New Procedures for Visualizing Data and Diagnosing Regression Models

by

Rajiv Menjoge

Submitted to the Sloan School of Management
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Operations Research

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2010

Author . . . .
Sloan School of Management
May 18, 2010

Certified by . . . .
Roy E. Welsch
Eastman Kodak Leaders for Global Operations Professor of Management
Professor of Statistics and Engineering Systems
Thesis Supervisor

Accepted by . . . .
Dimitris Bertsimas
Boeing Professor of Operations Research
Co-director, Operations Research Center

New Procedures for Visualizing Data and Diagnosing Regression Models

by

Rajiv Menjoge

Submitted to the Sloan School of Management
on May 18, 2010, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Operations Research

Abstract

This thesis presents new methods for exploring data using visualization techniques.

The first part of the thesis develops a procedure for visualizing the sampling variability of a plot. The motivation behind this development is that reporting a single plot of a sample of data without a description of its sampling variability can be uninformative and misleading in the same way that reporting a sample mean without a confidence interval can be.

Next, the thesis develops a method for simplifying large scatterplot matrices, using similar techniques as the above procedure. The second part of the thesis introduces a new diagnostic method for regression called backward selection search. Backward selection search identifies a relevant feature set and a set of influential observations with good accuracy, given the difficulty of the problem, and additionally provides a description, in the form of a set of plots, of how the regression inferences would be affected with other model choices, which are close to optimal. This description is useful, because an observation, that one analyst identifies as an outlier, could be identified as the most important observation in the data set by another analyst. The key idea behind backward selection search has implications for methodology improvements beyond the realm of visualization. This is described following the presentation of backward selection search.

Real and simulated examples, provided throughout the thesis, demonstrate that the methods developed in the first part of the thesis will improve the effectiveness and validity of data visualization, while the methods developed in the second half of the thesis will improve analysts' abilities to select robust models.

Thesis Supervisor: Roy E. Welsch
Title: Eastman Kodak Leaders for Global Operations Professor of Management
Professor of Statistics and Engineering Systems

Acknowledgments

This research was supported in part by the Singapore-MIT Alliance Program in Computation and Systems Biology and the MIT Center for Computational Research in Economics and Management Science.

I would also like to thank the following people:

My advisor, Professor Roy Welsch, for his guidance, patience, feedback, and support throughout the PhD.

My committee, consisting of Professor Alex Samarov and Professor Arnold Barnett. They have been excellent mentors. In addition, it was great working with Prof. Samarov in the context of research and Prof. Barnett in the context of teaching.

My colleagues at the Operations Research Center and related departments for their advice, and company: especially Douglas Fearing, Parikshit Shah, Jason Acimovic, Ruben Lobel, Deirdre Hatfield, Karima Nigmatulina, Andy Sun, Lavanya Marla, and Gareth Williams.

My family, for their love and guidance. Agustya for showing me around MIT when I arrived, and for his friendship through the years.

And the great friends I made along the way, including Aditya, Vikas, Kranthi, Laura, Theta, Tara, Premal, Angelin, Mihir, Angie, Sonya, Stephanie, Preet, Ky, Carolyn, Jen, Neha, Sheetal, Fatima, Sarah, Malavika, Mrin, and Gita.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Overview

This thesis introduces new procedures for visualizing data and diagnosing regression models. The advances in visualization will enable analysts to apply their own intuition to the data analysis process more effectively. Meanwhile, the advances in diagnostic methods will help analysts understand the extent to which small groups of variables and observations affect the inferences they make in regression. This will allow them to better select a final informative model, or a good set of models, depending on their motives. In the rest of this chapter, we provide a brief motivation and description of the specific methodology we develop in this thesis. We then outline the organization of the rest of the thesis.

1.2 Research Motivation and Description

In this section, we begin by describing the motivation for data visualization in general. We then describe the specific motivations and descriptions of the visualization and regression diagnostic methods we develop in this thesis.

1.2.1 Data Visualization in General

Data visualization is widely known to be extremely useful for several reasons. Below are a few of the several reasons why data visualization is an important part of data analysis:

- The "garbage in, garbage out" principal holds in statistics. If the data are erroneous, a statistician who analyzes the data and doesn't take this into account will report garbage, no matter how sophisticated his analysis techniques are. Visualization is a great way to identify garbage and will often reveal a surprising structure the analyst would have otherwise not taken into account.

- Data visualization can be a more intuitive summary of data than many statistical methods and can be much easier for the nonstatistician to analyze. This not only makes it a useful tool for presentation and communication, but also makes it a useful tool for combining domain knowledge and intuition with knowledge emerging from data

- Data visualization can indicate which transformations one would need to undertake in order to use a desired model and can often be used to indicate whether any sort of model should be used in the first place.

Data visualization spans several fields of study and has an extensive literature base. For a general overview of many of the successful visualization methods as well as some of the state-of-the-art methods, see, for instance, [66].

1.2.2 Data Visualization Methods in this Thesis

As mentioned above, data visualization is widely known to be an important area for development and extensive work has been conducted to create new methods and improve existing ones. Nevertheless, there is still room for improvement. In particular: 1. Most current plotting methods merely plot the data as is, and do not take into account the fact that the data itself is a "random" sample, and therefore encapsulates some uncertainty, and 2. Methods for plotting data with many variables can be improved; current methods for plotting multivariate data either lack interpretability or force the interpreter to concentrate on many things simultaneously. These three points are elaborated on below.

Describing the Sampling Variability of a Plot

The first area which needs to be further developed in visualization is the description of the sampling variability of a plot.

The data sets that analysts study are typically random samples from larger populations. Therefore, plots of these data sets depict the sample of data, but do not necessarily depict the larger population. In fact, if the sample size is small enough, the plot for the population could look completely different.

Figure 1-1, which shows four data sets drawn from the same population, illustrates our point. An analyst would only be given one of these four data sets; yet, if he were to rely only on the scatter plot of the data he was given to begin constructing models and making inferences, he would conclude very different things, depending on which one of the four data sets he had. In the case of the upper left scatter plot, he may conclude positive correlation. In the upper right scatter plot, he may infer nonlinearity. Meanwhile, in the bottom two scatter plots, he may respectively infer a negative correlation and a positive correlation with some outliers.

This caveat exists for any report that arises from a data set and hence, confidence intervals are typically reported for various summary statistics like the sample mean, in order to describe their sampling variability. This thesis proposes a new method to extend the idea and implementation of the confidence interval to describe a plot's sampling variability. The method works by using bootstrap methods to generate several plots, which could have arisen if the data had been sampled differently from the population, and then conveying the information given in the collection of plots by carefully selecting a few representative plots in the subset.

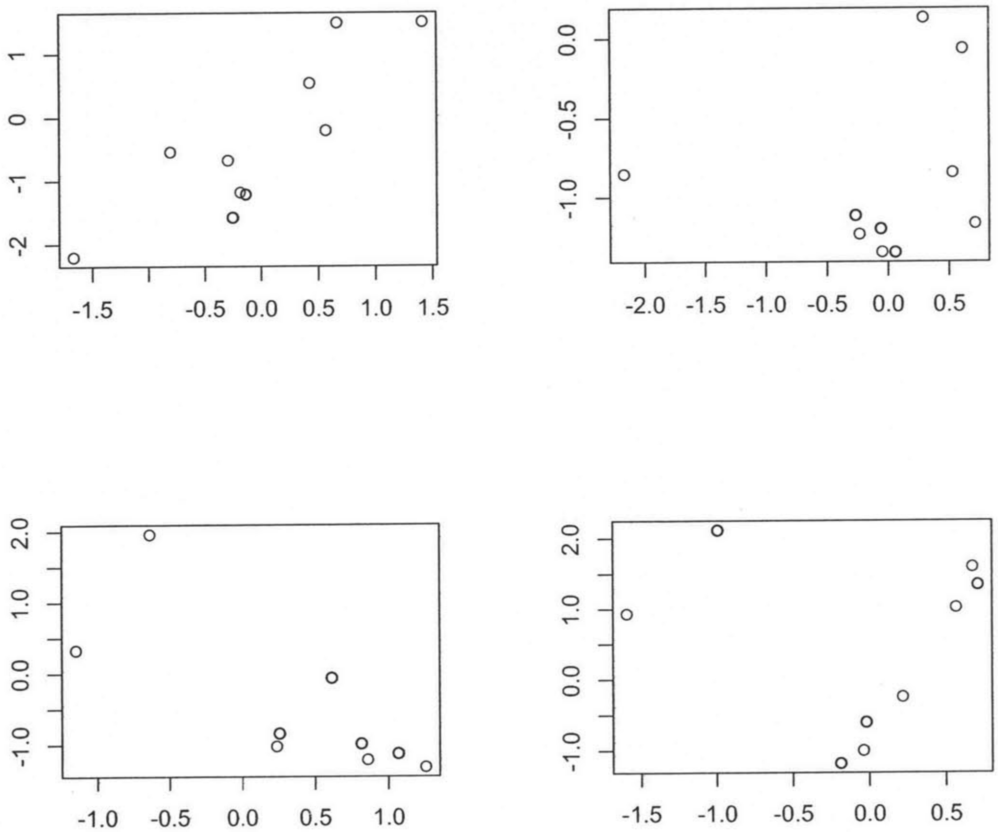


Figure 1-1: Scatter plots of four data sets from the same population

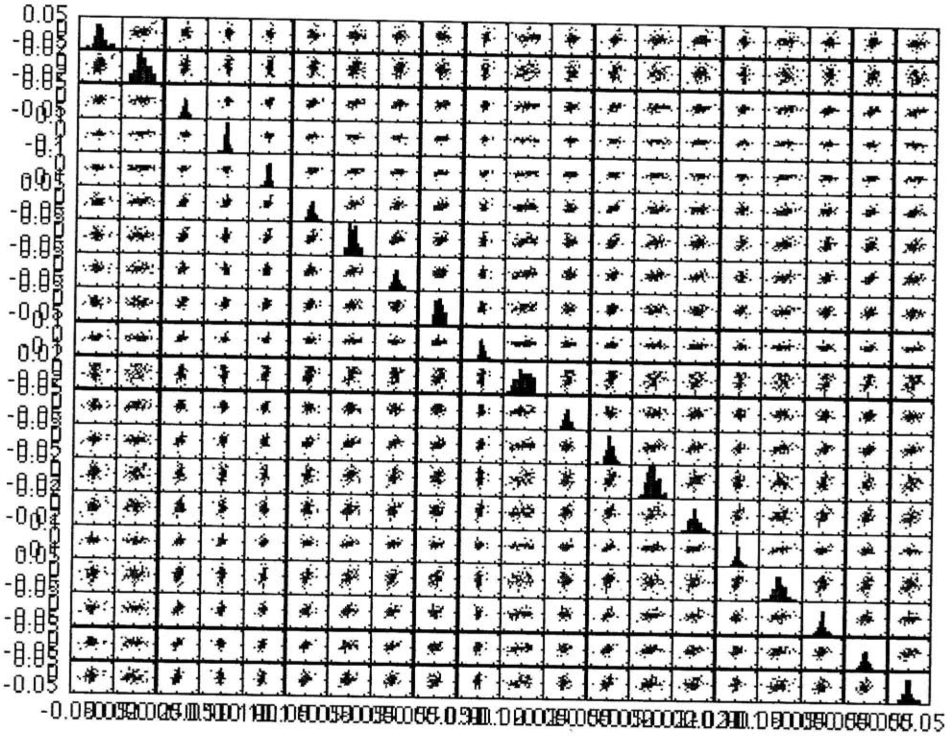


Figure 1-2: A scatterplot matrix of the first 20 variables in a data set of portfolio returns

Large Scatterplot Matrices

The other area, for which improvements would be helpful, is the visualization of high-dimensional data. Many methods currently exist for visualizing high-dimensional data (Chapter 3 mentions several of these), though each has its own limitation, whether it's oversimplification in some sense, undersimplification, or lack of interpretability. The scatterplot matrix is one effective method for displaying high-dimensional data. This method creates a grid of scatter plots and at cell (i, j), plots variable i against variable j. Nevertheless, even scatterplot matrices become too complicated after a given number of variables. See, for instance, Figure 1-2 for a scatterplot matrix of a the first 20 variables of a finance data set with 50 variables, which shows the daily returns of a set of stocks.

In this thesis, we propose a method for simplifying large scatterplot matrices, based on the observation that a few characteristic images often summarize all the images seen in a scatterplot matrix. The method developed in this paper converts a $k \times k$ scatterplot matrix into a smaller $l \times l$ scatterplot matrix where $l < k$, each of the l labels represents a set of variables rather than a single variable, and each cell exhibits a set of similar images rather than a single image. The idea behind the method is that if it is the case that all images in the scatterplot matrix are small variations on a small set of characteristic images, then simplifying the scatterplot matrix by exhibiting only those characteristic scatter plots wouldn't take away much information from the display. Our method uses heirarchical clustering to merge variables together and scatter plot distance measures to represent the diversity of variations on a characteristic image for a given cell.

Our method certainly doesn't solve the difficult problem of visualizing multivariate data by itself, and our method alone could very well miss important properties of a data set. However, like most visualization methods, it can be used in conjunction with other visualization methods.

1.2.3 Diagnostic Methods for Regression in this Thesis

Figures 1-3 and 1-4 provide a good motivation for both diagnostic methods in regression. Figure 1-3, a data set known as Anscombe's Quartet [2], displays four very different data sets which clearly should not be analyzed the same way. Figure 1-4, however, shows that when linear regression is applied to each of the four data sets, not only are the regression lines the same, but the detailed reports outputted by a typical statistical package (which includes an ANOVA analysis, R^2, and coefficient hypothesis tests) are completely identical!

The example these figures depict is naturally contrived, but clearly demonstrates the need to look more carefully at the data we are analyzing. It is this need, which diagnostic methods attempt to fulfill. Diagnostic methods provide reports that display the influence of data characteristics, such as influential observations, outliers, and collinearity (variables that are excessively correlated), on regression inferences. These reports, unlike the simple plots in the above figures, can generalize to cases

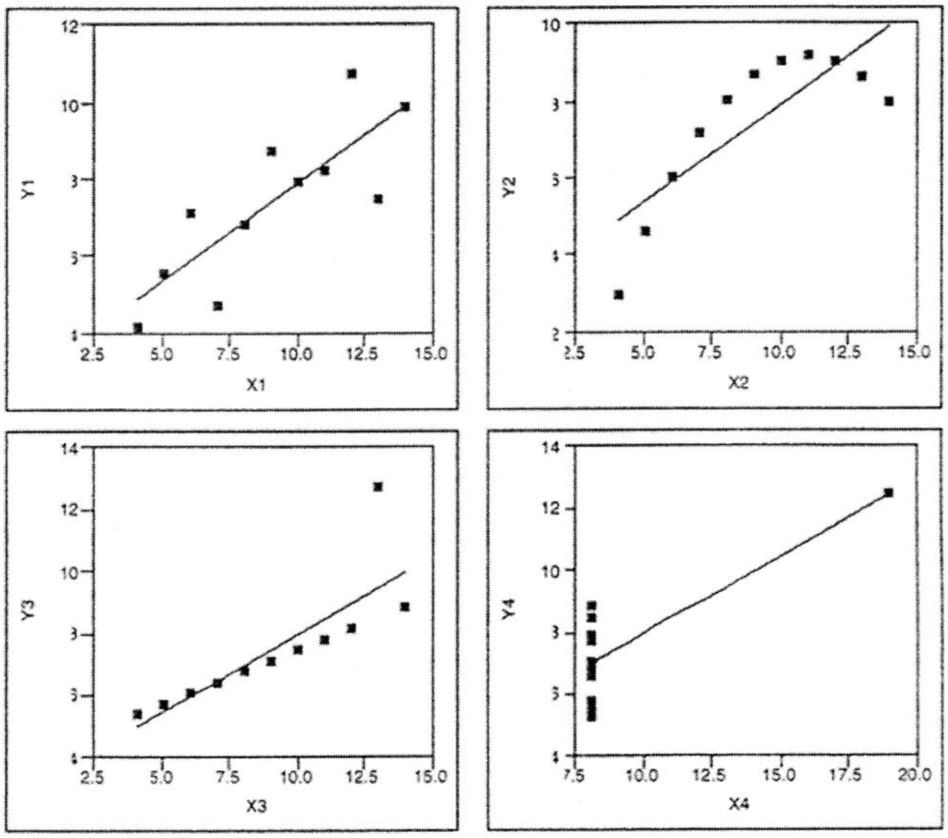
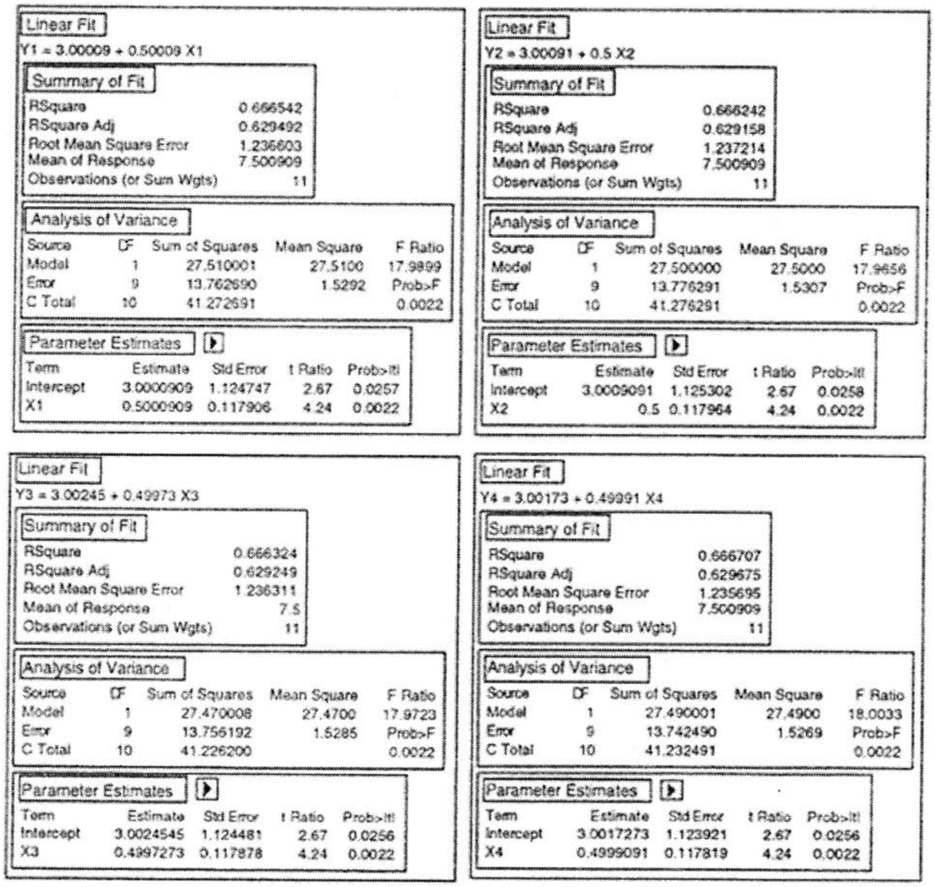


Figure
Regression
Lines for Four
Analysis

Figure 1-3: A plot of four data sets with a regression line superimposed

Figure
Statistical
Reports for
Four Analyses



```
Linear Fit
Y1 = 3.00009 + 0.50009 X1

Summary of Fit
RSquare                        0.666542
RSquare Adj                    0.629492
Root Mean Square Error         1.236603
Mean of Response               7.500909
Observations (or Sum Wgts)           11

Analysis of Variance
Source    DF  Sum of Squares  Mean Square   F Ratio
Model      1       27.510001      27.5100   17.9899
Error      9       13.762690       1.5292   Prob>F
C Total   10       41.272691                 0.0022

Parameter Estimates
Term        Estimate   Std Error  t Ratio  Prob>|t|
Intercept   3.0000909  1.124747      2.67    0.0257
X1          0.5000909  0.117906      4.24    0.0022
```

```
Linear Fit
Y2 = 3.00091 + 0.5 X2

Summary of Fit
RSquare                        0.666242
RSquare Adj                    0.629158
Root Mean Square Error         1.237214
Mean of Response               7.500909
Observations (or Sum Wgts)           11

Analysis of Variance
Source    DF  Sum of Squares  Mean Square   F Ratio
Model      1       27.500000      27.5000   17.9656
Error      9       13.776291       1.5307   Prob>F
C Total   10       41.276291                 0.0022

Parameter Estimates
Term        Estimate   Std Error  t Ratio  Prob>|t|
Intercept   3.0009091  1.125302      2.67    0.0258
X2                0.5  0.117964      4.24    0.0022
```

```
Linear Fit
Y3 = 3.00245 + 0.49973 X3

Summary of Fit
RSquare                        0.666324
RSquare Adj                    0.629249
Root Mean Square Error         1.236311
Mean of Response                    7.5
Observations (or Sum Wgts)           11

Analysis of Variance
Source    DF  Sum of Squares  Mean Square   F Ratio
Model      1       27.470008      27.4700   17.9723
Error      9       13.756192       1.5285   Prob>F
C Total   10       41.226200                 0.0022

Parameter Estimates
Term        Estimate   Std Error  t Ratio  Prob>|t|
Intercept   3.0024545  1.124481      2.67    0.0256
X3          0.4997273  0.117878      4.24    0.0022
```

```
Linear Fit
Y4 = 3.00173 + 0.49991 X4

Summary of Fit
RSquare                        0.666707
RSquare Adj                    0.629675
Root Mean Square Error         1.235695
Mean of Response               7.500909
Observations (or Sum Wgts)           11

Analysis of Variance
Source    DF  Sum of Squares  Mean Square   F Ratio
Model      1       27.490001      27.4900   18.0033
Error      9       13.742490       1.5269   Prob>F
C Total   10       41.232491                 0.0022

Parameter Estimates
Term        Estimate   Std Error  t Ratio  Prob>|t|
Intercept   3.0017273  1.123921      2.67    0.0256
X4          0.4999091  0.117819      4.24    0.0022
```

Figure 1-4: Reports for the four data sets in previous figure

with many explanatory variables. They are, however, difficult to produce for several reasons.

One reason, for instance, is that the presence of each of the characteristics we wish to measure influences our attempt to measure the others. We give a more thorough account of this in Chapter 4, but the plot of $X4$ vs $Y4$ shows a glimpse of this problem itself. The point in the upper right quadrant of the graph defines the entire relationship and is the reason why $X4$ and $Y4$ are perceived to be related. Hence, it would be valuable to discover and document this fact. However, if one ran a linear regression on this data set, the residual of the point would be zero, and so it would not be perceived to be an outlier. Here, the presence of an outlier actually influences our attempt to diagnose it as an outlier.

Forward Search [6] is a recently developed state-of-the-art method, which runs an algorithm through the data set of interest and plots certain measures as the algorithm progresses. The output is a plot, which highlights the effect of individual data points on the regression inference. By using bootstrap sampling, one can also figure out the variability of this plot.

In this thesis, we will extend this method to address not only observation-related diagnostics, but also feature-related diagnostics, by adding a feature selection component to the algorithm.

Forward search, as is, starts with a certain set of points, labeled "clean" points and at each step, increases the size of the set of "clean" points by one. Our algorithm, which we call backward selection search, instead augments the design matrix by appending an identity matrix to the right of the original design matrix, and performs backward selection (greedy, one-at-a-time feature elimination). Eliminating a feature of the original design matrix is equivalent to increasing the size of the set of "irrelevant" features by one, and eliminating a feature of the appended matrix is equivalent to increasing the size of the set of "clean" points by one. Hence, the result of our algorithm is an extension of forward search to simultaneous feature selection and outlier detection, as well as a new method for robust model selection.

