonality observed in response variables in order to reveal a relationship and discuss the drawbacks of this approach.

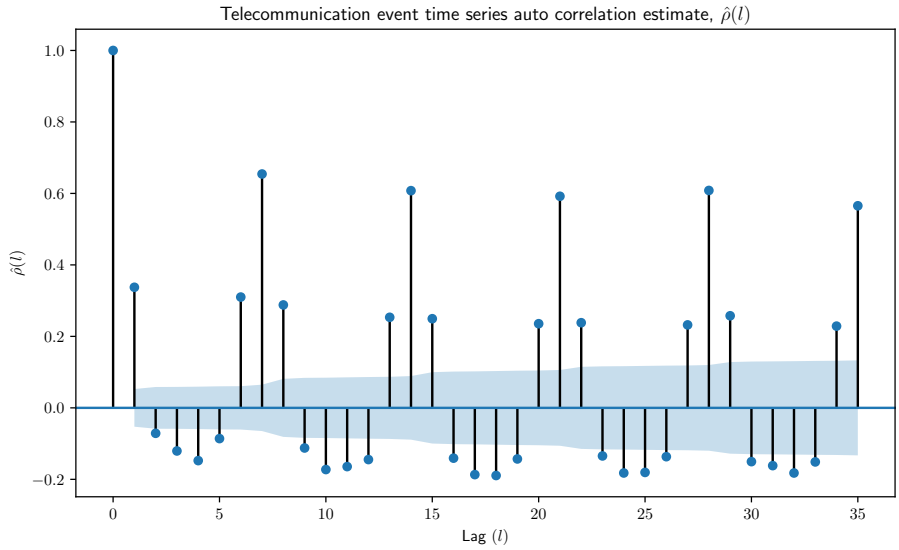


Figure 1.1.4: An estimate of the auto-correlation function for the telecommunications event rate data. The vertical lines show an estimate of the auto-correlation at lag l. The uncertainty cloud shows the 95% confidence intervals calculated using Bartlett's formula.

The events considered in the telecommunications event dataset are reported by customers. As such, fewer events are typically observed on UK bank holidays[1]. The bank holiday effect can not easily be seen by eye in Figure 1.1.1. However, by removing the weekly seasonality from the event data low events on bank holidays are made very clear. Figure 1.1.5 shows the events smoothed using a seven day symmetric moving average. This moving average calculates the 7-point running mean using one value from each day of the week, in doing so this smooths out the between day variation.

In this section we have highlighted a number of characteristics of this dataset that are also present in many of the telecommunications datasets. It is important to take these into consideration when estimating statistical models. We have also discussed a number of the problems BT encounter when producing models. Using the typical characteristics of the data and the problems often encountered we are able to create a list of requirements needed for a modelling approach. Firstly, we require a predictor selection algorithm that can perform favourably when many of the predictors are highly

[1]We note here that bank holidays may differ between Scotland, England, Northern Ireland and Wales.

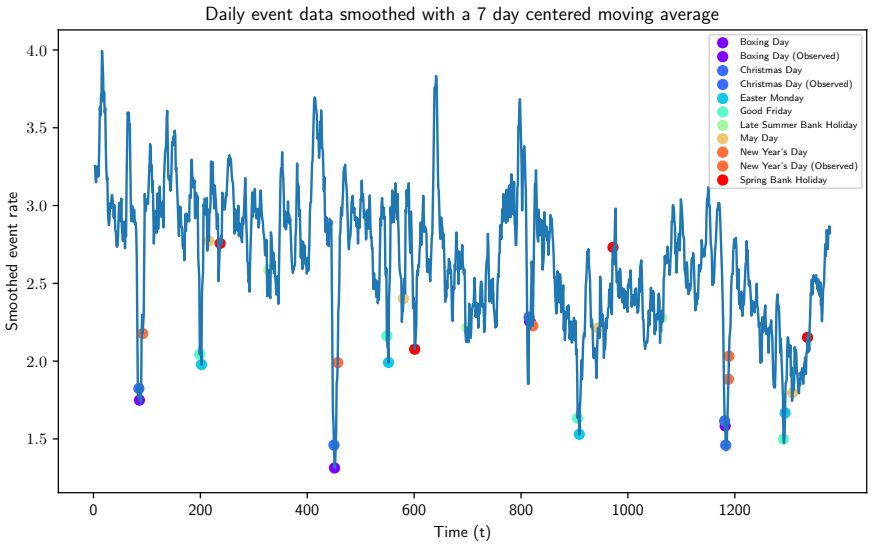


Figure 1.1.5: Daily event rates smoothed using a 7 day symmetric moving average. Low counts are often identified on bank holidays which are indicated by the coloured circles. An observed bank holiday indicates the additional bank holiday given in lieu of a bank holiday that falls on a weekend.

correlated. Secondly, we should consider approaches that jointly fit models for related response variables in order to encourage similarity amongst these models. Thirdly, the statistical models should be able to explain the serial correlation often observed in the response variables. And finally, the approach must produce sensible models with minimal human intervention so that a large number of models can be produced efficiently.

Now that we have introduced the modelling challenges often encountered when modelling telecommunications datasets we will provide the structure of the remainder of this thesis.

1.2 Thesis structure

In Chapter 2 we introduce the notation used throughout. We then review the most relevant literature relating to predictor selection in both univariate response and multivariate response linear regression and how the best set of predictors may be chosen. We will briefly describe the modelling approach used by our industrial collaborator for telecommunications data. By exploring the current approach, we can highlight the areas where this approach is undesirable and most in need of improvement. This will allow us to further motivate our methodology.

In Chapter 3 we present our semi-automated approach that simultaneously selects predictors for multivariate response linear regression. We then define a simultaneous shrinkage operator and show how it can be used to further improve parameter estimation. We integrate our simultaneous predictor

selection approach into a two-step procedure that iterates between learning the serial correlation of model errors and selects the best predictors to include into a model. We show empirically that simultaneous predictor selection can perform favourably when compared to univariate methods and that our two-step approach can greatly improve the performance of predictor selection. Finally we demonstrate our approach on a subset of the telecommunications dataset.

In Chapter 4 we apply our method to the *full* telecommunications event dataset. We explore the performance of our approach with the current approach rigorously, and provide insight into the gains of simultaneous predictor selection in practise. We also assess how well each approach satisfies the modelling assumptions that are specified a priori.

In Chapter 5 we investigate how the computational performance of the simultaneous predictor selection approach proposed in Chapter 3 is affected by various different implementations. In particular we consider solving the SBS problem by formulating a number of different MIQO programs. In addition to this, we compare the approach to a fast alternative that does not require an optimisation solver and can produce good quality models quickly.

In Chapter 6 we propose a number fast simultaneous predictor selection approaches. We discuss how these approaches relate to their univariate counterparts that have been proposed in the current body of literature. In a simulation study we compare how each approach performs across a range of practical performance criteria.

In Chapter 7 we carry out a detailed study to understand how the simultaneous shrinkage operator proposed in Chapter 3 performs in many different scenarios. In particular, we compare the performance of the shrinkage operator in *sparse*, *medium* and *dense* scenarios. Here sparse, medium and dense scenarios correspond to models with low to a high number of active predictors present.

Finally, we conclude this thesis in Chapter 8 by providing a summary of each chapter in turn. In addition to this, we also discuss areas for future research that may provide avenues for our industrial collaborator to explore and interesting areas for further academic research.

Chapter 2

Literature review and current procedures used to model telecommunications data

In this chapter we review the most relevant literature related to our modelling challenges. We consider both univariate and multivariate linear regression models and discuss a number of approaches used to estimate and select predictors for these models. We discuss some known properties of these methods and how they can be used to address our goals. Importantly, we also highlight areas where these approaches do not meet our needs and identify areas in the literature where significant contributions can be made. We follow by describing the current modelling approach employed by our industrial collaborator. But first, we introduce the notation used throughout this thesis.

2.1 Notation

Throughout this thesis we use Y to denote a response variable and X to denote a predictor. We are interested in how multiple predictors, $X_1, \ldots, X_P$ influence the response variable Y. In Chapter 1 we discussed the potential to jointly model response variables. When there are multiple response variables we will denote them by Y_m, for $m = 1, \ldots, M$. We will consider *grouping* related response variables. We use the notation $\mathcal{G}_i$ to denote the i^{th} group of response variables. The set $\mathcal{G}_i$ lists the indices of the response variables in Group i. For example, let Group 1 contain the variables $\{Y_1, Y_2, Y_3\}$. Then, we will refer to this group of response variables as $\mathcal{G}_1 = \{1, 2, 3\}$.

It will be useful to distinguish between predictors when there are multiple response variables.

For each response variable we assume that there are P predictors. We assume that each response variable has a realisation of each predictor such that $X_{p,m}$ denotes the realisation of predictor X_p for response variable Y_m. We denote all predictors for response variable Y_m as $X_{1,m}, \ldots, X_{P,m}$ for $m = 1, \ldots, M$. As an example, suppose that we consider only precipitation as a predictor. Here, $P = 1$ and the precipitation time series for each response variable is given by $X_{1,m}$ for $m = 1, \ldots, M$.

We will index the *observations* for both response and predictor variables by t. We use the convention that T denotes the total number of observations. Similarly, we use p to index the predictor variables and P to denote the total number of predictors. When we consider multiple response variables we use m to index each response variable and M to denote the total number of response variables. When considering only one response variable we denote the observations as,

$$\boldsymbol{y} = \begin{bmatrix} y_1 \\ \vdots \\ y_T \end{bmatrix}' \quad \text{and,} \quad \boldsymbol{x} = \begin{bmatrix} x_{1,1} & \ldots & x_{1,P} \\ \vdots & \ddots & \vdots \\ x_{T,1} & \cdots & x_{T,P} \end{bmatrix}. \tag{2.1.1}$$

Here, $'$ corresponds to the matrix transpose and we will use this throughout. We also use the convention that y corresponds to an observation of the random variable Y. We use x to denote an observation of predictor X but do not assume that these predictors are random variables. In (2.1.1) the matrix $\boldsymbol{y} \in \mathbb{R}^T$ is a matrix of dimension T and $\boldsymbol{x} \in \mathbb{R}^{T \times P}$ has dimension $T \times P$. When considering a group of response variables we generalise the notation in (2.1.1). When M response variables are considered we use the notation

$$\boldsymbol{y} \in \mathbb{R}^{T \times M} \quad \text{and} \quad \boldsymbol{x} \in \mathbb{R}^{T \times P \times M}$$

to represent the observations. Here, $y_{t,m}$ corresponds to observation t of response variables Y_m. The value $x_{t,p,m}$ corresponds to observation t of predictor $X_{p,m}$. In the presence of multiple response variables it will be useful to consider the data for only one response variable. In this case, we will use the notation

$$\boldsymbol{y}_{*,m} = \begin{bmatrix} y_{1,m} \\ \vdots \\ y_{T,m} \end{bmatrix}' \in \mathbb{R}^T \quad \text{and} \quad \boldsymbol{x}_{*,*,m} = \begin{bmatrix} x_{1,1,m} & \ldots & x_{1,P,m} \\ \vdots & \ddots & \vdots \\ x_{T,1,m} & \cdots & x_{T,P,m} \end{bmatrix} \in \mathbb{R}^{T \times P}.$$

Here, it is clear from the use of the asterisk that $\boldsymbol{y}$ is two dimensional and $\boldsymbol{x}$ is three dimensional. This indicates that we are considering data for multiple response variables. Specifying **only** the index m we indicate that we are considering **all data** for response variable Y_m and predictor variables $X_{1,m}, \ldots, X_{P,m}$ respectively.

The predictors used in an analysis may be produced by applying a series of transformations to some *base* predictor. We consider two transformations of a predictor. The first transformation is a

non-linear transformation that smooths a predictor and was used to produce the predictors shown in Figure 1.1.3. Given a base predictor, X_p the exponential smoothing function may be used to produce a predictor X_s, such that

$$x_{t,s} = \alpha x_{t,p} + (1-\alpha)x_{t-1,s}, \quad \text{for } t = 2, \ldots, T. \tag{2.1.2}$$

Here, we set $x_{1,s} = x_{1,t}$. The reason for applying such transformations will be made clear in Section 2.3. In equation (2.1.2) α is a parameter that is used to adjust how much the time series $X_{t,p}$ is smoothed. A value of α close to 1 will produce a time series very close to the original. A value of α close to 0 will produce a time series that evolves much more slowly. Suppose we applied the exponential smoothing function for $\alpha \in [0.1, 0.2, 0.3]$ to base predictor X_1. Then, we will produce three new predictors, call them X_2, X_3, X_4. These predictors are all produced from applying the exponential smoothing function to base predictor X_1 and give a group of predictors. When a group of predictors have been produced from a non-linear transformation of a base predictor we use the notation $\mathcal{T}_i$ to correspond to the i^{th} group. Suppose Non-linear Predictor Transformation Group 1 corresponds to the predictors $\{X_2, X_3, X_4\}$, then $\mathcal{T}_1 = \{2, 3, 4\}$.

It will also be useful to *lag* predictors. Given an observation of predictor X_p we will use the notation $LX_{t,p} = X_{t-1,p}$ to denote *lagging* the variable X_p by 1 lag. Here, L is known as the backward shift operator. Suppose we lag observations of predictor X_1, such that

$$\begin{aligned} x_{2,t} = Lx_{1,t} &= x_{1,t-1}, \quad \text{for } t = 2, \ldots, T, \\ x_{3,t} = L^2 x_{1,t} &= x_{1,t-2}, \quad \text{for } t = 3, \ldots, T, \\ x_{4,t} = L^3 x_{1,t} &= x_{1,t-3}, \quad \text{for } t = 4, \ldots, T. \end{aligned}$$

Here, we have created observations of predictor X_2, X_3 and X_4 by lagging predictor X_1 by 1, 2, and 3 respectively. It will be useful to group predictors that are produced from lagging a base predictor. Suppose Lagged Predictor Group 1 corresponds to the predictors X_2, X_3, X_4 then $\mathcal{L}_1 = \{2, 3, 4\}$.

2.2 Literature review

We will now explore the most important literature relevant to the work presented in this thesis. In Section 2.2.1 we introduce the linear regression model and methods used for estimation. We follow by introducing a generalisation of the linear regression model in Section 2.2.2 which is particularly useful when observations are ordered in time. In Section 2.2.3 we review the most popular methods used to select predictors. In Section 2.2.4 we discuss mathematical programming tools used to both estimate regression models and select predictors. In Section 2.2.5 we discuss methods used to estimate

models for a multivariate response variable and highlight why these methods are not suitable for our application. Finally, in Section 2.3 we outline the current modelling procedure typically used by our industrial collaborator and discuss the drawbacks of this approach that we aim to address.

2.2.1 Linear regression

Given a response variable Y and predictor variables $X_1, \ldots, X_P$, a linear regression model can be written in the form

$$Y = \beta_0 + \sum_{p=1}^{P} X_p \beta_P + \eta. \tag{2.2.1}$$

Here, $\beta_0, \beta_1, \ldots, \beta_P$ denote the regression coefficients and η denotes the model error. We refer to the errors of a linear regression model as *regression residuals*. It is assumed that the relationship between the response variable and the predictors is linear. Given observations, $\boldsymbol{y}$ of the response variable and, $\boldsymbol{x}$ for the predictor variables we wish to estimate the *best* model parameters $\beta_0, \beta_1, \ldots, \beta_P$. The *Ordinary Least Squares* (OLS) estimates find the *best* values by minimising the sum of squared residuals. The OLS estimates are given by the solution to the following minimisation problem,

$$\hat{\boldsymbol{\beta}} = [\hat{\beta_0}, \hat{\beta_1}, \ldots, \hat{\beta_P}] = \underset{\beta_0,\ldots,\beta_P}{\arg\min} \left[\sum_{t=1}^{T} \left(y_t - \beta_0 - \sum_{p=1}^{P} \beta_p x_{t,p} \right)^2 \right]. \tag{2.2.2}$$

Here, the residuals $\eta_t = y_t - \beta_0 - \sum_{p=1}^{P} x_{t,p}\beta_p$ for $t = 1, \ldots, T$. The coefficients can also be estimated using statistical inference. We can place an assumption on the distributional form of the regression residuals and use the method of maximum likelihood to estimate the regression coefficients. It is common to assume that the residuals are independent and identically distributed such that

$$\eta_t \sim \mathrm{N}(0, \sigma^2) \quad \text{for} \quad t = 1, \ldots, T.$$

Here, the residuals are assumed to be normally distributed with zero mean and common variance, σ^2. Under these assumptions, the least squares estimates are the same as the estimates obtained by the method of maximum likelihood (Rao and Toutenburg, 1999).

It is common to include the intercept term β_0 in a linear regression model unless there is good reason not to (Ryan, 2008). When an intercept term is included in the model in equation (2.2.1), we can append a column of 1's to the predictor matrix $\boldsymbol{x}$ so that,

$$\overset{*}{\boldsymbol{x}} = [\mathbf{1}\ \boldsymbol{x}] \in \mathbb{R}^{T \times (P+1)}.$$

Here, $\mathbf{1} \in \mathbb{R}^{T \times 1}$. Then, the least squares estimates of $\boldsymbol{\beta}$ in model (2.2.1) are available in closed from and given by,

$$\hat{\boldsymbol{\beta}} = (\overset{*}{\boldsymbol{x}}{}' \overset{*}{\boldsymbol{x}})^{-1} \overset{*}{\boldsymbol{x}}{}' \boldsymbol{y}. \tag{2.2.3}$$

As well as the closed form expression given in (2.2.3) the least squares estimator has a number of other desirable properties. The least squares estimator is consistent and optimal in the class of linear unbiased estimators (Rawlings et al., 1998). However, the first column of $\overset{*}{\boldsymbol{x}}$ can cause numerical instability. If one or more columns of $\boldsymbol{x}$ differ very little, then these columns will be near multiples of $\mathbf{1}$. In this case, the matrix $\overset{*}{\boldsymbol{x}}$ will be *ill-conditioned*. Ill-conditioned matrices can cause numerical instability and this was emphasised by Longley (1967). By *centering* the response variables we remove the need to include $\mathbf{1}$ in the predictor matrix and hence remove the intercept from the model. We *center* the response variable by subtracting the sample mean from each observed value. Centering the predictors can also be useful. When only the predictors are centered, the interpretation of the intercept, β_0 is the expected value of the response when the predictors X_p for $p = 1, \ldots, P$ are equal to their mean. Snee and Marquardt (1984) point out that the intercept is essentially a nuisance parameter as we are generally not interested in the value of the response when the predictors all take the value zero.

As well as centering, the data can also be scaled. The purpose of scaling the data is so that the arbitrariness in the choice of scale is eliminated (Mardia et al., 1994). For example, if X_1 measures the depth of rainfall and Y is the rate of telecommunication events, then Y will be the same whether X is measured in *mm* or *cm*. Scaling and centering of the response and predictor variables is accomplished as follows,

$$\tilde{Y} = \frac{Y - \mu_Y}{\sigma_Y} \quad \text{and} \quad \tilde{X}_p = \frac{X_p - \mu_{X_p}}{\sigma_{X_p}}. \tag{2.2.4}$$

Here, $\tilde{Y}$ and $\tilde{X}_p$ give the scaled and centered response and predictor variables. We will use the sample estimates, $\hat{\mu_Y} = \frac{1}{T}\sum_{t=1}^{T} y_t$, $\hat{\mu}_{X_p} = \frac{1}{T}\sum_{t=1}^{T} x_{t,p}$, $\hat{\sigma}_Y^2 = \frac{1}{T}\sum_{t=1}^{T}(y_t - \hat{\mu}_Y)^2$ and, $\hat{\sigma}_{X_p}^2 = \frac{1}{T}\sum_{t=1}^{T}\left(x_{t,p} - \hat{\mu}_{X_p}\right)^2$ and use these to center and scale the observed response and predictor variables. Unless otherwise stated we will estimate models of the form,

$$\tilde{Y} = \sum_{p=1}^{P} \tilde{X}_p \tilde{\beta}_p + \tilde{\eta}.$$

Here, $\tilde{\eta}$ is the regression residual obtained for the scaled and centered model. The model for the response variable on the original scale can be recovered as follows,

$$\begin{aligned}\frac{Y - \mu_Y}{\sigma_Y\sqrt{T}} = \sum_{p=1}^{P} \frac{X_p - \mu_{X_p}}{\sigma_{X_p}\sqrt{T}}\tilde{\beta}_p + \tilde{\eta} \iff Y &= \mu_Y - \sum_{p=1}^{P} \frac{\mu_{X_p}\tilde{\beta}_p\sigma_Y}{\sigma_{X_p}} + \sum_{p=1}^{P} X_p \frac{\tilde{\beta}_p\sigma_Y}{\sigma_{X_p}} + \tilde{\eta}\sigma_Y\sqrt{T} \\ &= \beta_0 + \sum_{p=1}^{P} X_p\beta_p + \eta.,\end{aligned}$$

where $\beta_0 = \mu_Y - \sum_{p=1}^{P} \frac{\mu_{X_p}\tilde{\beta}_p\sigma_Y}{\sigma_{X_p}}$, $\beta_p = \frac{\tilde{\beta}_p\sigma_Y}{\sigma_{X_p}}$ for $p = 1, \ldots, P$ and $\eta_t = \tilde{\eta}\sigma_Y\sqrt{T}$.

In the presence of highly correlated predictors Hoerl and Kennard (1970) showed that the least squares estimates can be unsatisfactory. In simulations, the estimated coefficients could even take

the wrong sign. That is, for a predictor which should have a positive effect on the response variable, estimates of the associated regression coefficient were found to be negative. Hoerl and Kennard (1970) proposed *shrinking* the coefficients towards a more stable solution, closer to the origin. The method proposed by Hoerl and Kennard (1970) is known as ridge regression. The ridge estimates are given in closed form by

$$\hat{\beta}_{\text{ridge}} = \arg\min_{\boldsymbol{\beta}} \left[\sum_{t=1}^{T} \left(y_t - \sum_{p=1}^{P} x_{t,p}\beta_p \right)^2 + \lambda \sum_{p=1}^{P} \beta_p^2 \right] = \left(\boldsymbol{x}'\boldsymbol{x} + \lambda \boldsymbol{I} \right)^{-1} \boldsymbol{x}'\boldsymbol{y}. \tag{2.2.5}$$

Here, λ is a tuning parameter and $\boldsymbol{I}$ is the $P \times P$ identity matrix. As λ increases the regression coefficients are *shrunk* towards zero. We can see this by considering the objective function in (2.2.5). As $\lambda \longrightarrow \infty$ the r.h.s term will start to dominate the objective function. As a consequence, small regression coefficients are required to minimise the objective function to keep the contributions of the r.h.s expression as small as possible. Shrinking the coefficients, although inducing bias, can improve prediction accuracy (Hastie et al., 2008).

Often, our industrial collaborator estimates regression models whereby the regression coefficients obtained are non-meaningful. The predictors used in the models are often highly correlated and this highlights the challenges of modelling with highly correlated predictors in our industrial setting. In addition to this, we are faced with correlation across time. We will now discuss a more general class of linear regression model where the residuals are assumed to be correlated across time.

2.2.2 Regression with correlated residuals

We stated earlier that the common assumptions placed on the residuals, η are that they are independent and normally distributed with zero mean and common variance σ^2. When regression residuals exhibit serial correlation they are no longer independent and the least squares estimates are inefficient, although they remain unbiased (Fang and Koreisha, 2004). Cochrane and Orcutt (1949) developed an approximate procedure for least squares estimation in the presence of serial correlation. The *Cochrane-Orcutt* procedure is suitable when the regression residuals can be written

$$\eta_t = \phi_1 \eta_{t-1} + e_t. \tag{2.2.6}$$

Model (2.2.6) is a special case of the more general Seasonal AutoRegressive Integrated Moving average (SARIMA) model,

$$\nabla^d \nabla_s^D \phi(L)\Phi(L)\eta_t = \theta(L)\Theta(L)e_t. \tag{2.2.7}$$

Here, it is often assumed that $e_t \sim \text{WN}(0, \sigma_e^2)$, a white noise process with zero mean and variance σ_e^2. For more details on white noise processes, the reader is refereed to Chatfield (2000). The SARIMA

model is composed of four components, the auto-regressive component $\phi(L) = 1 - \phi_1 L - \ldots - \phi_r L^r$ which we call the AutoRegressive (AR) polynomial. The backward shift operator is denoted, L such that $L\eta_t = \eta_{t-1}$. The Moving Average (MA) polynomial in (2.2.7) is given by $\theta(L) = 1 - \theta_1 L - \ldots - \theta_q L^q$. The integrated term relates to the differencing operator ∇ where $\nabla^d = (1 - L)^d$, and is applied d times. Finally, in a seasonal model there are seasonal counterparts of the AR, MA and differencing operator given by $\Phi(L) = 1 - \Phi_1 L^s - \ldots - \Phi^{Rs} L^{Rs}$, $\Theta(L) = 1 - \Theta_1 L^s - \ldots \Theta_q L^{Qs}$, and $\nabla_s^D = (1 - L^s)^D$ respectively. The seasonal polynomials differ as the lags are at multiples of the seasonal period, s. The residual model (2.2.6) considered by Cochrane and Orcutt (1949) is known as an AR(1) model where $\phi(L) = 1 - \phi_1 L$, with one autoregressive parameter.

Combining the general SARIMA model (2.2.7) with the linear regression model (2.2.1) gives a Regression-SARIMA (Reg-SARIMA) model

$$y_t = \sum_{p=1}^{P} x_{t,p}\beta_p + \eta_t \quad \text{where} \tag{2.2.8a}$$

$$\nabla^d \nabla_s^D \phi(L)\Phi(L)\eta_t = \theta(L)\Theta(L)e_t. \tag{2.2.8b}$$

It is known that if the SARIMA process is *invertible* (2.2.8) can be re-written as

$$\frac{\nabla^d \nabla_s^D \phi(L)\Phi(L)}{\theta(L)\Theta(L)} y_t = \sum_{p=1}^{P} \frac{\nabla^d \nabla_s^D \phi(L)\Phi(L)}{\theta(L)\Theta(L)} x_{t,p}\beta_p + e_t. \tag{2.2.9}$$

This can be seen as a linear regression model on the linearly transformed variables $\overset{*}{y}_t$ and $\overset{*}{x}_{t,p}$ where

$$\overset{*}{y}_t = \frac{\nabla^d \nabla_s^D \phi(L)\Phi(L)}{\theta(L)\Theta(L)} y_t \quad \text{and} \quad \overset{*}{x}_{t,p} = \frac{\nabla^d \nabla_s^D \phi(L)\Phi(L)}{\theta(L)\Theta(L)} x_{t,p}.$$

For more details on SARIMA models the reader is referred to Brockwell and Davis (2002). If the white noise process is assumed to be independent and normally distributed then the least squares estimator applied to the linear transformed data $(\overset{*}{\boldsymbol{y}}, \overset{*}{\boldsymbol{x}})$ will give us an efficient unbiased estimator of the regression coefficients.

In addition to serially correlated errors, the least squares estimates may be unsatisfactory for alternative reasons. One such reason is model interpretability (Hastie et al., 2008). When the number of predictors, P is large, having all predictors present will make the model complicated. It may be more beneficial to consider a smaller subset of predictors that exhibit the strongest effects. Interpretation is of considerable importance to our industrial application as we are trying to quantify the underlying physical relationship between a response variable and a set of predictors. In the following section we will discuss popular predictor selection methods that have been developed specifically to determine a good subset of predictors. A comprehensive review of classical methods is given by Hocking (1976) and for more recent developments see Hastie et al. (2008); Hutmacher and Kowalski (2014).

2.2.3 Predictor selection for linear regression

Statisticians have been concerned with predictor selection since the 1960's. Models with many predictors can be hard to interpret and misleading. For our industrial application we are trying to understand how a set of external predictors affect a set of response variables. It is of utmost importance that we can accurately estimate the effects of the predictors to better understand the underlying physical relationship between the response and predictor variables.

Early predictor selection methods such as the stepwise procedure first presented by Efroymson (1960) are based on simple principles but remain popular today. The idea here, is to add statistically significant predictors into the model, one-by-one, and remove any predictors that no longer remain significant. Statistical significance is determined by the F-statistic, see Miller (2002) or Ryan (2008) for further details. A predictor is added to the model if the F-statistic of the model with the addition of that predictor exceeds some value F_{in}, and a predictor is removed from the model if the F statistic of the model without the predictor exceeds F_{out}. This procedure is easy to implement, computationally efficient and has been shown by Miller (1996) to converge providing $F_{\text{out}} \leq F_{\text{in}}$.

The Efroymson stepwise algorithm uses two simple ideas, a *forwards* and *backwards* search. These ideas can be separated to give two additional stepwise methods, forwards stepwise and backwards stepwise. Oosterhoff (1963) observes that forward stepwise and backward stepwise need not agree. Mantel (1970) illustrates a scenario where forward stepwise could fail to identify an excellent model with two predictors because it may not include either of the predictors alone. These drawbacks of the forward stepwise approach are especially concerning in the presence of a greater number of predictors. Another criticism of all stepwise methods is that they may fail to identify the *best* subset of any given size. Consider at some point in a stepwise search there are k predictors in a model. There may exist another combination of k predictors which can further reduce the sum of squared residuals given in (2.2.2) than the current stepwise model. Stepwise selection may not be able to identify this model because of the iterative approach to adding or removing variables. The problem associated to finding the best k predictors for a regression model is known as the best-subset problem (Miller, 2002).

We refer to k, the number of predictors in the model as the model sparsity. It is computationally costly to fit every model of sparsity k, as given P predictors, there will be a total of $\frac{P!}{(P-k)!k!}$ models. The best-subset problem can be stated formally as

$$\min_{\boldsymbol{\beta}} \left[\sum_{t=1}^{T} \left(y_t - \sum_{p=1}^{P} x_{t,p}\beta_p \right)^2 \right] \quad \text{subject to} \quad ||\boldsymbol{\beta}||_0 \leq k. \tag{2.2.10}$$

Here, the l_0 pseudo-norm $||\boldsymbol{\beta}||_0 = \sum_{p=1}^{P} \mathbb{1}_{\beta_p \neq 0}$ counts the number of non-zero entries in $\boldsymbol{\beta}$. The best-subset problem is known to be a very hard problem (Natarajun, 1995). Many authors including

Hocking and Leslie (1967), Beale et al. (1967), Beale (1970a), LaMotte and Hocking (1970) and Furnival and Wilson (1974) have considered computationally efficient methods to find the best-subset of each size for $k \in \{1, 2, \ldots, P\}$. LaMotte (1972) incorporated the ideas from LaMotte and Hocking (1970) into a computer program called `select`. Hocking (1976) notes that an early version of `select` is inefficient for $P > 30$ and the program described by Furnival and Wilson (1974) is similar, although the computations are performed in a more efficient manner.

Until recently, selecting predictors using the best-subset method was not considered practical for problems where $P \geq 50$. A modern implementation of the best-subset approach is available in the `leaps` (Lumley, 2017) statistical software package for the `R` programming language (R Core Team, 2018). This best-subset implementation accepts up to $P = 49$ predictors. More than 50 predictors can be supplied but the software informs the user that computation may be slow. Recently, Bertsimas et al. (2016) have shown that with increases in computational power, advancements in specialised optimisation software and sophisticated mathematical programming models that the best-subset approach is now suitable for applications with P in the hundreds. We will consider the type of mathematical programs used to solve the best-subset problem in Section 2.2.4.

The best-subset and stepwise approaches are known in the literature as subset selection methods (Hastie et al., 2008). These approaches select predictors to include into a model and typically use the ordinary least squares estimator to estimate the associated coefficients of the predictors. Alternative approaches to estimate the regression coefficients use *shrinkage*. We have already seen the ridge estimator introduced by Hoerl and Kennard (1970). However, due to the form of the ridge estimator it is not capable of predictor selection because all regression coefficient estimates remain non-zero. The general form of a shrinkage operator penalises the ordinary least squares estimator as follows,

$$\min_{\boldsymbol{\beta}} \left[\sum_{t=1}^{T} \left(y_t - \sum_{p=1}^{P} x_{t,p} \beta_p \right)^2 + \lambda \mathcal{P}(\boldsymbol{\beta}) \right]. \tag{2.2.11}$$

Here, $\mathcal{P}(\cdot)$ is a penalty on the regression coefficients $\boldsymbol{\beta}$. Often, the penalty is chosen such that as λ increases, the values of the solution to (2.2.11) are *shrunk* towards zero. Tibshirani (1996) introduced the Least Absolute Shrinkage and Selection Operator (LASSO) which both shrinks coefficients and selects predictors. The LASSO penalty takes the form

$$\mathcal{P}(\boldsymbol{\beta}) = \sum_{p=1}^{P} |\beta_p|.$$

Here, $|\beta_p|$ denotes the absolute value of the regression coefficient. The LASSO approach has been generalised by many authors, including Zou and Hastie (2005), Tibshirani et al. (2005), Zou (2006) and Yuan and Lin (2006). However, Tibshirani (2011) notes that this approach did not receive much attention until Efron et al. (2004) developed an efficient algorithm to estimate the LASSO solutions.

Earlier implementations of the LASSO used an off-the-shelf quadratic solver that did not scale well (Tibshirani, 2011). A gradient-descent based method, later developed by Mazumder et al. (2011) can also be used to compute the LASSO solutions efficiently. The LASSO and variants can also be implemented with the Alternating Direction Method of Multipliers (ADMM) algorithms (Boyd et al., 2011; Gaines et al., 2018).

Under certain conditions, the LASSO benefits from desirable statistical properties, see for example Zhao and Yu (2006), Donoho (2006) Knight and Fu (2000) and Meinshausen and Buhlmann (2006). These, include the ability to correctly identify the true model. However, when these conditions are not satisfied the LASSO can be sub-optimal in model selection, see Zou (2006); Zhang and Huang (2008); Zou and Li (2008); Zhang (2010).

LASSO solutions can be computed efficiently because the LASSO penalty is convex (Efron et al., 2004). Alternative non-convex penalties have also been studied in the literature such as the MC$^+$ penalty of Zhang (2010). The general form of the non-convex penalty is

$$\mathcal{P}(\boldsymbol{\beta}) = \sum_{p=1}^{P} q(|\beta_p|; \lambda, \gamma).$$

Here, $q(|\beta_p|; \lambda, \gamma)$ is a non-convex function in $\boldsymbol{\beta}$ and λ and γ give the degree of regularisation and non-convexity of the penalty respectively. Mazumder et al. (2011) describe an algorithm to efficiently estimate the solutions of a family of non-convex penalties. The `sparsenet` package available for `R` implements the *Sparsenet* methodology described in Mazumder et al. (2011) using the MC$^+$ penalty of Zhang (2010).

Predictor selection can also be considered in a Bayesian framework. Park and Casella (2008) develop a fully Bayesian model for the LASSO problem. An advantage of Bayesian approaches to predictor selection is that standard errors of the regression coefficients are easily obtainable. The limiting distribution of the LASSO estimator is complex (Knight and Fu, 2000; Chatterjee and Lahiri, 2011) making it difficult to accurately quantify uncertainty in regression coefficients in the frequentist framework. Bayesian estimation can also incorporate expert knowledge (Jiang et al., 2016). Garthwaite and Dickey (1988) consider how to construct an informative prior so that expert opinion can be used efficiently.

A number of studies have taken place to compare the performance of subset selection and shrinkage approaches. Bertsimas et al. (2016) compare the best-subset method to the LASSO, Stepwise and Sparsenet methods concluding that the best-subset approach performs favourably by achieving sparse solutions with good predictive power. However, further investigation by Hastie et al. (2017) concludes that neither the LASSO or the best-subset approach uniformly dominate one another. These authors found that the best-subset approach performs best when the ratio between signal and

noise is high, whereas the LASSO is better in low ratio regimes. Hastie et al. (2017) conclude that a simplified version of the relaxed LASSO (Meinshausen, 2007) performed favourably overall. The relaxed LASSO implemented by Hastie et al. (2017) uses the least squares estimates to estimate the regression coefficients for predictors selected with LASSO.

In the following section we will describe a number of techniques that can be used to estimate regression models. In particular we focus on mathematical programming approaches. Much of our work has exploited the flexibility of mathematical programming and the general idea of mathematical programming is key to understanding the flexibility and power of these approaches.

2.2.4 Mathematical programming for regression

In Section 2.2.1 we provided the closed form expression for the OLS and ridge estimators. In Section 2.2.3 we provided references that focus on developing specialised algorithms for implementing the best-subset approach and the LASSO. Here, we consider a much more general approach that can be used to implement both, best-subset selection and the LASSO without the need to develop approach specific algorithms. The advantage here is that an approach can be modified, and providing modified approaches can be presented as one of a number of special mathematical programs, we can implement them easily using mathematical programming tools.

The OLS and ridge estimates are the solutions to the least-squares optimisation problems given in (2.2.2) and (2.2.5) respectively. Similarly, regression coefficients estimated using the LASSO and best-subset approach are solutions to optimisation problems, however numerical algorithms are needed to solve these problems. The original application of the LASSO Tibshirani obtained the LASSO estimates using a Quadratic Program (QP) solver (Tibshirani, 2011). The LASSO problem that is written in penalised form in (2.2.11) can be written as a Quadratic Program. Quadratic programs are special types of mathematical programs that can be formulated as follows (Nocedal and Wright, 2006)

$$\min \left[\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{Q}\boldsymbol{\beta} - \boldsymbol{a}'\boldsymbol{\beta}\right] \quad \text{subject to,} \tag{2.2.12a}$$

$$\boldsymbol{A}\boldsymbol{\beta} \leq \boldsymbol{C}. \tag{2.2.12b}$$

Here, $\boldsymbol{\beta} \in \mathbb{R}^P$ is the vector of *optimisation variables*, $\boldsymbol{Q} \in \mathbb{R}^{P\times P}$, $\boldsymbol{a} \in \mathbb{R}^P$, $\boldsymbol{A} \in \mathbb{R}^{n\times P}$, $\boldsymbol{C} \in \mathbb{R}^{n\times 1}$ and $\leq$ represents the element-wise less than or equal to inequality. The function, $\frac{1}{2}\boldsymbol{\beta}'\boldsymbol{Q}\boldsymbol{\beta} - \boldsymbol{a}'\boldsymbol{\beta}$ given in (2.2.12a) is known as the *objective function* and (2.2.12b) gives the *linear constraints*. When $\boldsymbol{Q}$ is a positive-semi-definite matrix (2.2.12) is a *Convex Quadratic Program* (CQP) (Boot, 1964). State-of-the-art optimisation solvers such as `Gurobi` Gurobi Optimization (2018) and `CPLEX` are capable of solving CQP's.

Bertsimas et al. (2016) discuss a number of Mixed Integer Quadratic Optimisation (MIQO) programs that can be formulated to solve the best-subset problem. The authors propose two formulations that provide good performance in practice. The recommended formulation is determined by the number of observations and the number of predictors under consideration. With these formulations Bertsimas et al. (2016) are able to solve best-subset problems with thousands of observations and hundreds of predictors within seconds. A MIQO program can be expressed as

$$\min \left[\frac{1}{2}\boldsymbol{x}'\boldsymbol{Q}\boldsymbol{x} + \boldsymbol{a}'\boldsymbol{x}\right] \quad \text{subject to,} \tag{2.2.13a}$$

$$\boldsymbol{A}\boldsymbol{x} \leq \boldsymbol{C}, \tag{2.2.13b}$$

$$x_i \in \{0, 1\}, \quad \text{for} \quad i \in \mathcal{I}, \tag{2.2.13c}$$

$$x_i \in \mathbb{R}^+ \quad \text{for} \quad i \notin \mathcal{I}. \tag{2.2.13d}$$

Here, $\boldsymbol{Q} \in \mathbb{R}^{D\times D}$, $\boldsymbol{A} \in \mathbb{R}^{n\times D}$, $\boldsymbol{C} \in \mathbb{R}^{n\times 1}$ and $\leq$ denotes the element-wise less than or equal to inequality. We optimise over the $\boldsymbol{x} \in \mathbb{R}^D$ containing both discrete $(x_i, i \in \mathcal{I})$ and continuous $(x_i, i \notin \mathcal{I})$ variables. Many optimisation solver are capable of solving MIQO programs. The ADMM algorithms that can be used to implement the LASSO are not able to implement the best-subset approach exactly (Boyd et al., 2011).

Constraining variables to take integer values makes mathematical programs very hard to solve (Natarajun, 1995). The literature for solving discrete optimisation problems is vast, but a good introduction to solving integer programming problems is given by Wolsey (1998). Many efficient approaches to integer programming problems implement the *branch-and-bound* method first proposed by Land and Doig (1960). The idea here is to create a tree that can be used to explore the solution space. Efficient algorithms prune this tree so that the entire solution space need not be explored.

Having introduced the linear regression model and estimation methods we now focus on the work of Bertsimas and King (2016). Approaches discussed thus far often produce undesirable models in the presence of highly correlated predictors. The method proposed by Bertsimas and King (2016) is able to exclude pairs of highly correlated predictors from entering a model. In Chapter 3 we generalise this approach to fit multiple linear regression models simultaneously and show how it can be used as an automated approach that can obtain good models with minimal effort.

An algorithmic approach to linear regression

Bertsimas and King (2016) proposed *an algorithmic approach to linear regression.* Bertsimas and King suggest using a MIQO to fit a model that satisfies a number of desirable attributes. These attributes are discussed in the following texts, Chatterjee et al. (2012); Draper and Smith (1998);

Seber and Lee (2003); Weisberg (2014). A simplification of the MIQO model presented by Bertsimas and King (2016) is given by

$$\min_{\boldsymbol{\beta},\boldsymbol{z}} \left[\sum_{t=1}^{T} \left(y_t - \sum_{p=1}^{P} x_{p,t}\beta_p \right)^2 \right] \quad \text{subject to,} \tag{2.2.14a}$$

$$z_p \in \{0,1\}, \quad \text{for } p = 1,\ldots,P, \tag{2.2.14b}$$

$$\beta_p \in \mathbb{R}, \quad \text{for } p = 1,\ldots,P, \tag{2.2.14c}$$

$$-Mz_p \leq \beta_p \leq Mz_p, \quad \text{for } p = 1,\ldots,P, \tag{2.2.14d}$$

$$\sum_{p=1}^{P} z_p \leq k, \tag{2.2.14e}$$

$$z_p + z_s \leq 1, \quad \forall (p,s) \in \mathcal{HC}, \tag{2.2.14f}$$

$$\sum_{p \in \mathcal{T}_j} z_p \leq 1, \quad \forall j, \tag{2.2.14g}$$

$$z_p = 1, \quad \forall p \in \mathcal{J}. \tag{2.2.14h}$$

For large enough M (2.2.14a) through (2.2.14e) provides a MIQO program that can be used to solve the best-subset problem (2.2.10). The binary variables z_p ensure that if $z_p = 0$ then $\beta_p = 0$, otherwise β_p can take any value within the range $[-M, M]$ by the constraints given in (2.2.14d). Constraint (2.2.14e) controls the sparsity of the model. As explained previously, this constraint is particularly useful when many predictors are available. By allowing no more than k of the binary variables, z_p to take the value one, constraint (2.2.14d) ensures that no more than k of the regression coefficients are non-zero. The remaining constraints help ensure that the models produced have a number of desirable properties. Define the set of pairs of highly correlated predictors,

$$\mathcal{HC} = \{(p,s) : \text{Cor}(X_p, X_s) > \rho, \ \forall (p,s) \in \{1,\ldots,P\} \times \{1,\ldots,P\}\}.$$

Using $\mathcal{HC}$ in constraint (2.2.14f) ensures that no pair of predictors with correlation exceeding ρ can be present in the model. The set $\mathcal{T}_j$ gives the indices of a set of predictors which are a result of applying non-linear transformations to one of the other predictor variables. Constraint (2.2.14g) ensures that at most one of the predictors from $\mathcal{T}_j$ is present in the model. Finally, the set $\mathcal{J}$ denotes the set of predictors that must be present in the model. The set of required predictors may be provided by expert knowledge.

The MIQO model presented in (2.2.14) simplifies the approach suggested by Bertsimas and King (2016) but promises a number of desirable properties in any model produced. Other properties provided by Bertsimas and King (2016) include avoiding particular combinations of predictors in a model and including groups of predictors, where all predictors in the group are either included or

not included. In addition to this, the objective can be modified to produce robust estimates of the regression coefficients in the presence of atypical observations.

The predictor selection and linear regression estimation approaches we have considered thus far are suitable when the response variable in a linear regression model is univariate. That is, we consider producing models for one response variable at a time. We will refer to the approach of modelling each response individually as *individual regression*. In the following section we consider multivariate response linear regression and approaches that have considered predictor selection for these models.

2.2.5 Multivariate response linear regression

Related response variables in our industrial applications may suggest that jointly modelling such variables could be favourable. Our industrial collaborator would like to explore the possibility of jointly estimating the system of models

$$\begin{aligned} Y_{t,1} &= \sum_{p=1}^{P} X_{t,p,1}\beta_{p,1} + \eta_{t,1}, \\ &\dots \\ Y_{t,M} &= \sum_{p=1}^{P} X_{t,p,M}\beta_{p,M} + \eta_{t,M}. \end{aligned} \tag{2.2.15}$$

In the literature, systems of models such as (2.2.15) are known as *seemingly unrelated regression models* (Nagabhushana Rao et al., 2013). Early work by Zellner (1962) sparked interest in producing efficient estimators for such models. Zellner (1962) gained efficiency by using a generalised least squares estimator that utilises correlation amongst the residuals between models for multiple response variables.

To the best of our knowledge, predictor selection for systems of linear regression models has not yet been considered in the literature. Systems similar to (2.2.15), where predictor selection methods are known, are known as multi-response regression models. These models are subtly different to the system (2.2.15). Multi-response models take the form

$$\begin{aligned} Y_1 &= \sum_{p=1}^{P} X_p\beta_{p,1} + \eta_1, \\ &\dots \\ Y_M &= \sum_{p=1}^{P} X_p\beta_{p,M} + \eta_M. \end{aligned} \tag{2.2.16}$$

Note that a single realisation of the predictors $X_1, \dots, X_P$ are present in each of the M regression models. Here, we do not assume that each response has its own unique realisation of the predictors

which is assumed in (2.2.15). The models presented in (2.2.16) are not entirely appropriate for our application, but some interesting literature in this area has inspired ideas for the work presented in later chapters.

Early work by Izenman (1975), van der Merwe and Zidek (1980) and Brown and Zidek (1980) used shrinkage estimation procedures to estimate multi-response models. However, the *curds and whey* method proposed by Breiman and Friedman (1997) performed more favourably in simulations performed by the authors. However, none of these approaches consider the problem of predictor selection.

Predictor selection for multi-response models has been considered by multiple authors, see for example Rothman et al. (2010), Lee and Liu (2012), Xin et al. (2017), and using a Bayesian framework, Lee et al. (2017). Similä and Tikka (2005) propose an extension of the LARS algorithm (Efron et al., 2004) which is generalised further by Similä and Tikka (2006). Turlach et al. (2005) and Similä and Tikka (2007) present algorithms for solving constrained optimisation problems related to multi-response models. These constrained optimisation problems take the form,

$$\min \left[\sum_{m=1}^{M} \sum_{t=1}^{T} \left(y_{t,m} - \sum_{p=1}^{P} x_{t,p} \beta_{m,p} \right)^2 \right] \quad \text{subject to} \quad \sum_{p=1}^{P} \left(||\boldsymbol{\beta}_{p,*}||_q \right) \leq \nu, \tag{2.2.17}$$

for some norm $||\cdot||_q$ on the regression coefficients. Here, we use the following notation for the regression coefficients,

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_{1,1} & \cdots & \beta_{1,M} \\ \vdots & \ddots & \vdots \\ \beta_{P,1} & \cdots & \beta_{P,M} \end{bmatrix} \quad \text{where} \quad \boldsymbol{\beta}_{p,*} = [\beta_{p,1}, \ldots, \beta_{p,M}] \quad \text{and} \quad \boldsymbol{\beta}_{*,m} = \begin{bmatrix} \beta_{1,m} \\ \vdots \\ \beta_{P,m} \end{bmatrix}.$$

In (2.2.17) the parameter ν is a tuning parameter. Turlach et al. (2005) consider the l_∞ norm, $||\boldsymbol{\beta}_{p,*}||_\infty = \max\{\beta_{p,1}, \ldots, \beta_{p,M}\}$ whereas Similä and Tikka (2007) consider the l_2 norm, $||\boldsymbol{\beta}_{p,*}||_2 = \sqrt{\sum_{m=1}^{M} \beta_{m,p}^2}$. We note here that the solutions obtained by Turlach et al. (2005) to problem (2.2.17) using the l_∞ norm are not sparse. The authors suggest a simple heuristic that determines which predictor coefficients to set to zero by considering the size of the coefficients in the solutions. The indices of the selected predictors from the heuristic are given by

$$\mathcal{I} = \{p : ||\boldsymbol{\beta}_{p,*}||_\infty > \nu 10^{-4} \quad \text{for } p = 1, \ldots, P\}.$$

Turlach et al. (2005) note that the coefficients in the solutions they obtain may not have any inherent meaning but may be useful for explanatory purposes.

For both the subset selection and shrinkage approaches used to estimate linear regression models a tuning parameter is needed. We will now discuss methods used to determine these tuning parameters and hence select an appropriate model.

2.2.6 Model selection

We have considered a number of methods that may determine a useful set of predictors to include into a linear regression model. Given k, the best-subset approach selects the predictors which minimises the least squares objective subject to at most k of the regression coefficients taking non-zero values. Given λ, the LASSO minimises a penalised form of the least squares objective. The form of the LASSO penalty both shrinks and selects predictors as some coefficients are set to zero exactly (Tibshirani, 1996). Here, we discuss how to determine λ and k. Each value of k and λ produces an estimate of a regression model using the best-subset approach and LASSO respectively. Selecting the tuning parameter will in effect select a linear regression model, so we use the terms *selecting a tuning parameter* and *selecting a model* interchangeably.

One approach to model selection is using information theory. Suppose we wish to select a model from a list $\mathcal{M}_1, \ldots, \mathcal{M}_N$. Each $\mathcal{M}_n$ is a set containing the indices of predictors in the model and we use $k_n = |\mathcal{M}_n|$ to denote the number of predictors in model $\mathcal{M}_n$. Under the normality assumptions of the regression residuals stated in Section 2.2.1 and given observed data $(\boldsymbol{y}, \boldsymbol{x})$, the likelihood function for a univariate response model, $\mathcal{M}_n$ is given by

$$L(\boldsymbol{\theta}_n) = \prod_{t=1}^{T} \left(\frac{1}{\sqrt{2\pi\sigma_n^2}} \exp \frac{\left(y_t - \sum_{p\in\mathcal{M}_n} x_{t,p}\beta_p\right)^2}{2\sigma_n^2} \right). \tag{2.2.18}$$

Here, $\boldsymbol{\theta} = [\beta_{p_1}, \ldots, \beta_{p_{k_n}}, \sigma_n^2]$ denotes the parameters for model $\mathcal{M}_n$ which include the regression coefficients and the variance of the residuals, σ_n^2. The likelihood function is maximised with the least squares estimates given by

$$\hat{\boldsymbol{\beta}}^{\mathrm{LS}} = (\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}'\boldsymbol{y},$$

where the residual variance is estimated as $\hat{\sigma_n}^{LS} = \frac{1}{T}\sum_{t=1}^{T}\left(y_t - \sum_{p\in\mathcal{M}_n} x_{t,p}\hat{\beta}_p^{LS}\right)^2$. The log-likelihood function is simply the log of the likelihood function. The log-likelihood for model $\mathcal{M}_n$ is given by

$$l(\boldsymbol{\theta}_n) = \frac{-T}{2}\log\left(2\pi\sigma_n^2\right) - \sum_{t=1}^{T} \frac{\left(y_t - \sum_{p\in\mathcal{M}_n} x_{t,p}\beta_p\right)^2}{2\sigma_n^2}.$$

We could choose the model, $\mathcal{M}_n$ which maximises the likelihood (2.2.18) and log-likelihood, but this will often choose one of the models with the largest number of parameters (Miller, 2002). Akaike (1973) suggested that if using the likelihood to select a model, a penalty should be deducted from the likelihood which penalises the number of parameters in the model. Akaike's Information Criterion (AIC) for model $\mathcal{M}_n$ is given by

$$\mathrm{AIC}_n = 2|\hat{\boldsymbol{\theta}}_n| - 2l(\hat{\boldsymbol{\theta}}_n).$$

Here, we denote the number of parameters in the likelihood $|\hat{\boldsymbol{\theta}}_n| = k_n + 1$, since there is a parameter for each predictor in the model and we must include the estimate of the residual variance. Given N models, $\mathcal{M}_1, \ldots, \mathcal{M}_N$ the model with the lowest AIC is selected. Several authors have proposed modifications of the AIC including Schwarz (1978), Rissanen (1978), Hannan and Quinn (1979) and Hurvich and Tsai (1989). The Schwarz criterion, also known as the Bayesian Information Criterion (BIC) is given by

$$\text{BIC}_n = |\hat{\boldsymbol{\theta}}_n| \log(T) - 2l(\hat{\boldsymbol{\theta}}_n).$$

The BIC is known to be asymptotically consistent for model selection (Hurvich and Tsai, 1989; Vrieze, 2012). However, the expected number of variables that should be omitted but are included in the model does not tend to zero as the sample size increases for the AIC (Miller, 2002).

