Author/s:
Miller, Hugh Richard

Title:
Statistical methods for the analysis of high-dimensional data

Date:
2010

Citation:
Miller, H. R. (2010). Statistical methods for the analysis of high-dimensional data. PhD thesis, Science, Department of Mathematics and Statistics, The University of Melbourne.

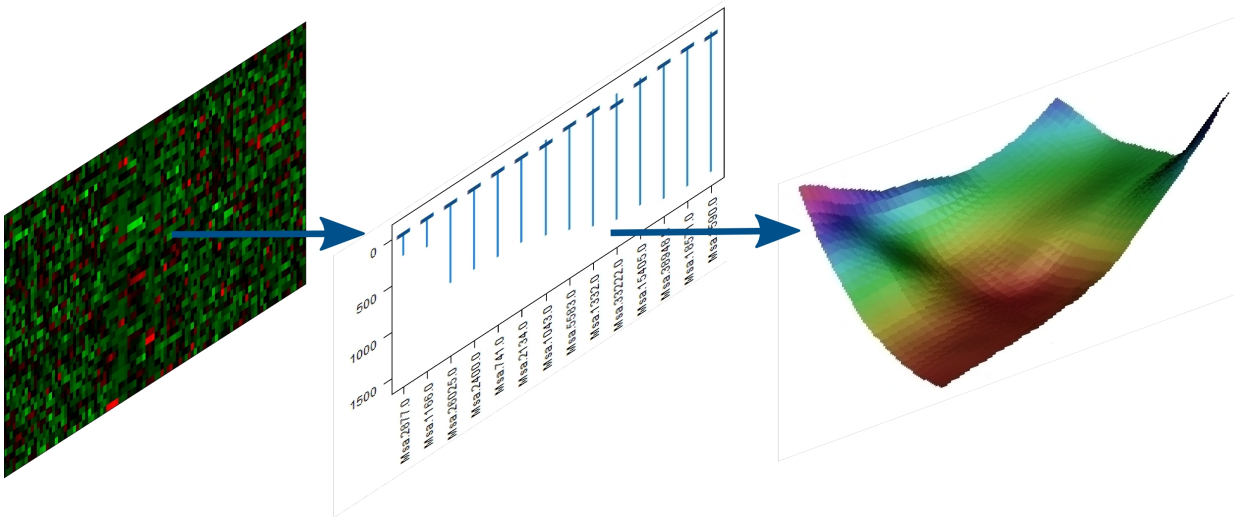
Persistent Link:
https://hdl.handle.net/11343/35462

Statistical Methods for the Analysis of High-Dimensional Data

Hugh Richard Miller

Department of Mathematics and Statistics
The University of Melbourne

August, 2010



Submitted in total fulfillment of the requirements
of the degree of Doctor of Philosophy

Printed on archival quality paper

Abstract

High-dimensional statistics has captured the imagination of many statisticians worldwide, because of its interesting applications as well as the unique challenges faced. This thesis addresses a range of problems in the area, integrated into an overall framework. Problems explored include how to effect feature selection and test variable relationships, particularly when important nonlinearities may be present; how to create a nonparametric model that adapts to the observed importance of different variables in a dataset; how to assess the reliability of a ranking, for example how to list genes in order of importance to a disease, and how to determine circumstances where we can expect parts of this ranking to be reliable; how to perform hypothesis tests on extremes of populations; and how to incrementally build a multivariate model in high-dimensional situations while simultaneously protecting against overfitting. These problems are addressed through both theoretical and numerical means.

Declaration

This is to certify that

1. the thesis comprises only my original work towards the PhD except where indicated in the Preface
2. due acknowledgment has been made in the text to all other material used,
3. the thesis is less than 100,000 words in length, exclusive of tables, figures and the Bibliography.

Hugh Miller

Preface

This thesis was written under the supervision of Prof. Peter Hall, with Dr Owen Jones and Dr Aurore Delaigle associate supervisors. Upon reading it, one may find that it reads as a sequence of semi-independent articles, tied together in the introduction into an overall framework. This is precisely the case. Each of Chapters 2 to 8 represent a paper at some stage of review and publication, and all are joint work with Peter Hall.

- Chapter 2 is a modified version of Hall and Miller (2009a), which has been published in the Journal of Computational and Graphical Statistics.
- Chapter 3 is a modified version of Hall and Miller (2010b), which is currently under peer review.
- Chapter 4 is a modified version of Miller and Hall (2010), which is currently under peer review.
- Chapter 5 is a modified version of Hall and Miller (2009b), which has been published in the Annals of Statistics.
- Chapter 6 is a modified version of Hall and Miller (2010c), which has been accepted for publication in the Annals of Statistics.
- Chapter 7 is a modified version of Hall and Miller (2010a), which has been accepted for publication in Biometrika.
- Chapter 8 is a modified version of Hall and Miller (2010d), which is currently under peer review.

The introduction addresses the general area of high-dimensional statistics, including an introduction to the specific problems treated in each chapter. Thereafter each chapter will typically be a combination of background material, description of the problem and approach, followed by theoretical and numerical results. Proofs will sometimes be deferred to the end of a chapter so as to not distract from the main argument. Relevant literature will be introduced throughout the whole thesis. A range of real data sets are used and are introduced when they first arise.

Acknowledgments

I would like to acknowledge my debt to my supervisor, Peter Hall. His vision, encouragement and tireless work ethic has been inspiring and greatly helped the speed and quality of the research.

Thanks also goes to the staff and students of the Department of Mathematics and Statistics at Melbourne University, for their company, insights and distractions.

To my laptop, which spent a fair amount of time simulating at full capacity. Thanks for only completely frying once.

Finally thanks go to my wife for her constant love and support. Hopefully the next baby, while less metaphorical, is even more fun!

Contents

List of Figures

List of Tables

Notation

Notation will be introduced as needed, so the following list of the most persistent notation is (hopefully) largely superfluous. However readers may still find it useful for reference.

I	Denotes the standard indicator function. For event $\mathcal{E}$, $I(\mathcal{E}) = 1$ if $\mathcal{E}$ is true, and zero otherwise.
$\text{mm}_{(j)}$	This expression is used to denote either the minimum or maximum of a set indexed by j. This permits extra generality in Chapter 7.
Φ	Denotes the cumulative distribution of the standard normal distribution.
ρ_A, ρ_S	Chapter 3 defines the asymmetric and symmetric versions of generalised correlation for assessing variable relationships.
θ_j	Denotes a parameter corresponding to the jth component, $1 \leq j \leq p$, which generally requires estimation. See Chapters 5, 6 and 7.
X, $X^{(j)}$	X will generally denote the main p-dimensional observation vector under study. The absence of a subscript usually denotes that X is regarded as a random variable, with the dimensions indexed by the superscript in brackets
X_i, X_{ij}	When we move from the random variable X to n training observations, we index these as p-vectors $X_1, \ldots, X_i, \ldots, X_n$. The jth element of X_i is denoted X_{ij}, but very occasionally as $X_i^{(j)}$, when it is more natural to use this notation.
$\mathcal{X}_j$	In Chapter 5 we allow the components, indexed by j for $1 \leq j \leq p$ to possibly have different sizes n_j, and so we introduce a modified notation to allow this.
Y	The response random variable, which we shall seek to explain using X, will always be 1-dimensional. In Chapters 2 and 4 this response will be continuous, while Chapter 8 examines the case of $Y \in \{0, 1\}$ categorical
Y_i	When we move from theoretical Y to training examples we denote this $Y_1, \ldots, Y_n$.
$O(a_n)$, $o(a_n)$	This is the usual order notation. If $b_n = O(a_n)$, for sequences a_n and b_n indexed by n, then there exists constant $C > 0$ and n_0 such that for $n > n_0$, $b_n < Ca_n$. Similarly if $b_n = o(a_n)$, then for any $\epsilon > 0$ there exists n_0 such that $n > n_0$ implies $b_n/a_n < \epsilon$.

Chapter 1

Introduction

1.1 What is high-dimensional statistics?

Many trends in statistics have been driven by the types of datasets produced by industry and science, particularly when they represent new and challenging problems. This is undoubtedly true of the topic referred to as "high-dimensional" statistics. It refers to situations where there are many variables or components available for use in any modelling or analysis. A model here is any statistical framework built using the data, and most commonly refers to a predictive model, where the variables are used to make predictions about a certain event, based on previous data.

The first main source of these high-dimensional datasets is industry, particularly organisations with large customer databases. The low cost of digital storage now allows massive amounts of information concerning each customer to be collected relatively cheaply. Features of this database are then extracted and statistically analysed. An organisation may seek to make use of a database in order to predict customer behaviour, better understand changes in their customer base, or even to specifically target products. Take, for example, the dataset used in the 2009 KDD Cup[1], which forms part of an annual data mining competition. The training dataset (that used for building statistical models) is comprised of 50,000 customers of a French mobile telephone company. For each of these customers there are 15,000 variables, or separate pieces of information, available. These variables encode virtually everything the company knows about the customer, from demographic details, to current and past products used, to telephone usage history. In this particular problem, the task was to use this information to predict a customer's likelihood of changing to a competitor, as well as whether a customer would upgrade their telephone plan or respond to

[1] www.kddcup-orange.com

marketing material. The abundance of information allowed these predictions to be performed with surprising accuracy.

Another key source of high-dimensional datasets is biology and related sciences. Modern equipment allows for simultaneous measurement of many different components, which are then collected in a single dataset. For example, genetic microarray experiments measure the relative activity (or "expression level") of thousands of genes from a single tissue sample. This allows a researcher to look for genes that are particularly active in one group of experiments compared to another. For instance, patients with a particular cancer may have a gene which shows much more activity than that observed in non-cancer patients. This would prompt investigators to focus on this particular gene with the hope of better understanding the cancer.

One key difference between the above two examples is the number of observations. While the order of the number of variables or components is similar for each (in the thousands), the customer database has 50,000 records to draw from, while a microarray experiment will rarely involve more than a hundred samples. This drastically affects what is able to be achieved for each problem. If there are many observations, a highly complex model incorporating hundreds of weak effects and interactions could be built, with reasonable confidence that most of them are genuinely important effects. Conversely, when there is a small number of observations the analysis is typically forced to be simpler, with an emphasis on reliably detecting just the main effects or variables.

It is worth mentioning here that the problems addressed in the context of high-dimensional statistics are rarely new; we are often attempting to solve traditional problems on new types of datasets. However, these datasets will often undermine a traditional technique (see the next section for further elaboration of this idea). Thus this thesis, as well as much of the existing literature, focuses on traditional topics such as supervised learning (using the variables to predict a continuous or categorical response), measuring the strength of variable relationships, looking at rankings of components and assessing the reliability of various procedures. These are all topics that have been examined in low-dimensional settings, but require reexamination for a new context.

A convention in statistics is to call the number of observations n, and the number of predictors p. Thus we may rephrase the previous examples; the first is called a "large n, large p" problem, while the latter is "small n, large p". The two examples sit at the opposite ends of the spectrum of high-dimensional datasets.

Of course, for serious study of high-dimensional problems, the types of issues found in real datasets must be incorporated into theoretical settings. One important consideration is that we must successfully capture the asymptotic relationship between n and p. Typically this is achieved by allowing p to grow with n. For ex-

ample many contexts in this thesis assume p is growing polynomially with n, so that $p = O(n^\alpha)$ for some $\alpha > 0$. Some authors have even studied the case where p grows exponentially in n (see for instance, Ng, 1998, Buhlmann, 2006, Fan and Lv, 2008). Some readers may find the idea of allowing p to grow with n somewhat bizarre; in a microarray experiment the number of genes does not grow as we analyse extra patients! However, the advantage of such an assumption is that it better captures the dynamics of our contemporary analyses. For example, if p grows quickly in n and we seek to measure some statistic for each variable (such as the mean) we would expect many of the variables to have close values and the correct ordering would be not determinable. Conversely in the case of fixed p we expect everything to be perfectly identifiable for large enough n. The first case is more akin to what we observe in reality, and so is more useful.

Finally for this section, we note that statistics is likely to see another rejuvenation in the field when so-called ultra-high dimensional datasets start to be collected and analysed. As with microarrays, technological advance in the biological sciences is causing another surge in the dimensionality of data able to be collected. In this case the culprit is high-throughput genome sequencing, which can detect the relative intensities of 500,000 different markers on the (human) genome. This means that particular genetic patterns can be analyses for relationships to disease and other conditions, with a hundred-fold increase in dimension compared to microarrays. Very little theory relevant to such datasets has permeated the statistical literature as yet, but rapid development seems likely.

1.2 Why is high-dimensional statistics special?

High-dimensional statistics is interesting not just because of the interesting applications, but because much of traditional statistical analysis must be rethought. We give a few examples of this below.

1.2.1 Significance and false positive rates. Suppose we have a high-dimensional dataset in which each observation belongs to one of two possible classes, and each of the p variables is continuous. A traditional t-test can be performed for each variable individually, assessing how significant the observed difference in mean is between the two classes. For a given significance level α, we expect that this proportion of variables, which have no actual relationship to the classes, will breach that level; that is, α of the redundant variables will appear as false positives. At a 5% level and $p = 10{,}000$, this would imply that up to 500 redundant variables will appear significant, in addition to any that are genuinely significant. Thus the possibility of many false positives hindering our ability to detect any true features is a major problem.

Literature exists on how to better choose the effective significance level in such

situations. One popular approach is to attempt to control the false discovery rate (the number of false positives divided by the number of rejections); see Benjamini and Hochberg (1995) and the review by Farcomeni (2008). Even though this gives a better feel for the level of error associated with a problem, the fundamental problem remains; the true effects have to appear very strongly to be clearly distinguished from the noise.

1.2.2 The problems of over-fitting. We begin this section with a small experiment. Take the leukemia microarray dataset, introduced in greater detail in Section 2.2.2. This dataset has $n = 72$, $p = 7{,}129$ and each observation belongs to one of two classes (types of acute Leukemia). Suppose we randomly choose 2/3 of the data for training, fitting a logistic regression (Hosmer and Lemeshow, 2000), using 20 randomly selected variables. We then test how well this model predicts for the original two thirds of the data, compared to the remaining one third. This experiment was performed 50 times and the results averaged. The models predicted correctly on the training data with 99% accuracy, while the accuracy on the testing third was under 70%, so not a long way from the 50% we would expect from a purely random model.

This illustrates a general principle, which is that when there are lots of variables, it is very easy to build a model that overfits the data. By this we mean that the apparent performance of the model on the training dataset is overly optimistic, and performance of the model on new, "unseen" data will be weak. This is a particular challenge in "small n, large p" situations, where it is very easy to construct a model that appears strong, but has disappointing future performance.

1.2.3 Computational complexity. Despite the inevitable truth that any comment on computational burdens will immediately be outdated, we make some remarks concerning practical limitations on high-dimensional analysis. The main observation is that if p is large, any method that takes $O(p^a)$ time for some $a > 1$ is likely to be infeasible. This limits certain particular types of approaches. For example, if we believed that some response depended on at most k of the variables, then the most natural (and optimal) way of finding these would be to test all subsets of k variables and choose the one that performs best, according to some measure. However, there are $\binom{p}{k} = O(p^k)$ such subsets, which grows rapidly in p and k. For instance, if $p = 10{,}000$ then there are nearly 50 million possibilities when k is only two. In these circumstances, finding ways to avoid such computation is necessary, such as the approach explored in Chapter 8.

1.3 The golden rules

Following on from the previous section, these are some guidelines that have characterised our research into the topic. They are a mix of common sense and a significant amount of experimentation.

No single approach will perform best in all high-dimensional situations

Just as the examples in Section 1.1 demonstrated a wide spectrum of types of problems, the best way to approach a problem, whether it is a prediction task or something else, will vary significantly. While this is obviously a boon for active researchers, who can continue to look for highly effective methods in various scenarios, it does mean that assignments of "best" or "worst" methods have diminished meanings in such settings.

As an example, Dettling (2004) compares the accuracy of seven classifiers on six microarray datasets. Four of the methods perform best on at least one of these, demonstrating that even when high-dimensional data is restricted to a single type, finding a best method is not possible.

The chances of detecting true effects accurately are often frighteningly small

This relates to the issue of false positives described above. Much of this thesis is concerned with assessing the accuracy of a ranking, where the variables of a dataset are ordered according to some definition of importance. Figure 1.1 gives an example of such a ranking in which 90% confidence intervals for the rank of each of the top 14 variables are included (calculated by means of the bootstrap), from a microarray dataset with $n = 30$ and $p = 6{,}319$. This particular example, and the corresponding approach, are described in detail in Chapter 4. For now, just observe the wide intervals for each variable, including those that are judged to be most important. Thus, if the experiment was performed a second time, the most important variable from the current ranking could reasonably be expected to rank anywhere in the top 200, meaning that it is very unlikely that it would be detected again as a particularly significant factor.

The assumption of sparsity is almost always unrealistic, but almost always useful

A sparse model is one that uses relatively few of the variables available. Penalisa-

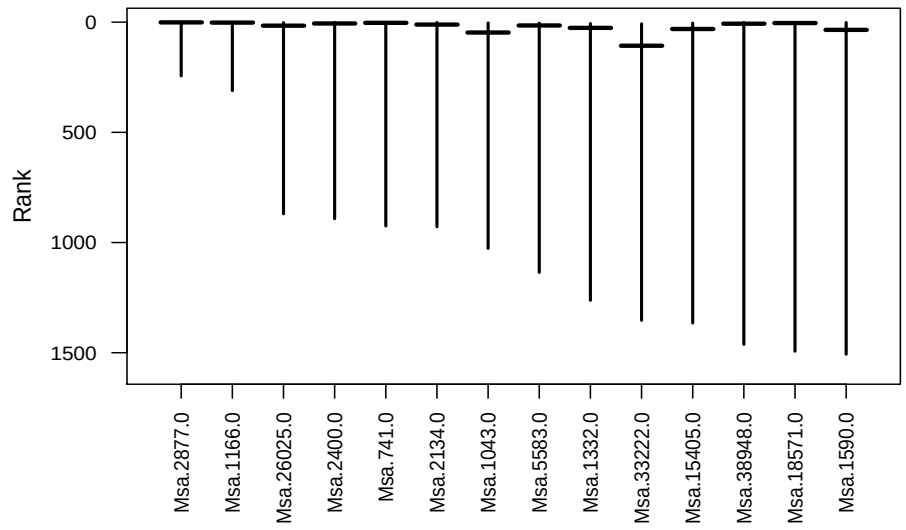


Figure 1.1: Important variables and ranking confidence intervals for the Ro131 example of Chapter 2.

tion methods, such as those discussed in the linear models section below, often give rise to sparse solutions and are increasingly popular today. The principle underlying the above rule is that even if the true situation is not sparse, there is little chance of correctly incorporating all these effects, and so a sparse model incorporating only the strongest variables will generally perform better. For instance, the University of Melbourne KDD Cup 2009 team built predictive models for the dataset that used only 200 variables, less than 2% of those available. Despite this sparsity, these models were powerful enough to win part of the competition (see Miller et al., 2009); thus the sparsity, rather than reducing the accuracy of a model, removed the noise associated with estimating the weak components.

As a second, somewhat more involved, example of this principle, consider the density plots in the left panel of Figure 1.2. This shows the distribution of scaled Mann-Whitney test scores for each of 7,129 genes in the leukemia microarray dataset, which is described in Section 1.2.2. The plot is taken from Section 6.4, which contains a more detailed description of the methodology. The dotted line represents the density of test scores assuming there was no relationship between the expression levels and the two categories in the response. The large departure of the actual density from this suggests that a good proportion of the genes, perhaps at least 30%, have some connection with the response. Suppose we wanted to build a predictive random forest model (Breiman, 2001a) using the top d ranked genes based on this Mann-Whitney test statistic. Based on the above comments, one may think that a good model may need at least $d = 2{,}000$. However the results in the right panel of Figure 1.2 (where we repeatedly broke the data into two third/one third train/test splits, selected genes,

built models using the training data, and measured performance on the test set), shows that a model size of 300, or an order of magnitude less, is preferred. Thus, erring on the side of sparsity can often improve predictive accuracy.

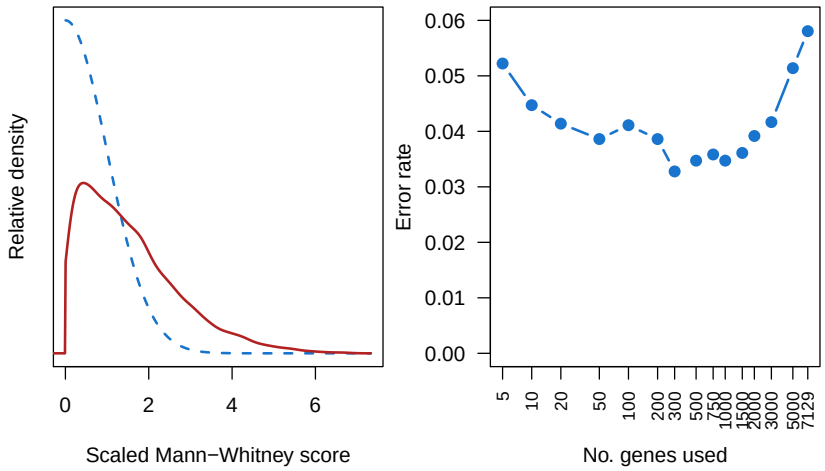


Figure 1.2: Model selection for the Leukemia dataset. The left panel shows the distribution of Mann-Whitney test statistics for the 7,129 genes in the dataset, compared to the null (dotted). The right panel shows the misclassification rates of random forest models where we use different numbers of the top-ranked genes, according to this test. About 300 genes appears optimal

As a slight caveat to the above argument, it is thought that many genomic problems may in fact be less sparse than originally thought (see for instance, Goldstein, 2009, Hirschhorn, 2009, and Kraft and Hunter, 2009). See Hall et al. (2010) for recent work attempting to allow for construction of effective models with lower degrees of sparsity.

Validate, and validate properly

Section 1.2.2 has already explored how easy it is to produce an overfitted model. In the example given, the models were clearly seen to be poor because they were validated. The most common way to do this is by train-test validation or cross-validation (see Hastie et al., 2001, Chapter 7), where part of the data is set aside to assess a model fit on the remainder. While simple, this remains one of the few effective ways to correctly ascertain how well a technique is performing.

It is also particularly important that validation be performed properly. An example illustrating this is where an initial variable selection step takes place, followed by a model selection step. If the variable selection is done using the whole dataset, then even if the model selection step is validated the final model will be overly optimistic. For this reason the idea of using two layers of cross-validation, introduced in Stone

(1974), is commonly employed in many scenarios (and has consequently been incorporated into the biostatistics R package Rmagpie[2] for the analysis of microarray data). The work in Chapter 8 provides another explicit scenario where this methodology is appropriate.

1.4 The role of numerical work

All subsequent chapters devote some attention to how methods perform on actual datasets, whether they are simulated or real. This is a useful safeguard; since applied statistics is generally motivated by problems associated with (real) datasets, any proposed methods should be validated by the same. However, we qualify the above statements with two comments. Firstly, measuring performance on datasets rarely gives deep insight to the performance of an approach, instead showing how it may compare to a competing method. Therefore strong numerical results are in no way a substitute for good theoretical results, and the information that these give. Secondly, because there are certain types of datasets that appear most commonly in the literature (for example, microarray data examples permeate much of the high-dimensional analysis literature), approaches can be biased towards solving those particular problems. While finding the best way to attack a particular problem is useful, there can be a risk of discarding a worthwhile approach due to its poor performance in a specific context.

The attempt in this work is to address a problem through a complementary mix of both theory and example, with the hope that results are both insightful and applicable.

1.5 A framework, and the structure of the remaining chapters

The schematic in Figure 1.3 represents an overarching framework for approaching a high-dimensional problem. Starting with the data, an analysis will often begin with feature selection, where dimension is drastically reduced, keeping only the most important variables. Once the feature selection is complete, this variable set may be used to create a final predictive model. Alternatively, some approaches fit a model directly, without an initial feature selection. Feature selection gives the analyst an indication of which variables are most important, and similar information may also come from the model. Once these are detected, some time may be spent investigating the variables and how they relate to each other. This may in turn feed back into a final model.