The idea for our algorithm also opens the door for other developments. First of all, we can easily incorporate prior knowledge and constraints into the algorithm. Additionally, the idea behind the algorithm implies a general methodology for detecting outliers in multivariate data (rather than merely regression data) via multivariate linear model feature selection methods on an augmented matrix.

1.3 Organization of the Thesis

The rest of this thesis is organized as follows: Chapter 2 describes the method for visualizing the sampling variability of plots, Chapter 3 describes the method for simplifying scatterplot matrices, Chapter 4 describes backward selection search, and finally, Chapter 5 concludes and discusses directions for future work.

Chapter 2

Visualizing the Sampling Variability of Plots

In this chapter, we give a fuller account for the general method for providing a description of the sampling variability of a plot of data, which was sketched out in the introduction.

As mentioned in the introduction, the data sets statisticians analyze are samples from larger populations of interest in many cases. Therefore, when a plot is made of a given data set, it must be understood that the plot could have looked entirely different if the data had been a different sample from the population of interest. This caveat exists for any report that arises from a data set, and hence confidence intervals are typically reported for various summary statistics like the sample mean, in order to describe their sampling variability. The goal of this chapter is to extend the idea and implementation of the confidence interval to describe a plot's sampling variability.

In the frequentist realm, this makes plots more valid as inferential tools. Meanwhile, in the Bayesian realm where the data is fixed, the procedures for conveying the information given in several plots through a few representative plots can help verify models by indicating if the data are a reasonable draw from the posterior, and indicating the way that the data differs if they are not a reasonable draw from the posterior.

As will be further discussed in the literature review, attempts have been made to do this for certain specific types of plots. These methods typically involve resampling from the data (via bootstrap methodology) and then overlaying the bootstrap samples on top of each other. The key limitation of previous literature is that the methods don't apply to some of the plots which are most often used in practice such as histograms, scatter plots, and scatterplot matrices. Indeed, many common plots can not be sensibly drawn on top of each other. In this chapter, we aim to produce a method which works even when it is not possible to overlay plots on top of each other and where the bootstrap samples represented are chosen in a somewhat systematic manner.

We create and represent the confidence interval in three steps, which we describe in more detail in section 2.2. The first step involves taking k (a large number) bootstrap samples from the data set and creating plots of each of these samples. The second step involves computing distances between each scatter plot pair, and letting the envelope consist of the $(1 - \alpha) \times k$ plots which are most similar to a central plot, where $(1 - \alpha)$ is an approximate confidence level between 0 and 1. Lastly, the final step consists of representing the many plots in this envelope. In section 2.2, a few informative methods for doing this are proposed.

The method includes the capacity to incorporate distributional assumptions and can be implemented for a large variety of plots. We focus on the method as it relates to the case of a scatter plot. However, we also present extensions to other types of plots.

The rest of this chapter is organized as follows: Section 2.1 provides a literature review. Section 2.2 describes our method more thoroughly. Section 2.3 provides a few illustrations. Lastly, Section 2.4 concludes with a discussion about the contributions of this paper and the directions for future research.

2.1 Literature Review

This section presents background material and related literature. The background material discusses bootstrap procedures and the Earth Mover's Distance. Bootstrap procedures are related to our method because we use them to complete the first step, and because reasonably similar work has been done with them. The Earth Mover's Distance is a distance measure between histograms. We cover this below because we view scatter plots as bivariate histograms and use the Earth Mover's Distance to find the distance between two scatter plots.

2.1.1 Background

Bootstrap Procedures

Bootstrap procedures were introduced in [14] and have been used extensively since then to estimate standard errors, confidence intervals, and sampling distributions for statistics of interest, in cases where analytical methods would be too cumbersome. A thorough review of the bootstrap procedure is given in [15].

A brief description of the procedure is as follows:

1. Assume a distributional form for the population. In the nonparametric case, the assumed distributional form would be that the population can only take the n values that the data set takes, and it takes those n values with certain probabilities.

2. Estimate the parameters for the assumed distributional form, usually by maximum likelihood. In the nonparametric case, one gets that the resulting estimated population distribution is the empirical distribution, where the population can only take the n values that the data set takes and it takes each of those with probability $\frac{1}{n}$.

3. Draw a large number of samples from the estimated population distribution, and for each of those samples, compute the statistic of interest. In the nonparametric case, drawing samples from the estimated population distribution corresponds to sampling the data with replacement.

The several values that the statistic of interest took in the bootstrap samples are samples from its approximate sampling distribution and can be used to compute standard errors, confidence intervals, and to visualize the sampling distribution.

Earth Mover's Distance

The Earth Mover's Distance, introduced in [50], measures the distance between two histograms (or in the more general case, distributions). It is used frequently in the field of content based image retrieval.

It is called the Earth Mover's Distance because if each histogram is viewed as a pile of dirt, the Earth Mover's Distance between the histogram is the minimal cost of turning one pile of dirt into the other, where the cost is assumed to be the amount of dirt moved times the distance by which it is moved.

The optimization problem that needs to be solved is a transportation problem: in particular, an uncapacitated minimum cost flow problem, whose computation time is $O(n^3 log(n))$, where n is the number of bins in the histogram [35]. There are, however, approximate solutions whose computation time is linear in n such as in [60].

The Earth Mover's Distance satisfies the properties of a distance metric and has additional properties, such as an equivalence to Mallow's Distance [39]. Additionally, it has been shown to have very good empirical performance (in terms of agreeing with measures of distinction based on the human eye) [58].

2.1.2 Related Literature

With the exception of Gelman (2004) [18], who makes a case for using plots as inferential tools in Bayesian analysis, we are not aware of any literature suggesting the inferential nature of basic plots. Additionally, we are not aware of any methods for assessing the sampling variability of plots of the data in its raw form, which is the focus of this paper. However, various attempts have been made to assess the variability of plots which evolve from functions of the data. In these cases, the general procedure is to resample from the data, create several of the plot objects (usually 20-30) and then find a way of representing the plot objects, such that they can be

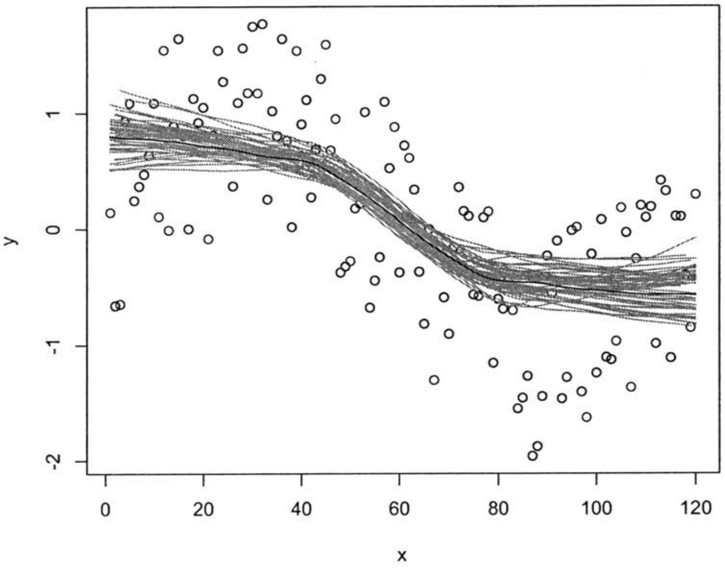


Figure 2-1: The bootstrap sampling variability of a loess curve

layered on top of each other.

One well-known example of this is representation of the sampling variability of the loess curve, a nonparametric regression curve formed by the data (see [21] and [22] for more details). Figure 2-1 shows an example. Here, the original loess curve for the data are plotted in black and the bootstrapped curves were overlaid in red on the same plot to show the sampling variability of the curves.

This procedure has also been used for the representation of the sampling variability of classification and regression trees (see [64] for more details). In addition, it has been used to describe the sampling variability of principal components plots (the projection of the data onto a plane which describes much of the variation). In these representations, the projections of various points emerging from different bootstrap samples are overlayed on top of each other and ellipses are drawn for each observation, which contain 95% of its realizations [11]). Kiers and Groenen (2006) [31] note that this representation lacks the ability to detect how different points depend on each other, and hence created a movie through the representations instead.

Our methodology is related to the literature above, but distinguishes itself in the types of plots, whose sampling variability it tries to assess. It also assesses and represents sampling variability in a way, which doesn't force one to restructure the representation of the plot object so that multiple plot objects can be overlayed on top of each other, thereby allowing one to generalize to other classes of plots.

2.2 Methodology

In this section, we flesh out the details of the three steps which were outlined in the introduction in the case of a scatter plot and discuss the incorporation of distributional assumptions and prior knowledge.

2.2.1 Step 1

In step 1, we take k (a large number) bootstrap samples from our data set and form a plot with each of these samples, so that k plots are produced. Where n is the size of the data set, each bootstrap sample is a sample of size n from an estimated distribution.

This estimated distribution is created by assuming a distributional form for the data generating process and then estimating the parameters of the distributional form, usually by maximum likelihood. In the nonparametric bootstrap, the assumed distributional form is simply discrete with a probability p_i that the sample observation will take the value of observation i in the given data set, and the parameters are $p_1, p_2, ..., p_n$. In the case of no prior knowledge, maximizing the likelihood for the nonparametric case yields $p_i = \frac{1}{n} \forall i$.

If prior knowledge exists, we merely modify our estimate of the distribution parameter and then proceed as before. The parameter estimate can be modified by replacing the estimate, which doesn't incorporate prior knowledge, with the posterior mean or the posterior mode.

As an example, if we were to make no parametric assumptions and impose the prior knowledge that the mean of the first variable, X_1, is less than or equal to 0.01, we could solve the following optimization problem in order to give us the posterior mode when the prior is uniform: $\max \sum_{i=1}^{n} \log(p_i)$ subject to the constraints: $\sum_{i=1}^{n} p_i = 1$, $p_i \geq 0, \forall i$ (in other words, the values p_i are valid probabilities), and $\sum_{i=1}^{n} p_i X_{1i} \leq 0.01$ (our prior knowledge is imposed), where x_{1i} represents the realization i of variable X_1.

2.2.2 Step 2

In this step, we create a "distance metric", which describes the distance between two plots. We use the created distance metric to create a k by k matrix of each plot's distance from each of the other plots. Following this, we define a central plot as the plot whose summed distances to other plots is minimized. Lastly, we collect the $(1 - \alpha) \times k$ plots which are closest to the central plot and let these be the plots in our confidence envelope of interest. It is this envelope of plots which we seek to visualize in the next step.

To find the distance metric between two scatter plots, we treat them as bivariate histograms, where each point in a scatter plot represents a bar of height 1 in a bivariate histogram (see Figure 2-2 for a picture of what such a bivariate histogram with five points would look like). We then use the Earth Mover's Distance, which was described in the literature review, between histograms to find the distance between two scatter plots.

The Earth Mover's distance metric is suggested because it is intuitive and has interesting properties as mentioned in the literature review. In addition, it doesn't require us to specify what we will be looking at before we see the plot, and it can be used for both parametric and nonparametric inference. The Earth Mover's distance metric also has demonstrated good empirical performance in previous literature. We demonstrate the types of plots the Earth Mover's Distance reports as similar using a couple of examples in the appendix.

In our case, the Earth Mover's distance simplifies so that one can solve it using an "Assignment Problem", a class of problems which can be solved very efficiently using, for instance, the Hungarian Algorithm [36]. The name of the class of problems refers to its applicability in the case where there are n agents, who need to be assigned to

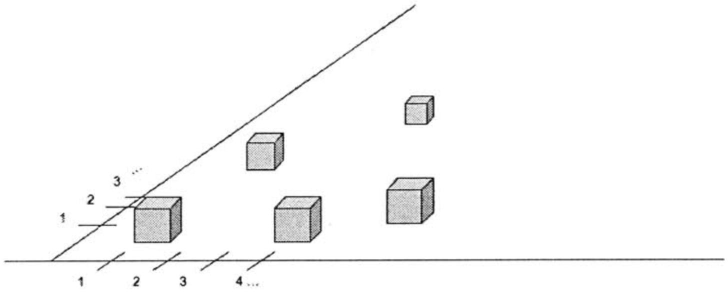


Figure 2-2: A bivariate histogram with five points.

n tasks in a way that minimizes the cost emerging from the assignments.

Even computing the Earth Mover's distance via an assignment problem can get time consuming when n and/or k is large, since k^2 computations need to be made. One remedy to this is to instead view the scatter plot as a regular histogram over the variable-space shown in the scatter plot, where each bin in the histogram is one of the squares in a 10×10 grid, which partitions the space. In this case, one will need to solve the usual Earth Mover's distance problem, but the problem size would never exceed 100 even when there are more data points, because the Earth Mover's distance's complexity is a function of the number of bins, rather than the number of data points.

In order to find the distance between other types of plots in general, the distance metric needs to be modified. However, the Earth Mover's distance itself generalizes to several types of plots. As we mentioned, the Earth Mover's distance simplifies to an "Assignment Problem" in the case of a scatter plot. In one dimension, where we are finding the distance between histograms with equal bin length, it simplifies even further to merely sorting two data sets and then finding the summed distance between the quantiles.

The Earth Mover's distance can also be used to find the distance between plots of multidimensional data. For parallel coordinate plots, we just compute an assignment problem, like in the case of scatter plots. For scatterplot matrices, one can merely use the summed distance between each pair of scatter plots in the matrix. Meanwhile, for biplots, one can add the distance between the point representations on the two plots and the distance between the variable representations.

Another interesting case where the Earth Mover's distance can be used is in the case where certain observations are tagged. This could occur, for instance, in a case where we would want to visualize two different groups of data and see how they interact with each other. In this case, one could use the Earth Mover's distance in exactly the same form as before, but merely change the point-wise distance to include whether there is a change in whether a point is tagged.

In some cases, such as the box plot and the dendrogram, the Earth Mover's Distance doesn't directly generalize. However, intuitive distance measures are not too difficult to develop in those situations. For instance, in the case of a box plot, an intuitive distance measure is the sum of the distances between the lower quartiles, the upper quartiles, the lower whiskers, and the upper whiskers of two box plots. Meanwhile, in the case of the dendrogram, an intuitive distance measure is the sum of the absolute differences in the heights at which each pair of points is joined into a cluster. We don't explore these specific plots in this chapter, but do give examples beyond the case of a scatter plot.

2.2.3 Step 3

In this step, we attempt to represent the multitude of scatter plots in the confidence envelope. Given that we have distances between each pair of plots, this can be done in several different ways. In the illustrations given in this chapter, we merely report the two plots which are farthest from one another for simplicity. This gives a sense of the "border" of the set of $(1-\alpha) \times k$ plots in the confidence envelope.

Nevertheless, several other summaries exist. A representation of all the plots in the envelope would involve ordering them in such a way that the distance between each plot and the adjacent plots is minimized. The problem of ordering the plots can be formulated as a case of the traveling salesman problem, an optimization problem which finds the tour through a given set of cities with the smallest distance (see [38] for more details). Several heuristics exist, which effectively do this.

Other alternatives include clustering methods and the use of multidimensional scaling. Clustering methods would cluster the plots based on their distance metric and then report the plots that correspond the the cluster centers. Multidimensional scaling, meanwhile, would start with a matrix of object to object dissimilarities (in this case the objects are plots), and then would assign each object a location in $N-$dimensional space. The resulting locations can be displayed graphically to understand the space of plots. More information about multidimensional scaling can be found, for instance, in [9].

2.3 Illustration

In this section, we apply methodology to four examples involving scatter plots and two examples involving other types of plots.

2.3.1 Scatter Plot Examples

We apply the methodology on four scatter plot examples in this section in order to demonstrate the output of the method. In our examples, we use In this example, we use $k = 1000$ and $\alpha = 5\%$. Our first illustration uses a sample of 10 points in two variables and exhibits no relationship. Figure 2-3 shows a scatter plot of the entire population with the sample filled in.

In order to get a sense of the variability, the two bootstrap plots, in the confidence envelope of $(1 - \alpha) \times k$ plots, which are farthest apart from each other are shown in Figure 2-4. As mentioned in the figure caption, in this and other plots in the chapter, a jitter is added, so that one can see each individual point. One will notice that the two plots appear quite different from each other and the original sample.

We contrast this with an illustration of a case where a scatter plot consists of 40

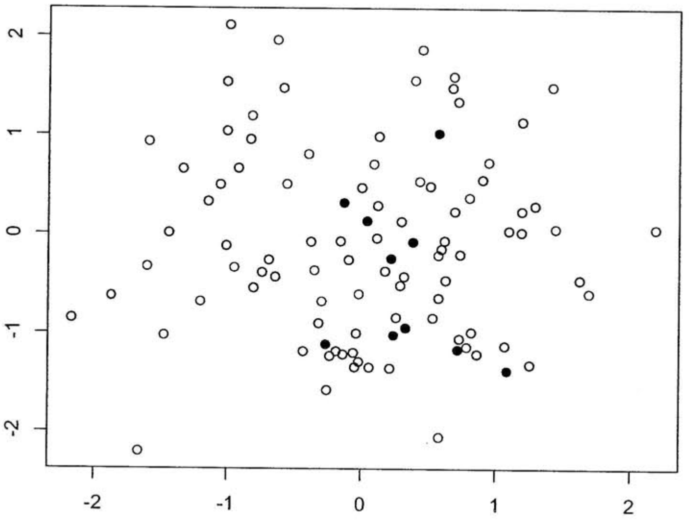


Figure 2-3: A scatter plot of a population of 100 observations. A sample data set of 10 observations are filled in.

Figure 2-4: The sample of 10 points (left) and the two bootstrap sample scatter plots farthest from each other. A jitter is added, so that multiple points covering the same space can be seen

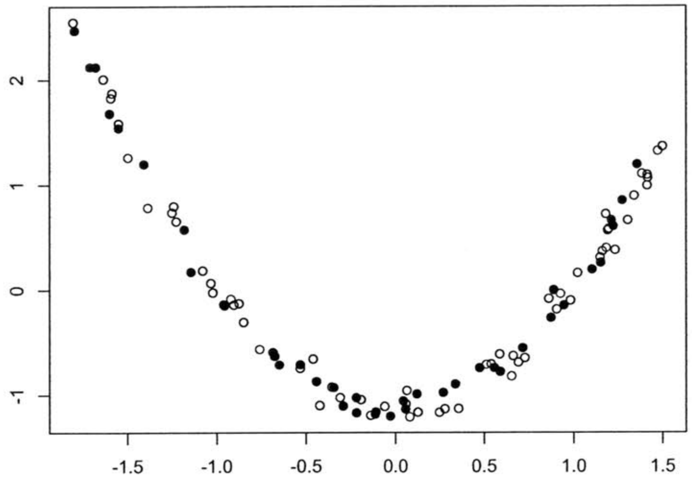


Figure 2-5: A scatter plot of a population of 100 observations. A sample data set of 40 observations are filled in. A jitter is added, so that multiple points covering the same space can be seen

points and exhibits a clear relationship. Figure 2-5 shows the sample embedded in the population and Figure 2-6 shows the two bootstrap plots in the envelope which are farthest apart. In this case, the two scatter plots which are farthest apart still tell a similar story to each other and to the original plot, in part because of the sample size, and in part because the relationship is more clear.

In our third and fourth scatter plot example, we illustrate two interesting ways of using our method for data analysis. Our third example uses the Hertzsprung-Russell Star Data [57], which is discussed in more detail in Chapter 4. Its plot is shown in Figure 4-1. The data set consists of four gross outliers that lie to the upper left of the plot.

Figure 2-7 shows the two extreme plots outputted by our method on this data set. There is mild variability in the relationship between log light and log temperature, but the fact that there is a relationship is clear.

In contrast, had one tried to assess the variability of the relationship based on the correlation coefficient, one would have obtained very different results. Figure 2-8

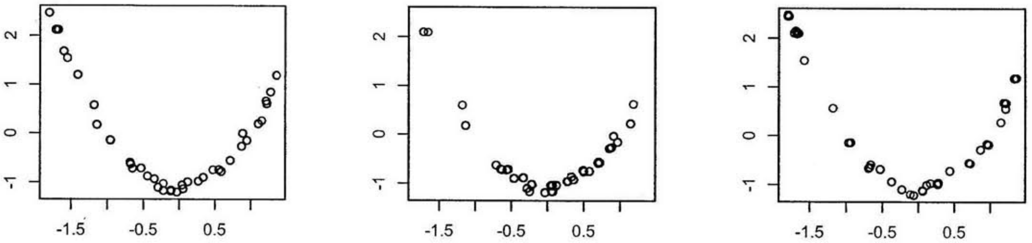


Figure 2-6: The sample of 40 points (left) and the two bootstrap sample scatter plots farthest from each other

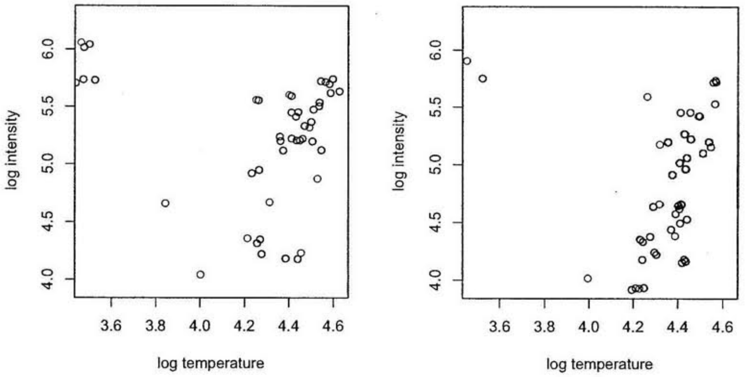


Figure 2-7: The two bootstrap plots farthest from each other.

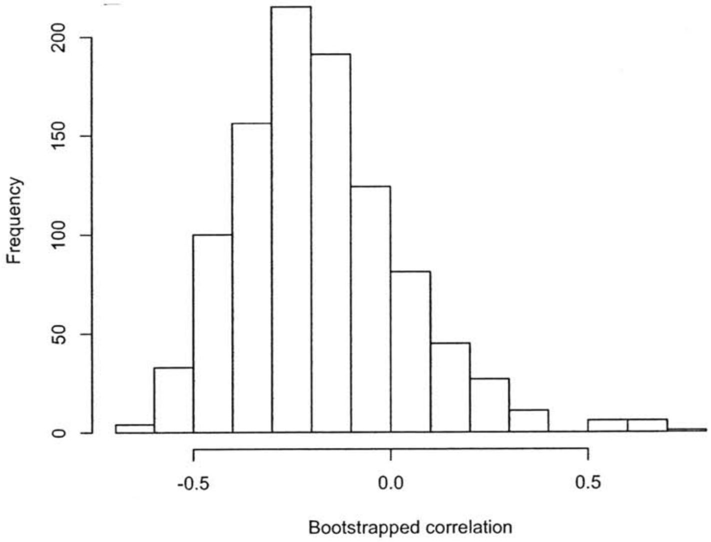


Figure 2-8: The bootstrapped sampling variability of correlation.

shows a histogram of the sampling variability of correlation. Note that the standard error is huge and much of the bulk of the histogram lies below zero. The reason the correlation behaves this way is that the four outliers to the upper left of the plot reverse the sign of the correlation between the majority of the points.

In this specific case, an informed statistician could have successfully used robust correlation to find the variability of the relationship. However, the point of this illustration is to show that our method has the advantage that one need not specify a possibly erroneous numeric quantity to understand the relationships in the data and their variability.