Stone (1977) showed the asymptotic equivalence of model selection by AIC and cross-validation. Cross-validation is an alternative to using information criterion for model selection. Cross-validation can be used when the data is permutable (Ding et al., 2019) meaning that there is no inherent order for the data. It works as follows. The data is split into a *training* and *validation* set. The training data is used to estimate each of the candidate models. Then, each of these models is used to make predictions for the validation data. For each model, some measure of predictive performance is recorded. The model with the best predictive performance is selected and then the whole dataset is used to re-estimate the selected model for future predictions.

As this form of cross-validation approach is suitable only for permutable data using this approach to select time series models is not appropriate, as each item of data is time ordered. A variant of cross-validation for time series is available from the literature and the interested reader is referred to Hyndman and Athanasopoulos (2019) for further details. Cross-validation has been used to select amongst models produced by the LASSO and best-subset approaches by Bertsimas et al. (2016), Bertsimas and King (2016) and Hastie et al. (2017). This works as follows. The best-subset and LASSO approaches are used to estimate a set of models for pre-specified values of the tuning parameters. For the best-subset approach typically $k = 1, 2, \ldots, P$ is used. For the LASSO, the default used in the `glmnet` package (Friedman et al., 2010) is for 50 values of λ ranging from $\lambda_{\max} = ||\boldsymbol{x}'\boldsymbol{y}||_\infty$ to a small fraction of $\lambda_{\max}$ on a log-scale. Then, the model selected for each approach is that which minimises the prediction error on the validation set.

The maximum likelihood estimates of the regression coefficients are not used to estimate the model in the LASSO approach. This is because the LASSO minimises a penalised form of the sum of squared residuals. Therefore, an alternative form for the information criteria is needed for the

LASSO. The AIC and BIC for the LASSO is derived by Zou et al. (2007) as

$$\text{AIC}_{\text{LASSO}}(\hat{\boldsymbol{\mu}}) = \frac{||\boldsymbol{y} - \hat{\boldsymbol{\mu}}||^2}{T\sigma^2} + \frac{2}{T}\hat{df}(\hat{\boldsymbol{\mu}}) \quad \text{and} \quad \text{BIC}_{\text{LASSO}}(\hat{\boldsymbol{\mu}}) = \frac{||\boldsymbol{y} - \hat{\boldsymbol{\mu}}||^2}{T\sigma^2} + \frac{\log(T)}{N}\hat{df}(\hat{\boldsymbol{\mu}}).$$

Here, $\hat{\boldsymbol{\mu}}$ are the fitted values from a LASSO model, σ^2 is the variance of the residuals and $\hat{df}(\hat{\boldsymbol{\mu}})$ are the *degrees of freedom* of the LASSO fit. It was shown by Zou et al. (2007) that an unbiased estimate of the degrees of freedom for the LASSO fit is given by the number of non-zero coefficients.

In this section we have considered a range of literature that considers estimating regression models and selecting predictors for these models. We have also considered generalisations of the linear regression model that include correlated residuals and multivariate responses. In the following section we provide details of the current modelling approach typically employed by our industrial partner.

2.3 Current procedures

Our aim is to develop statistical models that can explain the relationship between a response variable, Y_m and predictor variables, $X_{1,m}, \ldots, X_{P,m}$. These models may take the form

$$Y_m = f(X_{m,1}, \ldots, X_{p,m}) + \eta. \tag{2.3.1}$$

Here, f denotes some function and η denotes some variation in Y_m not attributed to the predictors $X_{m,1}, \ldots, X_{p,m}$. In Chapter 1 we discussed that the response variables in the telecommunications event dataset exhibit weekly seasonality. Also, bank holidays appear to adversely affect the variation in the responses. This behaviour of the response variables is not thought to be attributed to weather predictors, which are of primary interest for our industrial collaborator in the telecommunications event dataset. The current approach estimates the variation in the response variables caused by weekly seasonality and bank holiday affects and removes it from the response variables. This is seen as a data pre-processing step. The procedure for doing this follows. For ease of notation we shall drop the response index, m as this procedure is an individual regression procedure which is applied to each response variable separately.

It is possible to decompose the response variable into the sum of components. Hyndman and Athanasopoulos (2019) present an additive decomposition model of a time series as the sum of three components, these consist of a seasonal component, S_t a trend-cycle component, T_t and a remainder component R_t such that

$$Y_t = S_t + T_t + R_t.$$

There are a number of ways to estimate the components. Hyndman and Athanasopoulos (2019)

discuss a classical method using moving averages, however our industrial collaborator uses simple averages as follows.

First, we identify the *seasons*. Figure 2.3.1 shows a seasonal sub-series plot. Here, the events are plotted for each day of the week separately. It is clear that the level of events on Saturdays and Sundays are unique and lower than the level of each weekday. There is slight variation between levels of events for each weekday. As the level of events for each day of the week appears to vary it may be argued that we should estimate a seasonal component for each day of the week. Further, we observed in Figure 1.1.5 that events are typically lower on bank holidays. In Chapter 1 we discussed that the events appear to deviate much further from past values on Christmas and Boxing Day in comparison to all other bank holidays. This suggests that a single *seasonal component* for a Christmas Day and Boxing Day, and a seasonal component for all other bank holidays may be suitable.

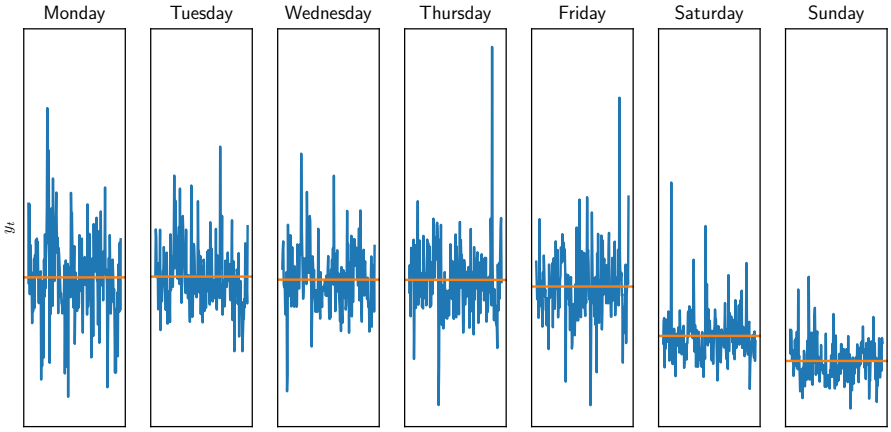


Figure 2.3.1: A seasonal sub-series plot highlighting the weekday levels of the telecommunication event data.

The seasonal components for each season are estimated in the following way. Let the sets of indices be defined

$$\begin{aligned}
\mathcal{S}_1 &= \{t : t \text{ corresponds to Christmas Day, Boxing Day or substitute}\},\\
\mathcal{S}_2 &= \{t : t \notin \mathcal{S}_1 \text{ and } t \text{ corresponds to a bank holiday}\},\\
\mathcal{S}_3 &= \{t : t \notin \mathcal{S}_1 \cup \mathcal{S}_2 \text{ and } t \text{ corresponds to a Monday}\},\\
&\vdots\\
\mathcal{S}_9 &= \{t :\notin \mathcal{S}_1 \cup \mathcal{S}_2 \text{ and } t \text{ corresponds to a Sunday}\}.
\end{aligned}$$

Recall that a bank holiday *substitute* is the additional bank holiday given in lieu of one that falls

on a Saturday or Sunday. Then, the estimates of the seasonal components corresponding to $\mathcal{S}_i$ are calculated as

$$\hat{S}_i = \frac{1}{|\mathcal{S}_i|} \sum_{t \in \mathcal{S}_i} Y_t.$$

The next step is to estimate the trend component. When there are long-term increases or decreases in a time series we say that the time series exhibits trend. Our industrial partner estimates the trend component by applying a 365 day centered moving average to the de-seasonalised data as follows,

$$\hat{T}_t = \frac{1}{\min\{t+183, T\} - \max\{1, t-183\}} \sum_{t=\max\{1,t-183\}}^{\min\{T,t+183\}} \left(Y_t - \hat{S}_t\right).$$

Note that for $t \in [1, 183]$ and $t \in [T-182, T]$ T_t is not strictly symmetric.

Once the trend and seasonal components have been estimated they can be removed and an estimate of the remainder component obtained as

$$\hat{R}_t = Y_t - \hat{S}_t - \hat{T}_t.$$

We let $\tilde{Y}_t = \hat{R}_t$ denote our pre-processed response data. It is possible that the predictor variables also have long-term increases or decreases. Therefore, a centered moving average is also applied to the predictor variables to obtain the pre-processed predictor variables,

$$\tilde{X}_{t,p} = X_{t,p} - \sum_{t=\max\{1,t-183\}}^{\min\{T,t+183\}} X_{t,p}.$$

Relating back to the model given in (2.3.1), we now seek a model of the form,

$$\tilde{Y}_t = \tilde{f}(\tilde{X}_1, \ldots, \tilde{X}_P) + \tilde{\eta}, \tag{2.3.2}$$

for some error $\tilde{\eta}$ and some function $\tilde{f}$. Our industrial partner assumes that $\tilde{f}$ is a linear function in the predictors.

Following the pre-processing of data, our industrial collaborator applies a stepwise search algorithm to select predictors. A number of undesirable properties of the resulting models are often observed, some of which we have already discussed. Typically, combinations of highly correlated predictors are selected for the models where the coefficients of the associated predictors have conflicting signs. This leads one to question the validity of such a model as one would expect strongly correlated predictors to affect the response variable in either a positive or negative way, but not in opposing ways. Hastie et al. (2008) note that this problem is often observed with the least squares estimates and motivates the application of ridge regression (Hoerl and Kennard, 1970).

The stepwise algorithm used by our industrial collaborator is implemented using the `stats::step` (R Core Team, 2018) function in `R`. This procedure iteratively adds the predictor which produces

a model with the lowest AIC, until the AIC of a model can not be reduced further by adding an additional predictor.

In this chapter we have introduced linear regression models and a number of methods used to estimate them. In particular, we focused on procedures that could produce sparse models where a number of the regression coefficients are estimated to be zero. Often these procedures use a tuning parameter and we discussed methods that can can be used to determine them. We introduced literature for predictor selection in multi-response models and described the procedure that our industrial collaborator uses to model telecommunications data. In the next chapter we describe the procedure that we have developed to model telecommunications data.

Chapter 3

Semi-automated simultaneous predictor selection for Regression-SARIMA models: An application to telecommunications events

Abstract: Deciding which predictors to use plays an integral role in deriving statistical models in a wide range of applications. Motivated by challenges of predicting events across a telecommunications network, we propose a semi-automated, joint model fitting procedure for linear regression models. Our approach can model and account for serial correlation in the regression residuals, produce sparse and interpretable models and can be used to jointly select models for a group of related response variables. We achieve this by fitting linear models under constraints on the number of non-zero coefficients using a generalisation of the Mixed Integer Quadratic Optimisation approach developed by Bertsimas and King (2016). Our approach can produce models with better predictive performance on the telecommunications data than methods currently used by industry.

This chapter is structured as follows. In Section 3.1 we start with an introduction to the industrial setting that motivated our methodology. In Section 3.2 we state our problem formally and review the existing literature for predictor selection in linear regression. We then discuss how to use the MIQO program presented by Bertsimas and King (2016) to develop a semi-automated modelling procedure. In Section 3.3 we introduce our MIQO program and extensions that can improve the performance of the models. Section 3.4 highlights the advantages of our approach over standard methods in the literature through a simulation study. We apply our approach to a motivating data application in Section 3.5 before concluding this chapter in Section 3.6.

3.1 Introduction

The use of statistical models to drive business efficiency is becoming increasingly wide spread (Proost and Fawcett, 2013). Consequently, organisations are recording more and more data for subsequent analysis, see for example Katal et al. (2013) and Jordan and Mitchel (2015). As a result, traditional (*manual*) approaches for building statistical models are often infeasible for the ever increasing volumes of data. Automating these approaches is necessary, and will allow principled statistical methods to continue driving business efficiency.

Telecommunication companies routinely collect a variety of data so as to better understand the physical relationship between their network and external influences. In practice, data is collected for a response variable (from the network) along with associated (external) predictor variables. Using this data, the goal is to obtain an interpretable statistical model that explains the behaviour between the response and most important predictors. Whilst historically statisticians have fitted such models by hand, this is costly. The work in this chapter is motivated by a current problem of this form by an industrial collaborator. They have data from many different locations within a network, and wish to develop appropriate models for how the rates of certain events depend on a range of external factors. The statistical challenges include how to fit sparse and interpretable models for each response, whilst allowing for the serial correlation in the data and ensuring we borrow information across the response variables. This all needs to be accomplished with minimal human input.

We propose a multivariate response implementation of the best-subset problem. The idea is to fit the *same* model for each response variable, but allow for the coefficients associated with a particular predictor to vary across each model. We show how the Mixed Integer Quadratic Optimisation (MIQO) approach of Bertsimas et al. (2016) can be used to automatically fit such a model in the presence of a known serial correlation structure for the time-series of responses, and propose an iterative procedure that alternates between learning the serial correlation structure and fitting the

model. Our approach can also shrink the coefficients associated with a particular predictor towards a common value. The model fits can be performed under constraints that avoid including highly correlated predictors, this helps with the interpretability of the final models. We reduce the human input by modelling characteristics of the response variables, instead of determining subjective steps to remove these characteristics. The only input needed is through choosing an appropriate set of predictors and potential non-linear transformations of the predictors. Here, we estimate the serial correlation by pre-specifying a suitable list of time series models, although iterative approaches outlined in Hyndman and Khandakar (2008) could be adopted. The predictor selection approach is computationally feasible for hundreds of predictors and tens of response variables.

There are many articles in the literature devoted to predictor selection in univariate response models see for example, Hastie et al. (2017), Bertsimas et al. (2016), Zou and Hastie (2005), Tibshirani (1996), and Hocking (1976) and the references therein. Hastie et al. (2008) collate many of the methods developed in the literature. Breiman and Friedman (1997) and Srivastava and Solanky (2003) have shown that simultaneous model estimation has advantages over individual modelling procedures. Turlach et al. (2005), Similä and Tikka (2007) and Simon et al. (2013) consider selecting variables for the multi-response models used by Breiman and Friedman (1997) and Srivastava and Solanky (2003). To the best of our knowledge simultaneous predictor selection for multiple separate linear regression models has not been considered in the literature. We show how MIQO can be used to automate model estimation and propose a two-step procedure to fit more general Regression Seasonal AutoRegressive Integrated Moving Average (Reg-SARIMA) models. We find that a more accurate specification of the model for the regression residuals can lead to a significant reduction in the variance of the predictor selection routine. Using the generalised least squares objective (Rao and Toutenburg, 1999) we can improve estimation accuracy of the regression coefficients and predictor selection accuracy.

3.2 Problem statement & existing approaches

First, we introduce the linear regression model and existing methods for choosing suitable predictors. We then outline our proposal to automate a modelling procedure for one response variable and show how expert opinion can be incorporated into our model.

The linear regression model is able to describe the relationship between a response variable, Y and dependent variables, $X_1, \ldots, X_P$ as follows

$$Y = \sum_{p=1}^{P} X_p \beta_p + \eta. \tag{3.2.1}$$

Here, the coefficient β_p tells us how much we should expect Y to change when we observe a unit change in X_p. If the set of predictors $\{X_1, \ldots, X_P\}$ is known, the coefficients, $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_P]$ can be estimated with the *Ordinary Least Squares* (OLS) estimates. Given observations of the response $\boldsymbol{y} \in \mathbb{R}^T$ and predictors $\boldsymbol{x} \in \mathbb{R}^{T \times P}$ the OLS estimates are given by

$$\hat{\boldsymbol{\beta}}_{OLS} = \arg\min_{\boldsymbol{\beta}} \left[\sum_{t=1}^{T} \left(y_t - \sum_{p=1}^{P} x_{t,p} \beta_p \right)^2 \right]. \tag{3.2.2}$$

Here, the aim is to find the values of $\boldsymbol{\beta}$ that minimise the sum of squared residuals. When P is large and contains redundant predictors, the OLS estimates can be unsatisfactory. Prediction accuracy can be improved by shrinking or setting some of the coefficients to zero (Hastie et al., 2008). Setting coefficients to zero removes the corresponding predictors from the model, leading to a simpler, more interpretable model. Throughout, we refer to the number of non-zero coefficients in the model as the model sparsity, which we denote k.

Often, practitioners can offer insight into which predictors may be suitable. The linear regression model assumes a linear relationship between predictors and response variable, but this may not be suitable (Rawlings et al., 1998). For example, some telecommunication events are caused by long periods of heavy rainfall, causing underground cables to flood. Exponential smoothing can be applied to daily precipitation measurements to provide a surrogate predictor for ground water levels. But this introduces the question of how best to choose the smoothing parameter. One option is to obtain such surrogate predictors for a grid of smoothing parameters. But this both increases substantially the number of potential predictors to choose from, and can lead to highly correlated predictors.

Selecting predictors can be achieved in several ways. One popular approach is shrinkage (Tibshirani, 1996), where the regression coefficients are shrunk towards zero. For a suitable penalty, $\mathcal{P}(\boldsymbol{\beta})$ and tuning parameter $\lambda \in \mathbb{R}^+$, some regression coefficients can be set to zero exactly. The penalty can be added to the least squares objective in the following way,

$$\sum_{t=1}^{T} \left(y_t - \sum_{p=1}^{P} x_{t,p} \beta_p \right)^2 + \lambda \mathcal{P}(\boldsymbol{\beta}). \tag{3.2.3}$$

The *Least Absolute Shrinkage and Selection Operator* (LASSO) penalty, $\mathcal{P}_{\text{LASSO}}(\boldsymbol{\beta}) = \sum_{p=1}^{P} |\beta_p|$, introduced by Tibshirani (1996) has received much attention in the literature. It has been applied and generalised by a variety of authors including Yuan and Lin (2006), Zou (2006) and, Tibshirani et al. (2005). Efron et al. (2004) developed an efficient algorithm that can obtain LASSO solutions very quickly. Tibshirani (1996) observed empirically that the LASSO performed unfavourably when high pairwise correlations exist between the predictors. In such cases the LASSO was dominated by the ridge penalty, $\mathcal{P}_{\text{ridge}}(\boldsymbol{\beta}) = \sum_{p=1}^{P} \beta_p^2$. The ridge penalty was developed by Hoerl and Kennard (1970) to improve prediction when predictors are highly correlated. Although a shrinkage method,

the ridge penalty does not act as a predictor selector as all coefficients remain non-zero in a ridge estimate. To improve the performance of the LASSO when predictors are highly correlated Zou and Hastie (2005) proposed the *elastic net* penalty given by, $\mathcal{P}_{\text{e-net}} = (1-\alpha)\sum_{p=1}^{P}\beta_p^2 + \alpha\sum_{p=1}^{P}|\beta_p|$. This penalty is a mixture of the LASSO and ridge penalties.

Alternatively, subset selection methods can be used to select predictors. By determining which subset of predictors to retain, subset methods use the least squares objective to estimate the coefficients of the retained predictors (Hastie et al., 2008). A number of classical subset methods are described in detail by Hocking (1976). The forward-stepwise routine is the current algorithm of choice for selecting predictors in our telecommunications application. This algorithm is usually initialised with an intercept term, iteratively adding the predictor most improving the least squares objective. This gives a fitted model with k predictors for $k = 1, \ldots, P$. However, the model produced by stepwise methods for any $k \geq 2$ are not guaranteed to be the best model with k predictors; in terms of having the smallest value of the least squares objective. Despite the sub-optimal stepwise models and issues raised by Mantel (1970), Beale (1970b), Berk (1978) and Hocking (1976), fast and easy implementation of these algorithms may explain why they remain popular.

Finding the model with sparsity k which minimises the least squares objective is known as the *best-subset* problem (Miller, 2002). The best-subset problem is stated formally as

$$\min_{\boldsymbol{\beta}} \left[\sum_{t=1}^{T}\left(y_t - \sum_{p=1}^{P} x_{t,p}\beta_p\right)^2\right] \quad \text{subject to} \quad \sum_{p=1}^{P}\mathbb{1}_{\beta_p\neq 0} \leq k. \tag{3.2.4}$$

Here, $\mathbb{1}_{\beta_p\neq 0}$ is an indicator variable taking the value 1 if coefficient β_p is non-zero and zero otherwise. An implementation of the best-subset method is available in the statistical package `leaps` (Lumley, 2017) in `R` (R Core Team, 2018) and capable of choosing from up to 49 predictors efficiently. A larger number of predictors may be provided although the computational time may be excessive. Bertsimas et al. (2016) showed that the combined improvements of computational power, mathematical optimisation algorithms, and sophisticated mathematical formulations, that the best-subset method is suitable for choosing amongst hundreds of predictors.

3.2.1 Our proposed automation procedure

Automated procedures can limit the control over the output. We do not seek a fully automated approach, but one that can produce sensible outputs with minimal input for hundreds of response variables. Bertsimas et al. (2016) have shown that the best-subset method tends to produce sparser models than the LASSO. Although the best-subset approach can be more computationally demanding than stepwise approaches, it tends to perform better when it can be applied (Berk, 1978). It is straightforward to implement a stepwise algorithm using MIQO and this can result in a significant

speed up due to the absence of the combinatorics of predictor inclusion. This idea is further explored in Chapter 6.

The best-subset problem with sparsity k can be solved by finding the optimal solution to the following MIQO program (Bertsimas et al., 2016),

$$\min_{\boldsymbol{\beta},\boldsymbol{z}} \left[\sum_{t=1}^{T} \left(y_t - \sum_{p=1}^{P} x_{t,p}\beta_p \right)^2 \right] \quad \text{subject to,} \tag{3.2.5a}$$

$$(1 - z_p, \beta_p) \in \mathcal{SOS}_1, \quad p = 1, \ldots, P, \tag{3.2.5b}$$

$$\sum_{p=1}^{P} z_p \leq k, \tag{3.2.5c}$$

$$\text{s.t. } z_p \in \{0, 1\}, \quad p = 1, \ldots, P, \tag{3.2.5d}$$

$$\beta_p \in \mathbb{R}, \quad p = 1, \ldots, P. \tag{3.2.5e}$$

Here, we use $\mathcal{SOS}_1$ to indicate specially ordered sets of type 1. At most one variable in a specially ordered set constraint can take a non-zero value. If the binary variable z_p takes the value 1 then necessarily, the continuous variable β_p must be zero as $(1 - z_p)$ and β_p form a specially ordered set (3.2.5b). Constraint (3.2.5c) controls the sparsity of the models by restricting the maximum number of predictors to k. The MIQO program can be solved for $k = 1, \ldots, P$. The value k can be chosen with model selection criteria such as the AIC (Akaike, 1973) or BIC (Schwarz, 1978). Alternatively, cross validation methods can be used (Stone, 1974).

Expert knowledge

Bertsimas and King (2016) show that we can easily add constraints to the MIQO program to avoid including pairs of highly correlated predictors into the model. We can add the constraints

$$z_p + z_s \leq 1, \quad \forall (p, s) \in \mathcal{HC}. \tag{3.2.6}$$

Constraints of the form (3.2.6) will allow at most one of the binary variables z_p or z_s to take the value 1. This ensures that at most one of the regression coefficients, β_p or β_s are non-zero so that only one of X_p or X_s will be present in the model. Adding constraints of the form (3.2.6) for all pairs of highly correlated variables, $\mathcal{HC} = \{(p, s) : \text{Cor}(X_p, X_s) > \rho\}$ will ensure that no two predictors with correlation exceeding ρ will enter the model.

Expert knowledge may suggest predictors that must be present in the model. This may be suitable to account for known outliers or other known external influences. Let the set $\mathcal{J}$ denote the indices of predictors that must be present in a model, Bertsimas and King (2016) show that these

predictors can be forced into the model with the constraints

$$z_p = 1, \quad \forall\, p \in \mathcal{J}.$$

Expert knowledge may also suggest how the predictors should affect the response variables. For example, some predictors may be known to have a positive effect on the response variable. Highly correlated predictors can lead to high variance of the least squares coefficients. Hastie et al. (2008) note that it is even possible for the coefficients to take the wrong sign. We propose to include expert knowledge as follows. Let the sets $\mathcal{P}$ and $\mathcal{N}$ denote the sets of predictor indices that should have positive and negative effects on the response variables respectively. Then, the constraints

$$\beta_p \geq 0, \quad \forall p \in \mathcal{P} \quad \text{and} \quad \beta_p \leq 0, \quad \forall p \in \mathcal{N}, \tag{3.2.7}$$

ensure that the regression coefficients take the correct sign according to expert opinion.

In Section 3.1 we discussed the need to determine the best parameter from a set of non-linear transformations. To ensure the best parameters are found, in terms of minimising the least squares objective, we can use the following constraints. Let $\mathcal{T}_i$ denote the set of predictors obtained by applying a non-linear transformation to an observed predictor for a grid of parameter values. Then, the constraints

$$\sum_{p \in \mathcal{T}_j} z_p \leq 1, \quad \text{for } \mathcal{T}_1, \ldots, \mathcal{T}_J, \tag{3.2.8}$$

ensure at most one of the predictors from each group $\mathcal{T}_j$ will appear in the model.

Although it may now be feasible to apply the best-subset method to problems with the number of predictors in the hundreds thanks to the work of Bertsimas et al. (2016) and advances in computational power, the best-subset approach can still be more computationally demanding than alternative methods. We now describe techniques to reduce the computational burden of the best-subset approach.

Computational considerations

The cardinality constraints in the best-subset problem (3.2.4) make it a difficult problem to solve. In fact, formulations of the best-subset problem (3.2.5) using integer variables make the problem NP-hard (Natarajun, 1995). When using constraints of the form (3.2.7) we have noticed a computational advantage. There appears to be considerable speed-up in the total runtime of the solver when the sign of the regression coefficients are restricted to either the positive half-line or negative half-line. Figure 3.2.1 shows the comparison of solving the best-subset problem for $k = 1, \ldots, 35$ using MIQO program (3.2.5) including constraints of the form (3.2.7) where $\mathcal{P} = \{1, \ldots, 35\}$ and $\mathcal{N} = \emptyset$.

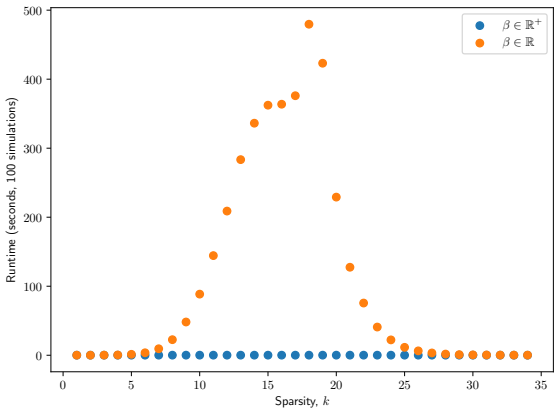


Figure 3.2.1: Average time taken to solve the best-subset problem for $k = 1, \ldots, 35$. The orange marks indicate the average time taken to solve the best-subset problem when $\boldsymbol{\beta} \in \mathbb{R}^P$ compared to the blue marks which constrain $\beta_p \geq 0$ for $p = 1, \ldots, P$. The time taken was averaged over 100 simulations.

In a typical implementation of the best-subset method using formulation (3.2.5), the computational burden of solving the best-subset problem appears to be when solving problems with $k \approx \frac{P}{2}$. This may be explained by the ${}^{P}C_{k}$ feasible combinations of predictors that a solver must *consider* to prove a solution is optimal. In a quest to reduce the computational burden of the best-subset approach an obvious question to ask is, *is solving the best-subset problems with sparsity levels* $k \approx \frac{P}{2}$ *necessary?* In our application, sparse models are desired in order to illustrate the strongest effects of a few predictors. Here, and possibly in many other applications, setting a maximum level of sparsity $K_{\max}$ may be a practical step to reduce the computational burden of the best-subset method.

A maximum level of sparsity could be chosen arbitrarily. However, in our application using constraints of the form (3.2.6) and (3.2.8) the value $K_{\max}$ can be determined automatically. Presence of the constraints (3.2.6) and (3.2.8) suggests that there exists a maximum level of model sparsity where at least one constraint of the form (3.2.6) or (3.2.8) will be violated if an additional predictor is included into the model. We have found that Gurobi (Gurobi Optimization, 2018) will inform the user if an MIQO program is infeasible very quickly. We propose to modify the sparsity constraint (3.2.5c) as follows,

$$\sum_{p=1}^{P} z_p = k.$$

Now, if $k > K_{\max}$ a feasible solution to the *modified* best-subset problem does not exist and the solver will inform the user of an infeasible MIQO program. Thus, we are no longer required to search for models with a greater number of predictors.

We have presented a MIQO program for the best-subset problem that can be used to automate

fitting linear regression models and discussed some techniques that can reduce the computational burden of the best-subset method. In the following section we will describe how we have extended this formulation to model multiple response variables simultaneously and describe a number of extensions that can improve estimation accuracy.

3.3 Simultaneous predictor selection for a system of linear regression models

Interpretability and consistency of models is important in an industry setting. If a model is not easy to interpret then it is of little use for practitioners trying to understand the dynamics of the system being modelled. When models contradict expert opinion or take very different forms for a number of related response variables, the reliability of the models may be questioned. We now describe how we extend the MIQO program (3.2.5) used to solve the best-subset problem. This MIQO program allows us to simultaneously select predictors and obtain models for multiple related response variables to ensure consistency in the selected predictors for each response variable.

3.3.1 Multiple datasets

Many of the response variables in telecommunication applications are correlated and often this is expected. However, due to the high correlation between the predictor variables associated with each response, models produced using the current procedure do not always suggest similarity amongst the response variables. This can be due to both the combination of predictors selected in the models and their estimated coefficient.

We now consider estimating regression models for M response variables simultaneously. We assume that these response variables are suitable for joint analysis. We write the system of models

$$\begin{aligned} Y_1 &= \sum_{p=1}^{P} X_{p,1}\beta_{p,1} + \eta_1, \\ &\dots \\ Y_M &= \sum_{p=1}^{P} X_{p,M}\beta_{p,M} + \eta_M. \end{aligned} \tag{3.3.1}$$

Here, we assume that each response variable has a unique realisation of the P predictor variables. For example, suppose predictor X_1 corresponds to precipitation. Then, predictor $X_{1,m}$ corresponds to the precipitation for response Y_m. Let $\mathcal{S}_m$ denote the set of selected predictors for response Y_m. The current procedure used by our industrial collaborator often produces models where $\mathcal{S}_{m_1} \neq \mathcal{S}_{m_2}$,

contrary to expert opinion. This motivates the following problem which we call the *Simultaneous Best-Subset* (SBS) problem,

$$\min_{\boldsymbol{\beta}} \left[\sum_{m=1}^{M} \sum_{t=1}^{T} \left(y_{t,m} - \sum_{p=1}^{P} x_{t,p,m} \beta_{p,m} \right)^2 \right] \quad \text{subject to } \bigcup_{m=1}^{M} \mathcal{S}_m \leq k. \tag{3.3.2}$$

The union $\bigcup_{m=1}^{M} \mathcal{S}_m$ gives the selected predictors across all models. If all models contain the same predictors then each model may include up to k predictors.

As well as consistency in predictor selection some similarity in the coefficients $\beta_{p,1}, \ldots, \beta_{p,M}$ may be expected. We can penalise for large dissimilarities in the coefficients by introducing auxiliary variables $\bar{\beta}_1, \ldots, \bar{\beta}_P$ and adding the penalty

$$\mathcal{P}(\boldsymbol{\beta}) = \lambda \sum_{m=1}^{M} \sum_{p=1}^{P} \left(\bar{\beta}_{p,m} - \beta_{p,m} \right)^2 \tag{3.3.3}$$

to the objective appearing in (3.3.2). A similar approach has been used by Tibshirani et al. (2005), Barbaglia et al. (2016) and Wilms et al. (2018) using l_1 penalties on the difference between coefficients. The *tuning parameter*, λ must be determined. For large λ the penalty (3.3.3) will dominate the objective and force the solver to encourage $\beta_{p,1}, \ldots, \beta_{p,M}$ close to $\bar{\beta}_p$ for $p = 1, \ldots, P$. In practise, a suitable range of λ must be determined. We have used a sequence of λ equally spaced on the log scale between $2g_k$ and a small fraction of $2g_k$. Let $\boldsymbol{\beta}^*$ denote the optimal solution to the SBS problem (3.3.2) with sparsity k. Then, we denote the value of the objective function to the SBS problem at $\boldsymbol{\beta}^*$ as g_k. We observed that coefficients become more stable for large values of λ and that the coefficients $\beta_{p,1}, \ldots, \beta_{p,M}$ become sufficiently close to $\bar{\beta}_p$ for $p = 1, \ldots, P$ when $\lambda = 2g_k$.

The number of binary variables in the optimisation model need not increase for simultaneously estimating multiple regression models. The number of binary variables remains to be the number of predictor variables, P. However, the number of constraints in the optimisation model must be increased to ensure a feasible solution of (3.3.2) is obtained. We use the $\mathcal{SOS}_1$ constraints

$$(1 - z_p, \beta_{p,m}) \in \mathcal{SOS} - 1, \quad \text{for } p = 1, \ldots, P, \ m = 1, \ldots, M. \tag{3.3.4}$$

These constraints, along with the sparsity constraint (3.2.5c), ensure that no more than k predictors are present across each of the M regression models. Lastly, we specify the range of coefficient values

$$\beta_{p,m} \in \mathbb{R}^+, \quad \text{for } p = 1, \ldots, P, \ m = 1, \ldots, M. \tag{3.3.5}$$

To prevent pairs of highly correlated predictors entering the models we define the set $\mathcal{HC}$ as follows,

$$\mathcal{HC} = \left\{ (p, s) : \text{ for } (p, s) \in \{1, \ldots, P\} \times \{1, \ldots, P\} \ \text{ if } \sum_{m=1}^{M} \sum_{p \neq s} \mathbb{1}_{cor(X_{m,p}, X_{m,s} > \rho)} > 0 \right\}.$$

By using the constraints of the form (3.2.6) we prevent any model in the system (3.3.1) containing pairs of predictors with correlation that exceeds ρ.

3.3.2 Application to serially correlated data

Fitting linear regression models to time ordered data often produces models where the observed residuals appear serially correlated (Brockwell and Davis, 2002). We propose a two-step algorithm similar to the Cochrane and Orcutt (1949) procedure, that implements a predictor selection step to a Generalised Least Squares (GLS) transform of the data. Here, we give an example of the GLS transform, before describing how we incorporate predictor selection. Suppose we have response variable Y and predictors $X_1, \ldots, X_P$ and the true model is

$$Y_t = \sum_{p=1}^{P} X_{t,p}\beta_P + \eta_t, \quad \text{where,} \tag{3.3.6a}$$

$$\eta_t = \phi\eta_{t-1} + e_t. \tag{3.3.6b}$$

Here, the regression residuals, η_t are serially correlated. Ignoring serial correlation in observed residuals may not only mis-specify the model but ignores potentially valuable information. Minimising the least squares objective (3.2.2) no longer gives the most efficient estimator (Rao and Toutenburg, 1999) for the regression coefficients. Providing (3.3.6b) is *stationary* (see Brockwell and Davis, 2002) we can write (3.3.6) as a regression model with residuals that are not serially correlated

$$\frac{Y_t}{1-\phi L} = \sum_{p=1}^{P} \frac{X_{t,p}}{1-\phi L}\beta_p + e_t. \tag{3.3.7}$$

Here, L is the backward-shift operator such that $L\eta_t = \eta_{t-1}$. The linear filter can be applied to the response and predictor variables to obtain $transformations$ of the original variables, $\tilde{Y}_t = \frac{Y_t}{1-\phi L}$ and $\tilde{X}_{t,p} = \frac{X_{t,p}}{1-\phi L}$. We show empirically in Section 3.4.2 that predictor selection accuracy can be improved by transforming the response and predictor variables appropriately.

In practise, neither the predictor variables present in the model or the serial correlation structure of the regression residuals are known. We assume a general Regression Seasonal AutoRegressive Integrated Moving Average (Reg-SARIMA) model of the form

$$y_{t,m} = \sum_{p=1}^{P} x_{t,p,m}\beta_{p,m} + \eta_{t,m}, \quad \text{where,} \tag{3.3.8a}$$

$$\eta_{t,m} = \frac{\theta_m(L)\Theta_m(L^s)}{\nabla^{d_m}\nabla_s^{D_m}\phi_m(L)\Phi_m(L^s)}\epsilon_{t,m}. \tag{3.3.8b}$$

The SARIMA model is composed of four components, the auto-regressive component $\phi(L) = 1 - \phi_1 L - \ldots - \phi_r L^r$ which we call the AutoRegressive (AR) polynomial. The backward shift operator is denoted, L such that $L\eta_t = \eta_{t-1}$. The Moving Average (MA) polynomial in (2.2.7) is given by $\theta(L) = 1 - \theta_1 L - \ldots - \theta_q L^q$. The integrated term relates to the differencing operator ∇ where $\nabla^d = (1-L)^d$, and is applied d times. Finally, in a seasonal model there are seasonal counterparts of the AR, MA

and differencing operator given by $\Phi(L) = 1 - \Phi_1 L^s - \ldots - \Phi^{Rs} L^{Rs}$, $\Theta(L) = 1 - \Theta_1 L^s - \ldots \Theta_q L^{Qs}$, and $\nabla_s^D = (1 - L^s)^D$ respectively. The seasonal polynomials differ as the lags are at multiples of the seasonal period, s. We propose the following two-step algorithm to determine the best predictors and serial correlation structure of the regression residuals.

First, we seek suitable predictors for the model. Fix the sparsity k and use the data

$$(Y_1, X_{1,1}, \ldots, X_{P,1}), \ldots, (Y_M, X_{1,M}, \ldots, X_{P,M})$$

to determine a suitable set of predictors by solving the SBS problem. Given initial estimates of the coefficients $\hat{\beta}_{1,1}^{k,0}, \ldots, \hat{\beta}_{P,M}^{k,0}$, obtain the observed residuals for each model

$$\hat{\eta}_{t,m}^{k,0} = y_{t,m} - \sum_{p=1}^{P} x_{t,p,m} \hat{\beta}_{p,m}^{k,0}.$$

Now we need to estimate the serial correlation structure of the regression residuals. Given a list $\mathcal{L}$ of suitable SARIMA models, these models can be fit to the observed regression residuals $\hat{\eta}_{t,m}^{k,0}$ for $m = 1, \ldots, M$. The best SARIMA model can be identified, for example, based on information criteria. We require the transformed data

$$\frac{\nabla^{\hat{d_m}} \nabla_s^{\hat{D_m}} \hat{\phi}_m(L) \hat{\Phi}_m(L^s)}{\hat{\theta}_m(L) \hat{\Theta}_m(L^s)} y_{t,m} = \tilde{y}_{t,m} \quad \text{and} \quad \frac{\nabla^{\hat{d_m}} \nabla_s^{\hat{D_m}} \hat{\phi}_m(L) \hat{\Phi}_m(L^s)}{\hat{\theta}_m(L) \hat{\Theta}_m(L^s)} x_{t,p,m} = \tilde{x}_{t,p,m} \tag{3.3.9}$$

for $m = 1, \ldots, M$ and $p = 1, \ldots, P$. Consider fitting the the SARIMA model (3.3.8b) to obtain the observed model errors $\hat{\epsilon}_{t,m}$,

$$\hat{\eta}_{t,m} \frac{\hat{\nabla}^{d_m} \hat{\nabla}_s^{D_m} \hat{\phi}_m(L) \hat{\Phi}_m(L^s)}{\hat{\theta}_m(L) \hat{\Theta}_m(L^s)} = \hat{\epsilon}_{t,m}.$$

This process can be applied to (3.3.9) to obtain $\tilde{y}_{t,m}$ and $\tilde{x}_{t,p,m}$ for $m = 1 \ldots, M$ and $p = 1, \ldots, P$. Then, the predictors can be re-selected by solving the SBS problem again, but with the filtered data $\tilde{y}_{t,m}$ and $\tilde{x}_{t,p,m}$. This procedure can be iterated until convergence in the regression estimates, selected predictors, and the models for serial correlation. Let $\boldsymbol{\beta}^i$, and $p^i, d^i, q^i, P^i, D^i, Q^i$ denote the estimates of the regression coefficients and SARIMA model order at iteration i. In addition to this, let $\mathcal{I}^i$ denoted the indices of the selected predictors at iteration i then, the we say that the algorithm converges if the following hold

- $\sum_{m=1}^{M} \sum_{p=1}^{P} |\beta_p^i - \beta_p^{i-1}| \leq \epsilon$.
- $\{p^i \equiv p^{i-1}\} \bigcap \{d^i \equiv d^{i-1}\} \bigcap \{q^i \equiv q^{i-1}\} \bigcap \{P^i \equiv P^{i-1}\} \bigcap \{D^i \equiv D^{i-1}\} \bigcap \{q^i \equiv Q^{i-1}\}$.
- $\mathcal{I}^i \equiv \mathcal{I}^{i-1}$.

If the procedure does not converge an upper limit to the number of iterations can be considered. We have observed that convergence often occurs after two iterations.

In the following section we demonstrate the improvements in predictor selection using our simultaneous approach and show how the two-step method can improve the variance and accuracy of predictor selection in the presence of serial correlation.

3.4 Simulation study

In this section we evaluate the performance of the Simultaneous Best-Subset (SBS) and two-step approaches for predictor selection. In particular, we compare how the SBS approach, which estimates a system of linear regression models (3.3.1) compares to the best-subset approach which estimates each model in a system individually. We then show how predictor selection accuracy can be improved using the two-step approach when the regression residuals are serially correlated. Following this, we compare the SBS approach to some of the methods discussed in Section 3.2. In the final part of this section we consider the computational demands of the SBS approach.

We generate synthetic data from the system of linear regression models (3.3.1) where we fix the regression coefficients such that

$$\beta_{p,m} = \begin{cases} 0.3, & \text{for } p = 17, \\ 1, & \text{for } p = 18, \\ 0.6, & \text{for } p = 19, \\ 0, & \text{otherwise,} \end{cases} \quad \text{for all } m.$$

The predictors associated to response variable Y_m are generated such that

$$\boldsymbol{X}_m \sim \text{MVN}_{35}(\mathbf{0}, \boldsymbol{\Sigma}) \quad \text{where} \quad \mathbf{0} \in \mathbb{R}^{35} \quad \text{and} \quad \boldsymbol{\Sigma} := (\boldsymbol{\Sigma})_{i,j} = \rho^{|i-j|} \quad \text{for } m = 1, \ldots, M.$$

The total number of regression models in the system, M will be made clear. We use $P = 35$ predictor variables as provably optimal solutions to the SBS problem can be obtained within seconds for all $k \in \{1, \ldots, 35\}$. We include predictors X_{17}, X_{18} and X_{19} so that for large values of ρ these predictors are highly correlated and hard to distinguish amongst the other predictors. This makes the task of identifying the true predictors challenging. Unless otherwise stated the regression residuals are generated such that

$$\eta_{t,m} \sim \text{N}(0, \sigma_\eta^2) \quad \text{for } m = 1, \ldots, M.$$

The variance of the regression residuals σ_η^2 will be made clear where relevant.

For each approach, we are particularly interested in the number of correct predictors which have been selected. In a simulation of size N, we may compute the proportion of times that an approach correctly identifies the correct subset of predictors for the system. For each simulation we will

produce a value in the interval $[0, 1]$. A value of 1 indicates that the approach correctly identified the predictors for each model in the system. A value of $\frac{m}{M}$ indicates that the approach identified the correct subset of predictors for m of the M models in the system. In addition to the proportion of times an approach correctly identifies predictors, we are also concerned with with the predictive performance of the approach. We measure this using the mean-squared error of prediction. Let $\hat{\boldsymbol{\beta}}$ denote an estimate of the regression coefficients for the system of models 3.3.1. We define the mean-squared prediction error for model m as

$$\text{MSE}^m_{\text{pred}}(\hat{\boldsymbol{\beta}}) = \frac{1}{T}\sum_{t=1}^{T}\left(y_{t,m} - \hat{y}_{t,m}\right)^2.$$

Here, $\hat{y}_{t,m}$ is the predicted value of $y_{t,m}$. We define the mean-squared prediction error of the system as

$$\text{MSE}_{\text{pred}}(\hat{\boldsymbol{\beta}}) = \frac{1}{M}\sum_{m=1}^{M}\text{MSE}^m_{\text{pred}}(\hat{\boldsymbol{\beta}}).$$

By considering the error in estimating the coefficients we can determine both the predictor selection accuracy and also the predictive power. Small values in estimation error will only be obtained if the coefficients that should be zero are zero, and the coefficients that shouldn't be zero are close to their true values. We define the mean-squared estimation error for model m, and the mean-squared error in estimation of the system respectively as

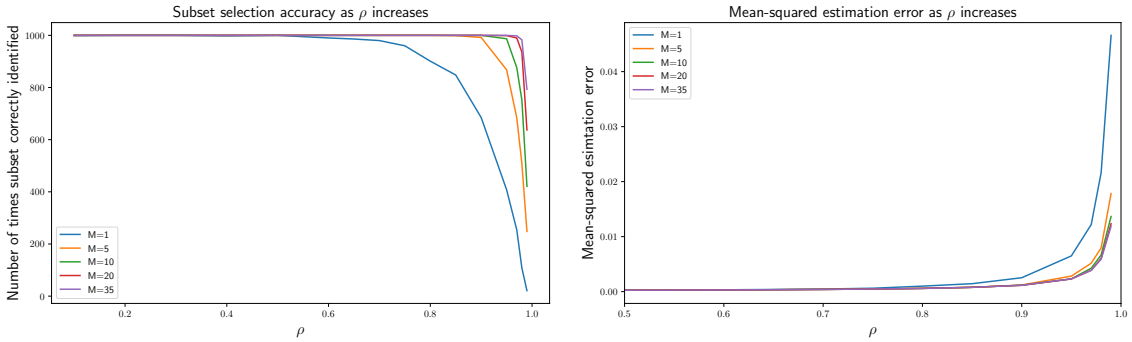
$$\text{MSE}^m_{\text{est}}(\boldsymbol{\beta}) = \frac{1}{P}\sum_{p=1}^{P}\left(\beta_{p,m} - \hat{\beta}_{p,m}\right)^2 \quad \text{for } m = 1, \ldots, M \quad \text{and } \text{MSE}_{\text{est}}(\boldsymbol{\beta}) = \frac{1}{M}\sum_{m=1}^{M}\text{MSE}^m_{\text{est}}(\boldsymbol{\beta}).$$

Now that we have discussed the main criteria used to assess the performance of the approaches we proceed with the evaluation of each approach. First, we consider the gains from simultaneous predictor selection.

3.4.1 Simultaneous selection

The SBS approach was proposed to jointly estimate and select predictors for a system of linear models. Here, we compare the performance of the SBS approach to the best-subset approach as we increase M. We show that both predictor correlation and the variance of the regression residuals affects the performance of the approaches and highlight the extent to which each approach is affected. We generate 1000 synthetic datasets as described in Section 3.4 and fix the residual variance such that $\text{Var}(\eta_{t,m}) = 1$ for $m = 1, \ldots, M$. We observe the predictor selection accuracy and the mean-squared estimation of the system when both approaches are applied with $k = 3$. This corresponds to the true model sparsity.

Figure 3.4.1a shows that selection accuracy for the best-subset method (M=1) deteriorates rapidly as the predictor correlation (ρ) exceeds 0.5. However, simultaneous predictor selection with



(a) Selection accuracy.

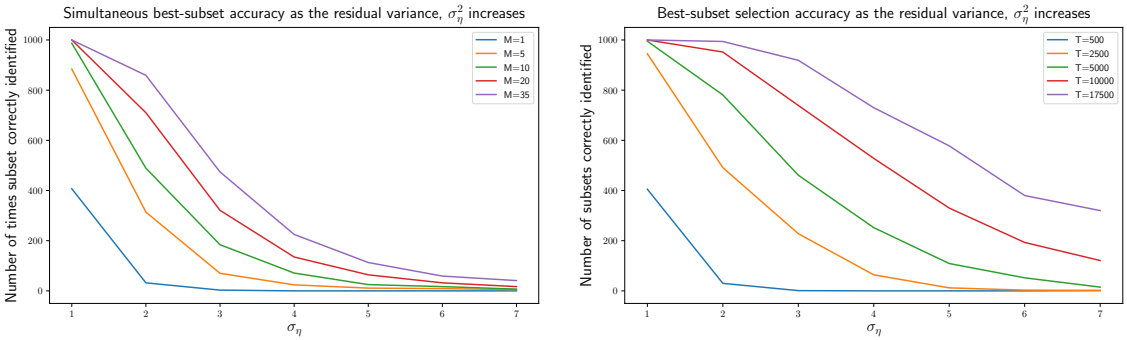
(b) Mean-squared estimation error.

Figure 3.4.1: Predictor selection accuracy and the mean-squared estimation error for the system as the predictor correlation increases.

$M = 5$ appears to accurately select predictors until ρ exceeds 0.87. As the number of models in the system increases the threshold at which predictor selection accuracy deteriorates appears to increase. There appears to be some consistency in the selection accuracy of the SBS approach as the number of models in the system increases. As a consequence of improved selection accuracy, the mean-squared estimation error for the system decreases as M increases. The mean-squared estimation error is shown in Figure 3.4.1b.

We now compare the performance of the SBS approach and the best-subset approach as the variance of the regression residuals increases. The same residual variance is used for each model in the system. Here, we fix the predictor correlation $\rho = 0.95$. Again, we simulate 1000 synthetic datasets and observe the predictor selection accuracy and mean-squared estimation error of the system when $k = 3$. Figure 3.4.2a shows that the best-subset approach is unable to identify the correct subset of predictors when $\sigma_{\eta_m} > 3$ for $m = 1, \ldots, M$. As the variance of the residuals increases the accuracy of the SBS approach deteriorates. However, as the number of models in the system increases the accuracy of the SBS approach improves.