[2] http://bioconductor.org/packages/2.5/bioc/html/Rmagpie.html

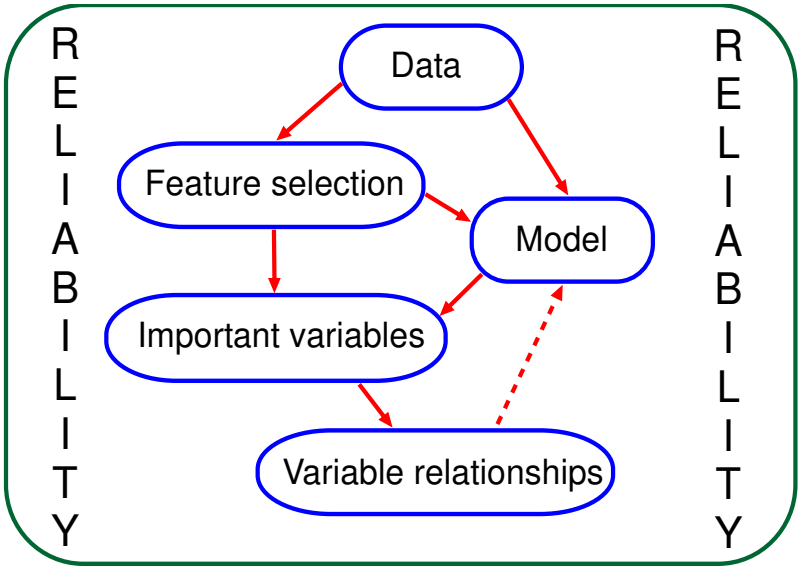


Figure 1.3: A possible framework for high-dimensional statistics

Surrounding all these different stages of analysis is the idea of reliability. The preceding sections should have impressed the importance of ensuring any results are properly validated and that the limitations are clearly understood. Thus quantifying the effectiveness of a model, feature selection procedure, or ordering will be an important issue addressed in the thesis. We shall see that the bootstrap is an important tool for exploring this.

The remaining sections in this chapter introduce some topics relevant to multiple chapters, so are included here for general reference. The material included is not new, but is instead intended to make the overall exposition clearer. The rest of the thesis then addresses various topics from the above framework. Chapter 2 looks at the feature selection problem in a particular context, when the response is continuous and there is a belief that not all effects are qualitatively linear. It also begins to address the question of how to assess the reliability of a ranking of variables. Chapter 3 focuses on how to analyse a relatively small set of variables for interesting relationships, suitable for after the important variables have been detected. Chapter 4 investigates one method for constructing a nonparametric model from a relatively small set of variables, such as those chosen through feature selection. Chapters 5, 6 and 7 address the reliability question relating to a ranking (such as a feature selection) in detail, each from a different perspective; Chapter 5 looks at correctly diagnosing the uncertainty in terms of the distribution of the ranks; Chapter 6 investigates contexts in which the top few variables may be detected correctly, even when correct ranking is not possible for the bulk of variables; and Chapter 7 explores conservative tests which can be used to assess interesting hypotheses regarding a ranking. Finally, Chapter 8 looks at a robust means of moving directly from the dataset to a model. The relationship

between each topic and the overarching framework should be clear.

The reader may have noticed that some of these Chapters address problems that are not strictly high-dimensional. In particular, Chapters 3 and 4 actually focus on situations where the dimensionality is moderate, and some of the work on rankings does not explicitly mention nor necessitate a high-dimensional context. However, we point out that this work still forms part of the overall picture described above, since working with a reduced number of variables is relatively common.

1.6 Moderate deviation properties

Suppose $U_1, \ldots, U_n$ are independent and identically distributed random variables with zero mean and variance σ^2. A moderate deviation for the mean of these random variables refers to a deviation of order $(n^{-1} \log n)^{1/2}$. Thus we are interested in quantities such as

$$P\left\{\left|\frac{U_1 + \ldots + U_n}{n}\right| > c\sigma\sqrt{\frac{\log n}{n}}\right\}.$$

This may be compared to an "ordinary" deviation, which is of order $n^{-1/2}$ and leads to results such as the the central limit theorem, in the case of the mean. Results for probabilities of moderate deviations, such as those found in Rubin and Sethuraman (1965) and Amosova (1972), are useful for ensuring some uniform convergence results in our work. Of particular note is Theorem 4 from Rubin and Sethuraman (1965). If $E(|U_i|^q) < \infty$ for some $q > d + 2$ with $d > 0$, then

$$P\left\{\left|\frac{U_1 + \ldots + U_n}{n}\right| > c\sigma\sqrt{\frac{\log n}{n}}\right\} \sim \frac{2}{\sqrt{2\pi d \log n}} n^{-d/2}. \tag{1.1}$$

Similar results generally hold for other asymptotically normal statistics besides the mean. See for instance Inglot et al. (1992).

1.7 The linear model

Suppose that for observations $i = 1, \ldots, n$ we have a continuous response Y_i, as well as p-dimensional predictors $X_i = (X_{i1}, \ldots, X_{ip})$. The typical regression problem is to find function f on p-dimensional space such that

$$Y_i = f(X_i) + \text{error}. \tag{1.2}$$

One of the most enduring parametric forms for f is the linear model,

$$f(X_i) = \beta_0 + \beta_i X_{i1} + \ldots + \beta_p X_{ip}. \tag{1.3}$$

The key reason for its popularity over time is that in many real problems the true function f is well approximated by the linear model, and it also serves as the basis for many nonlinear extensions, such as spline models (see Wahba, 1990 and De Boor, 2001) and local polynomial regression (see Simonoff, 1996 and Loader, 1999). There are many texts on linear models, including that by Mardia et al. (1979). Estimates for β_j are made by minimising a loss function, typically squared error, $\sum_i \{Y_i - f(X_i)\}^2$.

Linear models have played a particularly important role in high-dimensional statistics. Being among the simplest of models, they are suited to situations where describing complex nonlinear behaviours is not possible, and when most effects are qualitatively linear, or at least monotonic. This is particularly the case for typical "small n, large p" datasets.

Readers familiar with linear models will recognise a major identifiability problem with the linear model when $p \gg n$. There are infinitely many choices of β_j which perfectly fit the data, almost all of which will represent gross overfitting of the data, as discussed in Section 1.2.2. The most common way to overcome this problem is through some form of penalisation, which biases the model away from the overfit solution. The lasso (Tibshirani, 1996) is a good example of this, where instead of simply minimising the sum of squares we choose β_j to minimise

$$\sum_{i=1}^{n} \left\{ Y_i - \left(\beta_0 + \sum_{j=1}^{p} \beta_j X_{ij} \right) \right\}^2 + \lambda \sum_{j=1}^{n} |\beta_j| \,.$$

This is often referred to as an L_1 penalty, since the extra penalty term is the L_1 norm of the p-dimensional coefficient vector β. One feature of the lasso is that it automatically produces sparse models; if λ is sufficiently large, many of the estimates for β_j will be zero and the corresponding variables are effectively dropped from the model. This contrasts with ridge regression (Hoerl and Kennard, 1970) which uses an L_2 penalty, where all variables remain in the model while the coefficients are shrunk towards zero.

The linear model extends into situations where the response Y_i is not continuous. In particular if Y_i is a categorical binary variable, taking values 0 or 1 only, then logistic regression fits a linear model to the log-odds ratio. This is part of the generalised linear model framework, which has been covered in book form by McCullagh and Nelder (1989) and Dobson (2001).

There is a large literature on variable-selection methods relating to the linear model. Many of these contributions relate to exploring the effects of different penalties. See Chen et al. (1998), Zou (2006), Candes and Tao (2007) and Bickel et al. (2009), among many other efforts. Other work on the linear model includes, but is by no means restricted to, work on the nonnegative garotte (e.g. Breiman, 1995; Gao,

1998), on soft thresholding (e.g. Donoho et al., 1995), and related work (e.g. Donoho and Huo, 2001; Fan and Li, 2001; Donoho and Elad, 2003; Tropp, 2005; Donoho, 2006a; Donoho, 2006b).

1.8 The bootstrap

With the continuing growth of ever-cheaper computing, the bootstrap has increased in popularity, due to its ability to give significant insight into the sampling properties of a dataset. In following chapters we use the bootstrap extensively, to create confidence intervals, estimate p-values, diagnose variable selection uncertainty and to correctly estimate distributions of empirical rankings.

Suppose we are interested in a statistic θ with estimate $\hat{\theta}$ made from independent and identically distributed observations $X_1, \ldots, X_n$. The standard (nonparametric) bootstrap samples the observations with replacement, creating a pseudo-dataset $X_1^*, \ldots, X_n^*$. From this a bootstrapped version of the statistic, $\hat{\theta}^*$, can be calculated. Since the empirical cumulative density function of the observations may be viewed as an approximation of the true density for X_i, the relationship between $\hat{\theta}^*$ and $\hat{\theta}$ will resemble in many ways the relationship between $\hat{\theta}$ and θ. For example, if $\hat{\theta} - \theta$ is asymptotically normal, such as the mean of the observations, then the bootstrap distribution of $\hat{\theta}^* - \hat{\theta}$, conditional on the data, will also be asymptotically normal. Also, repeated bootstrap simulation of $\hat{\theta}^*$ allows calculation of the corresponding distribution function, which can be used to give a nominal $1 - \alpha$ confidence interval for $\hat{\theta}$.

A bootstrap metatheorem argues that, in a range of settings, bootstrap methods give consistent results for estimating distributions of parameter estimators "if and only if" the limiting distribution is normal (see e.g. Mammen, 1992). Some situations explored in this present work, notably rankings, are highly non-normal, and so methods of overcoming the inconsistency of the bootstrap are investigated.

We cannot hope to give proper coverage to the bootstrap in this introduction. Interested readers are referred to the texts by Hall (1992), Davison and Hinkley (1997) and Efron and Tibshirani (1997).

Chapter 2

Generalised correlation for feature selection

2.1 Background

A variety of linear model-based methods have been proposed for variable selection, as introduced in Section 1.7. In this approach it is argued that a response variable, Y_i, might be expressible as a linear form in a long p-vector, X_i, of explanatory variables, plus error, as in (1.2) and (1.3). Many, indeed the majority, of applications of this linear model (1.3) represent cases where the response is unlikely to be an actual linear function of X_i, for example where Y_i is a zero-one variable but the fitted response takes values that often lie outside the unit interval. However, inconsistency of prediction does not necessarily detract from the usefulness of such methods as devices for determining the components X_{ij} that most influence the value of Y_i. For example, inconsistency is often not a significant problem if the response of Y_i to an influential component X_{ij} is qualitatively linear, in particular if it is monotone and the gradient does not change rapidly.

In other settings, however, there is a risk that fitting an incorrect linear model will cause us to overlook some important components altogether. Theoretical examples of this type are identical to those used to show, by counter example, that the existence of conventional correlation does not equate to absence of a relationship. In Section 2.2, Example 2.2.1 will discuss a practical instance of this difficulty, and Example 2.2.2 there will treat another real dataset where challenges of a different nature arise. More generally, using an ill-fitting model to solve a variable-selection problem can result in reduced performance.

A little more subtly, even if the linear model is perfectly correct, fitting it can

conceal components that potentially influence linearly the value of Y_i. For instance, genes whose expression levels are strongly linearly associated with Y_i, and so would be of biological interest, can be confounded or not uniquely represented. In particular, if $X_{i1} = X_{i3} + X_{i4}$ and $X_{i2} = X_{i3} + X_{i5}$ then the linear models $Y_i = X_{i1} - X_{i2} + \text{error}$ and $Y_i = X_{i4} - X_{i5} + \text{error}$, and of course infinitely many others, are equally valid. This non-identifiability issue arises because the variable-selection problem is posed as one of model fitting, or prediction, which in our view is not necessarily a good idea. Thus, even nonlinear extensions to variable selection methods that focus on prediction, such as the group lasso or group LARS (Yuan and Lin, 2006), may still be inadequate in detecting all influential variables. Example 2.4.3 will explore this type of behaviour in greater detail. Also, Example 2.2.3 explores a real dataset where this masking interferes with variable selection.

These examples, and others that we shall give, argue in favour of methods for variable selection that focus specifically on that problem, without requiring a restrictive model such as that at (1.3). In this chapter we suggest techniques based on ranking generalised empirical correlations between components of X and the response Y. Section 2.2 discusses real-data examples which motivate our approach, Section 2.3 introduces our methodology, and Section 2.4 extends the discussion in Section 2.2 and also presents simulation studies which explore properties of the methodology. Section 2.5 provides theory that demonstrates the methodology's general properties.

2.2 Motivating examples

Here we discuss three real datasets which motivate the methodology we shall introduce in Section 2.3.

2.2.1 Example: Cardiomyopathy microarray data. This dataset was used by Segal et al. (2003) to evaluate regression-based approaches to microarray analysis. The aim was to determine which genes were influential for overexpression of a G protein-coupled receptor, designated Ro1, in mice. The research related to understanding types of human heart disease. The Ro1 expression level, Y_i, was measured for $n = 30$ specimens, and genetic expression levels, X_i, were obtained for $p = 6{,}319$ genes.

Our analysis will be based on ranking, over j, the maximum over h of the correlation between $h(X_{ij})$ and Y_i, where the correlation is computed from all data pairs (X_i, Y_i) for $i = 1, \ldots, n$. Here h is confined to a class $\mathcal{H}$ of functions. Taking $\mathcal{H}$ to consist entirely of linear functions gives the (absolute value of the) conventional correlation coefficient, but using a larger class enables us to explore nonlinear relationships. We shall take $\mathcal{H}$ to be a set of cubic splines. See Example 2.4.1 in Section 2.4 for further technical detail.

This approach leads us to rank two genes, Msa.2877.0 and Msa.1166.0, first and second, respectively. The first of these genes was identified by the linear-regression approach adopted by Segal et al. (2003), but the second was not. Figure 2.1 indicates why this is the case, by showing the scatterplots and corresponding cubic-spline fits. While Msa.2877.0 shows an essentially linear relationship, which is identified by many existing techniques, Msa.1166.0 exhibits clear nonlinear behaviour, where the response "flatlines" once the expression reaches a certain threshold. Another factor is the strong correlation of -0.75 between the two variables. This "masking effect" confounds standard linear modeling approaches to variable selection, and was discussed in Section 2.1. See also Examples 4, 5 and 6 in Section 2.4.

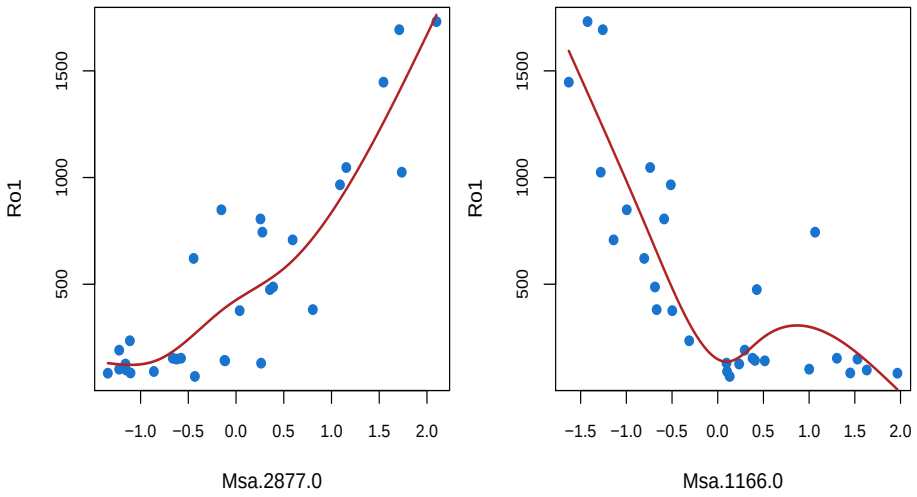


Figure 2.1: Top two variables with cubic spline fits for Example 2.2.1

2.2.2 Example: Acute Leukemia microarray data. This dataset comes from a study by Golub et al. (1999), where the aim was to use microarray evidence to distinguish between two types of acute leukemia (ALL/AML). There were $p = 7{,}129$ genes and $n = 38$ observations in the training data (27 ALL and 11 AML). There were also 34 observations in a separate test dataset with 20 ALL and 14 AML.

Methods based on linear correlation, of which those proposed in this chapter are a generalisation, are analogous to minimising the deviance of a normal model with identity link under the generalised linear model framework (McCullagh and Nelder, 1989). This suggests that binary data could be treated by minimising the deviance formula for Bernoulli data with a logistic link for each X_i, and using this to rank the values of

$$\inf_{h \in \mathcal{H}} \sum_{i=1}^{n} \left\{ -Y_i \log \left(e^{h(X_{ij})} \right) + \log \left(1 + e^{h(X_{ij})} \right) \right\} , \tag{2.1}$$

where each Y_i equals zero or one and $\mathcal{H}$ is a class of functions, for example the class

of polynomials of a given degree. In the analysis reported below we took $\mathcal{H}$ to be the set of all linear functions. For further detail see Example 2.4.2.

There is considerable overlap between the genes we found using this approach, and those discovered in other studies (Golub et al., 1999; Tibshirani et al., 2002; Fan and Fan, 2008; Hall et al., 2009; Fan and Lv, 2008). However, we argue that the set found in the present analysis represents an improvement over choices made by alternative methods. To address this point, a simple classifier was constructed. For the genes giving the five largest values of the quantity at (2.1), a classifier was chosen that minimised the misclassification rate on the training data, weighted so that the two classes had equal authority. These classifiers all had one decision value, above which the classification would be one class and below which it would be the other. Whichever class had the most "votes" out of the five would then be the overall predicted class. Although this was a very simple classifier it performed perfectly on the training data and had only one misclassification on the test set. This means the classifier performed at least as well as other approaches in the literature and, in most cases, used considerably fewer genes. We again stress that our purpose was not to build a predictive model, but to identify influential variables. If the latter problem, rather than prediction, is the ultimate aim, and it generally is, then it can be advantageous to focus on it from the start.

2.2.3 Example: Breast tumor X-ray data. This dataset was used as the training dataset in the 2008 KDD Cup data mining competition[1]. It consists of 102,294 observations, each corresponding to a potential malignant tumor spot on an X-ray. Each observation has 117 continuous variables identifying different attributes of the spot and a binary response identifying whether the spot is malignant, which is the case for 623 observations. For convenience here we disregard dependencies caused by spots resulting from the same patient.

This dataset actually forms a "large n, large p" problem and it is possible to build a fairly accurate classification model using the entire dataset. Suppose, however, that we had access to only 5% of the data. Then the roughly 30 positive responses would be insufficient to build a reasonable model, so detecting which variables are most important might be a more appropriate goal. With this in mind, consider the simulation experiment where we examine how effectively generalised correlation detects variables compared to a predictive method. The top twelve variables in the entire dataset were determined using a weighted random forest model (Breiman, 2001a). The random forest was chosen to be a reasonably "model-neutral" method for determining variable importance. We sampled 5% of the data and attempted to determine the 12 most influential variables using a given approach. Then we compared the results to the top 12 variables derived from the entire dataset and

[1] www.kddcup2008.com

calculated the number in common. Table 2.1 shows the results of 100 simulations for a generalised correlation approach using (2.1) and the logistic group lasso (Meier et al., 2008), each based on cubic splines with knots at the quartiles. The group lasso is a penalised regression method that allows for groups of variables, such as a collection of splines, and so is an appropriate candidate for comparison in the simulation study. The results for the top variables from a random forest applied to the sample are also included.

	No. of top 12 effects detected	
group lasso	5.86	(0.08)
generalised corr.	10.02	(0.08)
random forest	9.44	(0.10)

Table 2.1: Average number of variables detected under 5% sampling in for Example 2.2.3.

Generalised correlation performed better than both the random forest and group lasso models, the latter picking up less than half the variables on average. These results show that predictive methods are not necessarily the optimal way to approach the variable selection problem, if variable selection is the ultimate aim. In this particular case it is possible to show that correlations among variables are hindering variable selection for the group lasso and random forest procedures.

2.3 Methodology

2.3.1 Generalised correlation. Let $\mathcal{H}$ denote a vector space of functions, which for simplicity we take to include all linear functions. By restricting $\mathcal{H}$ to just its linear elements we obtain, in (2.2) below, the absolute values of conventional correlation coefficients, but more generally we could take $\mathcal{H}$ to be the vector space generated by any given set of functions h.

Assume that we observe independent and identically distributed pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ of p-vectors X_i and scalars Y_i. A generalised measure of correlation between Y_i and the jth component X_{ij} of X_i is given and estimated by

$$\sup_{h \in \mathcal{H}} \frac{\mathrm{cov}\{h(X_{1j}), Y_1\}}{\sqrt{\mathrm{var}\{h(X_{1j})\}\,\mathrm{var}(Y_1)}}\,, \quad \sup_{h \in \mathcal{H}} \frac{\sum_i \{h(X_{ij}) - \bar{h}_j\}\,(Y_i - \bar{Y})}{\sqrt{\sum_i \{h(X_{ij}) - \bar{h}_j\}^2 \cdot \sum_i (Y_i - \bar{Y})^2}}\,, \tag{2.2}$$

respectively, where $\bar{h}_j = n^{-1} \sum_i h(X_{ij})$. Since each of the factors $\mathrm{var}(Y_1)$ and $\sum_i (Y_i - \bar{Y})^2$, in the denominators at (2.2), do not depend on j, each may be replaced by any constant without affecting our ranking-based methodology. Therefore

we shall work instead with

$$\psi_j = \sup_{h\in\mathcal{H}} \frac{\mathrm{cov}\{h(X_{1j}), Y_1\}}{\sqrt{\mathrm{var}\{h(X_{1j})\}}}\,, \quad \hat{\psi}_j = \sup_{h\in\mathcal{H}} \frac{\sum_i \{h(X_{ij}) - \bar{h}_j\}\,(Y_i - \bar{Y})}{\sqrt{n\,\sum_i \{h(X_{ij}) - \bar{h}_j\}^2}}\,. \tag{2.3}$$

These measures of association reflect the approach suggested by Grindea and Postelnicu (1977). However, a variety of alternative measures could be used. See, for example, Griffiths (1972), Csörgő and Hall (1982) and Schechtman and Yitzhaki (1987). At first it might appear that the challenge of computing $\hat{\psi}_j$ in (2.3), for large p, might be onerous, even by modern computing standards. However the following theorem, with an simple proof, simplifies the problem in a wide range of cases.

Theorem 2.1. *Assume $\mathcal{H}$ is a finite-dimensional function space including the constant function, and that there exists $h \in \mathcal{H}$ that achieves $\hat{\psi}_j$ in the definition at* (2.3)*. Then*

$$\operatorname*{argmin}_{h\in\mathcal{H}} \sum_{i=1}^{n} \{Y_i - h(X_{ij})\}^2 \subseteq \operatorname*{argmax}_{h\in\mathcal{H}} \hat{\psi}_j\,.$$

That is, the maximiser of $\hat{\psi}_j$ is the solution to the least-squares problem in $\mathcal{H}$.