Our fourth example illustrates an entirely different problem in data analysis which our method enables one to address. Figure 2-9 shows the plot from our example, which emerges from a microarray data set [12], consisting of 38 observations and 3,051 variables. The plot shows the expressions of two genes (the x and y axes) for Leukemia patients (in white) and healthy patients (in black).

The plot seems to indicate that two genes very clearly distinguish patients with Leukemia from healthy patients, but one could argue that a data set consisting of

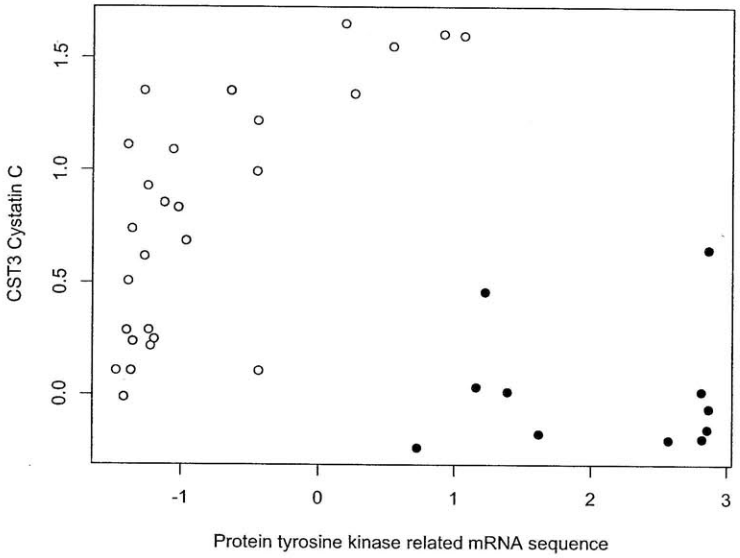


Figure 2-9: The expression levels of two genes (the x and y axes) for Leukemia patients (in white) and healthy patients (in black)

random noise would always contain some pair of variables which exhibited an interesting relationship, if the data set contained enough variables. This is an issue in data analysis in general, because the plots given in papers are often the most interesting plots of a group of different plots that were inspected.

In our case, there were 3,051 variables, which implies that about 4.5 million bivariate plots could have been considered before this plot was chosen. Hence, the argument above certainly applies to our data set.

In this illustration, we use our method to inspect whether we should be concerned about the fact that there were 3,051 variables in the data set.

For a set of 1000 iterations, we permuted the labels, which indicate whether or not a patient had cancer (this creates a resampled data set that is, for our purposes, random noise). In each iteration, we use forward selection to select the two variables which will yield the most interesting plot (we use forward selection because that was the heuristic which outputted the variables exhibited in the plot in Figure 2-9). We then used steps 2 and 3 of our method to summarize the plots arising from the

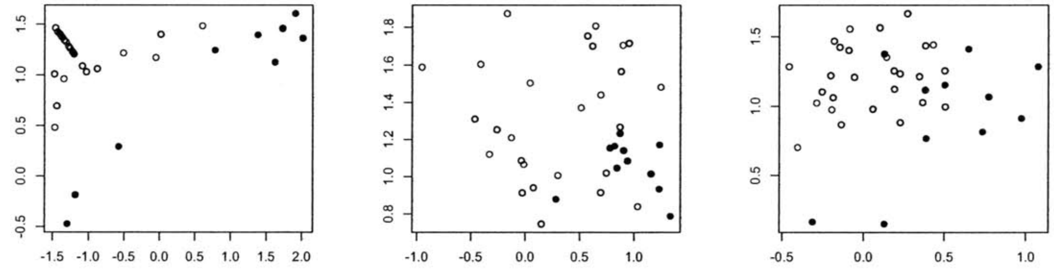


Figure 2-10: The two plots farthest from each other (left and center) and the central plot (right), among those in the confidence envelope for the set of plots arising from the permutation iterations

iterations.

Figure 2-10 shows the output of our method. Indeed, any of the three plots in the figure would seem to exhibit an interesting relationship if presented on its own. Hence, it is clear that there is some optimism in Figure 2-9. However, the relationship in 2-9 still stands out, despite the optimism. Our methodology therefore allows one to search through several plots to find interesting plots to report while remaining statistically sound.

It should be noted that our methodology for this illustration was slightly different than that for the previous illustrations, though the underlying idea is the same. We used permutation sampling instead of bootstrap sampling for this example and the distance metric was modified. Since each scatter plot, among those created in the iterations, represents a different pair of variables, the distance method we used between two scatter plots first put the scatter plot on the same scale, then made sure the signs were aligned properly, and then used the Assignment problem with the constraint that a black point could not be assigned to a white point and visa versa.

2.3.2 Other Examples

In this section, we give two examples outside the realm of the usual scatter plot in order to demonstrate the versatility of our methodology and to show that for some plots, even the astute statistician who can eyeball the sampling variability of a typical plot will struggle to assess the uncertainty of an image.

The data for the first example are simulated from the Barrow Wheel Benchmark [40], which is a challenging synthetic data set, often used to evaluate robust covariance estimation methods. It is described more fully in Section 4.5, and Figure 4-14 shows a plot of the data set when it was generated with 2 variables and 50 observations.

Figure 2-11 shows four plots emerging from this data set, when it was generated with 6 variables and 50 observations. The plot in the upper left shows the biplot of the data set. A biplot essentially projects a data set onto the hyperplane defined its first few principal components. The principal components are the eigenvectors of the covariance matrix of the data, and the first few principal components define the hyperplane which explains the most variability of the data. The biplot projects both the observations and the variables onto this hyperplane and can be used to assess the structure of the data, as well as the relationship between latent factors, which characterize the resulting principal components.

We applied our methodology to the biplot, computing the distance only as a function of the points for simplicity. The upper right of Figure 2-11 shows the central plot. Meanwhile, the two plots at the bottom show the two extremes. As is shown, they look completely different from both the original plot and each other. The representation of the variables, which would usually suggest their correlation to each other and their relevance, is also entirely different.

The data in the second example contain 50 industries among the MSCI US Equity Indices. Returns are daily, beginning 01/03/1995 and ending 02/07/2005. In our example, we use the first 100 of these days to compute the portfolio weights that maximize the Sharpe ratio and then apply these portfolio weights for the next 100 days and plot the histogram of the returns on these days to judge the distribution of future returns, given the investment strategy. The methodology used to find the portfolio weights which maximize Sharpe ratio is the methodology proposed by Markowitz, who initiated the mathematical framework for portfolio optimization [41].

Figure 2-12 demonstrates the sampling variability of such a histogram, due both

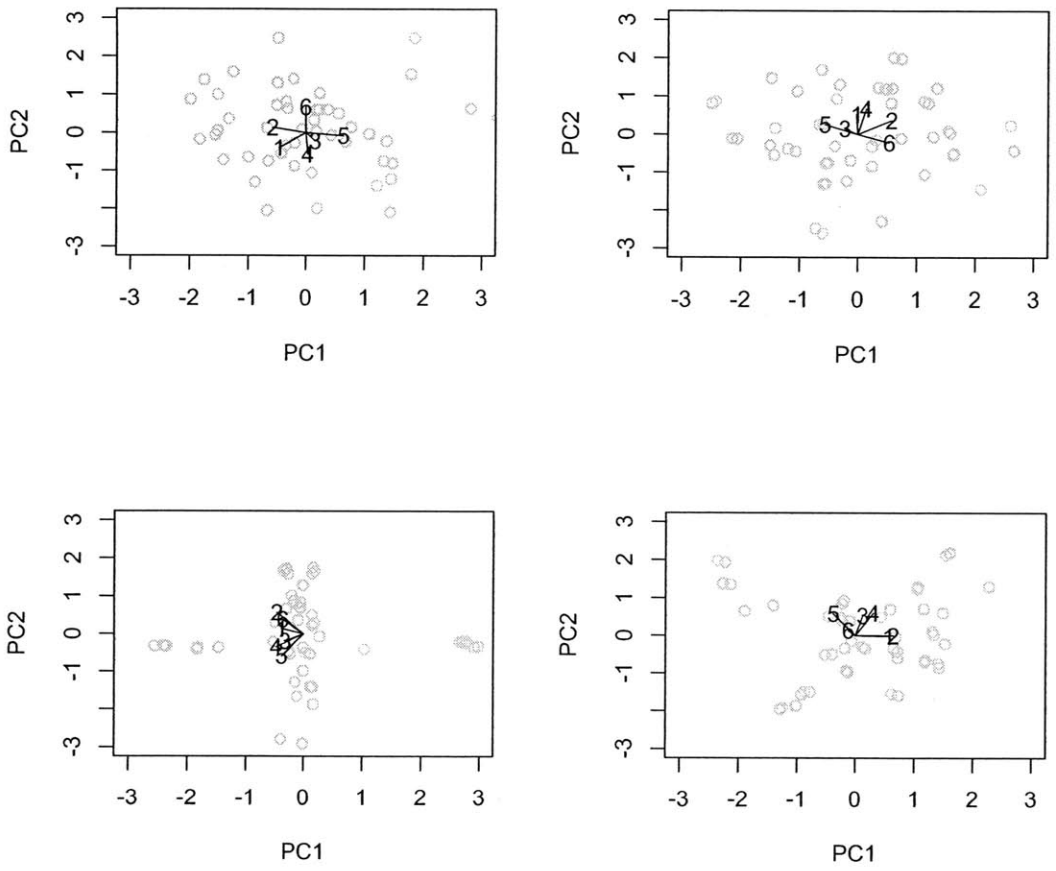


Figure 2-11: Biplot of the original data set (upper right), central biplot (upper left), and the two extremes (lower two plots) for the Barrow Wheel Benchmark

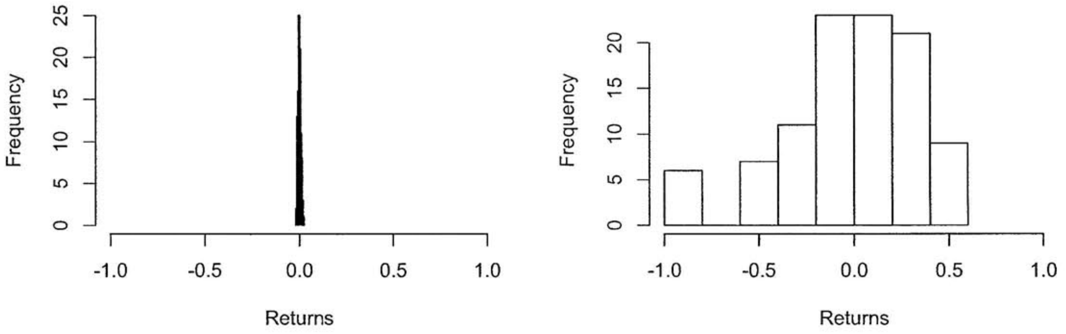


Figure 2-12: The two extremes for return histograms with maximum Sharpe ratio portfolios

to the sampling variability in the first 100 days used to determine the investment strategy and the sampling variability of the next 100 days. The two histograms in the figure show the two extreme histograms based on our procedure. Not only is the variability of the distribution, as represented by the histogram, entirely different, but the histogram in the right panel is skewed to the left. The analysis demonstrates that one must be careful when using the highly variable portfolio weights returned by portfolio optimization to invest in assets.

2.4 Discussion

In this chapter, we developed a way to visualize the sampling variability of a plot, which has the capacity to help all data analysts, from the novice analyst without a statistical background, to the expert statistician who may be able to picture the variability of a simple plot such as a scatter plot of data, but not the variability in a more complicated plot such as a biplot. Additionally, although we have not demonstrated this in the illustration section, our methodology allows an analyst to make inferences in the reverse direction. If we were interested in a Bayesian paradigm, we would assume that our data is fixed (and hence does not have sampling variability). However, we would eventually produce a predictive distribution from the Bayesian model. In order to test whether our data, which has a finite sample size n strongly differs from that predictive distribution in some particular way, one would need to understand what typical samples of size n would look like (the maximum likelihood sample of size n would be n observations at the mode, which would produce a false inference). Our methodology provides a way to do that.

This paper opens the door for many future research topics. Such topics include an analysis of the extent to which incorporating prior knowledge is useful in the frequentist nonparametric bootstrap and an empirical study of the extent to which plots in the envelope are closer to the population plot than plots outside the envelope. In addition, there are many limitations and biases that bootstrap methods produce, which have not been addressed in this paper. One, for instance, is that bias corrections can produce better bootstrapped confidence intervals for many types of point estimates. We would expect this to be the case for scatter plots as well.

Chapter 3

Simplifying Large Scatterplot Matrices

As mentioned in the introduction and demonstrated in Figure 1-2, scatterplot matrices can be notoriously difficult to interpret for high dimensional data, in part because of the sheer size of the grid they produce, and in part because of the amount of information they present.

The goal of this chapter is a method for simplifying scatterplot matrices, which sacrifices some additional information, but in exchange, produces a much smaller grid. Figure 3-1 depicts the objective of our method. We wish to convert a $k \times k$ scatterplot matrix into a smaller $l \times l$ scatterplot matrix where $l < k$, each of the l labels represents a set of variables rather than a single variable, and each cell exhibits a set of similar images rather than a single image.

The rest of this chapter is organized as follows: Section 3.1 provides a literature review, Section 3.2 describes the methodology used to simplify the scatter plots, Section 3.3 provides an illustration on a simulated data set, and Section 3.4 concludes.

3.1 Literature Review

As a review, a scatterplot matrix consists of a square grid of cells, where the cell with row index i and column index j exhibits a scatter plot of variable i against variable

j. The scatterplot matrix only allows the analyst to view one and two-dimensional relationships, thus sacrificing the ability to view higher order relationships between variables. However, the benefit is a display which is much more interpretable than displays which try to visualize all of the variables simultaneously.

As mentioned before, however, even visualizing a full scatterplot matrix itself can be cumbersome if the dimensionality is high enough, sacrificing considerable interpretability. A variety of different methods have been brought forth to address this, each sacrificing some of the information of a full scatterplot matrix.

One popular methodology is to perform a scatterplot matrix of a few factors which describe a large part of the variability of the original data. The factors are often assumed to be the first few principal components of the data matrix (see [28], [34], [37], [13], [47]). When the factors themselves are interpretable, this can be very useful. However, in many cases, the principal components from the data set, or other factors derived as a function of the data set, may not be easily interpretable. Additionally, in many of these cases, the factors are restricted to be linear combinations of the variables.

Another interesting methodology is scagnostics (scatterplot diagnostics), introduced in [65] and further developed in [63]. Scagnostics considers a set of characteristics of a scatter plot, such as whether it's monotonic, whether it's convex, whether there are outlying points, and etc., and then outputs a scatterplot matrix where each variable is one of the characteristics and each plot in the original scatterplot matrix is a point in the scatterplot matrix of characteristics. This method involves the specification of characteristics and is quite useful when detecting the unusual graphs in a collection of graphs.

Nevertheless, surprisingly few methods have been brought forth to simplify the presentation of the scatter plots in the original variables, which is desirable because the original variables are typically much more interpretable than functions involving several of them. The methods developed have generally dealt with choosing an optimal ordering for the variables.

The methods most similar to the remedy suggested in this paper are [27], which

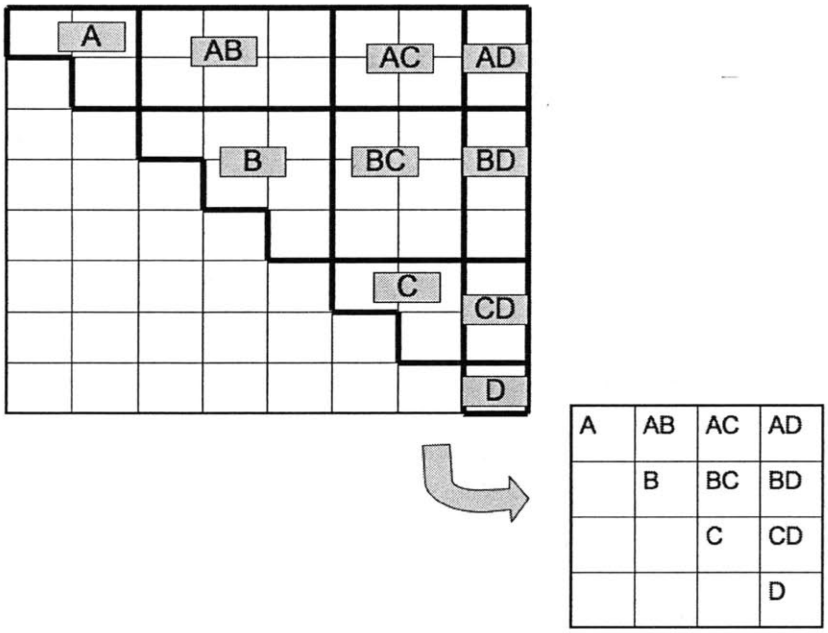


Figure 3-1: The process of simplifying a scatterplot matrix

tries to order the variables so that the most interesting panels are placed in prominent positions near the diagonal, and [1], which attempts to place similar variables close to each other in related displays. These methods can simplify the display considerably without sacrificing additional information, but in many cases, the sheer size of a scatterplot matrix can still make these methods of limited use.

The idea behind the method is that if it is the case that all images in the scatterplot matrix are small variations on a small set of characteristic images, then simplifying the scatterplot matrix by exhibiting only those characteristic plots wouldn't take away much information from the display. Our method includes the capacity to incorporate some types of domain knowledge (such as which variables should and should not be in the same cluster) and has the ability to display its own sampling variability.

3.2 Methodology

Let k be the number of variables in the data set and $l < k$ be the desired length and width of the grid in the simplified scatterplot matrix. The larger l is, the less information is lost, but the more difficult the display is to interpret. Our method first determines how exactly clusters of plots of the $k \times k$ scatterplot matrix should be placed into the cells of the $l \times l$ grid, and then determines how the multitude of plots in a single cell of the $l \times l$ grid should be expressed. Our method has the capacity of incorporating certain types of domain knowledge as well as the capacity to incorporate the sampling uncertainty of the data.

The first step of our method uses heirarchical clustering to cluster the variables into l groups. Heirarchical clustering is a simple clustering algorithm which starts with each observation as a separate cluster. At a given step, the two clusters which are closest are merged until there is only one cluster. Various statistics are recorded at each stage and are summarized into a diagram called a dendrogram. Heirarchical clustering is chosen because the dendrogram allows the user to more intuitively choose the parameter l and to understand whether there is enough clustering to merit the use of the method in this paper to begin with.

Heirarchical clustering requires the specification of a distance measure between the variables that we are clustering. In this case, the distance measure we use, between a variable s and a variable t, is the sum of the distances between the plot of s vs v and the plot of t vs v across all variables v. The distance measure between two plots, meanwhile, first scales the variables in the plots, by subtracting the mean and dividing by the standard deviation, and then outputs the Earth Mover's Distance between the two scaled plots. This metric is chosen because it allows variables that are unrelated to everything else to be clustered together.

Given l clusters, the next step is to figure out exactly what should be displayed in cell i, j of the smaller $l \times l$ scatterplot matrix. Assuming that cluster i contains k_i variables and cluster j contains k_j variables, cell i, j will contain a visual summary of $k_i \times k_j$ plots.

In order to construct possible visual summaries, we first find the distance between each pair of plots. We again use the distance measure developed in the previous chapter here. The variables in both plots are first scaled.

At this point, we have a few options for what to put in cell i, j. One is to present the plot, among the $k_i \times k_j$ plots, whose summed distance to all other plots is minimal. This could be considered the central plot. Another is to present the two plots, which are farthest from each other. This gives a notion of the "border" of the set of plots. A final possibility is to create an animation through the $k_i \times k_j$ plots. In order to make the animation smooth, we need to order the frames in such a way that each frame is as close as possible to its neighboring frames. As mentioned in the previous section, the optimization problem which solves this permutation problem is the "Traveling Salesman Problem", for which many reasonable heuristics exist [38].

This completes the description of the procedure, though it should be noted that our method has the capacity of incorporating certain types of domain knowledge, as well as sampling uncertainty. The domain knowledge that can be incorporated includes claims on which variables are believed to be related and which variables are of particular interest. If variables u and v are believed to be related, one can change the clustering process by increasing the measure of similarity between variables u and v. Meanwhile, if variables s and t are both of particular interest, we would prefer to make them in separate clusters, so that their effect can be elucidated. This can be achieved by decreasing their measure of similarity.

Meanwhile, sampling uncertainty can be added in a way similar to the previous section. For variable clusters, i and j, bootstrap samples of individual pairs of variables can be added to the set of plots that needs to be summarized in cell i,j of the reduced scatterplot matrix.

3.3 Illustration

In this section, we provide an illustration of the procedure above on a simulated data set. Figure 3-2 shows the data set. Although this data set is not overwhelmingly large by itself, it is nevertheless difficult to digest immediately. Figure 3-3 shows the plot with the variables optimally reordered by the methods of [27]. Reordering the variables provides an improvement, though the scatterplot matrix still depicts a lot of information. Figure 3-4 shows the output of scagnostics on the data set. Scagnostics is an interesting method when there are a few plots which are dissimilar to many of the others. However, our the plots for our data set don't have that property, so the output of scagnostics on our data set is about equally difficult to interpret as the original plot itself.

Figure 3-5 shows the dendrogram produced by the heirarchical clustering algorithm. The heights at which clusters come together in the dendrogram represent the distances of those clusters when they were merged in the algorithm. Hence, if the dendrogram didn't present reasonably clear clusters, our method wouldn't be of great help. In this case, the dendrogram indicates that if we wanted a 4×4 simplified scatterplot matrix, we would cluster the variables into the following groups: $\{(3,6),(4,9,5),(1,8),(2,7)\}$.

Figure 3-6 shows the central simplified scatterplot matrix. This plot shows the characteristic images found in the larger plot. Meanwhile, Figures 3-7 and 3-8 show the two border scatterplot matrices. They demonstrate that the 4 by 4 scatterplot matrix, in this case, doesn't completely replace the 8 by 8 original scatterplot matrix, but also doesn't sacrifice an unacceptable amount of information. The exception to this is the plot at the lower right, which represents the plot when one of the variables $\{2,7\}$ is plotted against one of the variables $\{2,7\}$. This occurs because variables 2 and 7 are random noise which exhibit no interesting relationship to any of the other variables.

3.4 Discussion

We presented a method for simplifying large scatterplot matrices based on clustering variables and summarizing groups of plots. Our methods will be helpful for data whose scatter plot matrix consists of a smaller number of recurring images. Future research on this topic could consider extensions of this methodology to other high dimensional plots, such as parallel coordinates plots and star plots.