The SBS approach was proposed to improve estimation accuracy for a system of related linear regression models. We have seen in Figures 3.4.1a and 3.4.2a that the accuracy of the SBS approach appears to improve with the number of regression models in the system. The improved accuracy of the SBS approach over the best-subset approach may be explained as the SBS approach uses *more information* to fit a single regression model. We now compare the performance of the SBS approach to the best-subset approach where each method uses the same number of observations. Figure 3.4.2b shows the selection accuracy of the best-subset approach using MT observations. Consider the line in Figure 3.4.2b corresponding to $T = 2500$. This can be compared to the line in Figure 3.4.2a corresponding to $M = 5$ as 500 observations were generated for each of the 5 response variables.



(a) The selection accuracy of the SBS method. (b) Selection accuracy of the best-subset method.

Figure 3.4.2: Predictor selection accuracy as the variance of the residuals increases. We compare the SBS approach using MT observations where there are M response variables each with T observations to the best-subset approach with one response variable which has MT observations.

Effectively, each approach uses 2500 observations but the selection accuracy for the SBS approach is not as accurate. In practise, we are typically limited to the number of observations for each response variable. We now show how our simultaneous shrinkage operator proposed in Section 3.3.1 may further improve estimation accuracy.

Simultaneous shrinkage

Here, we investigate the impact of the simultaneous shrinkage estimator on the estimates of the regression coefficients and the predictive performance of the models. We fix $M = 5$ and simulate 750 observations for each response variable and their associated predictors from the model defined in Section 3.4. We split the data randomly into two sets. We use 500 observations for each response variable as a training set to estimate the models. The remaining 250 observations for each response variable are used to determine the predictive accuracy of the models. We fix $\rho = 0.95$, $k = 3$ and $\sigma^2_{\eta_m} = 2$ for $m = 1, \ldots, M$.

Figure 3.4.3 shows the trace-plots of the regression coefficients for each of the five models in the system as the value of the simultaneous shrinkage penalty increases. As λ increases, the simultaneous best-subset changes a total of three times. Initially a noisy predictor, 21 is included into the model (shown by non-zero red trace). Then, predictor 21 is dropped for 27, this is then reversed, before predictor 21 is then dropped for the true predictor, 17. The horizontal lines show the coefficients of predictors 17, 18 and 19. Some coefficient estimates start far from the true value, see for example $\beta_{19,1}, \beta_{18,2}$ and $\beta_{18,5}$. But, as the shrinkage penalty increases the estimates of each coefficient appear to eventually approach the true values.

Shrinking the coefficients from each model to a common value increases the in-sample error.

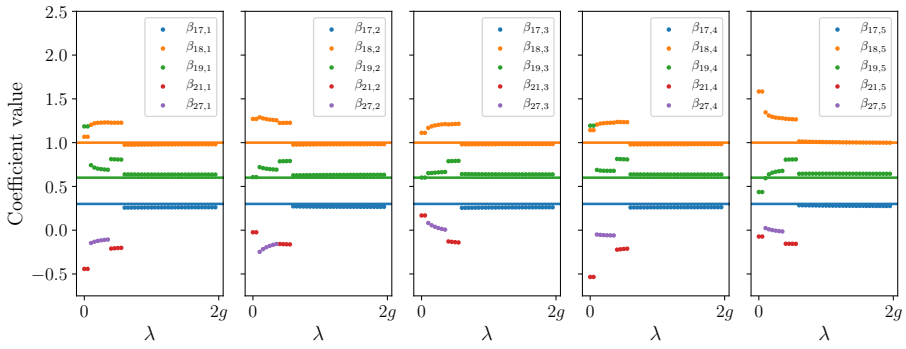
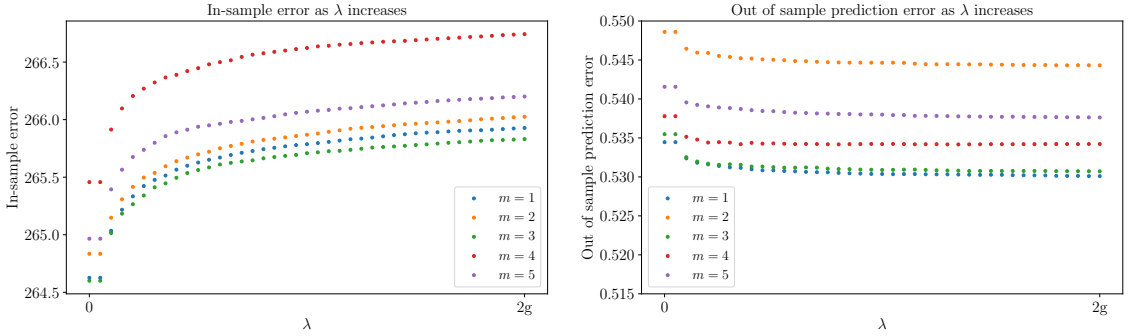


Figure 3.4.3: Trace plot of the regression coefficients as the shrinkage parameter λ is increased, penalising dissimilarities in $\beta_{p,1}, \ldots, \beta_{p,M}$ for $p = 1, \ldots, 35$. Here, we solve the SBS problem with simultaneous shrinkage with $k = 3$.

Figure 3.4.4a shows the mean-squared prediction error for the system on the data used to estimate the model as the shrinkage penalty increases. However, as the coefficients approach the true values this reduces the mean-squared prediction error of the held-out sample. This is shown in Figure 3.4.4b.



(a) In-sample prediction error (b) Out-of-sample prediction error

Figure 3.4.4: In-sample and out-of-sample mean-squared prediction error of the system as the shrinkage penalty increases.

In this section we have shown that estimation accuracy of the regression coefficients can be improved with simultaneous predictor selection. This in part may be as the simultaneous shrinkage operator can help identify the predictors that should be in the model. Consequently, this can lead to a reduction in prediction error. Now we investigate how the autocorrelation in regression residuals can affect predictor selection accuracy and how the two-step procedure can be used to improve selection accuracy.

3.4.2 Application to serially correlated data

In Section 3.3.2 we motivated the need to consider serial correlation in the regression residuals. Here, we compare the predictors selected using the SBS approach where the serial correlation in the observed residuals is ignored, to using the SBS approach in the two-step procedure discussed in Section 3.3.2. We simulate data from the system of models (3.3.1) where $M = 5$ and impose the following correlation structure on the residuals

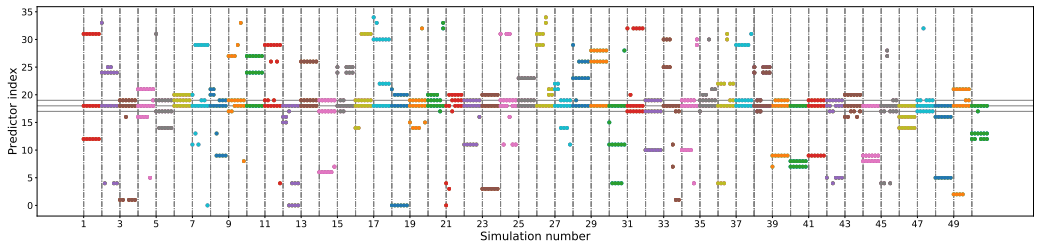
$$\eta_{t,m} = 0.9\eta_{t-1,m} + e_{t,m} \quad \text{for } m = 1, \ldots, 5. \tag{3.4.1}$$

Here, $e_{t,m} \sim \mathrm{N}(0, 1)$ are simulated independently for $m = 1, \ldots, 5$. The regression coefficients and predictors are the same as those given in Section 3.4. We simulate 600 observations and observe the predictors selected for each approach using the first 500 observations, the first 520 observations, and so on, until all 600 observations are used. Each method will be applied a total of 6 times. Our industrial collaborator observed that the selected predictors often change with small changes in the data. By increasing the number of observations used by the SBS approach we can determine if the high variation in the selected predictors can be reduced with the two-step approach.

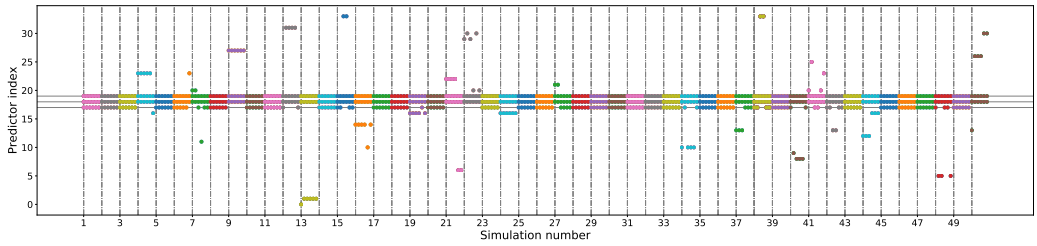
Figure 3.4.5 shows the trace-plots of selected predictors for each approach using the first 50 datasets. We simulated a total of 500 datasets but Figure 3.4.5 shows only the first 50 for clarity. The results of the remaining datasets were found to be similar. The dots between each pair of vertical lines corresponds to a single dataset. The vertical triplet of dots indicates the selected predictors using T observations. The left-most vertical triplet between each pair of vertical lines indicates the predictors selected using 500 observations. From left to right, the vertical triplets indicate the selected predictors using $T = 500, 520, \ldots, 600$ observations from that dataset.

Figure 3.4.5a shows that when the serial correlation in the regression residuals is ignored there is large variation in the selected predictors. There are few datasets where the best-subset approach could correctly identify the true predictors. The true predictors are indicated by the horizontal lines at 17,18 and 19. Further, the subset selected by the SBS approach frequently changes as the number of observations increases. This can be seen by the level of the dots changing from left to right within each pair of vertical lines. When the serial correlation in the regression residuals is addressed variation in the selected predictors is much reduced. Figure 3.4.5b shows the predictors selected using the SBS approach within the two-step algorithm. We can clearly see that the two-step approach correctly identifies the true subset more often. Further, there are fewer datasets where the predictors selected change as the number of observations is increased. Although this behaviour is still observed in the two-step approach, it is far less frequent.

It is possible to recover the true correlation structure of the regression residuals. Recall from



(a) Unfiltered predictor selections



(b) Filtered predictor selections

Figure 3.4.5: A comparison of the iterative approach which adjusts for serial correlation (b), and the standard approach that ignores serial correlation in the regression residuals (a).

Section 3.3.2 that we fit multiple SARIMA models to the regression residuals observed after estimating the regression coefficients. We select the SARIMA model with the lowest value of the BIC (Schwarz, 1978). Figure 3.4.6 shows that we can often recover the true correlation structure. Each of the 5 rows in Figure 3.4.6 indicates the results for each of the 5 response variables. The vertical axis indicates each of the SARIMA p, d, q, P, D, Q orders. Each pair of vertical lines indicates a dataset, similar to Figure 3.4.5. If '.' appears in Figure 3.4.6 on the row corresponding to p it indicates that p was correctly identified. If an integer appears in place of '.', then this is the value fit in error. Occasionally the wrong SARIMA model was fit to the residuals, but this did not appear to adversely affect the selected predictors.

In the following section we compare the SBS approach to alternative approaches from the literature. In addition to this, we modify one approach to select predictors simultaneously for a system of linear regression models to give a comparison to an alternate simultaneous procedure.

3.4.3 Comparison to other approaches

In this section we generate data for a system of models (3.3.1) where $M = 5$ and compare the models fit by the LASSO (Tibshirani, 1996), the elastic-net (Zou and Hastie, 2005) and stepwise selection (Miller, 2002), to the SBS approach and a modified version of the Simultaneous Variable Selection (SVS) approach proposed by Turlach et al. (2005). In each simulation we generate 1000 observations

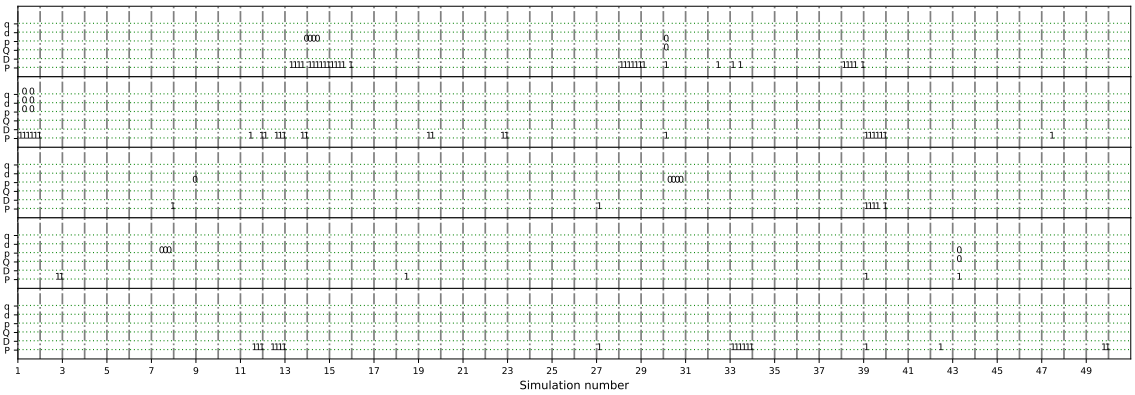


Figure 3.4.6: Indicating if the true time SARIMA model orders were identified in each application of the two-step SBS algorithm.

of each response variable and the associated predictors. The observations are randomly divided into train/test/validation sets to the proportions 50%/25%/25% respectively.

We use the training data to estimate the models. The stepwise method is implemented using the `stats::step` (R Core Team, 2019) function and automatically selects a model using the AIC (Akaike, 1973). The LASSO (Tibshirani, 1996) and its generalisation, the elastic-net (Zou and Hastie, 2005) require tuning parameters to be determined. We determine the tuning parameters as follows. First, we apply each method to the training data for each response variable for a range of tuning parameter values. We then select the model for each response variable that has the lowest mean-squared prediction error on the test dataset. We apply the elastic-net using $\alpha = 0, 0.1, \ldots, 1$, and for 100 values of the shrinkage penalty, λ. Note that $\alpha = 1$ gives the LASSO. The elastic-net is applied using the `glmnet` (Zou and Hastie, 2018) package in `R`.

Details of how we modified the SVS method are given in Appendix 3.A. The SVS approach was proposed for exploratory analysis in selecting predictors for multi-response models (Breiman and Friedman, 1997), but we modify this approach to estimate a system of linear regression models (3.3.1) and consider this modified approach in its own right. We apply the modified SVS method with 100 values of the tuning parameter. We can force the coefficients estimated using the modified SVS method to take positive values only. The results for this approach will be presented as $\texttt{SVS}^{+}$. The SBS approach is implemented by generalising the MIQO program (3.2.5) described in Section 3.3. We include the constraints (3.2.6) that exclude pairs of highly correlated predictors with correlation exceeding 0.8. We consider $k = 1, \ldots, K_{\max}$ where $K_{\max}$ is determined automatically using the procedure described in Section 3.2.1. The mathematics programs formulated for both the modified SVS and SBS approach were solved using Gurobi (Gurobi Optimization, 2018).

The modified SVS and SBS approaches fit the models for each response variable simultaneously.

Therefore, we select the models for each response variable simultaneously. Consider the 5 regression models obtained simultaneously for each value of tuning parameter for both the SBS and SVS approaches, as a model for the system. Then, we select the model for the system by selecting the models with the lowest mean-squared prediction error for the system on the test data.

In this simulation we use groups of highly correlated predictors as we expect groups of highly correlated predictors in our telecommunications application. The predictors are denoted

$$\boldsymbol{X} = [\boldsymbol{X}_{(1)}, \boldsymbol{X}_{(2)}, \boldsymbol{X}_{(3)}, \boldsymbol{X}_{(4)}, \boldsymbol{X}_{(5)}] \in \mathbb{R}^{35}.$$

Here, $\boldsymbol{X}_{(b)}$ corresponds to the predictor group b. Group b of predictors contains $b+4$ predictors such that

$$\boldsymbol{X}_{(b)} = [X_{(b),1}, \ldots, X_{(b),b+4}] \quad \text{for } b = 1, 2, 3, 4, 5.$$

The group sizes are 5,6,7,8 and 9 respectively. Each group contains highly correlated predictors such that

$$\boldsymbol{X}_{(b)} \sim \text{MVN}(\boldsymbol{0}_{b+4}, \boldsymbol{\Sigma}_{(b)}) \quad \text{where} \quad \boldsymbol{\Sigma}_{(b),i,j} := 0.95^{|i-j|}.$$

Here, $\boldsymbol{0}_{b+4} \in \mathbb{R}^{b+4}$ is a vector of zeros. We use a predictor from each group to generate the response variables such that the regression coefficients are given by

$$\beta_{p,m} = \begin{cases} 1, & \text{if } p = 30, \\ 0.775, & \text{if } p = 25, \\ 0.55, & \text{if } p = 14, \\ 0.325, & \text{if } p = 5, \\ 0.1, & \text{if } p = 2, \\ 0, & \text{otherwise,} \end{cases} \quad \text{for } m = 1, \ldots, 5. \tag{3.4.2}$$

The variance of the regression residuals is such that $\text{Var}(\eta_{t,m}) = 2$ for $m = 1, \ldots, 5$.

We average a number of performance criteria over 50 simulations. The average mean-squared prediction error for the system on the validation data is represented by *MSE*. The average model sparsity is shown by *Sparsity*. This sparsity measure may be misleading in terms of indicating whether an approach could often identify the true model sparsity. For this reason we also consider the average number of false negatives (F^-), the number of predictors that should have been selected but were not. In addition to this we show the average number of false positives (F^+), the average number of predictors that should have been selected, but were not. We define a *naive* similarity measure

$$\frac{1}{MP} \sum_{m=1}^{M} \sum_{p=1}^{P} \left(\beta_{p,m} - \frac{1}{M} \sum_{m=1}^{M} \beta_{p,m} \right)^2$$

which provides an estimate of the across model variation in the regression coefficients. The average time to implement each approach is indicated by *Time* and the average number of models containing the true subset of predictors is shown by *True Subset*. Finally, the proportion of models containing negative coefficients is shown by *Neg Coef*.

Table 3.4.1: Summary measures of the predictor selection algorithms

	MSE	Sparsity	Time	Similarity	False$^-$	False$^+$	Neg Coef	True Subset
LASSO	4.09	12.82	**0.54**	0.11	1.07	8.89	0.61	0
SBS	**4.01**	**4.84**	18.33	**0.001**	0.84	**0.68**	**0**	**0.25**
`SVS`	4.04	17.20	2.81	0.010	**0.47**	12.67	0.90	0
`SVS`$^+$	4.03	13.96	3.05	0.008	0.52	9.48	**0**	0
Stepwise	6.23	7.28	2.01	0.149	1.73	4.01	0.83	0

Table 3.4.1 shows the results. Note that in each application of the elastic-net the best performing model corresponds to $\alpha = 1$ giving the LASSO. The SBS approach appears to produce the sparsest models. In addition to this, the SBS approach also appears to include the lowest number of false positives which may suggest that the SBS approach can accurately select a subset of the true predictors. After some investigation we did notice that often predictor 2 was not selected. Considering the coefficients of the models given in 3.4.2 we can see that the coefficient of predictor 2 is the smallest and may be hard to identify given the noise.

The SBS approach did however take the longest time to implement on average but does produce the models with the lowest prediction accuracy. The mathematical programming approach allows us to only accept models with positive coefficients for the SBS and `SVS`$^+$ methods and we can see that all other approaches contain a high proportion of models with negative coefficients. Only the SBS approach was able to identify the correct subset and it did this only 25% of the time. Despite low false negative values of the other approaches the high false positive values may explain why the other approaches were not able to identify the correct subset. Finally, the SBS approach also provided system of models whereby the coefficients for each model were most similar. The univariate stepwise and LASSO approaches produced system of models with highly varied coefficients across models. This may be explained by large variations in the selected predictors across the models.

3.4.4 Computational aspects

In Section 3.2.1 we discussed a number of approaches that can be used to ensure the SBS approach is computationally feasible. Here, we are interested in a worst case scenario and consider how the SBS approach scales with the number of predictors and number of models in the system. We solve

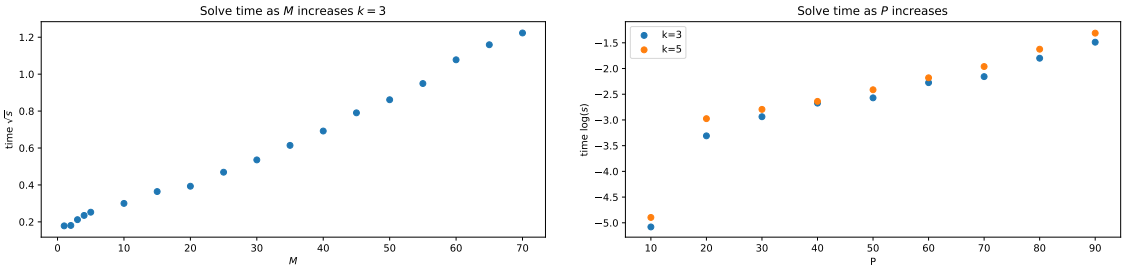
the SBS problem (3.3.2) by generalising the MIQO program (3.2.5) but do not consider any of the extensions discussed in Section 3.3.1. In this simulation study all data is generated as follows,

$$\boldsymbol{X}_m \sim \text{MVN}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \quad \text{where } \boldsymbol{\mu} = [0, \ldots, 0] \in \mathbb{R}^P \quad \text{and } \boldsymbol{\Sigma}_{i,j} = 0.25^{|i-j|}.$$

The number of response variables and the number of predictors will be made clear where relevant. The regression coefficients are given by

$$\beta_{p,m} = \begin{cases} 1, & \text{if } p = 1, 3, 5, \\ 0, & \text{otherwise,} \end{cases} \qquad \text{for } m = 1, \ldots, M.$$

We generate $T = 500$ observations for each response variable and its associated predictor variables and average the results over 50 simulations.



(a) SBS approach scaling with M (b) SBS approach scaling with P

Figure 3.4.7: Scaling of the SBS approach as the number of regression models in the system (M) increases and the number of predictors (P) increases, respectively.

Figure 3.4.7a shows how the SBS approach scales as the number of models in the system increases. Here, we fix $P = 35$ and solve the SBS problem when $k = 3$. There is a near linear trend for the solve time on the square root scale. This suggests that the time to solve the SBS problem scales quadratically with M. Figure 3.4.7b shows how the time to solve the SBS problem scales with P. Here, we fix $M = 5$ and solve for both $k = 3$ and $k = 5$. The SBS approach appears to scale exponentially with P. This can be seen by the near linear trend when $M > 10$ with time on a logarithmic scale. We also see an increase in the solve time when $k = 5$ in comparison to $k = 3$ in agreement with our observations in Figure 3.2.1.

In this section we have shown empirically that the SBS approach seems to have consistency in predictor selection as the number of regression models in a system increases. We have shown that the simultaneous shrinkage operator can improve the coefficient estimates by encouraging the coefficients between models in the system to common values. We have also shown that better selection accuracy can be obtained with a two-step procedure that accounts for serial correlation in

the regression residuals. We will now apply our approach to an example from the telecommunication events dataset.

3.5 Data study

The daily events in a telecommunications network are recorded by *type* and *location* within the network. Each type of event may be influenced by a different set of predictors. In the application presented here, location corresponds to a geographic location, but more detailed information, such as the location within the network, is available in other datasets. We use three response variables of the same type, from locations considered to be suitable for joint modelling. We use five groups of predictor variables. The first four groups of predictors are derived from transformations applied to external predictors. The last group relates to indicator variables to adjust for calendar affects.

We present three approaches for modelling the event data. The first approach, which we refer to as *Automated*, is our joint approach for selecting predictors simultaneously for multiple response variables using our two-step procedure to estimate a model for the regression residuals. The second approach (*Individual Automated*) uses the *Automated* approach but is applied to each response variable individually. Consequently the *Individual Automated* approach cannot take advantage of simultaneous predictor selection. We present the *Individual Automated* approach to clearly highlight the gains in simultaneous predictor selection. The final approach (*Baseline*) is the current approach adopted by our industrial collaborator. This approach removes the weekly seasonality and calendar effects from the response variables as part of a data pre-processing step. It can be quite a time-consuming process to determine how best to remove the seasonality and calendar effects. There are various ways of achieving this, see for example Hyndman and Athanasopoulos (2019). It is up to the analyst to determine the *best* procedure to employ. This is subjective and assumes that the weekly seasonality and bank holiday effects are estimated without error. Ignoring estimation error may cause predictions made from the models to be misleadingly accurate. The current procedure is included as a baseline comparison. There are a total of 1396 daily observations, corresponding to about 3 years 9 months of data.

The estimated regression coefficients for the three approaches are given in Table 3.5.1. It is clear from Table 3.5.1 that the *Automated* and *Individual Automated* approaches produce models that are much sparser than those produced by the *Baseline* approach, not considering the calendar effects. All coefficients produced from the *Automated* and *Individual Automated* approaches have positive coefficients which would be expected from these variables, with the exception of the calendar effect variables which are negative. The *Baseline* approach includes highly correlated predictors from the

same group and with opposing effects. All six transformations of Predictor 3 are included. Both large negative and large positive coefficients appear for the predictors in Group 3 for the *Baseline* approach. This appears to be the behaviour of the least squares estimator, discussed by Hastie et al. (2008). Using simultaneous predictor selection and constraining the sign of the coefficients we are able to select the single best transformation of the base predictor used to produce Group 3.

Table 3.5.1: Regression coefficients for the Automated, Individual Automated and Baseline procedures. Each column represents the three different response variables for each method. The rows determine the predictor variables. The dashes indicate that the coefficient was exactly zero and hence the associated predictor was not selected.

		Automated			Individual Automated			Baseline		
Predictor	Coefficient	(m)			(m)			(m)		
Group		1	2	3	1	2	3	1	2	3
1	$\beta_{1.1,m}$	-	-	-	-	-	-	-	-	-
	$\beta_{1.2,m}$	-	-	-	-	-	0.01	-	-	0.01
	$\beta_{1.3,m}$	0.01	0.01	0.01	0.01	0.01	-	0.01	0.01	-
2	$\beta_{2.1,m}$	-	-	-	-	-	-	-	-	-
	$\beta_{2.2,m}$	0.02	0.02	0.01	0.02	0.02	0.01	0.03	0.02	0.03
	$\beta_{2.3,m}$	-	-	-	-	-	-	-0.02	-0.02	-0.03
3	$\beta_{3.1,m}$	-	-	-	-	-	-	-0.03	-0.01	-0.02
	$\beta_{3.2,m}$	-	-	-	-	-	-	0.21	1.12	0.13
	$\beta_{3.3,m}$	0.06	0.05	0.05	0.06	0.05		-1.96	-4.55	-0.86
	$\beta_{3.4,m}$	-	-	-	-	-	-	7	6.49	1.59
	$\beta_{3.5,m}$	-	-	-	-	-	-	-9.87	-3.03	-0.77
	$\beta_{3.6,m}$	-	-	-	-	-	0.09	4.82	-0.00	-
4	$\beta_{4.1,m}$	-	-	-	-	-	-	-	-	0.01
	$\beta_{4.2,m}$	-	-	-	-	-	-	-	-	-
	$\beta_{4.3,m}$	0.03	0.02	0.01	0.03	0.02	0.01	0.02	0.03	-
5	$\beta_{5.1,m}$	-0.77	-0.78	-0.65	-0.77	-0.78	-0.64	-	-	-
	$\beta_{5.2,m}$	-0.73	-0.79	-0.68	-0.73	-0.79	-0.68	-	-	-
	$\beta_{5.3,m}$	-0.27	-0.27	0.24	-0.27	-0.27	0.24	-	-	-

The mean squared errors for the 14 day-ahead predictions for the three approaches are given in Table 3.5.2. Recall that the Reg-SARIMA models explain the seasonality and calendar affects. They also describe the effects of other predictors. By selecting predictors simultaneously the *Automated*

approach provides more accurate forecasts of the response variables. Table 3.B.1 shows the estimates of the SARIMA coefficients for the *Automated* approach.

Table 3.5.2: MSE for the 14 day-ahead predictions for each of the three response variables and the three methods described in Section 3.5.

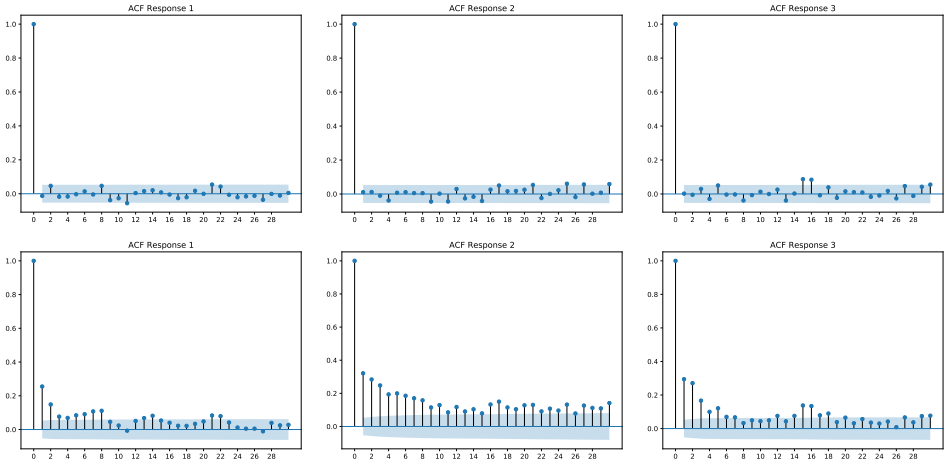
MSE Prediction (m)	Automated	Individual Automated	Baseline
(1)	**0.204**	**0.204**	0.280
(2)	**0.172**	0.173	0.314
(3)	**0.173**	0.182	0.212

We model the response variables using Reg-SARIMA models. The regression part of the model can explain the effect of predictors and the SARIMA part can explain seasonality and serial correlation. To determine whether the models produced by the *Automated* approach have adequately captured the serial correlation and seasonality within the data we can inspect the sample autocorrelation and sample partial autocorrelation functions of the model errors. The sample autocorrelation functions for the *Automated* and *Baseline* approaches are shown in Figure 3.5.1. There appears to be very little significant serial correlation in the model errors for the *Automated* approach. Modelling the regression residuals as a SARIMA process appears to account for most of the serial correlation. The *Baseline* approach would appear to violate the typical regression assumptions of independent regression residuals as there appears to be significant serial correlation at many lags in the regression residuals for all three response variables. Similar conclusions for the sample partial autocorrelation functions can be made, these are shown in Figure 3.5.2.

When serial correlation in the regression residuals is ignored the standard errors for each of the regression coefficients may be severely underestimated (Rawlings et al., 1998). This would raise suspicions about the significance of any predictor in the model. Further, prediction intervals are likely to be too narrow.

3.6 Conclusions and further work

Motivated by a real world industrial problem we have proposed a procedure to help automate the modelling process of telecommunications data. More specifically, we have developed a MIQO program to solve the simultaneous best-subset problem proposed in Section 3.3, to simultaneously select predictors when jointly modelling multiple response variables. We have integrated predictor selection within a two-step procedure, that iterates between selecting predictors for a regression model and modelling the serial correlation of the regression residuals. Automation is achieved by adding



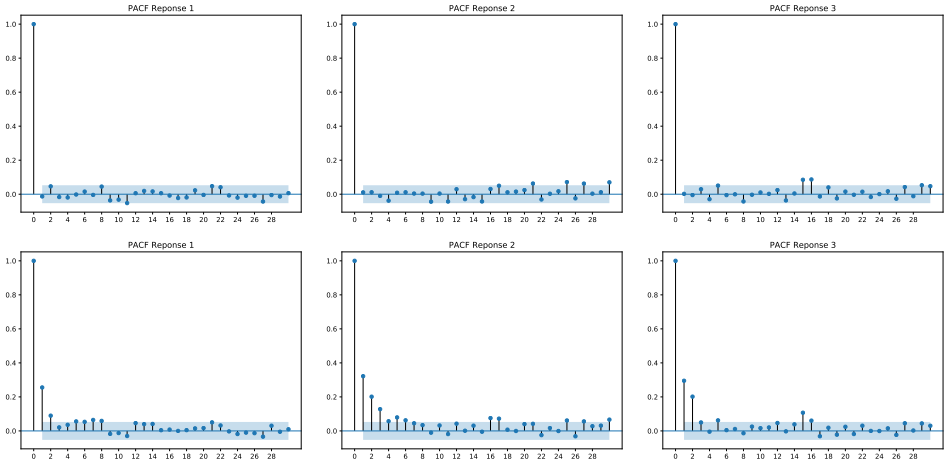
(a) $m = 1$ (b) $m = 2$ (c) $m = 3$

Figure 3.5.1: An estimate of the autocorrelation function for the fitted model errors for each of the three response variables. The estimates for both the *Automated* approach (top) and the *Baseline* approach (bottom) are shown. The vertical lines show an estimate of the autocorrelation at lag l. The uncertainty cloud shows the 95% confidence intervals where the standard deviation is calculated according to Bartlett's formula.

constraints to the MIQO program to ensure sensible models are produced and by eliminating the need to pre-process the data by modelling calendar affects and seasonality.

We have shown that predictor selection accuracy can be improved by simultaneously selecting predictors for multiple response variables. Selection accuracy and coefficient estimation can further be improved by shrinkage. The shrinkage we introduced is only possible when joint estimation of models is considered. In contrast to LASSO like penalties that shrink coefficients towards zero our shrinkage method forces coefficients between models towards a common value.

An interesting avenue for future research would investigate the impact of modelling the regression residuals simultaneously. We may consider modelling the regression residuals as a Vector Auto-Regression (VAR) which could explain both serial and cross correlations between the regression residuals from multiple models. We anticipate that prediction error may be reduced further as well as give a consistent form for the regression residuals between models.



(a) $m = 1$ (b) $m = 2$ (c) $m = 3$

Figure 3.5.2: An estimate of the partial autocorrelation function for the fitted model errors for each of the three response variables. The estimates for both the *Automated* approach (top) and the *Baseline* approach (bottom) are shown. The vertical lines show an estimate of the partial autocorrelation at lag l. The uncertainty cloud shows the 95% confidence intervals where the standard deviation is calculated as $\frac{1}{\sqrt{1396}}$.

3.A Implementing the modified SVS method

In this appendix we introduce the Convex Quadratic Program (CQP) introduced by Turlach et al. (2005) to solve the Simultaneous Variable Selection (SVS) problem. The SVS approach was proposed by Turlach et al. (2005) as an explanatory tool to help determine sets of suitable predictors for multi-response models (Breiman and Friedman, 1997). We modify the CQP used by Turlach et al. (2005) to produce feasible solutions to the SBS problem. The SVS problem is given by

$$\begin{aligned} \min_{\boldsymbol{\beta}} \left[\sum_{m=1}^{M} \sum_{t=1}^{T} \left(y_{t,m} - \sum_{p=1}^{P} x_{t,p} \beta_{p,m} \right)^2 \right] \quad \text{subject to,} \\ \sum_{p=1}^{P} \max \left(|\beta_{p,1}|, \ldots, |\beta_{p,M}| \right) \leq \nu. \end{aligned} \tag{3.A.1}$$

Here, M denotes the number of response variables considered for joint analysis. When $M = 1$, (3.A.1) gives the LASSO in *constrained form* (Tibshirani, 1996). We propose to modify the SVS problem given in (3.A.1) by replacing the objective with that used for the SBS problem. This gives

the following,

$$\begin{aligned} \min_{\boldsymbol{\beta}} \left[\sum_{m=1}^{M} \sum_{t=1}^{T} \left(y_{t,m} - \sum_{p=1}^{P} x_{t,p,m} \beta_{p,m} \right)^2 \right] \quad \text{subject to,} \\ \sum_{p=1}^{P} \max \left(|\beta_{p,1}|, \ldots, |\beta_{p,M}| \right) \leq \nu. \end{aligned} \tag{3.A.2}$$

The CQP formulated by Turlach et al. (2005) to solve the SVS problem is given by

$$\begin{aligned} \min_{\boldsymbol{\beta},\boldsymbol{z}} \left[\sum_{m=1}^{M} \sum_{t=1}^{T} \left(y_{t,m} - \sum_{p=1}^{P} x_{t,p} \beta_{p,m} \right)^2 \right] \quad \text{subject to,} \\ \boldsymbol{u}_M \otimes \boldsymbol{z} - \boldsymbol{\beta} \geq 0, \\ \boldsymbol{u}_M \otimes \boldsymbol{z} + \boldsymbol{\beta} \geq 0, \\ \nu - \boldsymbol{u}_P \boldsymbol{z} \geq 0. \end{aligned} \tag{3.A.3}$$

Here, $\boldsymbol{u}_j \in \mathbb{R}^j$, with each entry equal to 1 and $\boldsymbol{z} \in \mathbb{R}^P$ are auxiliary variables. We modify formulation (3.A.3) to solve problem (3.A.2) as follows,

$$\begin{aligned} \min_{\boldsymbol{\beta},\boldsymbol{z}} \left[\sum_{m=1}^{M} \sum_{t=1}^{T} \left(y_{t,m} - \sum_{p=1}^{P} x_{t,p,m} \beta_{p,m} \right)^2 \right] \quad \text{subject to,} \\ \boldsymbol{u}_M \otimes \boldsymbol{z} - \boldsymbol{\beta} \geq 0, \\ \boldsymbol{u}_M \otimes \boldsymbol{z} + \boldsymbol{\beta} \geq 0, \\ \nu - \boldsymbol{u}_P \boldsymbol{z} \geq 0. \end{aligned} \tag{3.A.4}$$

We must determine the maximum value of ν. We set $\nu_{\max} = \sum_{m=1}^{M} \sum_{p=1}^{P} |\hat{\beta}_{p,m}|$ where

$$\hat{\boldsymbol{\beta}} = \arg\max \left[\sum_{m=1}^{M} \sum_{t=1}^{T} \left(y_{t,m} - \sum_{p=1}^{P} x_{t,p,m} \beta_{p,m} \right)^2 \right].$$

All coefficients given by a solution to formulation (3.A.4) are non-zero. We apply the heuristic proposed by Turlach et al. (2005) to determine which predictors should be zero. Let

$$\mathcal{I} = \{p : \max\{|\beta_{p,1}|, \ldots, |\beta_{p,M}|\} > \nu \times 1e^{-4}, \text{ for } p = 1, \ldots, P\}.$$

Then $\mathcal{I}$ denotes the indices of the non-zero coefficients.

3.B Parameter estimates for the SARIMA residual models

Here, we provide the parameter estimates for the SARIMA models fitted to the regression residuals. A SARIMA (2,0,1)(1,0,1,7) model was selected for the residuals for Y_1, the coefficients are given in Table 3.B.1. For Y_2 and Y_3 the order of the SARIMA models for the residuals was (1,0,1)(1,1,1,7).

Table 3.B.1: Parameter estimates for the SARIMA models, fitted to the regression residuals for response variables, Y_1, Y_2 and Y_3. The dashes indicate the coefficient was exactly zero so the parameter was not present in the model used.

Parameter	Estimate		
	$(m = 1)$	$(m = 2)$	$(m = 3)$
$\phi_{m,1}$	0.914	0.797	0.221
$\phi_{m,2}$	-0.082	–	–
$\theta_{m,1}$	-0.708	-0.570	0.180
$\Phi_{m,1}$	0.057	0.043	0.022
$\Theta_{m,1}$	-0.979	-0.964	-0.972
σ_{η_m}	0.293	0.298	0.397

Chapter 4

Telecommunications event data case study

In Chapter 3 we compared the performance of models produced by our semi-automated two-step approach against the models produced by our industrial collaborator. This was achieved by fitting models to one group of response variables and observing various properties of the models fit. By excluding pairs of highly correlated predictors we were able to produce more interpretable models with the *Automated* approach, in which the sign of each regression coefficient was as expected. This approach typically produced models with fewer weather related predictors, greater predictive accuracy, and agreement amongst the selected predictors used in each model of response variables within a group. In this chapter we further validate the performance of the *Automated* approach by fitting models to all other groups of response variables.

In Section 4.1 we provide details of the dataset. In particular, we will present the groups of response variables and give some details on how the groups are determined. We will then discuss the predictors. In this chapter we have increased the total number of predictors considered by including lagged predictors which may provide useful information for predicting telecommunication events.

In Section 4.2 we give details on the four approaches used to model the telecommunication events. In Section 4.3 we will reiterate the aims and motivation for developing the *Automated* approach. We will assess how the *Automated* approach and others compare at satisfying these aims. We end this chapter by concluding our findings in Section 4.4.

4.1 Data description

Recall that we denote response variables, Y_m corresponding to the telecommunication event rates for a given event *type*, recorded at location m in the network. Associated with each response variable we have predictor variables denoted $X_{p,m}$ corresponding to the realisation of predictor p for response variable m. Further details of the response variables follow.

Response variables: The telecommunications event dataset comprises 36 response variables. Each response variable measures the rate of telecommunication events at a given location within the network. The number of events per day for a given event type is recorded and scaled by the number of active lines. Since the number of (*active*) lines providing a particular service changes through time it is important to scale the number of events by the number of *active lines*.

Each response variable is allocated to a *response group*. We have a total of seven response groups which we denote $\mathcal{G}_i$ for $i = 1, \ldots, 7$. The set $\mathcal{G}_i$ contains the indices of the related response variables within Group i. Table 4.1.1 shows each of the seven groups. In Section 1.1 we discussed that the effect of a predictor on events may vary depending on where the events occur. Here, each response group is determined by a geographical region in the UK.

Table 4.1.1: Allocation of the 36 response variables to the 7 response groups.

Response group	Number in group
$\mathcal{G}_1 = \{1, 2, 3, 4, 5, 6\}$	6
$\mathcal{G}_2 = \{7, 8, 9, 10, 11, 12\}$	6
$\mathcal{G}_3 = \{13, 14, 15\}$	3
$\mathcal{G}_4 = \{16, 17, 18, 19, 20\}$	5
$\mathcal{G}_5 = \{21, 22, 23, 24, 25\}$	5
$\mathcal{G}_6 = \{26, 27, 28, 29, 30, 31, 32\}$	7
$\mathcal{G}_7 = \{33, 34, 35, 36\}$	4

Predictor variables: Many of the predictors that we consider including in the models for response variables are derived from the measurements of weather variables. In Chapter 1 we discussed the importance of understanding the effect of external variables on the telecommunications network for the industry. The following weather variables are considered here,

1. Humidity: The mean relative humidity (gm^{-3}) over a 24-hour period.

2. Wind speed: The maximum recorded wind speed (mph) within a 24-hour period.

3. Precipitation: The total amount of rainfall (mm) within a 24-hour period.

4. Lightning: The total number of lighting strikes within a 24-hour period.

Non-linear transformations of the weather variables can often be more suitable to include in models for telecommunication events. Here, the non-linear transformation applied to the weather variables is the exponential smoothing function defined in equation (2.1.2). Suitable smoothing parameters for the exponential smoothing function were chosen such that the maximum pairwise correlation between the resulting predictors is around 0.97. We denote the set of predictors produced from applying the exponential smoothing function to a weather variable for a range of smoothing parameters by $\mathcal{T}_i$. The groups of predictor variables are shown in Table 4.1.2. We can see that group $\mathcal{T}_1$ consists of four predictors derived from applying the exponential smoothing function (given in Section 2.1) to the Humidity weather variable.

Table 4.1.2: The predictors used by our automated procedure grouped by transformation for the telecommunications event data.

Predictor Group	Predictor index	Number in group
$\mathcal{T}_1$ (Humidity)	1,2,3,4	4
$\mathcal{T}_2$ (Wind Speed)	5,6,7,8	4
$\mathcal{T}_3$ (Precipitation)	9,10,11,12,13,14	6
$\mathcal{T}_4$ (Lightning)	15,16,17	3
$\mathcal{L}_1$ (Humidity)	18,19,20,21,22,23,24	7
$\mathcal{L}_2$ (Lighting)	25, 26	2
$\mathcal{L}_3$ (Wind Speed)	27,28,29	3
$\mathcal{L}_4$ (Precipitation)	30,31,32,33,34,35,36	7
$\mathcal{B}$ (Bank holidays)	37,38,39	3

In addition to non-linear transformations of the weather variables, we also include lagged weather variables as predictors. For each weather variable we always include the lag 0. A group of predictors derived from lagging a weather variable is denoted $\mathcal{L}_i$. We can see from Table 4.1.2 that the lags included for precipitation are, 0,1,2,3,4,5 and 6. The number of predictors within each predictor group varies. The size of predictor group is determined by the associated weather variables along with the duration of its effect. For example, ground water levels may rise after prolonged rainfall. The effect of lightning strikes are thought to be more immediate. Therefore, there will be a large number of lagged variables for precipitation, and fewer for the lightning variable.

Finally, we use bank holiday indicators as predictors for the *Automated* approach. Bank holiday predictors are not provided for all other approaches as bank holiday effects are accounted for in

the pre-processing step. We consider a total of three bank holiday indicators. The bank holiday indicators are predictors 37, 38 and 39. Predictor 38 indicates Christmas bank holidays such that

$$x_{t,38,m} = \begin{cases} -1, & \text{if } t \text{ corresponds to Christmas Day, Boxing Day or any substitute,} \\ 0, & \text{otherwise.} \end{cases}$$

Predictor 39 indicates the Christmas-New Year period such that

$$x_{t,39,m} = \begin{cases} -1, & \text{if } t \text{ corresponds to any date between 27/12/yyyy to 1/1/(yyyy + 1),} \\ 0, & \text{otherwise.} \end{cases}$$

Here, we use *yyyy* to denote the years included in the telecommunications dataset. Finally, Predictor 37 takes the value -1 for any other bank holidays and zero otherwise.

In the following section we give details on the methods used to produce models for the telecommunication events.

4.2 Details of the implemented approaches

We compare the *Automated* approach to three alternatives. The current approach used by our industrial partner will be used to give a baseline comparison. We refer to this approach as the *Baseline* approach. We observed in Chapter 3 that the *Baseline* approach often includes many pairs of highly correlated predictors where the sign of the associated regression coefficients oppose. Therefore, we modify the *Baseline* approach by increasing the penalty used in the stepwise selection procedure. We call this modified approach the *Modified baseline* approach. The *Baseline* approach uses the AIC (Akaike, 1973) to terminate the stepwise algorithm, whereas the *Modified baseline* approach will use the BIC (Schwarz, 1978). We also apply the SBS approach to the data after pre-processing. The *Automated* approach differs from the *Baseline* approach in a number of ways. Firstly, it estimates models for the fault rates directly. Secondly, it fits a more general Reg-SARIMA model. We apply the SBS approach to the data after pre-processing, to highlight the differences between stepwise selection used in the baseline approach and simultaneous predictor selection. We call this approach the *Simultaneous baseline* approach. Further details of these approaches follow.

Baseline approach: This approach is currently used by our industrial collaborator. The telecommunication event rates and the associated predictor variables are pre-processed using the procedure described in Section 2.3. We denote the pre-processed response variable $\tilde{Y}_m$, and the observations $\tilde{\boldsymbol{y}}_m \in \mathbb{R}^{T\times 1}$. The associated observations of the predictors after pre-processing are denoted

$\tilde{\boldsymbol{x}}_m \in \mathbb{R}^{T\times P}$. A forward stepwise selection approach is used to estimate models of the form

$$\tilde{y}_{t,m} = \beta_{0,m} + \sum_{p=1}^{36} \tilde{x}_{t,p,m}\beta_{p,m} + \tilde{\eta}_{t,m} \tag{4.2.1}$$

for $m = 1, \ldots, 36$. Here, predictors $X_1, \ldots, X_{36}$ from the 39 predictors listed in Table 4.1.2 are considered for inclusion into the model. The forward stepwise algorithm is implemented in the `R` language using the `stats::step` function. The algorithm is initialised using a model including only the intercept term $\beta_{0,m}$. Then, the algorithm iteratively adds predictors that most reduce the AIC (Akaike, 1973). The algorithm terminates when it is no longer possible to reduce the AIC by adding another predictor. This approach is applied to each of the 36 response variables individually.

Modified baseline approach: This approach is the same as the *Baseline* approach although the stepwise algorithm terminates when it is no longer possible to improve the BIC (Schwarz, 1978) of the model by including another predictor.

Simultaneous baseline approach: This approach solves the SBS problem to estimate models of the form (4.2.1) for each response variable in a group simultaneously. In the MIQO program used to solve the SBS problem we include the constraints that exclude pairs of highly correlated predictors from entering the models. We also include the constraints that permit at most one of the predictors from the groups $\mathcal{T}_j$ for $j = 1, 2, 3, 4$ given in Table 4.1.2. We solve the following MIQO problem,

$$\begin{aligned}
\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta},\boldsymbol{\eta}}{\arg\min} \left[\sum_{m=1}^{M}\sum_{t=1}^{T}\left(\tilde{y}_{t,m} - \sum_{p=1}^{P}\tilde{x}_{t,p,m}\beta_{p,m}\right)^2\right] \quad & \text{subject to,} \\
(\beta_{p,m}, \eta_p) \in \mathcal{SOS} - 1, \quad & \text{for } m = 1, \ldots, M, \text{ for } p = 1 \ldots, P, \\
-\sum_{p=1}^{P}\eta_p = k - P, \quad & \\
\beta_{p,m} \in \mathbb{R}^+, \quad & \text{for } m = 1, \ldots, M, \text{ for } p = 1 \ldots, P, \\
\eta_p \in \{0, 1\}, \quad & \text{for } p = 1, \ldots, P, \\
-\eta_p - \eta_s \leq -1, \quad & \forall (p, s) \in \mathcal{HC}, \qquad (4.2.2a) \\
-\sum_{p\in\mathcal{T}_i}\eta_p \leq 1 - |\mathcal{T}_i|, \quad & \text{for } 1 = 1, 2, 3, 4, \qquad (4.2.2b)
\end{aligned}$$

given the response data $\tilde{\boldsymbol{y}} \in \mathbb{R}^{T\times M}$ and predictor data $\tilde{\boldsymbol{x}} \in \mathbb{R}^{T\times P\times M}$ that has been pre-processed. Here, we use M to denote the total number of response variables in a group. Constraints of the form (4.2.2a) ensure that pairs of highly correlated predictors are not present in the model. Constraints (4.2.2b) ensure that at most one predictor from the set $\mathcal{T}_j$ for $j = 1, 2, 3, 4$ are present in each of the regression models estimated by solving (4.2.2).

For each group of response variables we solve (4.2.2) for $k = 1, \ldots, K_{\max}$. The value $K_{\max}$ is determined automatically for each group of response variables. It may differ between groups of response variables as it depends on the sample estimates of the correlation between the predictors in each group of response variables. In Section 3.2.1 we discussed how to automatically determine $K_{\max}$ based on the feasibility of the MIQO program.

For each response variable, $\tilde{Y}_m$ where $m \in \mathcal{G}_i$ we will have $K_{\max}$ models of the form (4.2.1). We will select the best $k \in \{1, \ldots, K_{\max}\}$ for each $\tilde{Y}_m$ where $m \in \mathcal{G}_i$, simultaneously. Let the model for response variable $\tilde{Y}_m$ estimated using (4.2.2) be denoted $\mathcal{M}_m^k$. The superscript k denotes the sparsity level used in the SBS model (4.2.2). Let, $\texttt{BIC}(\mathcal{M}_m^k)$ denote the BIC (Schwarz, 1978) of model $\mathcal{M}_m^k$. Then, we simultaneously select the models for $\tilde{Y}_m$ corresponding to the sparsity

$$\hat{k} = \arg\min_k \left\{ \sum_{m \in \mathcal{G}_i} \texttt{BIC}(\mathcal{M}_m^k) \right\}.$$

The final approach that we consider is our proposed automated approach.