Proof: Without loss of generality let $\bar{Y} = 0$, and define $S_h = \sum_i \{h(X_{ij}) - \bar{h}_j\}^2$ to be the sum of squares associated with a choice of h. Let $b(X_{ij})$ be a (finite) basis expansion of X_{ij} in $\mathcal{H}$. The the least-squares problem may be expressed as choosing β to minimise $\sum_i \{Y_i - b(X_{ij})^T\beta\}^2$. Provided the matrix $b(X_j)$, with rows $b(X_{ij})$, has full rank (the basis may be constrained to satisfy this), there is the usual least-squares solution. Identifying $h(X_{ij})$ with $b(X_{ij})^T\beta$ it is not hard to show that $\bar{h}_j = \bar{Y} = 0$. Let K be the value of S_h where h corresponds to the least squares solution. Then the least squares problem can then be expressed as

$$\operatorname*{argmin}_{h\in\mathcal{H}} \sum_i \{Y_i - h(X_{ij})\}^2 \;=\; \operatorname*{argmin}_{h\in\mathcal{H}\,|\,\bar{h}_j=0, S_h=K} \sum_i -h(X_{ij})Y_i\,.$$

Since correlations are invariant under constant shifts and scalar multiplication we have

$$\begin{aligned} \operatorname*{argmax}_{h\in\mathcal{H}} \hat{\psi}_j \;&\supseteq\; \operatorname*{argmax}_{h\in\mathcal{H}\,|\,S_h=K, \bar{h}_j=0} \frac{\sum_i \{h(X_{ij})Y_i - \bar{h}_j Y_i\}}{\sqrt{S_h}} \\ &=\; \operatorname*{argmin}_{h\in\mathcal{H}\,|\,\bar{h}_j=0, S_h=K} \sum_i -h(X_{ij})Y_i\,, \end{aligned}$$

which completes the proof.

Thus for the case described in Theorem 2.1 (for example, polynomials up to some degree d), the least-squares problem has an explicit analytic solution. This avoids a potentially cumbersome optimisation problem and allows "basis expansions" of X_{ij}.

Global modeling techniques generally preclude basis expansions on the grounds that they create an even larger dimensionality problem and make it difficult to assess the influence of the underlying variables.

One implication of Theorem 2.1 is that the ranks of the $\hat{\psi}_j$'s are the same whether we consider $\hat{\psi}_j$ itself or the reduction in the size of squared error,

$$\hat{\varphi}_j = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 - \inf_{h \in \mathcal{H}} \sum_{i=1}^{n} \{Y_i - h(X_{ij})\}^2 .$$

This is particularly useful when some of the components of X_i are categorical. In such a case the correlation (simple or generalised) cannot be easily defined, but $\hat{\varphi}_j$ can be measured by taking h to be the mean response of each category. Restricting $\mathcal{H}$ to a space of constant and linear functions recovers the ranking based on conventional correlations.

2.3.2 Correlation ranking. We order the estimators $\hat{\psi}_j$ at (2.3) as $\hat{\psi}_{\hat{j}_1} \geq \ldots \geq \hat{\psi}_{\hat{j}_p}$, say, and take

$$\hat{j}_1 \succeq \ldots \succeq \hat{j}_p \tag{2.4}$$

to represent an empirical ranking of the component indices of X in order of their impact, expressed through a generalised coefficient of correlation. In (2.4), the notation $j \succeq j'$ means formally that $\hat{\psi}_j \geq \hat{\psi}_{j'}$, and informally that "our empirical assessment, based on correlation, suggests that the jth coefficient of X has at least as much influence on the value of Y as does the j' th coefficient." Using this criterion, the ranking $r = \hat{r}(j)$ of the jth component is defined to be the value of r for which $\hat{j}_r = j$.

The authority of the ranking at (2.4) can be assessed using bootstrap methods, as follows. For each j in the range $1 \leq j \leq p$, compute $\hat{\psi}_j^*$, being the bootstrap version of $\hat{\psi}_j$ and calculated from a resample $(X_1^*, Y_1^*), \ldots, (X_n^*, Y_n^*)$, drawn by sampling randomly, with replacement, from the original dataset $\mathcal{D} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$. Compute the corresponding version of the ranking at (2.4), denoted by $\hat{j}_1^* \succeq \ldots \succeq \hat{j}_p^*$, and calculate too the corresponding bootstrap version, $r^*(j)$ say, of $r(j)$. Given a value α, such as 0.05, 0.10 or 0.20, compute a nominal $(1 - \alpha)$-level, two-sided, equal-tailed, percentile-method prediction interval for the ranking, i.e. an interval $[\hat{r}_-(j), \hat{r}_+(j)]$ where

$$P\big\{r^*(j) \leq \hat{r}_-(j) \,\big|\, \mathcal{D}\big\} \approx P\big\{r^*(j) \geq \hat{r}_+(j) \,\big|\, \mathcal{D}\big\} \approx \tfrac{1}{2}\,\alpha\,.$$

We indicate approximations in these formulae since the discreteness of ranks restricts the smoothness of the bootstrap distribution.

Display these intervals as lines stacked one beside the other on the same figure, each plotted on the same scale and bearing a mark showing the respective value

of $\hat{r}(j)$. Convenient orderings for the lines include the one indicated in (2.4), or the ordering in terms of increasing $\hat{r}_+(j)$. The second choice generally provides greater insight since it emphasises variables that consistently rank strongly in the bootstrap simulations. Only lines for relatively low values of $\hat{r}$ or $\hat{r}_+$ would be depicted; see the next section for examples. If two prediction intervals (represented by the lines) failed to overlap, this would provide empirical evidence that the more highly ranked component did indeed enjoy greater impact on Y than its competitor, at least in terms of the way we have measured impact. Thus bootstrap methods allow us to generate confidence intervals. This standard use of the bootstrap is often fine, although in certain situations, most notably when the correlation scores for different components are (nearly) tied, the standard bootstrap can fail. This behaviour, and means of overcoming it, is examined in detail in Chapter 5, but for now we assume these confidence intervals are reasonable.

An important consideration of the approach presented is determining at what level significance is drawn; that is, how to decide which variables are influential and which are not. One proposed criterion is to regard a variable as influential if $\hat{r}_+(j) < \frac{1}{2}\,p$. This rule assumes that the number of influential variables is considerably less than the total number p; if all components were genuinely related to Y then the rule would reject at least half of them. There are many circumstances, such as genetic microarray data, where this assumption is reasonable. The rationale is if all the variables were independent of Y, then the rank of each would randomly fluctuate across 1 through p, with an average rank of $\frac{1}{2}\,p$. If the prediction interval of a variable's rank does not breach $\frac{1}{2}\,p$ for a given significance level, then it is unlikely to be independent of Y. However, there may be an undesirably high rate of false positives under this criterion, particularly for small n. A natural way to tune the rule to alleviate this problem is to replace $\frac{1}{2}\,p$ by some smaller fraction of p. Rather than trying to predict a suitable level, it is generally easier to plot the results and allow the data to suggest a suitable level. This principle is further explored in Example 2.4.3 of the numerical work.

2.3.3 Ranking conventional correlations. Since conventional correlation measures the strength of a linear relationship then, in many cases, component ranking in terms of conventional correlation gives results not unlike those obtained by linear model fitting, for example using the lasso. In particular, if the linear model at (1.2) and (1.3) holds in a form where only a fixed number, q say, of the coefficients β_j are nonzero, and if the coefficients of correlation of Y_i with all other components of X_i are bounded away from ± 1, then under moment conditions on the components (see Section 2.5), the probability that the q special components appear first in a ranking of the absolute values of conventional correlation coefficients converges to 1 as n and p diverge.

However, linear methods such as the lasso can be challenged when it comes to

identifying, purely empirically, the q special components. The conventional lasso can fail to correctly choose the components, even if p is kept fixed as n diverges. Component ranking, based on the absolute values of conventional correlation coefficients, can be used for an initial "massive dimension reduction" step, reducing dimension in one hit from p to a relatively low value, larger than q, from which dimension can be further reduced to q by implementation of an existing adaptively penalised form of the lasso.

Another potential advantage of ranking methods based on conventional correlation coefficients is that they overcome problems with errors in variables. For example, suppose that, in a generalisation of (1.2) and (1.3),

$$Y_i = g(W_i) + \text{error}\,, \tag{2.5}$$

where g is a potentially nonlinear function, W_i denotes the p-vector of actual (but hidden) explanatory variables, and the error is independent of W_i. In errors-in-variables problems we observe only Y_i and $X_i = W_i + \delta_i$, where the p-vector δ_i is a second source of error with zero mean, independent of W_i and of the error in (2.5). There is a large literature on problems framed in this way, in cases where p is substantially smaller than n; usually, $p = 1$. This work can be accessed through the monograph by Carroll et al. (2006). The effect of the errors δ_i vanishes entirely from the correlation between X_{ij} and Y_i:

$$\text{cov}(X_{ij}, Y_i) = \text{cov}\{W_{ij} + \delta_{ij}, g(W_i)\} = \text{cov}\{W_{ij}, g(W_i)\} = \text{cov}(W_{ij}, Y_i)\,.$$

In particular, the conventional correlation between X_{ij} and Y_i is exactly equal to the conventional correlation between W_{ij} and Y_i. Generalised correlations will not in general retain this property; if the distribution of δ_i were known then the effect of the error could be at least partially reduced by "deconvolution", but this approach is not attractive when $p \gg n$.

Therefore, component ranking in terms of the absolute values of conventional correlations is an effective way of removing the effects of errors in variables, even if, as in (2.5), the response is a nonlinear function of the hidden explanatory variable. Example 2.4.4 will address problems of this type.

2.4 Numerical properties

2.4.1 Example: Continuation of Example 2.2.1. We used natural cubic splines, with three interior knots on the quartiles of the variable's observed values, because, unlike quadratic splines, such functions model both nonlinear monotone

functions and multimodal functions. This gives them significant flexibility. To implement the bootstrap method described in Section 2.3.2 we used 400 resamples, $\alpha = 0.02$ and a $\frac{1}{4}\,p$ cutoff for $\hat{r}_+$. This resulted in the selection of 14 genes, of which two, the genes Msa.2877.0 and Msa.1166.0 discussed in Section 2.2, were particularly influential. This can be deduced from the marked jump in the length of the prediction intervals, represented by vertical lines in Figure 2.2, between the second and third most highly ranked genes. Examples 2.4.4 and 2.4.5, below, will summarise the results of simulation studies motivated by the findings above.

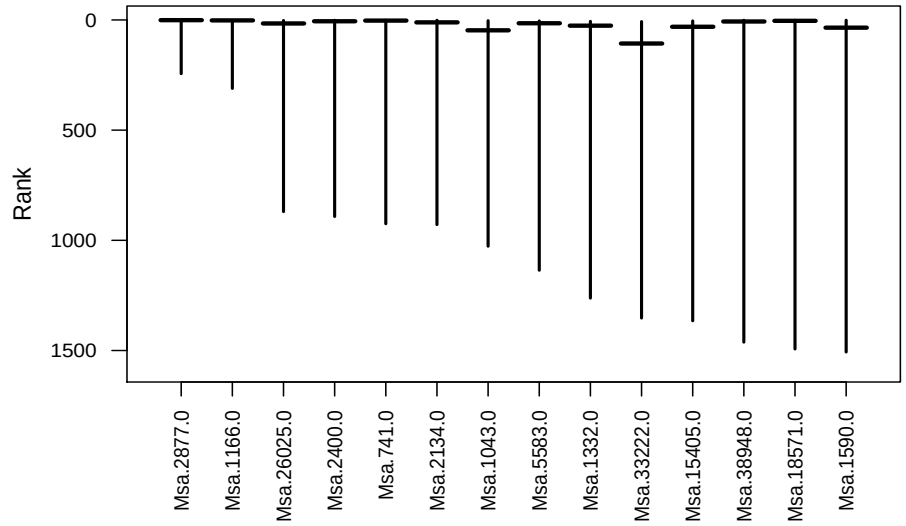


Figure 2.2: Variables ordered by $\hat{r}_+$ for Example 2.4.1

2.4.2 Example: Continuation of Example 2.2.2. When $\mathcal{H}$ is constrained to include linear functions of X_{ij}, as was the case in our treatment of this example in Section 2.2, the approach is analogous to ranking the absolute values of conventional correlation coefficients. Our bootstrap implementation used 200 resamples and $\alpha = 0.05$. All variables were standardised to have sample mean zero and sample variance one. Figure 2.3 shows the influential genes using a $\frac{1}{8}\,p$ cutoff for $\hat{r}_+$. The first two or three genes are seen to stand out, in terms of influence, and then influence remains approximately constant until genes 9 or 10. From that point there is another noticeable drop in influence, to a point from which it tails off fairly steadily.

2.4.3 Example: Variable masking. Motivated by an example discussed in the Introduction we look at a linear model where variables are highly correlated, and we compare the variable selection performance of our method and the lasso (Tibshirani, 1996).

First we describe the model generating the data. For $1 \le j \le 5$ let $\{X_{ij}, X_{i,j+5}\}$ be independent pairs of normal random variables with zero means, unit variances

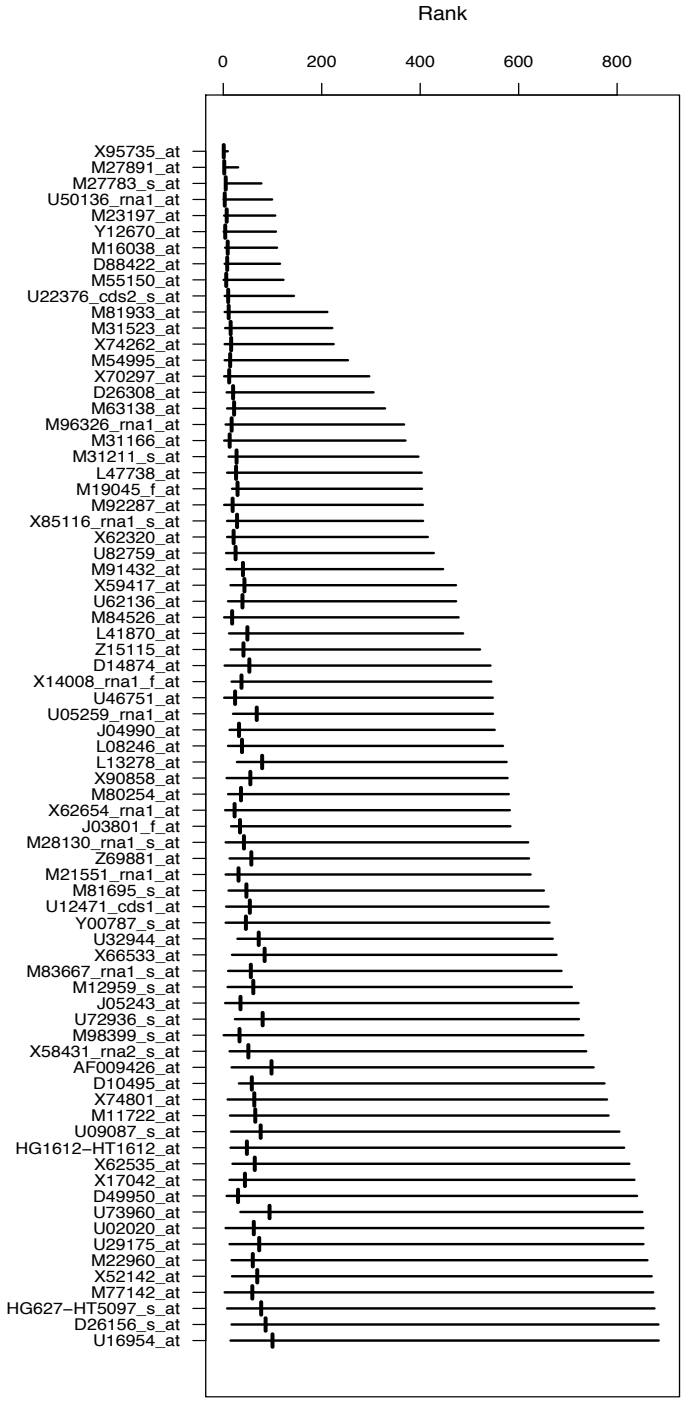


Figure 2.3: Top 67 variables by $\hat{r}_+$ for Example 2.4.2

and correlation equal to 0.85. Let

$$Y_i = \sum_{j=1}^{5} \tfrac{6-j}{5} \left(X_{ij} + X_{i,j+5}\right) + \epsilon_i \,,$$

where ϵ_i is a normal error with zero mean and standard deviation 5. Thus the pairs make a decreasing contribution to Y_i as j increase. Also, let X_{ij} be an independent standard normal random variable, for $11 \leq j \leq 5000$. Thus, Y_i is a linear function of just the first 10 components in a vector of 5000 $\mathrm{N}(0,1)$ components.

To apply the lasso we used the least angle regression (LARS) implementation (Efron et al., 2004), and to implement our method we ranked the absolute values of conventional correlation coefficients. These two approaches were compared by examining the top ten variables that each suggested. For the correlation-based approach this meant taking the ten variables with lowest $\hat{r}_+$, while for the lasso it involved gradually relaxing the penalisation condition until just 10 variables were admitted (note that the cross-validated, lowest-error lasso model under the "one standard error rule" generally admitted fewer than ten variables). For each set we then counted how many main effects were detected (that is, for how many $j \in [1,5]$ did one of $\{X_{ij}, X_{i,j+5}\}$ appear in the set), as well as how many surrogate effects were detected (the number of $j \in [1,5]$ for which both X_{ij} and $X_{i,j+5}$ were in the set). Even though the effects were linear, we have also included results for detections using generalised correlation and group LARS using cubic splines. Group LARS was used rather than the group lasso, for reasons of computational feasibility, but the two methods generally show comparable performance. The experiment was repeated 100 times and the average results are presented in Table 2.2 for various n.

The main feature of the results is that while the lasso and group LARS are better at detecting weaker main effects compared to conventional and generalised correlation respectively, they fail to select the second of each correlated pair of variables. Of course, this is a consequence of using model fitting as a surrogate for variable selection; adding a highly correlated random variable does not greatly improve predictive accuracy, but it nevertheless produces influential variables which, from most practical viewpoints, should be detected by a good variable selector. Thus the results highlight the risk of using a prediction-based method as a means of detecting influential variables. Also, some loss of detection power is observed when moving to nonlinear methods, particularly for lower sample sizes. However, given that this enables the user to detect genuine nonlinear patterns should they exist, the loss appears tolerable.

Figure 2.4 shows typical, randomly chosen results for our bootstrapped ranking approach for various n in this simulation. For this purpose we used 100 bootstrap resamples and took $\alpha = 0.1$. Of note is the increased ability with which weaker

		No. of main effects detected	No. of surrogate effects detected
$n = 100$	lasso	1.76	0.39
	corr	1.58	1.11
	gLARS	1.06	0.49
	gcorr	0.98	0.49
$n = 200$	lasso	2.91	0.99
	corr	2.57	2.06
	gLARS	2.09	1.21
	gcorr	2.25	1.67
$n = 500$	lasso	3.98	2.45
	corr	3.54	3.28
	gLARS	3.43	1.95
	gcorr	3.23	2.94
$n = 1000$	lasso	4.32	3.27
	corr	4.10	3.87
	gLARS	3.91	2.36
	gcorr	3.93	3.66

Table 2.2: Average number of variables detected under simulation

trends are identified, and the increased stability in the ranking of genuine variables as n increases. The theoretical basis for these ideas is covered in Section 2.4.

As discussed in Section 2.3.2, there are practical considerations when choosing the level at which variables are classified as significant. Figure 2.5 gives the number of variables admitted when $n = 500$. It shows four variables appearing very strongly, seen in the flat Section of the curve before 3% of p, and then the number of variables admitted grows exponentially. The proposed $p/2$ level admits only a moderate number of variables (70), but at fractions larger than this the number of variables tends to be unwieldy. Although any choice of cutoff between 3% and 50% might be considered reasonable, and would largely be driven by a user's tolerance of false positives, any presentation should highlight the relative strength of the top four variables. We emphasise that the final choice for the cutoff should ideally be based on the dataset itself.

2.4.4 Example: A non-linear situation. For this simulation study we took the first component to have a nonlinear impact on Y_i and to have contamination of errors-in-variables type: $X_{i1} = W_i + \delta_i$ and $Y_i = W_i^2 - 1 + \epsilon_i$. Here each W_i was taken to be uniform on $[-2, 2]$, and the two error terms, δ_i and ϵ_i, were both normal with zero mean and standard deviation $\frac{3}{4}$. Also, $X_{i2}, X_{i3}, \ldots, X_{i,5000}$ were taken to be independent $\mathrm{N}(0, 1)$ random variables. The simulations were run with $n = 200$, prediction bands for the ranking used $\alpha = 0.02$, and 500 bootstrap simulations were performed.

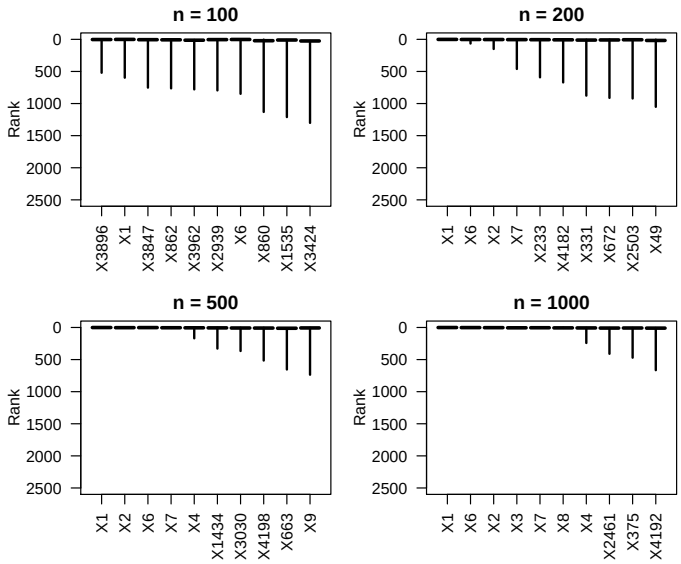


Figure 2.4: Top ten variables by $\hat{r}_+$ for Example 2.4.3 with various n

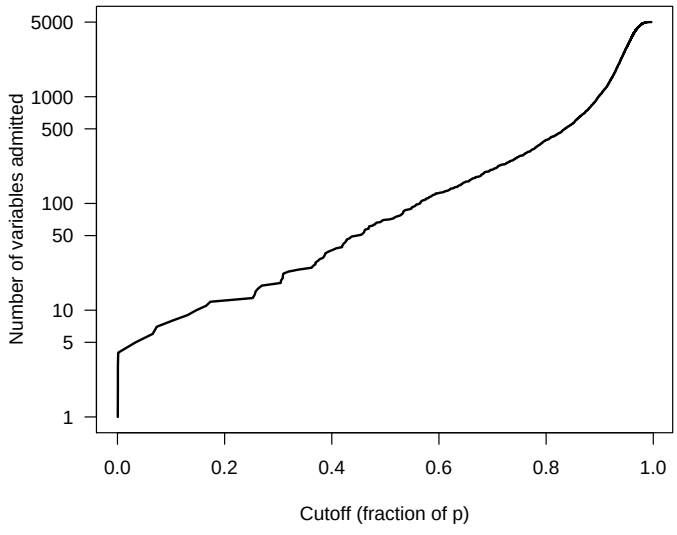


Figure 2.5: Number of variables admitted at various cutoffs for Example 2.4.3 with $n = 500$.

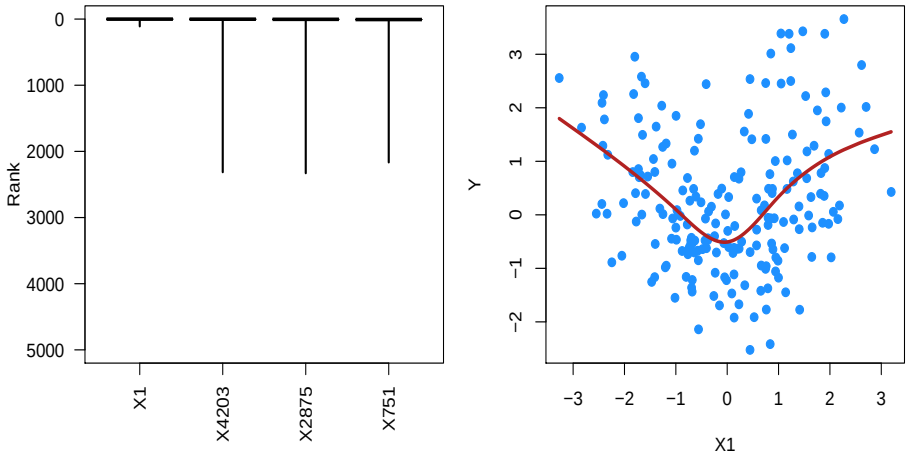


Figure 2.6: Top variables by $\hat{r}_+$ for Example 2.4.4 and the cubic spline fit for X_1

In this case, if ranking is based on conventional correlation then X_{i1} does not appear influential, due to its nonlinear relationship with Y_i. This is true of other linear based approaches; for instance, the lasso fails to detect X_{i1}. Thus the generalised correlation of (2.3) was used, where $\mathcal{H}$ was a basis of natural cubic splines constructed in the same way as in Example 2.4.1. As Figure 2.6 demonstrates, under the second criterion, X_{i1} emerges strongly as the top variable, with only three false positives if we use a cutoff at $\frac{1}{2}\,p$. The natural cubic spline fit captures the relationship between X_{i1} and Y_i, although the plot in Figure 2.6 suggests there is some bias at the limits of X_{i1}.