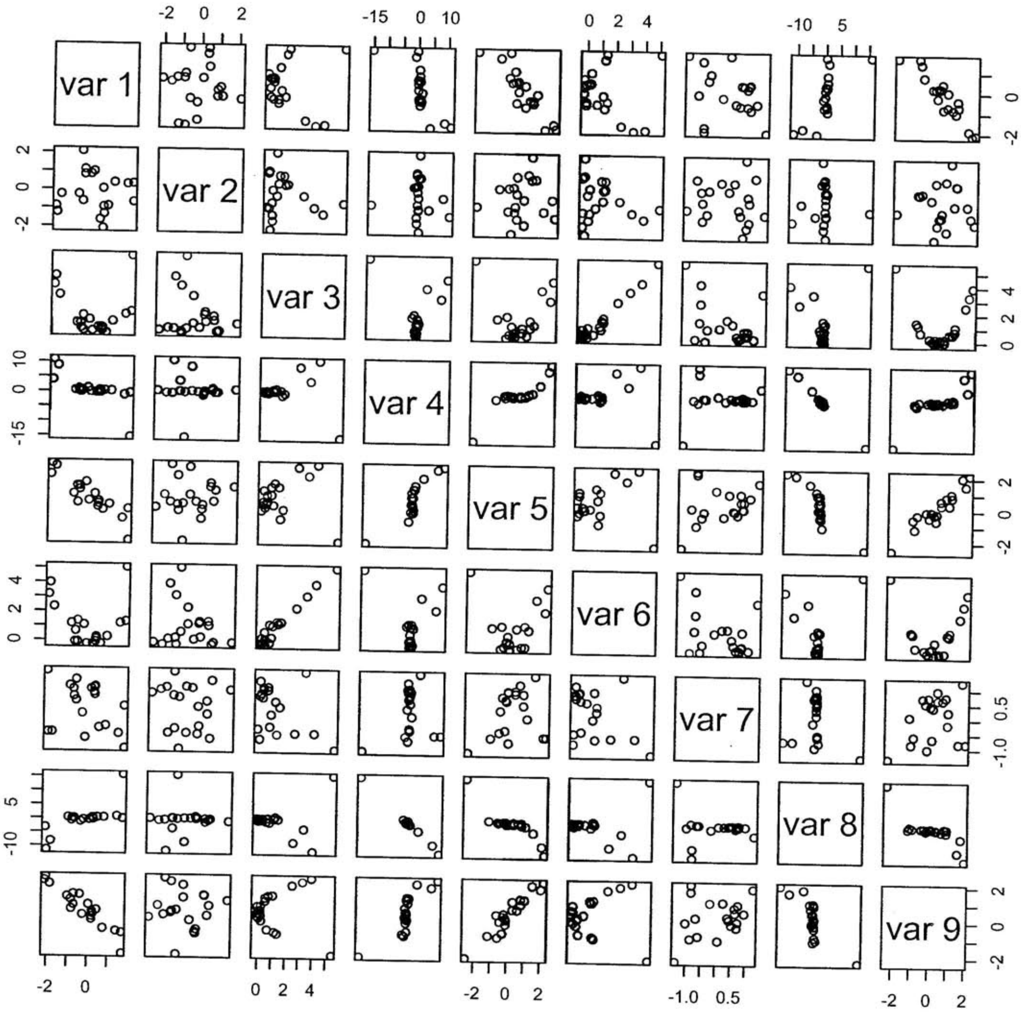


Figure 3-2: An Illustration of a scatterplot matrix

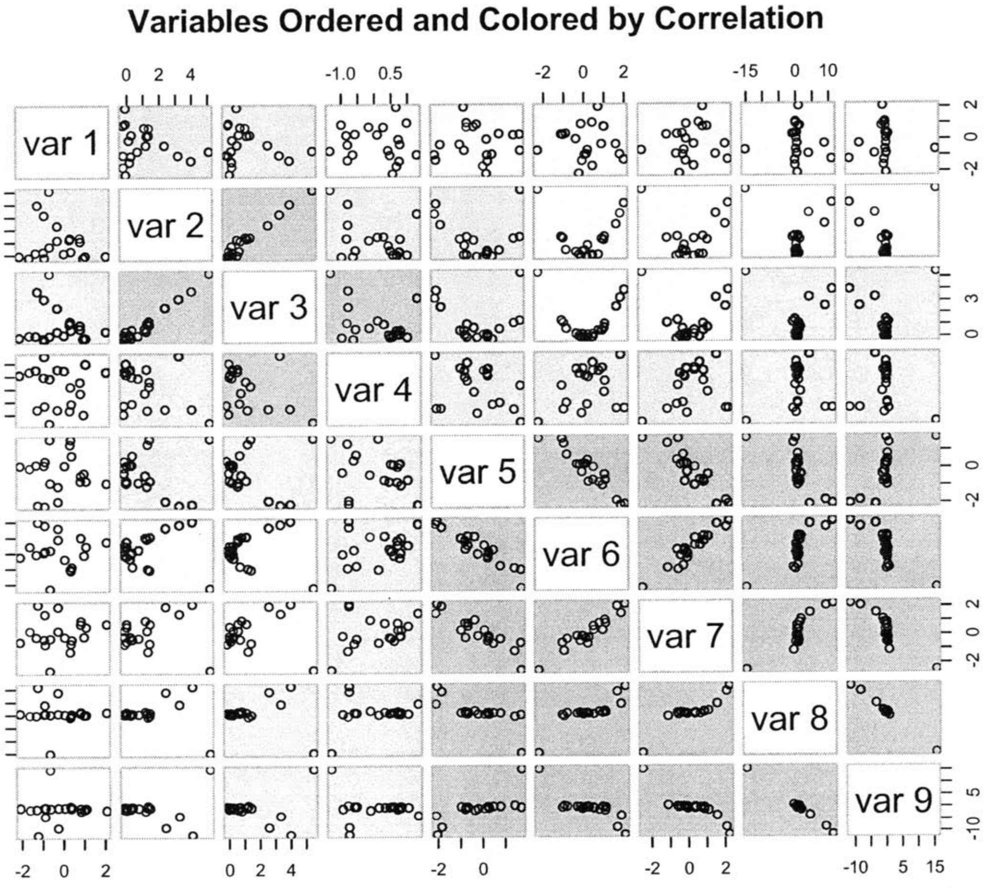


Figure 3-3: A scatterplot matrix with optimized variable order

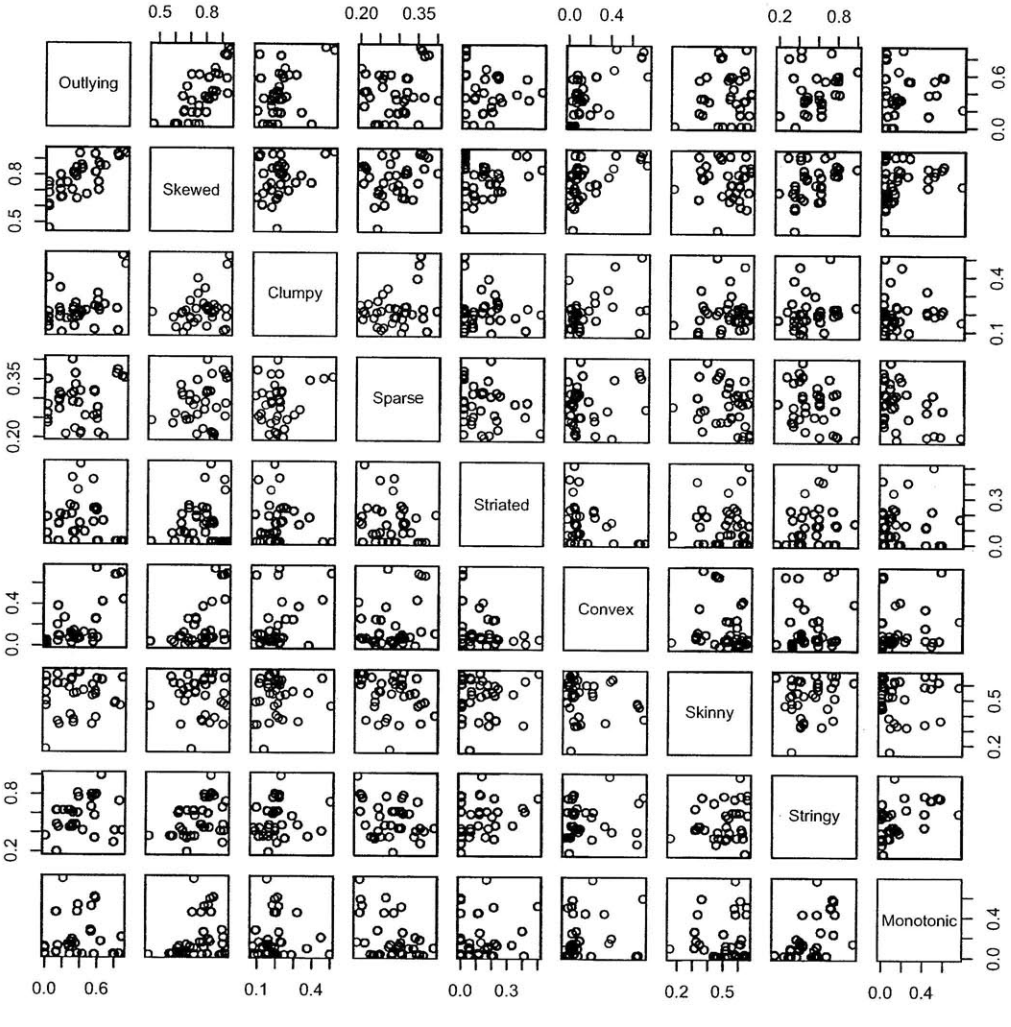


Figure 3-4: Output of scagnostics for the illustrated scatterplot matrix

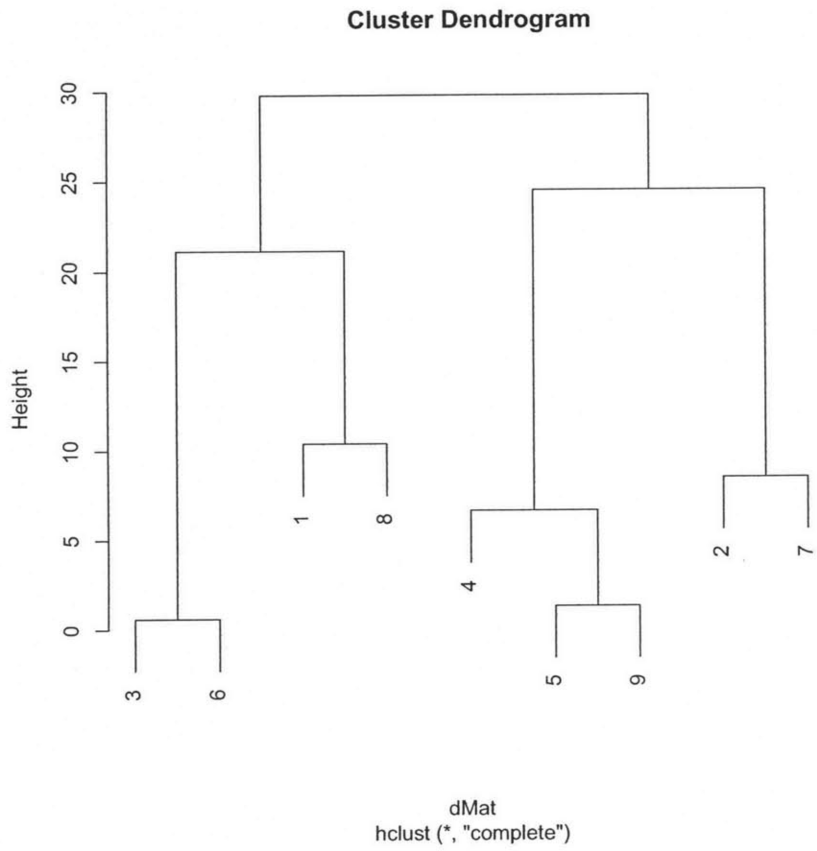


Figure 3-5: A dendrogram for the illustrated example

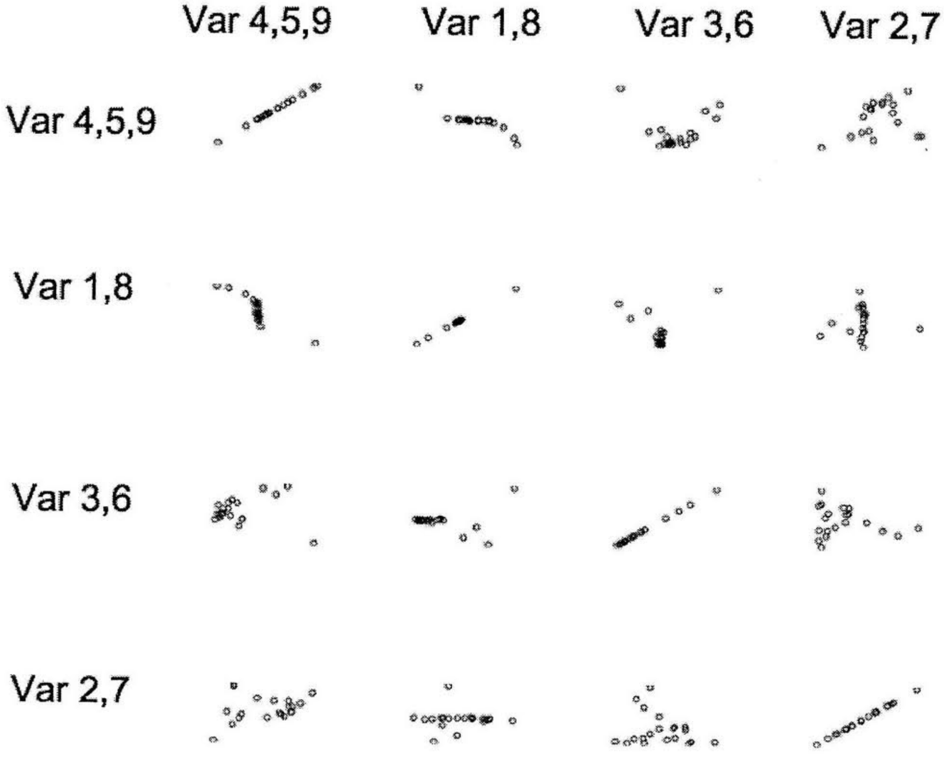


Figure 3-6: The central simplified plot for the illustrated example

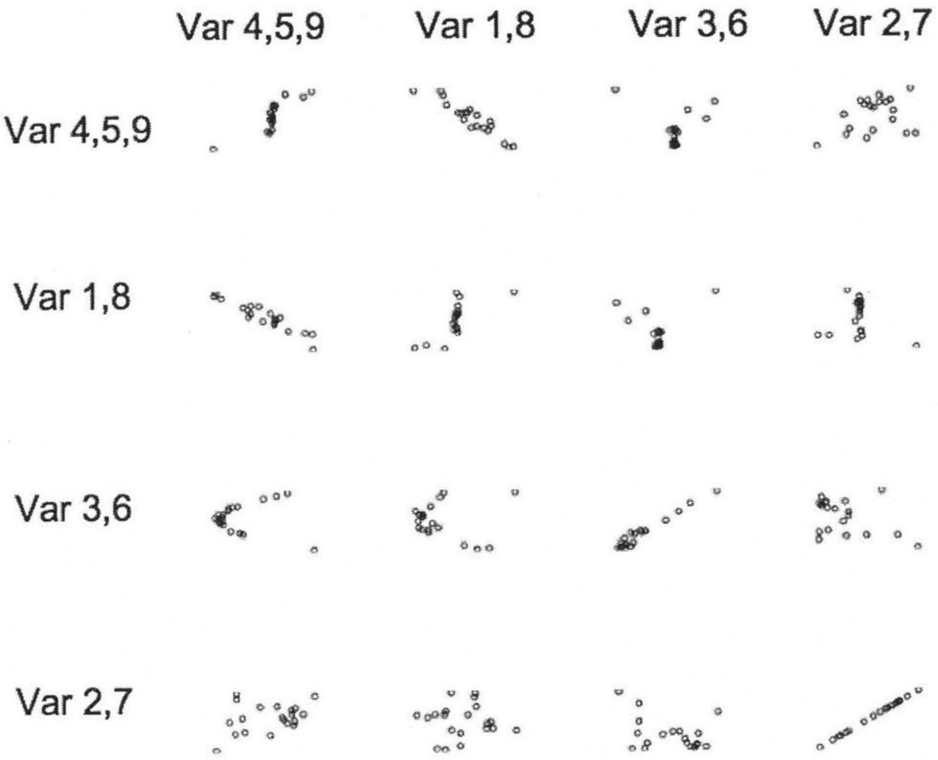


Figure 3-7: An extreme simplified plot for the illustrated example

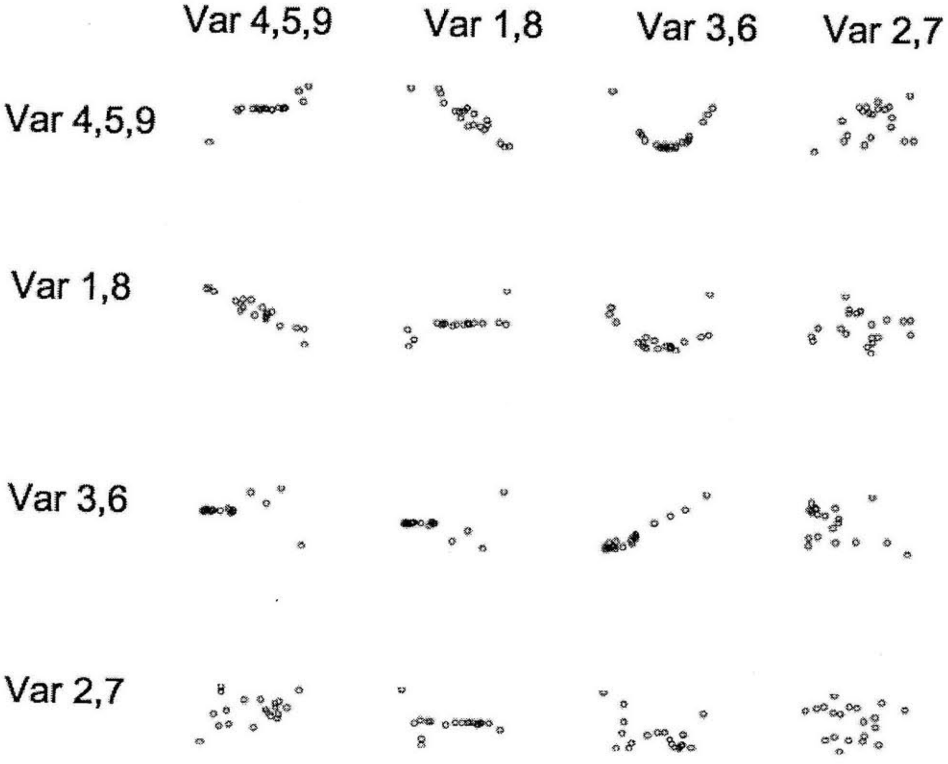


Figure 3-8: Another extreme simplified plot for the illustrated example

Chapter 4

Backward Selection Search for Diagnosing Regression Models and its Extensions

At this point of the thesis, we switch gears to the specific case of diagnosing regression models, which we partly motivated in the introduction. In this chapter, we present backward selection search in more depth. As stated in the introduction, backward selection search is a diagnostic method along the lines of forward search to simultaneously study the effect of individual observations and features on the inferences made in linear regression.

The method operates by appending dummy variables to the data matrix and performing backward selection on the augmented matrix. It outputs sequences of feature-outlier combinations which can be evaluated by similar plots to those of forward search and includes the capacity to incorporate prior knowledge, in order to mitigate issues such as collinearity. It also allows for alternative ways to understand the selection of the final model.

We organize the rest of the chapter as follows: Section 4.1 elaborates on the motivation for diagnostic methods for regression, which was described briefly in the introduction, Section 4.1.1 provides a brief literature review, discussing other ways the problem of simultaneous feature selection and outlier detection has been tackled and how diagnostic methods have evolved. Section 4.2 provides a review of forward search. Section 4.3 discusses the method that we propose in this paper, which we call backward selection search. Section 4.4 presents the output of our method on five well-known data sets. Following this, a notable extension of the idea of backward selection search, which enables outlier detection in multivariate data to be formulated as a feature selection problem, is presented in Section 4.5. Lastly, conclusions are in section 4.6.

4.1 Motivation

Outliers substantially complicate the already difficult task of model selection in linear regression. The question of which features to select as well as how many of the chosen features to select can both be grossly influenced by outliers, and to make things even tougher, the features that are selected in a model will influence which observations are considered outliers.

Robust model selection, however, brings complications even beyond its statistical framework. For instance, one complication of outlier detection is that a point that statistical methods deem an outlier could in fact be the most important observation in the data set depending on the application and the cause for the outlier. Forward search is one remedy for this in that it identifies outlying points of various magnitudes and creates plots that highlight the effect of each observation on various inferences made in linear regression. It thereby provides several possible good models and gives the analyst a way to visualize these. The goal of this paper is to extend these ideas to the case of simultaneous feature selection and outlier detection, which has arisen more recently in the literature.

4.1.1 Literature Review

In the sections below, we provide a literature review for robust model selection and diagnostic methods for outlier detection. We then discuss how our method fits in with the existing literature.

Diagnostic Methods for Outlier Detection

Diagnostic methods for outlier detection aim to separate observations into a "clean" set and a set of possible outliers [19], [20], as well as to highlight the effects of these possible outliers. Some of the earliest diagnostic methods include leave-one-out deletion techniques, which study the effect of each individual observation on the inference [8] by removing one of the data points and calculating the statistics of interest with and without this point to see if there is a large difference. However, this tends to fail when outliers come in groups. In general, the phenomena of outliers going undetected because of the presence of another set of outliers (Masking) and "good" observations being misidentified as outliers because of the presence of a set of outliers (Swamping) [19], [20], [4], [17], make the aim of diagnostic methods difficult to accomplish.

Over time, however, progress has been made and a popular recent algorithm, proposed by Atkinson and Riani [6] called forward search does a good job of detecting outliers and highlighting the influence of these outliers on various regression statistics. The forward search starts with a small subset of q "good" points, where q is the number of parameters (generally $q = p + 1$ where p is the number of features). Iteratively, the points which adhere most to the pattern that the good points follow are added to the set of "good" points, until all points are in the set of "good" points. The output is a sequence of sets of points of sizes $q, q + 1, ...$, and n along with the statistics of interest for these sets. Plots are then made where the y-axis is a statistic of interest and the x-axis is the size of the subset. Details are given in section 2.

Robust Model Selection

Some of the earliest papers in simultaneous outlier detection and feature selection focus on developing criteria for an optimal robust model: Ronchetti [53] and Ronchetti and Staudte [55] proposed robust versions of the selection criteria AIC and C_p respectively. Meanwhile, Ronchetti, Field, and Blanchard [54] proposed robust model selection by cross-validation.

More recent papers have explored the additional issue of selecting a sequence of relevant features in a robust manner: Khan, Van Aelst and Zamar [29] proposed a way of replacing various statistics with their robust counterparts in the LARS algorithm to perform robust LARS; Morgenthaler, Welsch, and Zenide [46] formulated outlier detection as a variable selection problem on an augmented matrix. McCann [43], McCann and Welsch [45], and Kim, Park, and Krzanowski [33] have built extensions on this using LARS and best subsets respectively to perform the variable selection. McCann and Welsch [44] also used the idea of feature selection on an augmented matrix to propose an alternative way to select sequences of points for forward search. In addition, Atkinson and Riani [5] enhanced their idea of monitoring variable coefficients and significance in the forward search through an added variable t-test for variable selection in the context of regression.

Our Approach

We seek to combine the two objectives of diagnosing outliers and selecting features into one method which simultaneously selects one parameter for both feature set size and observation set size, and produces a sequence of reasonable feature-observation models, which can be combined with prior knowledge and visualized for further inspection. Our method is most similar to [46] and its extensions, while our analysis mechanism tries to replicate that of forward search and [44].