Automated approach: This approach differs from the three approaches given previously as it does not require the data to be pre-processed and fits models to the telecommunication event rates directly. We estimate Reg-SARIMA models of the form

$$y_{t,m} = \sum_{p=1}^{39} x_{t,p,m}\beta_{p,m} + \eta_{t,m}, \quad \text{where,} \tag{4.2.3a}$$

$$\nabla_m \nabla_m^s \phi_m(L)\Phi_m(L)\eta_{t,m} = \theta_m(L)\Theta_m(L)e_{m,t}, \tag{4.2.3b}$$

simultaneously for all $m \in \mathcal{G}_i$. We repeat this approach for $i = 1, \ldots, 7$. We follow with a description of how models of the form (4.2.3) are estimated.

The following steps are taken for each group of response variables, $\mathcal{G}_i$ for $i = 1, \ldots, 7$. To avoid cumbersome notation given group $\mathcal{G}_i$ we use M to denote the total number of response variables in a given group. Rather than refer to the observed response variable data $\boldsymbol{y}_{\mathcal{G}_i}$ we simply use $\boldsymbol{y}$, and similarly for the predictor variables. We let $\mathcal{L}$ denote the list of length N giving the order SARIMA models considered for the regression residuals

$$\mathcal{L} = [(p_1, d_1, q_1, P_1, D_1, Q_1, s_1), \ldots, (p_N, d_N, q_N, P_N, D_N, Q_N, s_N)].$$

For each level of sparsity, $k = 1, \ldots, K_{\max}$ the following steps are taken.

1. Use the response observations, $\boldsymbol{y} \in \mathbb{R}^{T \times M}$ and predictor observations, $\boldsymbol{x} \in \mathbb{R}^{T \times P \times M}$ in the optimisation model (4.2.2) to estimate the regression coefficients $\hat{\boldsymbol{\beta}}^0 \in \mathbb{R}^{P \times M}$ for models of the form

$$y_{t,m} = \sum_{p=1}^{39} x_{t,p,m}\beta_{p,m} + \eta_{t,m}.$$

2. For, $m = 1, \ldots, M$ determine the best SARIMA models for the regression residuals (4.2.3b)

   (a) Calculate the residuals

   $$\hat{\boldsymbol{\eta}}^0_{*,m} = \boldsymbol{y}_{*,m} - \boldsymbol{x}_{*,*,m}\hat{\boldsymbol{\beta}}^0_{*,m}.$$

   (b) For, $n = 1, \ldots, N$, fit the SARIMA model of order $(p_n, d_n, q_n, P_n, D_n, Q_n, s_n)$ to $\hat{\boldsymbol{\eta}}^0_{*,m}$. Denote the N models fitted as $\hat{\mathcal{R}}^0_{m,n}$.

   (c) Select the best SARIMA model using the BIC (Schwarz, 1978). Let, $\texttt{BIC}(\hat{\mathcal{R}}^0_{m,n})$ denote the value of the BIC of model $\hat{\mathcal{R}}^0_{m,n}$, then we select the best n associated to response variable m as

   $$\hat{n}_m = \arg\min_n \{\hat{\mathcal{R}}^0_{m,n}\}.$$

   (d) Now we apply the Generalised Least Squares (GLS) transformation to the response and predictor variables. Let

   $$\hat{\phi}^0_m(L), \hat{\Phi}^0_m(L), \hat{\theta}^0_m(L), \hat{\Theta}^0_m(L), \nabla_{\hat{d}^0_m}, \nabla^{\hat{s}^0_m}_{\hat{D}^0_m},$$

   denote the SARIMA polynomials and differencing operators corresponding to model $\hat{\mathcal{R}}^0_{m,\hat{n}}$. Then, the we apply the GLS transformation to the data

   $$\bar{y}_{t,m} = \frac{\nabla_{\hat{d}^0_m}\nabla^{\hat{s}^0_m}_{\hat{D}^0_m}\hat{\phi}^0_m(L), \hat{\Phi}^0_m(L)}{\hat{\theta}^0_m(L)\hat{\Theta}^0_m(L)} y_{t,m}, \quad \bar{x}_{t,p,m} = \frac{\nabla_{\hat{d}^0_m}\nabla^{\hat{s}^0_m}_{\hat{D}^0_m}\hat{\phi}^0_m(L), \hat{\Phi}^0_m(L)}{\hat{\theta}^0_m(L)\hat{\Theta}^0_m(L)} x_{t,p,m},$$

   where $\bar{y}_{t,m}$ and $\bar{x}_{t,p,m}$ are the GLS transformed response and predictor variables.

3. The following steps now iterate until we obtain convergence in the two-step algorithm or some maximum iteration number, $maxit$ is reached. We discuss what it means for the algorithm to have converged shortly. For $it = 1, \ldots, maxit$:

   (a) We re-estimate the regression coefficients using the response data $\bar{\boldsymbol{y}} \in \mathbb{R}^{T\times M}$ and for the predictors, $\bar{\boldsymbol{x}} \in \mathbb{R}^{T\times P\times M}$ that has had the GLS transformation applied. Using $\bar{\boldsymbol{y}}$ and $\bar{\boldsymbol{x}}$ in the optimisation model (4.2.2) we estimate the regression coefficients $\hat{\boldsymbol{\beta}}^{it} \in \mathbb{R}^{P\times M}$ for models of the form

   $$\bar{y}_{t,m} = \sum_{p=1}^{39} \bar{x}_{t,p,m}\beta_{p,m} + \eta_{t,m}.$$

   (b) For, $n = 1, \ldots, N$ fit the SARIMA model of order $(p_n, d_n, q_n, P_n, D_n, Q_n, s_n)$ to $\hat{\boldsymbol{\eta}}^{it}_{*,m}$. Denote the N models fitted as $\hat{\mathcal{R}}^{it}_{m,n}$.

   (c) Select the best SARIMA model using the BIC (Schwarz, 1978). Select the n for response variable Y_m such that

   $$\hat{n}_m = \arg\min_n \{\hat{\mathcal{R}}^{it}_{m,n}\}.$$

Let, $(p^{it}_{\hat{n}_m}, d^{it}_{\hat{n}_m}, q^{it}_{\hat{n}_m}, P^{it}_{\hat{n}_m}, D^{it}_{\hat{n}_m}, Q^{it}_{\hat{n}_m}, s^{it}_{\hat{n}_m})$ denote the selected SARIMA model order corresponding to Y_m at iteration it.

(d) Now we re-apply the Generalised Least Squares (GLS) transformation to the response and predictor variables. Let

$$\hat{\phi}^{it}_m(L), \hat{\Phi}^{it}_m(L), \hat{\theta}^{it}_m(L), \hat{\Theta}^{it}_m(L), \nabla_{\hat{d}^{it}_m}, \nabla^{\hat{s}^{it}_m}_{\hat{D}^{it}_m},$$

denote the SARIMA polynomials and differencing operators corresponding to model $\hat{\mathcal{R}}^{it}_{m,\hat{n}}$. Then, the GLS transformation of the data becomes

$$\bar{y}_{t,m} = \frac{\nabla_{\hat{d}^{it}_m} \nabla^{\hat{s}^{it}_m}_{\hat{D}^{it}_m} \hat{\phi}^{it}_m(L), \hat{\Phi}^{it}_m(L)}{\hat{\theta}^{it}_m(L)\hat{\Theta}^{it}_m(L)} y_{t,m}, \quad \bar{x}_{t,p,m} = \frac{\nabla_{\hat{d}^{it}_m} \nabla^{\hat{s}^{it}_m}_{\hat{D}^{it}_m} \hat{\phi}^{it}_m(L), \hat{\Phi}^{it}_m(L)}{\hat{\theta}^{it}_m(L)\hat{\Theta}^{it}_m(L)} x_{t,p,m}.$$

(e) Check for convergence. If the algorithm has converged, stop. Otherwise, return to 3(a).

4. If $maxit$ has been reached the two-step algorithm has not converged.

Steps 1-4 will produce $K_{\max}$ models of the form (4.2.3) for each response variable, giving a total of $K_{\max} \times M$ models for response Group i. The sparsity of the Reg-SARIMA models is also chosen simultaneously. For each level of sparsity k, denote the associated Reg-SARIMA model for response m as $\mathcal{M}_{k,m}$. Let $BIC(\mathcal{M}_{k,m})$ denote the value of the BIC (Schwarz, 1978) for model $\mathcal{M}_{k,m}$. Then, we select the model sparsity as

$$\hat{k} = \min_k \sum_{m \in \mathcal{G}_i} BIC(\mathcal{M}_{k,m}).$$

For each level of sparsity we sum the BIC of each model fit to the response variables in $\mathcal{G}_i$, and select the k for which this sum is minimal. The criteria used to determine whether the two-step algorithm has converged is as follows.

At iteration, $it > 0$ in the two-step algorithm we check

1. If the same subset of predictors has been selected between two consecutive iterations,

$$\{p : \beta^{it}_{p,m} \neq 0 \text{ for } p = 1, \ldots, P, \ \forall m \in \mathcal{G}_i\} = \{p : \beta^{it-1}_{p,m} \neq 0 \text{ for } p = 1, \ldots, P, \ \forall m \in \mathcal{G}_i\}.$$

2. If the sum of the absolute differences in coefficient estimates between two consecutive iterations is less than some convergence tolerance ϵ,

$$\sum_{m \in \mathcal{G}_i} \sum_{p=1}^{P} \left(|\hat{\boldsymbol{\beta}}^{it}_{p,m} - \hat{\boldsymbol{\beta}}^{it-1}_{p,m}| \right) \leq \epsilon.$$

3. Finally, check whether the same SARIMA models were selected consecutively for the Reg-SARIMA models for each $m \in \mathcal{G}_i$,

$$(p^{it}_{\hat{n}_m}, d^{it}_{\hat{n}_m}, q^{it}_{\hat{n}_m}, P^{it}_{\hat{n}_m}, D^{it}_{\hat{n}_m}, Q^{it}_{\hat{n}_m}, s^{it}_{\hat{n}_m}) = (p^{it-1}_{\hat{n}_m}, d^{it-1}_{\hat{n}_m}, q^{it-1}_{\hat{n}_m}, P^{it-1}_{\hat{n}_m}, D^{it-1}_{\hat{n}_m}, Q^{it-1}_{\hat{n}_m}, s^{it-1}_{\hat{n}_m}).$$

If the logical conditions 1-3 are all satisfied, then the two-step algorithm has converged. We will now evaluate the performance of each of the approaches.

4.3 Evaluation of the approaches

The *Automated* procedure discussed in Chapter 3 was developed to address many of the challenges often encountered when modelling telecommunications data. We now recall these challenges and explain how we will assess if each of these challenges has been addressed.

Firstly, an approach with minimal user input was required. We have discussed the *Automated* approach in detail in Chapter 3 and note that significantly less time is required to produce statistical models. This is achieved by modelling the response variables directly. The advantage here is that behaviour observed in the response variables, not thought to be attributed to weather variables, can be incorporated into the models themselves. By modelling the events directly we can incorporate non-weather related variation into the models. This allows us to model the seasonality using the Reg-SARIMA model and the calendar effects can be estimated using indicator variables. Using indicator variables to account for calendar affects has two benefits. Firstly, the approach will decide automatically if a calendar affect exists by including the respective indicator variable into the model. This removes judgemental elements of the modelling procedure that may differ amongst different analysts. The second advantage in using indicator variables is that the effects of the weather related predictors are estimated at the same time as calendar effects. The pre-processing stage assumes the calendar effects are estimated without error. This can lead to over-confident predictions.

The second requirement of the approach is to produce interpretable models when selecting amongst a large number of highly correlated predictor variables. Our approach guarantees that the sign of the regression coefficients obtained using our automated procedure agrees with expert opinion by placing constraints on the regression coefficients. Interpretable models will also be assessed by the consistency of selected predictors amongst response variables from the same group. It is thought that models for response variables within the same group should contain similar predictors. We will inspect the predictors chosen for each response variable looking for consistency among the chosen predictors.

The third requirement of the modelling approach was to adequately capture serial correlation in the response variables. We will investigate the ability of each method to capture serial correlation by observing the sample autocorrelation plots of the model errors. Let $\hat{y}_{t,m}$ denote the fitted values of the telecommunication events. We will inspect the sample autocorrelation $\hat{\rho}(\hat{e}_{t,m})$, where $\hat{e}_{t,m} = y_{t,m} - \hat{y}_{t,m}$ are the model errors. A model that fails to capture serial correlation should yield

significant peaks in the autocorrelation plot of the model errors.

Finally, a procedure that can jointly model response variables was sought. The *Simultaneous baseline* and *Automated* approaches estimate the impact of predictors jointly. The *Simultaneous baseline* approach has been proposed directly to show improvements in simultaneous predictor selection in comparison to the *Baseline* approach. We will evaluate the prediction error for 14 day-ahead and 365 day-ahead forecasts for each approach. Comparing the prediction errors between the *Baseline*, *Modified baseline* and *Simultaneous baseline* approaches will give a direct insight into the improvements obtained by using simultaneous predictor selection. A comparison of the selected predictors now follows.

4.3.1 Comparison of selected predictors

In this section we will investigate the predictors selected between the approaches described in Section 4.2. A common challenge when selecting amongst highly correlated predictors is obtaining models whereby coefficients for highly correlated predictors are not contradictory. Table 4.3.1 shows the regression coefficients for the models produced by the *Modified baseline* approach for the six response variables in Group 1. Predictors 16 and 17 are present in four of the six models. The correlation between predictors 16 and 17 is at least 0.94 across all response variables. Observing the coefficients in Table 4.3.1 we can see that for each response variable where predictors 16 and 17 are present, the coefficients take opposing signs. Predictors 16 and 17 correspond to a transformation of the lightning variable. The four models aforementioned would suggest that lighting has both positive and negative effects on telecommunication events. Such observations are seen for all other groups of response variables. Conflicting signs for coefficients amongst highly correlated predictors is more common for models produced by the *Baseline* approach.

There are a few predictors present in most of the models within Group 1. Predictor 17 appears in each model, and predictors 4 ($\mathcal{T}_1$, humidity) and 30 ($\mathcal{L}_4$, precipitation) appear in all models except for response variable 5. In order to access the differences using simultaneous predictor selection we compare this to the models produced by the *Simultaneous baseline* approach. The coefficients for the selected predictors using the *Simultaneous baseline* approach are shown in Table 4.3.2. Predictor 17 which appeared in all models produced by the *Modified baseline* approach appear in the models produced by the *Simultaneous baseline* approach. All predictors selected by the *Simultaneous baseline* approach appear in at least one of the models produced by the *Modified baseline* approach. By design, none of the coefficients take negative values and the average model sparsity for the *Simultaneous baseline* approach is 6. The average sparsity of the models produced by the *Modified baseline* is 7.5.

Table 4.3.1: Regression coefficients produced using the *Modified baseline* approach for the predictors selected in the models for response variables in Group 1. Negative coefficients are highlighted in red. The dashes indicate the coefficient was zero exactly, hence the associated predictor was not selected.

		Coefficient estimate for response variable,					
Predictor Group	Coefficient	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
$\mathcal{T}_1$ (Humidty)	$\beta_{3,m}$	–	–	–	–	0.089	–
-	$\beta_{4,m}$	0.05	0.008	0.079	0.027	–	0.07
$\mathcal{T}_2$ (Windspeed)	$\beta_{6,m}$	0.054	–	0.009	–	–	–
$\mathcal{T}_3$ (Precipitation)	$\beta_{9,m}$	–	–	–	–	0.01	–
-	$\beta_{11,m}$	–	–	–	–	–	0.003
-	$\beta_{12,m}$	0.01	0.009	–	–	–	0.006
-	$\beta_{13,m}$	–	–	0.025	0.01	0.017	–
-	$\beta_{14,m}$	–	0.015	–	0.009	–	–
$\mathcal{T}_4$ (Lightning)	$\beta_{15,m}$	–	–	–	–	0.008	–
-	$\beta_{16,m}$	-0.027	–	0.006	0.005	–	0.028
-	$\beta_{17,m}$	0.01	0.012	-0.014	-0.02	-0.004	-0.017
$\mathcal{L}_2$ (Lightning)	$\beta_{25,m}$	–	–	–	–	–	0.011
-	$\beta_{26,m}$	-0.003	–	–	–	–	–
$\mathcal{L}_3$ (Windspeed)	$\beta_{27,m}$	–	0.008	–	-0.012	-0.008	–
$\mathcal{L}_4$ (Precipitation)	$\beta_{30,m}$	0.007	0.12	0.006	0.115	–	-0.029

The regression coefficients for the Regression-SARIMA models obtained using the *Automated* approach are shown in Table 4.3.3. There is some agreement in the selected predictors with the *Simultaneous baseline* and *Automated* approaches. In particular, all predictors selected by the *Simultaneous baseline* approach appear in the models produced by the *Automated* approach, with the exception of predictor 27 ($\mathcal{L}_3$, windspeed). In addition to the predictors selected by the *Simultaneous baseline* approach predictors, 6, 37, 38 and 39 appear in the models.

Figure 4.3.1 shows how the simultaneous best-subset of predictors changes as the two-step algorithm progresses. Given a sparsity level k, recall from Section 4.2 that we obtain estimates of the regression coefficients for models of the form (4.2.1). We then proceed by finding a suitable SARIMA model for the regression residuals and then re-select the best-subset of predictors on a GLS transform of the response and predictor variables. Therefore, for each level of sparsity, we will have at least two best-subsets of predictor variables.

Figure 4.3.1 shows that for sparsity levels, 1, 3, 9, 10 and 11 the two-step algorithm converges in two steps. Each set of horizontal dots indicates the presence of a predictor in the two-step algorithm.

We can see that predictor 37 was selected initially as the best predictor. This is shown by the left-most blue dot in Figure 4.3.1. Once suitable SARIMA models are selected for the regression residuals, the best single predictor to include for the linear regression models on the GLS transform of the data is selected. This is again predictor 37. For all other levels of sparsity the two-step algorithm required three iterations for this group of response variables. It appears that as soon as predictor 39 is included into the models, predictor 27 is replaced with predictor 6.

Table 4.3.2: Regression coefficients produced using the *Simultaneous baseline* approach for the predictors selected in the models for response variables in Group 1.

		Coefficient estimate for response variable,					
Predictor Group	Coefficient	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
$\mathcal{T}_1$ (Humidity)	$\beta_{4,m}$	0.009	0.009	0.005	0.010	0.011	0.005
$\mathcal{T}_3$ (Precipitation)	$\beta_{13,m}$	0.076	0.071	0.082	0.088	0.075	0.051
$\mathcal{T}_4$ (Lighting)	$\beta_{17,m}$	0.013	0.010	0.008	0.003	0.007	0.006
$\mathcal{L}_2$ (Lightning)	$\beta_{25,m}$	0.002	0.001	0.000	0.001	0.002	0.003
$\mathcal{L}_3$ (Windspeed)	$\beta_{27,m}$	0.007	0.008	0.004	0.005	0.006	0.002
$\mathcal{L}_4$ (Precipitation)	$\beta_{30,m}$	0.009	0.014	0.009	0.007	0.004	0.010

Table 4.3.3: Regression coefficients produced using the *Automated* approach for the predictors selected in the models for response variables in Group 1.

		Coefficient estimate for response variable,					
PredictorGroup	Coefficient	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
$\mathcal{T}_1$ (Humidity)	$\beta_{4,m}$	0.010	0.010	0.007	0.010	0.011	0.006
$\mathcal{T}_2$ (Windspeed)	$\beta_{6,m}$	0.011	0.011	0.010	0.006	0.007	0.005
$\mathcal{T}_3$ (Precipitation)	$\beta_{13,m}$	0.073	0.059	0.065	0.077	0.051	0.037
$\mathcal{T}_4$ (Lightning)	$\beta_{17,m}$	0.010	0.006	0.005	0.001	0.007	0.003
$\mathcal{L}_3$ (Lightning)	$\beta_{25,m}$	0.003	0.001	0.000	0.001	0.002	0.003
$\mathcal{L}_4$ (Precipitation)	$\beta_{30,m}$	0.009	0.014	0.010	0.008	0.005	0.010
$\mathcal{B}$ (Bank Holidays)	$\beta_{37,m}$	0.760	0.780	0.656	0.644	0.755	0.756
-	$\beta_{38,m}$	0.723	0.794	0.686	0.750	0.883	0.735
-	$\beta_{39,m}$	0.240	0.243	0.210	0.100	0.233	0.217

The substitution of predictors conditional on the inclusion of another predictor highlights the importance of considering the best-subset approach, or one of the hybrid variants we proposed in Chapter 5. In a forward stepwise approach, once a predictor is selected it cannot be removed or

substituted with another predictor. This substitution of predictor behaviour is again observed as the model sparsity changes from 9 to 10, as predictor 6 is dropped from the models.

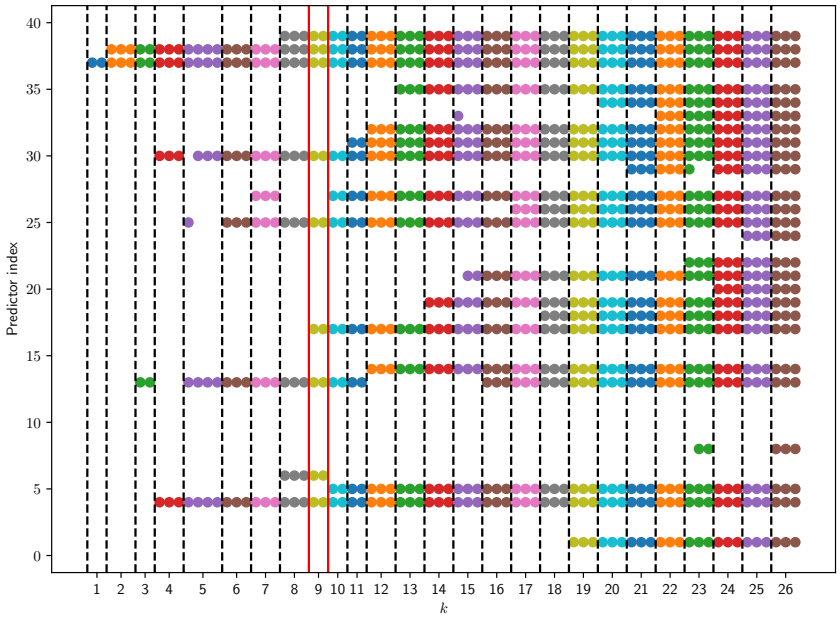


Figure 4.3.1: Trace-plot indicating the predictors selected for Group 1 at each iteration of the two-step algorithm (*Automated approach*). The red vertical lines indicate the selected model.

Predictors 37, 38, and 39 appear in all of the Reg-SARIMA models produced using the *Automated* approach for the telecommunications event dataset. The two-step algorithm converges in no more than eight iterations for all response groups. The number of predictors present in all models is approximately 8. An exception to this is response Group 2. The trace of selected predictors for Group 2 is shown in Figure 4.3.2. The models for Group 2 contained only four predictors, in which only one is weather related. In consultation with our industrial partner the response variables in Group 2 are expected to be less influenced by weather variables due to their location in the network.

In the following we will discuss the serial correlation captured by the *Automated* approach with the *Baseline* approach.

4.3.2 Modelling serial correlation

Significant serial correlation is observed in the model errors for all response variables fit using the *Baseline*, *Modified baseline*, and *Simultaneous baseline* approaches. The presence of serial correlation indicates that the model for each response variable fails to adequately explain the serial correlation observed. In comparison, almost all significant serial correlation in the response variables appears to be captured in the models fit by the *Automated* approach. Marginally significant serial correlation

appears in the sample autocorrelation estimates of the model errors fit using the *Automated* approach. The auto-correlation plot of the model errors appear very similar to those observed in Chapter 3 for both the *Automated* and *Baseline* approaches. The autocorrelation and partial autocorrelation plots are provided in Appendix 4.A.

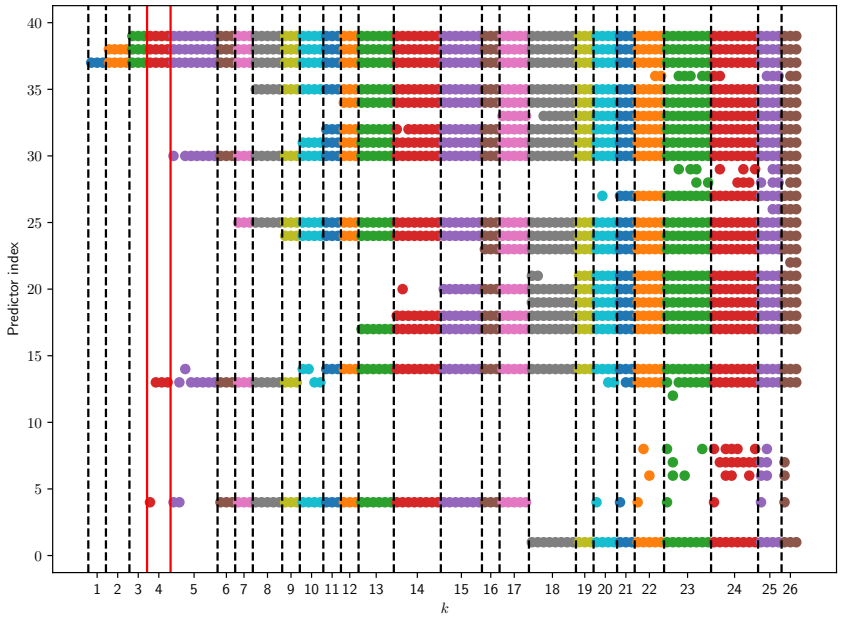


Figure 4.3.2: Trace-plot indicating the predictors selected for Group 2 at each iteration of the two-step algorithm. The red vertical lines indicate the selected model.

The estimates of the SARIMA parameters for Group 5 are shown in Table 4.3.4. An interesting observation here is that the same order of SARIMA model was selected for each response variable in Group 1. Similar parameter estimates were obtained for each SARIMA model. The SARIMA models were estimated and selected separately for each response variable. In Chapter 8 we discuss the potential of estimating these models simultaneously.

A common SARIMA model for the regression residuals was found for many of the response variables in the telecommunications dataset. In fact, the Regression-SARIMA(1,0,1)(1,0,1,7) model was fitted to all but one response variable in Group 4. In this single exception, the order of the Reg-SARIMA model fit was (1,0,0)(1,0,1,7). The difference here is that the model did not include a moving average term.

Table 4.3.4: The SARIMA coefficients (given to 2.dp) for each Reg-SARIMA model fitted using the *Automated* approach for all response variables in Group 1 .

	SARIMA Parameters for response variable,					
Parameter	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6
$\phi_{1,m}$	0.66	0.85	0.73	0.92	0.98	0.93
d_m	0	0	0	0	0	0
$\theta_{1,m}$	-0.43	-0.65	-0.49	-0.81	-0.92	-0.81
$\Phi_{1,m}$	0.06	-0.01	-0.01	0.03	-0.01	-0.01
D_m	1	1	1	1	1	1
$\Theta_{1,m}$	-0.97	-0.97	-0.97	-1.01	-0.97	-0.96
$\sigma^2_{e_m}$	0.06	0.06	0.08	0.052	0.06	0.04

Finally we shall assess the performance of the models numerically.

4.3.3 Predictions

To quantify the performance of the models numerically we consider the mean-squared error of the 14 day-ahead and 365 day-ahead predictions. It may not be possible to obtain accurate values of some predictors which are required to predict the response variables. In particular, we may be unable to obtain accurate predictions of the weather predictors for more than a couple of days ahead. However, the main purpose of these models is for explanatory purposes, so we compare predictions based on observed values of the predictors. By comparing the predictive performance of the models on both a short and long term horizon we can better understand how each of the approaches perform.

Table 4.3.5: Mean squared error of the 14 day-ahead predictions for each response variable in Group 1 by the four approaches given in Section 4.2.

	Mean-squared prediction error for response variable,						
Approach	Y_1	Y_2	Y_3	Y_4	Y_5	Y_6	Average
Baseline	0.247	0.293	0.167	0.058	0.266	0.096	0.188
Modified baseline	0.254	0.254	0.184	0.056	0.238	0.092	0.180
Simultaneous baseline	0.176	0.110	0.345	0.045	**0.104**	0.082	0.144
Automated	**0.140**	**0.104**	**0.111**	**0.041**	0.130	**0.029**	**0.093**

The mean-squared prediction error for the 14 day-ahead predictions for each response variable in Group 1 is shown in Table 4.3.5. The *Automated* approach produced the lowest mean-squared prediction error for five of the six response variables. The *Simultaneous baseline* approach produced

the lowest prediction error for response variable Y_5, followed by the Automated approach. The mean-squared prediction error averaged over all response variables was at least 35% lower for models produced by the *Automated* approach in comparison to the *Simultaneous baseline* approach, which followed in second place. The 14 day-ahead predictions made from the models produced by the *Modified baseline*, *Simultaneous baseline* and *Automated* approaches are shown in Figure 4.3.3 for response variables in Group 1.

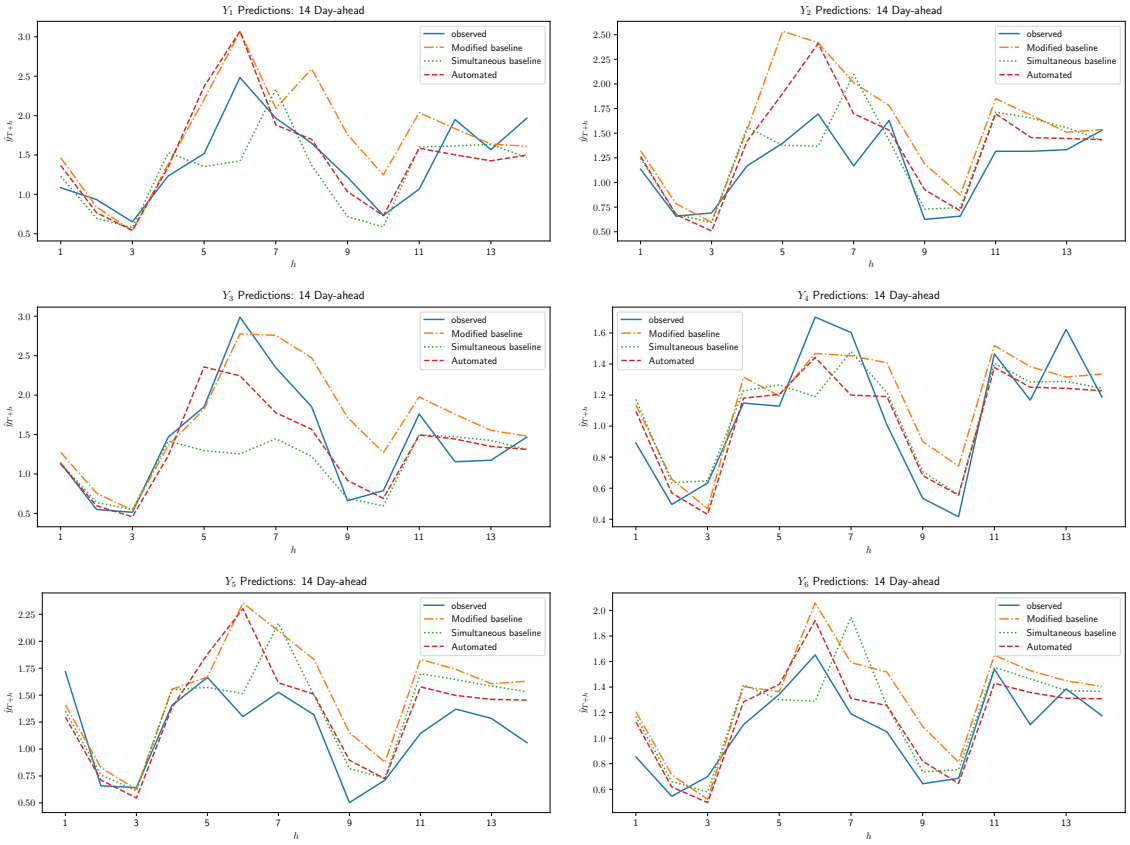


Figure 4.3.3: Prediction plots for all response variables in Group 1 for the *Modified baseline*, *Simultaneous baseline* and *Automated* approaches.

Across all nine response groups we found that the Reg-SARIMA models produced by the *Automated* approach produced the most accurate predictions for the 14 day-ahead predictions. Slight reductions in prediction errors were observed with the *Simultaneous baseline* approach over the *Baseline* and *Modified baseline* approaches. The improvement in prediction accuracy of the *Simultaneous baseline* compared to the *Baseline* and *Modified baseline* approaches may be due to selecting predictors simultaneously. There appeared to be no overall winner between the *Baseline* and *Modified baseline* approaches for the 14 day-ahead predictions. However, the models produced by the *Modified baseline* approach were typically much sparser than models produced by the *Baseline* approach.

Table 4.3.6: Mean squared error of the 365 day-ahead prediction for each response variable in Group 4 by the four approaches given in Section 4.2.

	m					
Approach	1	2	3	4	5	Average
Baseline	0.075	**0.054**	0.058	0.068	**0.050**	0.061
Modified baseline	**0.074**	**0.054**	**0.057**	**0.063**	0.051	**0.060**
Simultaneous baseline	0.117	0.062	0.061	0.125	0.104	0.094
Automated	0.085	0.062	0.085	0.078	0.069	0.076

The 365 day-ahead mean-squared prediction error averaged across each response variable in a response group was lowest for models produced by the *Automated* approach for five of the seven response groups. Table 4.3.6 shows the results for Group 4 where the models produced by the *Automated* approach were not the most accurate over a 365 day period. Despite the *Automated* approach not being the most accurate, the prediction errors are comparable to the *Baseline* approaches and far less effort was needed to implement this approach.

4.4 Conclusion

In this chapter we have applied our *Automated* simultaneous predictor selection approach to the full telecommunications event dataset. We compared the performance of our approach to the *Baseline* approach currently used by our industrial collaborator. We found the models produced by the *Automated* approach generally more favourable.

Firstly, the automated approach does not require the data to be pre-processed, so can produce models with less effort in comparison to all other approaches. The MIQO framework used to fit models in the *Automated* approach excludes pairs of highly correlated predictors. Consequently, the models fit by the *Automated* approach do not contain pairs of highly correlated predictors, for which the corresponding coefficients are opposing in sign. This is guaranteed by enforcing positive regression coefficients.

The models produced by the *Automated* approach often resulted in models with fewer weather related predictor variables. Despite including fewer predictors, the models produced by the *Automated* approach performed comparably. The 14 day-ahead predictions were most accurate, in terms of mean-squared prediction error for models produced by the *Automated* approach. The 365 day-ahead predictions averaged over each response variable within a group were most accurate for the Automated approach for five of the seven response groups. The *Automated* approach performed the

best for a short horizon. This was most likely due to the Reg-SARIMA's ability to capture the serial correlation in the response variables.

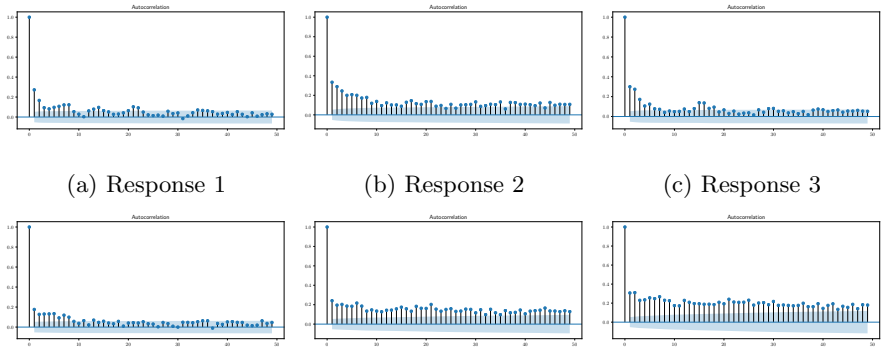
In conclusion, the Reg-SARIMA models fit by the *Automated* approach are more favourable over a number of criteria, and significantly less time is required by an analyst to estimate non-weather related effects. This is desirable for our industrial collaborator as datasets grow in size and large numbers of models are required. Secondly, the *Automated* approach jointly selects predictors for groups of response variables. This allows us to achieve consistency amongst groups of response variables.

4.A Supplementary ACF and PACF plots

This appendix contains the ACF and PACF plots of the model errors from the *Modified Baseline*, *Simultaneous Baseline* and *Automated* approaches for Group 1.

4.A.1 Modified Baseline Approach

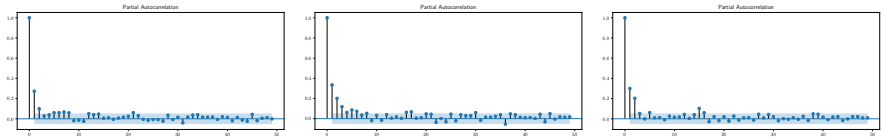
Both, significant autocorrelation and partial autocorrelation was found in the model errors from the *Modified Baseline* approach. These plots are shown in Figures 4.A.1 and 4.A.2 respectively.



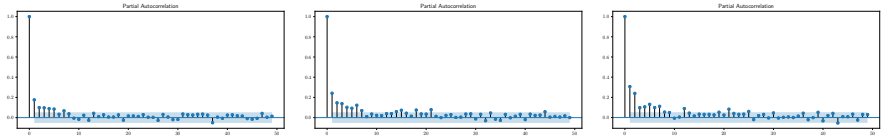
(a) Response 1 (b) Response 2 (c) Response 3

(d) Response 4 (e) Response 5 (f) Response 6

Figure 4.A.1: ACF of the model errors from the *Modified Baseline* approach for Group 1. The uncertainty cloud shows the 95% confidence interval calculated using Bartlett's formula.



(a) Response 1 (b) Response 2 (c) Response 3

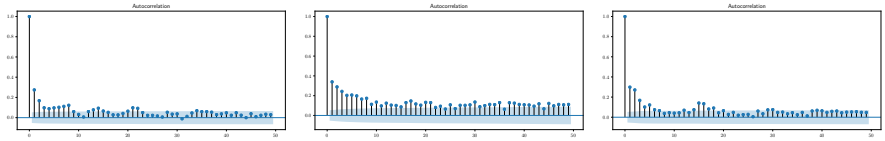


(d) Response 4 (e) Response 5 (f) Response 6

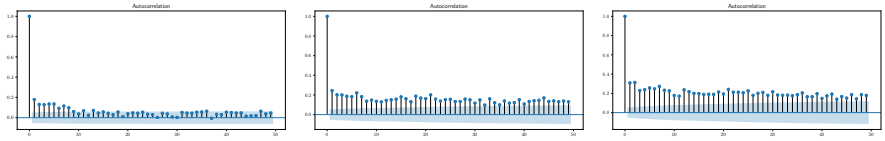
Figure 4.A.2: PACF of the model errors from the *Modified Baseline* approach for Group 1.

4.A.2 Simultaneous Baseline Approach

Both, significant autocorrelation and partial autocorrelation was found in the model errors from the *Simultaneous Baseline* approach. These plots are shown in Figures 4.A.3 and 4.A.4 respectively.

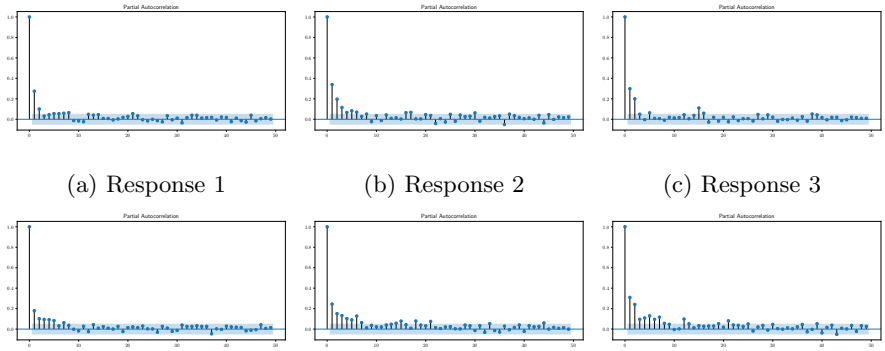


(a) Response 1 (b) Response 2 (c) Response 3



(d) Response 4 (e) Response 5 (f) Response 6

Figure 4.A.3: ACF of the model errors from the *Simultaneous Baseline* approach for Group 1. The uncertainty cloud shows the 95% confidence interval calculated using Bartlett's formula.



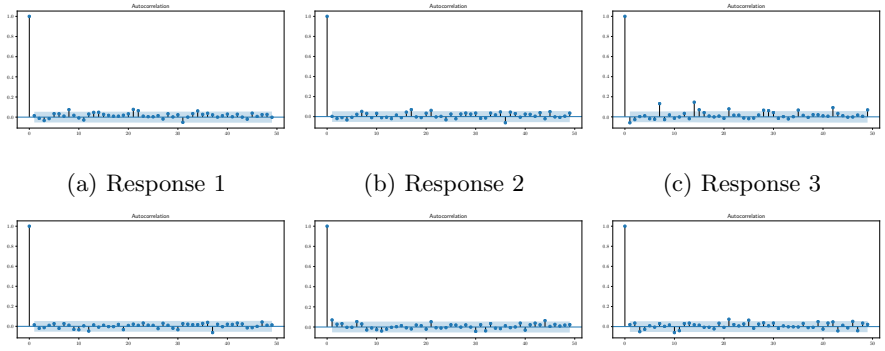
(a) Response 1 (b) Response 2 (c) Response 3

(d) Response 4 (e) Response 5 (f) Response 6

Figure 4.A.4: PACF of the model errors from the *Simultaneous Baseline* approach.

4.A.3 Automated Approach

The model errors from the *Automated* approach typically contain far less significant autocorrelation compared to the *Modified Baseline* and *Simultaneous Baseline* approaches. At the 95% confidence level very few lags show signs of significant autocorrelation for all response variables in Group 1 with the exception of Response 3, this can be seen in Figure 4.A.5. Significant autocorrelation is found at the lags which are multiples of seven for Response 3. This may be explained by the *Automated* approach failing to include a seasonal autoregressive term for the regression residuals.

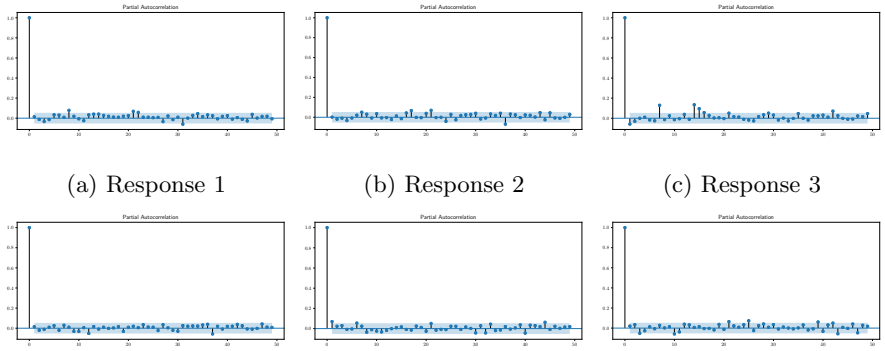


(a) Response 1 (b) Response 2 (c) Response 3

(d) Response 4 (e) Response 5 (f) Response 6

Figure 4.A.5: ACF of the model errors from the *Automated* approach for Group 1. The uncertainty cloud shows the 95% confidence interval calculated using Bartlett's formula.

Similarly, the partial autocorrelation appears significant at very few lags at the 95% level for all response variables in Group 1 with the exception of Response 3, this can be seen in Figure 4.A.6. Again, this may be explained by the failure of the *Automated* approach to specify a seasonal autoregressive term for the regression residuals.



(a) Response 1 (b) Response 2 (c) Response 3

(d) Response 4 (e) Response 5 (f) Response 6

Figure 4.A.6: PACF of the model errors from the *Automated* approach for Group 1.

Chapter 5

Simultaneous best-subset implementation study

In Chapter 3 we observed that the time taken to solve the SBS problem exactly using MIQO programs can be excessive. In this chapter we show that it is possible to reduce the time required to solve the SBS problem. We achieve this by using data driven parameters to improve the performance of the optimisation solver. We develop a discrete first-order algorithm to obtain good feasible solutions to the SBS problem and show that these solutions can be used produce good statistical models in practice.

The structure of this chapter follows. In Section 5.1 we specify three systems of linear regression models that we use in the remainder of this thesis to assess the performance of simultaneous predictor selection algorithms. In Section 5.2 we re-visit the SBS problem and derive a number of data-driven MIQO programs that can be used to solve the SBS problem exactly and demonstrate that this can lead to a reduction in the time required to solve the SBS problem. In Section 5.4 we develop a discrete first-order algorithm to provide good feasible solutions to the SBS problem in practise. We show how feasible solutions to the SBS problem can be used to estimate the data-driven parameters in Section 5.5. In Section 5.6 we perform a simulation study to investigate how our methods perform.

5.1 Models

This chapter is concerned with approaches that can simultaneously select predictors for systems of linear regression models taking the form

$$\begin{aligned} y_{t,1} &= \sum_{p=1}^{P} x_{t,p,1}\beta_{p,1} + \eta_{t,1}, \\ &\dots \\ y_{t,M} &= \sum_{p=1}^{P} x_{t,p,M}\beta_{p,M} + \eta_{t,M}. \end{aligned} \tag{5.1.1}$$

Here, we have M response variables and a realisation of each of the P predictor variables for each response variable. The regression residuals, $\eta_{t,m}$ for all models are independently distributed such that

$$\eta_{t,m} \sim \mathrm{N}(0, \sigma^2_{\eta_m}) \quad \text{for } m = 1, \dots, M. \tag{5.1.2}$$

Here, we refer to $\sigma^2_{\eta_m}$ as the residual variance for model m. The structure of the regression coefficients and distributions of the predictor variables are given for four models as follows.

Uniformly spaced model: In this model the indices of the non-zero regression coefficients are uniformly spaced such that

$$\beta_{p,m} = \begin{cases} 1, & \text{if } p \in \{1, 8, 15, 22, 29\}, \\ 0, & \text{otherwise,} \end{cases} \qquad \text{for } m = 1, \dots, 5. \tag{5.1.3}$$

Here, the predictor variables $\boldsymbol{X}_m$ are distributed such that

$$\boldsymbol{X}_m \sim \mathrm{MVN}(\mathbf{0}_{35}, \boldsymbol{\Sigma}) \quad \text{where} \quad \mathbf{0}_P = [0, \dots, 0] \in \mathbb{R}^{35} \quad \text{and} \quad \boldsymbol{\Sigma}_{i,j} \in \mathbb{R}^{35\times 35} := \Sigma_{i,j} = \rho^{|i-j|}$$

for $m = 1, \dots, 5$.

Adjacent model: The *Adjacent* model, whereby the position of the non-zero coefficients are adjacent such that,

$$\beta_{p,m} = \begin{cases} 0.3, & \text{if } p = 17, \\ 1, & \text{if } p = 18, \\ 0.6, & \text{if } p = 19, \\ 0, & \text{otherwise,} \end{cases} \qquad \text{for } m = 1, \dots, 5. \tag{5.1.4}$$

Here, the predictor variables $\boldsymbol{X}_m$ are distributed such that

$$\boldsymbol{X}_m \sim \mathrm{MVN}(\mathbf{0}_{35}, \boldsymbol{\Sigma}) \quad \text{where} \quad \mathbf{0}_P = [0, \dots, 0] \in \mathbb{R}^{35} \quad \text{and} \quad \boldsymbol{\Sigma}_{i,j} \in \mathbb{R}^{35\times 35} := \Sigma_{i,j} = \rho^{|i-j|}$$

for $m = 1, \dots, 5$.

Application model: In this model the predictors behave similarly to the predictors in our telecommunications application. We generate the predictors such that blocks of predictors can be generated where there is pair-wise correlation amongst the predictors from each block. The regression coefficients are such that

$$\beta_{p,m} = \begin{cases} 1, & \text{if } p = 30, \\ 0.775, & \text{if } p = 25, \\ 0.55, & \text{if } p = 14, \\ 0.325, & \text{if } p = 5, \\ 0.1, & \text{if } p = 2, \\ 0, & \text{otherwise,} \end{cases} \quad \text{for } m = 1, \ldots, 5. \tag{5.1.5}$$

The predictors $\boldsymbol{X}_m$ are distributed such that

$$\boldsymbol{X}_m \sim \text{MVN}_{35}(\mathbf{0}_{35}, \boldsymbol{\Sigma}) \quad \text{for } m = 1, \ldots, 5.$$

The covariance matrix of the predictors, $\boldsymbol{\Sigma}$ has the block diagonal structure such that

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}^{(1)} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \boldsymbol{\Sigma}^{(5)} \end{bmatrix} \in \mathbb{R}^{35 \times 35}.$$

Here, $\boldsymbol{\Sigma}^{(b)} \in \mathbb{R}^{(b+4)\times(b+4)} := \Sigma^{(b)}_{i,j} = \rho^{|i-j|}$ for $b = 1, \ldots, 5$. In the *Uniformly spaced*, *Adjacent*, and *Application* models, we will present results for two values of $\rho \in \{0.5, 2\}$. The results for each model will be presented *ModelName-*ρ and the variance of the regression residuals, $\sigma^2_{\eta_m}$ for $m = 1, \ldots, 5$, will be made clear.

Scaling model: To determine how the approaches scale with P and M we simulate data from the *Scaling* model, so-called as we use it to investigate the scaling of approaches. In this model the regression coefficients are given such that,

$$\beta_{p,m} = \begin{cases} 1, & \text{if } p = 1, 3, 5, \\ 0, & \text{otherwise,} \end{cases} \quad \text{for } m = 1, \ldots, M. \tag{5.1.6}$$

Here, the predictor variables $\boldsymbol{X}_m$ are distributed such that

$$\boldsymbol{X}_m \sim \text{MVN}(\mathbf{0}_P, \boldsymbol{\Sigma}) \quad \text{where} \quad \mathbf{0}_P = [0, \ldots, 0] \in \mathbb{R}^P \quad \text{and} \quad \boldsymbol{\Sigma}_{i,j} \in \mathbb{R}^{P\times P} := \Sigma_{i,j} = 0.25^{|i-j|}.$$

The residuals, $\eta_{t,m}$ are independently distributed such that

$$\eta_{t,m} \sim \text{N}(0, 0.5).$$

Now that we have introduced the models that will be used in simulation studies throughout the remainder of this this thesis we will consider the possibility of reducing the time required to solve the SBS problem.

5.2 Introduction

In Chapter 3 we introduced the Simultaneous Best-Subset (SBS) problem to select predictors simultaneously for systems of linear regression models of the form

$$\begin{aligned} y_{t,1} &= \sum_{p=1}^{P} x_{t,p,1}\beta_{p,1} + \eta_{t,1}, \\ &\dots \\ y_{t,M} &= \sum_{p=1}^{P} x_{t,p,M}\beta_{p,M} + \eta_{t,M}. \end{aligned} \tag{5.2.1}$$

Recall that the SBS problem is defined as,

$$\min \left[\sum_{m=1}^{M} \sum_{t=1}^{T} \left(y_{t,m} - \sum_{p=1}^{P} x_{t,p,m}\beta_{p,m} \right)^2 \right] \quad \text{subject to} \quad \left| \bigcup_{m=1}^{M} \mathcal{S}_m \right| \leq k. \tag{5.2.2}$$

Here, $\mathcal{S}_m = \{p : \beta_{p,m} \neq 0 \text{ for } p = 1, \dots, P\}$ denotes the predictors selected in model m. The SBS problem seeks to minimise the sum of squared residuals across M regression models, providing that at most k unique predictors are included across all models. We denote k as the overall model sparsity. The sparsity of each regression model fit in this joint approach cannot exceed k and each model typically assumes the same predictors. In Chapter 3, we described a number of practical procedures that ensured solving the SBS problem is feasible. These procedures included selecting a maximum level of sparsity, $K_{\max} << P$ when P is large. In addition to this a maximum runtime for the optimisation solver can be provided as good solutions are often found very quickly. We explained how $K_{\max}$ can be determined automatically when constraints that exclude predictors with high pairwise correlation are used. We now consider whether using MIQO models with data specific parameters can reduce the time to solve the SBS problem.