2.4.5 Example: A highly non-linear situation. Here we report the results of simulating a model with highly non-linear structure. Let $W_{i1}, \ldots, W_{i6}$ and $X_{i5}, \ldots, X_{i,5000}$ be independent standard normal random variables, and put

$$Y_i \;=\; 2\sin\left\{\frac{\pi}{2}\,(W_{i1} + 0.5\,W_{i2})\right\} + \sum_{j=3}^{5} W_{ij}^2 + 0.4\,e^{W_{i6}} + Z_{i0}\,,$$

$X_{i1} = 2\,W_{i1}^2 + Z_{i1}$, $X_{i2} = 2\,W_{i2} + Z_{i2}$, $X_{i3} = W_{i3}\,W_{i4} + Z_{i3}$ and $X_{i4} = W_{i6} + Z_{i4}$, with each of the Z_{ij} being normal random variables with mean 0 and standard deviation 0.1. This simulation was run using natural cubic splines for $\mathcal{H}$, as in Example 2.4.4, with 500 observations, 500 bootstrap simulations and used prediction level of $\alpha = 0.02$.

The variables with the lowest 99% percentile ranking are plotted in Figure 2.7. A comparison of the lengths of prediction intervals shows immediately that just two variables, X_{i3} and X_{i4}, appear influential. Two marginal false positives also have a markedly smaller degree of influence. What is interesting is that X_{i1} and X_{i2} do not appear influential; this is due to the heavy codependence of X_{i1} and X_{i2} in producing

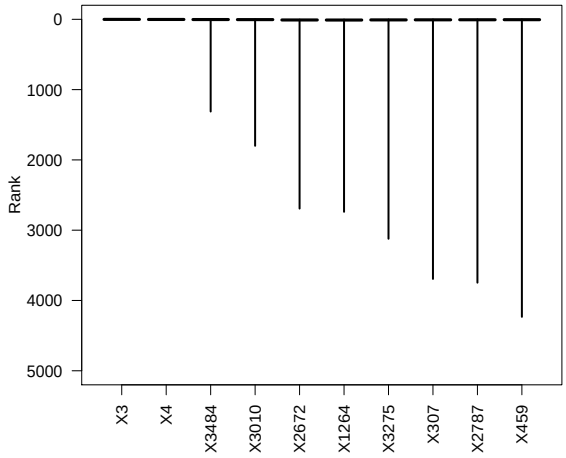


Figure 2.7: Top ten variables by $\hat{r}_+$ for Example 2.4.5

Y_i. This highlights a drawback of measuring the correlation of individual variables; sometimes the combination of several variables may be influential, while individually they are not. Note that if a variable $X_{i,5001} = W_{i1} + 0.5\,W_{i2} + Z_{i5}$, with Z_{i5} normal with mean zero and standard deviation 0.1, were constructed then this would present as influential in the simulation. Interestingly, the lasso (weakly) detects X_{i2} but fails to detect X_{i3}; its inconsistency resulting from the highly nonlinear behaviour of the system.

2.5 Theoretical properties

We shall state and prove a result describing the sensitivity of the rankings given by the method described in Section 2.2. Let $h = h_j$ denote the function for which the supremum in the definition of ψ_j, in (2.3), is achieved. We take $\mathcal{H}$ to be a class of polynomials — see assumption (2.6)(b) below — and in that case the supremum is achieved at a particular element of $\mathcal{H}$. Since our methodology is invariant under changes to the scales of Y_i and to the components of X_i, then in formulating our assumptions below we may assume without loss of generality that $\mathrm{var}\{h_j(X_{ij})\} = \mathrm{var}(Y_i) = 1$ for each i and j; see (2.6)(c) below. In all other respects, except where constrained by (2.6)(e), we allow the distribution of (X_i, Y_i) to vary with n. We think of p, too, as a function of n, diverging to infinity as n increases, but diverging at no faster than a polynomial rate; see (2.6)(d). Our main other assumption is a moment condition, (2.6)(e):

(a) the pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are independent and identically distributed;
(b) $\mathcal{H}$ is the class of polynomial functions of degree up to but not exceeding the positive integer $d \geq 1$;
(c) $\text{var}\{h_j(X_{ij})\} = \text{var}(Y_i) = 1$ for each i and j;
(d) for a constant $\gamma > 0$ and all sufficiently large n, $p \leq \text{const.}\, n^\gamma$; and
(e) for a constant $C > 4\,d\,(\gamma + 1)$, $\sup_n \max_{j \leq p} E|X_{1j}|^C < \infty$ and $\sup_n E|Y_1|^C < \infty$.

(2.6)

Given constants $0 < c_1 < c_2 < \infty$, write $\mathcal{I}_1(c_1)$ and $\mathcal{I}_2(c_2)$ for the sets of indices j for which $|\,\text{cov}(X_i, Y_i)| \leq c_1\,(n^{-1}\,\log n)^{1/2}$ and $|\,\text{cov}(X_i, Y_i)| \geq c_2\,(n^{-1}\,\log n)^{1/2}$, respectively.

Theorem 2.2. *Assume* (2.6). *If, in the definitions of* $\mathcal{I}_1(c_1)$ *and* $\mathcal{I}_2(c_2)$, *the constants* c_1 *and* c_2 *are chosen sufficiently small and sufficiently large, respectively, then, in the correlation-based ranking at* (2.5), *with probability converging to 1 as* $n \to \infty$ *all the indices in* $\mathcal{I}_2(c_2)$ *are listed before any of the indices in* $\mathcal{I}_1(c_1)$.

Before proving Theorem 2.2, we discuss its implications. The theorem argues that the sensitivity point for component ranking based on correlation, or covariance, is on the scale of $(n^{-1}\,\log n)^{1/2}$. In particular, components for which the covariances are at least as large as sufficiently large constant multiples of $(n^{-1}\,\log n)^{1/2}$, are very likely to be ranked ahead of covariances which are of smaller order than this. To appreciate the clarity of the implications of this result, assume for simplicity that $\mathcal{H}$ is the set of linear functions, suppose that exactly q (a fixed number) components of X are correlated with Y, and have correlation coefficients whose absolute values are bounded above a positive constant; and that all the other components have correlations with Y which are uniformly of smaller order than $(n^{-1}\,\log n)^{1/2}$. For example, this would be the case if all the latter components of X were uncorrelated with Y. Then, with probability converging to 1 as p increases, all the q correlated components are listed together in the first q places of the ranking at (2.4), and all the other components are listed together in the last $p - q$ places.

Proof of Theorem 2.2: Using moderate-deviation formulae for probabilities associated with sums of independent random variables (see Section 1.6), it can be shown that if $b > 0$ is given, and if $\sup_n \max_{j \leq p} E|X_{1j}|^C < \infty$ for some $C > 4\,d\,(b + 1)$, then

$$P\Big\{\big|\hat{\psi}_j - \psi_j\big| > c_0\,\big(n^{-1}\,\log n\big)^{1/2} \quad \text{for some} \quad 1 \leq j \leq p\Big\} = O(\delta)\,,$$

where c_0 is a constant and $\delta = p\,n^{-b}\,(\log n)^{-1/2}$. Hence, with probability equal to

$1 - O(\delta)$, $|\hat{\psi}_j| \le 2\, c_0\, (n^{-1} \log n)^{1/2}$ for all j such that $|\psi_j| \le c_0\, (n^{-1} \log n)^{1/2}$, and $|\hat{\psi}_j| > 2\, c_0\, (n^{-1} \log n)^{1/2}$ for all j for which $|\psi_j| > 3\, c_0\, (n^{-1} \log n)^{1/2}$. It follows that if, in the definitions of the sets $\mathcal{I}_1(c_1)$ and $\mathcal{I}_2(c_2)$ of indices, $c_1 \le c_0$ and $c_2 > 3\, c_0$, then, in the ranking at (2.4), with probability equal to $1 - O(\delta)$, all the indices in $\mathcal{I}_1$ are placed ahead of all the indices in $\mathcal{I}_2$. Provided $p \le \text{const.}\, n^\gamma$ (as specified in (2.6)(c)), and $b \ge \gamma$, we have $\delta \to 0$ as $p \to \infty$.

Chapter 3

Generalised correlation for variable relationships

3.1 Background

A standard statistical approach to solving variable-selection problems involves determining, for a long data vector X, a relatively small number of components (or variables) on which correct classification or prediction depends. This was precisely the aim of the previous chapter, which explored ways to achieve this when the response was continuous and nonlinear relationships were believed to exist. These identified components might represent a small number of genes (say, in genomic problems), typically between a few and a few tens, out of thousands or tens of thousands of components in X. Once this feature selection has been effected, a further key task is understanding what these selected features represent. In some instances the influential genes might be selected primarily because they each represent, in different ways, the same phenomenon. In this case, if a final predictive model was required, some of the components could, in the presence of others, be essentially redundant, in that they might be deleted without appreciably changing the performance of a classifier or predictor. Alternatively the small set of components may represent different effects, and so each contribute meaningful information to a final model.

In theoretical terms the potential for these phenomena is clear. For example, if two components are highly correlated, where correlation is measured conventionally, then one of them can often be deleted without the final result being greatly influenced. More generally, if one of the components is a function of the other then it might be possible to drop either of the components without affecting the overall performance of a classifier or predictor. These occurrences are not limited to very high-dimensional

variable selection problems; even when the number of variables is only moderate there is potential for highly influential variables to be closely related, and potentially redundant.

However, while avoiding redundancy is important, there are still more compelling motivations for understanding the relationships among the "significant" components of X. Indeed, in genomic problems there are good scientific reasons for wishing to understand the joint behaviour of different, influential components. For example, it is important to comprehend the manner in which the components identified by variable selection operate together to exert their influence. We might ask whether the expression levels of two particular variables tend to increase and decrease together, or whether there is a more subtle and complex connection between them. Can we quantify and explain this complexity, thereby gaining a better understanding of the problem than just that the fact that the d selected variables seem to be more "significant" than others for classification or prediction? In particular, insight into the way in which gene expression levels vary jointly, to produce an overall significant effect, can be of greater value than simply knowing which set of genes is the most significant.

Here we suggest a simple way of answering questions such as these. We continue to use techniques based on generalised correlation, and show how to apply them to explore relationships among different components. We introduce graphical methods that enable us to access this sort of information quickly and reliably, thereby guiding the experimenter towards further pertinent questions that could be asked of the data.

Real-data examples where these issues are important arise in a diverse range of situations. We illustrate this point using multiple real-data examples, with a medical emphasis. Analysis of these examples can be found in Section 3.3.1. The first example is the Leukemia dataset of Golub et al. (1999) introduced in Section 2.2.2. The majority of approaches for this type of dataset seek to first reduce the number of influential genes to a manageable set, and then use this set to draw conclusions and make predictions. The previous chapter provides one method to perform this first step, but other mean of dimension reduction exist as well. See, for example, Golub et al. (1999), Tibshirani et al. (2002) and Fan and Lv (2008). It is useful to understand how the genes produced by the selection process relate to one another.

Secondly, the Wisconsin breast cancer dataset[1] contains nine predictor variables characterising the properties of an X-ray breast mass, together with a categorical response variable indicating whether the mass is a malignant tumour or not. The data were first discussed by Street et al. (1993), using 569 observations, and another 130 observations have been added since that time. We ran a standard random forest model (Breiman, 2001a) on the dataset. This model achieves good prediction (97%

[1]Downloadable from the UCI machine learning database http://archive.ics.uci.edu/ml/index.html

accuracy) when applied to out-of-bag observations, and the variable importance output suggests that the second predictor, uniformity of cell size, is the most important. However, removing this predictor and again creating a random forest model gives marginally better prediction. This suggests heavy dependence between the second predictor and other variables.

Thirdly we examine the hepatitis survival dataset[2] analysed by Diaconis and Efron (1983) and Cestnik et al. (1987). These approaches used logistic regression on 19 predictors to estimate the survival or non-survival of 155 hepatitis patients. The methodology was critiqued by Breiman (2001b), who pointed to interactions between variables as hindering the logistic techniques and reducing predictive accuracy. A semi-automatic means of detecting such interactions, for example the methods introduced in this chapter, is therefore important in this situation.

3.2 Methodology

3.2.1 Generalised correlation for measuring strength of association and the potential for prediction.

Assume that a variable selection method has narrowed the number of "significant" components from a very large value, equal to the length p of X, to just d. For simplicity we designate these by $X^{(1)}, \ldots, X^{(d)}$. We define the (symmetric) generalised correlation between $X^{(j_1)}$ and $X^{(j_2)}$ to be

$$\rho_S(j_1, j_2) = \sup_{g_1, g_2 \in \mathcal{G}} \operatorname{cor}\left\{g_1\big(X^{(j_1)}\big), g_2\big(X^{(j_2)}\big)\right\}, \tag{3.1}$$

where $\mathcal{G}$ represents a class of functions (for example, the class of polynomials of a given degree), and $\operatorname{cor}(U, V)$ denotes the standard correlation coefficient between random variables U and V. We interpret $\rho_S(j_1, j_2)$ as the extent of association between $X^{(j_1)}$ and $X^{(j_2)}$, in the sense of generalised correlation. Note that, if $\mathcal{G}$ denotes the class of linear functions, then $\rho_S(j_1, j_2) = |\operatorname{cor}(X^{(j_1)}, X^{(j_2)})|$.

We also define the asymmetric, or predictive, version of generalised correlation,

$$\rho_A(j_1, j_2) = \sup_{g \in \mathcal{G}} \operatorname{cor}\left\{X^{(j_1)}, g\big(X^{(j_2)}\big)\right\}, \tag{3.2}$$

which can be interpreted as a measure of the potential for predicting $X^{(j_1)}$ from a function of $X^{(j_2)}$, when that function comes from $\mathcal{G}$. In particular, if $\mathcal{G}$ is closed under addition of scalars and under scalar multiplication, then $\rho_A(j_1, j_2) = 1$ if and only if $X^{(j_1)} = g(X^{(j_2)})$ with probability 1, for some $g \in \mathcal{G}$; and $\rho_A(j_1, j_0) = 0$ if $X^{(j_1)}$ and $X^{(j_2)}$ are statistically independent. Note that, if $\mathcal{G}$ is the class of linear

[2]Downloadable from the UCI machine learning database http://archive.ics.uci.edu/ml/index.html

functions, then $\rho_S(j_1, j_2) = \rho_A(j_1, j_2) = \rho_A(j_1, j_2)$. The subscripts in ρ_S, ρ_A denote the symmetric and asymmetric versions of the generalised correlation measure.

The predictive correlation $\rho_A(j_1, j_2)$ is precisely the tool introduced in Chapter 2, while $\rho_S(j_1, j_2)$ represents a further extension. However, $\rho_A(j_1, j_2)$ and $\rho_S(j_1, j_2)$ can be used to explore relationships among components in very general settings, for example when $X^{(1)}, \ldots, X^{(r)}$ are selected using a conventional linear model-based method such as the lasso (Tibshirani, 1996; Chen et al., 1998) or the Dantzig selector (Candes and Tao, 2007), or when $X^{(1)}, \ldots, X^{(r)}$ are identified in terms of their leverage for classification (e.g. Hall et al., 2009) rather than via more conventional variable selection.

3.2.2 Estimators of $\rho_S(j_1, j_2)$ and $\rho_A(j_1, j_2)$. Assume, as in Section 3.2, that X has already been reduced to a much shorter vector of length d, using a variable selection method such as those related to prediction or classification. However, in an abuse of notation we shall refer to the shorter vector as X, and in particular we shall suppose that we observe data vectors $X_i = (X_{i1}, \ldots, X_{id})$ for $1 \le i \le n$, where each X_i is distributed as $X = (X^{(1)}, \ldots, X^{(d)})$. Define $\bar{X}_{\cdot j} = n^{-1} \sum_i X_{ij}$ and $\bar{X}_{\cdot j}(g) = n^{-1} \sum_i g(X_{ij})$. Estimators of $\rho_A(j_1, j_2)$ and $\rho_S(j_1, j_2)$ are given respectively by

$$\begin{aligned}
\hat{\rho}_S(j_1, j_2) &= \sup_{g_1, g_2 \in \mathcal{G}} \frac{\sum_i \{g_1(X_{ij_1}) - \bar{X}_{\cdot j_1}(g_1)\} \{g_2(X_{ij_2}) - \bar{X}_{\cdot j_2}(g_2)\}}{\left[\sum_i \{g_1(X_{ij_1}) - \bar{X}_{\cdot j_1}(g_1)\}^2 \sum_i \{g_2(X_{ij_2}) - \bar{X}_{\cdot j_2}(g_2)\}^2\right]^{1/2}}, \\
\hat{\rho}_A(j_1, j_2) &= \sup_{g \in \mathcal{G}} \frac{\sum_i (X_{ij_1} - \bar{X}_{\cdot j_1}) \{g(X_{ij_2}) - \bar{X}_{\cdot j_2}(g)\}}{\left[\sum_i (X_{ij_1} - \bar{X}_{\cdot j_1})^2 \sum_i \{g(X_{ij_2}) - \bar{X}_{\cdot j_2}(g)\}^2\right]^{1/2}}.
\end{aligned}$$

Compare (3.1) and (3.2). If the class $\mathcal{G}$ is determined by only a finite number of parameters (for example, as in the case where $\mathcal{G}$ is the set of all polynomials of given degree), then generally, under mild additional assumptions (for instance, moment conditions in the polynomial example), the estimators $\hat{\rho}_S(j_1, j_2)$ and $\hat{\rho}_A(j_1, j_2)$ are root-n consistent for $\rho_S(j_1, j_2)$ and $\rho_A(j_1, j_2)$, respectively.

3.2.3 Graphical methods for depicting $\hat{\rho}_S(j_1, j_2)$ and $\hat{\rho}_A(j_1, j_2)$. There are several ways of depicting graphically the values of $\hat{\rho}_S(j_1, j_2)$ and $\hat{\rho}_A(j_1, j_2)$. We describe two of them here. First, we suggest representing $\hat{\rho}_S(j_1, j_2)$ and $\hat{\rho}_A(j_1, j_2)$ in terms of the darkness of a grey shade, or the warmth of a colour in a spectrum, using a square matrix. Specifically, construct an $r \times r$ array of square boxes, and colour box (j_1, j_2) to reflect the value of $\hat{\rho}_A(j_1, j_2)$, where j_1 and j_2 are indicated on the vertical and horizontal axes, respectively.

In this depiction the boxes down the main diagonal would be black, if using grey shade to represent the potential for prediction, or dark red, if using colour for that

purpose. (Of course, $\hat{\rho}_A(j\,,j) = 1$ for each j.) A representation of $\hat{\rho}_S(j_1, j_2)$ would be similar, except that this quantity is symmetric and so does not, in principle, require entries both above and below the main diagonal. However, it is helpful to be able to view the relationships between one variable and all the others in a single row, or single column, and so we suggest retaining all the boxes both above and below the diagonal.

A second way of depicting $\hat{\rho}_A(j_1, j_2)$ graphically is to place d points in the plane, numbered from 1 to d, and link points j_1 and j_2 by an arrow leading from j_2 to j_1. In the resulting diagram the value of $\hat{\rho}_A(j_1, j_2)$ can be represented by the thickness of the arrow (if using only the colour black), or the darkness of the grey shade or the colour of the arrow. It is helpful to locate the numbered points in the plane strategically, so that the graphical representation is as uncluttered as possible. Relatively weak predictive potential can be ignored, for example by confining attention to pairs (j_1, j_2) for which $\hat{\rho}_A(j_1, j_2)$ exceeds a given threshold. Strength of association can be represented similarly, on a separate diagram. In this instance, in recognition of the symmetry of $\hat{\rho}_S(j_1, j_2)$, the arrows should be double-ended.

3.2.4 Graphing predictive relationships. Provided sample size, n, is not too small, it is possible to go beyond the simple numerical descriptor $\hat{\rho}_A(j_1, j_2)$ when assessing the potential that $X^{(j_2)}$ has for predicting $X^{(j_1)}$. For example, we can investigate a simple regression model,

$$X^{(j_1)} = g(X^{(j_2)}) + \text{error}\,, \tag{3.3}$$

where $g = g_{j_1 j_2}$ is estimated nonparametrically rather than through being constrained to lie in a predetermined function class $\mathcal{G}$. If we observe data vectors X_i distributed as $X = (X^{(1)}, \ldots, X^{(d)})$ (see Section 3.2.2) then we can estimate $g_{j_1 j_2}$ using standard methods, for instance employing local-linear regression with a bandwidth chosen by cross-validation, and construct a lattice, or set of trellis plots, of graphs of function estimates. Examples will be given in Section 3.3. The concept of trellis plot has its roots in work of Chambers and Hastie (1992) and Becker et al. (1996).

While this method has the potential to provide a relatively high amount of information in visual form, in genomic applications it can be seriously inhibited by the small sample sizes that commonly occur in practice, or by the difficulty that a reader has absorbing, and placing into context, all the information that arises when d is relatively large. In such cases the cruder representations discussed in Section 3.2.3 can be more useful.

3.3 Examples

3.3.1 Real-data examples. We begin with the leukemia dataset and include plots showing how the relationships may be presented. Here we first take the top ten genes, as found in Section 2.4.2, thus reducing the problem of finding variable relationships to one with manageable dimension. The associative and predictive potential between each pair of genes were calculated. Firstly the associative potentials $\rho_S(j_1, j_2)$ are presented in Figure 3.1. Recall that this measure is symmetric, so the matrix diagram is symmetric about its main diagonal and double-headed arrows are used in the arrow diagram. The order of variables was taken from the original ranking, but the points were rearranged in the arrow diagram to reduce clutter. The darkness of the arrows indicates how much higher the relationship is above the threshold.

Unless otherwise stated, we took the class of functions $\mathcal{G}$, in (3.1) and (3.2), to be the set of all natural cubic splines with knots at data quartiles. This enables our definition of "generalised correlation" to capture a simple wiggle or turning point in the middle of an otherwise monotone function, and in this respect generalised correlation is more appropriate than classes of linear or quadratic functions.

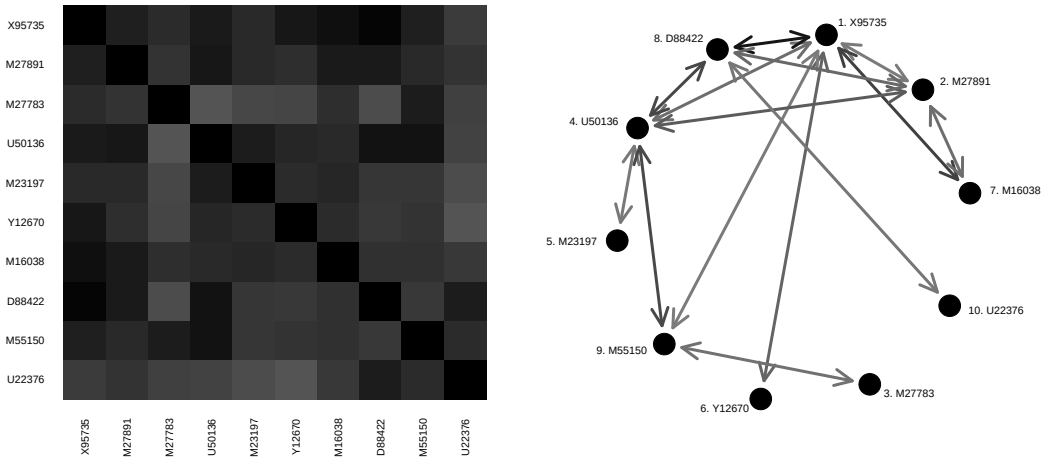


Figure 3.1: Associative potential for AML/ALL genes

The most interesting feature of Figure 3.1 is that the associations are generally very high. It can be seen quickly from the matrix diagram that the first variable is quite strongly associated with all the others, and, on the other hand, that the third variable is generally weakly related. Similar results are obtained in the case of prediction; see Figure 3.2. They suggest that, for example, the first variable, X95735, might perform as well on its own as it does together with the variables

that are associated with it, or which predict it. In particular, the variables that are good predictors of X95735 might be removed without predictive performance being appreciably affected. We shall explore this possibility four paragraphs below.