4.2 A Description of Forward Search

Forward search is a diagnostic method that produces plots which help identify outliers, their structure, and their effect on various statistics of interest. As mentioned before, it starts with a small subset of q "clean" points, where q is the number of parameters (generally $q = p + 1$ where p is the number of features). Iteratively, the points which adhere most to the pattern that the "clean" points follow are added to the set of "clean" points, until all points (including outliers) are in the set of "clean" points. The output is a sequence of sets of points of sizes $q, q + 1, ...,$ and n along with the statistics of interest for these sets. Plots are then made where the y-axis is a statistic of interest and the x-axis is the size of the subset. The following is the procedure in detail:

1. Start by identifying an initial subset of q "clean" points using a high breakdown robust method. Atkinson and Riani suggest using Least Median of Squares in order to find an initial fit and then selecting the q points with the smallest squared residuals with respect to the initial fit to be the q "clean" points. Least Trimmed Squares is a reasonable alternative to Least Median of Squares, but both methods have been found to pick good initial subsets.

2. In a typical iteration, where the "clean" subset contains m points: conduct ordinary least squares using those m observations. Then compute all statistics of interest, say θ_m. Find the squared residuals for all points (not just the ones in the good set) and let the updated "clean" subset contain only the $m + 1$ points with the smallest squared residuals to this fit. It should be noted that these sequences of points are not necessarily nested.

3. Repeat step 2 until all the points are added, so that we are left with a vector $\theta = \{\theta_{q+1}, \theta_{q+2}, ..., \theta_n\}'$. The output of this procedure is typically a plot of θ_m as a function of m.

The purpose of the plot is to study how various outliers affect the statistics of interest. As a simple example, Figure 4-1 shows a plot of the Hertzsprung-Russell Star data set, that will be discussed further in section 4. It is an example where leave-one-out deletion techniques fail. Figures 4-2 and 4-3 illustrate the visual output that forward search provides for this data set. Figure 4-2 shows a plot of subset size vs R^2 and indicates a sharp decline around subset size 43. This suggests that there are about four outliers since there are 47 observations in total. Looking at Figure 4-3, which is a plot of subset size vs scaled residuals we see that the outliers mask each other since their scaled residuals move together throughout the search.

In addition to standardized residuals, Atkinson and Riani also suggest other possible statistics to consider, such as coefficient size, coefficient t-statistics, squared

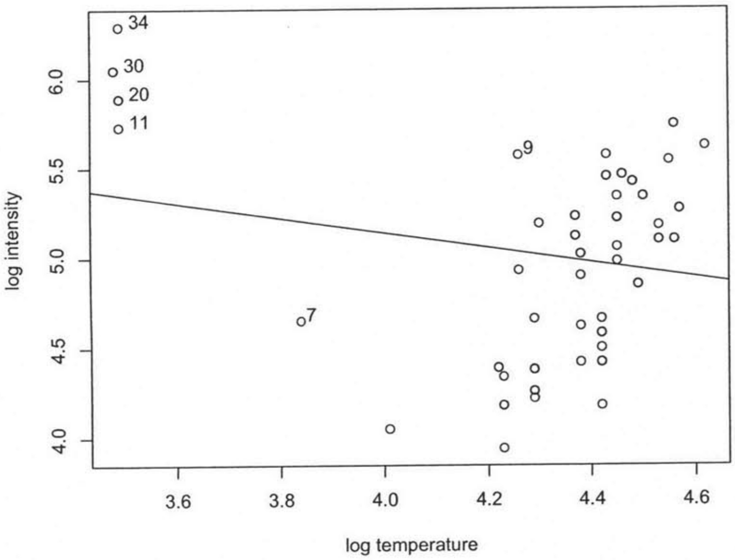


Figure 4-1: Plot of Hertzsprung-Russell Star data with superimposed regression line based on OLS fit

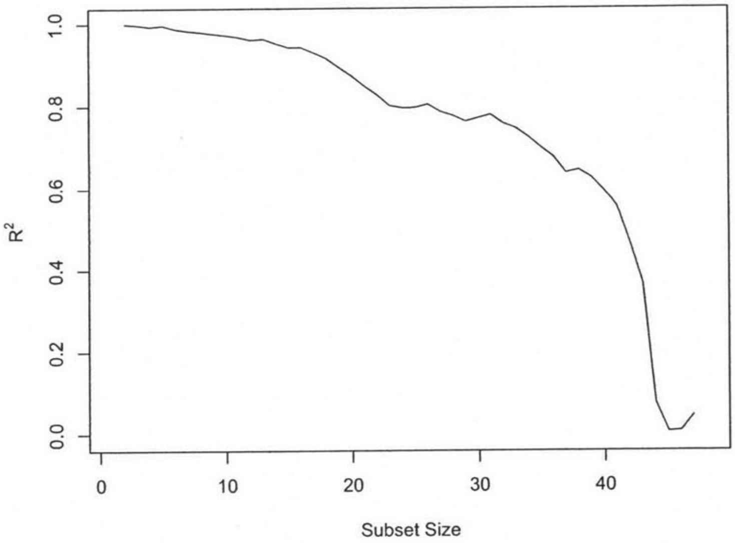


Figure 4-2: Hertzsprung-Russell Star data set: forward plot for R^2

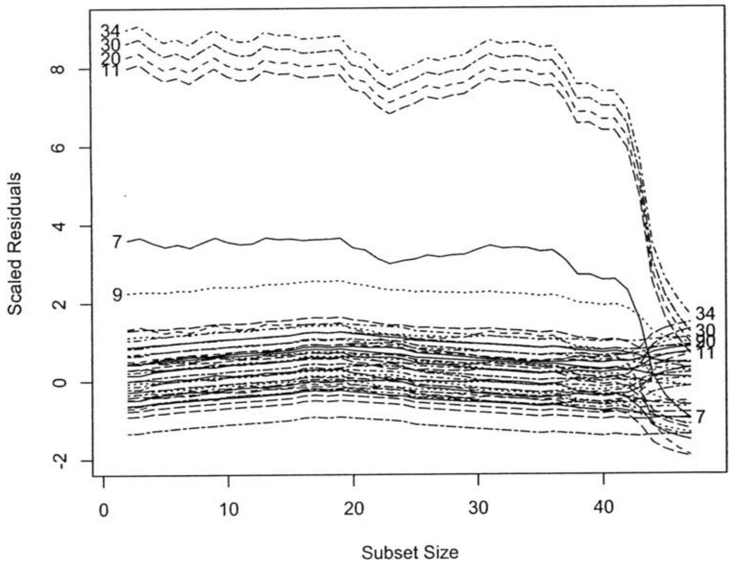


Figure 4-3: Hertzsprung-Russell Star data set: forward plot for scaled residuals

standardized residuals, Cook's distance, the maximum studentized residual, the minimum deletion residual, and leverage, all of which can be transferred over to our method.

4.3 Backward Selection Search

Our method is based on the well-known principles underlying the mean-shift outlier model [8], which state that an equivalent alternative to excluding j points from a regression model is to append j extra columns to our design matrix, with each column a dummy variable that has value 1 at the index of the outlier it represents, and 0 elsewhere. We begin this section by reviewing the details of this result and then proceed to the description of our procedure. Following this, we discuss computational tricks to make our procedure more efficient, and then we illustrate our procedure on a data set to demonstrate the conclusions one can draw from its output.

4.3.1 Mean-Shift Outlier Model

To give a more precise description of the statement above, we will use the notation below:

Let X denote the n by p matrix of realizations of the p explanatory variables, y denote the n by 1 vector of realizations of the response variable, I denote the n by n identity matrix, and J denote a set of indices. Furthermore, let "|" denote horizontal matrix concatenation.

We denote submatrices and subvectors as follows. Where K and L are sets of indices, $*$ is the set which contains every index, A is a given matrix, and b is a given vector, we let $A_{[K,L]}$ denote the submatrix of A containing the rows with indices in K and the columns with indices in L. On the other hand, we let $A_{(K,L)}$ be the submatrix of A which contains all rows and columns of A except rows with indices in K and columns with indices in L respectively. $A_{[K,L)}$ and $A_{(K,L]}$ are defined similarly. Similarly, let $b_{[K]}$ denote the elements of b with indices in K and let $b_{(K)}$ denote the elements of b whose indices are not in K. Lastly, we use the shorthand of a_i' to denote row i of matrix A and b_j to denote element j of vector b. Hence, based on our notation, $a_i' = A[i,*]$ and $b_j = b[j]$. Meanwhile, $A(i,*]$ would denote all rows of A except a_i'.

The proposition below is a well-known result (see for instance [6]), but we provide a simple proof in order for this thesis to be self-containted.

Proposition: The estimated OLS coefficient for each of the p explanatory variables in the regression of y onto $[X|I_{[*,J]}]$ equals the corresponding estimated OLS coefficient in the regression of $y_{(J)}$ onto $X_{(J,*]}$.

Proof:

Let $\hat{\beta}$ denote the OLS vector of explanatory variable coefficients for the regression of y onto $[X|I_{[*,J]}]$ (Note that $\hat{\beta}$ does not include the coefficients for dummy variables). Then, where β and α are vectors of decision variables that correspond to the coefficients to be chosen for the explanatory variables and the coefficients to be chosen for the dummy variables respectively:

$$
\begin{aligned}
\hat{\beta} &= \arg\min_{\beta}[\min_{\alpha}(y - X\beta - I_{[*,J]}\alpha)'(y - X\beta - I_{[*,J]}\alpha)] \\
&= \arg\min_{\beta}(\sum_{i \notin J}(y_i - x_i'\beta - 0)^2 + \min_{\alpha}\sum_{i \in J}(y_i - x_i'\beta - \alpha_i)^2) \\
&= \arg\min_{\beta}\sum_{i \notin J}(y_i - x_i'\beta)^2 \\
&= \arg\min_{\beta}(y_{(J)} - X_{(J,*]}\beta)'(y_{(J)} - X_{(J,*]}\beta)
\end{aligned}
$$

The first equality follows from the equation for a subvector of the estimated coefficient vector under OLS. Here, the full estimated coefficient vector would be the concatenation of $\hat{\beta}$ and $\hat{\alpha}$, and the subvector would be $\hat{\beta}$.

The third equality follows from noting that $(y_i - x_i'\beta - \alpha_i)^2 \geq 0$ and equals 0 when α_i is set to $y_i - x_i'\beta$. This completes the proof, since the last equation is the optimization problem for OLS regression of $y_{(J)}$ onto $X_{(J,*]}$.

Corollary:

The t-statistics for the features in the two regression models above are the same.

Proof:

In both regression models, $\hat{\beta} = (X'_{(J,*]}X_{(J,*]})^{-1}X_{(J,*]}y_{(J)}$. Additionally, in both models, $y_{(J)}$ has the same distribution. It follows that the estimate, expected value, and variance are the same in both cases. Hence, the coefficient t-statistics are the same.

4.3.2 Procedure

Producing a Sequence of Models

Based on the results above, we can perform a procedure similar to forward search, but using the backward selection algorithm. In particular, assume we have a matrix $E_d = [e_{i_1}, e_{i_2}, ..., e_{i_d}]$ where e_j is a column vector with j^{th} element 1 and all other elements 0, and where $\{i_1, ..., i_d\}$ is some subset of $\{1, ..., n\}$. As a result of the principles above, comparing a regression of y upon $[X|E_d]$ and y upon $[X|E_d|e_{i_{d+1}}]$

is equivalent to comparing a regression of y upon X with clean set $\{i_{d+1}, ..., i_n\}$ (meaning that the other observations are excluded from the model) and a regression of y upon X with clean set $\{i_{d+2}, ..., i_n\}$ respectively. Hence, if we preselect good points $\{i_{n-q-1}, ..., i_n\}$ and apply backward selection to $[X|E_{n-q-2}]$, we have a method which is very similar to forward search. One advantageous difference here, however, is that we can remove a variable of X in backward selection as well as a dummy variable. This procedure hence can perform simultaneous feature selection and outlier detection and has a reasonable way of comparing the actions of adding an extra observation and removing an extra feature.

Our procedure is thus as follows:

1. Start by identifying an initial subset of $q + 1$ "good" points. The procedure to do so can be the same as in forward search. We then append dummy variables for all other points to the X matrix to form a matrix $Z = [X|E_{n-q-1}]$.

2. Perform backward selection on the matrix Z with respect to y, keeping track of the statistics along the way at each iteration, say θ_i for iteration i. This means that at each step, we iteratively delete the least relevant variable. We choose the variable to delete as that which has the coefficient with the lowest absolute t-statistic, and we add the constraint that the intercept should not be deleted. It should be noted that the need to calculate this t-statistic is why we need the initial $q + 1$ points in the first step.

3. We are left with a vector θ which gives values of the statistic of interest for different models.

Although it may not be necessary to specify one specific model at the end of the analysis, the model selection problem with respect to the output of this algorithm would be to select an iteration at which to stop. Given a specified stopping point (and hence model), the model would deem all variables that were deleted before the stopping point irrelevant, and all points whose corresponding dummy variable was not deleted before the stopping point to be outliers.

Incorporating Prior Knowledge

An additional advantage of our model is that we can relatively easily incorporate prior knowledge via mixed estimation [62]. This technique uses the prior information by augmenting the data directly instead of using a prior distribution as in the Bayesian Model. Denote our current model as $y = Zv + \epsilon = X\beta + E_d\alpha + \epsilon$. In mixed estimation we assume that we can write a set of restrictions on the coefficients v of the form $a = Dv + \delta$ where δ is multivariate normal with $E[\delta] = 0$ and $\text{var}(\delta) = V$, D is a matrix of known constants and a is a vector of random variables. We can treat these equations as new data and solve for the significance of the coefficients of v via the methods of generalized least squares. As our algorithm only requires the computation of the significance of coefficients, this fits into the framework of our model.

A typical case of adding prior knowledge occurs when we add prior knowledge that the coefficients have an independent normal distribution around zero. Where $Dv = V\beta + W\alpha$, this would correspond to either making V, W, or D equal to the identity matrix times some constant $\sqrt{\lambda}$ chosen by the user. This prior, combined with our model, can mitigate problems such as collinearity and cases where $p > n$.

Selecting a Final Model

Given the trace our method produces there are several possible ways to select the final model. First, one could use any one of the statistics produced in the early work of simultaneous feature selection and outlier detection (cited in the literature review) and compute the statistics for each model produced in the trace.

Second, one could look for drastic changes in the diagnostic plots produced to explore a variety of possible models. To assist in identifying drastic changes for forward search, Riani and Atkinson [51] produce and plot confidence bounds for their statistic of interest by running forward search on many bootstrap samples. In each bootstrap sample, X is generated as an n by p matrix of standard normal realizations and y is generated as an n by 1 vector of standard normal realizations. Riani, Atkinson, and Cerioli [52] also develop a hypothesis test for the number of outliers based on these plots. We use similar plots on the data sets tested, running backward selection search on 1000 bootstrap samples.

In this paper we also study a third way of understanding what the final model should be. These ideas are based on the developments in [61], which propose to stop a feature selection process by augmenting the data matrix with an artificial random feature and stopping the process once that random feature is selected. Since our method is a special case of a feature selection process, we attempt to apply this procedure to the data sets in section 4, simulating 1000 routines and we track where in the trace a random additional feature is eliminated.

4.3.3 Computational Tricks

In the previous sections, we described backward selection search as performing backward selection on an augmented design matrix. However, if we don't add prior knowledge, the algebra ends up simplifying, so that we do not need to actually append the identity matrix when computing the output, thus reducing computation times. We state the details of this below, and then discuss how the computational efficiency of our method compares to that of forward search.

Computing t-statistics

We first describe how to compute the t-statistics without actually appending the identity matrix. Since this is all we need to do at each step in order to determine the sequence of models, we can conclude that, in the case of no prior knowledge, we need not carry around the identity matrix in the computation.

Suppose at a given iteration, J is the set of observations which have not been added. Let D be the current design matrix (without the identity matrix appended) containing only those features which still remain in the model. Then, by the corollary to the proposition earlier in this section, the t-statistics of the features in the current model are merely the t-statistics of the features of the regression of $y_{(J)}$ onto $D_{(J,*]}$.

Meanwhile, to compute the t-statistics for a dummy variable corresponding to an observation in J, we first note that the corollary implies that the regression of y onto $[D|I_{[*,J]}]$ gives the same t-value for dummy variable i as the regression of $y_{(K)}$ onto $[D_{(K,*]}|I_{[*,i]}]$, where K is the set of all observations in J except i. [6] shows that this ends up simplifying to $(y_i - d_i'\hat{\beta})/(s\sqrt{1 + d_i'(D'_{(J,*]}D_{(J,*]})^{-1}d_i})$.

Although this is all that is necessary for the sequencing of the models, similar arguments show that we can use the added t-statistic formula derived in [6] to compute the t-statistics for variables which have already been eliminated and observations not in J.

Computational Differences to Forward Search

In the case of no prior knowledge, the bulk of the computation in our algorithm that one needs to do at every step is the computation of $(D'_{(J,*]}D_{(J,*]})^{-1}$. However, in order to compute the various diagnostics of forward search, one requires these computations as well. It follows that the only computational differences of our algorithm to that of forward search are the p extra steps our algorithm takes. This extra complexity is in part alleviated by tricks to update an inverse of a matrix when one column is deleted, based on block matrix formulas [30].

In order to give an idea of how long these extra steps would take in practice, we provide the extra computation times for a few different problem sizes in table 4.1. In each cell, the first number is the amount of time in seconds that the p extra steps took when all observations were present (this will be longer than the amount of time the p steps will take in general). The second number, meanwhile, is the amount of time forward search took. All computations were done using R version 2.9.0 on a Mac Book with 4GB RAM. The differences in computation times are not large, as is expected.

When prior knowledge is added, an upper bound on the computation time is the computation time of backward selection for a generalized linear model on an $n \times n$ design matrix. However, if the artificial data points, whose form the prior knowledge takes, have a special form such as independence, which is ordinarily the case, we can use the same computational tricks as above.

Table 4.1: Computation times in seconds for different problem sizes. The first number is the time for p extra steps and the second number is the time for forward search.

	$n = 100$	$n = 200$	$n = 500$
$p = 10$	0.01, 0.42	0.02, 1.30	0.08, 7.22
$p = 30$	0.03, 0.76	0.07, 1.98	0.35, 9.69
$p = 50$	0.08, 1.48	0.16, 3.15	0.66, 13.55
$p = 70$	0.14, 2.55	0.30, 4.65	1.07, 19.90
$p = 90$	0.25, 4.17	0.45, 6.69	1.51, 22.71

4.3.4 Illustration of the Procedure

In order to demonstrate the utility of our method and its differences to forward search, we add three features to the Hertzsprung-Russell Star data set displayed in the previous section. In the augmented data set, the first feature is the first feature of the original data set, which is the log of temperature. The second feature is simulated from a standard normal distribution and hence should be considered irrelevant in the evaluation. The third feature takes value 1 at observations 7 and 9, and takes value 0 elsewhere. It would be expected that with this feature in the data set, observations 7 and 9 wouldn't be considered as outliers. Finally, the fourth feature is set to be $0.999X_1 + 0.001Z$, where Z is a standard normal random variable and X_1 is the first variable. It is designed to be collinear to the first variable.

Figures 4-4 and 4-5 show two plots in the output of the R function fwdlm [49], which conducts forward search. Figure 4-4 suggests what appear to be four masked outliers. This makes sense, since observations 7 and 9 are not expected to be outliers given the presence of the third variable. Figure 4-5, which plots t-statistics, interestingly suggests that even well before observations 11, 20, 30, and 34 are taken out, variable 1 is irrelevant. Although this inference is false, it is not surprising that this inference was mistakenly made, as this is one of the well known effects of collinearity.

Figures 4-6 and 4-7 show the corresponding plots for backward selection search. Meanwhile, Table 4.2 shows the last 12 steps that the algorithm takes. Here, X_j represents the j^{th} feature among the original variables, while y_i represents the i^{th}

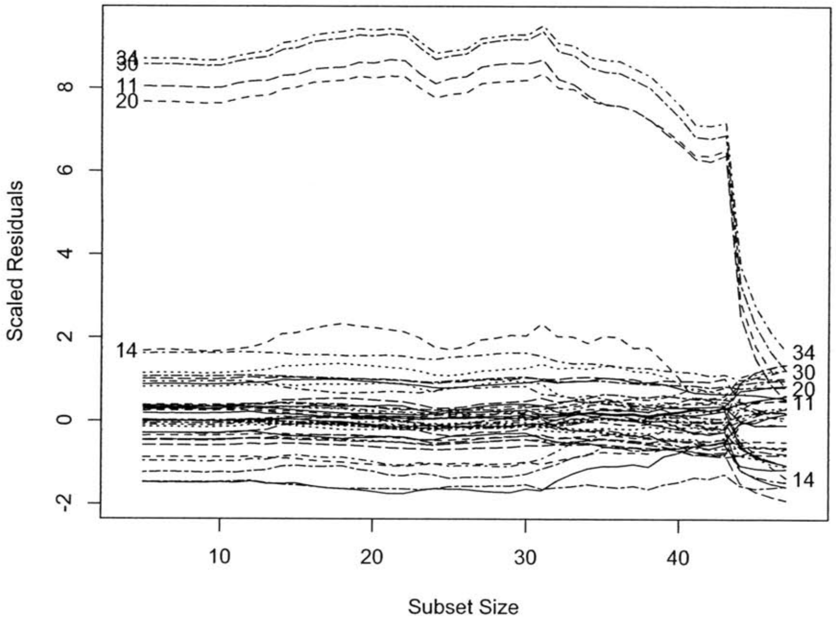


Figure 4-4: Hertzsprung-Russell Star data set with three additional features: Forward plot for standardized residuals based on forward search

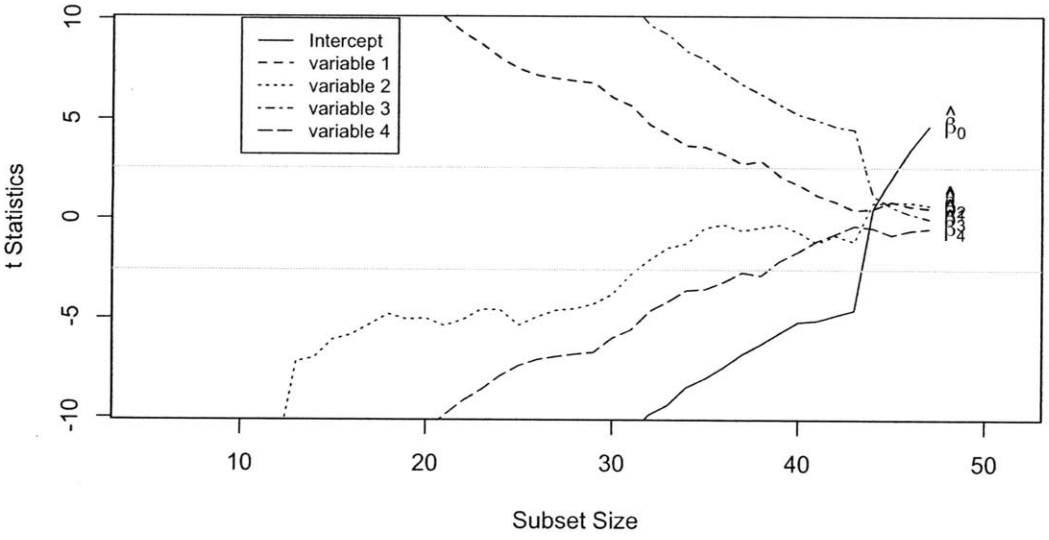


Figure 4-5: Hertzsprung-Russell Star data set with three additional features: Forward plot for coefficient t-statistics based on forward search

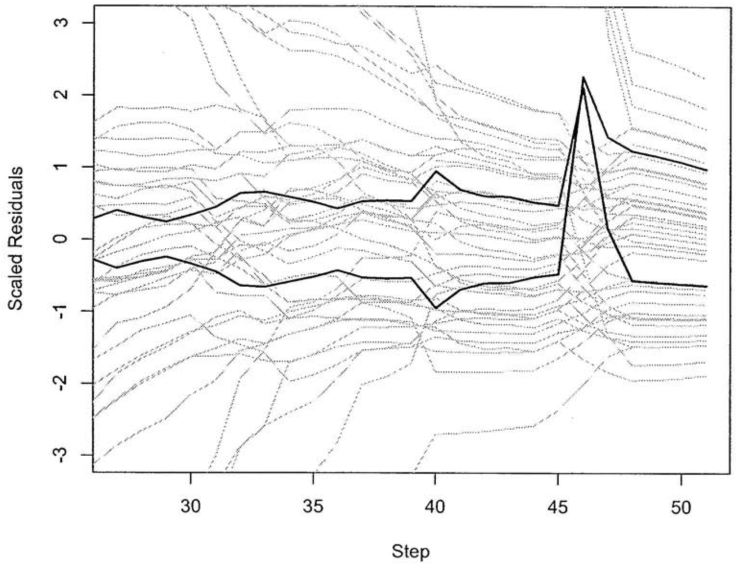


Figure 4-6: Hertzsprung-Russell Star data set with three additional features: Forward plot for scaled residuals based on backward selection search. The plot for observations 7 and 9 are highlighted in black

observation. Hence, after feature 4 is deleted, observations 14, 3, 40, 5, and 18 are added. Then, feature 3 is deleted, observation 11 is added, feature 1 is deleted, and finally observations 20, 30, and 34 are added. This means, for instance, that in a model where observation 20 is not considered an outlier, feature 1 is deemed irrelevant (since feature 1 will be deleted in a model which stops the trace after observation 20 is added). However, the claim that feature 1 is relevant is stronger than the claim that observations 7 and 9 are outliers. If we stopped at the sixth from last step, observations 11, 20, 30, and 34 would be considered outliers, and observations 2 and 4 would be considered irrelevant, which is what we would expect based on the way we simulated these features.