It is possible to formulate optimisation problems using parameterised formulations. Parameterised formulations are used to solve many mixed integer optimisation problems. An example includes the use of big-M parameters to *activate* constraints (Dai et al., 2019). Consider the two constraints

$$\beta_{p,m} \leq \eta_p M \quad \text{and} \quad -\eta_p M \leq \beta_{p,m}, \tag{5.2.3}$$

where $\beta_{p,m} \in \mathbb{R}$ and $\eta_p \in \{0, 1\}$. If $\eta_p = 0$ then the constraints given in (5.2.3) are satisfied only if $\beta_{p,m} = 0$. The idea is to choose M large enough such that if $\beta^*_{p,m}$ denotes the optimal value of $\beta_{p,m}$ in

a solution, then $M > \beta_{p,m}$ and $-M < \beta_{p,m}$. When $\eta_p = 0$, the constraints in (5.2.3) are *active*, but inactive otherwise. Soltysik and Yarnold (2010) present a formulation for the multivariable optimal discriminant analysis model using big-M parameters and discuss an approach to obtain a lower bound on M, thereby improving the computational efficiency in solving the associated problem.

For the simultaneous best-subset formulations big-M parameters can reduce the feasible solution space. The idea is to reduce the solution space of the formulation when the integer constraints are relaxed so that the integer variables are *closer* to integer solutions in the relaxed problem. Bertsimas et al. (2016) use this idea to improve the performance of the optimiser for solving the best-subset problem.

The parameters used in a formulation may be data dependent. This means that a parameter used in a formulation for one dataset may not give optimal solutions to the mathematical problem when used with another dataset. This could be because a parameter provided for the optimisation problem is too small hence the optimal solution to the mathematical problem is not feasible for the optimisation formulation. We will now demonstrate this with an example. Consider the best-subset problem

$$\min \left[\sum_{t=1}^{T} \left(y_t - \sum_{p=1}^{P} x_{t,p}\beta_p \right)^2 \right] \quad \text{subject to } ||\boldsymbol{\beta}||_0 \leq k. \tag{5.2.4}$$

The objective is to minimise the sum of squared residuals subject to at most k predictors present in the model. Bertsimas et al. (2016) present a more structured version of the best-subset problem,

$$\begin{aligned} \min \left[\sum_{t=1}^{T} \left(y_t - \sum_{p=1}^{P} x_{t,p}\beta_p \right)^2 \right] \quad & \text{subject to,} \\ -M_U \leq \beta_p \leq M_U, \quad & \text{for } p = 1, \ldots, P, && \text{(5.2.5a)} \\ ||\boldsymbol{\beta}||_1 \leq M_l, & && \text{(5.2.5b)} \\ ||\boldsymbol{\beta}||_0 \leq k. & \end{aligned}$$

Here, two additional types of constraint have been added. Firstly, constraint (5.2.5a) bounds the maximum absolute value of all regression coefficients. Secondly, the l_1 norm, $||\boldsymbol{\beta}||_1 = \sum_{p=1}^{P} |\beta_p|$ is bounded above by M_l in constraint (5.2.5b). Provided M_l and M_U are chosen to be significantly large then solutions to problem (5.2.5) will also be solutions to problem (5.2.4). Parameters M_l and M_U chosen for one dataset may not be large enough to obtain optimal solutions to problem (5.2.4) for another dataset.

In Section 3.4.4 we observed how significant improvements to the time to solve the SBS problem can be achieved by setting the lower bound for all regression coefficients to zero. The motivation for this was to exploit application specific knowledge. Here we assumed that an increasing value

in all predictors must increase the value of the response variable. This assumption may not be appropriate in general. Bertsimas and King (2016) provide a number of techniques that could be used to estimate data-specific parameters. Despite showing some improvements in the performance of the solver such as increasing the rate at which the lower-bound to the optimisation formulation increases, the authors failed to illustrate the practical improvements in the time to solve the best-subset problem. We shall now investigate the impact of parameterised formulations for solving the SBS problem.

5.3 Parameterised formulations for the SBS problem

We can generalise the *more structured* problem (5.2.5) to obtain a more structured problem associated to the SBS problem. Solutions to the SBS problem (7.1.1) can be obtained by solving

$$\begin{aligned} \min_{\boldsymbol{\beta}} \left[\sum_{m=1}^{M} \sum_{t=1}^{T} \left(y_{t,m} - \sum_{p=1}^{P} x_{t,p,m} \beta_{p,m} \right)^2 \right] \quad & \text{subject to,} \\ -M_U \leq \beta_{p,m} \leq M_U, \quad & \text{for } p = 1, \ldots, P, \ m = 1, \ldots, M, \\ ||\boldsymbol{\beta}||_1 \leq M_l, & \\ ||\boldsymbol{\beta}||_0 \leq k. & \end{aligned} \tag{5.3.1}$$

Here, $\boldsymbol{\beta} \in \mathbb{R}^{P \times M}$, and provided M_l and M_U are chosen sufficiently large the optimal solution to the problem given in (5.3.1) will provide the optimal solution to the SBS problem. Solving (5.3.1) will estimate a system of M regression models. Therefore, we could constrain the maximum absolute value and l_1 norm of the regression coefficients for each model. This suggests solving the following problem,

$$\begin{aligned} \min_{\boldsymbol{\beta}} \left[\sum_{m=1}^{M} \sum_{t=1}^{T} \left(y_{t,m} - \sum_{p=1}^{P} x_{t,p,m} \beta_{p,m} \right)^2 \right] \quad & \text{subject to,} \\ -M_U^m \leq \beta_{p,m} \leq M_U^m, \quad & \text{for } p = 1, \ldots, P, \ m = 1, \ldots, M, \\ ||\boldsymbol{\beta}||_1 \leq M_l^m, & \\ ||\boldsymbol{\beta}||_0 \leq k. & \end{aligned} \tag{5.3.2}$$

Again, provided M_l^m and M_U^m are chosen sufficiently large solving (5.3.2) will give us an optimal solution to the SBS problem given in (7.1.1). In problem (5.3.2) we have the absolute value of the regression coefficients and l_1 norm of the regression coefficients dependent on m. It is possible to formulate problems (5.3.1) and (5.3.2) as MIQO models. However, a little work is needed to include the constraint on the l_1 norm of the regression coefficients. The constraint

$$||\boldsymbol{\beta}||_1 \leq M_l^m,$$

can be satisfied by introducing variables $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{MP}$ and the constraints,

$$\begin{aligned} u_{p,m} - v_{p,m} = \beta_{p,m}, &\quad \text{for } p = 1, \ldots, P,\ m = 1, \ldots, M, \\ u_{p,m} \geq 0, &\quad \text{for } p = 1, \ldots, P,\ m = 1, \ldots, M, \\ v_{p,m} \geq 0, &\quad \text{for } p = 1, \ldots, P,\ m = 1, \ldots, M, \\ \sum_{m=1}^{M} \sum_{p=1}^{P} u_{p,m} + v_{p,m} \leq M_l. & \end{aligned}$$

We no longer require the $\mathcal{SOS}_1$ constraints to control $\beta_{p,m}$ or η_p, both taking non-zero values. We can control the zero-valued regression coefficients with M_U directly,

$$\begin{aligned} \beta_{p,m} + M_U \eta_p \leq M_U \iff \beta_{p,m} \leq (1 - \eta_p) M_U, &\quad \text{for } m = 1, \ldots, M,\ p = 1, \ldots, P, \text{ and,} \\ -\beta_{p,m} + M_U \eta_p \leq M_U \iff -M_U (1 - \eta_p) \leq \beta_{p,m}, &\quad \text{for } m = 1, \ldots, M,\ p = 1, \ldots, P. \end{aligned}$$

It is known generally in the optimisation literature that increasing the number of variables in a formulation to an optimisation problem can increase the time taken to solve the problem (Chen et al., 2010). Therefore, we will consider solving a more structured formulation of the SBS both with and without the l_1 constraints on the regression variables. We will compare the time taken to solve the following MIQO models.

Formulation 1: We used this formulation in Chapter 3 to solve the SBS problem. This formulation does not require any data specific parameters. Formulation 1 is given by

$$\min_{\boldsymbol{\beta}, \boldsymbol{\eta}} \left[\sum_{m=1}^{M} \sum_{t=1}^{T} \left(y_{t,m} - \sum_{p=1}^{P} x_{t,p,m} \beta_{p,m} \right)^2 \right] \quad \text{subject to,} \tag{5.3.3}$$

$$(\beta_{p,m}, \eta_p) \in \mathcal{SOS}_1, \quad \text{for } p = 1, \ldots, P,\ m = 1, \ldots, M, \tag{5.3.4}$$

$$-\sum_{p=1}^{P} \eta_p \leq k - P, \tag{5.3.5}$$

$$\beta_{p,m} \in \mathbb{R}, \quad \text{for } p = 1, \ldots, P,\ m = 1, \ldots, M, \tag{5.3.6}$$

$$\eta_p \in \{0, 1\}, \quad \text{for } p = 1, \ldots, P. \tag{5.3.7}$$

Formulation-l_∞-l_1: By adding the constraints for the maximum absolute value and l_1 norm of all regression coefficients we arrive at the following formulation,

$$\min_{\boldsymbol{\beta},\boldsymbol{\eta},\boldsymbol{u},\boldsymbol{v}} \left[\sum_{m=1}^{M} \sum_{t=1}^{T} \left(y_{t,m} - \sum_{p=1}^{P} x_{t,p,m} \beta_{p,m} \right)^2 \right] \quad \text{subject to,} \tag{5.3.8}$$

$$\beta_{p,m} + M_U \eta_p \leq M_U, \quad \text{for } p = 1, \ldots, P, \ m = 1, \ldots, M, \tag{5.3.9}$$

$$-\beta_{p,m} + M_U \eta_p \leq M_U, \quad \text{for } p = 1, \ldots, P, \ m = 1, \ldots, M, \tag{5.3.10}$$

$$-\sum_{p=1}^{P} \eta_p \leq k - P, \tag{5.3.11}$$

$$\beta_{p,m} \in \mathbb{R}, \quad \text{for } p = 1, \ldots, P, \ m = 1, \ldots, M, \tag{5.3.12}$$

$$\eta_p \in \{0, 1\}, \quad \text{for } p = 1, \ldots, P, \tag{5.3.13}$$

$$u_{p,m} \in \mathbb{R}^+, \quad \text{for } p = 1, \ldots, P, \ m = 1, \ldots, M, \tag{5.3.14}$$

$$v_{p,m} \in \mathbb{R}^+, \quad \text{for } p = 1, \ldots, P, \ m = 1, \ldots, M, \tag{5.3.15}$$

$$u_{p,m} - v_{p,m} = \beta_{p,m}, \quad \text{for } p = 1, \ldots, P, \ m = 1, \ldots, M, \tag{5.3.16}$$

$$\sum_{m=1}^{M} \sum_{p=1}^{P} u_{p,m} + v_{p,m} \leq M_l, \tag{5.3.17}$$

$$-M_U \leq \beta_{p,m} \leq M_U, \quad \text{for } p = 1, \ldots, P, \ m = 1, \ldots, M. \tag{5.3.18}$$

The name for this formulation indicates that both the l_∞ norm, $||\boldsymbol{\beta}||_\infty = \max_{p,m} \boldsymbol{\beta}$, and l_1 norm constraints are included in this formulation.

Formulation-l_∞^m-l_1^m: Model specific bounds, M_U^m and M_l^m for $m = 1, \ldots, M$, can be used to provide the following formulation,

$$\min_{\boldsymbol{\beta},\boldsymbol{\eta},\boldsymbol{u},\boldsymbol{v}} \left[\sum_{m=1}^{M} \sum_{t=1}^{T} \left(y_{t,m} - \sum_{p=1}^{P} x_{t,p,m} \beta_{p,m} \right)^2 \right] \quad \text{subject to,} \tag{5.3.19}$$

$$\beta_{p,m} + M_U^m \eta_p \leq M_U^m, \quad \text{for } p = 1, \ldots, P, \ m = 1, \ldots, M, \tag{5.3.20}$$

$$-\beta_{p,m} + M_U^m \eta_p \leq M_U^m, \quad \text{for } p = 1, \ldots, P, \ m = 1, \ldots, M, \tag{5.3.21}$$

$$-\sum_{p=1}^{P} \eta_p \leq k - P, \tag{5.3.22}$$

$$\beta_{p,m} \in \mathbb{R}, \quad \text{for } p = 1, \ldots, P, \ m = 1, \ldots, M, \tag{5.3.23}$$

$$\eta_p \in \{0, 1\}, \quad \text{for } p = 1, \ldots, P, \tag{5.3.24}$$

$$u_{p,m} \in \mathbb{R}^+, \quad \text{for } p = 1, \ldots, P, \ m = 1, \ldots, M, \tag{5.3.25}$$

$$v_{p,m} \in \mathbb{R}^+, \quad \text{for } p = 1, \ldots, P, \ m = 1, \ldots, M, \tag{5.3.26}$$

$$u_{p,m} - v_{p,m} = \beta_{p,m}, \quad \text{for } p = 1, \ldots, P, \ m = 1, \ldots, M, \tag{5.3.27}$$

$$\sum_{m=1}^{M} \sum_{p=1}^{P} u_{p,m} + v_{p,m} \leq M_l^m, \tag{5.3.28}$$

$$-M_U^m \leq \beta_{p,m} \leq M_U^m, \quad \text{for } p = 1, \ldots, P, \ m = 1, \ldots, M. \tag{5.3.29}$$

The name for this formulation indicates that the l_∞ norm constraints that are *response specific* are included on the regression coefficients.

Formulation-l_∞: Finally, we consider constraining the maximum absolute values of the regression coefficients without the constraint on the l_1 norm as follows,

$$\min_{\boldsymbol{\beta},\boldsymbol{\eta}} \left[\sum_{m=1}^{M} \sum_{t=1}^{T} \left(y_{t,m} - \sum_{p=1}^{P} x_{t,p,m} \beta_{p,m} \right)^2 \right] \quad \text{subject to,} \tag{5.3.30}$$

$$\beta_{p,m} + M_U \eta_p \leq M_U \quad \text{for } p = 1, \ldots, P, \ m = 1, \ldots, M, \tag{5.3.31}$$

$$-\beta_{p,m} + M_U \eta_p \leq M_U \quad \text{for } p = 1, \ldots, P, \ m = 1, \ldots, M, \tag{5.3.32}$$

$$-\sum_{p=1}^{P} \eta_p \leq k - P, \tag{5.3.33}$$

$$\beta_{p,m} \in \mathbb{R}, \quad \text{for } p = 1, \ldots, P, \ m = 1, \ldots, M, \tag{5.3.34}$$

$$\eta_p \in \{0, 1\}, \quad \text{for } p = 1, \ldots, P, \tag{5.3.35}$$

$$-M_U \leq \beta_{p,m} \leq M_U \quad \text{for } p = 1, \ldots, P, \ m = 1, \ldots, M. \tag{5.3.36}$$

The name for this formulation indicates that only the l_∞ constraints are included in this formulation.

5.3.1 Estimating the parameters: A demonstration

To determine the values M_U, M_l, and M_U^m for $m = 1, \ldots, M$, we can solve the unparameterised formulation for the SBS problem given by Formulation-1 in (5.3.3). Denote the optimal solution obtained by solving Formulation-1 by $\boldsymbol{\beta}^*$ then the parameters for the parameterised formulations can then be estimated using $\boldsymbol{\beta}^*$ as follows,

$$M_U = ||\boldsymbol{\beta}^*||_\infty \quad \text{and,} \quad M_U^m = ||\boldsymbol{\beta}^*_{*,m}||_\infty.$$

Ordering the values $|\boldsymbol{\beta}^*_{(1)}| \geq |\boldsymbol{\beta}^*_{(2)}| \geq \ldots \geq |\boldsymbol{\beta}^*_{(P)}|$ the parameter M_l can be set

$$M_l = \sum_{i=1}^{k} |\boldsymbol{\beta}^*_{(i)}|.$$

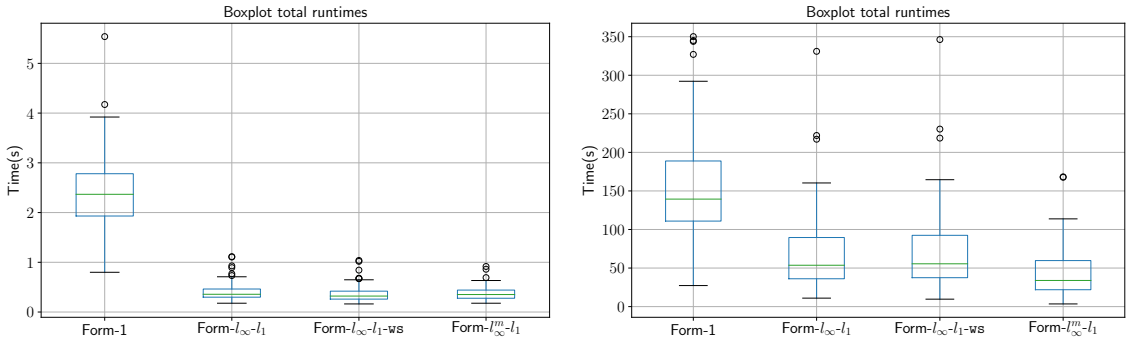
Here, k denotes the sparsity of the SBS problem. Estimating parameters for the parameterised formulations by first solving the unparameterised formulation is not practically sensible. Once the optimal solution to the SBS problem is obtained there is no value in solving an alternative formulation of the problem. However, by estimating the parameters in this way we can ensure two things. Firstly, the parameters are valid. This means that using them will ensure we can obtain the optimal solution to the SBS problem as the parameters will be *sufficiently large*. Secondly, these parameters will be as small as possible. This means that by solving a formulation with parameters any smaller will not produce optimal solutions to the SBS problem. Further, if we do not observe an improvement in the time to solve the parameterised formulations when the parameters are derived from the optimal solution, it is unlikely that we would observe a reduction in time to solve the SBS problem if the parameters were estimated by any other means.

5.3.2 Motivating demonstration

The purpose of this section is to determine whether it is possible to reduce the time required to solve the SBS problem using the parameterised formulations given in Section 5.3. We generate 100 datasets from the *Application* model defined in Section 5.1. We consider the time taken to solve the SBS problem for $k \in \{5, 10\}$. Using smaller values of k is unlikely to show large differences in the time to solve the SBS problem as Gurobi can solve these problems in a very short amount of time.

Figure 5.3.1 shows the box-plots for the total time to solve the SBS problem using the four formulations described in Section 5.3. The results are presented abbreviating *Formulation* to *Form*. Figure 5.3.1a shows that the total time to solve the SBS problem using Formulation-l_∞-l_1, Formulation-l_∞-l_1-ws and Formulation-l_∞^m-l_1 appear very similar and much lower in comparison to the unparameterised formulation (Formulation-1). The total solve time for solving the SBS problem with $k = 10$ is shown in Figure 5.3.1b. Again, the parameterised formulations appear to perform more favourably.

Formulation-l^m_∞-l_1, which includes the model specific parameters, provides the shortest times to solve the SBS problem. The largest time taken by Formulation-l^m_∞-l_1 is nearly half the largest times of any of the other three methods when $k = 10$.



(a) $k = 5$ (b) $k = 10$

Figure 5.3.1: Box-plots for the total time to solve the SBS problem, using the formulations proposed in Section 5.3. The boxes indicate the lower-quartile, median and upper quartile of 100 runtimes for each formulation. The points identify runtimes greater than 1.5 times the inter-quartile range.

In practise, it is not feasible to determine the parameters for parameterised formulations by first solving the SBS problem using an unparameterised formulation. We have shown here that a reduction in the total time to solve the SBS problem can be reduced with parameterised formulations. We will now consider how to estimate the formulation parameters practically. The idea here is to produce a good feasible solution to the SBS problem quickly and then use the feasible solution to estimate the formulation parameters.

5.4 A discrete first-order approach to the SBS problem

Bertsimas et al. (2016) develop a discrete extension of first-order methods in convex optimisation (Nesterov, 2005) to obtain near optimal solutions to problems of the form,

$$\min_{\boldsymbol{\beta}} g(\boldsymbol{\beta}) \quad \text{subject to } ||\boldsymbol{\beta}||_0 \leq k. \tag{5.4.1}$$

Here, we consider one linear regression model only where $\boldsymbol{\beta} \in \mathbb{R}^{P\times 1}$ and the response and predictor observations are given by,

$$\boldsymbol{y} \in \mathbb{R}^{T\times 1} \quad \text{and} \quad \boldsymbol{x} \in \mathbb{R}^{T\times P}.$$

We consider modifying the methods proposed by Bertsimas et al. (2016) to obtain good solutions to

the SBS problem. Recall from Chapter 2 that the objective in (5.4.1) for the best-subset problem is

$$g(\boldsymbol{\beta}) = \sum_{t=1}^{T} \left(y_t - \sum_{p=1}^{P} x_{t,p}\beta_p \right)^2 .$$

Bertsimas et al. (2016) proposed Algorithm 1 to obtain good feasible solutions to the best-subset problem. This algorithm uses a convergence tolerance, ϵ and a parameter, L which must be greater than or equal to the largest eigenvalue of $\boldsymbol{x'x}$. The hard-thresholding operator (Donoho and Johnstone, 1994), $\boldsymbol{H}_k(\boldsymbol{c})$ used in Algorithm 1, is defined as follows. Let $\hat{\boldsymbol{\beta}} \in \boldsymbol{H}_k(\boldsymbol{c})$ and order the values $|c_{(1)}| \geq |c_{(2)}| \geq \ldots \geq |c_{(P)}|$ then

$$\hat{\beta} = \begin{cases} c_i, & \text{if } i \in \{(1) \ldots, (k)\}, \\ 0, & \text{otherwise.} \end{cases}$$

Here, $\hat{\beta}_i$ is the i^{th} coordinate of $\hat{\boldsymbol{\beta}}$. Algorithm 1 applies the hard-thresholding operator to a gradient descent update of the regression coefficients. Note the dependence of the hard-thresholding operator on k as it uses the k largest values in absolute value of the input. We could apply Algorithm 1 with the SBS objective,

$$g(\boldsymbol{\beta}) = \sum_{m=1}^{M} \sum_{t=1}^{T} \left(y_{t,m} - \sum_{p=1}^{P} x_{t,m,p}\beta_{p,m} \right)^2 .$$

However, for $k = 1$ only one of the MP coefficients will be non-zero. A feasible solution to the SBS problem will allow up to M coefficients to be non-zero, providing each of these coefficients corresponds to the same predictor. We modify Algorithm 1 so that at sparsity k, we obtain a solution of $\boldsymbol{\beta} \in \mathbb{R}^{P \times M}$ with kM non-zero coefficients. We use the superscript i to denote the i^{th} estimate of the regression coefficients obtained in our iterative algorithm.

Algorithm 1 Discrete first-order algorithm proposed by Bertsimas et al. (2016) to obtain good feasible solutions to the best-subset problem.

1: Initialise with $\boldsymbol{\beta}^0 \in \mathbb{R}^{P \times 1}$ such that $||\boldsymbol{\beta}^0||_0 \leq k$.
2: **for** $i \geq 1$ **do**
$\quad \boldsymbol{\beta}^i \in \boldsymbol{H}_k \left(\boldsymbol{\beta}^{i-1} - \frac{1}{L} \nabla g(\boldsymbol{\beta}^{i-1}) \right).$
3: $\quad$ **if** $\left| g(\boldsymbol{\beta}^i) - g(\beta^{i-1}) \right| \leq \epsilon$ **then**
4: $\quad\quad$ **return** $\boldsymbol{\beta}^i$
5: $\quad$ **end if**
6: **end for**

We propose to apply a gradient decent update to the individual model coefficients and modify the hard-thresholding operator such that the same indices of the non-zero coefficients are chosen for each

model. Applying a gradient descent on each model should ensure we obtain good coefficient updates for each model. A modified hard-thresholding operator will ensure that the non-zero coefficients in each of the M models correspond to the same predictors.

We will first introduce the notation used to describe our Discrete First-Order Algorithm (DFOA) for the SBS problem. The idea is to initialise an algorithm with a feasible solution to the SBS problem, and then combine a gradient descent algorithm with a hard-thresholding operator. The gradient decent step will produce new values for the regression coefficients that reduce the value of the objective function. However, the new values for the regression coefficients are not guaranteed to satisfy the sparsity constraint of the SBS problem. We use the hard-thresholding operator to determine which coefficients should be set to zero in order to satisfy the sparsity constraints. By initialising the algorithm with a number of feasible solutions we can improve the chance of obtaining good feasible solutions.

Recall that we are trying to obtain good regression coefficients for the system of linear regression models (5.2.1). All $M \times P$ regression coefficients are denoted $\boldsymbol{\beta} \in \mathbb{R}^{P \times M}$. Let the sets, $\mathcal{I}_m(\boldsymbol{\beta})$ be defined as,

$$\mathcal{I}_m(\boldsymbol{\beta}) = \{p : \beta_{p,m} \neq 0, \text{ for } p = 1, \ldots, P\}, \text{ for } m = 1, \ldots, M.$$

Here, $\mathcal{I}_m(\boldsymbol{\beta})$ gives the non-zero coefficient indices for model m. If $|\bigcup_{m=1}^{M} \mathcal{I}_m(\boldsymbol{\beta})| \leq k$, then each model has at most k non-zero coefficients since $\mathcal{I}_m(\boldsymbol{\beta}) \subseteq \bigcup_{m=1}^{M} \mathcal{I}_m(\boldsymbol{\beta})$, for $m = 1, \ldots, M$. The SBS objective can be re-written,

$$g(\boldsymbol{\beta}) = \sum_{m=1}^{M} g_m(\boldsymbol{\beta}),$$

where $g_m(\boldsymbol{\beta}) = \sum_{t=1}^{T} \left(y_{t,m} - \sum_{p=1}^{P} x_{t,p,m} \beta_{p,m} \right)^2$, gives the residual sums of squares for model m. The first derivatives with respect to $\beta_{p,m}$ follow,

$$\frac{d}{d\beta_{p,m}} g_m(\boldsymbol{\beta}) = \nabla g_m(\boldsymbol{\beta}) = -2 \sum_{t=1}^{T} (x_{t,p,m}) \sum_{t=1}^{T} \left(y_{t,m} - \sum_{p=1}^{P} x_{t,p,m} \beta_{p,m} \right).$$

Finally, we propose a modification of the hard-thresholding operator, which we call the *modified hard-thresholding operator* to ensure a solution, $\boldsymbol{\beta} \in \tilde{H}_k(\boldsymbol{\beta})$ is feasible for the SBS problem with sparsity k[1]. If we order the coefficients as follows,

$$\sum_{m=1}^{M} |\beta_{(1),m}| \geq \sum_{m=1}^{M} |\beta_{(2),m}| \geq \cdots \geq \sum_{m=1}^{M} |\beta_{(P),m}|,$$

then let

$$\hat{\beta}_{p,m} = \begin{cases} \beta_{p,m}, & \text{if } p \in \{(1), (2), \ldots, (k)\}, \\ 0, & \text{otherwise,} \end{cases}$$

[1]Note that the data has been standardised as described in Chapter 2

so that $\hat{\boldsymbol{\beta}} \in \tilde{H}_k(\boldsymbol{\beta})$. The algorithm requires a convergence tolerance, ϵ and parameters, L_m for $m = 1, \ldots, M$. We set L_m equal to the largest eigenvalue of the matrix $\boldsymbol{x}'_{*,*,m}\boldsymbol{x}_{*,*,m}$, for $m = 1, \ldots, M$ and $\epsilon = 1e^{-6}$. Pseudo-code for our algorithm is given by Algorithm 2.

Algorithm 2 A discrete first order algorithm to obtain feasible solutions to the simultaneous best-subset problem.

Initialise with $\boldsymbol{\beta}^0 \in \mathbb{R}^{P\times M}$ such that $\left|\bigcup_{m=1}^{M} \mathcal{I}(\boldsymbol{\beta}_{*,m})\right| \leq k$
for $i \geq 1$ **do**,
  for $m = 1, \ldots, M$ **do**
    $\boldsymbol{\beta}_m = \boldsymbol{\beta}^{i-1}_{*,m} - \frac{1}{L_m}\nabla g_m(\boldsymbol{\beta})$
  end for
  $\boldsymbol{\beta}^i \in H_k([\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_M])$
  if $|\sum_{m=1}^{M} g_m(\boldsymbol{\beta}) - \sum_{m=1}^{M} g_m(\boldsymbol{\beta}^{i-1})| \leq \epsilon$ **then**
    return $\boldsymbol{\beta}^i$
  end if
end for

Algorithm 2 can be initialised by randomly selecting k predictors from $\{1, \ldots, P\}$ to be present in the regression models, then estimating the associated coefficients by minimising the SBS objective. We will investigate the quality of solutions obtained from Algorithm 2 by comparing them to the optimal solutions to the SBS problem in Section 5.6.

Having considered how to obtain feasible solutions to the SBS problem we will now consider how to estimate parameters for the parameterised formulations given in Section 5.3.

5.5 Estimating formulation parameters

Given a feasible solution, $\boldsymbol{\beta}^*_k$ to the SBS problem with sparsity k, we can estimate the parameters for the parameterised formulations. Let $g(\boldsymbol{\beta}^*_k) = \mathcal{UB}_k$, denote the value of the SBS objective at the feasible solution $\boldsymbol{\beta}^*_k$. The objective value $g(\boldsymbol{\beta}^*_k)$ is an upper bound to the objective value of the SBS problem at an optimal solution. Alternative solutions (maintaining sparsity k) may exist that reduce the objective further. We estimate parameters for the parameterised formulations using the following idea. The idea is to consider the maximum and minimum value of all regression coefficients subject to the objective of the SBS problem not exceeding $\mathcal{UB}_k$. We consider ensuring the objective remains below $\mathcal{UB}_k$ ignoring the sparsity of the solutions so we can determine a valid upper bound to all of the regression coefficients. Finding the maximum and minimum values of each $\beta_{p,m}$ subject

to $g(\boldsymbol{\beta}^*) \leq \mathcal{UB}_k$ can be stated using two convex quadratically constrained optimisation problems,

$$u^+_{p,m} := \max_{\boldsymbol{\beta}} \beta_{p,m} \quad \text{subject to,} \qquad u^-_{p,m} := \min_{\boldsymbol{\beta}} \beta_{p,m} \quad \text{subject to,}$$
$$\sum_{m=1}^{M}\sum_{p=1}^{P}\left(y_{t,m} - \sum_{p=1}^{P} x_{t,p,m}\beta_{p,m}\right)^2 \leq \mathcal{UB}_k, \quad \text{and,} \quad \sum_{m=1}^{M}\sum_{p=1}^{P}\left(y_{t,m} - \sum_{p=1}^{P} x_{t,p,m}\beta_{p,m}\right)^2 \leq \mathcal{UB}_k. \tag{5.5.1}$$

Here, $u^+_{p,m}$ and $u^-_{p,m}$ are the largest and smallest values respectively of $\beta_{p,m}$ providing $g(\boldsymbol{\beta}) \leq \mathcal{UB}_k$, whilst allowing all other variables to vary. Since all other variables are allowed to take non-zero values, i.e there is no sparsity constraints on $\boldsymbol{\beta}$, the value

$$M_U = \max_{p,m}\{|u^+_{p,m}|, |u^-_{p,m}|\},$$

gives a valid upper bound to the maximum absolute value of all regression coefficients to the SBS problem with sparsity k. To see this, consider solving the SBS problem with sparsity k_i. Let

$$g_{k_i} := \min_{\boldsymbol{\beta}} \left[\sum_{m=1}^{M}\sum_{t=1}^{T}\left(y_{t,m} - \sum_{p=1}^{P} x_{t,p,m}\beta_{p,m}\right)^2\right] \quad \text{subject to } \left|\bigcup_{m=1}^{M} \mathcal{I}_m(\boldsymbol{\beta})\right| \leq k_i,$$

then $g_{k_j} \leq g_{k_i}$ if $k_j \geq k_i$. This can be seen as the optimal solution giving g_{k_i} is a feasible solution to the SBS problem with sparsity k_j. Hence g_{k_j} will not exceed g_{k_i}. Therefore, seeking the minimum and maximum of all $\beta_{p,m}$ subject to

$$\sum_{m=1}^{M}\sum_{t=1}^{T}\left(y_{t,m} - \sum_{p=1}^{P} x_{t,p,m}\beta_{p,m}\right)^2 \leq \mathcal{UB}_k, \tag{5.5.2}$$

will produce a smaller and larger values for $\beta_{p,m}$ when minimising and maximising respectively, than if sparsity constraints were also placed on $\boldsymbol{\beta}$.

The parameter M_l can be determined by ordering the maximum absolute values of all bounds of the regression coefficients. Taking the largest kM regression coefficients in absolute value as we have M regression models each with sparsity k. Order the bounds as follows,

$$\left|u^{(sgn_1)}_{(p_1,m_1)}\right| \geq \left|u^{(sgn_2)}_{(p_2,m_2)}\right| \geq \ldots \geq \left|u^{(sgn_{MP})}_{(p_{2MP},m_{2MP})}\right|.$$

Then, $M_l = \sum_{i=1}^{kM} u^{(sgn_i)}_{(p_i,m_i)}$ gives a valid upper bound for the l_1 norm of the regression coefficients. We have not determined whether it is possible to determine valid bounds for the model specific l_1 norm, M^m_l and maximum absolute bound, M^m_U used in Formulation-l^m_∞-l_1. We leave further details of this for the discussion in Section 5.7.

Bertsimas et al. (2016) suggest solving the convex quadratic programs given in 5.5.1 (with only one response variable) with an optimisation solver. However these solutions are available analytically. This is considered in the following section.

5.5.1 An analytical approach

We will now explain how the solutions to the convex quadratic programs given in (5.5.1) are available in closed form. The idea here is to fix $\beta_{p,m}$ and find an analytical solution for all other coefficients, β_{p^*,m^*} where $p^* \neq p$ and $m^* \neq m$. Then, fixing the values of β_{p^*,m^*} for all $p^* \neq p$ and for all $m^* \neq m$, we solve a quadratic equation to find both the largest and smallest values of $\beta_{p,m}$, providing the objective of the SBS problem does not exceed $\mathcal{UB}$.

It is well known that the solution to the least squares problem

$$\min_{\boldsymbol{\beta}} \left[\sum_{t=1}^{T} \left(y_t - \sum_{p=1}^{P} x_{t,p}\beta_p \right)^2 \right] = (\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{x}\boldsymbol{\beta})$$

is given by the least squares estimator $\hat{\boldsymbol{\beta}} = (\boldsymbol{x}'\boldsymbol{x})^{-1}\boldsymbol{x}'\boldsymbol{y}$. Consider the following modified least squares problem,

$$g(\boldsymbol{\beta}_{(-i)}) := \min_{\beta_1,\ldots,\beta_{i-1},\beta_{i+1},\beta_P} \sum_{t=1}^{T} \left(y_t - \sum_{p\neq i} x_{t,p}\beta_p \right)^2$$

where we remove predictor X_i and variable β_i from the minimisation problem. This is available in closed form by removing the column of $\boldsymbol{x}$ relating to predictor X_i. Let $\boldsymbol{x}_{(-i)} \in \mathbb{R}^{T\times(P-1)}$ denote the predictor matrix with predictor X_i removed. Then, the solution to the modified least squares problem is given by $\tilde{\boldsymbol{\beta}} = (\boldsymbol{x}'_{(-i)}\boldsymbol{x}_{(-i)})^{-1}\boldsymbol{x}'_{(-i)}\boldsymbol{y}$. We find the values of β_p, where $p \neq i$, that minimise the modified least squares objective. Given the objective of the modified least squares problem, $g(\boldsymbol{\beta}_{(-p)})$ we solve the quadratic equation

$$\sum_{t=1}^{T} x_{t,i}\beta_i^2 + \sum_{t=1}^{T} x_{t,i}y_t\beta_i + \sum_{t=1}^{T} y_t = \mathcal{UB}_k - g(\boldsymbol{\beta}_{(-i)}). \tag{5.5.3}$$

Here, $\mathcal{UB}_k$ is a valid upper bound to the best-subset problem with sparsity k. The quadratic equation (5.5.3) can be solved using the quadratic formula. We will now write the SBS objective using matrix notation and show how bounds for the variables for the SBS problem can be obtained analytically. Let,

$$\boldsymbol{\beta}_{-(p,m)} = \begin{bmatrix} \beta_{1,1} \\ \cdots \\ \beta_{p-1,m} \\ \beta_{p+1,m} \\ \cdots \\ \beta_{p,m} \end{bmatrix} \quad \text{and,} \quad \boldsymbol{x}_{-(m,p)} = \begin{bmatrix} x_{1,1,1} & & & & & & \mathbf{0} \\ & \ddots & & & & & \\ & & x_{1,p-1,m} & x_{1,p+1,m} & & & \\ & & \vdots & \vdots & & & \\ & & x_{T,p-1,m} & x_{T,p+1,m} & & & \\ & & & & & \ddots & \\ \mathbf{0} & & & & & & x_{T,P,M} \end{bmatrix}.$$

The above notation indicates that the row corresponding to $\beta_{p,m}$ is removed from all regression coefficients and the column corresponding to predictor p for model m from $\boldsymbol{x}$ respectively. We define the *modified simultaneous least squares estimator*

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{-(p_j,m_i)} &= \arg\min \left[\sum_{m \neq m_i} \sum_{t=1}^{T} \left(\tilde{y}_{t,m} - \sum_{p \neq p_j} x_{t,p,m} \beta_{p,m} \right)^2 \right] \\ &= (\boldsymbol{x}'_{-(p_j,m_i)} \boldsymbol{x}_{-(p_j,m_i)})^{-1} \boldsymbol{x}'_{-(p_j,m_i)} \tilde{\boldsymbol{y}},\end{aligned}$$

where $\tilde{y}_{t,m} = y_{t,m} - x_{t,p,m}\beta_{p,m}$. The closed form expression for the modified least squares estimator is given by

$$\begin{aligned}g_{-(p,m)}(\hat{\boldsymbol{\beta}}_{-(p,m)}) =& ||\tilde{\boldsymbol{y}} - \boldsymbol{x}_{-(p,m)} \hat{\boldsymbol{\beta}}_{-(p,m)}||_2^2 \\ =& ||\tilde{\boldsymbol{y}} - \boldsymbol{x}_{-(p,m)} (\boldsymbol{x}'_{-(p,m)} \boldsymbol{x}_{-(p,m)})^{-1} \boldsymbol{x}_{-(p,m)} \tilde{\boldsymbol{y}}||_2^2 \\ =& (\tilde{\boldsymbol{y}} - \boldsymbol{x}_{-(p,m)} (\boldsymbol{x}'_{-(p,m)} \boldsymbol{x}_{-(p,m)})^{-1} \boldsymbol{x}_{-(p,m)} \tilde{\boldsymbol{y}})' \times \\ & (\tilde{\boldsymbol{y}} - \boldsymbol{x}_{-(p,m)} (\boldsymbol{x}'_{-(p,m)} \boldsymbol{x}_{-(p,m)})^{-1} \boldsymbol{x}_{-(p,m)} \tilde{\boldsymbol{y}}) \\ =& \tilde{\boldsymbol{y}}' (\boldsymbol{I}_{MT} - \boldsymbol{x}_{-(p,m)} (\boldsymbol{x}'_{-(p,m)} \boldsymbol{x}_{-(p,m)})^{-1} \boldsymbol{x}_{-(p,m)})' \times \\ & (\boldsymbol{I}_{MT} - \boldsymbol{x}_{-(p,m)} (\boldsymbol{x}'_{-(p,m)} \boldsymbol{x}_{-(p,m)})^{-1} \boldsymbol{x}_{-(p,m)}) \tilde{\boldsymbol{y}} \\ =& \tilde{\boldsymbol{y}}' \boldsymbol{A}_{p,m} \tilde{\boldsymbol{y}} \\ =& (\boldsymbol{y} - \boldsymbol{x}_{p,m}\beta_{p,m})' \boldsymbol{A}_{p,m} (\boldsymbol{y} - \boldsymbol{x}_{p,m}\beta_{p,m}) \\ =& \boldsymbol{x}'_{p,m} \boldsymbol{A}_{p,m} \boldsymbol{x}_{p,m} \beta_{p,m}^2 - 2\boldsymbol{y}' \boldsymbol{A}_{p,m} \boldsymbol{x}_{(p,m)} \beta_{p,m} + \boldsymbol{y}' \boldsymbol{A}_{p,m} \boldsymbol{y}.\end{aligned} \tag{5.5.4}$$

Setting (5.5.4) equal to $\mathcal{UB}_k$, gives us a quadratic equation in $\beta_{p,m}$ as follows,

$$\boldsymbol{x}'_{(p,m)} \boldsymbol{A}_{p,m} \boldsymbol{x}_{(p,m)} \beta_{p,m}^2 - 2\boldsymbol{y}' \boldsymbol{A}_{p,m} \boldsymbol{x}_{(p,m)} \beta_{p,m} + \boldsymbol{y}' \boldsymbol{A}_{p,m} \boldsymbol{y} = \mathcal{UB}_k. \tag{5.5.5}$$

Here,

$$\boldsymbol{A}_{p,m} = (\boldsymbol{I}_{MT} - \boldsymbol{x}_{-(p,m)} (\boldsymbol{x}'_{-(p,m)} \boldsymbol{x}_{-(p,m)})^{-1} \boldsymbol{x}_{-(p,m)})' (\boldsymbol{I}_{MT} - \boldsymbol{x}_{-(p,m)} (\boldsymbol{x}'_{-(p,m)} \boldsymbol{x}_{-(p,m)})^{-1} \boldsymbol{x}_{-(p,m)}).$$

Solving (5.5.5), gives us the minimum and maximum values of $\beta_{p,m}$, such that the equation given in (5.5.2) holds.

So far in this chapter we have presented a number of parameterised MIQO formulations for the SBS problem and shown that by using these formulations it is possible to reduce the time required to solve the SBS problem. We have developed a DFOA to quickly determine feasible solutions to the SBS problem and shown how the SBS objective value at these feasible solution can be used to determine bounds on the regression coefficients. In the following section we investigate the quality of the solutions obtained using our DFOA and whether the MIQO formulations presented in Section 5.3 can be used in practice to reduce the time required to solve the SBS problem.

5.6 Simulation study

In this simulation study we will generate synthetic data from the models presented in Section 5.1. We aim to determine how feasible solutions to the SBS problem provided by the DFOA presented in Section 5.4 compare to the optimal SBS solution. In addition to this, we evaluate the practical reduction in time to solve the SBS problem using the MIQO formulations presented in Section 5.3.

5.6.1 Performance of the DFOA algorithm

In Section 5.4 we described a DFOA to quickly obtain feasible solutions to the SBS problem. In this section we investigate how quickly the DFOA can obtain solutions. We simulate data from the *Scaling* model given in Section 5.1 and determine how the DFOA scales with M, the number of response variables and P, the number of predictor variables. We run each simulation 50 times and present the average time to implement the DFOA. To investigate how the algorithm scales with P and M we fix $M = 5$ and $P = 30$ respectively. Figure 5.6.1 shows how the algorithm scales. Figure 5.6.1a shows that the DFOA appears to scale linearly with M. In contrast, Figure 5.6.1b shows that the DFOA appears to scale quadratically with P. Note the average time for Figure 5.6.1b is shown on the square-root scale.

We now consider how the DFOA performs in comparison to the SBS approach. The DFOA is used to quickly obtain solutions to the SBS problem but the *SBS approach* finds the optimal solutions to the SBS problem. We compare the performance using a number of criteria. We generate 750 observations from each model and split randomly to create a training set of 500 and validation set of 250 observations. We average all results over 50 simulations. We record the total time to solve both the SBS problem and implement the DFOA which we present as *Total Runtime*. We record the mean-squared estimation error of the regression coefficients defined as

$$\text{MSE}_e(\boldsymbol{\beta}) = \frac{1}{MT}\sum_{m=1}^{M}\sum_{p=1}^{P}\left(\beta_{p,m} - \hat{\beta}_{p,m}\right)^2.$$

Here, $\hat{\beta}_{p,m}$ denotes the estimate of the true regression coefficient value $\beta_{p,m}$. We also compare the mean-squared prediction accuracy of the system given by,

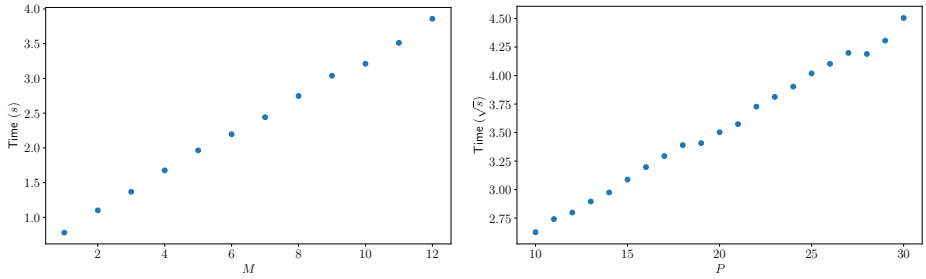
$$\text{MSE}_p(\boldsymbol{\beta}) = \frac{1}{MP}\sum_{m=1}^{M}\sum_{t=1}^{T}\left(y_{t,m} - \hat{y}_{t,m}\right)^2.$$

Here, $\hat{y}_{t,m}$ is the predicted value of $y_{t,m}$ where $y_{t,m}$ corresponds to an observation in the test set. We also consider how many of the true predictors the approach selects. Finally, denote the SBS objective value at an optimal solution as g^*. Then, we compare the relative accuracy of the objective

(Bertsimas et al., 2016) to the best solution provided by the DFOA as,

$$\text{Relative accuracy} = \frac{g^* - g}{g^*}.$$

Here, g denotes the lowest objective value of SBS problem given the 50 solutions provided by the DFOA.



(a) Scaling with M (b) Scaling with P

Figure 5.6.1: Scaling of the discrete first order algorithm as the number of predictors, P and the number of regression models, M increases.

We generate data from the *Adjacent*, *Application* and *Uniformly spaced* models given in Section 5.1 and apply the DFOA using three levels of sparsity. The true model sparsity is indicated in bold and we choose a value both greater and less than this value to compare. The relative accuracy of the DFOA is shown in Table 5.6.1. We can see that when $\rho = 0.25$ the DFOA was able to identify the optimal value when $k \leq k^*$, where k^* is the true model sparsity for all three models. When $\rho = 0.95$, the DFOA was able to find the optimal solution for the *Adjacent* model when $k = 1$. Clearly, the DFOA performs more favourably as a predictor selector when the correlation amongst the predictors is low.

Table 5.6.2 shows the performance measures for the models estimated using the SBS and DFOA approaches when applied to data generated from the *Uniformly spaced* model when $\rho = 0.95$. The average time to implement the DFOA is 50 times is less than 0.2 seconds for each level of sparsity. The time to solve the SBS problem is under 0.5 seconds when $k \in \{2, 5\}$. However, when $k = 25$ the time to solve the SBS problem takes 34 seconds on average. Despite the SBS approach taking over 180 times longer on average than the DFOA approach, the mean-squared estimation of the system is identical. Further, the mean-squared prediction error of the system is slightly lower for solutions to the SBS problem provided by the DFOA, and the DFOA correctly identified all five predictors used to generate the response data.

Table 5.6.3 shows the performance measures for the models estimated using the SBS and DFOA approaches when applied to data generated from the *Application* model when $\rho = 0.25$. We can

Table 5.6.1: Relative accuracy of the solutions to the SBS problem produced by the DFOA algorithm.

	Adjacent model			*Application* model			*Uniformly spaced* model		
	k = 1	**k = 3**	k = 7	k = 3	**k = 5**	k = 8	k = 2	**k = 5**	k = 25
$\rho = 0.95$	0	0.077	0.0276	0.3241	0.3179	0.1533	0.1253	0.6478	0.0068
$\rho = 0.25$	0	0	0.0027	0	0	0.0025	0	0	0.0043

see that the DFOA takes less than 2.5 seconds for each level of sparsity whilst the SBS approach actually solves the SBS problem in under one second, for all three levels of sparsity. Here, the time to solve the SBS problem to optimality is faster than implementing the DFOA approach 50 times. In this example the DFOA found the optimal solution in each simulation for all levels of sparsity.

Table 5.6.2: Performance of the DFOA when data is generated from the *Uniformly spaced* model and $\rho = 0.95$.

	$k = 2$		**k = 5**		$k = 25$	
	SBS	DFOA	SBS	DFOA	SBS	DFOA
Total Runtime	0.20	0.09	0.07	0.10	34.44	0.19
MSE Estimation	0.16	0.28	0.00	0.16	0.01	0.01
MSE Prediction	1.64	1.88	0.25	0.73	0.27	0.26
# True Predictors	2	1.08	5	2.04	5	5

In Section 5.A, we provide the summary results comparing the DFOA and the SBS approach for the *Adjacent*, *Application* and *Uniformly spaced* models when $\rho = 0.95$ and $\rho = 0.25$ that are not presented here. We find that the DFOA is consistent in the time to provide solutions to the SBS problem taking under three seconds on average in all simulations. In contrast, we found that the time to obtain the optimal solution to the SBS problem could vary much more, taking over 30 seconds on average in some cases. When $\rho = 0.25$, the DFOA appears to be able to identify the predictors used to generate the response variables. Hence, the mean-squared error in estimation and prediction for the system is very similar to that obtained from the optimal solution to the SBS problem. However, when $\rho = 0.95$ the predictors become more correlated and are more indistinguishable between one-and-other and the DFOA is not able to identify the predictors generating the response variables as accurately. The mean-squared estimation error of the system is typically worse for models estimated using the DFOA, although since the predictors are highly correlated, the mean-squared prediction error of the two approaches is very similar.

We can conclude that the DFOA can provide very good solutions to the SBS problem when

Table 5.6.3: Performance of the DFOA when data is generated from the *Application* model and $\rho = 0.25$.

	k = 3		k = 5		k = 8	
	SBS	DFOA	SBS	DFOA	SBS	DFOA
Total Runtime	0.09	2.04	0.15	2.41	0.82	2.42
MSE Estimation	0.00	0.00	0.00	0.00	0.00	0.00
MSE Prediction	0.37	0.37	0.25	0.25	0.26	0.26
# True Predictors	3	3	5	5	5	5

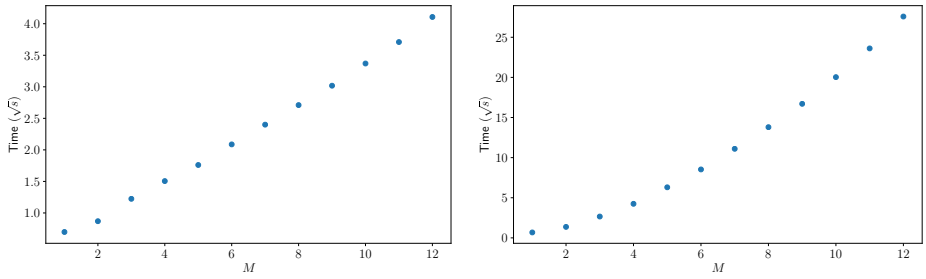
the correlation amongst predictors is low. The DFOA is not able to identify the predictors used to generate the response variables when the correlation amongst the predictors is high. However, when the correlation amongst the predictors is high, the DFOA can estimate models with predictive performance comparable to the optimal solution of the SBS approach. Given a good solution to the SBS problem, we now compare the performance of the two methods used to estimate the parameters of the formulations presented in Section 5.3.