The arrow diagram in Figure 3.1 uses a threshold of 0.85 to exclude a reasonable number of pairs. In fact, the lowest association is 0.67 (the relationship between the third and fourth genes), which is still strikingly high. This issue is discussed further below. The large number of close associations between the first variable and the others, and the small number of associations involving the third variable, are also reflected in the arrow diagram. However, that diagram displays relatively little information about strength, although it gives a clearer picture than does the matrix diagram of the pattern of linkages.

Figure 3.2 presents the asymmetrical predictive relationships $\rho_A(j_1, j_2)$. Generally features are similar to the associative results. For example, there is a clearly a relatively large number of predictive relationships involving the first variable, and a relatively small number involving the third variable. The threshold for the arrow diagram has been kept at 0.85 for direct comparison to Figure 3.1, although the diagram is noticeably more cluttered. Interesting features are those that display one way predictiveness. For instance the fourth variable (U50136) is strongly predictive for the eighth (D88422) but not vice versa. Figure 3.3 contains a scaled plot of the D88422 against U50136, and it is evident why this one-directional relationship exists; D88422 has increased expression only when U50136 is above a particular threshold. This means that while U50136 is a good predictor for D88422, it is difficult to use D88422 to separate low to medium expression levels of U50136, limiting its predictiveness. It can be observed that this is a strong relationship that would not be fully captured by traditional linear correlation measures.

As described in Section 3.2.4 it is possible to investigate, using regression models, the ability to predict variables from others. Figure 3.4 shows the trellis plots for each pairwise combination of standardised variables. The fitted line is a local linear fit using a bandwidth that minimises the generalised cross-validation statistic (see Loader, 1999). The trellis plot gives further insight into previous results. For instance, the lighter third row seen in the matrix representations in Figures 3.1 and 3.2 are largely explained by an outlier in the third variable (M27891) that is not reflected in any of the other gene expressions.

As mentioned earlier, it is possible to use the idea of variable predictiveness to eliminate redundant variables. Suppose we use the arrow plot in Figure 3.2 to remove variables, by eliminating the variables whose dots are pointed to. If we decide to keep the first variable, judged the most important, then removing the second, fourth, seventh and eight variables would remove all arrows from the diagram. A basic random forest model to predict leukemia type using all ten genes has one out-of-bag

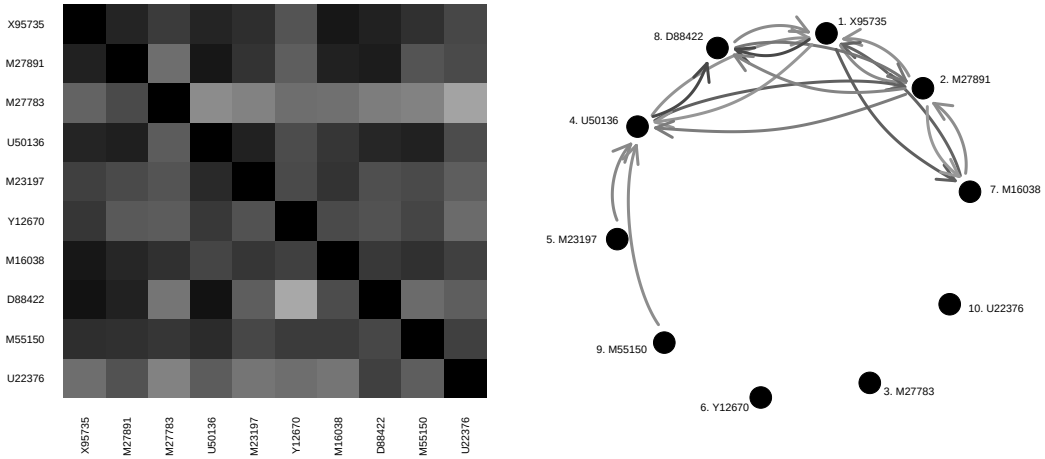


Figure 3.2: Predictiveness potential for AML/ALL genes

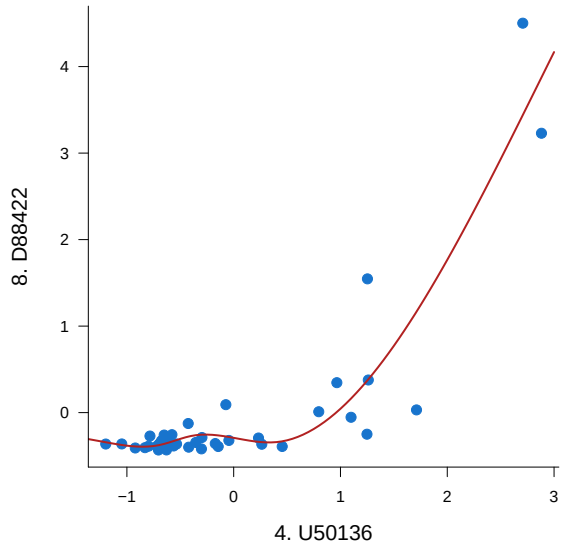


Figure 3.3: Plot of 4th variable against 8th with natural cubic spline fit

error (analogous to a cross-validated error) of 1/38 on the learn set and two errors on the test set. However a random forest excluding the second, fourth, seventh and eighth variables generates one out-of-bag error on the learn set and just one error on the test set. Although the improvement may partly be due to random noise in the data, it certainly suggests that performance of the six gene model is at least as good as that of the ten gene model. The analysis has successfully removed unnecessary variables.

Some further comment should be made regarding the high levels of correlation found in the AML/ALL example. Much of this is due to how the genes were originally selected; if two genes are powerful in separating the two cancer groups, then they are both likely to correlate closely with the response and hence with each other. This phenomenon is easily reproducible in theoretical examples, as the final example in Section 3.3.2 demonstrates. However, the effect is unlikely to account for all the correlation, and there is the alternative explanation that the gene responses are correlated for biological reasons; for example, the gene pathways may overlap.

Other gene selections show similar high correlations between genes. For instance, consider the set of fifty genes selected by Golub et al. (1999). Twenty-five of these genes had high expression levels for AML, and for these the average predictive potential $\rho_A(j_1, j_2)$ was 0.60. The other twenty-five genes had high expressions for ALL and these had an average predictive potential of 0.58. The implications of these relationships for prediction are significant. For instance, a scheme where each gene "votes" for the Leukemia type has fewer effective votes than the number of genes due to the high correlations.

The association plots for the Wisconsin breast cancer dataset are presented in Figure 3.5. The arrow diagram uses a (relatively high) threshold of 0.7. Here $\mathcal{G}$ contained only linear functions, so associations were equivalent to standard correlations. The clearest feature is the heavy association between the second and third variables. This explains why the removal of uniformity of cell size, the second variable, did not impact model prediction at all, as discussed in Section 3.1. Comparing $\rho_A(2, 3)$ and $\rho_A(3, 2)$ with $\mathcal{G}$, including cubic splines, does not add much further insight as both variables appear equally good at predicting each other.

The hepatitis data were slightly more difficult to analyse than the previous two, due to the presence of missing values as well as two-level categorical predictor variables. The first of these issues was addressed by considering only observations that contained no missing values for each pairwise choice of variables. To address the second, each categorical variable was coded as a 0-1 variable and was treated numerically for correlations. The categorical variables had only the identity function in the set of functions $\mathcal{G}$, while the other variables contained the space formed by natural cubic splines with three interior knots. Aside from the strong associations

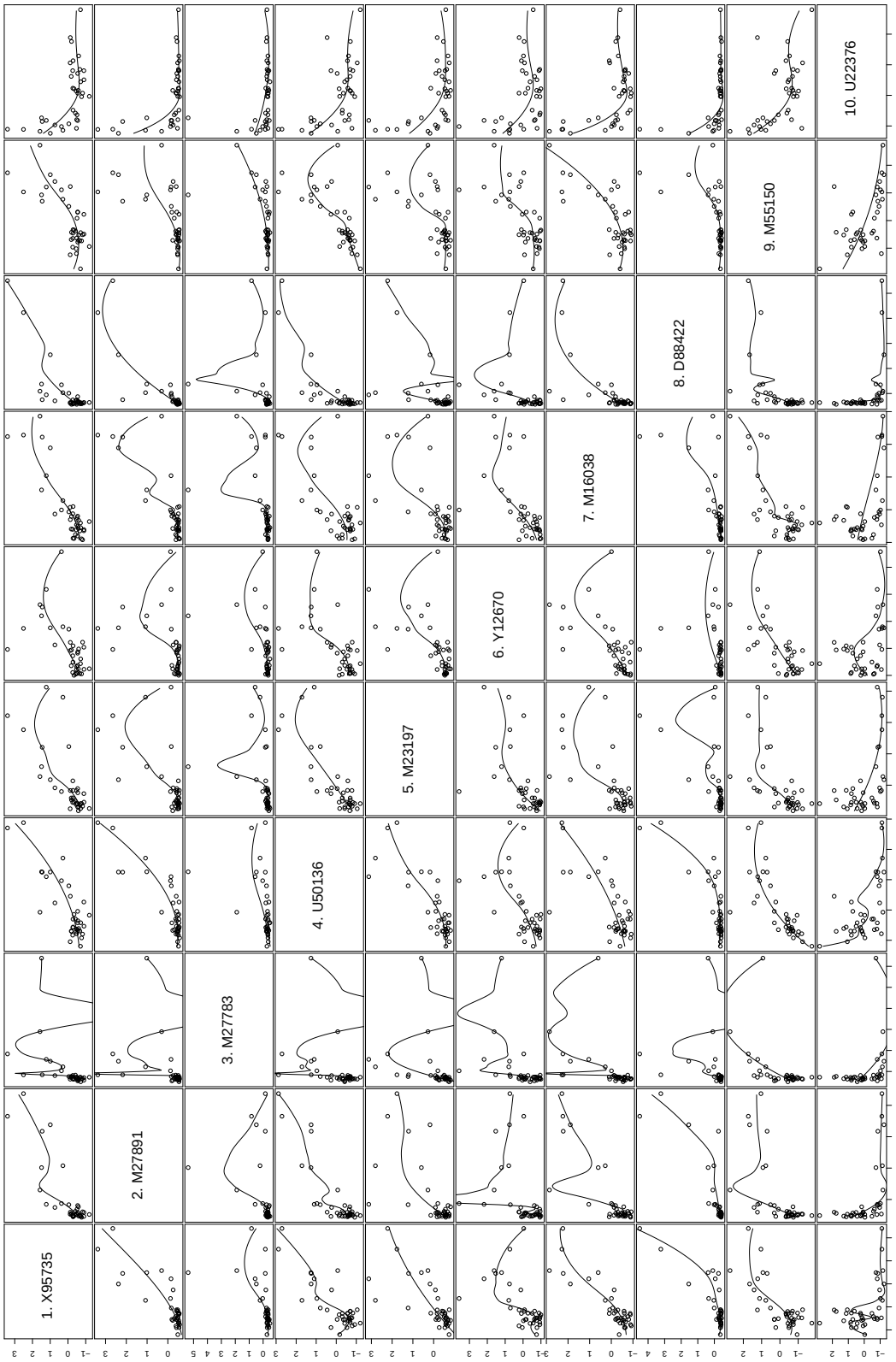


Figure 3.4: Trellis plots of Leukemia genes with local linear fit

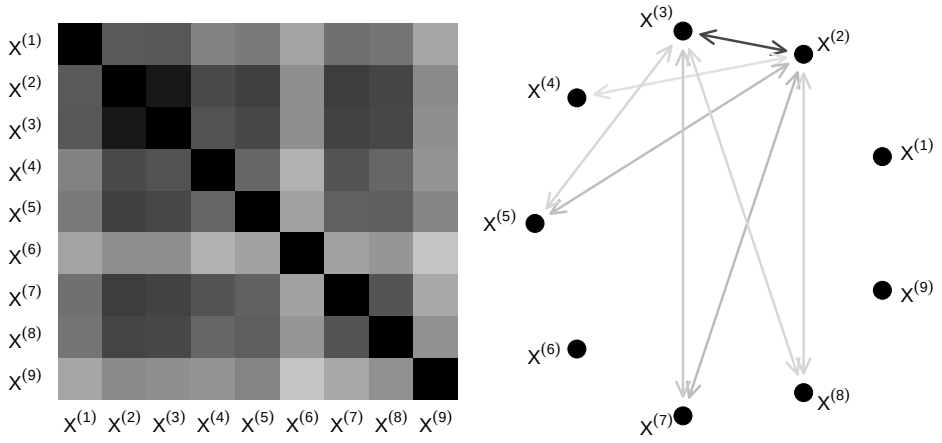


Figure 3.5: Association plots for the Wisconsin breast cancer data

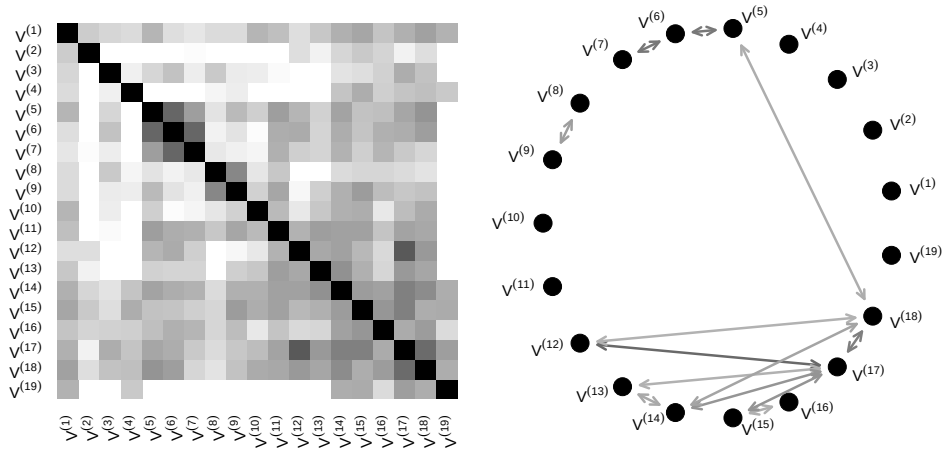


Figure 3.6: Association plots for hepatitis data

of the sixth variable with both the fifth and seventh, the dominant feature in the figure is the heavy relationship between the twelfth and seventeenth variables. This is exactly the relationship detected by Breiman (2001b) and cited as the key problem with prior attempts at logistic regression. Variable 17 is a much stronger predictor of variable 12 than the reverse ($\rho_A(12,\, 17) = 0.65$ while $\rho_A(17,\, 12) = 0.55$), so a case could be made for excluding variable 12 from a final model.

3.3.2 Theoretical examples based on random-effects models. In the examples below, and in graphical depictions of the strength of association or prediction, each arrow that links two components represents a random effect that they have in common. In a matrix depiction, the darkness that corresponds to a pair of components is proportional to the number of random effects that the components share. We shall use standard correlation to measure relationship, and in this case the corre-

lation that explains strength of prediction is exactly the same as the correlation that explains association. Therefore we shall refer simply to association.

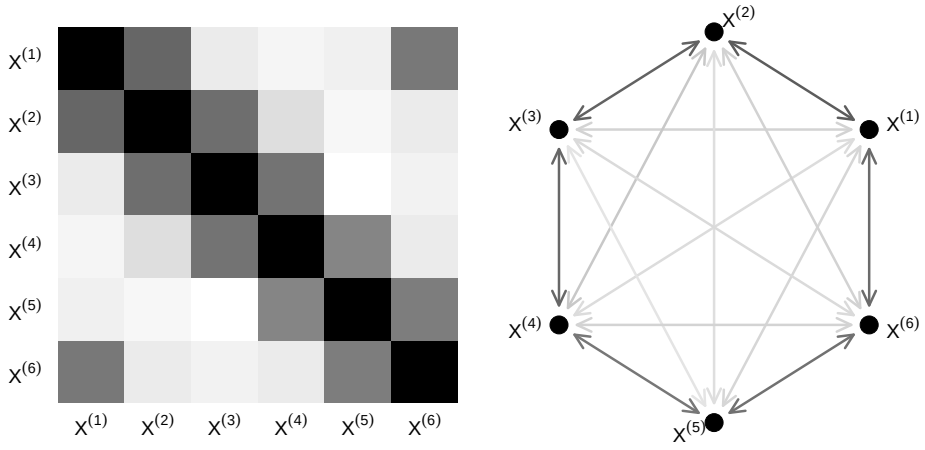


Figure 3.7: Association plots for periodic case $(r, s) = (6, 2)$

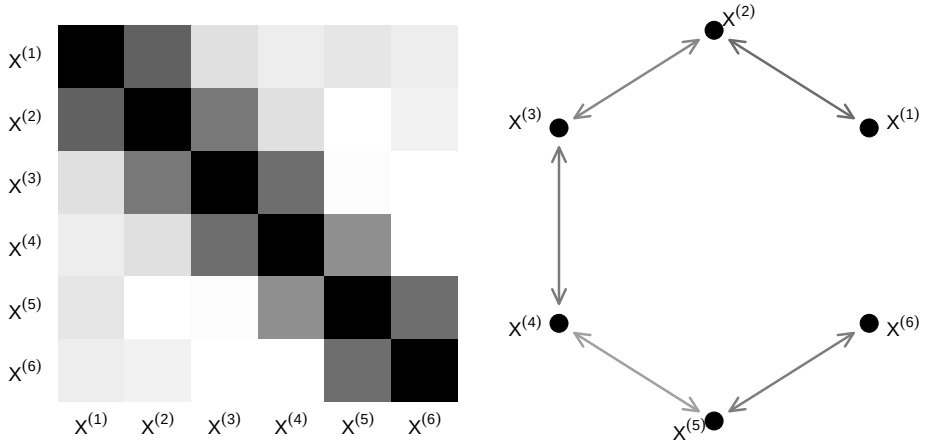


Figure 3.8: Association plots for aperiodic case $(r, s) = (6, 2)$

First we consider a "chain" structure, where each component is associated with just two neighbours in a simple closed circuit. Let $X^{(j)} = U_j + U_{j+1}$ for $1 \leq j \leq r-1$, and $X^{(r)} = U_r + U_1$, where the variables U_j are independent and identically distributed with finite variance. Then $\mathrm{cor}(X^{(j)}, X^{(j+1)}) = \frac{1}{2}$ for $1 \leq j \leq r-1$, $\mathrm{cor}(X^{(1)}, X^{(r)}) = \frac{1}{2}$, and $\mathrm{cor}(X^{(j_1)}, X^{(j_2)}) = 0$ for all distinct pairs $\{j_1, j_2\}$ not covered by this scheme. More generally we could take

$$X^{(j)} = U_j + \ldots + U_{j+s} \tag{3.4}$$

for all $j \in [1, r-s]$, and either complete the sequence (for values $j = r-s+1, \ldots, r$) in a periodic fashion, as suggested in the previous, simpler example, or define the

sequence nonperiodically, interpreting (3.4) as holding for all $j \in [1, r]$. Figures 3.7 and 3.8 show, in the case $(r, s) = (6, 2)$, graphical depictions in periodic and non-periodic cases. Normal random variables with mean zero and standard deviation one were used for the U_j with $n = 100$. In the periodic case, the pairwise association is evidenced by both the matrix diagram, with dark shades on the diagonals closest to the main diagonal, as well as the arrow diagram, which has dark arrows between neighbouring points. Here the arrow diagram includes all arrows, to give a sense of the relative associations. In the aperiodic case a threshold of 0.2 was used and only associations above this level were included in the diagram. Again the pairwise associations are clear.

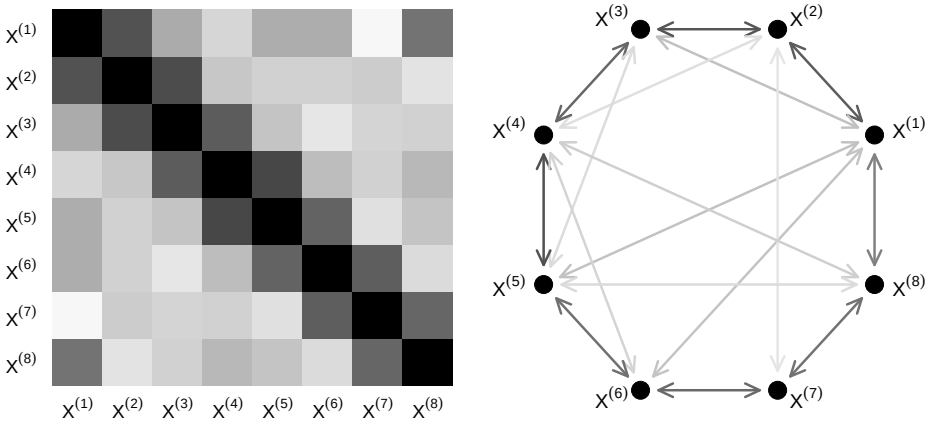


Figure 3.9: Association plots for periodic case $(r, s) = (8, 3)$

As a slightly more complex example the periodic case when $(r, s) = (8, 3)$ is presented, using the same assumptions for U_j and n. Then the theoretical associations are $2/3$ for pairs (X_j, X_{j+1}) (interpreted cyclically), $1/3$ for pairs (X_j, X_{j+2}) and zero for other pairs. Figure 3.9 displays the results. The arrow diagram used a threshold of 0.2. The strong associations are clearly visible, while the weaker associations (those with a theoretical association of $1/3$) have been partially obscured by noise; not all of these associations appear in the arrow diagram and some false associations appear in preference. While increasing n would better resolve the associations, the example demonstrates how noise can hide weaker associations.

If $r = 4$ and we define $X^{(1)} = U_1 + U_4$, $X^{(2)} = U_1 + U_2 + U_5$, $X^{(3)} = U_2 + U_3$ and $X^{(4)} = U_3 + U_4 + U_5$ then, in a graphical representation, the pattern of arrows can be depicted as a square with vertices $X^{(1)}, \ldots, X^{(4)}$ and with an additional line drawn as the diagonal between $X^{(2)}$ and $X^{(4)}$. Figure 3.10 plots the results with $n = 100$, using standard normal random variables and a threshold of 0.2 for the arrow diagram. As expected, associations exist for each pair of variables excluding the pair (X_1, X_3). Many more complex examples can be constructed along similar lines.

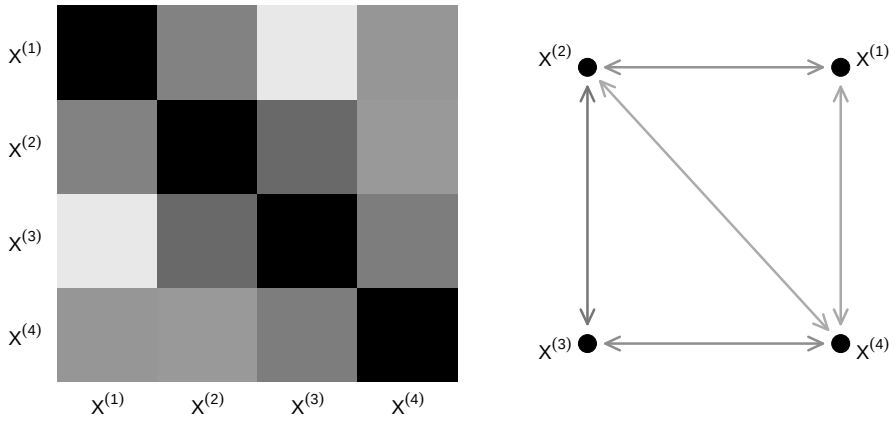


Figure 3.10: Association plots for $r = 4$ example

The final example in this section demonstrates the artificially inflated relationships that can occur when selecting variables from a large set, using a response vector. For example, suppose we have $p = 10{,}000$ uncorrelated standard normal random variables and a response variable which is also an independent standard normal variable. When we take $n = 30$ and choose the five variables that best correlate with the response, there is an average correlation (taking absolute values) of 0.41 between these variables, much higher than the 0.15 observed over all variables. This demonstrates that selected subsets tend to overstate the amount of correlation.