Figure 4-6 highlights observations 7 and 9 and shows a sharp increase in their scaled residuals when feature 3 is deleted. This implies that the presence of feature 3 makes observations 7 and 9 appear nonoutlying. Meanwhile, figure 4-7 shows a dramatic increase in the t-statistic for the first feature when the fourth feature is

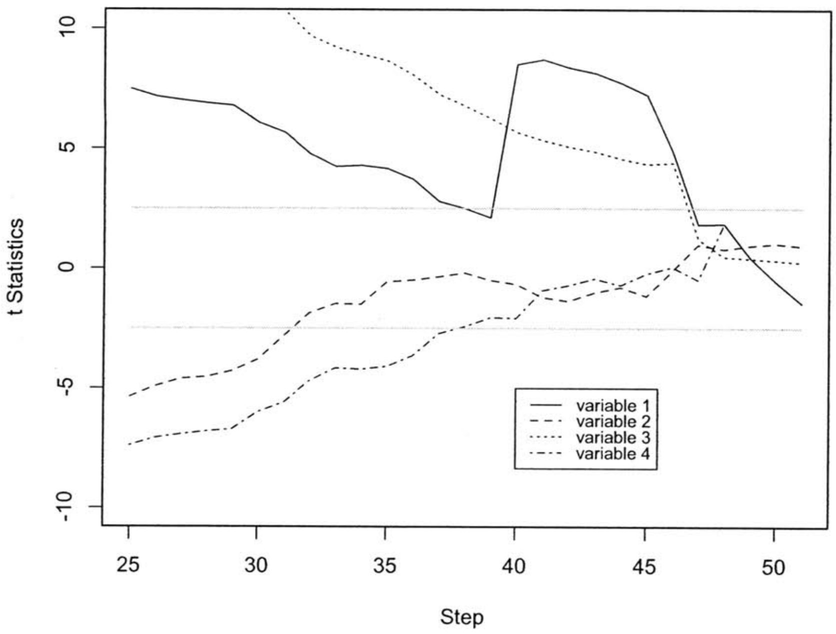


Figure 4-7: Hertzsprung-Russell Star data set with three additional features: Forward plot for coefficient t-statistics based on backward selection search

Table 4.2: Last twelve steps of backward selection search in the modification of the Hertzsprung-Russell Star data

Data		Last Twelve Steps											
Hertzsprung-Russell	step	40	41	42	43	44	45	46	47	48	49	50	51
Star Data	add	-	y_{14}	y_3	y_{40}	y_5	y_{18}	-	y_{11}	-	y_{20}	y_{30}	y_{34}
Modification	drop	X_4	-	-	-	-	-	X_3	-	X_1	-	-	-

eliminated from the data set. This implies that features 1 and 4 are collinear, in that whether one is present will have a large effect on the inferences about the other one.

It should be noted that this is not a criticism of forward search, as forward search still does the job it intended to do on this data set. Instead, it is an example of where it may be informative to extend the job that forward search intends to do. It should also be noted that we constructed this data set for the purpose of illustration, so we will save the validation, as well as a more in depth analysis of the sequencing and model selection capabilities of backward selection search, for the next section.

4.3.5 Limitations of the Procedure

The procedure has a similar limitation to forward search in that it requires a set of "clean" starting points. However, while forward search can sometimes recover from a poor set of starting points, backward selection search can't necessarily recover since the sets of clean points are nested. We thereby also implemented a modification of the algorithm which deletes the least significant feature, adds the most significant feature based on added variable t-tests, and then deletes the least significant feature from this modified subset in a typical iteration. This modified algorithm recovered from more starting points than backward selection search, but still failed in some cases where forward search also failed. Since backward selection search has a structure which is more intuitive and easier to analyze, and since it works well given that we can indeed come up with a clean set of starting points, we focus on the case where we have a clean set of starting points in the paper. It should also be pointed out that several quick heuristics with relatively good performance are now available. For instance, Nguyen and Welsch [48] use semidefinite programming, and Kim and Krzanowski [32] use clustering to identify potential outliers in linear regression.

4.4 Output on Data Sets

4.4.1 Description of Data Sets

We used the following five data sets to evaluate our method: Stackloss data [10], Scottish Hill Racing data [3], Modified Wood Gravity data [56], Modified Hertzsprung Russell Star data, and Modified Belgian Phone data [57]. The first three of these are the data sets that are used by Hoeting *et al.* [25] and Kim *et al.* [33] to evaluate their respective methods. The last two data sets are based on two simple one-dimensional data sets discussed by Rousseeuw and Leroy [57], but modified to contain additional irrelevant features. The original data sets have become standard benchmark data sets for robust methods in regression.

The Stackloss data [10] have been studied by several authors in the context of identifying outliers. The data set contains three variables and 21 observations and the general consensus is that the third predictor variable should be dropped and that observations 1, 3, 4, and 21 are potential outliers if the data are analyzed on the original scale. On the other hand, if the data are analyzed on the square root scale, only observations 4 and 21 are potential outliers [6].

The Modified Wood Gravity data are based on real data but are modified by Rousseeuw [56] to contain outliers at cases 4, 6, 8, and 19. The data set has five variables and 20 observations. The two best subset models are (X_1, X_2, X_3) and (X_1, X_2, X_3, X_5) based on C_p and adjusted R^2 respectively, while the best subset model with outliers (4,6,8,19) removed is (X_1, X_3, X_4, X_5). Therefore, this is a data set where the optimal subset of features depends on the outliers.

The Scottish Hill Racing data [3] contain the record-winning times for 35 hill races in Scotland. The data set contains two independent variables, distance and climb, with the dependent variable, time and has 35 observations. Observation 33 is known to be an error in the data and when a linear model is applied, there is consensus that observations 7 and 18 are outlying. Both variables are significant.

The Hertzsprung-Russell Star data set is an example of a data set used to make a scatter plot of a star's luminosity against its temperature on a log scale. In the

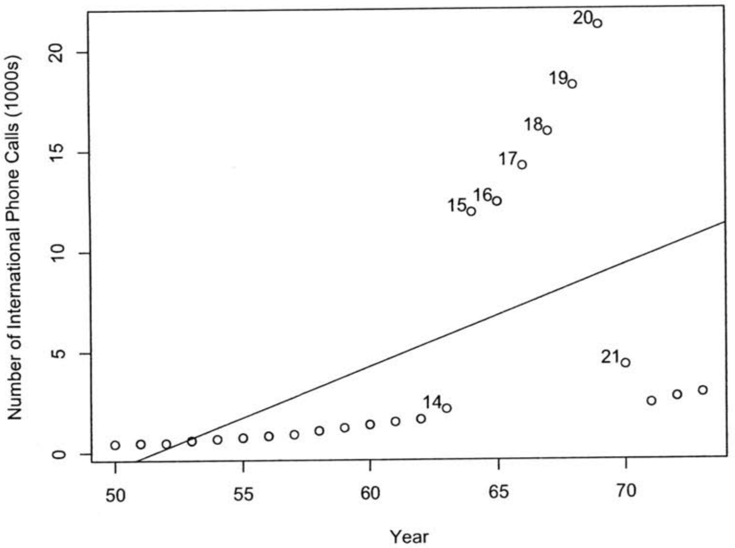


Figure 4-8: Plot of Belgian Phone data with superimposed regression line based on OLS fit

original data set, there are 47 observations and one explanatory variable. Figure 4-1 shows a plot of the data set with the OLS fit. Observations 11, 20, 30, and 34 are gross outliers and observations 7 and 9 are moderate outliers. We modify these data by simulating five additional variables from a standard normal distribution and appending them to the original data.

Finally, the Belgian Phone data are a record of the number of international phone calls from Belgium, taken from the Belgian Statistical Survey. The response is the number of phone calls, and the explanatory variable is the year. Figure 4-8 shows a plot of the data with the OLS fit. Observations 15-20 are known errors in the data and observations 14 and 21 may be considered outliers. We modify these data by simulating three additional variables from a standard normal distribution and appending them to the original data.

4.4.2 Performance

All computations and plots were constructed in R with randomization seed 2.

Table 4.3 shows the last 10 features eliminated in each of the data sets. Here again, X_j represents the j^{th} feature among the original variables, while y_i represents the i^{th} observation. For example, in the case of the Hertzsprung-Russell Star data set, the table indicates that if we stopped at the seventh from last step, then our model would say that variable 1 is relevant, the five other simulated variables are irrelevant, and that observations 7, 9, 11, 20, 30, and 34 are outliers. Based on the data description, this is the model we would want to select.

The table reads the same way for the other data sets and we see that our method very successfully sequences the variables and outliers, since for each data set, the optimal sequence occurs at one of the possible stopping times.

In addition to testing the sequencing of our method in data with known properties, we also tested the case where we add prior information for the Modified Wood Gravity data. The last two rows of Table 4.3 show the sequencing of the models when we add the prior information that $\sqrt{\lambda}\alpha_i + \epsilon_i = 0$ for all i, where α_j is the coefficient of the j^{th} dummy variable and ϵ_j is the error term, for $\lambda = 1, 2$. This corresponds to increasing degrees in belief that we should focus more on feature selection at the expense of outlier detection. The table shows the last ten features picked in each of these cases and we get that (X_1, X_2, X_3) and (X_1, X_2, X_3, X_5) show up as the relevant features, which is what we expected based on the description of the data provided.

Next, we investigated the actual question of model selection itself. As mentioned before, there are several ways of doing this, including computing robust model selection statistics such as C_p for each of the models in the trace. Here, we focus on three plots which assist us in the process. Figure 4-9 shows, for each data set, a plot of the iteration of the algorithm we are at versus the absolute value of the t-value of the coefficient of the feature being eliminated at that step. The gray lines in the plot show the 1, 2, and 3 standard deviation bands for this t-statistic based on the parametric bootstrap procedure described earlier. Figure 4-10 is a generic forward plot of the iteration we are at in the algorithm versus the model's R^2. Lastly, we performed 1000 simulations for each data set, where we appended a random feature to the data matrix, ran backward selection search, and observed where in the sequence the ran-

Table 4.3: Last ten steps of backward selection search in each data set

Data		Last Ten Steps									
Hertzsprung-Russell	step	44	45	46	47	48	49	50	51	52	53
Star Data	add	y_{40}	y_5	y_{18}	y_9	y_7	y_{11}	-	y_{20}	y_{30}	y_{34}
	drop	-	-	-	-	-	-	X_1	-	-	-
Belgian Phone Data	step	19	20	21	22	23	24	25	26	27	28
	add	-	y_{14}	y_{21}	-	y_{15}	y_{16}	y_{17}	y_{18}	y_{19}	y_{20}
	drop	X_3	-	-	X_1	-	-	-	-	-	-
Stackloss Data	step	15	16	17	18	19	20	21	22	23	24
Original Scale	add	y_{14}	y_{13}	y_{20}	y_2	-	y_1	y_3	y_4	y_{21}	-
	drop	-	-	-	-	X_2	-	-	-	-	X_1
Stackloss Data	step	15	16	17	18	19	20	21	22	23	24
Square Root Scale	add	y_{16}	y_3	y_{14}	y_{13}	y_{20}	y_2	-	y_4	y_{21}	-
	drop	-	-	-	-	-	-	X_2	-	-	X_1
Scottish Hill Racing	step	28	29	30	31	32	33	34	35	36	37
Data	add	y_{35}	y_{30}	y_{14}	y_6	y_{19}	y_{33}	y_7	y_{18}	-	-
	drop	-	-	-	-	-	-	-	-	X_2	X_1
Modified Wood	step	16	17	18	19	20	21	22	23	24	25
Gravity Data	add	y_9	y_{18}	-	-	-	-	y_4	y_6	y_8	y_{19}
	drop	-	-	X_4	X_5	X_1	X_3	-	-	-	-
Modified Wood	step	16	17	18	19	20	21	22	23	24	25
Gravity Data	add	y_7	-	y_5	y_{19}	y_1	y_{14}	y_{12}	-	-	-
$\lambda = 1$	drop	-	X_5	-	-	-	-	-	X_3	X_1	X_2
Modified Wood	step	16	17	18	19	20	21	22	23	24	25
Gravity Data	add	y_1	y_7	y_5	y_{19}	y_{14}	y_{12}	-	-	-	-
$\lambda = 2$	drop	-	-	-	-	-	-	X_5	X_3	X_1	X_2

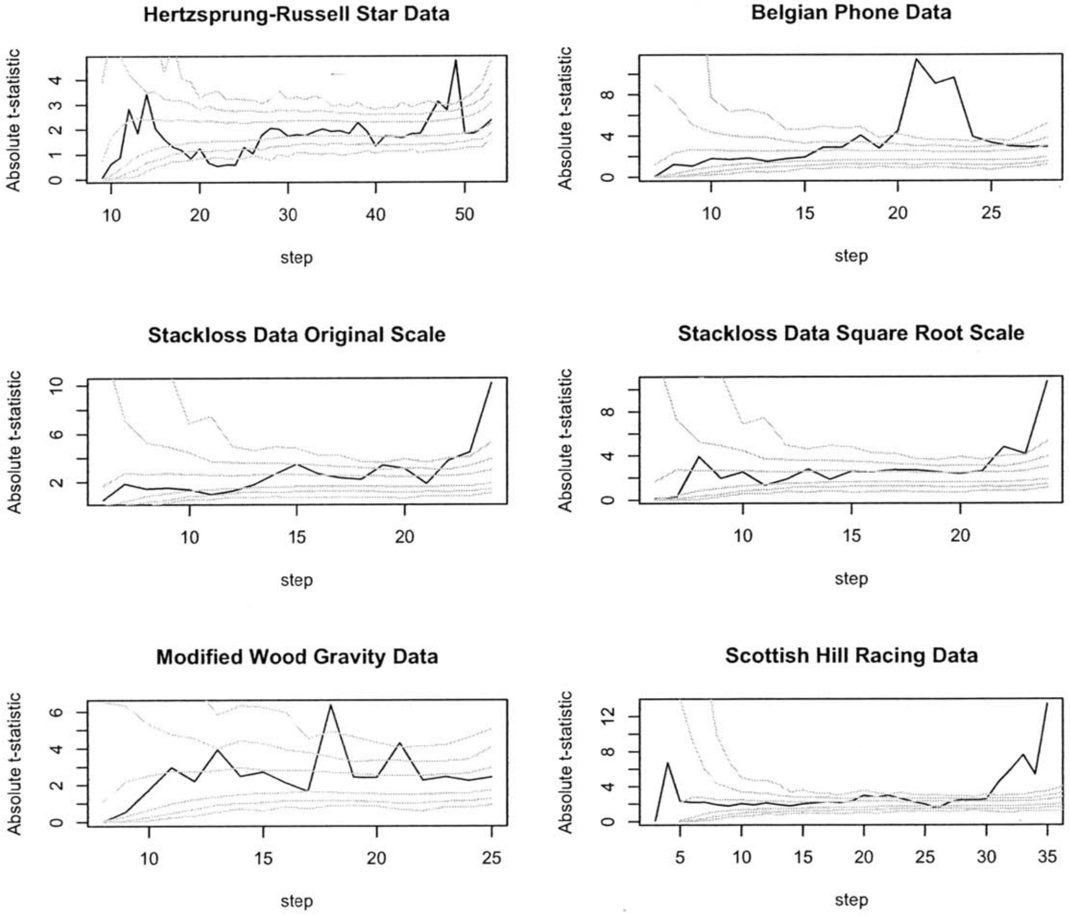
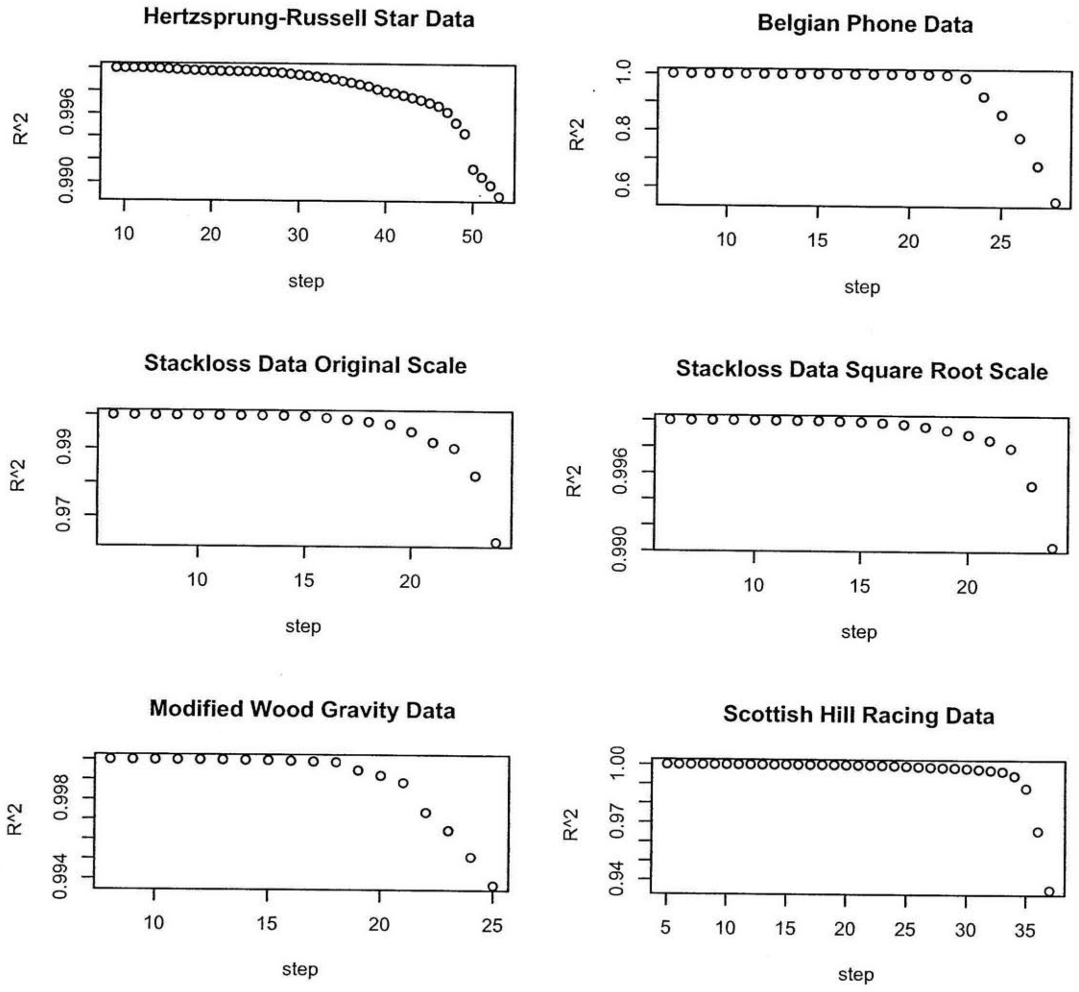


Figure 4-9: Forward plot of the absolute value of the t-statistic of the exiting feature for each data set with envelopes

dom feature was eliminated. Figure 4-11 shows the empirical cumulative distribution function of the stopping times (the time the random feature was eliminated) for each data set.

Figure 4-9 is the most effective of the rules on the data sets we considered. One would infer a stopping point from Figure 4-9 by observing the points at which the forward plot crosses its envelopes and choosing a reasonable point among these. As shown in the plots, this method agrees with previous literature in selecting a model, in all cases, except the Stackloss data. However, the square root transformation resolves this problem.

Figure 4-11 is also an informative plot. One would infer a stopping point from



Hertzsprung-Russell Star Data

Belgian Phone Data

Stackloss Data Original Scale

Stackloss Data Square Root Scale

Modified Wood Gravity Data

Scottish Hill Racing Data

Figure 4-10: Forward plot of the R^2 for each data set

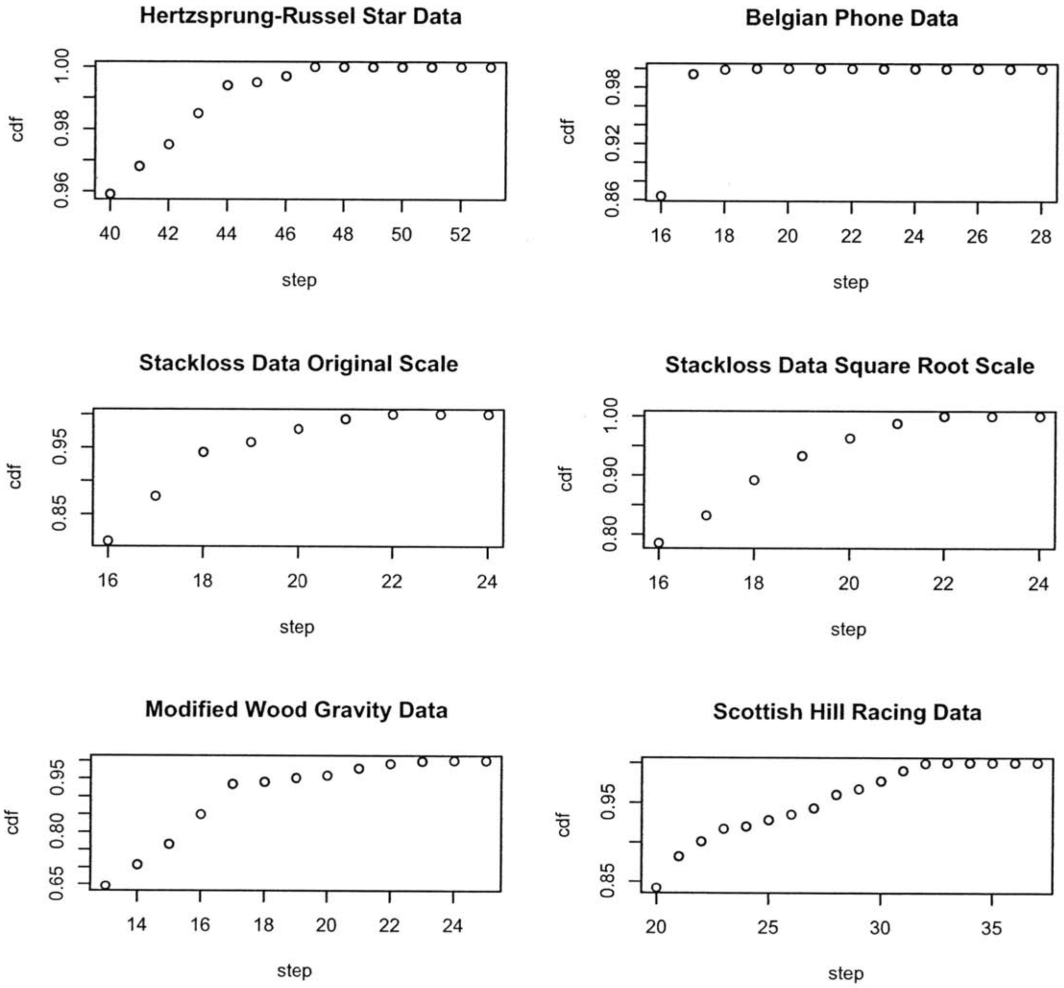


Figure 4-11: Empirical cumulative distribution function of elimination time of a random feature for each data set

Figure 4-11 by either observing sharp increases or pattern changes in the cdf, or by setting some sort of a threshold, α, and deciding that the correct model will be that for which the cdf equals α. This would imply that the probability that an irrelevant feature would be deemed relevant should be at most $1 - \alpha$. In each data set, the cdf flattens around the stopping time that previous literature agrees on. However, the appropriate cut-offs tend to vary across data sets, so while this method produces a useful, informative depiction of the model selection process, it is not as strong as the previous method as a stand-alone model selection method.