5.6.2 Estimating formulation parameters

In Section 5.5 we discussed two methods for estimating the parameters for the MIQO models presented in Section 5.3. The first method was based on solving the convex quadratic programming problems proposed by Bertsimas et al. (2016). The second method used a closed form solution. In order to determine how best to estimate the parameters for the parameterised SBS formulations we now consider how each method scales with $M$and P. We use Gurobi (Gurobi Optimization, 2018) to solve all MP convex quadratic programs given in (5.5.1) giving the bounds on each regression coefficient. The closed form method was implemented in `Python3.6` (Python Software Foundation, 2017) using `numpy1.16.1` (Oliphant, 2006). We average our results over 25 simulations for each value of M and P. We simulate data from the *Scaling* model. To determine how the two approaches scale with M we fix $P = 30$. To determine how the two approaches scale with P we fix $M = 5$.

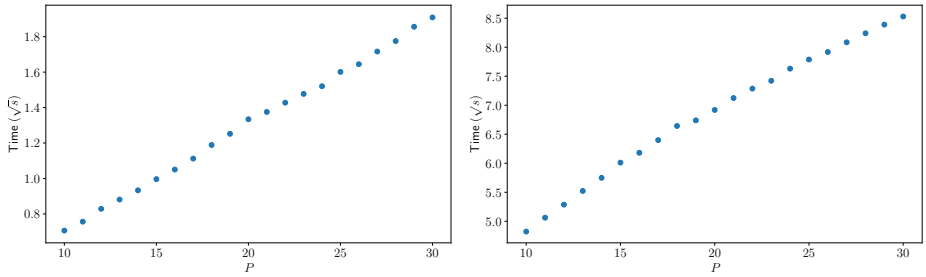
Figure 5.6.2 shows how the algorithms scale as M increases. The vertical axis shows the square-root of the total time taken. Figure 5.6.2a shows that solving all convex quadratic programs appear to scale quadratically with M. However, the parabola in Figure 5.6.2b shows that the closed form implementation scales at a worse than quadratic rate.

The closed form and convex quadratic program methods appear to scale quadratically with the number of predictors. Figure 5.6.3 shows the total time (on square root sale) to estimate the formulation parameters as P increases using the convex quadratic programs and the closed form



(a) CQB (b) CF

Figure 5.6.2: The runtime of the Convex Quadratic Programming and Closed Form methods for estimating the SBS formulation parameters, as M increases.



(a) CQB (b) CF

Figure 5.6.3: The runtime of the Convex Quadratic Programming and Closed Form methods for estimating the SBS formulation parameters, as P increases.

expression. Despite both algorithms scaling quadratically with P, the total runtime of the closed form method is considerably higher than using the convex quadratic programs. With $P = 35$ and $M = 5$ the closed form approach took over 70 seconds on average to estimate all of the parameters. In comparison, the convex quadratic program approach took a little over 3 seconds on average. It may be possibly to improve the computational time of the closed form approach by using specially designed routines that take advantage of the block diagonal matrices. We do not consider implementing this as Gurobi appears to solve all of the convex quadratic programs quickly.

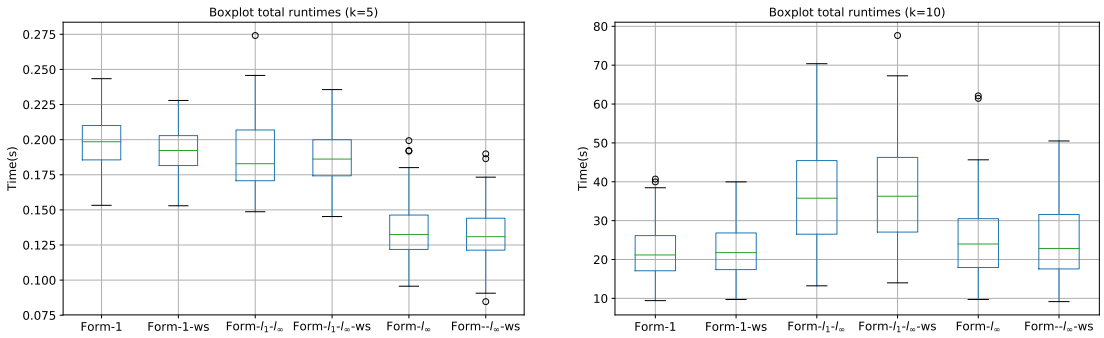
Now that we have determined which method we will use to estimate the formulation parameters, we will investigate the time to solve the SBS problem using the parameters estimated from the data.

5.6.3 Practical impact of warmstarts and parameterised formulations

We now consider the *practical* advantages of estimating the parameters used in parameterised formulations of the SBS problem. Particularly, we investigate if we can solve the SBS problem quicker using the parameterised formulations given Section 5.3 where we estimate the parameters using the

methods developed in Section 5.5.

Here, we compare the time required to solve the SBS problem using the following formulations, Formulation-1, Formulation-l_∞-l_1, and Formulation-l_∞. Formulation-1 does not use any parameters. Formulation-l_∞-l_1 requires the parameters $M_U \geq ||\boldsymbol{\beta}||_0$, and $M_l \geq ||\boldsymbol{\beta}||_1$. This formulation contains $M \times P$ additional variables in comparison to Formulation-1 due to the l_1 norm constraint. Finally, Formulation-l_∞ requires the parameter $M_U \geq ||\boldsymbol{\beta}||_0$. Each formulation is solved with and without a warmstart. These results will be presented as Form-*-ws and Form-* respectively. We do not compare Formulation-l_∞^m-l_1^m as we have not considered whether it is possible to obtain provable bounds on the model specific parameters and leave this to a discussion in Chapter 8. We simulate data from the *Application* model and consider the time to solve the SBS problem using sparsity levels $k \in \{5, 10\}$. We use the same simulation study given in Section 5.3.2 but here the parameters are estimated from the data rather than determined from optimal solutions to the SBS problem.



(a) $k = 5$. (b) $k = 10$.

Figure 5.6.4: The time to solve the SBS problem using three formulations discussed in Section 5.3. Each formulation is used with and without warmstarts, and the parameters are estimated using the CQP approach discussed in Section 5.5.

Figure 5.6.4 shows the box-plots of the time to solve the SBS problem with each formulation with and without using warmstarts. It appears that Formulation-l_∞ can produce the optimal solution for the SBS problem fastest at the true level of sparsity ($k = 5$). The spread of total solve times appears similar across all methods and there does not appear to be any significant advantage to providing the solver with a warmstart solution. However, observing the median times for each formulation, there appears to be less than 0.075 seconds difference. When $k = 10$, the difference between median times to solve the SBS problem using each formulation is greater. Here, Formulation-1, that does not require any parameters appears to solve the SBS problem the fastest. Further, this formulation appears to have the smallest variance in the solve times. Formulation-l_∞ follows Formulation-1 in the ranking for the fastest solve times although the spread of solve times is slightly larger. Formulation-

l_∞ appears to take around 3 seconds more to solve the SBS problem. These results are surprising as the feasible range of the continuous variables has been reduced. The solve times for Formulation-l_∞-l_1 are much higher than the other two formulations. This is likely to be caused by a greater number of variables in the optimisation problem. When $k = 10$, there appears to be no significant improvement by supplying the solver with a warmstart solution.

5.7 Conclusion

In this chapter we have determined that it is possible to reduce the time to solve the SBS problem using MIQO programs with data driven parameters. Initially we achieved this by obtaining the optimal solution to the SBS problem. This approach to estimating the parameters was of little practical use. In Section 5.4 we developed a DFOA that could produce good solutions to the SBS problem. The DFOA performed well as an approach to simultaneously select predictors when the predictor correlation is low and produced models with predictive performance comparable to models estimated using the optimal solution to the SBS problem. In Section 5.5 we proposed two methods for estimating the parameters for the MIQO models given in Section 5.3, given good solutions to the SBS problem. When estimating the parameters using a solution obtained from the DFOA we found that the time to solve the SBS was not necessarily reduced.

5.A Additional results for the performance of the DFOA

In this appendix we provide additional results to Section 5.6. These tables show the results for the models estimated using 50 random initialisations of the DFOA which provide feasible solutions to the SBS problem.

Table 5.A.1: The performance of the models estimated using the DFOA and optimal solutions to the SBS problem. The data is generated from the *Adjacent* model with $\rho = 0.95$.

	$k = 1$		$\mathbf{k = 3}$		$k = 7$	
	SBS	DFOA	SBS	DFOA	SBS	DFOA
Total Runtime	0.07	0.04	0.10	0.10	5.51	0.12
MSE Estimation	0.03	0.03	0.00	0.02	0.00	0.00
MSE Prediction	0.30	0.30	0.25	0.27	0.26	0.26
# True Predictors	1	1	3	1.56	3	2.31

Table 5.A.2: The performance of the models estimated using the DFOA and optimal solutions to the SBS problem. The data is generated from the *Adjacent* model with $\rho = 0.25$.

	$k = 1$		$\mathbf{k = 3}$		$k = 7$	
	SBS	DFOA	SBS	DFOA	SBS	DFOA
Total Runtime	0.07	0.54	0.08	2.26	3.47	2.34
MSE Estimation	0.01	0.01	0.00	0.00	0.00	0.00
MSE Prediction	0.67	0.67	0.25	0.25	0.26	0.26
# True Predictors	1	1	3	3	3	3

Table 5.A.3: The performance of the models estimated using the DFOA and optimal solutions to the SBS problem. The data is generated from the *Application* model with $\rho = 0.95$.

	$k = 3$		$\mathbf{k = 5}$		$k = 8$	
	SBS	DFOA	SBS	DFOA	SBS	DFOA
Total Runtime	0.13	0.09	0.21	0.11	3.14	0.12
MSE Estimation	0.00	0.06	0.00	0.03	0.00	0.02
MSE Prediction	0.37	0.56	0.25	0.38	0.26	0.30
# True Predictors	3	1.08	4.88	2.08	4.84	2.92

Table 5.A.4: The performance of the models estimated using the DFOA and optimal solutions to the SBS problem. The data is generated from the *Uniformly-spaced* model with $\rho = 0.25$.

	$k = 2$		$\mathbf{k = 5}$		$k = 25$	
	SBS	DFOA	SBS	DFOA	SBS	DFOA
Total Runtime	0.07	1.04	0.07	2.05	38.39	2.51
MSE Estimation	0.09	0.09	0.00	0.00	0.00	0.00
MSE Prediction	3.27	3.27	0.25	0.25	0.27	0.26
# True Predictors	2	2	5	5	5	5

Chapter 6

Fast simultaneous predictor algorithms

In Chapter 5 we found that good predictive performance can be achieved by estimating systems of linear regression models using good feasible solutions to the SBS problem where these solutions were obtained quickly from a DFOA. In this chapter we consider simultaneous predictor selection approaches that can be implemented much faster than the SBS approach and show that these approaches perform well in practise.

This chapter is organised as follows. In Section 6.1 we describe the simultaneous predictor selection approaches and how to implement them. In Section 6.2 we carry out a simulation study to compare the performance of these approaches. We conclude this chapter in Section 6.3.

6.1 The approaches

The first implementation we consider is a stepwise algorithm, naturally extending the popular stepwise algorithm used for a single linear regression model. The second approach is a hybrid, mixing the best-subset approach with stepwise selection. Finally, we consider adapting the Simultaneous Variable Selection (SVS) method proposed by Turlach et al. (2005). Details of the three approaches follow.

6.1.1 A stepwise approach:

An obvious fast alternative to the best-subset implementation of simultaneous variable selection is a stepwise algorithm. The idea is to iteratively add (or remove) the predictor that most improves

(or worsens) the simultaneous least squares objective. A forward stepwise implementation can be formulated as a MIQO optimisation problem. The advantages of formulating a MIQO program is that the automation constraints that we introduced in Chapter 3 can be added easily, resulting in a fast approach that could be used to automate a modelling procedure. The initial MIQO formulation could be

$$\begin{aligned} \min_{\boldsymbol{\beta},\boldsymbol{\eta}} \left[\sum_{m=1}^{M} \sum_{t=1}^{T} \left(y_{t,m} - \sum_{p=1}^{P} x_{t,p,m}\beta_{p,m} \right)^2 \right] \quad & \text{subject to,} \\ -\sum_{p=1}^{P} \eta_p \leq k - P, & \\ (\beta_{p,m}, \eta_p) \in \mathcal{SOS}_1, & \\ \beta_{p,m} \in \mathbb{R}, \quad p = 1, \ldots, P, \ m = 1, \ldots, M, & \\ \eta_p \in \{0, 1\}, \quad p = 1, \ldots, P. & \end{aligned} \tag{6.1.1}$$

This formulation is equivalent to Formulation 1 given in Section 5.3 with $k = 1$. When a predictor is selected we can remove the associated binary variable from the formulation that has sparsity $k + 1$. Let the set $\mathcal{S}_k$ denote the selected predictors for the stepwise implementation at sparsity k. Then, a stepwise formulation for sparsity k may be given by

$$\min_{\boldsymbol{\beta},\boldsymbol{\eta}} \sum_{m=1}^{M} \left[\sum_{t=1}^{T} \left(y_{t,m} - \sum_{p=1}^{P} x_{t,p,m}\beta_{p,m} \right)^2 \right] \quad \text{subject to,} \tag{6.1.2a}$$

$$-\sum_{p \in \mathcal{P} \setminus \mathcal{S}_{k-1}} \eta_p \leq 1 - |\mathcal{P} \setminus \mathcal{S}_{k-1}|, \tag{6.1.2b}$$

$$(\beta_{p,m}, \eta_p) \in \mathcal{SOS}_1, \quad p \in \mathcal{P} \setminus \mathcal{S}_{k-1}, \quad m = 1, \ldots, M, \tag{6.1.2c}$$

$$\beta_{p,m} \in \mathbb{R}, \quad p = 1, \ldots, P, \quad m = 1, \ldots, M, \tag{6.1.2d}$$

$$\eta_p \in \{0, 1\}, \quad p \in \mathcal{P} \setminus \mathcal{S}_{k-1}. \tag{6.1.2e}$$

Formulation (6.1.2) has a computational advantage over the SBS formulations. Firstly, at stage k there are only $P - k + 1$ possible combinations of predictors to select. Secondly, the number of integer variables decreases as the sparsity level increases.

It is not entirely necessary to formulate a MIQO problem to implement a stepwise algorithm. Alternatively, a greedy search algorithm that fits all of the models iteratively, and finds the best predictor to add to the models simultaneously could be used. However, as was previously mentioned, the MIQO approach allows us to include our automated constraints introduced in Chapter 3.

6.1.2 A hybrid approach:

By formulating a stepwise approach as a MIQO program, it is easy to see how to create a hybrid between a stepwise selection procedure and the best-subset selection procedure. A forward stepwise approach does not guarantee to find the optimal solution to the SBS problem for any level of sparsity when $k > 1$. This is because all selected predictors must remain in the model as the sparsity of the model increases. However, an approach need not be this restrictive. With increases in computational power and methods for optimisation highlighted by Bertsimas et al. (2016), it is possible to *trade-off* the combinatorial aspect of the problem with a stepwise approach. We propose a hybrid stepwise/best-subset approach that allows r previously selected variables to be replaced. Miller (1984) considered an approach where selected predictors are replaced in turn. This differs from our proposed approach as we consider replacing r predictors and are guaranteed to find the best substitutes. When r is set to zero, we have standard stepwise. When seeking a model with sparsity k and $r = k$ we have the standard best-subset implementation. Any value $0 < r < k$ gives a hybrid approach at which the computational demands should be lower than a best-subset approach, but consequently higher than a stepwise approach. We could formulate such an approach as follows,

$$\begin{aligned}
\min_{\boldsymbol{\beta},\boldsymbol{\eta}} \left[\sum_{m=1}^{M} \sum_{t=1}^{T} \left(y_{t,m} - \sum_{p=1}^{P} x_{t,p,m} \beta_{p,m} \right)^2 \right] \quad & \text{subject to,} \\
-\sum_{p=1}^{P} \eta_p \leq k - P, & \\
-\sum_{p \in \mathcal{S}_{k-1}} \eta_p = k - r - 1 - |\mathcal{S}_{k-1}|, & \qquad (6.1.3a) \\
(\beta_{p,m}, \eta_p) \in \mathcal{SOS}_1, \quad \text{for } p = 1, \ldots, P, \ m = 1, \ldots, M, & \\
\eta_p \in \{0, 1\}, \quad p = 1, \ldots, P. &
\end{aligned}$$

Here, constraint (6.1.3a) ensures that $k - r$ of the previously selected predictors are included in the model. Stepwise approaches have typically been criticised as they do not guarantee the best-subset for any given level of sparsity, see for example Beale (1970b) and Mantel (1970). By allowing at most r previously selected variables to be *exchanged*, we are *more likely* to obtain the best-subset for a given level of sparsity k.

6.1.3 Modified simultaneous variable selection:

Turlach et al. (2005) proposed a Simultaneous Variable Selection (SVS) approach for selecting predictors in multi-response models. Multi-response models have been used by Breiman and Friedman

(1997) and Similä and Tikka (2005) to improve predictive performance in multivariate response regression models. Multi-response models take the form

$$y_{t,m} = \sum_{p=1}^{P} x_{t,p}\beta_{p,m} + \eta_{t,m}, \quad \text{for } m = 1, \ldots, M.$$

Here, one predictor matrix $\boldsymbol{x} \in \mathbb{R}^{T\times P}$ is used to predict values in all M response variables. We assume that a predictor matrix $\boldsymbol{x}_m \in \mathbb{R}^{T\times P}$ is available for each of the M regression models, where $\boldsymbol{x}_m$ can be thought of as a realisation of the P predictors for each of the M models. Following the LASSO method of Tibshirani (1996), Turlach et al. (2005) arrived at the following problem,

$$\min_{\boldsymbol{\beta}} \left[\sum_{t=1}^{T}\sum_{m=1}^{M} \left(y_{t,m} - \sum_{p=1}^{P} x_{t,p}\beta_{p,m} \right)^2 \right] \quad \text{subject to} \quad \sum_{p=1}^{P} \max\{|\beta_{p,1}|, \ldots, |\beta_{p,m}|\} \leq \nu. \tag{6.1.4}$$

Problem (6.1.4) for $M = 1$ gives the LASSO (Tibshirani, 1996) problem in constrained form. Using a convex quadratic optimisation formulation problem (6.1.4) can be written as

$$\min_{\boldsymbol{\beta},\boldsymbol{z}} \left[\sum_{m=1}^{M}\sum_{t=1}^{T} \left(y_{t,m} - \sum_{p=1}^{P} x_{t,p}\beta_{p,m} \right)^2 \right] \quad \text{subject to,} \tag{6.1.5a}$$

$$\boldsymbol{u}_M \otimes \boldsymbol{z} - \boldsymbol{\beta} \geq 0, \tag{6.1.5b}$$

$$\boldsymbol{u}_M \otimes \boldsymbol{z} + \boldsymbol{\beta} \geq 0, \tag{6.1.5c}$$

$$\nu - \boldsymbol{u}_P \boldsymbol{z} \geq 0. \tag{6.1.5d}$$

Here, $\boldsymbol{z} \in \mathbb{R}^P$ and $\boldsymbol{u}_M \in \mathbb{R}^M$ are auxiliary variables and $\boldsymbol{\beta} \in \mathbb{R}^{MP}$. The constraints (6.1.5b) ensure that $\beta_{p,m} \leq z_p$ for $m = 1, \ldots, M$ and constraints (6.1.5c) ensure that $-\beta_{p,m} \leq z_p$ for $m = 1, \ldots, M$. Collectively, (6.1.5b) and (6.1.5c) ensure that $-\beta_{p,m} \leq z_p \leq \beta_{p,m}$ for $m = 1, \ldots, M$ so that with (6.1.5d) we have $\sum_{p=1}^{P} \max\{|\beta_{p,1}|, \ldots, |\beta_{p,m}|\} \leq \nu$. All regression coefficients in a solution to (6.1.5) will have a non-zero value. Turlach et al. (2005) propose the following to select predictors. Let

$$\mathcal{J} = \{p : \max\{\beta_{p,1}, \ldots, \beta_{p,m}\} > \nu 10^{-4} \quad \text{for } p = 1, \ldots, P\}, \tag{6.1.6}$$

then the coefficients $\beta_{m,p}$ for $p \notin \mathcal{J}$ and for $m = 1, \ldots, M$ should be set to zero.

The SVS approach was proposed as an exploratory tool. We propose to modify this approach to determine a suitable subset of predictors and then use the simultaneous least squares objective to estimate the coefficients of the selected variables much like the idea of the relaxed LASSO (Meinshausen, 2007). The convex quadratic program we solve to determine the subsets of predictors

follows,

$$\min_{\boldsymbol{\beta},\boldsymbol{z}} \left[\sum_{m=1}^{M} \sum_{t=1}^{T} \left(y_{t,m} - \sum_{p=1}^{P} x_{t,p,m} \beta_{p,m} \right)^2 \right] \quad \text{subject to,}$$

$$\begin{aligned} \boldsymbol{u}_M \otimes \boldsymbol{z} - \boldsymbol{\beta} &\geq 0, \\ \boldsymbol{u}_M \otimes \boldsymbol{z} + \boldsymbol{\beta} &\geq 0, \\ \nu - \boldsymbol{u}_P \boldsymbol{z} &\geq 0. \end{aligned} \tag{6.1.7}$$

Here, the objective is modified to use the simultaneous least squares objective. We use the same heuristic as Turlach et al. (2005) for selecting the non-zero coefficients. A suitable range of values for ν can be determined easily. For $\nu > \sum_{m=1}^{M} \sum_{p=1}^{P} \beta^*_{p,m}$ the solution to (6.1.7) will be $\boldsymbol{\beta}^*$, the simultaneous least squares estimate. So solving (6.1.7) for a range of $\nu \in (0, \sum_{m=1}^{M} \sum_{p=1}^{P} \beta^*_{p,m})$ will help us produce a range of suitable subsets. Let $\mathcal{J}$ denote the selected predictors using some value of the tuning parameter ν, then the coefficients are estimated as

$$\hat{\boldsymbol{\beta}} = \arg\min \left[\sum_{m=1}^{M} \sum_{t=1}^{T} \left(y_{t,m} - \sum_{p=1}^{P} x_{t,p,m} \beta_{p,m} \right)^2 \right] \quad \text{subject to,}$$

$$\beta_{j,m} = 0, \quad \text{for } j \notin \mathcal{J}, \text{ for } m = 1, \ldots, M.$$

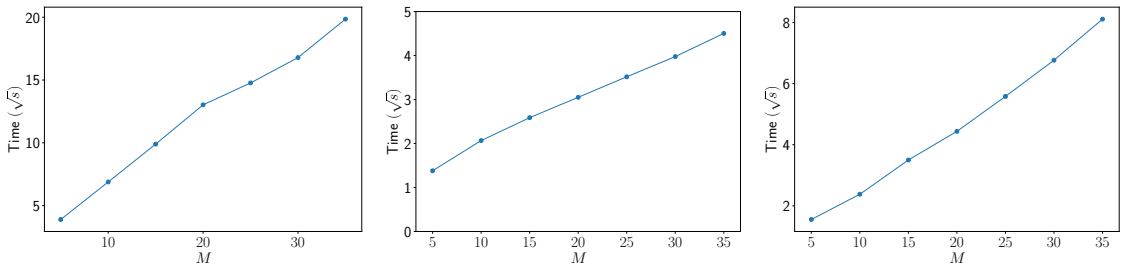
We could also consider applying the simultaneous shrinkage operator here.

Now that we have introduced each of the simultaneous predictor selection approaches we will study how they perform using a simulation study. We will investigate the total time needed to implement each approach and how the models estimated using each approach perform, when compared to models estimated using the SBS approach.

6.2 Simulation study

Firstly, we will consider how each approach scales with M and P. The Stepwise approach produces at most P models, a model is produced at each stage of the Stepwise approach. We will consider the time to implement the Hybrid approach when $r = 1$ which we denote *Hybrid-1*. The Hybrid approach also produces at most P models, one for each level of sparsity, $k = 1, \ldots, P$. The number of models produced by the SVS approach is not easy to determine a priori. This is because it depends on the size of the coefficients in a solution to the convex quadratic problem given in (6.1.7) and the heuristic given in (6.1.6). We will implement the modified SVS approach using 100 values of ν. We have found that the number of predictors selected using the heuristic changes more frequently for small values of ν. In order to obtain the largest number of unique subsets of selected predictors for the modified SVS approach we space ν on a logarithmic scale.

We generate 100 datasets from the *Scaling* model given in Section 5.1, and present the average time taken to implement each approach. To investigate how the approaches scale with P we fix $M = 5$ and to investigate how they scale with M we fix $P = 35$.

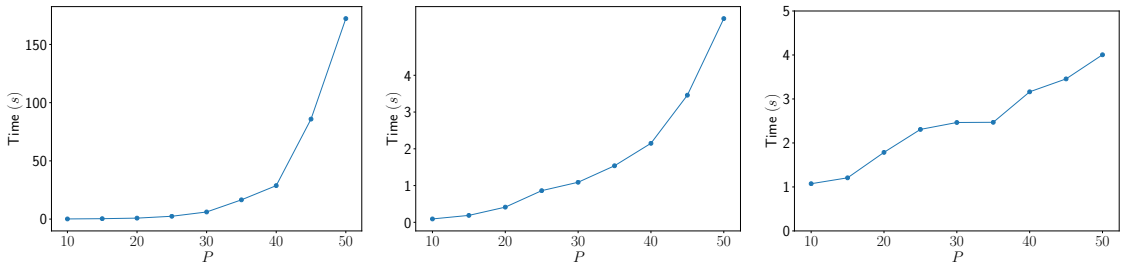


(a) Hybrid-1 (b) Stepwise (c) SVS

Figure 6.2.1: Scaling of the Hybrid, Stepwise, and SVS simultaneous predictor selection approaches with M, the number of response variables

Figure 6.2.1 shows how the approaches scale with M. All approaches appear to scale quadratically. The Hybrid-1 approach takes on average 400 seconds to solve problems with $M = 35$ and $P = 35$. In Section 3.2.1 we observed that with $M = 5$ and $P = 35$ the SBS approach took around 400 seconds to solve just one SBS problem with $k = \frac{35}{2}$. Here, 35 MIQO programs have been solved with 30 more response variables. The Stepwise approach is considerably faster, taking only 20 seconds to produce models for each level of sparsity. With $M = 5$ and $P = 35$ the SVS approach takes around one minute to solve all 100 problems given each value of the tuning parameter.

The SBS approach scaled poorly with P. Figure 6.2.2 shows how the approaches scale with P. The Hybrid approach appears to scale exponentially with P, but we were able to obtain the models for all levels of sparsity in under 3 minutes with $P = 50$. This can be seen in Figure 6.2.2a. The Stepwise method appears to scale quadratically with P, (see Figure 6.2.2b) and the SVS approach appears to scale approximately linearly with P (see Figure 6.2.2c).



(a) Hybrid 1 (b) Step (c) SVS

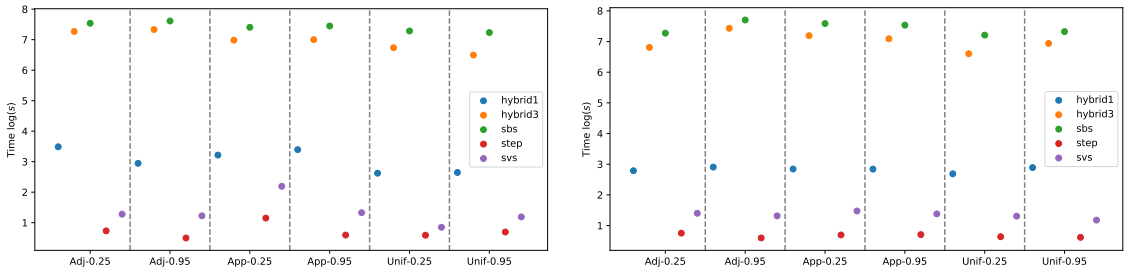
Figure 6.2.2: Scaling of the Hybrid, Stepwise, and SVS simultaneous predictor selection approaches with P the number of predictors variables.

The simultaneous predictor selection approaches given in Section 6.1 can be implemented much faster than the SBS approach. It is now important for us to determine if the quality of the regression models estimated using these approaches is as good as the models estimated using the SBS approach. We simulate 25 datasets from the *Adjacent*, *Application* and *Uniformly-spaced* models defined in Section 5.1. The results for each dataset will be provided using the abbreviated names, *Adj*, *App* and *Unif* respectively. The value of ρ used follows the abbreviated names. Each dataset will consist of 1000 observations for each response variable and we split the observations randomly into training/test/validation sets to the ratios 50%/25%/25% respectively. The following applies for each approach. We apply the approach to the training data, using each value of the associated tuning parameters. For each value of the tuning parameter we calculate the mean-squared prediction error of the associated system of linear regression models using the test data. Then, we select the model associated to the tuning parameter that gives the lowest prediction error on the validation data.

We compare the selected models for each approach using a number of criteria. These include the mean-squared prediction error of the system on the validation data, the mean-squared estimation error of the system, and the average sparsity of each of the models. We will also provide the time to implement each approach as a comparison. Each of these criteria is now discussed in turn.

6.2.1 Average time to implement each approach:

The motivation for developing alternatives to the SBS approach was to obtain more computationally efficient methods for simultaneously selecting predictors. Figure 6.2.3 compares the natural logarithm of the time in seconds to implement each approach. We can see that the Stepwise approach is consistently the fastest, followed by SVS, Hybrid-1, Hybrid-3 and then the SBS approaches.



(a) $\sigma_{\eta_m} = 0.5$ for $m = 1, \ldots, 5$ (b) $\sigma_{\eta_m} = 2$ for $m = 1, \ldots, 5$

Figure 6.2.3: The average time to implement each simultaneous predictor selection method. Each group of five points shows the average time to implement each method on a log scale, for of the synthetic data models, and for each of the five simultaneous predictor selection methods

It is not easy to determine the actual times to implement each approach by eye in Figure 6.2.3.

Therefore, we will now consider the average time taken to implement each approach for the implementation times shown in Figure 6.2.3. On average, both the Stepwise and SVS approaches were implemented in under four seconds. The Hybrid-1 approach took less then 20 seconds and both the Hybrid-3 and SBS approaches exceeded 1100 seconds. The value of r, that determines how many previously selected predictors in the Hybrid can be substituted, plays a significant role in the total time to implement the Hybrid approach. In the examples presented in Figure 6.2.3, we observe in excess of a 55 times factor speedup from the Hybrid-3 approach to the Hybrid-1 approach.

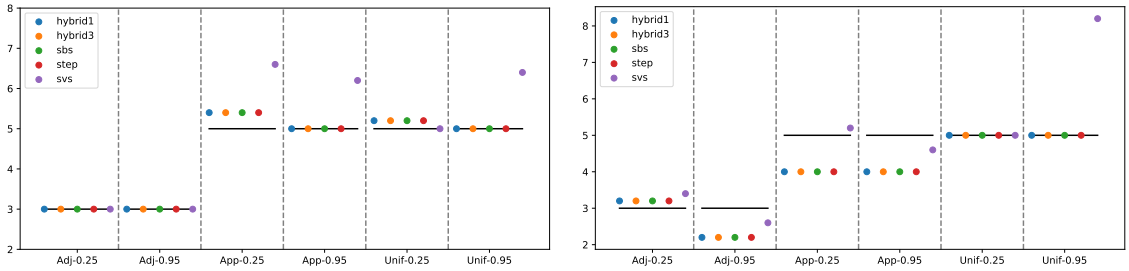
6.2.2 Average model sparsity:

We calculate the average model sparsity for the selected models for each approach. We define the estimated model sparsity as

$$\hat{k} = \frac{1}{5}\sum_{m=1}^{5}\sum_{p=1}^{35}\mathbb{I}_{\hat{\beta}_{p,m}\neq 0}.$$

Here, indicator $\mathbb{I}_{\hat{\beta}_{p,m}}$ takes the value 1 if the estimated coefficient $\hat{\beta}_{p,m}$ is not equal to zero. The estimated model sparsity is then averaged over the 25 simulations. Figure 6.2.4 shows the average model sparsity for each of the approaches. The black horizontal lines indicate the true model sparsity.

The models selected for the SBS, Stepwise, Hybrid-1 and Hybrid-3 approaches were identical. In most simulations, the models selected for the SVS approach typically included slightly more predictors on average than all other approaches. With the exception of *Uniformly spaced* model, with $\rho = 0.95$ and $\sigma^2_{\eta_m} = 2$, for $m = 1, \ldots, 5$ the average sparsity of the model selected for the SVS approach typically contained only one more predictor than the other approaches.



(a) $\sigma_{\eta_m} = 0.5$ for $m = 1, \ldots, 5$ (b) $\sigma_{\eta_m} = 2$ for $m = 1, \ldots, 5$

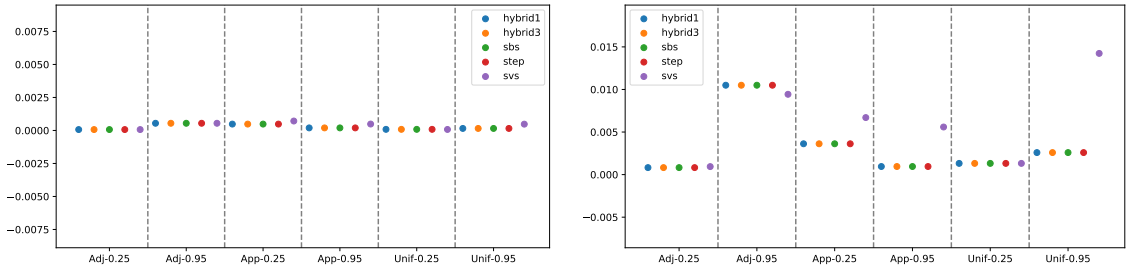
Figure 6.2.4: The average sparsity of the models fit by the simultaneous predictor selection approaches.

6.2.3 Mean-squared estimation error:

For each system of regression models estimated using the simultaneous predictor selection approaches we calculate the mean-squared estimation error of the system as

$$\frac{1}{35 \times 5} \sum_{m=1}^{5} \sum_{p=1}^{35} \left(\beta_{p,m} - \hat{\beta}_{p,m} \right)^2 .$$

Here, $\beta_{p,m}$ is the true value of the coefficients given in Section 5.1, and $\hat{\beta}_{p,m}$ is the corresponding estimate. The mean-squared estimation errors are shown in Figure 6.2.5. The mean-squared estimation error of the SVS approach is typically higher than all other approaches. We have seen that the average sparsity of the model for the SVS approach was higher than all other approaches. The increased estimation error of the SVS approach may be caused by non-zero coefficient estimates that should be zero. To determine if this it the case we need to determine how often the approaches identified the correct subset of predictors.



(a) $\sigma_\eta = 0.5$ (b) $\sigma_\eta = 2$

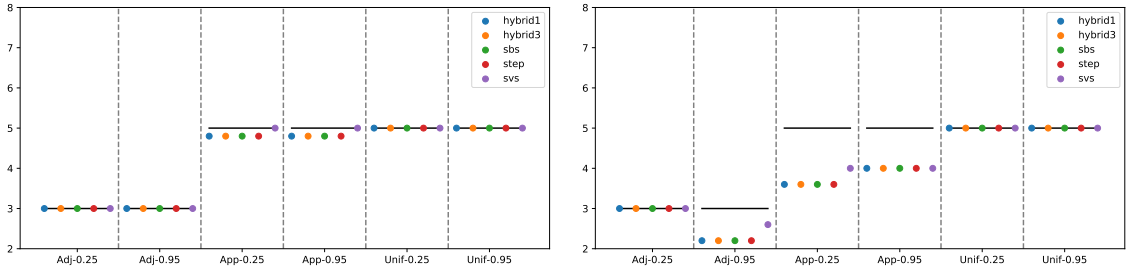
Figure 6.2.5: The average mean-squared estimation error of the system for each simultaneous predictor selection approach.

6.2.4 Average number of correctly identified predictors:

The average number of correctly identified predictors, for each approach, is shown in Figure 6.2.6. The true model sparsity is again shown by the black horizontal line. Despite the SVS approach producing models with more predictors than the other approaches, it appears that the predictors selected by the SVS approach often contained the true predictors. Here, the results for the SVS approach appear more favourable as the models estimated using the SVS approach contained the true predictors more often.

6.2.5 Mean-squared error in prediction:

Finally, we consider how the models estimated using each approach compare in predicting values for the validation dataset. For each of the selected models we calculate the mean-squared prediction



(a) $\sigma_\eta = 0.5$ (b) $\sigma_\eta = 2$

Figure 6.2.6: The average number of correctly identified predictors for each of the simultaneous predictor selection approaches.

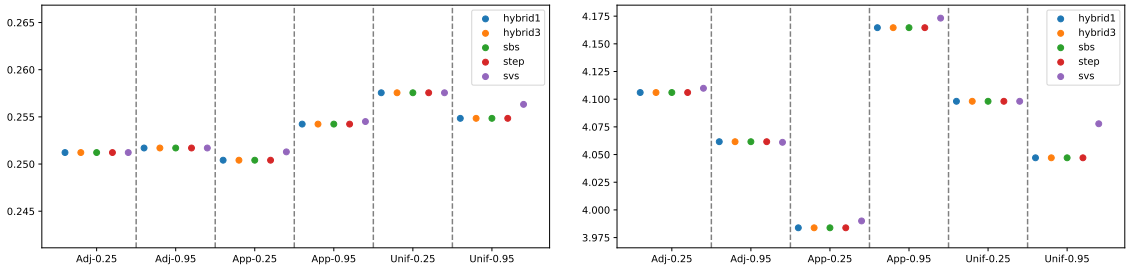
error of the system as

$$\frac{1}{5 \times 250} \sum_{m=1}^{5} \sum_{t=1}^{250} \left(y_{t,m}^{\text{validation}} - \hat{y}_{m,t}^{\text{validation}} \right)^2 .$$

Here, $y_{t,m}^{\text{validation}}$ is the t^{th} observation of the m^{th} response variable from the validation dataset and $\hat{y}_{m,t}^{\text{validation}}$ is the associated fitted value. The mean-squared prediction error averaged over the 25 simulations is shown in Figure 6.3.1. We can see that the average prediction error for the SVS approach is at least that of all other approaches. This could again be explained by the inclusion of noisy predictors.

6.3 Conclusion

In this chapter we have proposed simultaneous predictor selection approaches that can be implemented in significantly less time than the SBS approach. Whilst these alternative approaches can be seen to give approximate solutions to the SBS problem, these approaches perform well in practise. In addition to this, our Hybrid approach is capable of trading-off the combinatorial challenges of obtaining the optimal solution to the SBS problem with fast runtimes. This is achieved by a parameter r which allows at most r of the predictors to be replaced as the algorithm progresses. In our simulation studies we found that the Stepwise and Hybrid methods produced the same solution as the SBS approach, despite taken significantly less time to produce the solution.



(a) $\sigma_\eta = 0.5$ (b) $\sigma_\eta = 2$

Figure 6.3.1: The average mean-squared prediction error for each of the simultaneous predictor selection approaches

Chapter 7

Simultaneous shrinkage study

In Chapter 3 we introduced a simultaneous shrinkage operator that could improve predictor selection accuracy and significantly improve model estimation accuracy. In this chapter we apply the SBS problem with simultaneous shrinkage to the data generating models defined in Section 5.1 to better understand the behaviour of this operator under different generating processes. In addition to this, we consider how the shrinkage operator performs under different model sparsity's.

7.1 Introduction

Recall that the SBS problem with simultaneous shrinkage is defined as

$$\min \left[\sum_{m=1}^{M} \sum_{t=1}^{T} \left(y_{t,m} - \sum_{p=1}^{P} x_{t,p,m} \beta_{p,m} \right)^2 + \lambda \sum_{m=1}^{M} \sum_{p=1}^{P} \left(\beta_{p,m} - \bar{\beta}_{p,m} \right)^2 \right] \quad \text{subject to,} \tag{7.1.1}$$

$$\left| \bigcup_{m=1}^{M} \mathcal{S}_m \right| \leq k. \tag{7.1.2}$$

Here, $\mathcal{S}_m = \{p : \beta_{p,m} \neq 0 \text{ for } p = 1, \ldots, P\}$ denotes the predictors selected in model m. In Section 3.4.1 we applied simultaneous shrinkage to the *Adjacent* model, given in Section 5.1 when $k = 3$, the true level of sparsity. We also found that the true subset of predictors was not initially selected in the solution to the SBS problem but as λ increased, the subset of predictors selected changed to the true subset. We did not consider the effect of the operator on the SBS solution when k was greater than, or less than, the true level of sparsity.

In the following sections we select a subset of the results and summarise the effect of the shrinkage operator on the SBS solution when it is applied when the level of sparsity, k, is greater than, equal to, and less than the true model sparsity. We simulate 750 observations from each model and randomly allocate 500 to a training dataset and use the remaining observations for validation dataset. Each

time we solve the SBS problem with shrinkage we observe the trace-plot of the SBS solution as the penalty λ increases. We also observe the effect of increasing λ on the mean-squared estimation error of the system and the mean-squared prediction error of the system on the validation data. We define the mean-squared estimation estimation error of the system as

$$\text{MSE}_e = \sum_{m=1}^{M} \sum_{p=1}^{P} \left(\beta_{p,m} - \hat{\beta}_{p,m} \right)^2 .$$

Here, $\beta_{p,m}$ for $p = 1, \ldots, P$ and $m = 1, \ldots, M$ denote the true value of the regression coefficients and $\hat{\beta}_{p,m}$ for $p = 1, \ldots, P$ and $m = 1, \ldots, M$ denote the estimates obtained from a solution of the SBS problem with shrinkage. The mean-squared prediction error of the system is given by

$$\text{MSE}_p = \sum_{m=1}^{M} \sum_{t=1}^{T} \left(y_{t,m} - \hat{y}_{t,m} \right)^2 .$$

Here, $y_{t,m}$ denotes an observation from the validation dataset and $\hat{y}_{t,m}$ denotes the fitted value obtained from a model estimated by solving the SBS problem with shrinkage. We now discuss the performance of the simultaneous shrinkage operator for the three cases of sparsity in turn.

7.2 True level of sparsity

In Chapter 3, we applied the simultaneous shrinkage operator to data generated from the Adjacent model with $\rho = 0.95$. We observed that as λ increases, the regression coefficients can be pushed towards the true values. Figure 7.2.1 shows the effect of increasing the simultaneous shrinkage penalty on the estimates of the regression coefficients for data generated from the Application model where $\rho = 0.25$. Here, $k = 5$ the true sparsity of the model. We can see that many of the regression coefficients are pushed closer towards the true values with the largest changes in the coefficients observed for small values of λ. However, some regression coefficients are pushed away from the true values. This occurs for coefficients $\beta_{31,3}$, and $\beta_{31,4}$ for example.

Figure 7.2.2a shows that the mean-squared estimation error is initially improved for all response variables. The largest gains in estimation error are observed for Response 1. Improving the estimation error appears to have a subsequent improvement in prediction error. Again, the most significant improvements appear to be for Response 1. We observe slight improvements in prediction error for response variables, 2, 4, and 5. However, the prediction accuracy for Response 3 is reduced slightly. The prediction accuracy averaged across all response variables does increase as the shrinkage penalty increases.

Appendix 5.A shows the results for the application of the shrinkage operator to all other datasets. At the true level of sparsity the shrinkage operator does typically improve the mean-squared error in both estimation and prediction.

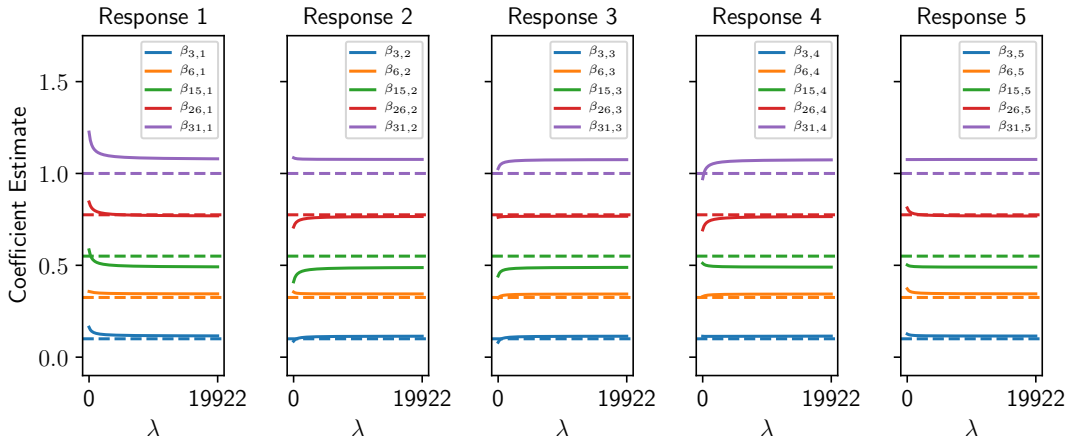
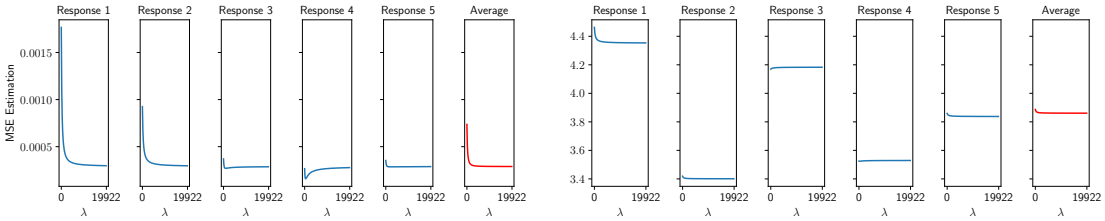


Figure 7.2.1: Trace-plot of the regression coefficients for each response variable as λ increases. The data is generated from the Application model where $\rho = 0.25$ and $\text{var}(\eta_m) = 2$ for $m = 1, \ldots, 5$.



(a) MSE Estimation

(b) MSE Prediction

Figure 7.2.2: The effect of the simultaneous shrinkage operator on the mean-squared estimation and prediction error. The data is generated from the Application model, where $\rho = 0.25$ and $\text{var}(\eta_m) = 2$ for $m = 1, \ldots, 5$.

7.3 Noisy models

Here we shall investigate the effect of the simultaneous shrinkage operator when the value of k is set higher than the true level of model sparsity. In practice, the true level of model sparsity is unknown so it is of interest to us to observe how the operator behaves in general.

Figure 7.3.1 shows the trace of regression coefficient estimates for each response variable as the simultaneous shrinkage penalty increases. The data is generated from the Uniformly spaced model with $\rho = 0.95$. Here, each non-zero coefficient assumes the value one. The coefficients corresponding to predictors 1,8,15,22 and 29 are non-zero. We observe that the estimates of the non-zero coefficients are around one and that they appear to improve slightly with shrinkage. An interesting observation is how the shrinkage operator affects the estimates of the coefficients that should be zero. The sparsity level $k = 25$ indicates that up to 20 additional regression coefficients are allowed to take non-zero values in a solution provided from the mixed-integer quadratic optimisation problem. Here, it appears that as the value of the shrinkage penalty increases, the values of the coefficients for many

of the predictors not present in the true model are pushed towards zero.

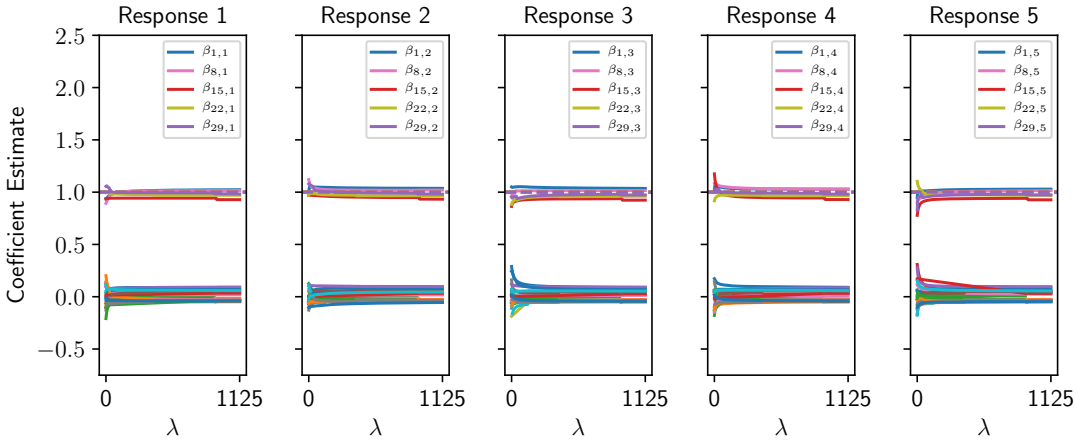
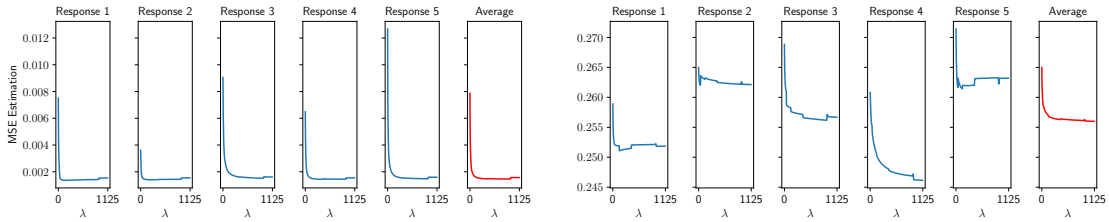


Figure 7.3.1: Trace of the regression coefficients for each response variable as λ increases. The data is generated from the Uniformly spaced model with $\rho = 0.95$ and $\text{var}(\eta_m) = 0.5$, for $m = 1, \ldots, 5$. Due to space constraints, the legend shows only the coefficients that are non-zero.

The solution to the SBS problem with $k = 25$ includes many more predictors into a model when compared to the true model. Erroneously estimating coefficients that should be zero as non-zero affects the mean-squared error in estimating the regression coefficients. Shrinking the regression coefficients towards a common value shows that the error in estimation can be dramatically reduced. This is shown in Figure 7.3.2a. The simultaneous shrinkage operator appears to push many of the coefficients that should take zero values, towards zero, having a great impact on the overall estimation accuracy of the simultaneous best-subset method. Estimation accuracy appears to be improved by over 80% in comparison to the solution provided by the SBS approach without shrinkage. As a consequence of improved estimation, we observe an improvement in the prediction error. Figure 7.3.2b shows that the prediction error for each response variable improves significantly.



(a) MSE Estimation (b) MSE Prediction

Figure 7.3.2: The effect of the simultaneous shrinkage operator on the mean-squared estimation and prediction errors. The data is generated from the Uniformly spaced model with $\rho = 0.95$ and $\text{var}(\eta_m) = 0.5$, for $m = 1, \ldots, 5$.

When the simultaneous shrinkage operator is added to the objective of the SBS problem and when

noisy predictors are present in the model we typically observe an improvement in both estimation error and prediction error. We believe that this may be due to forcing many of the coefficients of the noisy variables towards zero. This may be caused by the coefficients of a given noisy predictor taking a mixture of values above and below zero in each of the regression models. As the penalty in the simultaneous shrinkage operator increases the coefficients are pushed closer to a common value, which may be close to zero.

7.4 Sparse models

Finally, we discuss the effect of the simultaneous shrinkage operator on sparse models. Here, we set the sparsity, k to a value less than the true model sparsity. Figure 7.4.1 shows the trace-plot of the regression coefficients as the simultaneous shrinkage penalty increases for the Adjacent model with $\rho = 0.25$. We observe a slight movement in the regression coefficient estimates as the penalty increases.