3.3.3 Comparisons with partial correlation. An alternative approach to standard or generalised correlations for detecting variable relationships is using partial correlation. The partial correlation between two variables is the (standard) correlation between the two with the effect of a set of other variables removed. Applying partial correlation to the situation described in Section 3.2, the partial correlation between $X^{(i)}$ and $X^{(j)}$ is the correlation of the residuals of these variables when each is linearly regressed on $X^{-(i,j)}$, the set of variables excluding the ith and jth. This may be denoted as $\rho_{i,j\cdot-(i,j)}$. The estimation of partial correlation is closely related to estimating the inverse covariance matrix and the original work may be found in the paper by Dempster (1972). Recent work applying partial correlation to high-dimensional settings includes that of Meinshausen and Bühlmann (2006) and Peng et al. (2009).

The key benefit in using partial correlation is that causal relationships are extracted. For instance, if two variables are highly correlated due only to their dependence on a third controlling variable, then this correlation will be ignored when partial correlations are taken, leaving only the relationships with the controlling variable. This allows a network of variables to be generated with controlling variables

clearly visible.

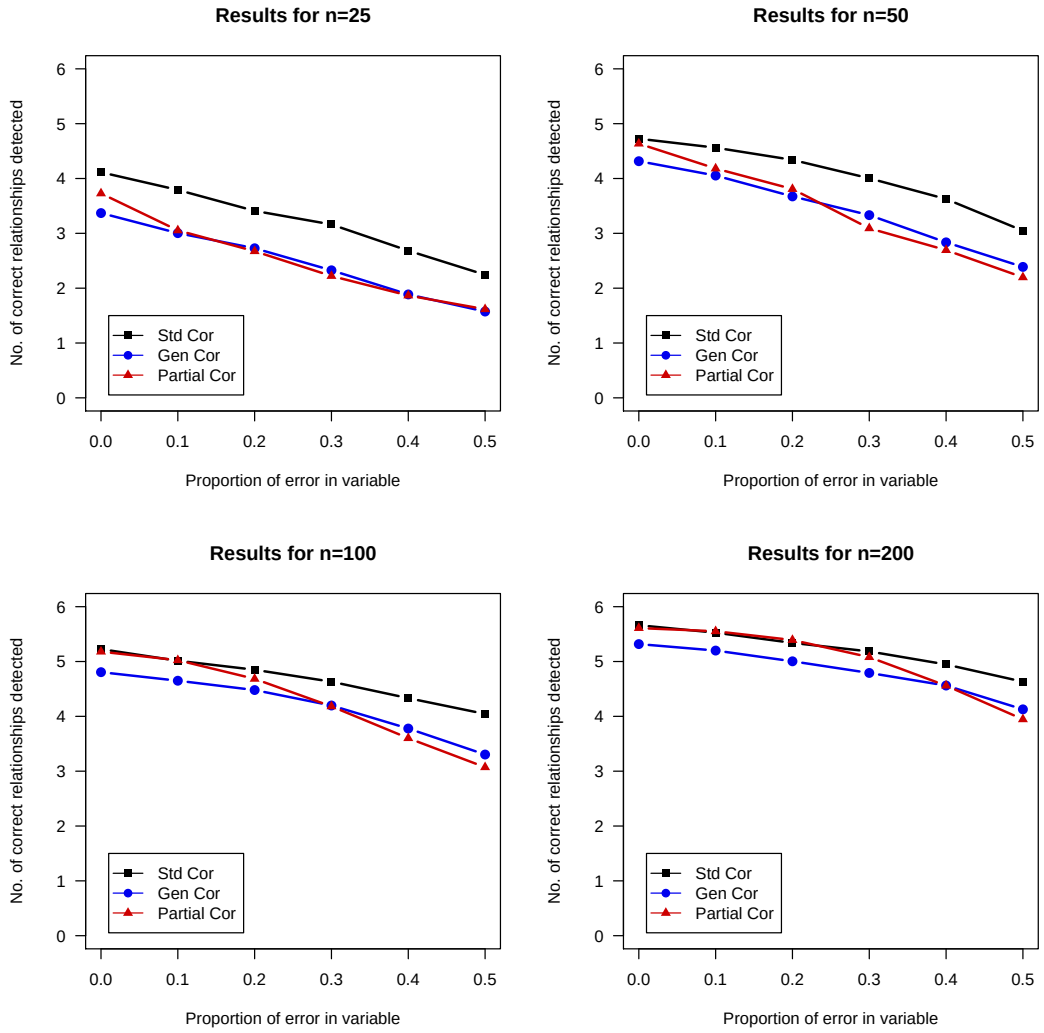


Figure 3.11: Comparison of relationship detection power for standard, generalised and partial correlations in the presence of errors in variables.

There are, however some drawbacks to using partial correlation to detect significant variable relationships, when compared to generalised correlation. The first relates to linearity. Partial correlation slightly outperforms generalised correlation in settings where the main relationships are linear (and where, therefore, standard correlation generally outperforms both methods). The explanation for this is clear – partial correlation is still, at its core, based on standard linear correlation, and does not "spend" information in the sample to estimate nonlinear behaviour. Conversely, when relationships are nonlinear then, as might be expected, generalised correlation performs much better than either standard or partial correlation.

A second issue is that of errors in variables. Even in linear cases, partial cor-

relation is rather seriously affected by errors in variables, as the following small theoretical example demonstrates. Suppose that we have three observed variables $X^{(1)}, X^{(2)}, X^{(3)}$ and that a "true" underlying variable $W^{(1)}$ exists such that:

- $X^{(1)}$ is an attempt to measure $W^{(1)}$, but may have some error in it. Thus let $\text{cor}(W^{(1)}, X^{(1)}) = \rho_1$.
- $W^{(1)}$ controls $X^{(2)}$ and $X^{(3)}$, that is $X^{(j)} = W^{(1)} + \epsilon_j$ for $j = 2, 3$ and some error ϵ_j. For simplicity assume that $\text{cor}(W^{(1)}, X^{(2)}) = \text{cor}(W^{(1)}, X^{(3)}) = \rho_2$.

If there is no error-in-variable for $X^{(1)}$ then $\rho_1 = 1$ and the theoretical partial correlation between $X^{(2)}$ and $X^{(3)}$ with respect to $X^{(1)}$ is zero ($\rho_{23\cdot 1} = 0$). This gives the desired situation when the non-zero partial correlations relate each of $X^{(2)}$ and $X^{(3)}$ to $X^{(1)}$, but $X^{(2)}$ and $X^{(3)}$ are not themselves related. However, if $\rho_1 < 1$ then

$$\rho_{23\cdot 1} = 1 - \frac{1 - \rho_2^2}{1 - \rho_1^2 \rho_2^2},$$

which implies a non-zero partial correlation between $X^{(2)}$ and $X^{(3)}$ is detected. Observe that this value does not disappear as $n \to \infty$; it is intrinsic to the problem.

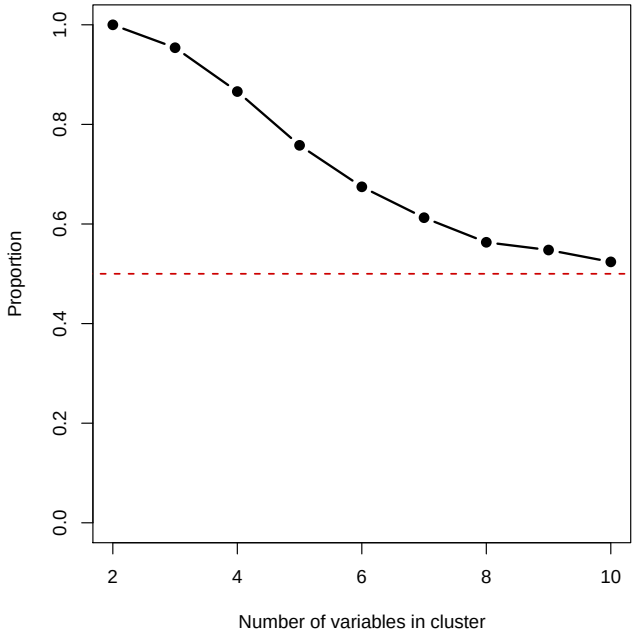


Figure 3.12: Proportion of partial correlations above average random noise level.

In small samples, and even when the extent of errors is as small as 10% and the relationships are linear, the partial correlation deteriorates to the extent that it loses all of its potential advantages over generalised correlation. To show this, a simulated dataset containing ten variables and six (nonzero) linear relationships of strengths

0.9, 0.6, 0.6, 0.5, 0.3 and 0.2 was created. Figure 3.11 shows the average number of these six relationships detected by the three correlation methods for various sample sizes and error in variables. Since the relationships are linear, generalised correlation would be expected to be the weakest method, but it is clear from the graphs that performance of partial correlation degrades quickly with the presence of errors.

A third issue relates to the existence of "clusters" of strongly related variables in the data. Suppose that there are d variables, all with pairwise (standard) correlation ρ. This is a situation where there is no controlling variable, either because it is hidden or because the variables are equally controlling. In this case it is possible to show that the pairwise partial correlation is $\rho/(1+d\rho)$. This means that if d is larger than the range three to five, approximately, the relationships between these variables will be obscured when partial correlations are taken, and noise may prevent some of these genuine relationships from being seen. Figure 3.12 demonstrates this behaviour; it shows, for $\rho = 0.8$ and various d, the proportion of partial correlation relationships that are above the average "noise" level. Sample size was 50. It is clear in this example that as the cluster size approaches 10, the size of a partial correlation is almost indistinguishable from random noise. This effect is detrimental, particularly when clusters of this type are of significance.

Chapter 4

Local regression and variable selection

4.1 Background

This chapter is designed to address the problem of building a final regression model, as one might do after an initial dimension reduction along the lines of Chapter 2. The classical regression problem is concerned with predicting a noisy continuous response using a d-dimensional predictor vector with support on some d-dimensional subspace. This functional relationship is often taken to be smooth and methods for estimating it range from parametric models, which specify the form of the relationship between predictors and response, through to nonparametric models, which have fewer prior assumptions about the shape of the fit. An important consideration for fitting such a regression model, particularly if the d variables are a subset of a much larger set, is whether all d predictors are in fact necessary. If a particular predictor has no relationship to the response, the model will be made both simpler and more accurate by removing it. This of course is reason for recent interest in sparse models, as discuss earlier. Most attention has been given to models with parametric forms, and in particular the linear model of Section 1.7, where the response is assumed to vary linearly with the predictors. However, there has also been some investigation into variable selection for nonlinear models, notably through the use of smoothing splines and local regression.

One common feature of the existing sparse methods is that the variable selection is "global" in nature, attempting to universally include or exclude a predictor. Such an approach does not naturally reconcile well with some nonparametric techniques, such as local polynomial regression, which focus on a "local" subset of the data to

estimate the response. In this local context it would be more helpful to understand local variable influence, since predictors that are irrelevant in some regions may in fact be important elsewhere in the subspace. Just as in the global setting, such information would allow us to improve the accuracy and parsimony of a model, but at a local level.

However, this approach to variable selection can be problematic. Most notably, variable significance affects the definition of "local". To illustrate concretely, suppose that two data points are close in every dimension except one. In typical local regression these points would not be considered close, and so the response at one point would not impact the other. If, however, we establish that the one predictor they differ by is not influential over a range that includes both these points, then they should actually be treated as neighbouring, and be treated as such in the model. Any methodology seeking to incorporate local variable influence needs to accommodate such potential situations.

Understanding local variable significance can also give additional insight into a dataset. If a variable is not important in certain regions of the support, knowledge of this allows us to discount it in certain circumstances, simplifying our understanding of the problem. For example, if none of the variables are relevant in a region, we may treat the response as locally constant and so know that we can ignore predictor effects when an observation lies in this region.

A final consideration is theoretical performance. In particular we shall present an approach that is "oracle"; that is, its performance is comparable to that of a particularly well-informed statistician, who has been provided in advance with the correct variables. It is interesting to note that variable interactions often cause sparse parametric approaches to fail to be oracle, but in the local nonparametric setting this is not an issue, because such interactions vanish as the neighbourhood of consideration shrinks.

In this chapter we propose a flexible and adaptive approach to local variable selection using local polynomial regression. The key technique is careful adjustment of the local regression bandwidths to allow for variable redundancy. The method has been named LABAVS, standing for "locally adaptive bandwidth and variable selection". Section 4.2 will introduce the LABAVS algorithm, including a motivating example and possible variations. Section 4.3 will deal with theoretical properties and in particular presents a result showing that the performance of LABAVS is better than oracle when the dimension remains fixed. Section 4.4 presents numerical results for both real and simulated data, showing that the algorithm can improve prediction accuracy and is also a useful tool in arriving at an intuitive understanding of the data. Technical details have been relegated to Section 4.5.

LABAVS is perhaps best viewed as an improvement to local polynomial regres-

sion, and will retain some of the advantages and disadvantages associated with this approach. In particular, it still suffers the "curse of dimensionality," in that it struggles to detect local patterns when the dimension of genuine variables increases beyond a few. It is not the first attempt at incorporating variable selection into local polynomial regression; the papers by Lafferty and Wasserman (2008) and Bertin and Lecué (2008) also do this. We compare our approach to these in some detail in Section 4.2.6. LABAVS can also be compared to other nonparametric techniques in use for low to moderate dimensions. These include generalised additive models, MARS and tree based methods (see Hastie et al., 2001).

The earliest work on local polynomial regression dates back to that of Nadaraya (1964) and Watson (1964). General references on the subject include Wand and Jones (1995), Simonoff (1996) and Loader (1999). An adaptive approach to bandwidth selection may be found in Fan and Gijbels (1995), although this was not in the context of variable selection. Tibshirani (1996) studies the LASSO, one of the most popular sparse solutions for the linear model; more recent related work on the linear model includes that of Candes and Tao (2007) and Bickel et al. (2009). Zou (2006) created the adaptive version of the LASSO and proved oracle performance for it. Lin and Zhang (2006) and Yuan and Lin (2006) have investigated sparse solutions to smoothing spline models. The work of Tropp (2004), Fuchs (2005), Zhao and Yu (2007), Meinshausen et al. (2007), Meinshausen and Yu (2009) and Wasserman and Roeder (2009) is also relevant here.

The LABAVS algorithm also bears some similarity to the approach adopted by Hall et al. (2004). There the aim was to estimate the conditional density of a response using the predictors. Cross-validation was employed and the bandwidths in irrelevant dimensions diverged, thereby greatly downweighting those components. In the present work the focus is more explicitly on variable selection, as well as attempting to capture local variable dependencies.

4.2 Methodology

4.2.1 Model and definitions. Suppose that we have a continuous response Y_i and a d-dimensional random predictor vector $X_i = (X_{i1}, \ldots, X_{id})$ which has support on some subspace $\mathcal{C} \subset \mathbb{R}^d$. Further, assume that the observation pairs (Y_i, X_i) are independent and identically distributed for $i = 1, \ldots, n$, and that X_i has density function f. The response is related to the predictors through a function g,

$$Y_i = g(X_i) + \epsilon_i \,, \tag{4.1}$$

with the error ϵ_i having zero mean and fixed variance. Smoothness conditions for f and g will be discussed in the theory section.

Local polynomial regression makes use of a kernel and bandwidth to assign increased weight to neighbouring observations compared to those further away, which will often have zero weight. We take $K(u) = \prod_{1 \le j \le d} K^*(u^{(j)})$ to be the d-dimensional rectangular kernel formed from a one dimensional kernel K^* such as the tricubic kernel,

$$K^*(u^{(j)}) = (35/32)(1 - x^2)^3 \mathbb{I}(|x| < 1)\,.$$

Assume K^* is symmetric with support on $[-1, 1]$. For $d \times d$ bandwidth matrix H the kernel with bandwidth H, denoted K_H, is

$$K_H(u) = \frac{1}{|H|^{1/2}} K(H^{-1/2}u)\,. \tag{4.2}$$

We assume that the bandwidth matrices are diagonal, $H = \mathrm{diag}(h_1^2, \ldots, h_d^2)$, with each $h_j > 0$, and write $H(x)$ when H varies as a function of x. Asymmetric bandwidths can be defined as having both a lower and an upper (diagonal) bandwidth matrix, H^L and H^U respectively, for a given estimation point x, rather than a single bandwidth H for all x. The kernel weight of an observation X_i at estimation point x with asymmetrical bandwidth matrices $H^L(x)$ and $H^U(x)$, is

$$\begin{aligned} K_{H^U(x), H^L(x)}(X_i - x) \;=\; & \prod_{j\,:\,X_{ij} < x^{(j)}} \frac{1}{h_j^L(x)} K^*\left(\frac{X_{ij} - x^{(j)}}{h_j^L(x)}\right) \\ & \times \prod_{j\,:\,X_{ij} \ge x^{(j)}} \frac{1}{h_j^U(x)} K^*\left(\frac{X_{ij} - x^{(j)}}{h_j^U(x)}\right)\,. \end{aligned} \tag{4.3}$$

This amounts to having (possibly) different window sizes above and below x in each direction. Although such unbalanced bandwidths would often lead to undesirable bias properties in local regression, here they will be used principally to extend bandwidths in dimensions considered redundant, so this issue is not a concern.

We also allow the possibility of infinite bandwidths $h_j = \infty$. In calculating the kernel in (4.2) when h_j is infinite, proceed as if the jth dimension did not exist (or equivalently, as if the jth factor in rectangular kernel product is always equal to 1). If all bandwidths are infinite, consider the kernel weight to be 1 everywhere. Although the kernel and bandwidth conditions above have been defined fairly narrowly to promote simplicity in exposition, many of these assumptions are easily generalised.

Local polynomial regression estimates of the response at point x, $\hat{g}(x)$, are found by fitting a polynomial q to the observed data, using the kernel and bandwidth to weight observations. This is usually done by minimising the weighted sum of squares,

$$\sum_{i=1}^{n} \{Y_i - q(X_i - x)\}^2 K_H(X_i - x)\,. \tag{4.4}$$

Once the minimisation has been performed, $q(0)$ becomes the point estimate for $g(x)$. The polynomial is of some fixed degree p, with larger values of p generally decreasing bias at the cost of increased variance. Of particular interest in the theoretical section will be the local linear fit, which minimises

$$\sum_{i=1}^{n}\left[\left\{Y_i-\gamma_0-\sum_{j=1}^{d}(X_{ij}-x^{(j)})\gamma_j\right\}^2 K_H(X_i-x)\right], \tag{4.5}$$

over γ_0 and $\gamma=(\gamma_1,\ldots,\gamma_d)$.

4.2.2 The LABAVS Algorithm. Below is the LABAVS algorithm that will perform local variable selection and vary the bandwidths accordingly. The choice of H in the first step can be local or global and should be selected as for a traditional polynomial regression, using cross-validation, a plug-in estimator or some other standard technique. Methods for assessing variable significance in Step 2, and the degree of shrinkage needed in Step 4, are discussed below.

LABAVS Algorithm

1. Find a starting $d\times d$ bandwidth $H=\mathrm{diag}(h^2,\ldots,h^2)$.
2. For each point x of a representative grid in the data support, perform local variable selection to determine disjoint index sets $\hat{\mathcal{A}}^+(x),\hat{\mathcal{A}}^-(x)$, with $\hat{\mathcal{A}}^+(x)\cup\hat{\mathcal{A}}^-(x)=\{1,\ldots,d\}$, for variables that are considered relevant and redundant respectively.
3. For any given x, derive new local bandwidth matrices $H^L(x)$ and $H^U(x)$ by extending the bandwidth in each dimension indexed in $\hat{\mathcal{A}}^-(x)$. The resulting space given nonzero weight by the kernel $K_{H^L(x),H^U(x)}(u-x)$ is the rectangle of maximal area with all grid points x_0 inside the region satisfying $\hat{\mathcal{A}}^+(x_0)\subset\hat{\mathcal{A}}^+(x)$. Here $\hat{\mathcal{A}}^+(x)$ is calculated explicitly as in Step 2, or taken as the set corresponding the closest grid point to x.
4. Shrink the bandwidth slightly for those variables in $\hat{\mathcal{A}}^+(x)$ according to the amount that bandwidths have increased in the other variables. See Section 4.2.4 for details.
5. Compute the local polynomial estimator at x, excluding variables in $\hat{\mathcal{A}}^-(x)$ and using adjusted asymmetrical bandwidths $H^L(x)$ and $H^U(x)$. The expression to be minimised is

$$\sum_{i=1}^{n}\left\{Y_i-q(X_i-x)\right\}^2 K_{H^L(x),H^U(x)}(X_i-x),$$

where the minimisation runs over all polynomials q of appropriate degree. The value of $q(0)$ in the minimisation is the final local linear estimator.

We refer to a rectangle in Step 3 of the algorithm since we are using a product kernel bandwidth, which has nonzero support on a rectangle. The key feature of the algorithm is that variable selection directly affects the bandwidth, increasing it in the direction of variables that have no influence on the point estimator. If a variable has no influence anywhere, it has the potential to be completely removed from the local regression, reducing the dimension of the problem. For variables that have no influence in certain areas, the algorithm achieves a partial dimension reduction. The increased bandwidths reduce the variance of the estimate and Step 4 swaps some of this reduction for a decrease in the bias to further improve the overall estimator.

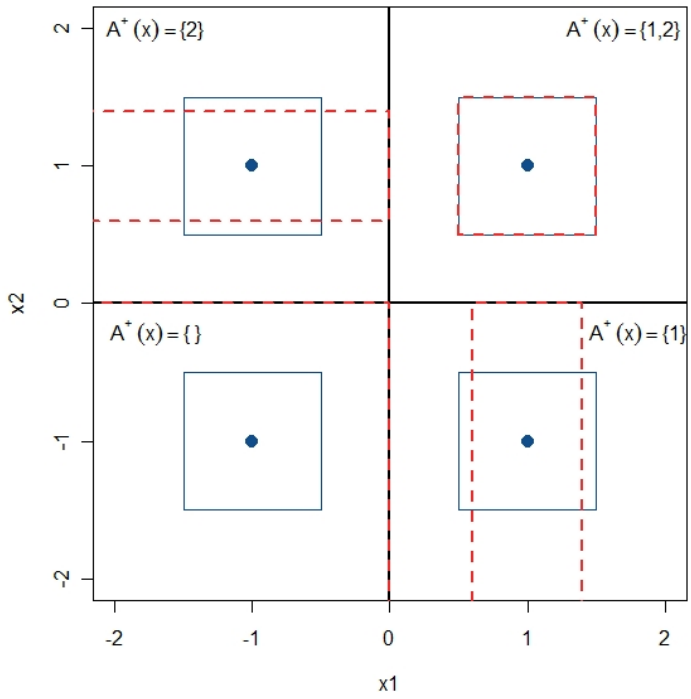


Figure 4.1: Bandwidth adjustments under ideal circumstances in illustrative example.

As a concrete example of the approach, define the following one-dimensional "Huberised" linear function:

$$g(x) = x^2\mathbb{I}(0 < x \leq 0.4) + (0.8x - 0.16)\mathbb{I}(x > 0.4)\,, \tag{4.6}$$

and let $g(X) = g\{([X^{(1)}]_+^2 + [X^{(2)}]_+^2)^{1/2}\}$ for 2-dimensional random variable $X = (X^{(1)}, X^{(2)})$. Assume that X is uniformly distributed on the space $[-2, 2] \times [-2, 2]$.