Lastly, Figure 4-10, which indicates a potential model by a sharp movement in the plot, is the least informative among the three plots, but demonstrates the application of forward search plots in our method.

While we had, for each of these data sets, a high breakdown method for selecting an initial set of data points, it would be informative to see what would happen if the initial set were perturbed. We study this by considering the non-outlying observations in each of the data sets we tested (based on our knowledge of the data sets), and then letting the initial set be a random sample of the non-outlying observations. For each step, Table 4.4 shows the number of times, out of 100, that the chosen points and features matched those chosen when a high breakdown method was used to select the initial observations.

The resulting stability of our method varies across data sets from excellent to less than optimal, depending on the characteristics of the data set. However, since our method is greedy and won't recover if an outlier is placed in the initial subset, there is certainly room for improvement in future work.

Table 4.4: Sensitivity to the starting set, measured by the number of random starts (out of 100) that result in the same model as the model with an optimized start. As described in the paper, outliers are excluded from the random sampling.

Data		Sensitivity by Step									
Hertzsprung-Russell	step	44	45	46	47	48	49	50	51	52	53
Star Data	sensitivity	1	6	42	53	65	79	79	81	83	88
Belgian Phone Data	step	19	20	21	22	23	24	25	26	27	28
	sensitivity	61	92	100	100	100	100	100	100	100	100
Stackloss Data	step	15	16	17	18	19	20	21	22	23	24
Original Scale	sensitivity	7	9	26	27	70	87	87	91	94	97
Scottish Hill Racing	step	28	29	30	31	32	33	34	35	36	37
Data	sensitivity	14	32	39	44	54	56	72	79	79	98
Modified Wood	step	16	17	18	19	20	21	22	23	24	25
Gravity Data	sensitivity	16	22	83	83	84	95	95	95	95	95

4.5 An Extension: Outlier Detection for Covariance Estimation via Feature Selection Algorithms

The idea of modeling an observation as an artificial feature, as presented in the previous sections and used by ([46], [43]), has established a connection between outlier detection in linear regression and feature selection on an augmented design matrix. It turns out that this connection carries over to the case of multivariate data, where we are interested in the estimation of a robust covariance matrix.

Robust covariance estimation is a central problem in robust statistics and has an extensive literature base (see for instance [42]). Finding an algorithm with desirable properties that works in every situation has proven to be very difficult, but several good algorithms which are useful in practice exist (for example [26]). The extension that we speak of in this section also does not attempt to work in every situation, but establishes the connection between robust covariance estimation and the mean-shift outlier model, thereby opening the door for more robust covariance algorithms to be developed. In particular, the connection makes it possible to use feature selection and dimension reduction algorithms of multivariate linear models in order to detect and control for outliers in multivariate data.

As this is not directly related to the central topic of the thesis, we do not delve fully into the implications of this. We do, however, briefly explore the procedure with backward selection used as one example of a feature selection algorithm and discuss the preliminary results we obtained. Our procedure, in this case, performs surprisingly well on some difficult data sets and resembles forward search for multivariate data with a simpler starting algorithm. In the rest of this section, we provide our methodology and a discussion of results on a selection of interesting and challenging data sets.

4.5.1 Methodology

We pose the robust covariance estimation problem as a feature selection problem as follows: We set our data matrix to be Y, construct X to be a column of 1's with an identity matrix appended to the right, and perform feature selection in the multivariate linear model of Y onto X, where Y is considered to be the matrix of realizations of the response variables, and X is the design matrix. The estimate of the covariance matrix of the error term for the regression is the algorithm's estimate of a robust covariance matrix of the original data.

The justification for our algorithm lies in the following two key relationships: 1. The estimated classical covariance matrix for multivariate data is the same as the estimated conditional covariance matrix of the multivariate linear model of Y onto X, where Y is the data matrix, and X is a column of 1's. 2. The mean-shift outlier model establishes that performing a regression with deleted observations yields the same results (estimated coefficient vector and estimated conditional covariance matrix) as performing a regression onto an augmented X matrix, where the augmented columns are dummy columns corresponding to the observations deleted in the first model (The proof of this is similar to the proof earlier in this chapter for the univariate case and is provided in the appendix).

One could, in principal, use this methodology to apply any feature selection algorithm to estimate a robust covariance matrix, though in this section, we explore backwards elimination as the feature selection algorithm of choice. In particular, we scale the data (Y matrix) by subtracting the column medians and dividing by the column mads, and then iteratively eliminate the least relevant feature of the X matrix, producing a ranked list of outliers.

The criterion for feature relevance is the determinant (product of eigenvalues) of the conditional covariance matrix when the feature is eliminated (the justification for this criterion is the likelihood ratio test). When there are ties, we use instead the product of the nonzero eigenvalues (this equals the determinant when all eigenvalues are nonzero). When there are still ties, we use the L1 norm of the coefficient matrix. As ties occur for the first few features eliminated, this specification is important.

4.5.2 Performance on Data Sets

We evaluate backward elimination on four real data sets - The Hertzsprung-Russell Star Data [57], Biochemical Data [59], Wine Data [24], and Swiss Heads Data [16] - as well as two realizations of the synthetic Barrow Wheel Benchmark [40].

The Hertzsprung-Russell Star Data set was described earlier in the chapter. The data set is typically used as a regression example, but we treat the response and explanatory data together as multivariate data in this paper, because it is nonetheless a good demonstration of influential, masked outliers, which can easily be visualized.

The Biochemical data set consists of two variables which are measurements of phosphate and chloride in the urine of 12 men with similar weights. Figure 4-13 contains a plot of the data. The data set is explored in [42]. Observation 3 (marked 12 in the plot) is considered to be an outlier.

The wine data is also explored in [42]. It contains, for each of 59 wines grown in the same region in Italy, the quantities of 13 constituents. There are known to be at least 7 masked outliers in this data set, the corresponding observation numbers being: $\{20, 22, 47, 40, 44, 46, 42\}$.

The swiss heads data is analyzed in [7] and contains six readings on the dimensions of heads of 200 twenty year old Swiss soldiers. The analysis suggests that 198 observations are clean and two observations: 104 and 111 are outliers.

Lastly, the barrow wheel benchmark is simulated from a multivariate distribution,

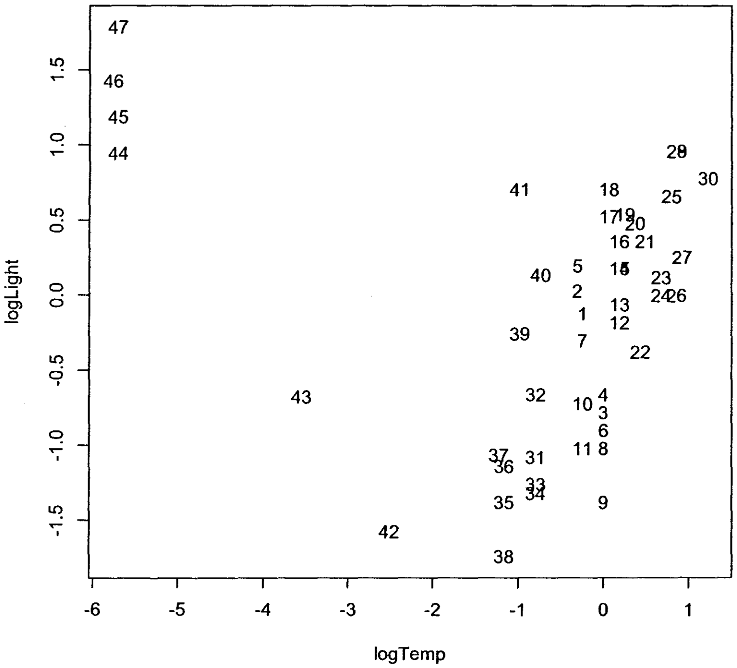


Figure 4-12: Plot of Hertzsprung-Russell Star Data. The number indicates when the corresponding feature was eliminated in the algorithm.

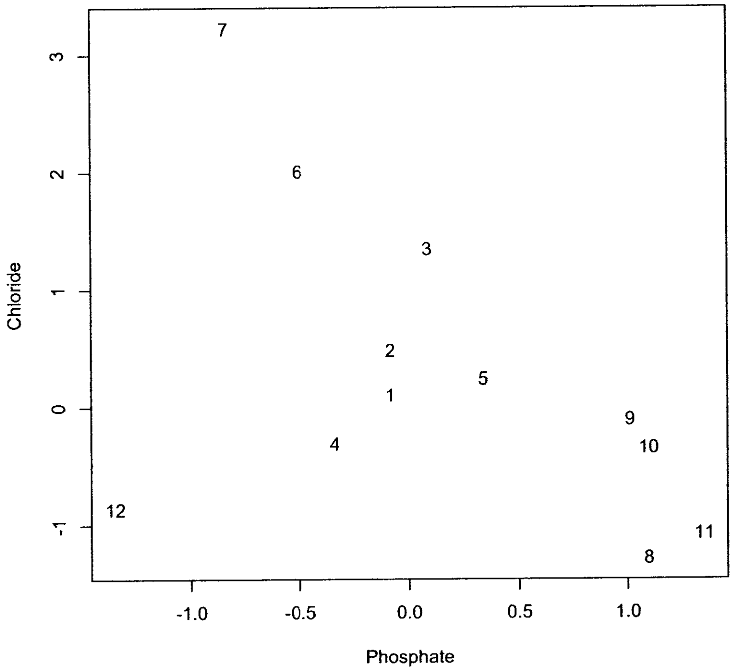


Figure 4-13: Plot of Biochemical Data. The number indicates when the corresponding feature was eliminated in the algorithm.

which has been proposed as a benchmark to evaluate robust multivariate methodologies. It is a mixture of a flat normal distribution, contaminated with a portion $\epsilon = 1/p$ of gross errors concentrated near a one-dimensional subspace, where p is the specified dimensionality. The entire data set is rotated in a way, such that except in the case where $p = 2$, bivariate plots would not reveal the structure of the data. We tested the barrow wheel data set with $n = 50, p = 2$ and $n = 500, p = 6$. In the barrow wheel data set, the outliers are the last n/p observations generated.

For each of the four real data sets and for the barrow wheel data set, the outliers specified above were among the last features eliminated in our algorithm, indicating that our algorithm correctly identified them as the most outlying points. Taking a closer look, Figures 4-12 and 4-13 show the plots of the Hertzsprung-Russell Star Data and the Biochemical with the number of the point representing when in the algorithm the dummy variable corresponding to the point was eliminated (hence, the highest number is the most outlying point, as deemed by the algorithm). Meanwhile, Figure 4-14 shows a similar plot for the barrow wheel benchmark, when $p = 2$. These figures demonstrate that our algorithm does well from the start (when the entire identity matrix is appended).

It should also be noted that although our algorithm identified the seven prominent masked outliers in the wine data, the description of the wine data indicated the presence of some other minor outliers, some of which the backwards selection algorithm eliminated early in its search. Hence, our algorithm is not perfect, but we tested it in the harder case with the assistance of no prior knowledge or heuristic. When we provided our algorithm with a hint by appending an extra row corresponding to adding prior knowledge that the data should be centered near the coordinate-wise median, the algorithm worked satisfactorily for the data set.

Based on the results, the extension we document needs more extensive thought in order to compete with premier outlier detection algorithms, but produces noteworthy results on its own and implies other procedures based on other feature selection algorithms, which may prove to be even better.

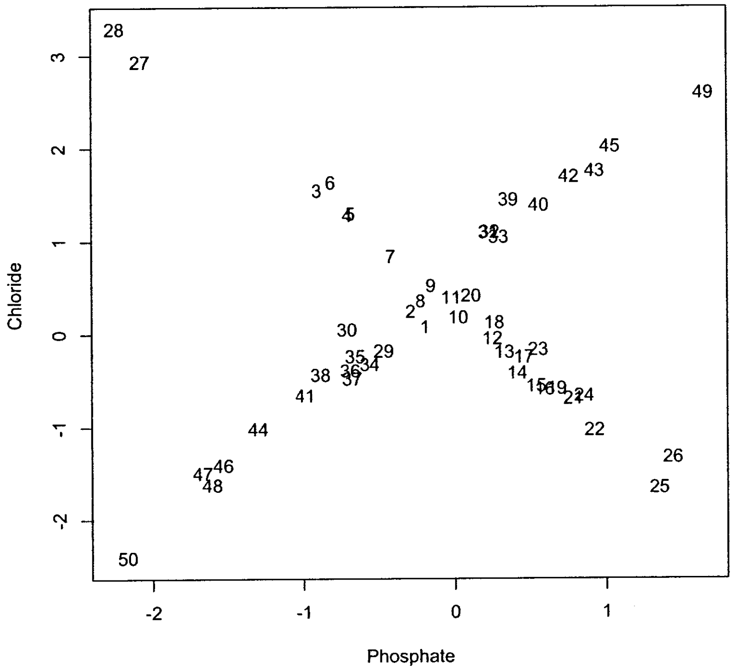


Figure 4-14: Plot of Barrow Wheel Data. The number indicates when the corresponding feature was eliminated in the algorithm.

4.6 Discussion

We have proposed a diagnostic method for simultaneous feature selection and outlier detection, which performs quite promisingly on the data sets we have used. Our method, like forward search, allows one to assess the effect of outliers on linear regression inferences. However, it also allows us to assess feature diagnostics, and allows us to visualize a sequence of models of interest. We can use any of the methods discussed in this paper to decide when to stop the sequencing. In addition, we can easily incorporate prior knowledge and hence control for issues arising because of rank-deficient matrices.

We have also described an interesting extension of the ideas of forward search, which poses outlier detection in multivariate data as a feature selection problem for a multivariate linear model. We hope that it will inspire the development of other effective methods for detecting outliers in multivariate data.

Our method still has the weakness that it is difficult to find a computationally efficient way to choose a good initial subset. This will be the direction of future work.

Chapter 5

Contributions and Future Work

In this thesis, we made contributions in data visualization and regression diagnostics.

In data visualization, we presented an alternative way to understand plots. The methodology we created highlights the notion that several common plots that are frequently used in every day life are in fact subject to sampling variability. We proposed a simple method for representing this sampling variability. This method will help make plots more valid inferential tools, in the same way that a confidence interval makes the sample mean a more valid inferential tool.

Additionally, our method can be generalized to several types of plots, including some, for which some methods for representing sampling variability have been developed. In these cases, our method provides an alternative representation, which is also simple and effective. Lastly, our method can be used to help verify distributional assumptions and models in Bayesian analysis.

Our alternative way to understand plots also brought forth a way to simplify large scatterplot matrices, when most of the cells in the scatterplot matrix contain small variations on one of a few key plots. Our contribution here will lead to simpler representations of scatterplot matrices and hence more tools for the visualization of multivariate data.

In regression diagnostics, we proposed a method for producing a sequence of valid selections of observations and features, in order to assist an analyst in robust model selection. Our method executes simultaneous feature selection and outlier detection, and creates plots which elucidate the effect of each individual observation and feature on the analysis. This is important, since a point that one analyst may consider to be an outlier, could be the most important point in the data set to another analyst. This contribution will allow for analysts to better select relevant variables in the presence of outliers. Additionally, the idea behind this contribution provides a new way of thinking about outlier detection in multivariate data, by formulating it as a feature selection problem.

In each of the methods that we have created in this thesis, several directions for future work have been created. Our visualization methodology probably extends to more plots than we have described in the corresponding chapters. Additionally, improvements for selecting the 95 % confidence intervals are likely possible. Meanwhile, researchers are still looking for an efficient way to select an initial subset, with which to start the forward search algorithm. Additionally, feature selection algorithms other than backward selection, for the problem of selecting features in a multivariate linear model, may yield improved computation time and better accuracy.

Appendix A

A More General Mean-Shift Outlier Model

We state, here, the elaboration and proof for the mean-shift outlier model, explained in Section 4.3.1, for a more general case. The results can be applied to justify the methodology developed in Section 4.5. As mentioned in the literature review, this equivalence has been proved in cases beyond linear regression (see for instance [23]).

Let our notation be defined as follows:

Y is a given n by r matrix of realizations for the response variable; X is a given n by p matrix of realizations for the predictor variables; U is a set of indices with cardinality t; $J = U^c$ is the set of indices complementary to U; β is a matrix of decision variables corresponding to the p by r matrix of the coefficients of the predictor variables; α is a matrix of decision variables corresponding to the t by r matrix of the coefficients of artificial variables which will be added in the result below; L_i is a loss function which maps two vectors onto a real number for each i; I is the identity matrix.

Furthermore, we let "|" denote horizontal matrix concatenation and we denote submatrices and row vectors as we did in Section 4.3.1. That is: where K and L are sets of indices, $*$ is the set which contains every index, A is a given matrix, and b is a given vector, we let $A_{[K,L]}$ denote the submatrix of A containing the rows with indices in K and the columns with indices in L. On the other hand, we let $A_{(K,L)}$ be the submatrix of A which contains all rows and columns of A except rows with indices in K and columns with indices in L respectively. $A_{[K,L)}$ and $A_{(K,L]}$ are defined similarly. Similarly, let $b_{[K]}$ denote the elements of b with indices in K and let $b_{(K)}$ denote the elements of b whose indices are not in K. Lastly, we use the shorthand of a_i' to denote row i of matrix A and b_j to denote element j of vector b. Hence, based on our notation, $a_i' = A[i,*]$ and $b_j = b[j]$. Meanwhile, $A(i,*]$ would denote all rows of A except a_i'.

The following result then holds:

Claim: If $\forall i$, we have that $\exists c_i \in R^r$ with $B_i = \lim_{b \to c_i} L_i(y_i, b) \leq L(y_i, z) \forall z$, then the following condition holds:

$$\arg\min_\beta \inf_\alpha \sum_{i=1}^n L_i(y_i^T, x_i^T\beta + I[i,U]\alpha) = \arg\min_\beta \sum_{i \in J} L_i(y_i^T, x_i^T\beta)$$

Proof:

$$\begin{aligned}
& \arg\min_\beta \inf_\alpha \sum_{i=1}^n L_i(y_i^T, x_i^T\beta + I[i,U]\alpha) \\
= \; & \arg\min_\beta (\inf_\alpha \sum_{i \in J} L_i(y_i^T, x_i^T\beta) + \inf_\alpha \sum_{i \in U} L_i(y_i^T, x_i^T\beta + \alpha_i)) \\
= \; & \arg\min_\beta (\sum_{i \in J} L_i(y_i^T, x_i^T\beta) + \sum_{i \in U} \inf_{\alpha_i} L_i(y_i^T, x_i^T\beta + \alpha_i)) \\
= \; & \arg\min_\beta (\sum_{i \in J} L_i(y_i^T, x_i^T\beta) + \sum_{i \in U} B_i) \\
= \; & \arg\min_\beta (\sum_{i \in J} L_i(y_i^T, x_i^T\beta))
\end{aligned}$$

The second equality follows by plugging in the actual values of z_i. Notice that if $B_i = 0 \forall i$, we also have that the minimums are the same.

The multivariate linear model (linear regression with multiple responses) is a special case of the above setup, where $L_i(y_i^T, u_i^T) = (y_i^T - u_i^T)\Sigma^{-1}(y_i - u_i)$ with Σ the conditional covariance of the response variables. Here, $c_i = y_i^T$, so $\alpha_i = y_i^T - x_i^T\beta$, and $B_i = 0$. It follows that deleting observations and adding dummy columns gives the same estimate of β in the case of multiple output linear regression. In addition, the minimums of the losses over the unknown coefficient vectors, which we will denote by $S(\Sigma)$ are the same in the two cases as well since $B_i = 0$.

The estimate of the covariance of the residual terms is generally estimated by $\hat{\Sigma}_\epsilon = \arg\min_\Sigma(S(\Sigma))/(n-p)$. Although the number of observations and variables differ in the case of deleting observations and the case of adding dummy columns, the number of observations minus the number of variables, $n-p$, is the same in both cases, so the estimate of $\hat{\Sigma}$ is also the same in both cases.

It should be noted, however, that the result stated in this section is a reasonably general result and hence, the idea of modeling observations as features can extend to several other types of models, such as logistic regression models.

Appendix B

Demonstration of Earth Mover's Distance

In order to demonstrate the Earth Mover's Distance's conception of distance, we show four sets of plots in this appendix. For the first two cases, we work with the data in Figure 2-3.

The first set of plots is Figure B-1. In this sequence of plots, we took the 950 bootstrapped plots of the data in Figure 2-3 which are closest to the central bootstrapped plot and then ordered them, so that they are as close to each other as possible, in terms of Earth Mover's Distance, using a heuristic to the traveling salesman problem. The sequence of plots show the first 25 of these plots.

For the sake of contrast, Figure B-2, the second set of plots, shows a random 25 bootstrapped plots. As one can see from this example, the Earth Mover's Distance tends to agree with our perception of distance.

To consider a second example, we consider the histogram example with the return data, shown in Figure 2-12. Figure B-3 shows the first 25 histograms, ordered so that they are as close to each other as possible, in terms of Earth Mover's Distance. Meanwhile, Figure B-4 shows 25 random bootstrapped histograms.