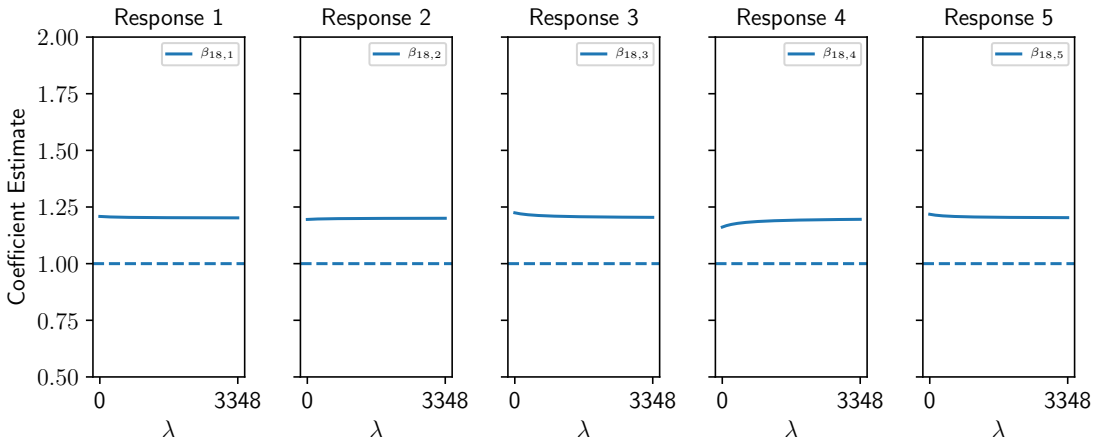
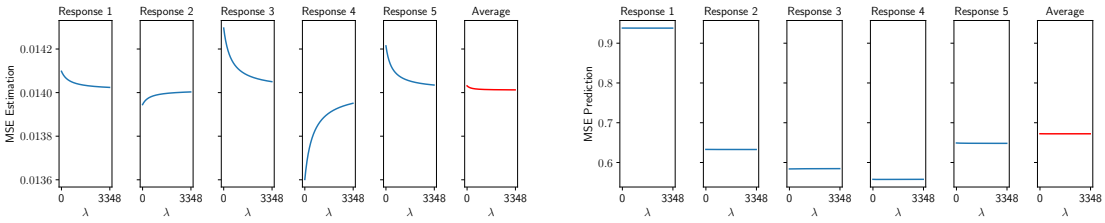


Figure 7.4.1: Trace of the regression coefficients for each response variable as λ increases. The data is generated from the Adjacent model with $\rho = 0.25$ and $\text{var}(\eta_m) = 0.5$, for $m = 1, \ldots, 5$.

As a consequence of slight changes in the regression coefficient estimates, the mean-squared estimation error changes very slightly. This is illustrated in Figure 7.4.2. The average mean-squared prediction error does not appear to change, but we can confirm it does decrease (see Figure 7.4.2b).

When applying the simultaneous shrinkage operator to sparse models, the effect on the regression coefficients is harder to generalise. Appendix 5.A shows that reasonably large changes can be observed in the regression coefficients estimates. This can have a large impact on the mean-squared estimation error, in contrast to our observations in this section. However, the gain in prediction accuracy is typically small.



(a) MSE Estimation (b) MSE Prediction

Figure 7.4.2: The effect of the simultaneous shrinkage operator on the mean-squared estimation and prediction error. The data is generated from the Adjacent model with $\rho = 0.25$ and $\text{var}(\eta_m) = 0.5$, for $m = 1, \ldots, 5$.

7.5 Conclusion

In this chapter we have observed that the simultaneous shrinkage operator can improve regression coefficient estimation for a system of linear regression models. When the sparsity of the SBS problem is set at least that of the true model, we typically find that both estimation and prediction error improves as the level of shrinkage increases. These effects are even stronger when k is much greater than the true level of sparsity. This is a consequence of many of the noisy coefficient estimates being driven towards zero.

7.A Additional results for the SBS problem with simultaneous shrinkage

In this appendix we present the remainder of the results for Chapter 7. Here, we use the solutions of the SBS problem with simultaneous shrinkage to estimate the system of linear regression models for the models defined in Section 5.1 and for three levels of k.

The results presented here are for the *Adjacent* model with $\rho = 0.95$, and $\sigma^2_{\eta_m} = 0.5$. Here, the level of sparsity k, is less than the true sparsity of the model.

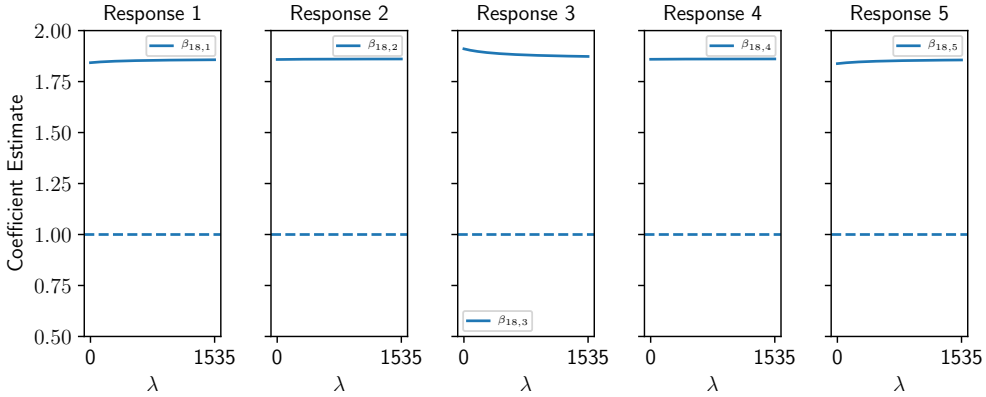


Figure 7.A.1: Regression coefficient trace-plots for an increasing shrinkage penalty.

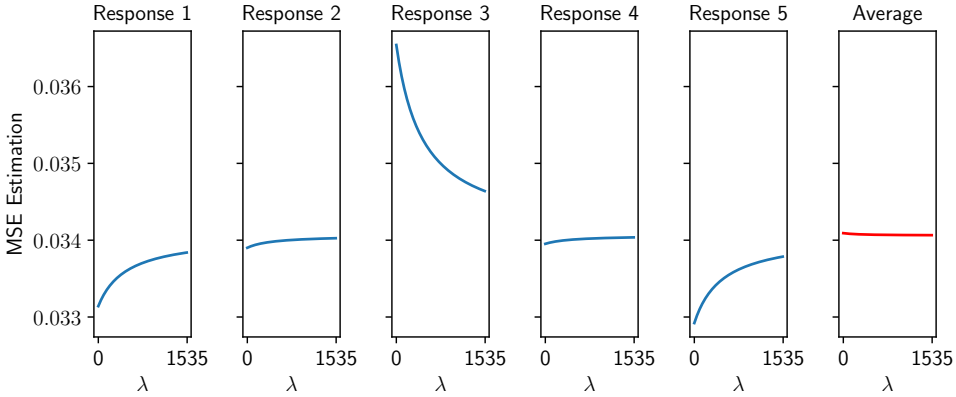


Figure 7.A.2: Mean-squared estimation error for each model. The mean-squared estimation error for the system is given in red.

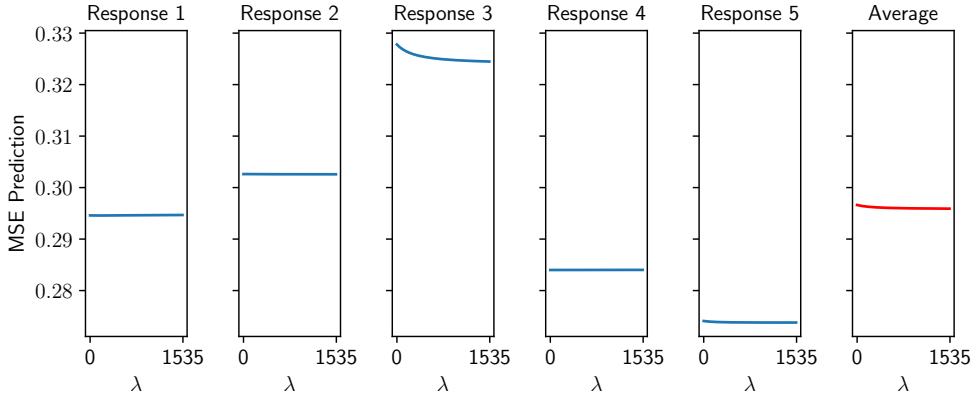


Figure 7.A.3: Mean-squared prediction error for each response variable in the hold-out dataset. The mean-squared prediction error for the system is given in red.

The results here are for the *Adjacent* model with $\rho = 0.95$, and $\sigma^2_{\eta_m} = 0.5$. Here, the level of sparsity k, is equal to the true sparsity of the model.

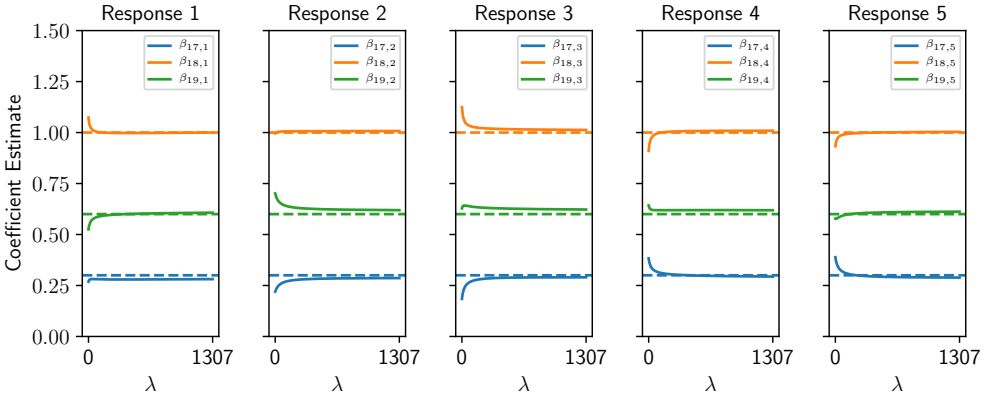


Figure 7.A.4: Regression coefficient trace-plots for an increasing shrinkage penalty.

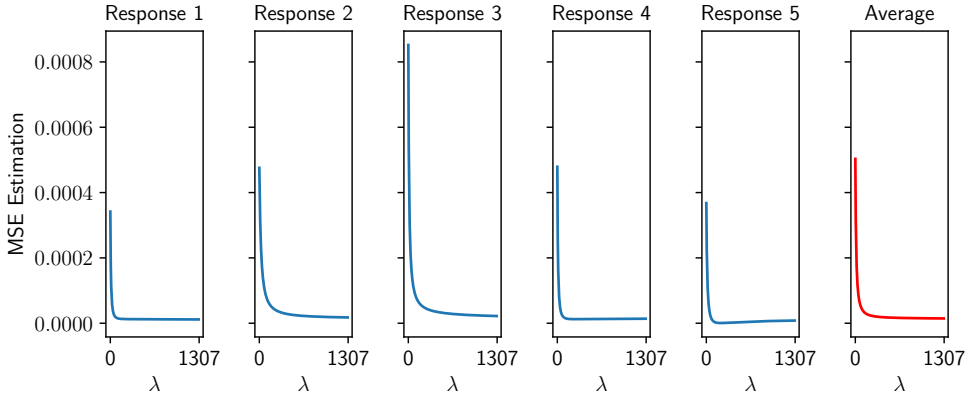


Figure 7.A.5: Mean-squared estimation error for each model. The mean-squared estimation error for the system is given in red.

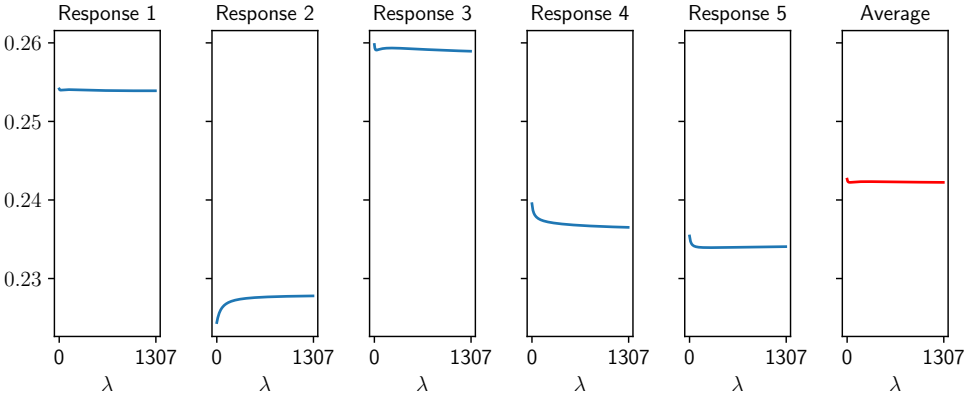


Figure 7.A.6: Mean-squared prediction error for each response variable in the hold-out dataset. The mean-squared prediction error for the system is given in red.

The results here are for the *Adjacent* model with $\rho = 0.95$, and $\sigma^2_{\eta_m} = 0.5$. Here, the level of sparsity k, is greater than the true sparsity of the model.

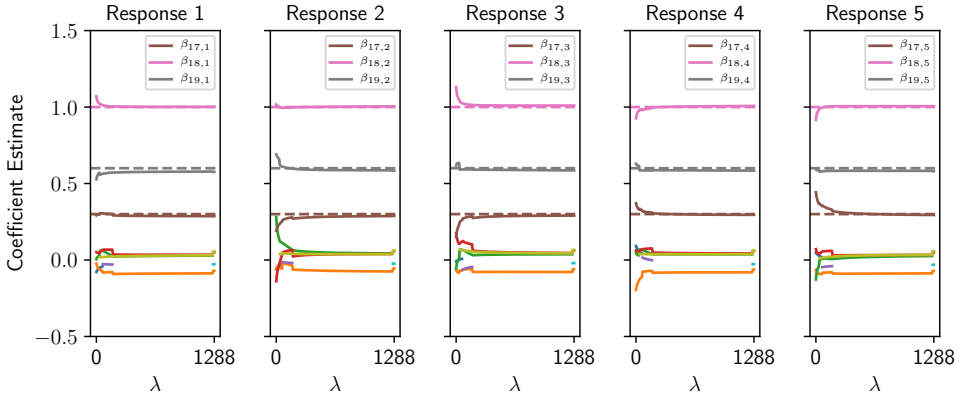


Figure 7.A.7: Regression coefficient trace-plots for an increasing shrinkage penalty.

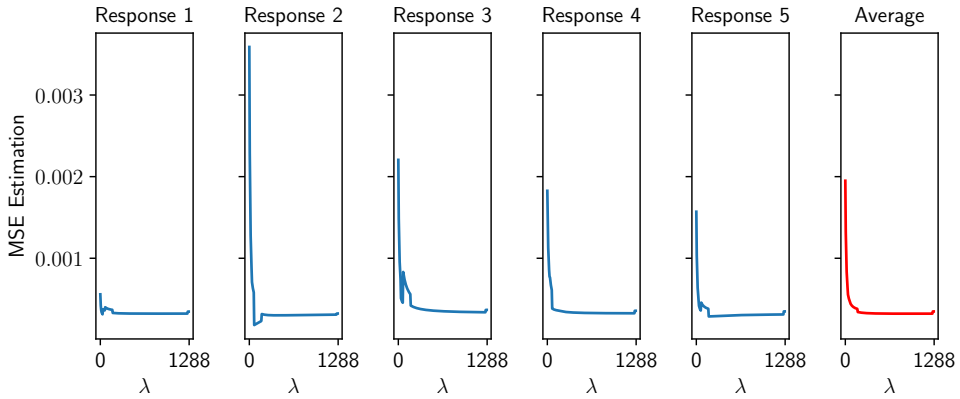


Figure 7.A.8: Mean-squared estimation error for each model. The mean-squared estimation error for the system is given in red.

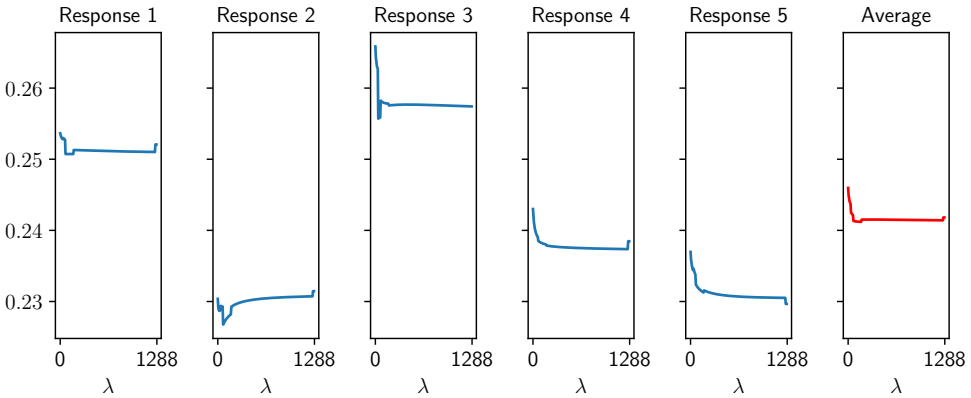


Figure 7.A.9: Mean-squared prediction error for each response variable in the hold-out dataset. The mean-squared prediction error for the system is given in red.

The results presented here are for the *Adjacent* model with $\rho = 0.95$, and $\sigma^2_{\eta_m} = 2$. Here, the level of sparsity k, is less than the true sparsity of the model.

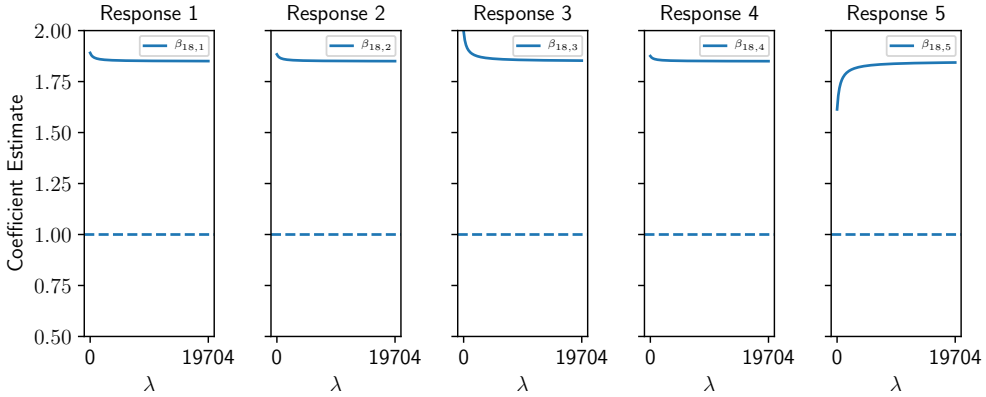


Figure 7.A.10: Regression coefficient trace-plots for an increasing shrinkage penalty.

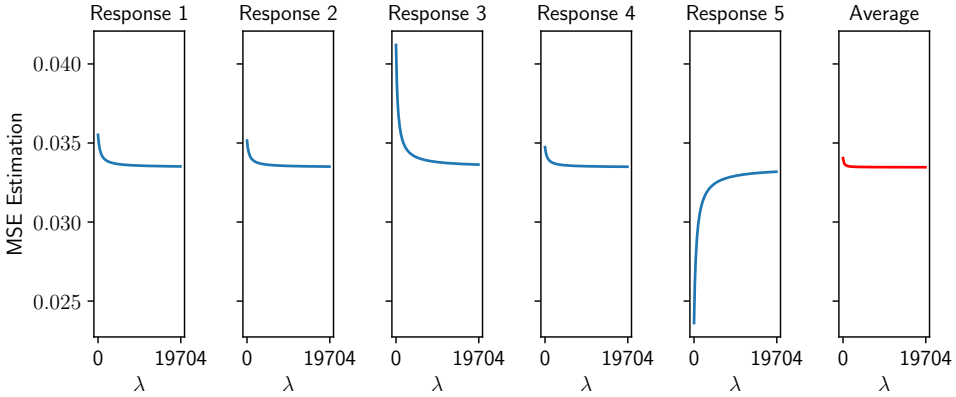


Figure 7.A.11: Mean-squared estimation error for each model. The mean-squared estimation error for the system is given in red.

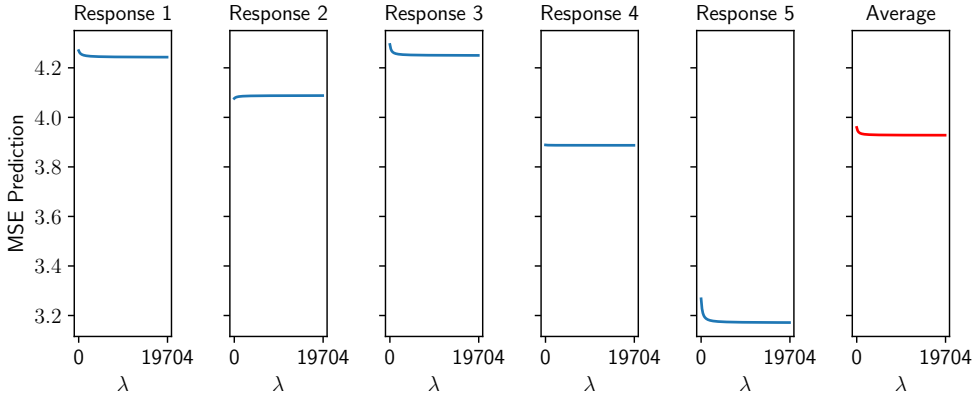


Figure 7.A.12: Mean-squared prediction error for each response variable in the hold-out dataset. The mean-squared prediction error for the system is given in red.

The results here are for the *Adjacent* model with $\rho = 0.95$, and $\sigma^2_{\eta_m} = 2$. Here, the level of sparsity k, is equal to the true sparsity of the model.

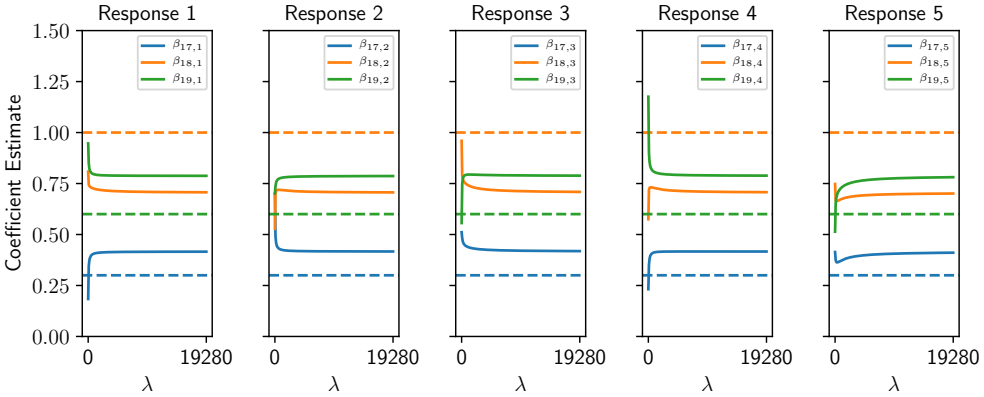


Figure 7.A.13: Regression coefficient trace-plots for an increasing shrinkage penalty.

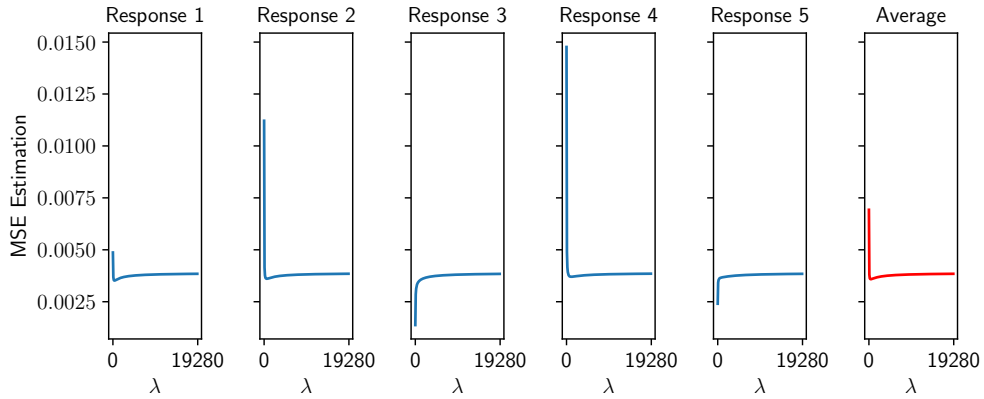


Figure 7.A.14: Mean-squared estimation error for each model. The mean-squared estimation error for the system is given in red.

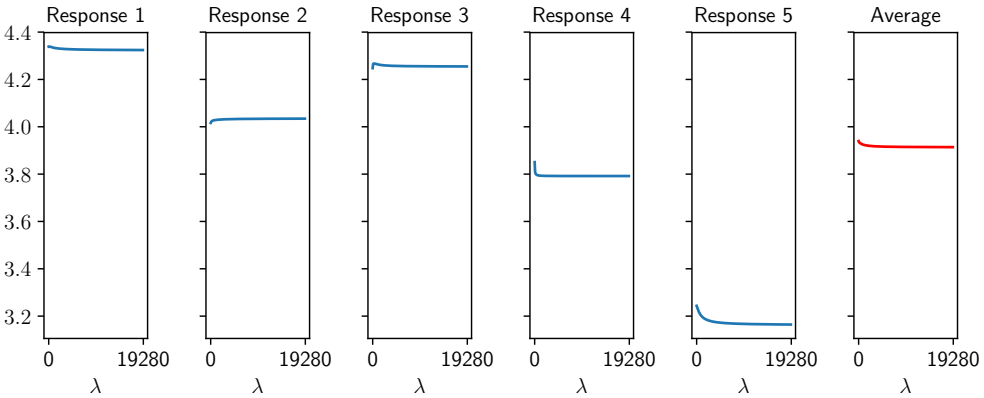


Figure 7.A.15: Mean-squared prediction error for each response variable in the hold-out dataset. The mean-squared prediction error for the system is given in red.

The results here are for the *Adjacent* model with $\rho = 0.95$, and $\sigma^2_{\eta_m} = 2$. Here, the level of sparsity k, is greater than the true sparsity of the model.

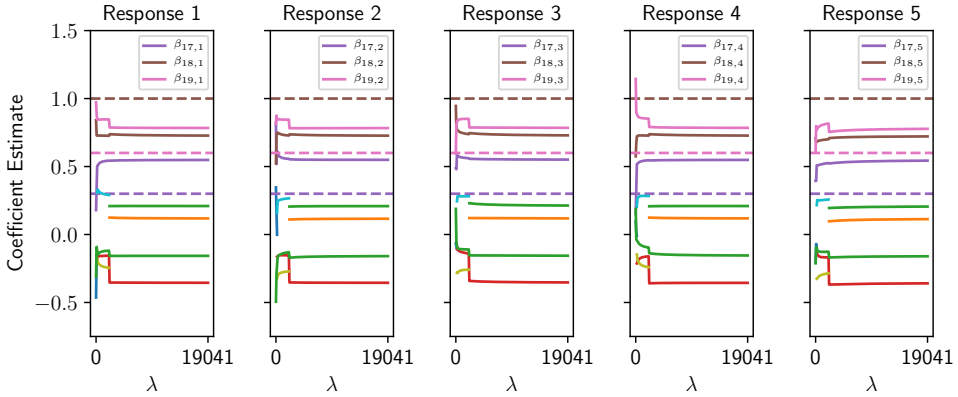


Figure 7.A.16: Regression coefficient trace-plots for an increasing shrinkage penalty.

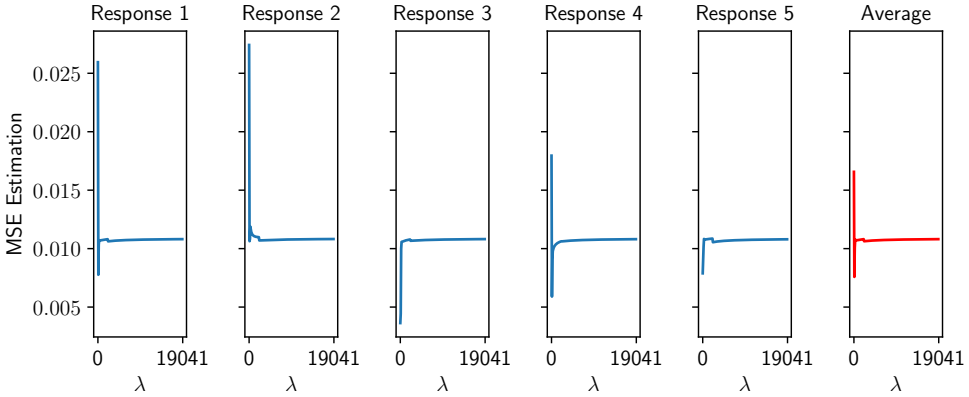


Figure 7.A.17: Mean-squared estimation error for each model. The mean-squared estimation error for the system is given in red.

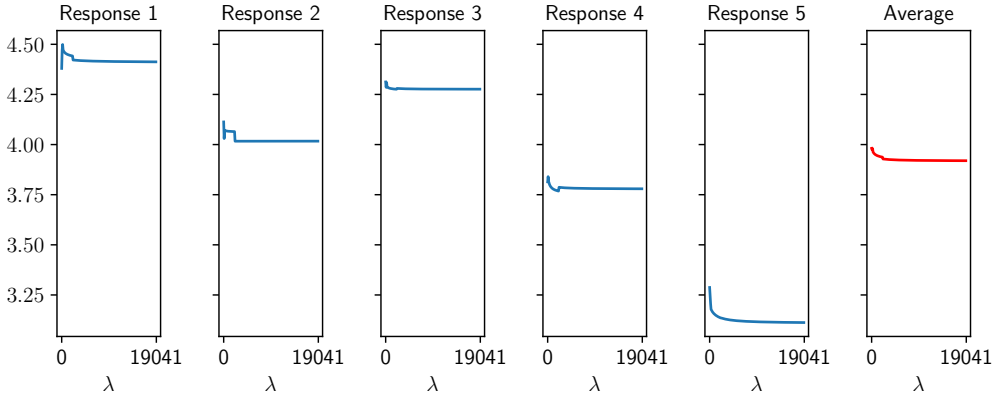


Figure 7.A.18: Mean-squared prediction error for each response variable in the hold-out dataset. The mean-squared prediction error for the system is given in red.

The results presented here are for the *Application* model with $\rho = 0.95$, and $\sigma^2_{\eta_m} = 0.5$. Here, the level of sparsity k, is less than the true sparsity of the model.

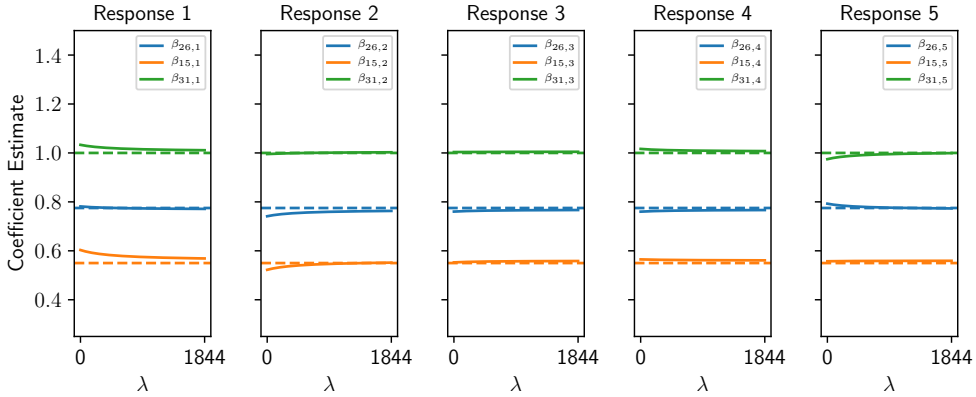


Figure 7.A.19: Regression coefficient trace-plots for an increasing shrinkage penalty.

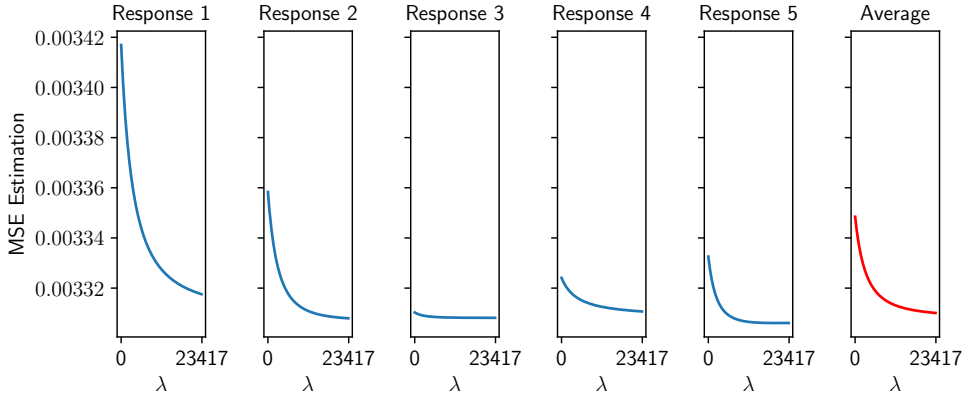


Figure 7.A.20: Mean-squared estimation error for each model. The mean-squared estimation error for the system is given in red.

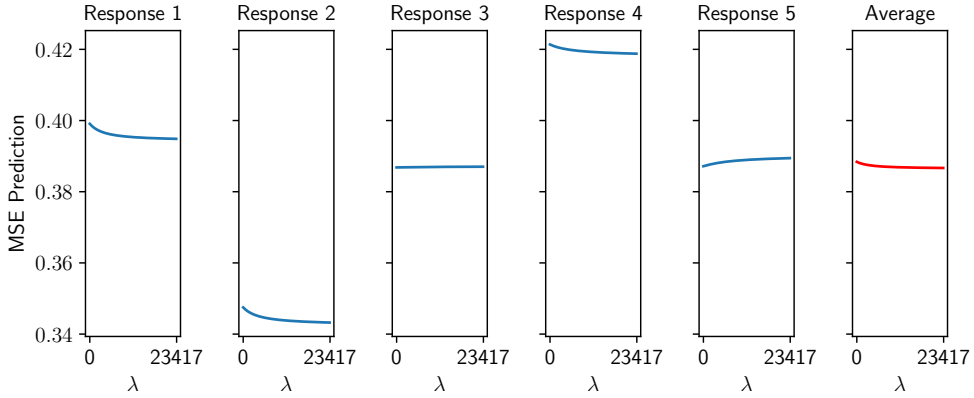


Figure 7.A.21: Mean-squared prediction error for each response variable in the hold-out dataset. The mean-squared prediction error for the system is given in red.

The results here are for the *Application* model with $\rho = 0.95$, and $\sigma^2_{\eta_m} = 0.5$. Here, the level of sparsity k, is equal to the true sparsity of the model.

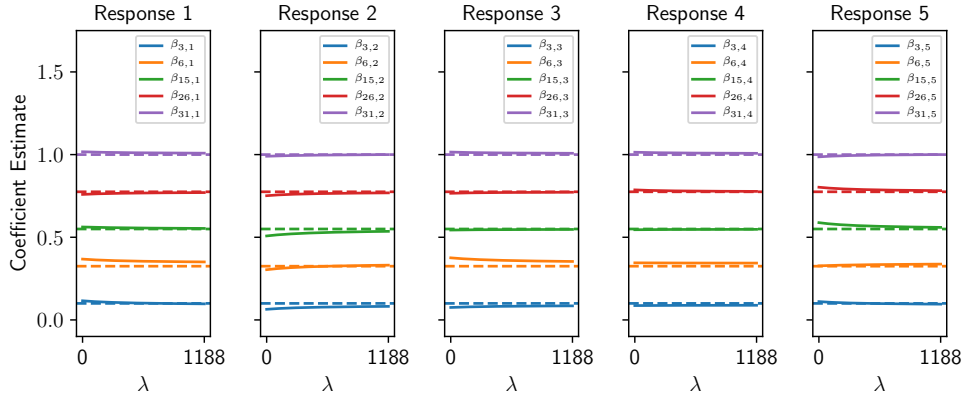


Figure 7.A.22: Regression coefficient trace-plots for an increasing shrinkage penalty.

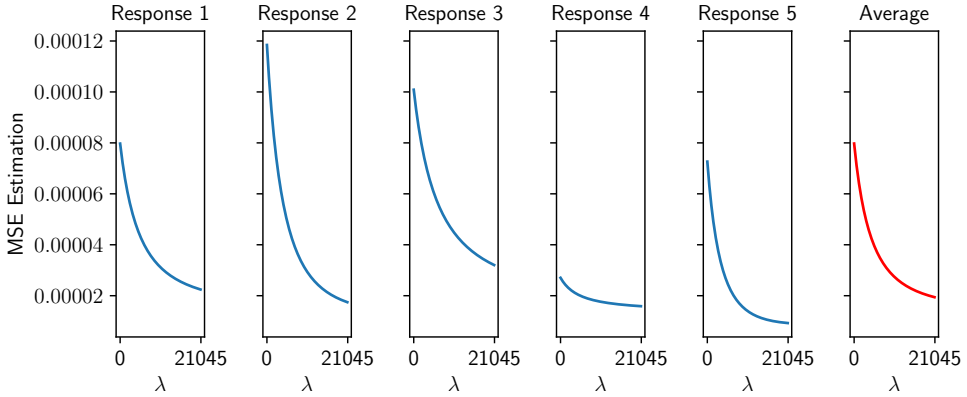


Figure 7.A.23: Mean-squared estimation error for each model. The mean-squared estimation error for the system is given in red.

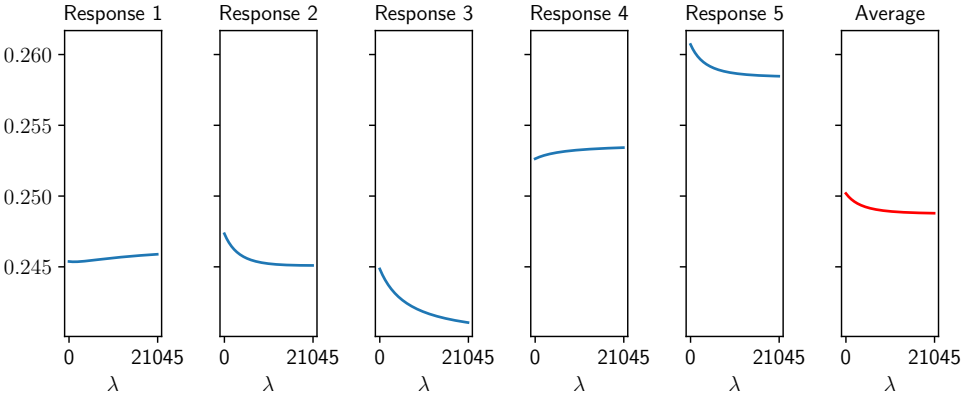


Figure 7.A.24: Mean-squared prediction error for each response variable in the hold-out dataset. The mean-squared prediction error for the system is given in red.

The results here are for the *Application* model with $\rho = 0.95$, and $\sigma^2_{\eta_m} = 0.5$. Here, the level of sparsity k, is greater than the true sparsity of the model.

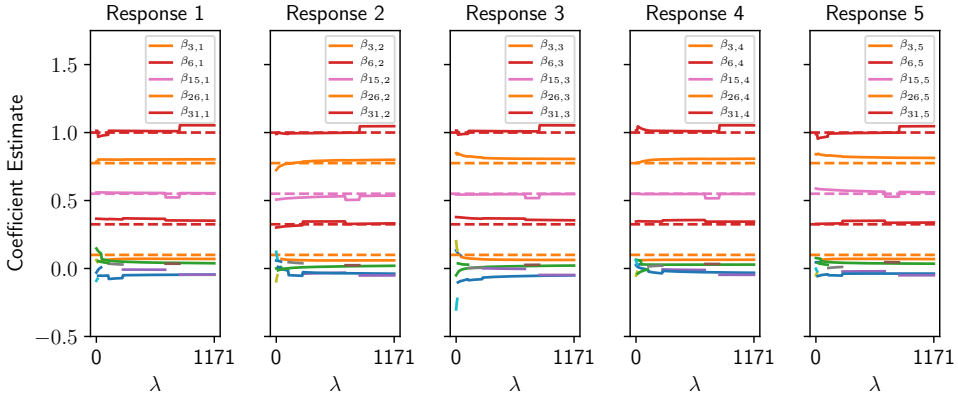


Figure 7.A.25: Regression coefficient trace-plots for an increasing shrinkage penalty.

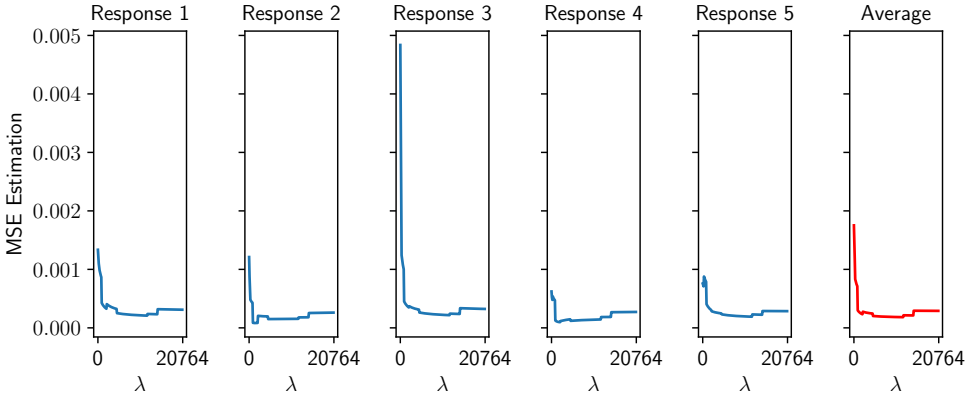


Figure 7.A.26: Mean-squared estimation error for each model. The mean-squared estimation error for the system is given in red.

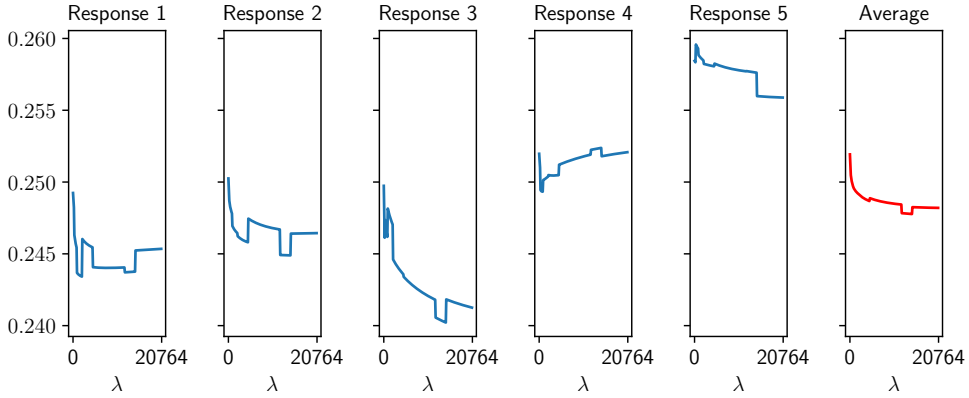


Figure 7.A.27: Mean-squared prediction error for each response variable in the hold-out dataset. The mean-squared prediction error for the system is given in red.

The results presented here are for the *Application* model with $\rho = 0.95$, and $\sigma^2_{\eta_m} = 2$. Here, the level of sparsity k, is less than the true sparsity of the model.

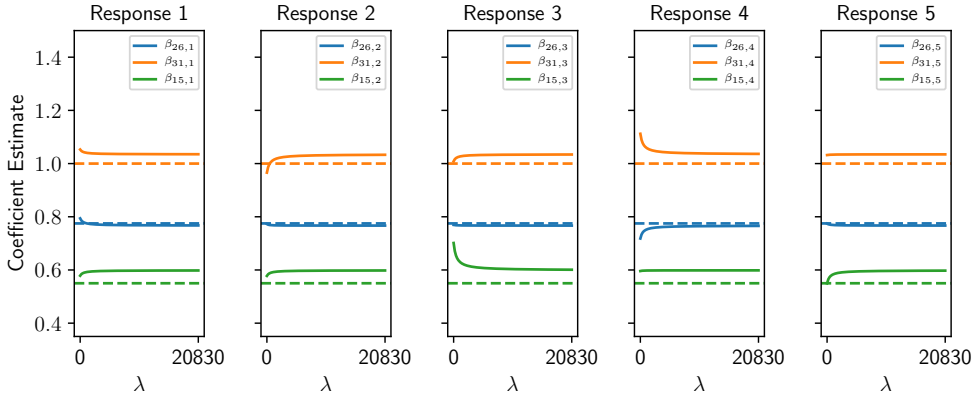


Figure 7.A.28: Regression coefficient trace-plots for an increasing shrinkage penalty.

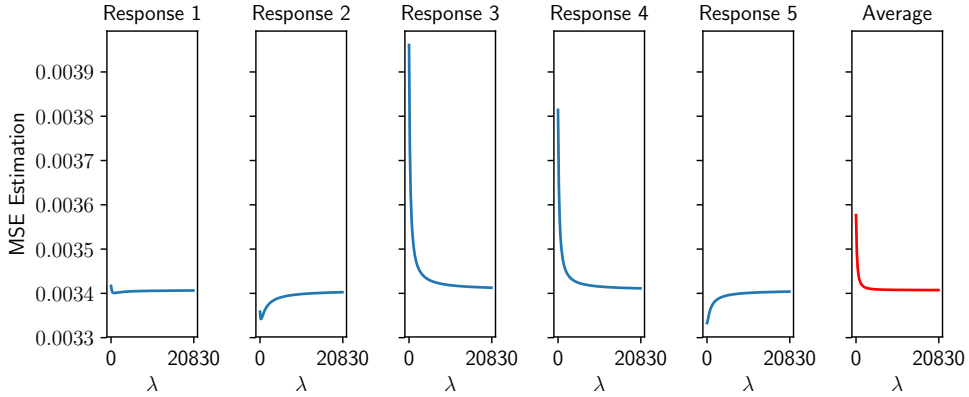


Figure 7.A.29: Mean-squared estimation error for each model. The mean-squared estimation error for the system is given in red.

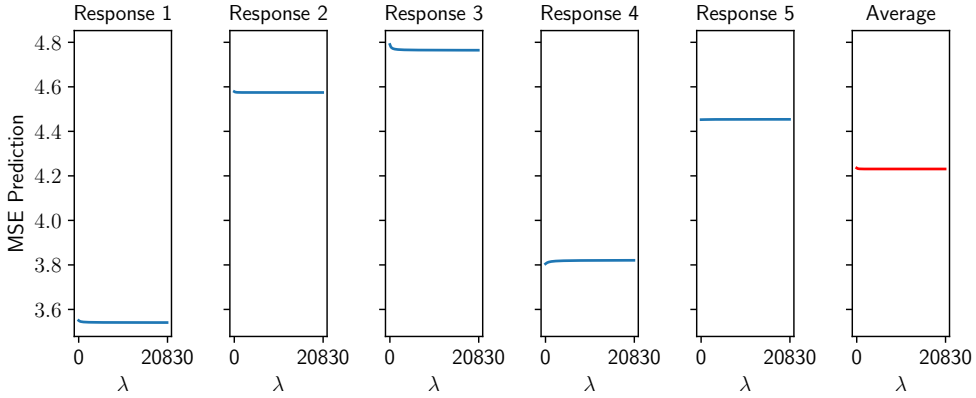


Figure 7.A.30: Mean-squared prediction error for each response variable in the hold-out dataset. The mean-squared prediction error for the system is given in red.

The results here are for the *Application* model with $\rho = 0.95$, and $\sigma^2_{\eta_m} = 2$. Here, the level of sparsity k, is equal to the true sparsity of the model.

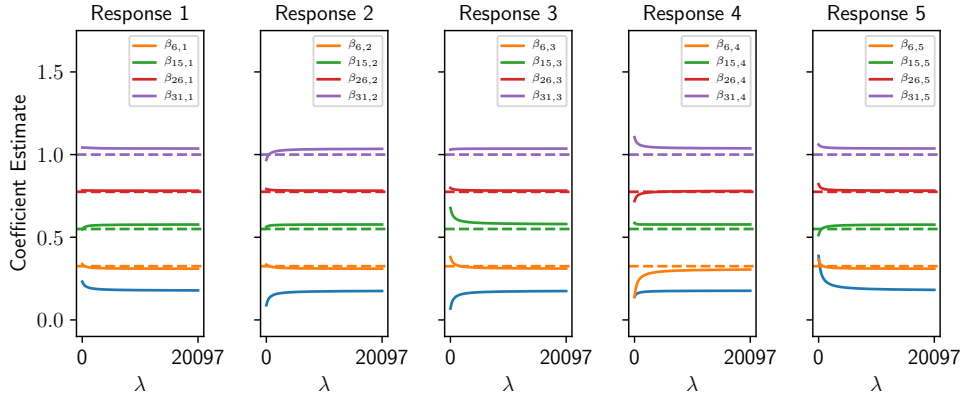


Figure 7.A.31: Regression coefficient trace-plots for an increasing shrinkage penalty.

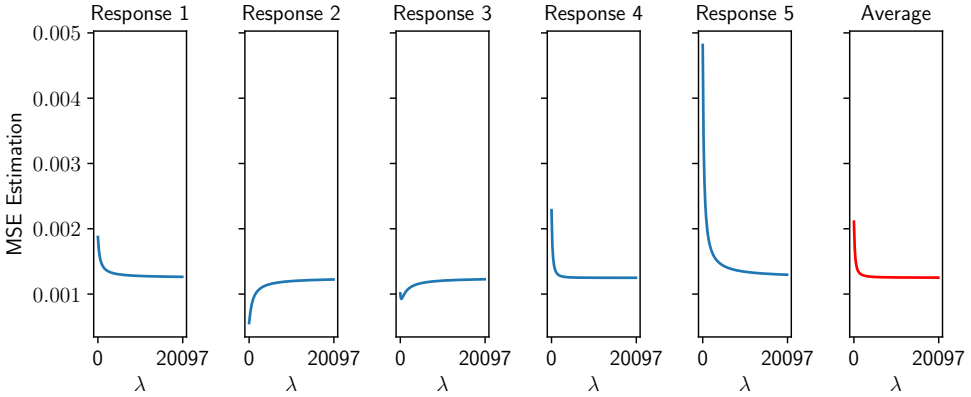


Figure 7.A.32: Mean-squared estimation error for each model. The mean-squared estimation error for the system is given in red.

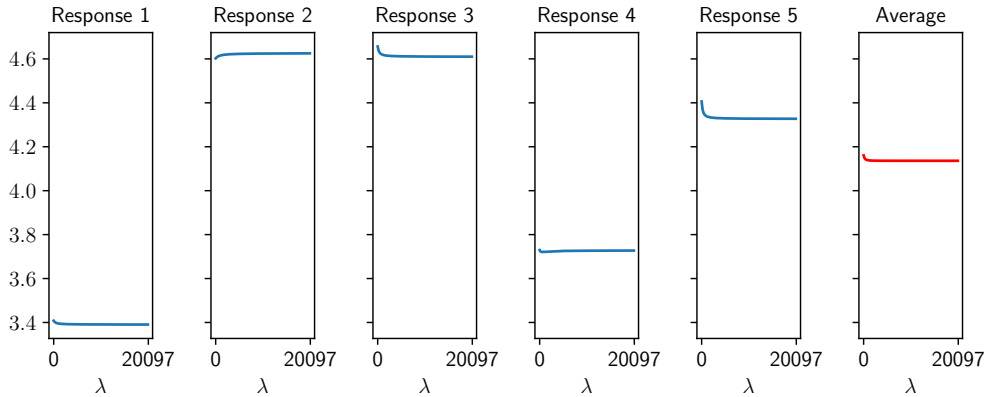


Figure 7.A.33: Mean-squared prediction error for each response variable in the hold-out dataset. The mean-squared prediction error for the system is given in red.