Notice that when $X^{(1)}$, $X^{(2)} < 0$ the response variable Y in (4.1) is independent of $X^{(1)}$ and $X^{(2)}$; when $X^{(1)} < 0$ and $X^{(2)} > 0$ the response depends on $X^{(2)}$ only; when $X^{(1)} > 0$ and $X^{(2)} < 0$ the response depends on $X^{(1)}$ only; when $X^{(1)}$, $X^{(2)} > 0$ the response depends on both $X^{(1)}$ and $X^{(2)}$. Thus in each of these quadrants a different subset of the predictors is significant. A local approach to variable significance can capture these different dependencies, while a global variable redundancy test would not eliminate any variables.

Now consider how the algorithm applies to this example, starting with a uniform initial bandwidth of $h = 0.5$ in both dimensions. Assuming that variable significance is estimated perfectly on a dense grid, figure 4.1 illustrates the adjusted bandwidths for each of the quadrants. The dots are four sample estimation points, the surrounding unit squares indicate the initial bandwidths and the dashed lines indicate how the bandwidths are modified. In the bottom left quadrant both variables are considered redundant, and so the bandwidth expands to cover the entire quadrant. This is optimal behaviour, since the true function is constant over this region, implying that the best estimator will be produced by including the whole area. In the bottom right quadrant the first dimension is significant while the second is not. Thus the bandwidth for the second dimension is "stretched", while the first is shrunken somewhat. Again, this is desirable for improving the estimator. The stretching in the second dimension improves the estimator by reducing the variance as more points are considered. Then the shrunken first dimension swaps some of this reduction in variance for decreased bias. Finally, in the top right quadrant, there is no change in the bandwidth since both variables are considered to be significant.

4.2.3 Variable selection step. Below are three possible ways to effect variable selection at x in Step 2 of the algorithm, presented in the context of local linear regression. They all make use of a tuning parameter λ which controls how aggressive the model is in declaring variables as irrelevant. Cross validation can be used to select an appropriate level for λ. So that the tuning parameters are comparable at different points in the data domain, it is useful to consider a local standardisation of the data at x. Define $\bar{X}_x = (\bar{X}_x^{(1)}, \ldots, \bar{X}_x^{(d)})$ and $\bar{Y}_x$ by

$$\bar{X}_x^{(j)} = \frac{\sum_{i=1}^n X_{ij} K_H(X_i - x)}{\sum_{i=1}^n K_H(X_i - x)}, \quad \bar{Y}_x = \frac{\sum_{i=1}^n Y_i K_H(X_i - x)}{\sum_{i=1}^n K_H(X_i - x)},$$

and define $\tilde{X}_i = (\tilde{X}_{i1}, \ldots, \tilde{X}_{id})$ and $\tilde{Y}_i$ by

$$\begin{aligned} \tilde{X}_{ij} &= \frac{(X_{ij} - \bar{X}_x^{(j)})\{K_H(X_i - x)\}^{1/2}}{\left\{\sum_{i=1}^n \left(X_{ij} - \bar{X}_x^{(j)}\right)^2 K_H(X_i - x)\right\}^{1/2}}, \\ \tilde{Y}_i &= (Y_i - \bar{Y}_x)\{K_H(X_i - x)\}^{1/2}. \end{aligned} \tag{4.7}$$

Notice that $\tilde{X}$ and $\tilde{Y}$ incorporate the weight $K_H(X_i - x)$ into the expression.

1. **Hard thresholding:** Choose parameters to minimise the weighted least squares expression,
$$\sum_{i=1}^{n} \left\{ \tilde{Y}_i - \beta_0 - \sum_{j=1}^{d} \tilde{X}_{ij}\beta_j \right\}^2 , \tag{4.8}$$
and classify as redundant those variables for which $|\hat{\beta}_j| < \lambda$. This can be extended to higher degree polynomials, although performance tends to be more unstable.

2. **Backwards stepwise approach:** For each individual j, calculate the percentage increase in the sum of squares if the jth variable is excluded from the local fit. Explicitly, if if $\hat{q}$ is the optimal local fit using all variables and $\hat{q}_j$ is the fit using all except the jth, we classify the jth variable as redundant if
$$\frac{\sum_{i=1}^{n} \{Y_i - \hat{q}_j(X_i - x)\}^2 K_H(X_i - x) - \sum_{i=1}^{n} \{Y_i - \hat{q}(X_i - x)\}^2 K_H(X_i - x)}{\sum_{i=1}^{n} \{Y_i - \hat{q}(X_i - x)\}^2 K_H(X_i - x)} < \lambda . \tag{4.9}$$
This approach is so named as it is analogous to the first step of a backwards stepwise procedure.

3. **Local lasso:** Minimise the expression
$$\sum_{i=1}^{n} \left\{ \tilde{Y}_i - \gamma_0 - \sum_{j=1}^{d} \tilde{X}_{ij}\gamma_j \right\}^2 + \lambda \sum_{j=1}^{d} |\gamma_j| . \tag{4.10}$$
Those variables for which γ_j are set to zero in this minimisation are then classified as redundant. While the normal lasso can have consistency problems (Zou, 2006), this local version does not since variables are asymptotically independent as $h \to 0$. The approach also scales naturally to higher order polynomials, provided all polynomial terms are locally standardised; a variable is considered redundant if all terms that include it have corresponding parameters set to zero by the lasso.

We have found that the first and second of the above approaches have produced the most compelling numerical results. The numerical work in Section 4.4 uses the first approach for linear polynomials, while the theoretical work in Section 4.3 establishes uniform consistency for both of the first two methods, guaranteeing oracle performance.

4.2.4 Variable shrinkage step. The variable shrinkage step depends on whether the initial bandwidth, and thus the shrunken bandwidth h', is chosen locally or

globally. Define

$$V[x,H] = \frac{\sum\{K_H(X_i - x)\}^2}{\{\sum K_H(X_i - x)\}^2}\,, \tag{4.11}$$

where the bandwidth term in the function V is allowed to be asymmetrical, in which case we write as $V[x,\{H^L(x),H^U(x)\}]$. Thus H has been replaced by the asymmetrical bandwidth $\{H^L(x),H^U(x)\}$, with H^L and H^U denoting the lower and upper bandwidths respectively. Then in the local case, letting $d'(x)$ denote the cardinality of $\hat{\mathcal{A}}(x)$, let

$$M(x) = V[x,\{H^L(x),H^L(x)\}]/V[x,H]\,. \tag{4.12}$$

The expression is asymptotically proportional to $\{h'(x)\}^{-d'(x)}$ and estimates the degree of variance stabilisation resulting from the bandwidth adjustment. Using this, the correct amount of bandwidth needed in step 4 is $h'(x) = h\{M(x)d'(x)/d\}^{1/4}$. Since both sides of this expression depend on $h'(x)$, shrinkage can be approximated in the following way. Let

$$M^*(x) = V[x,\{\tilde{H}^L(x),\tilde{H}^L(x)\}]/V[x,H]\,,$$

where $\tilde{H}^L(x)$ and $\tilde{H}^U(x)$ are the bandwidth matrices immediately after step 3. Then the shrunken bandwidths are $h'(x) = h\{M^*(x)d'(x)/d\}^{1/(d'(x)+4)}$.

In the global bandwidth case, we define

$$M[\{H^L(X),H^U(X)\},H] = \frac{E\left(V[X,\{H^L(X),H^L(X)\}]\right)}{E\{V[X,H]\}}\,. \tag{4.13}$$

This expression measures the average variance stabilisation across the domain. In this case, the shrinkage factor should satisfy

$$h' = h\left(M[\{H^L(X),H^U(X)\},H]E\{d'(X)\}/d\right)^{1/4}\,. \tag{4.14}$$

The theoretical properties in Section 4.3 deal with the global bandwidth scenario. The treatment for the local case is similar, except that care must be taken in regions of the domain where the function g behaves in a way that is exactly estimable by a local polynomial and thus has potentially no bias.

4.2.5 Further remarks.

1. The choice of distance between grid points in Step 2 is somewhat arbitrary, but should be taken as less than h so that all data points are considered in calculations. In the asymptotic theory we let this length decrease faster than the rate of the bandwidth, and in numerical experimentation the choice impacts only slightly on the results.

2. Step 5 of the algorithm forces the estimate at point x to exclude variables indexed in $\hat{\mathcal{A}}^{-}(x)$. An alternative is to still use all variables in the final fit. This may be advantageous in situations with significant noise, where variable admission and omission is more likely to have errors. Despite including these extra variables, the adjusted bandwidths still ensure that estimation accuracy is increased.

3. Finding the maximal rectangle for each representative point, as suggested in step 3 of the algorithm, can be a fairly intensive computational task. In our numerical work we simplified this by expanding the rectangle equally until the boundary met a "bad" grid point (i.e. a point x' such that $\hat{\mathcal{A}}^{+}(x') \not\subseteq \hat{\mathcal{A}}^{+}(x)$). The corresponding direction was then held constant while the others continue to increase uniformly. We continued until each dimension stopped expanding or grew to be infinite. This approach does not invalidate the asymptotic results in Section 4.3, but there may be some deterioration in numerical performance associated with this simplification.

4. If a variable is redundant everywhere, results in Section 4.3 demonstrate that the algorithm is consistent; the probability that the variable is classified as redundant everywhere tends to 1 as n grows. However, the exact probability is not easy to calculate and for fixed n we may want greater control over the ability to exclude a variable completely. In such circumstances a global variable selection approach may be appropriate.

5. As noted at the start of Section 4.2.2, the initial bandwidth in Step 1 does not necessarily have to be fixed over the domain. For instance, a nearest neighbour bandwidth, where h at x is roughly proportional to $f(x)^{-1}$, could be used. Employing this approach offers many practical advantages and the theoretical basis is similar to that for the constant bandwidth. The numerical work makes use of nearest neighbour bandwidths throughout. In addition, we could use an initial bandwidth that was allowed to vary for each variable, $H = \mathrm{diag}(h_1^2, \ldots, h_p^2)$. So long as, asymptotically, each h_j was equal to $C_j h$ for some controlling bandwidth h and constant C_j, the theory would hold, although details are not pursued here.

4.2.6 Comparison to other local variable selection approaches. As mentioned in the introduction, two recent papers take a similar approach to this problem. Firstly Lafferty and Wasserman (2008) introduce the rodeo procedure. This attempts to assign adaptive bandwidths based on the derivative with respect to the bandwidth for each dimension, $\partial \hat{g}(x)/\partial h_j$. This has the attractive feature of bypassing the actual local shape and instead focussing on whether an estimate is improved

by shrinking the bandwidths. It is also a greedy approach, starting with large bandwidths in each direction and shrinking only those that cause a change in the estimator at a point. The second paper is by Bertin and Lecué (2008), who implement a two step procedure to reduce the dimensionality of a local estimate. The first step fits a local linear estimate with an L_1 or lasso type penalty, which identifies the relevant variables. This is followed by a second local linear fit using this reduced dimensionality. The lasso penalty they use is precisely the same as the third approach suggested in Section 4.2.3.

We comment on the similarities and differences of these two approaches compared to the current presentation, which are summarised in Table 4.1. Firstly the theoretical framework of the two other papers focus exclusively on the performance at a single point, while the LABAVS approach ensures uniformly oracle performance on the whole domain (although uniformly oracle performance may be provable for other approaches). The framework for the other two also assumes that variables are either active on the whole domain or redundant everywhere, while we have already discussed the usefulness of an approach that can adapt to variables that are redundant on various parts of the data. We believe this is particularly important, since local tests of variable significance will give the same results everywhere. Related to this, our method does not require an assumption of nonzero gradients (whether with respect to the bandwidth or variables) to obtain adequate theoretical performance, in contrast to the other methods. On the other hand, ensuring uniform performance while allowing d to be increasing is quite challenging, so our presentation assumes d is fixed, in contrast to other treatments. It is also worth noting that the greedy approach of Lafferty and Wasserman potentially gives it an advantage in higher dimensional situations.

While all approaches work in a similar framework, the above discussion demonstrates that there are significant differences. Our methodology may be viewed as a generalisation of the work of Bertin and Lecué, save for imposing fixed dimensionality. It can also be viewed as a competitor to the rodeo, and some numerical examples comparing the two are provided.

	LABAVS	Rodeo	Bertin and Lecué (2008)
Oracle performance on entire domain	✔	✘	✘
Allows for locally redundant variables	✔	✘	✘
Relevant variables allowed to have zero gradient	✔	✘	✘
Theory allows dimension d to increase with n	✘	✔	✔
Greedy algorithm applicable for higher dimensions	✘	✔	✘

Table 4.1: Summary of locally adaptive bandwidth approaches

With regards to computation time, for estimation at a single point the rodeo is substantially faster, since calculating variable significance on a large grid of points is not required. If however we need to make predictions at a reasonable number of points, then Labavs is likely to be more efficient, since the grid calculations need only be done once, while rodeo requires a new set of bandwidth calculations for each point.

4.3 Theoretical properties

As mentioned in the introduction, a useful means of establishing the power of a model that includes variable selection is to compare it with an oracle model, where the redundant variables are removed before the modelling is undertaken. In the linear (and the parametric) context, we interpret the oracle property as satisfying two conditions as $n \to \infty$:

1. the probability that the correct variables are selected converges to 1, and
2. the nonzero parameters are estimated at the same asymptotic rate as they would be if the correct variables were known in advance.

We wish to extend this notion of an oracle property to the nonparametric setting, where some predictors may be redundant. Here there are no parameters to estimate, so attention should instead be given to the error associated with estimating g. Below we define weak and strong forms of these oracle properties:

Definition 1. *The weak oracle property in nonparametric regression is:*

1. *the probability that the correct variables are selected converges to 1, and*
2. *at each point x the error of the estimator $\hat{g}(x)$ decreases at the same asymptotic rate as it would if the correct variables were known in advance.*

Definition 2. *The strong oracle property in nonparametric regression is:*

1. *the probability that the correct variables are selected converges to 1, and*
2. *at each point x the error of the estimator $\hat{g}(x)$ has the same first-order asymptotic properties as it would if the correct variables were known in advance.*

Observe that the weak oracle property achieves the correct rate of estimation while the strong version achieves both the correct rate and the same asymptotic distribution. The first definition is most analogous to its parametric counterpart, while the second is more ambitious in scope.

Here we establish the strong version of the nonparametric oracle property for the LABAVS algorithm, with technical details found in Section 4.5. We shall restrict

attention to the case of fixed dimension. In the case of increasing dimension, we could add an asymptotically consistent screening method, such as that proposed in Chapter 2, to reduce it back to fixed d. The treatment here focuses on local linear polynomials, partly for convenience but also recognising that the linear factors dominate higher order terms in the asymptotic local fit. Thus our initial fit is found by minimising the expression (4.5). We impose further conditions on the kernel K:

$$\begin{aligned}&\int K(z)dz = 1,\ \int z^{(j)}K(z)dz = 0 \text{ for each } j,\ \int z^{(j)}z^{(k)}K(z)dz = 0 \text{ when}\\ &j \neq k \text{ and } \int (z^{(j)})^2K(z)dz = \mu_2(K) > 0, \text{ with } \mu_2(K) \text{ independent of } j.\end{aligned} \tag{4.15}$$

The useful quantity $R(K)$, depending on the choice of kernel, is defined as

$$R(K) = \int K(z)^2dz = \left\{\int K^*(z^{(j)})^2dz^{(j)}\right\}^d,$$

where K^* is the univariate kernel introduced on page 51. Let $a_n \asymp b_n$ denote the property that $a_n = O(b_n)$ and $b_n = O(a_n)$. We also require the following conditions (4.16), needed to ensure uniform consistency of our estimators.

1. The support $\mathcal{C} = \{x : f(x) > 0\}$ of the random variable X is compact. Further, f and its first order partial derivatives are bounded and uniformly continuous on the interior of $\mathcal{C}$, and $\inf_{x\in\mathcal{C}} f(x) > 0$. In cases where this is not true of f, we choose $\mathcal{C}$ to be a subset of the support of f satisfying the desired properties.
2. The kernel function K is bounded with compact support and satisfies $|p(u)K(u) - p(v)K(v)| \leq C_1||u - v||$ for some $C_1 > 0$ and all points u, v in $\mathcal{C}$. Here $p(u)$ denotes a single polynomial term of the form $\prod(u^{(j)})^{a_j}$ with the nonnegative integers a_j satisfying $\sum a_j \leq 4$. The bound C_1 should hold for all such choices of p.
3. The function g has bounded and uniformly continuous partial derivatives up to and including order p, with $p \geq 2$. If $(D^k g)(x)$ denotes the partial derivative
$$\frac{\partial^{|k|}g(x)}{\partial(x^{(1)})^{k_1}\cdots\partial(x^{(d)})^{k_d}},$$
with $|k| = \sum k_j$, then we assume that these derivatives exist on the interior of $\mathcal{C}$ and satisfy, for some constant C_2,
$$|h(u) - h(v)| \leq C_2||u - v||.$$
4. $E(|Y|^\xi) < \infty$ for some $\xi > 2$.
5. The conditional density $f_{X|Y}(x|y)$ of X_i, conditional on Y, exists and is bounded.

(4.16)

6. $$\text{For some } 0 < \rho < 1, \ \frac{n^{1-2/\xi} h^d}{\log n \{\log n (\log \log n)^{1+\rho}\}^{2/\xi}} \to \infty \, .$$

7. The Hessian of g, $\mathcal{H}_g$, is nonzero on a set of nonzero measure in $\mathcal{C}$.

The conditions in (4.16), except perhaps the first, are fairly natural and not overly constrictive. For example, the sixth will occur naturally for any reasonable choice of h, while the second follows easily if K has a bounded derivative. The last condition is purely for convenience in the asymptotics; if $\mathcal{H}_g$ was zero almost everywhere then g would be linear and there would be no bias in the estimate, improving accuracy. The first condition will not apply if the densities trail off to zero, rather than experiencing a sharp cutoff at the boundaries of $\mathcal{C}$. However, in such circumstances our results apply to a subset of the entire domain, chosen so that the density did not fall below a specified minimum. Performance inside this region would then conform to the optimal accuracies presented, while estimation outside this region would be poorer. This distinction is unavoidable, since estimation in the tails is usually problematic and it would be unusual to guarantee uniformly good performance there.

Step 1 of the LABAVS Algorithm allows the initial bandwidth to be chosen globally or locally. Here we shall focus on the global case, where an initial bandwidth $H = \text{diag}(h^2, \ldots, h^2)$ is used. Further, we assume that this h is chosen to minimise the mean integrated squared error (MISE):

$$E\left[\int \{\hat{g}(x) - g(x)\}^2 f(x) dx\right] ,$$

where the outer expectation runs over the estimator $\hat{g}$. It is possible to show that under our assumptions that

$$h = \left[\frac{d\sigma^2 R(K) A_{\mathcal{C}}}{n\mu_2(K)^2 A_{\mathcal{H}_g}}\right]^{-1/(d+4)} , \tag{4.17}$$

where $A_{\mathcal{C}}$ and $A_{\mathcal{H}_g}$ are constants, defined in the Appendix, depending only on $\mathcal{C}$ and the function g respectively. Notice in particular that $h \asymp n^{1/(d+4)}$. Details are given in Lemma 4.5 in Section 4.5.

A key result in establishing good performance, in Theorem 4.1 below, is uniform consistency of the local polynomial parameter estimates. It is a simplified version of a result by Masry (1996), and no proof is included.

Theorem 4.1. *Suppose the conditions in (4.16) hold and we use parameter estimates from a degree p polynomial regression to estimate the partial derivatives of g. Then for each k with $0 \le |k| \le p$ we have*

$$\sup_{x \in \mathcal{C}} |(\widehat{D^k g})(x) - (D^k g)(x)| = O\left[\left(\frac{\log n}{n h^{d+2|k|}}\right)^{1/2}\right] + O(h^{p-|k|+1}) \text{ almost surely.}$$

Since the partial derivative estimate at x is proportional to the corresponding local polynomial coefficient, Theorem 1 ensures that the local polynomial coefficients are consistently estimated uniformly for suitable h. The scaling applied in (4.7) does not impact on this, as the proof of Theorem 4.2 demonstrates.

Let $\mathcal{C}^-$ denote the points $x \in \mathcal{C}$ satisfying $\partial g(x)/\partial x^{(j)} = 0$ and $\partial^2 g(x)/\partial x^{(j)2} \neq 0$. That is, $\mathcal{C}^-$ denotes the points where the true set of relevant variables changes. Notice that in the illustrative example in Section 4.2.2 we had $\mathcal{C}^- = \{x \,|\, x^{(1)} = 0\,, x^{(2)} = 0\}$. The smoothness assumed of g implies that $\mathcal{C}^-$ has Lebesgue measure 0. Let $\delta > 0$ and let $\mathcal{O}_\delta$ be the smallest open set containing $\mathcal{C}^-$ such that

$$\inf_{x \in \mathcal{C} \setminus \mathcal{O}_\delta,\, j \in \mathcal{A}^+(x)} |\partial g(x)/\partial x^{(j)}| = \delta\,. \tag{4.18}$$

Intuitively this means that on the set $\mathcal{C} \setminus \mathcal{O}_\delta$ the relevant variables have the absolute value of their corresponding derivatives $|\partial g(x)/\partial x^{(j)}|$ bounded below by $\delta > 0$, while irrelevant variables have $\partial g(x)/\partial x^{(j)} = 0$. Thus we have a "gap" between the true and irrelevant variables in this region that we may exploit. The volume of $\mathcal{O}_\delta$ may be made arbitrarily small by choosing δ small. Call the set $\hat{\mathcal{A}}^+(x)$ in the algorithm correct if the variables in it are the same as the set of variables j with $\partial g(x)/\partial x^{(j)} \neq 0$. Denote the latter correct set by $\mathcal{A}^+(x)$.

Theorem 4.2. *Suppose δ is given, h is chosen to minimise squared error as in (4.17), $\hat{\mathcal{A}}^+(x)$ is formed using the first approach in Section 4.2.3, and λ has a growth rate between arbitrary constant multiples of $h^2(n \log n)^{1/2}$ and $hn^{1/2}$. If f has bounded and uniformly continuous derivatives of degree 2, then the probability that $\hat{\mathcal{A}}^+(x)$ is correct on the whole set $\mathcal{C} \setminus \mathcal{O}_\delta$ tends to 1 as $n \to \infty$. That is,*

$$P(\hat{\mathcal{A}}^+(x) = \mathcal{A}^+(x) \text{ for all } x \in \mathcal{C} \setminus \mathcal{O}_\delta) \to 1 \text{ as } n \to \infty.$$

Furthermore, variables that are genuinely redundant everywhere will be correctly classified as such with probability tending to 1.

The property (4.18) ensures that the coefficients in the local linear fit are consistently estimated with error of order $O\{h(\log n)^{1/2}\}$. The adjustment in (4.7) means that the actual coefficients estimated are of order $hn^{1/2}$ times this, so the range of λ given is correct for separating true and redundant variables. The definition of $\mathcal{O}_\delta$ ensures that the classification is correct on $\mathcal{C} \setminus \mathcal{O}_\delta$, while variables that are redundant everywhere will be recognised as such.