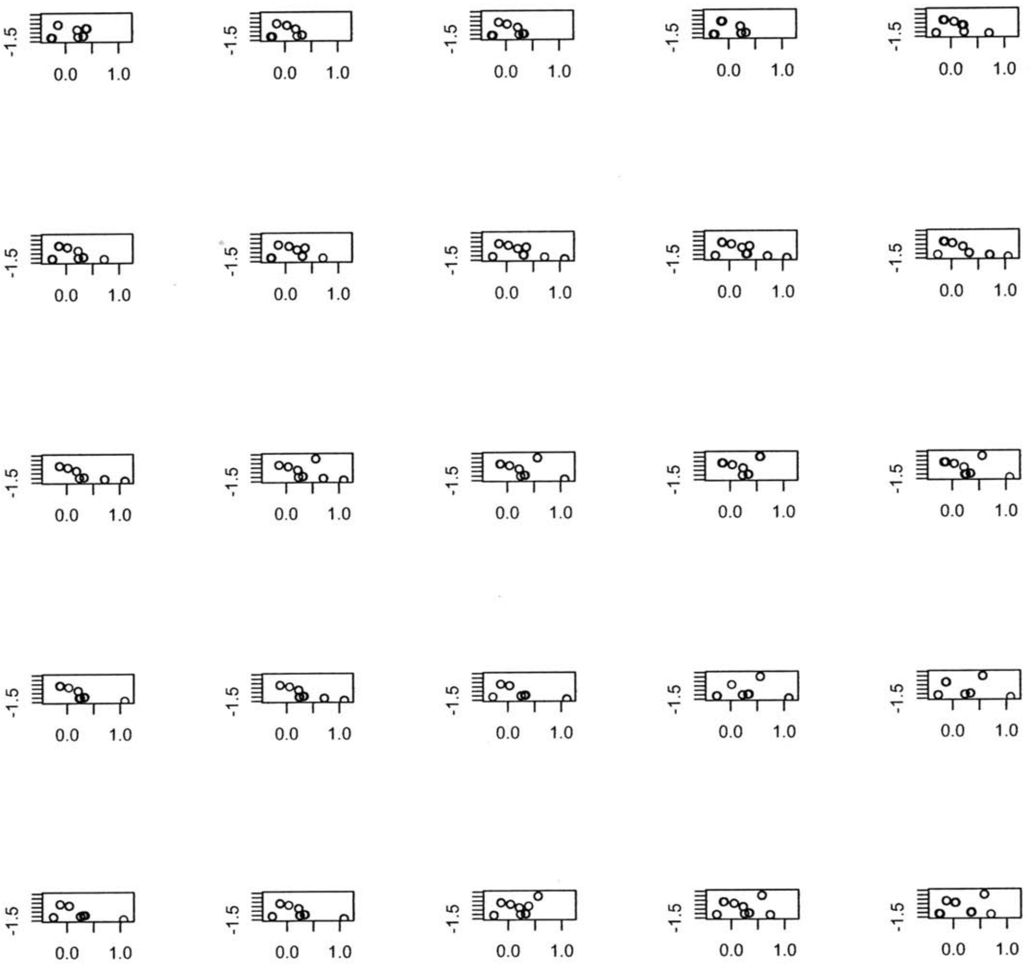


Figure B-1: The first 25 bootstrapped plots, when they are ordered by applying a heuristic to the travelling salesman problem with the Earth Mover's Distance metric

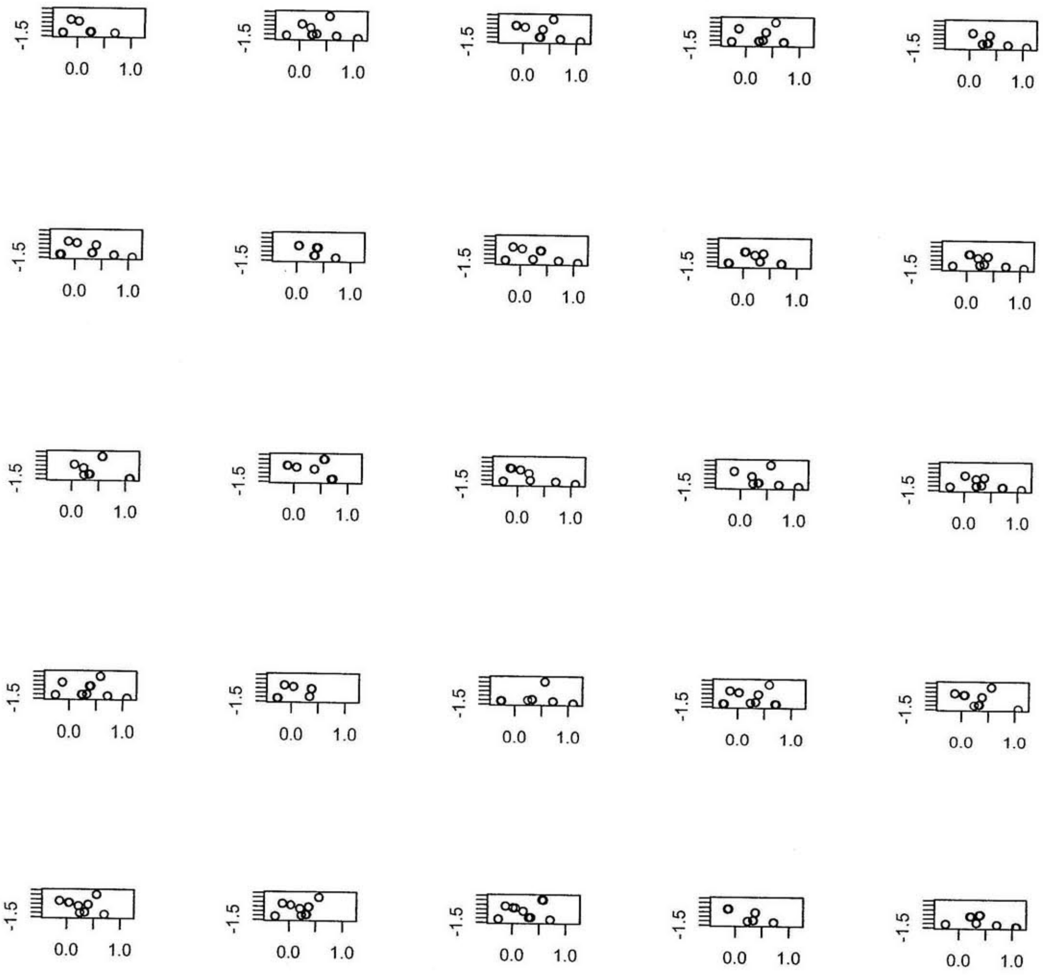


Figure B-2: A random sample of 25 bootstrapped plots

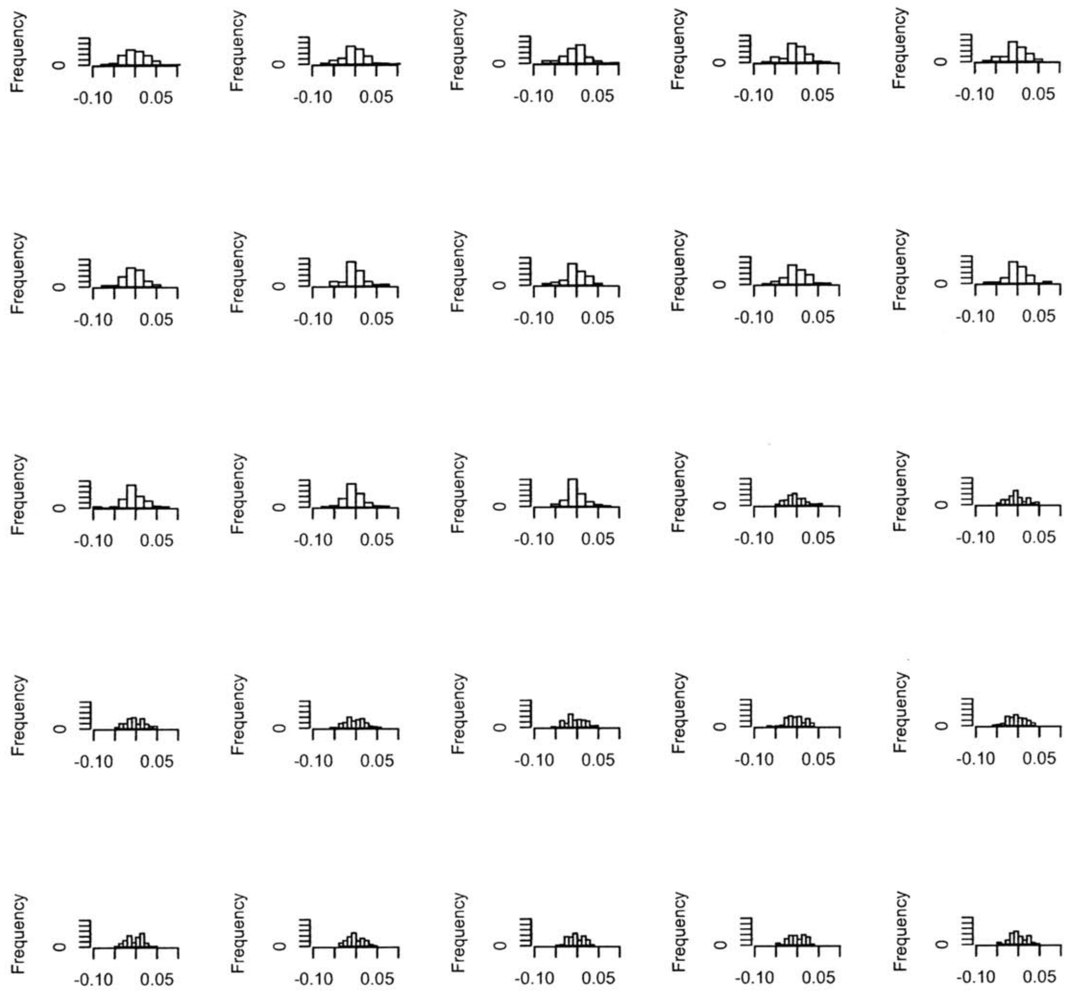


Figure B-3: An ordered sample of 25 bootstrapped histograms

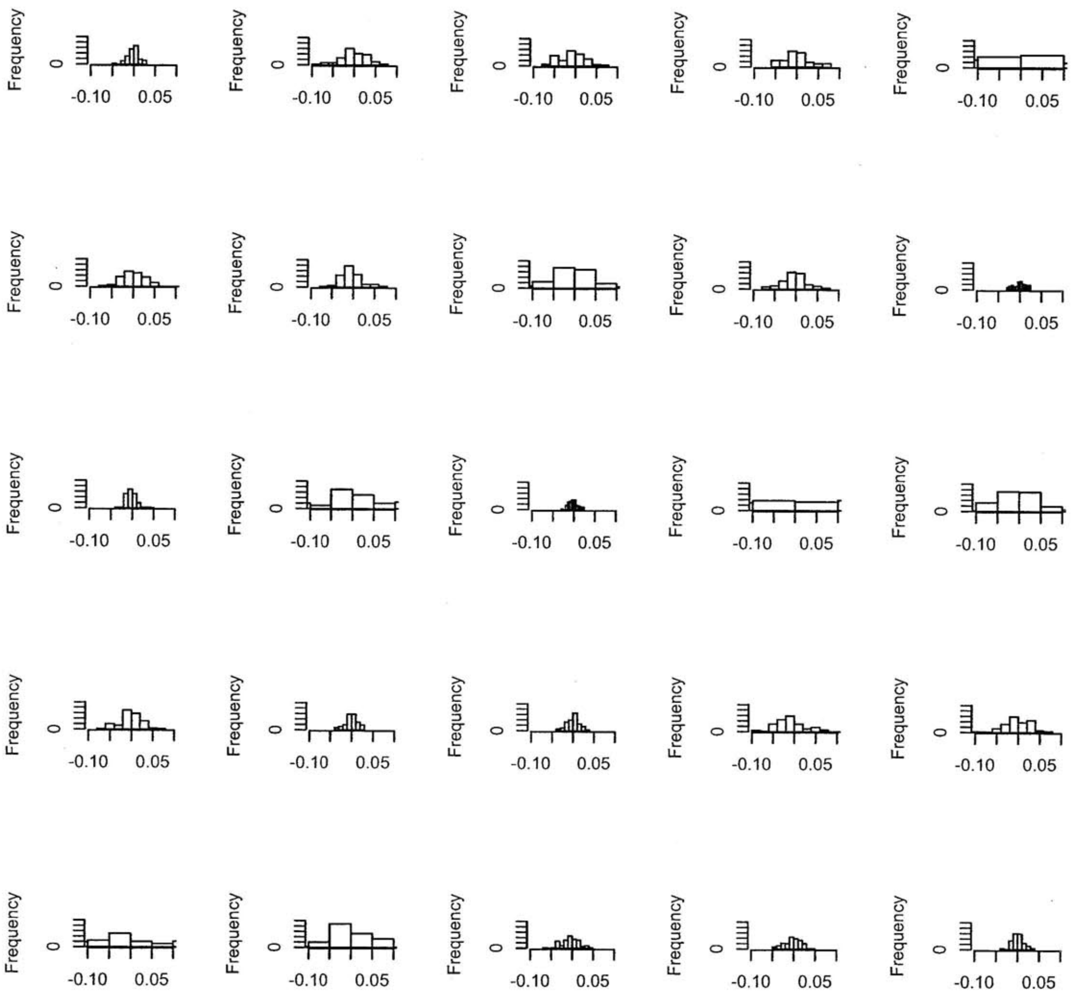


Figure B-4: A random sample of 25 bootstrapped histograms

Bibliography

[1] M. Ankerst, S. Berchtold, and D.A. Keim. Similarity clustering of dimensions for an enhanced visualization of multidimensional data. pages 52–60. IEEE Symposium on Information Visualization, 1998.

[2] F.J. Anscombe. Graphs in statistical analysis. *American Statistician*, 27:17–21, 1973.

[3] A.C. Atkinson. Comments on 'influential observations, high leverage points, and outliers in linear regression'. *Statistical Science*, 1(3):397–402, 1986.

[4] A.C. Atkinson. Masking unmasked. *Biometrika*, 73(3):533–541, 1986.

[5] A.C. Atkinson and Riani M. Forward search added-variable t-tests and the effect of masked outliers on model selection. *Biometrika*, 89(4):939–946, 2002.

[6] A.C. Atkinson and M. Riani. *Robust Diagnostic Regression Analysis.* Springer-Verlag, first edition, 2000.

[7] A.C. Atkinson and M. Riani. *Exploring Multivariate Data with the Forward Search.* Springer-Verlag, NY, NY, 2004.

[8] D.A. Belsley, E. Kuh, and R.E. Welsch. *Regression Diagnostics.* Wiley, Hoboken, NJ, 1980.

[9] I. Borg and P. Groenen. *Modern Multidimensional Scaling: theory and applications.* Springer, New York, NY, 2005.

[10] K.A. Brownlee. *Statistical Theory and Methodology in Science and Engineering.* Wiley, New York, NY, 2000.

[11] F. Chateau and L. Lebart. Assessing sample variability in visualization techniques related to principal component analysis: bootstrap and alternative simulation methods. In A. Prats, editor, *Computational Statistics*, pages 205–210, 1996.

[12] R. Diaz-Uriarte and S. Andres. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(3), 2006.

[13] S. dos Santos and K. Brodlie. Gaining understanding of multivariate and multidimensional data through visualization. *Computers and Graphics*, 28:311–325, 2004.

[14] B. Efron. Bootstrap methods: another look at the jackknife. *Annals of Statistics*, 7:1–26, 1979.

[15] B. Efron and R.J. Tibshirani. *An Introduction to the Bootstrap.* Chapman & Hall, Boca Raton, FL, 1994.

[16] B. Flury and H. Riedwyl. *Multivariate Statistics: A Practical Approach.* Chapman and Hall, London, 1988.

[17] W. Fung. Unmasking outliers and leverage points: A confirmation. *Journal of the American Statistical Association*, 88:515–519, 1993.

[18] A. Gelman. Exploratory data analysis for complex models. *Computational and Graphical Statistics*, 13(4):755–779, 2004.

[19] A.S. Hadi. Identifying multiple outliers in multivariate data. *Journal of Royal Statistical Society, Series B*, 54(3):761–771, 1992.

[20] A.S. Hadi and J.S. Simonoff. Procedures for the identification of multiple outliers in linear models. *Journal of the American Statistical Association*, 88(424):1264–1272, 1993.

[21] W. Haerdle. *Applied Non-parametric Regression.* Oxford University Press, 1990.

[22] P. Hall. *The Bootstrap and Edgeworth Expansion.* Springer-Verlag, 1992.

[23] Haslett and K. Hayes. Residuals for the linear model with general covariance structure. *J. Roy. Statist. Soc. Ser. B*, 60:201215, 1998.

[24] S. Hettich and S.D. Bay. *The UCI KDD Archive, kdd.ics.uci.edu.* University of California, Department of Information and Computer Science, Irvine, CA, 1999.

[25] J. Hoeting, A.E. Raftery, and D. Madigan. A method for simultaneous variable selection and outlier identification in linear regression. *Computational Statistics*, 22:252–270, 1996.

[26] M. Hubert, P.J. Rousseeuw, and S.V. Aelst. High-breakdown robust multivariate methods. *Statistical Science*, 23(1):92–119, 2008.

[27] C. B. Hurley. Clustering visualizations of multidimensional data. *Journal of Computational and Graphical Statistics*, 13:129–133, 2004.

[28] Yang J., Ward M. O., E. A. Rundensteiner, and S. Huang. Visual hierarchical dimension reduction for exploration of high dimensional datasets. Grenoble, France, 2003. Joint IEEE/EG Symposium on Visualization.

[29] J. Khan, S. Van Aelst, and R.H. Zamar. Robust linear model selection based on least angle regression. *Journal of the American Statistical Association*, 102(480):1289–1299, 2007.

[30] M.E. Khan. Updating inverse of a matrix when a column is added/removed. Technical report, University of British Columbia, 2008.

[31] H.A.L. Kiers and P.J.F. Groenen. Visualizing dependence of bootstrap confidence intervals for methods yielding spatial configurations. In Sergio Zani, Andrea Ceroli, Marco Riani, and Maurizio Vichi, editors, *Studies in Classification, Data Analysis, and Knowledge Organization*, pages 119–26, Vichi, Berlin, 2006. Springer.

[32] S. Kim and W.J. Krzanowski. Detecting multiple outliers in linear regression using a cluster method combined with graphical visualization. *Computational Statistics*, 22:109–119, 2007.

[33] S. Kim, S.H. Park, and W.J. Krzanowski. Simultaneous variable selection and outlier identification in linear regression using the mean-shift outlier model. *Journal of Applied Statistics*, 35(3):283–291, 2008.

[34] D. Komura, H. Nakamura, S. Tsutsumi, H. Aburatani, and S. Ihara. Multidimensional support vector machines for visualization of gene expression data. pages 175–179, Nicosia, Cyprus, 2004. ACM symposium on Applied computing.

[35] B. Korte and J. Vygen. *Combinatorial Optimization: Theory and Algorithms.* Springer, New York, NY, 2000.

[36] H.W. Kuhn. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2:83–97, 1955.

[37] J. Landgrebe, W. Wurst, and G. Welzl. Permutation-validated principal components analysis of microarray data. *Genome Biology*, 3:research 0019.1–0019.11, 2002.

[38] E.L. Lawler, J.K. Lenstra, A.H.G. Rinnooy Kan, and D.B. Shmoys. *The Traveling Salesman Problem.* Wiley, NY, 1985.

[39] E. Levina and P. Bickel. The earthmover's distance is the mallows distance: Some insights from statistics. pages 251–256, Vancouver, Canada, 2001. ICCV.

[40] M. Maechler and W. Stahel. Robust scatter estimators - the barrow wheel benchmark. Parma, Italy, 2009. International Conference on Robust Statistics.

[41] H. Markowitz. Portfolio selection. *Journal of Finance*, 7:77–91, 1952.

[42] R.A. Maronna, R.D. Martin, and V.J. Yohai. *Robust Statistics: Theory and Methods.* John Wiley & Sons Ltd., West Sussex, England, 2006.

[43] L. McCann. *Robust Model Selection and Outlier Detection in Linear Regression.* PhD thesis, MIT, 2005.

[44] L. McCann and R.E. Welsch. Diagnostic data traces using penalty methods. In J. Antoch, editor, *Comp. Stat*, pages 1481–1488, Heidelberg, 2004. Computational Statistics, Physica-Verlag.

[45] L. McCann and R.E. Welsch. Robust variable selection using least angle regression and elemental set sampling. *Computational Statistics and Data Analysis*, 52:249–257, 2007.

[46] S. Morgenthaler, R.E. Welsch, and A. Zenide. Algorithms for robust model selection in linear regression. In M. Hubert, G. Pison, A. Struyf, and S. Van Aelst, editors, *Theory and Applications of Recent Robust Methods*, pages 195 – 206. Birkhauser, Basel, Switzerland, 2004.

[47] W. Muller, T. Nocke, and H. Schumann. Enhancing the visualization process with principal component analysis to support the explorationof trends. APVIS, 2006.

[48] T. Nguyen. *O.R. Techniques in Portfolio Optimization, Robust Statistics, and Hedge Fund Strategies.* PhD thesis, MIT, 2009.

[49] *forward: Forward Search. R package version 1.0.2.*, 2009. Originally written for S-Plus by: Kjell Konis and Marco Riani. Ported to R by Luca Scrucca <luca@stat.unipg.it>.

[50] S. Peleg, M. Werman, and H. Rom. A unified approach to the change of resolution: Space and gray-level. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11:739–742, 1989.

[51] M. Riani and A.C. Atkinson. Fast calibrations of the forward search for testing multiple outliers in regression. *Advances in Data Analysis and Classification*, 1(2):123–141, 2007.

[52] M. Riani, A.C. Atkinson, and A. Cerioli. Finding an unknown number of multivariate outliers. *Journal of the Royal Statistical Society, series B*, 71(2):447–466, 2009.

[53] E. Ronchetti. Robust model selection in regression. *Statistics and Probability Letters*, 3:21–23, 1985.

[54] E. Ronchetti, C. Field, and W. Blanchard. Robust linear model selection by cross-validation. *Journal of the American Statistical Association*, 92:1017–1023, 1997.

[55] E. Ronchetti and R. G. Staudte. A robust version of mallows c_p. *Journal of the American Statistical Association*, 89:550–559, 1994.

[56] P.J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79:871–880, 1984.

[57] P.J. Rousseeuw and A.M. Leroy. *Robust Regression and Outlier Detection.* John Wiley & Sons, Inc., New York, NY, 1987.

[58] Y. Rubner, C. Tomasi, and L.J. Guibas. A metric for distributions with applications to image databases. pages 59–66, Bombay, India, 1998. IEEE International Conference on Computer Vision.

[59] G.A.F. Seber. *Multivariate Observations.* John Wiley & Sons, Inc., NY, NY, 1984.

[60] S. Shirdhonkar and D. Jacobs. Approximate earth movers distance in linear time. CVPR, 2008.

[61] H. Stoppiglia and G. Dreyfus. Ranking a random feature for variable and feature selection. *Journal of Machine Learning Research*, 3:1399–1414, 2003.

[62] H. Theil and A.S. Goldberger. On pure and mixed estimation in economics. *International Economic Review*, 2:65–78, 1961.

[63] J. W. Tukey and P.A. Tukey. Computer graphics and exploratory data analysis: An introduction. Fairfax, VA, 1985. Sixth Annual Conference and Exposition: Computer Graphics85.

[64] S. Urbanek. Visualizing trees and forests. In *Handbook of Data Visualization*, pages 243–264. Wiley, NY, 2008.

[65] L. Wilkinson, A. Anand, and R. Grossman. Graph-theoretic scagnostics. pages 157–174. IEEE Symposium on Information Visualization, 2005.

[66] F.W. Young, P.M. Valero-Mora, and M. Friendly. *Visual Statistics: Seeing Data with Dynamic Interactive Graphics.* Wiley Series in Probability and Statistics, Hoboken, NJ, 2006.