The results here are for the *Application* model with $\rho = 0.95$, and $\sigma^2_{\eta_m} = 2$. Here, the level of sparsity k, is greater than the true sparsity of the model.

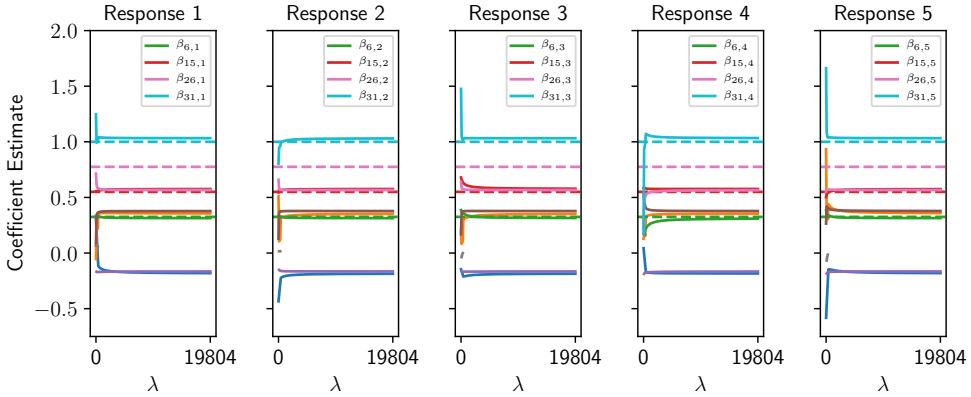


Figure 7.A.34: Regression coefficient trace-plots for an increasing shrinkage penalty.

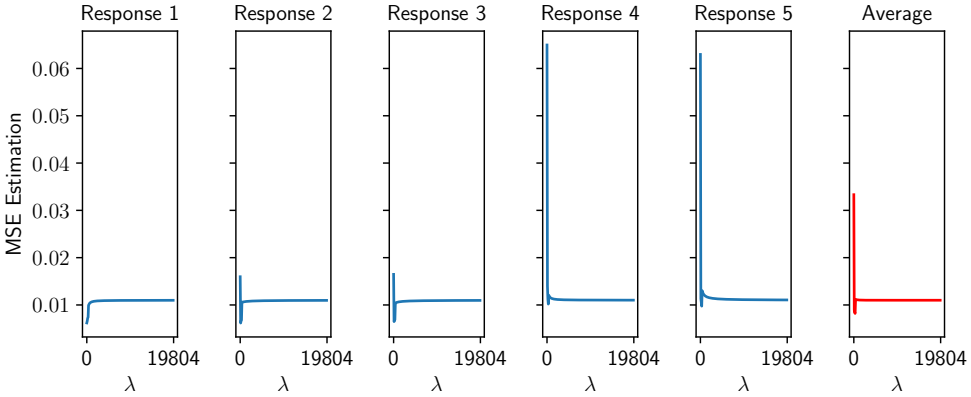


Figure 7.A.35: Mean-squared estimation error for each model. The mean-squared estimation error for the system is given in red.

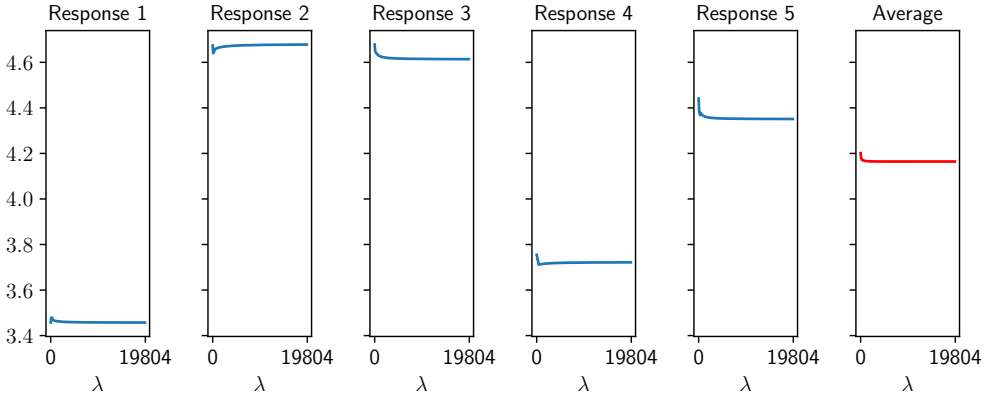


Figure 7.A.36: Mean-squared prediction error for each response variable in the hold-out dataset. The mean-squared prediction error for the system is given in red.

The results presented here are for the *Uniformly spaced* model with $\rho = 0.95$, and $\sigma^2_{\eta_m} = 0.5$. Here, the level of sparsity k, is less than the true sparsity of the model.

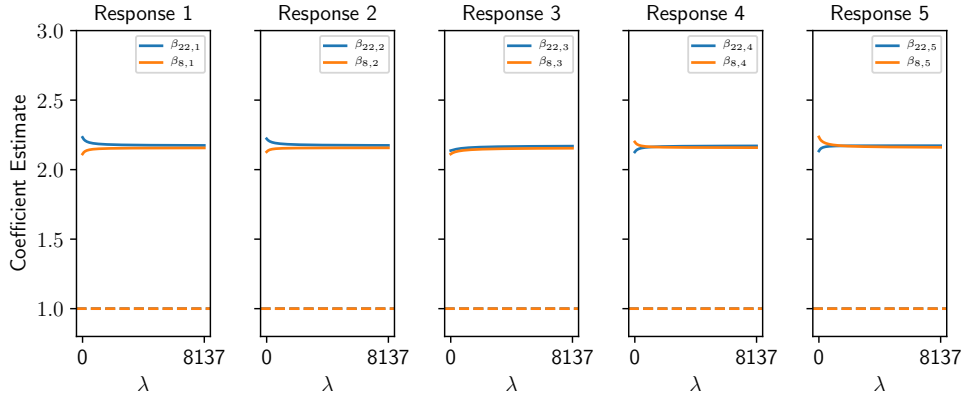


Figure 7.A.37: Regression coefficient trace-plots for an increasing shrinkage penalty.

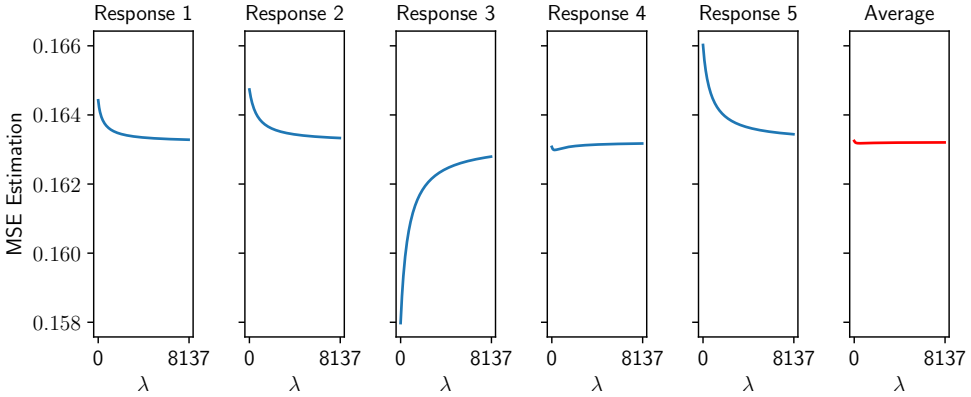


Figure 7.A.38: Mean-squared estimation error for each model. The mean-squared estimation error for the system is given in red.

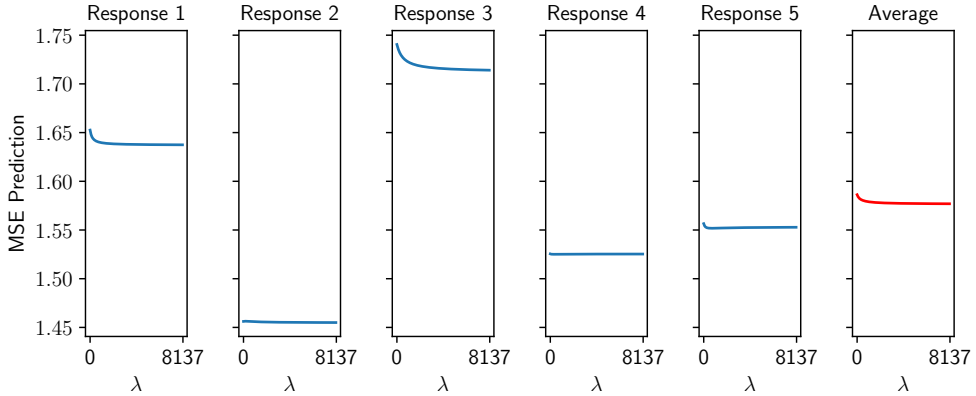


Figure 7.A.39: Mean-squared prediction error for each response variable in the hold-out dataset. The mean-squared prediction error for the system is given in red.

The results here are for the *Uniformly spaced* model with $\rho = 0.95$, and $\sigma^2_{\eta_m} = 0.5$. Here, the level of sparsity k, is equal to the true sparsity of the model.

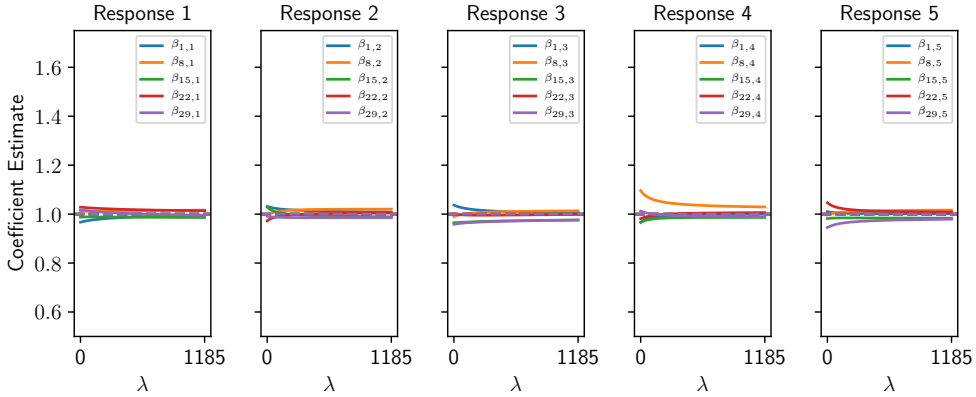


Figure 7.A.40: Regression coefficient trace-plots for an increasing shrinkage penalty.

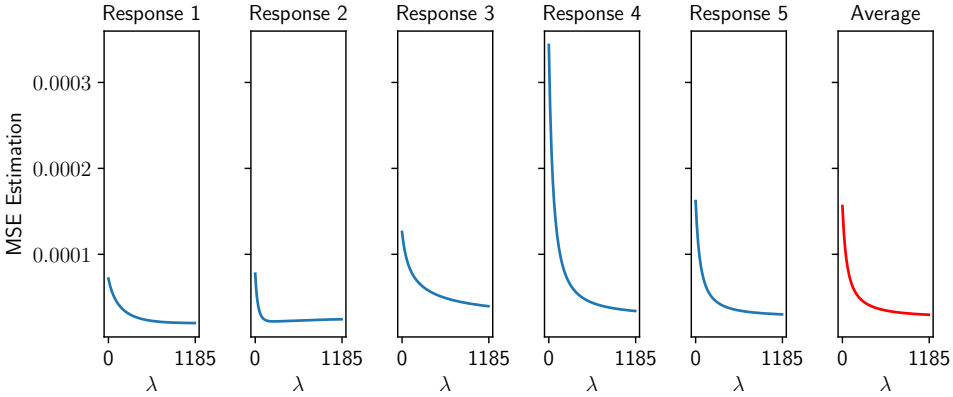


Figure 7.A.41: Mean-squared estimation error for each model. The mean-squared estimation error for the system is given in red.

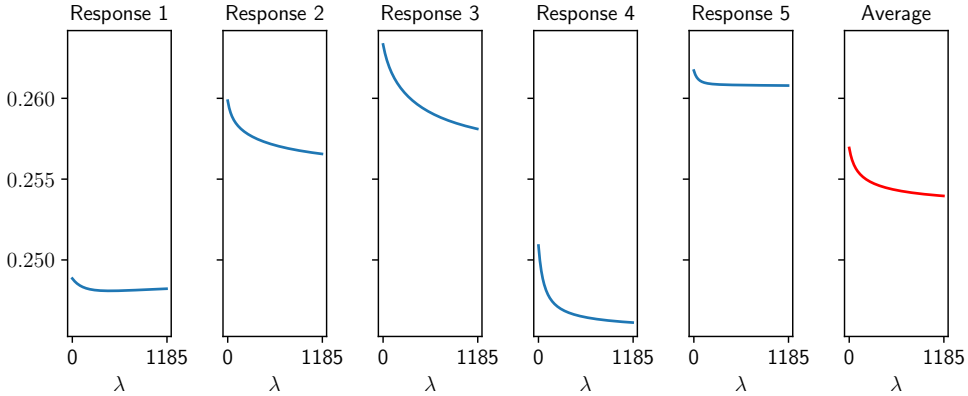


Figure 7.A.42: Mean-squared prediction error for each response variable in the hold-out dataset. The mean-squared prediction error for the system is given in red.

The results here are for the *Uniformly spaced* model with $\rho = 0.95$, and $\sigma^2_{\eta_m} = 0.5$. Here, the level of sparsity k, is greater than the true sparsity of the model.

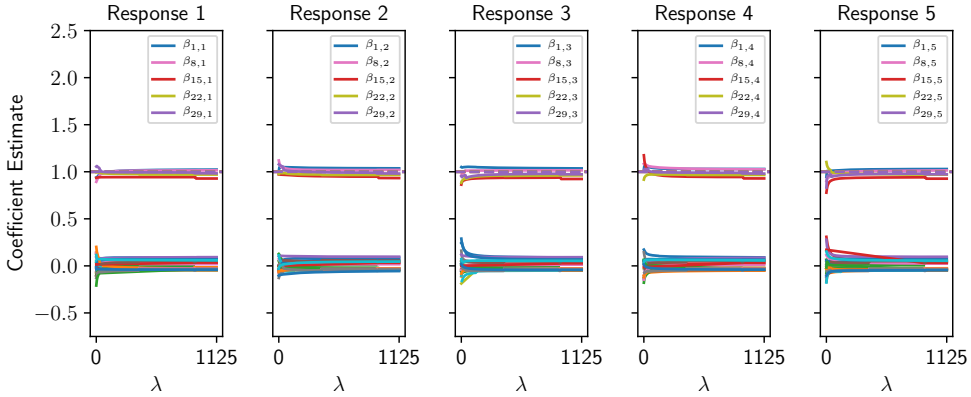


Figure 7.A.43: Regression coefficient trace-plots for an increasing shrinkage penalty.

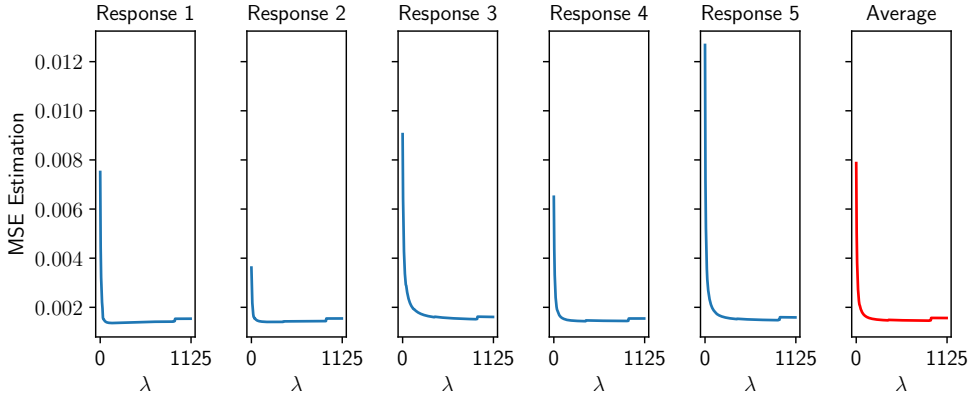


Figure 7.A.44: Mean-squared estimation error for each model. The mean-squared estimation error for the system is given in red.

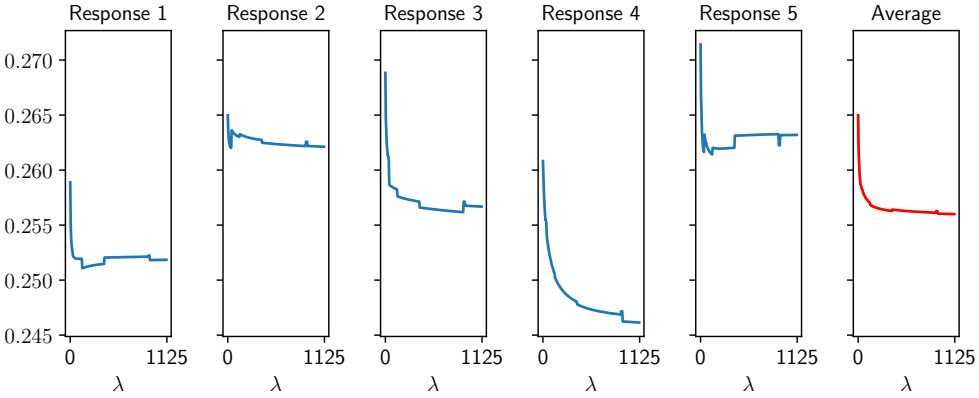


Figure 7.A.45: Mean-squared prediction error for each response variable in the hold-out dataset. The mean-squared prediction error for the system is given in red.

The results presented here are for the *Uniformly spaced* model with $\rho = 0.95$, and $\sigma^2_{\eta_m} = 2$. Here, the level of sparsity k, is less than the true sparsity of the model.

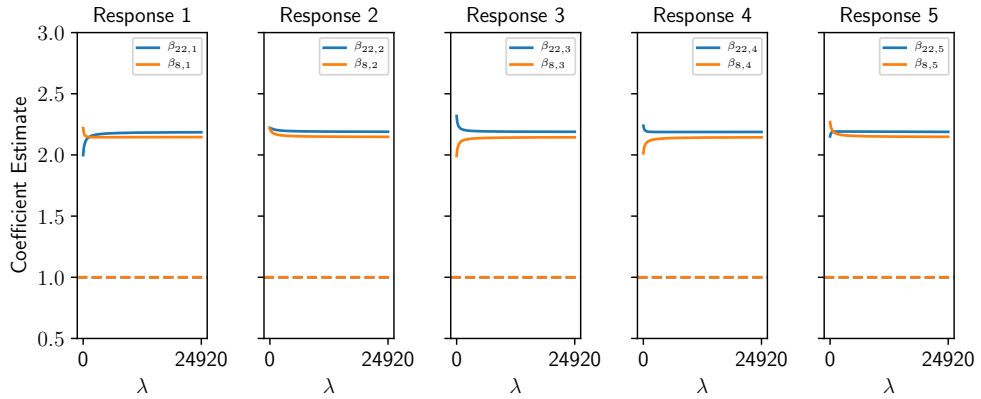


Figure 7.A.46: Regression coefficient trace-plots for an increasing shrinkage penalty.

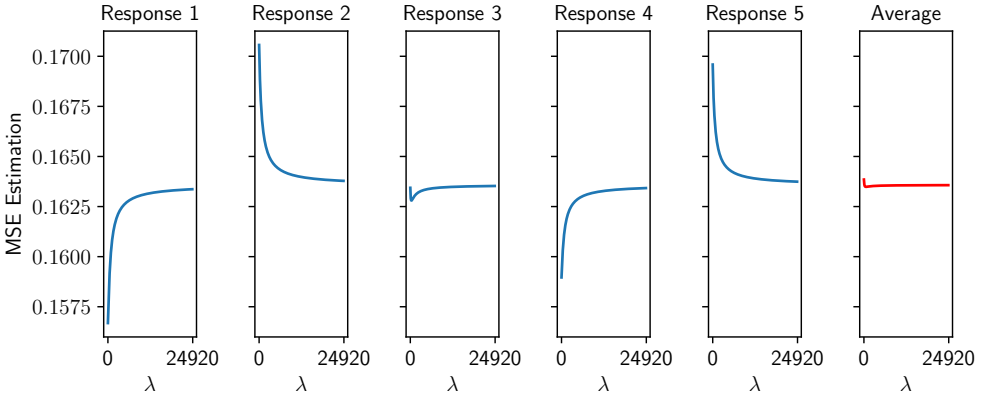


Figure 7.A.47: Mean-squared estimation error for each model. The mean-squared estimation error for the system is given in red.

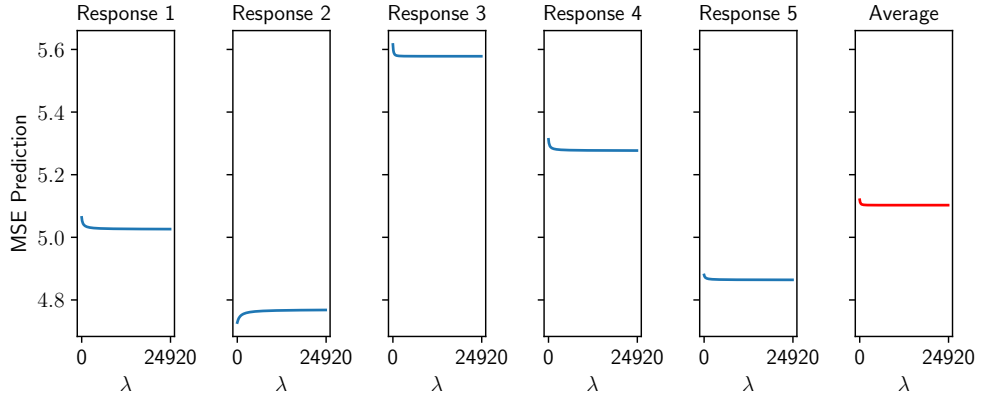


Figure 7.A.48: Mean-squared prediction error for each response variable in the hold-out dataset. The mean-squared prediction error for the system is given in red.

The results here are for the *Uniformly spaced* model with $\rho = 0.95$, and $\sigma^2_{\eta_m} = 2$. Here, the level of sparsity k, is equal to the true sparsity of the model.

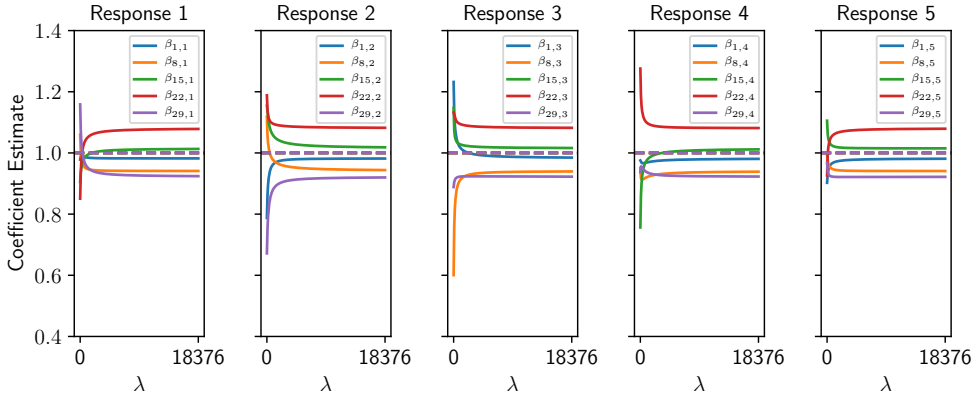


Figure 7.A.49: Regression coefficient trace-plots for an increasing shrinkage penalty.

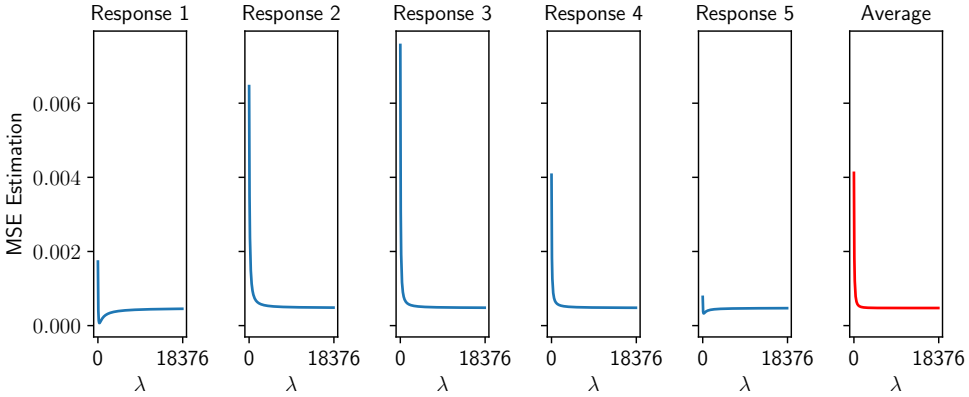


Figure 7.A.50: Mean-squared estimation error for each model. The mean-squared estimation error for the system is given in red.

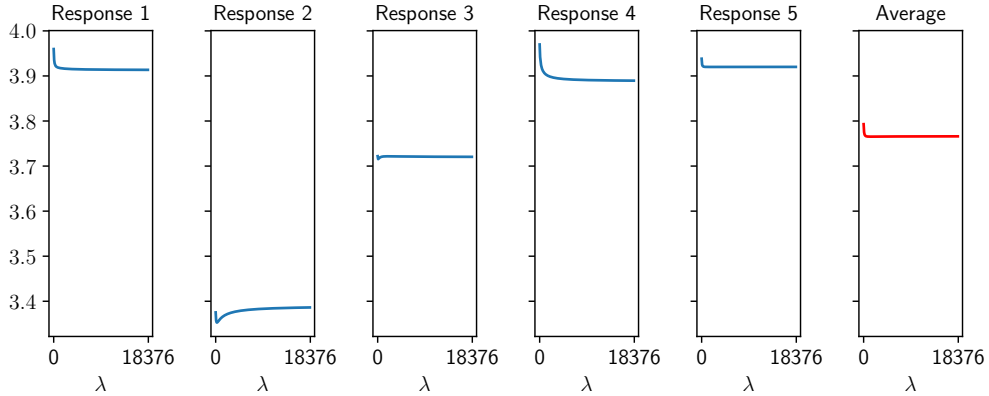


Figure 7.A.51: Mean-squared prediction error for each response variable in the hold-out dataset. The mean-squared prediction error for the system is given in red.

The results here are for the *Uniformly spaced* model with $\rho = 0.95$, and $\sigma^2_{\eta_m} = 2$. Here, the level of sparsity k, is greater than the true sparsity of the model.

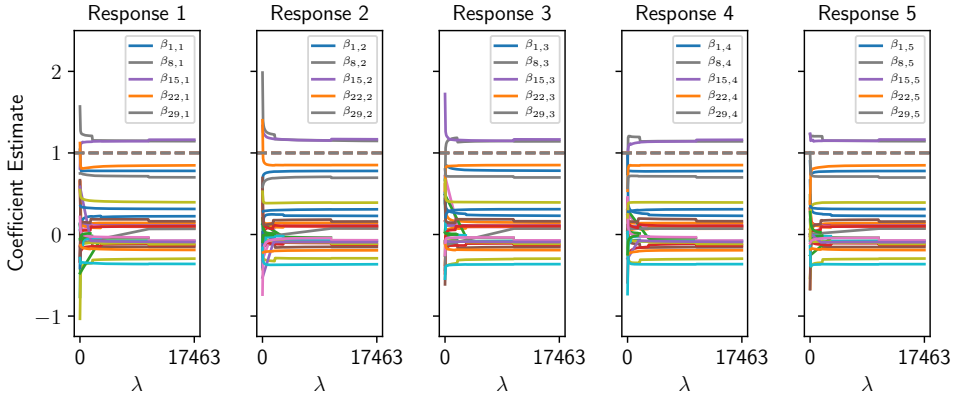


Figure 7.A.52: Regression coefficient trace-plots for an increasing shrinkage penalty.

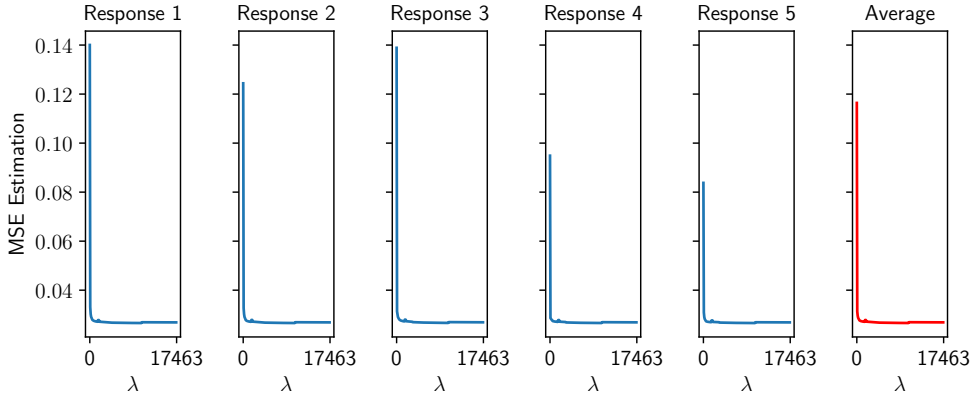


Figure 7.A.53: Mean-squared estimation error for each model. The mean-squared estimation error for the system is given in red.

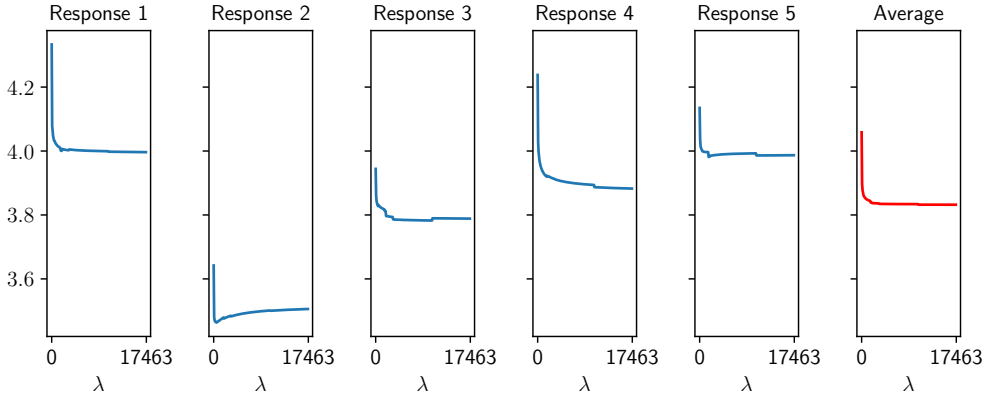


Figure 7.A.54: Mean-squared prediction error for each response variable in the hold-out dataset. The mean-squared prediction error for the system is given in red.

Chapter 8

Conclusions and further work

Within this thesis we have developed and implemented a range of simultaneous predictor selection methods to jointly estimate multiple linear regression models. Much of the work in this thesis has been motivated by the challenges encountered when modelling telecommunications data.

In Chapter 3, we proposed a generalisation of the best-subset problem (Miller, 2002) which we called the *Simultaneous Best-Subset* problem. The idea of solving the Simultaneous Best Subset problem is to simultaneously select predictors for multiple linear regression models. By allowing at most k unique predictors to be present across a set of regression models, we were able to obtain sparse models in which the same predictors were often present in each regression model. In addition to this, we were able to show empirically that the regression models obtained from solving the SBS problem were superior to those obtained from fitting each regression model *individually* using the best-subset approach. The solutions to the SBS problem more often contained the true predictors than solutions to the best-subset problem. Further, the SBS approach appeared to be consistent in predictor selection. As we increased the number of models jointly estimated the SBS method would more often identify the correct predictors. Consequently, the estimation error in the regression coefficients also reduced.

In Chapter 3, we determined the best SARIMA models for the regression residuals by fitting each model from a list of suitable models and selecting the model with the smallest value of the BIC (Schwarz, 1978). Whilst this approach is in a sense *automatic*, there is scope to improve it. Hyndman and Khandakar (2008) developed an algorithm to automatically identify SARIMA models. Their motivation for this approach was to obtain automatic forecasts for a large number of time series in a business setting. Their approach is implemented in the `forecast` (Hyndman and Khandakar, 2008) package for `R` and may provide a *more automated* approach than that we have implemented.

Additionally, we have identified the SARIMA models for the residuals individually. It may be

possible to improve the performance of the models by modelling the residuals as a multivariate process using the multi-class vector autoregressive models used by Barbaglia et al. (2016) and Wilms et al. (2018). This multi-class approach is even capable of encouraging similarity across the residual models. This idea is similar in nature to the idea we have used in our simultaneous shrinkage operator.

In Chapter 4 we applied the *Automated* approach developed in Chapter 3 to the whole telecommunications dataset. We found that the Reg-SARIMA models produced by this approach were more accurate for short-term predictions and were often more accurate for long-term predictions. In cases where the Reg-SARIMA models were less accurate the predictive accuracy was comparable to all other approaches. The models produced by the *Automated* approach often contained fewer weather related predictor variables. Consistency in the selected predictors across models within a response ensured that the models were far more interpretable. In addition to this, the effects of all predictors were inline with expert opinion and the models did not include pairs of highly correlated predictors with opposing effects. This provided a significant improvement over the current procedure.

In Chapter 5 we were able to show empirically that the time to solve the SBS problem could be reduced with formulations of the SBS problem that contained data-driven parameters. These parameters were derived from optimal solutions of the SBS problem. However, in practice we were not able to reduce the time to solve the SBS problem using parameters derived from solutions obtained by our Discrete First Order Algorithm (DFOA). Bertsimas et al. (2016) noted that the optimisation solver often found very good solutions to the best-subset problem quickly. It would be interesting to compare the solutions obtained from our DFOA to a solution obtained from an optimisation solver after a short amount of time. If very good solutions for the SBS problem are obtained in a short amount of time we may be able to improve the formulation parameter estimates and hence improve the time to solve the SBS problem in practise.

In Chapter 6 we proposed a number of alternative *fast* simultaneous predictor selection methods. With these methods we were able to jointly estimate multiple linear regression models significantly quicker than applying the SBS method. We found that these fast methods often fit the same models as the SBS approach, so could be used as a practical alternative to the SBS approach in problems where the number of response variables, or the number of predictors is much higher. Whilst the models produced by the Stepwise and Hybrid approaches were the same as the model produced by the SBS approach in our simulation study in Section 6.2, it would be of significant practical interest to see if these fast approaches perform as well as the SBS approach when applied to the telecommunications event data.

In Chapter 7 we further studied the performance of the simultaneous shrinkage operator that, to

the best of our knowledge, has not been considered in the literature. The idea behind simultaneous shrinkage was to *shrink* coefficients across multiple regression models towards a common value. We found that the operator could improve the selection accuracy of predictors and the estimation of regression coefficients. The operator was found to be particularly useful when the models produced by the SBS approach were not sparse. With shrinkage, we were able to drive many of the coefficients that should be zero towards zero as the penalty increased. This ultimately reduced the mean-squared prediction error of the models. Rather than using an l_2 shrinkage penalty, we could consider using an l_1 penalty of the form

$$\mathcal{P}(\boldsymbol{\beta}) = \lambda \sum_{m=1}^{M} \sum_{p=1}^{P} |\tilde{\beta}_p - \beta_{p,m}|,$$

where $\tilde{\beta}_p$ for $p = 1, \ldots, P$ are auxiliary variables used to produce similar values across models. The form of this l_1 penalty may set $\tilde{\beta}_p = \beta_{p,m}$ for $p = 1, \ldots, P$ and $m = 1, \ldots, M$ for some λ large enough.

In Chapter 5, we were able to identify a good value for the shrinkage parameter using cross-validation approaches (Stone, 1974). However, it may not always to appropriate to use cross-validation to determine parameters. One example is when only a small number of observations are available. Zou et al. (2007) were able to do this for the LASSO using the framework proposed by Stein (1981). This work may provide a good starting point for us to derive information criteria such as the AIC (Akaike, 1973) or BIC (Schwarz, 1978) for systems of linear regression models estimated with simultaneous shrinkage.

Our approach could also be extended in a number of ways. In some applications it may not be possible to observe all predictors for each response variable. We could easily modify the MIQO program to ensure that similarity in predictor selection is encouraged in this scenario. We could also implement a simulation study where the variance of some response variables is greater than others. We suspect that compared to individual regression methods, we may gain estimation accuracy in the response variables with the highest variance at the expense of losing accuracy in estimation for the response variables with the lowest variance. It would also be great to see our *Automated* approach applied to other datasets. Since providing the software package to BT, it has already been applied to investigate the electricity demand of different telecommunications buildings. The automated nature of the approach whilst providing good interpretable models shows that our approach can have significant impact in industry,

Bibliography

Akaike, H. (1973). *Information Theory and an Extension of the Maximum Likelihood Principle*, pages 199–213. Springer, New York, NY.

Barbaglia, L., Wilms, I., and Croux, C. (2016). Commodity dynamics: A sparse multi-class approach. *Energy Economics*, 60:62–72.

Beale, E. (1970a). Selecting an Optimum Subset. In *Integer and Nonlinear Programming*. North-Holland Publishing Company.

Beale, E. M. L. (1970b). Note on Procedures for Variable Selection in Multiple Regression. *Technometrics*, 12(4):909–914.

Beale, E. M. L., Kendall, M. G., and Mann, D. W. (1967). The Discarding of Variables in Multivariate Analysis. *Biometrika*, 54(3 and 4):357–366.

Berk, K. N. (1978). Comparing Subset Regression Procedures. *Technometrics*, 20(1):1–6.

Bertsimas, D. and King, A. (2016). OR Forum: An Algorithmic Approach to Linear Regression. *Operations Research*, 64(1):2–16.

Bertsimas, D., King, A., and Muzumder, R. (2016). Best Subset Selection Via a Modern Optimisation Lens. *The Annals of Statistics*, 44(2):813–852.

Boot, J. C. G. (1964). *Quadratic Programming: Algorithms-Anomolies-Applications*, volume 2. North-Holland Publishing Company, Amsterdam.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 3(1):1–122.

Breiman, L. and Friedman, J. H. (1997). Predicting Multivariate Responses in a Multiple Linear Regression. *Journal of the Royal Statistical Society, Series B*, 59(1):3–54.

Brockwell, P. J. and Davis, R. A. (2002). *Introduction to Time Series and Forecasting*. Springer, second edition.

Brown, P. J. and Zidek, J. V. (1980). Adaptive Multivariate Ridge Regression. *Annals of Statistics*, 8(1):64–74.

Chatfield, C. (2000). *Time-Series Forecasting*. Chapman & Hall CRC.

Chatterjee, A. and Lahiri, S. N. (2011). Bootstrapping lasso estimators. *Journal of the American Statistical Association*, 106(494):608–625.

Chatterjee, S., Price, A. S., and Price, B. (2012). *Regression Analysis by Example*. John Wiley & Sons, New York, 5th edition.

Chen, D.-S., Batson, R. G., and Dang, Y. (2010). *Applied Integer Programming: Modleling and Solution*. Wiley.

Cochrane, D. and Orcutt, G. H. (1949). Application of Least Squares Regression to Relationships Containing Auto- Correlated Error Terms. *Journal of the American Statistical Association*, 44(245):32–61.

Dai, H., Izatt, G., and Tedrake, R. (2019). Global Inverse Kinematics via Mixed-Integer Convex Optimization. *To appear in the International Journal of Robotics Research*.

Ding, J., Tarokh, V., and Yang, Y. (2019). Model Selection Techniques-An Overview. *To appear in IEEE Signal Processing Magazine*.

Donoho, D. L. (2006). Compressed Sensing. *IEE Transactions on Information Theory*, 52(4).

Donoho, D. L. and Johnstone, J. M. (1994). Ideal Spatial Adptation by Wavelet Shrinkage. *Biometrika*, 81(3):425–455.

Draper, N. R. and Smith, H. (1998). *Applied Regression Analysis*. Wiley Series in Probability and Statistics. Wiley, New York, 3rd edition.

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004). Least Angle Regression. *The Annals of Statistics*, 32(2):407–409.

Efroymson, M. A. (1960). Multiple Regression Analysis. *Mathematical Methods for Digital Computers*, 2(2).

Fang, Y. and Koreisha, S. G. (2004). Forecasting with Serially Correlated Regression Models. *Journal of Statistical Computation and Simulation*, 74(9):625–649.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software*, 33(1):1–22.

Furnival, G. M. and Wilson, R. W. J. (1974). Regressions by Leaps and Bounds. *Technometrics*, 16(4):499–511.

Gaines, B. R., Kim, J., and Zhou, H. (2018). Algorithms for fitting the constrained lasso. *Journal of Computational and Graphical Statistics*, 27(4):861–871.

Garthwaite, P. H. and Dickey, J. M. (1988). Quantifying expert opinion in linear regression problems. *Journal of the Royal Statistical Society Series B*, 50(3):462–474.

Gurobi Optimization, L. (2018). Gurobi optimizer reference manual.

Hannan, E. J. and Quinn, B. G. (1979). The Determination of the Order of an Autoregression. *Journal of the Royal Statistical Society, Series B*, 41(2):190–195.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer New York, 2nd edition.

Hastie, T., Tibshirani, R., and Tibshirani, R. (2017). Extended Comparisons of Best Subset Selection, Forward Stepwise Selection, and the Lasso. Submitted.

Hocking, R. (1976). A Biometrics Invited Paper: The Analysis and Selection of Variables in Linear Regression. *Biometrics*, 32(1):1–49.

Hocking, R. R. and Leslie, R. N. (1967). Selection of the Best Subset in Regression Analysis. *Technometrics*, 9(4):531–540.

Hoerl, E. and Kennard, R. W. (1970). Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1):55–67.

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and Time Series Model Selection in Small Samples. *Biometrika*, 76(2):297–307.

Hutmacher, M. M. and Kowalski, K. (2014). Covariate Selection in Pharmacometric Analyses: A Review of Methods. *British Journal of Clinical Pharmacology*, 79(1):132–147.

Hyndman, R. J. and Athanasopoulos, G. (2019). *Forecasting: Principles and Practice*. OTexts: Melbourne, Australia, 3rd edition. OTexts.com/fpp3 Accessed on 19/08/2019.

Hyndman, R. J. and Khandakar, Y. (2008). Automatic Time Series Forecasting: The forecast Package for R. *Journal of Statistical Software*, 27(3).

Izenman, A. J. (1975). Reduced-rank Regression for the Multivariate Linear Model. *Journal of Multivariate Analysis*, 5(2):248–264.

Jiang, Y., He, Y., and Zhang, H. (2016). Variable selection with prior information for generalized linear models via the prior lasso method. *Journal of the American Statistical Association*, 111(513):355–376.

Jordan, M. I. and Mitchel, T. M. (2015). Machine Leanring: Trends, Prespectives and Prospects. *Science*, 349(6245):255–260.

Katal, A., Wazid, M., and Goudar, R. (2013). Big Data: Issues, Challenges, Tools and Good Practices. In Parashar, M., Zomaya, A., Chen, J., Cao, J.-N., Bouvry, P., and Prasad, S., editors, *2013 Sixth International Conference on Contemporary Computing (IC3)*. Jaypee Institute of Information Technology, IEEE.

Knight, K. and Fu, W. (2000). Asymptotics for LASSO-Type Estimators. *The Annals of Statistics*, 28(5):1356–1378.

LaMotte, L. R. (1972). The SELECT Routines: A Program for Identifying Best Subset Regressions. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 21(1):92–93.

LaMotte, L. R. and Hocking, R. R. (1970). Computational Efficiency in the Selection of Regression Variables. *Technometrics*, 12(1):83–93.

Land, A. H. and Doig, A. (1960). An automatic method for solving discrete programming problems. *Econometrica*, 28(3):497–520.

Lee, K. H., Tadesse, M. G., Baccarelli, A. A., Schwartz, J., and Coull, B. A. (2017). Multivariate Bayesian Variable Selection Exploiting Dependence Structure among Outcomes: Application to Air Pollution Effects on DNA Methylation. *Biometrics*, 73:232–241.

Lee, W. and Liu, Y. (2012). Simultaneous Multiple Response Regression and Inverse Covariance Matrix Estimation via Penalized Gaussian Maximum Likelihood. *Journal of Multivariate Analysis*, 1(111):241–255.

Longley, J. W. (1967). An Appraisal of Least Squares Programs for the Electronic Computer from the point of View of Use. *Journal of the American Statistical Association*, 62:819–841.

Lumley, T. (2017). *leaps: Regression Subset Selection*. R package version 3.0.

Mantel, N. (1970). Why Stepdown Procedures in Variable Selection. *Technometrics*, 12(3):621–625.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1994). *Multivariate Analysis*. Academic Press.

Mazumder, R., Friedman, J. H., and Hastie, T. (2011). SparseNet: Coordinate Descent win Non-convex Penalties. *Journal of the American Statistical Association*, 106(495):1125–1138.

Meinshausen, N. (2007). Relaxed lasso. *Computational Statistics & Data Analysis*, 52:374–393.

Meinshausen, N. and Buhlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Miller, A. J. (1984). Selection of subsets of regression variables. *Journal of the Royal Statistical Society Series A*, 147(3):389–425.

Miller, A. J. (1996). The Convergence of Efroymson's Stepwise Algorithm. *The American Statistician*, 50(2):180–181.

Miller, A. J. (2002). *Subset Selection in Regression*. Monographs on statistics and applied probability 95. Chapman and Hall CRC, 2nd edition.

Nagabhushana Rao, R. V. S. S., Balasiddamuni, P., and Ramana Murthy, B. (2013). *Inference in Seemingly Unrelated Regression Equation Models: Feasible Estimation*. LAP Lambert Academic Publishing, first edition.

Natarajun, B. K. (1995). Sparse Approximate Solutions to Linear Systems. *Siam Journal of Computing*, 24(2):227–234.

Nesterov, Y. (2005). Smooth Minimization of Non-smooth Functions. *Mathematical Programming*, 103(1):127–152.

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer Series in Operations Research. Springer, second edition.

Oliphant, T. E. (2006). *A guide to NumPy*. Trelgol publishing, USA.

Oosterhoff, J. (1963). On the Selection of Independent Variables in a Regression Equation. Technical report, Amsterdam. Stichting Mathematisk Centrum.

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

Proost, F. and Fawcett, T. (2013). Data Science and its Relationship to Big Data and Data-Driven Decision Making. *Big Data*, 1(1):52–59.

Python Software Foundation (2017). *Python 3.6.3 Documentation*.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rao, C. R. and Toutenburg, H. (1999). *Linear Models: Least Squares and Alternatives*. Springer, 2nd edition.

Rawlings, J. O., Pantula, S. G., and Dickey, D. A. (1998). *Applied Regression Analysis: A Research Tool*. Springer, 2nd edition.

Rissanen, J. (1978). Modeling by Shortest Data Description. *Automatica*, 14(5):465–471.

Rothman, A. J., Levina, E., and Zhu, J. (2010). Sparse Multivariate Regression With Covariance Estimation. *Journal of Computational and Graphical Statistics*, 19(4):947–962.

Ryan, T. P. (2008). *Modern Regression Methods*. Wiley, 2nd edition.

Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2):461–464.

Seber, G. A. and Lee, A. J. (2003). *Linear Regression Analysis*. John Wiley & Sons, New York.

Similä, T. and Tikka, J. (2005). Multiresponse Sparse Regression with Application to Multidimensional Scaling. In Duch, W., Kacprzyk, J., Oja, E., and Zadrożny, S., editors, *Artificial Neural Networks: Formal Models and Their Applications – ICANN 2005*, pages 97–102, Berlin, Heidelberg. Springer Berlin Heidelberg.

Similä, T. and Tikka, J. (2006). Common Subset Selection of Inputs in Multiresponse Regression. In *The 2006 IEEE International Joint Conference on Neural Network Proceedings*, pages 1908–1915.

Similä, T. and Tikka, J. (2007). Input Selection and Shrinkage in Multiresponse Linear Regression. *Computational Statistics & Data Analysis*, 52:406–422.

Simon, N., Friedman, J., and Hastie, T. (2013). A Blockwise Descent Algorithm for Group-penalizes Multiresponse and Multinomial Regression. Submitted.

Snee, R. D. and Marquardt, D. W. (1984). Comment on "Demeaning conditioning diagnostics through centering". *The American Statistician*, 38:83–87.

Soltysik, R. C. and Yarnold, P. R. (2010). Two-Group MultiODA: A Mixed-Integer Linear Programming Solution with Bounded M. *Optimal Data Analysis*, 1(1):30–37.

Srivastava, M. S. and Solanky, T. K. S. (2003). Predicting Multivariate Response in Linear Regression Model. *Communications in Statistics-Simulation and Computation*, 32(2):389–409.

Stein, C. M. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, 9(6):1135–1151.

Stone, M. (1974). Cross-validatory Choice and Assessment of Statistical Predictions. *Journal of the Royal Statistical Society, Series B*, 36:111–147.

Stone, M. (1977). An Asymptotic Equivalence of Choice of Model by Cross-Validation and Akaike's Criterion. *Journal of the Royal Statistical Society, Series B*, 39(1):44–47.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the LASSO. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288.

Tibshirani, R. (2011). Regression Shrinkage and Selection via the LASSO: A Retrospective. *Journal of the Royal Statistical Society, Series B*, 3(73):273–282.

Tibshirani, R., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and Smoothness via the Fused LASSO. *Journal of the Royal Statistical Society, Series B*, 67(1):91–108.

Turlach, B. A., Venables, W. N., and Wright, S. J. (2005). Simultaneous Variable Selection. *Technometrics*, 47(3):349–363.

van der Merwe, A. and Zidek, J. V. (1980). Multivariate Regression Analysis and Canonical Variates. *The Canadian Journal of Statistics*, 8(1):27–39.

Vrieze, S. I. (2012). Model selection and psychological theory: A discussion of the differences between the akaike information criterion (aic) and the bayesian information criterion (bic). *Psychological Methods*, 17(2):228–243.

Weisberg, S. (2014). *Applied Linear Regression*. John Wiley & Sons, New York.

Wilms, I., Barbaglia, L., and Croux, C. (2018). Multiclass vector auto-regressive models for multi-store sales data. *Journal of the Royal Statistical Society, Series C: Applied Statistics*, 67:435–452.

Wolsey, L. A. (1998). *Integer Programming*. John Wiley & Sons.

Xin, X., Hu, J., and Liu, L. (2017). On the Oracle Property of a Generalized Adaptive Elastic-Net for Multivariate Linear Regression with a Divergin Number of Parameters. *Journal of Multivariate Analysis*, 162:16–31.

Yuan, M. and Lin, Y. (2006). Model Selection and Estimation in Regression With Grouped Variables. *Journal of the Royal Statistical Society, Series B*, 68(1):49–67.

Zellner, A. (1962). An Efficient Method of Estimating Seemingly Unrelated Regressions and Tests for Aggregation Bias. *Journal of the American Statistical Association*, 57(298):348–368.

Zhang, C.-H. (2010). Nearly Unbiased Variable Selection Under MinMax Concave Penalty. *The Annals of Statistics*, 38(2):894–942.

Zhang, C.-H. and Huang, J. (2008). The Sparsity and Bias of the LASSO Selection in High-Dimensional Linear Reression. *The Annals of Statistics*, 36(4):1567–1594.

Zhao, P. and Yu, B. (2006). On Model Selection Consistency of LASSO. *Journal of Machine Learning*, 7:2541–2563.

Zou, H. (2006). The Adaptive Lasso and Its Oracle Properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, pages 301–320.

Zou, H. and Hastie, T. (2018). *elasticnet: Elastic-Net for Sparse Estimation and Sparse PCA*. R package version 1.1.1.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the "Degress of Freedom" of the LASSO. *The Annals of Statistics*, 35(5):2173–2192.

Zou, H. and Li, R. (2008). One-Step Estimates in Nonconcave Penalized Likelihood Models. *The Annals of Statistics*, 36(4):1509–1533.