The next result ensures consistency for the second approach in Section 4.2.3. We make one further assumption, concerning the error ϵ_i. Observe that this holds

trivially if ϵ_i is bounded. Assume that:

$$\text{there exists } C_3 \text{ such that } E(|\epsilon_i|^\alpha) \leq C_3^\alpha \text{ for } \alpha = 1, 2, 3, 4, \ldots. \tag{4.19}$$

Theorem 4.3. *Suppose δ is given, h is chosen to minimise squared error as in (4.17), and $\hat{\mathcal{A}}^+(x)$ is formed using the second approach in Section 4.2.3. Provided that $\lambda = o(h^2)$ and $h^4 \log n = o(\lambda)$, the probability that $\hat{\mathcal{A}}^+(x)$ is correct on $\mathcal{C} \setminus \mathcal{O}_\delta$ tends to 1 as $n \to \infty$. Furthermore, variables that are genuinely redundant everywhere will be correctly classified as such with probability tending to 1.*

The previous two results ensure that we have consistent variable selection for the first two approaches in Section 4.2.3. Finally we can state and prove the strong oracle property for $\mathcal{C} \setminus \mathcal{O}_\delta$. Although the result does not cover the whole space $\mathcal{C}$, recall that we may make the area $\mathcal{O}_\delta$ arbitrarily small by decreasing δ. Furthermore, the proof implies that if we restricted attention to removing only those variables that are redundant everywhere, we would actually have the oracle property on the whole of $\mathcal{C}$; however we sacrifice this performance on $\mathcal{O}_\delta$ to improve the fit elsewhere by adjusting for locally redundant variables. In the following theorem the matrix $\ddot{H}$ is the diagonal bandwidth matrix with bandwidth ∞ for globally redundant variables and $\ddot{h}$ for the other variables, where

$$\ddot{h} = h \left(M(\ddot{H}, H) \ddot{d}/d \right)^{1/4} .$$

Here $\ddot{d}$ denotes the number of variables that are not globally redundant.

Theorem 4.4. *The estimates produced by the algorithm, where variable selection is performed using the first or second approach in Section 4.2.3, satisfy the strong definition of the nonparametric oracle property on $\mathcal{C}$. Further, when there are locally redundant variables, squared estimation error is actually less than the oracle performance by a factor of $M[\{H^L(X), H^U(X)\}, \ddot{H}] < 1$. That is,*

$$E\{\ddot{g}(x) - g(x)\}^2] = M[(H^L(X), H^U(X)), \ddot{H}] E\{\hat{g}(x) - g(x)\}^2] ,$$

where $\ddot{g}$ denotes the estimator arising from the LABAVS algorithm and $\hat{g}$ is the oracle estimator.

4.4 Numerical properties

The examples presented in this section compare the performance of two versions of the LABAVS algorithm with ordinary least squares, a traditional local linear fit, generalised additive models, tree-based gradient boosting and MARS. Table 4.2 describes the approaches used. The implementations of the latter four methods were

from the R packages *locfit*, *gam*, *gbm* and *polspline* respectively. Tuning parameters such as bandwidths for local methods, λ in LABAVS, number of trees in boosting, and MARS model complexity, were chosen to give best performance for each method. The LABAVS models used the first variable selection approach of Section 4.2.3. All the local methods used nearest neighbour bandwidths, with the initial bandwidth chosen each time so as to minimise cross-validated squared error. The OLS linear model was included as a standard benchmark, but obviously will fail to adequately detect nonlinear features of a dataset.

Name	Description
LABAVS-A	LABAVS with linear fit, all vars in final fit
LABAVS-B	LABAVS with linear fit, relevant vars only in final fit
LOC1	Local linear regression
OLS	Ordinary least squares linear regression
GBM	Boosting with trees, depth equal to three
GAM	Generalised additive models with splines
MARS	Multivariate adaptive regression splines

Table 4.2: Approaches included in computational comparisons

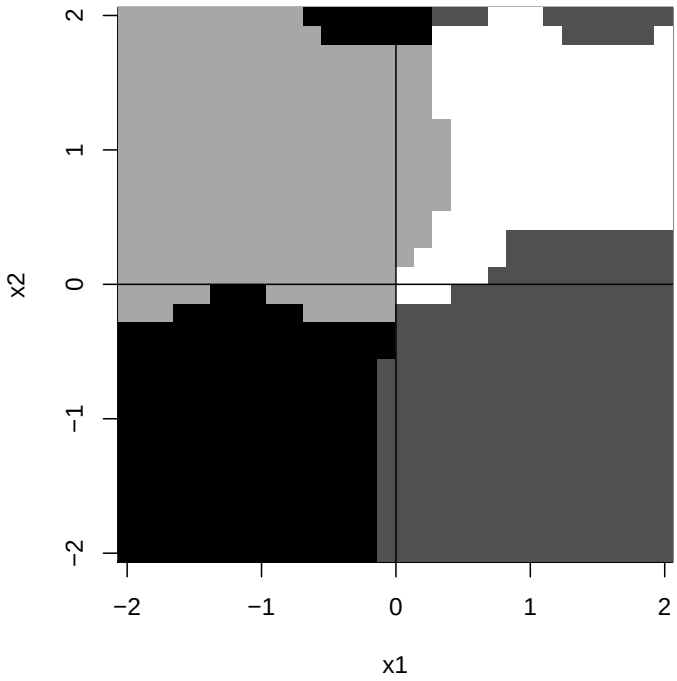


Figure 4.2: Plot of detected variable significance across subspace in Example 4.4.1.

4.4.1 Example: 2-dimensional simulation. The example introduced in Sec-

tion 4.2.2 was simulated with $n = 500$. The error for Y_i was normal with standard deviation 0.3. We first compare LABAVS to the rodeo and the methodology of Bertin and Lecué (2008) at the four representative points in Figure 4.1. Table 4.3 shows the mean squared error of the prediction compared to the true value over 100 simulations. In all cases parameters were chosen to minimise this average error. At all points the LABAVS approach performed strongest. The method of Bertin and Lecué (2008) performed poorly in situations where at least one variable is redundant; this is to be expected, since it excludes the variable completely and so will incorporate regions where it is actually important, causing significant bias. The rodeo also did not perform as well; we found it tended to overestimate the optimal bandwidths in redundant directions.

Test Point	LABAVS-A	LABAVS-B	rodeo	Bertin and Lecué
(1,1)	0.0022	0.0022	0.0065	0.0023
(1,-1)	0.0011	0.0013	0.0015	0.0018
(-1,1)	0.0009	0.0011	0.0015	0.0013
(-1,-1)	0.0006	0.0007	0.0008	0.0013

Table 4.3: Mean squared prediction error on sample points in Example 4.4.1

We then compared LABAVS with the other model approaches which are designed to make multiple predictions, rather than a specific point. For each simulation all the models were fitted and the average squared error was estimated using a separate test set of 500 observations. The simulation was run 100 times and the average error and its associated standard deviation for each model are recorded in Table 4.4.

Approach	Error	Std Dev
LABAVS-A	2.18	(0.71)
LABAVS-B	1.87	(0.65)
LOC1	2.31	(0.73)
OLS	42.85	(2.64)
GBM	2.47	(0.67)
GAM	5.93	(0.57)
MARS	2.35	(0.90)

Table 4.4: Mean squared error sum of test dataset in Example 4.4.1

Inspection of the results shows that the LABAVS models performed best, able to allow for the different dependencies on the variables. In particular the algorithm improved on the performance of the local linear model on which it is based. The local linear regression, the boosted model and MARS also performed reasonably, while GAM struggled with the nonadditive nature of the problem, and a strict linear model is clearly unsuitable here.

To show how effective variable selection is for LABAVS, Figure 4.2 graphically represents the sets $\hat{\mathcal{A}}^+$ at each grid point for one of the simulations, with the darkest representing $\{\}$, the next darkest $\{1\}$, the next darkest $\{2\}$ and finally the lightest $\{1, 2\}$. Here the variable selection has performed well; there is some encroachment of irrelevant variables into the wrong quadrants but the selection pattern is broadly correct. The encroachment is more prevalent near the boundaries since the bandwidths are slightly larger there, to cover the same number of neighbouring points.

4.4.2 Example: p-dimensional simulation. We next show that LABAVS can effectively remove redundant variables completely. Retain the setup of Example 4.4.1, except that we add $d^* = d - 2$ variables similarly distributed (uniform on $[-2, 2]$), which have no influence on the response. Also, keep the parameters relating to the LABAVS fit the same as the previous example, except that the cutoff for hard threshold variable selection, λ is permitted to vary. Table 4.4.2 shows the proportion of times from 500 simulations that LABAVS effected complete removal of the redundant dimensions, for various λ and p^*. Note that the cutoff level of 0.55 is that used in the previous example, and the two genuine variables were never completely removed in any of the simulations. The results suggest that to properly exclude redundant variables, a higher threshold is needed than would otherwise be the case. This causes the final model to be slightly underfitted when compared to the oracle model, but this effect is not too severe; Figure 4.3 shows how the variable significance plots change for a particular simulation with different values of the cutoff. It is clear that the patterns are still broadly correct, and the results still represent a significant improvement in traditional linear regression.

	Number of redundant dimensions			
λ	1	2	3	4
0.55	0.394	0.086	0.034	0.038
0.65	0.800	0.542	0.456	0.506
0.75	0.952	0.892	0.874	0.864
0.85	0.996	0.984	0.994	0.974
0.95	0.998	1.000	1.000	0.992

Table 4.5: Proportion of simulations where redundant variables completely removed by LABAVS

4.4.3 Example: ozone dataset. The first real data example used is the ozone dataset from Hastie et al. (2001), p.175. It is the same as the air dataset in S-PLUS, up to a cube root transformation in the response. The dataset contains meteorological measurements for New York collected from May to September 1973. There are 111 observations in the dataset, a fairly moderate size. Our aim here is to predict the

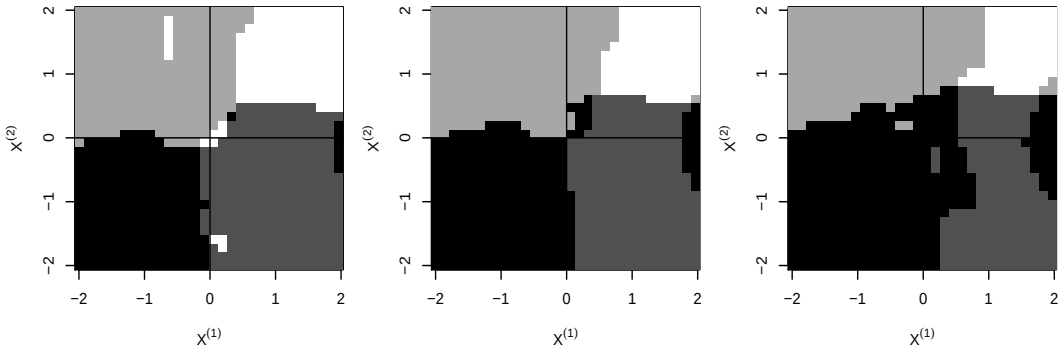


Figure 4.3: Plot of detected variable significance across subspace in Example 4.4.2, under various choices for λ.

ozone concentration using two of the other variables, temperature and wind, and scaled to unit variance when fitting the models. The smoothed perspective plot of the data in Figure 4.4 shows strong dependence on each of the two variables in some parts of the domain, but some sections appear flat in one or both directions in other parts. For example, the area surrounding a temperature of 70 and wind speed of 15 appears to be flat, implying that for reasonably low winds and high temperatures the ozone concentration is fairly stable. This suggests that LABAVS, by expanding the bandwidths here, could be potentially useful in reducing error. We performed a similar comparative analysis to that in Example 4.4.1, except that error rates were calculated using leave-one-out cross validation, where an estimate for each individual observations was computed after using all other observations to build the model. The resulting mean squared errors and corresponding standard deviations are presented in Table 4.6.

Approach	Error	Std Dev
LABAVS-A	277	(53)
LABAVS-B	284	(55)
LOC1	290	(55)
OLS	491	(110)
GBM	403	(118)
GAM	391	(98)
MARS	457	(115)

Table 4.6: Cross-validated mean squared error sum for the ozone dataset

The results suggest that the data is best modelled using local linear methods, and that LABAVS offers a noticeable improvement over a traditional local fit, due to its ability to improve the estimate in the presence of redundant variables. The perspective plot in the left panel of Figure 4.4 suggests a highly non-additive model,

which may explain why GAM performs poorly. There is also a large amount of local curvature, which hinders the OLS, GBM and MARS fits. The right panel of Figure 4.4 shows the variable selection results for the linear version of LABAVS across the data support, using the same shading as in Figure 4.1. We see that variable dependence is fairly complex, with all combinations of variables being significant in different regions. In particular, notice that the procedure has labelled both variables redundant in the region around $(70, 15)$, confirming our initial suspicions. This plot is also highly suggestive, revealing further interesting features. For instance, there is also little dependence on wind when temperatures are relatively high. Such observations are noteworthy and potentially useful.

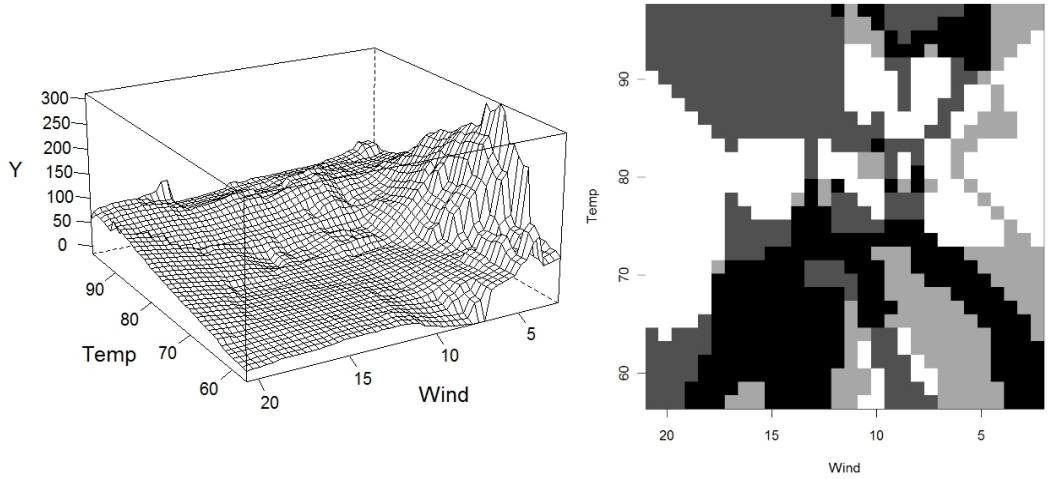


Figure 4.4: Ozone dataset smoothed perspective plot and variable selection plot.

4.4.4 Example: ethanol dataset. As a second low-dimensional real data example, we use the ethanol dataset which has been studied extensively, for example by Loader (1999). The response is the amount of a certain set of pollutants emitted by an engine, with two predictors: the compression ratio of the engine and the equivalence ratio of air to petrol. There are 88 observations, a fairly moderate size. Inspection of the data shows strong dependence on the equivalence ratio, but the case for the compression ratio is less clear. This suggests LABAVS could be potentially useful in reducing error. We performed a similar analysis to that in Example 4.4.3, with the results are presented in Table 4.7.

The results in Table 4.7 show that this problem is particularly suited to MARS, which performed the best. After MARS, LABAVS produced the next strongest result, again improving on the traditional local linear model. The GBM and GAM models were inferior to the local linear fit.

Approach	Error	Std Dev
LABAVS-A	0.075	(0.011)
LABAVS-B	0.085	(0.014)
LOC1	0.090	(0.012)
OLS	1.348	(0.128)
GBM	0.104	(0.020)
GAM	0.098	(0.012)
MARS	0.045	(0.008)

Table 4.7: Cross-validated mean squared error sum for the ethanol dataset

4.5 Technical arguments

We first prove the following lemma concerning the asymptotic behaviour of h.

Lemma 4.5. *The choice of h that minimises the mean integrated squared error is asymptotically the minimiser of*

$$(1/4)h^4\mu_2(K)^2 A_{\mathcal{H}_g} + \sigma^2(nh^d)^{-1}R(K)A_{\mathcal{C}}\,, \tag{4.20}$$

where $R(K) = \int K(x)^2dx$ for the function K, $A_{\mathcal{H}_g} = \int \mathrm{tr}\{\mathcal{H}_g(x)\}^2 f(x)dx$ and $A_{\mathcal{C}} = \int_{\mathcal{C}} 1dx$. Further,

$$h = \left[\frac{d\sigma^2 R(K)A_{\mathcal{C}}}{n\mu_2(K)^2 A_{\mathcal{H}_g}}\right]^{1/(d+4)}\,. \tag{4.21}$$

Proof: Ruppert and Wand (1994) show that for x in the interior of $\mathcal{C}$ we have bias and variance expressions

$$E\{\hat{g}(x) - g(x)\} = (1/2)\mu_2(K)h^2\,\mathrm{tr}\{\mathcal{H}_g(x)\} + o_P(h^2)\ \text{, and}$$

$$Var\{\hat{g}(x)\} = \{n^{-1}h^{-d}R(K)f(x)^{-1}\sigma^2\}\{1 + o_P(1)\}\,.$$

Substituting these into the mean integrated squared error expression yields

$$\begin{aligned}
MISE &= \int E\{\hat{g}(x) - g(x)\}^2 f(x)dx \\
&= \int \left[E\{\hat{g}(x)\} - g(x)\right]^2 + \mathrm{Var}\{\hat{g}(x)\}\big]\, f(x)dx \\
&= \int (1/4)\mu_2(K)^2h^4\,\mathrm{tr}\{\mathcal{H}_g(x)\}^2 f(x)dx + o_P(h^4) \\
&+ \int n^{-1}h^{-d}R(K)f(x)^{-1}\sigma^2 f(x)dx + o_P(n^{-1}h^{-d}) \\
&= (1/4)h^4\mu_2(K)^2 A_{\mathcal{H}_g} + \sigma^2(nh^d)^{-1}R(K)A_{\mathcal{C}} + o_P(h^4 + n^{-1}h^{-d})\,.
\end{aligned}$$

This establishes the first part of the Lemma. Notice that assumptions (4.15) and

(4.16) ensure that the factors $\mu_2(K)^2 A_{\mathcal{H}_g}$ and $R(K)A_{\mathcal{C}}$ are well defined and strictly positive. Elementary calculus minimising (4.20) with respect to h completes the Lemma. ■

Observe that we may express Y_i using a first order Taylor expansion for g:

$$g(x) + \mathcal{D}_g(x)^T(X_i - x) + \epsilon_i + T(x)\,,$$

where the remainder term is $T(x) = \sum_{j,k} e_{j,k}(x)(X_{ij} - x^{(j)})(X_{ik} - x^{(k)})$ with the terms $e_{j,k}$ are uniformly bounded. For local linear regression we aim to show that our local linear approximation $\hat{\gamma}_0 + \hat{\gamma}^T(X_i - x)$ is a good approximation for this expansion and that the remainder behaves. The following two results are needed before proving the Theorem 4.2 and Theorem 4.3. Firstly, the following version of Bernstein's Inequality may be found in Ibragimov and Linnik (1971), p169.

Theorem 4.6 (Bernstein's Inequality)**.** *Suppose U_i are independent random variables, let $A_2 = \sum_{i=1}^n \mathrm{Var}(U_i)$ and $S_n = \sum_{i=1}^n U_i$. Suppose further that for some $L > 0$ we have*

$$|E[\{U_i - E(U_i)\}^k]| \leq \tfrac{1}{2}\,\mathrm{Var}(U_i)L^{k-2}k!\,.$$

Then

$$P\{|S_n - E(S_n)| \geq 2t\sqrt{A_2}\} < 2e^{-t^2}\,.$$

Secondly, the following lemma contains a proof which is applicable to many uniform convergence type results. The structure is similar to that of Masry (1996), although it is simplified considerably when using independent observations and Bernstein's Inequality. In the proof, let $C_4 = \sup f(x) < \infty$ and $C_5 = \inf f(x) > 0$ for $x \in \mathcal{C}$.

Lemma 4.7. $\sup_{x\in\mathcal{C}} |n^{-1}\sum_i \epsilon_i K_H(X_i - x)| = O\left\{\left(n^{-1}h^{-d}\log n\right)^{1/2}\right\}$.

Proof: Since ϵ_i is independent of X_i and $E(\epsilon_i) = 0$, we have $E\{\epsilon_i K_H(X_i - x)\} = 0$. As $\mathcal{C}$ is compact we may cover it with $L(n) = (n/h^{d+2}\log n)^{d/2}$ cubes $I_1, \ldots, I_{L(n)}$, each with the same side length, proportional to $L(n)^{-1/d}$. Then

$$\begin{aligned}\sup_{x\in\mathcal{C}} |n^{-1}\textstyle\sum_i \epsilon_i K_H(X_i - x)| &\leq \max_m \sup_{x\in\mathcal{C}\cap I_m} |n^{-1}\textstyle\sum \epsilon_i K_H(X_i - x) - n^{-1}\sum \epsilon_i K_H(X_i - x_m)| \\ &+ \max_m |n^{-1}\textstyle\sum \epsilon_i K_H(X_i - x_m)| \;=\; Q_1 + Q_2\end{aligned}$$

From the second condition of (4.16) we know that

$$\begin{aligned}|\epsilon_i K_H(X_i - x) - \epsilon_i K_H(X_i - x_m)| &\leq \frac{C_1\epsilon_i}{h^d} \parallel h^{-1}(x - x_m) \parallel \\ &\leq \frac{C_1'\epsilon_i}{h^{d+1}}\left(\frac{h^{d+2}\log n}{n}\right)^{1/2} = C_1'\epsilon_i\left(\frac{\log n}{nh^d}\right)^{1/2}\end{aligned}$$

This expression is independent of x and m, and so $Q_1 \leq C_1' \left(\frac{\log n}{nh^d}\right)^{1/2} \left|n^{-1}\sum \epsilon_i\right|$, which implies that $Q_1 = O[(\frac{\log n}{nh^d})^{1/2}]$. Now with regard to Q_2, notice that

$$P(Q_2 > \eta) \leq L(n) \sup_x P\{|n^{-1}\textstyle\sum \epsilon_i K_H(X_i - x)| > \eta\}\,. \tag{4.22}$$

Letting $B_2 = \sup_x K(u)$ and using the first property in (4.19) we see that for $k = 3, 4, \ldots,$

$$\begin{aligned} |E[\{\epsilon_i K_H(X_i - x)\}^\alpha]| &\leq \sigma^2 C_3^{\alpha-2} \int K_H(u-x)^\alpha f(u) du \\ &\leq \mathrm{Var}\{\epsilon_i K_H(X_i - x)\}(B_2 C_3)^{\alpha-2}\,. \end{aligned}$$

Also, if $B_3 = \int K(u)^2 du$ we can show that $\mathrm{Var}\{K_H(X_i - x)\} \leq C_4 \sigma^2 B_3 h^{-d}$. We may let n be large enough so that

$$(B_4 \log n)^{1/2} \leq \frac{\sqrt{\sum E\{\epsilon_i^2 K_H(X_i - x)^2\}}}{2B_2 C_3}\,,$$

for some B_4 to be determined below. Then by Bernstein's inequality

$$\begin{aligned} P\Big\{ & \left|n^{-1}\sum \epsilon_i K_H(X_i - x)\right| \geq 2(B_4 \log n)^{1/2}\Big(\frac{\sigma^2 C_4 B_3^2}{nh^d}\Big)^{1/2}\Big\} \\ &\leq P\left\{|\sum \epsilon_i K_H(X_i - x)| \geq 2(B_4 \log n)^{1/2}\sqrt{\sum E(\epsilon_i^2 K_H(X_i - x)^2)}\right\} \\ &\leq 2e^{-B_4 \log n} \leq \frac{2}{n^{-B_4}}\,. \end{aligned}$$

Comparing this inequality to (4.22) and choosing B_4 large enough then the expression $2L(n)n^{-B_4}$ is summable, by the Borel-Cantelli lemma we may conclude that $Q_2 = O[(\frac{\log n}{nh^d})^{1/2}]$ and the lemma is proved. ■

In a similar fashion it is also possible to prove, letting $Z_i = X_i - x$ and $\zeta = (n^{-1}h^{-d}\log n)^{1/2}$,

$$\sup_x |n^{-1}\textstyle\sum_i K_H(Z_i) - E\{K_H(Z_i)\}| = O(\zeta) \tag{4.23}$$

$$\sup_x |n^{-1}\textstyle\sum_i K_H(Z_i)^2 - E\{K_H(Z_i)^2\}| = O(h^{-d}\zeta) \tag{4.24}$$

$$\sup_x |n^{-1}\textstyle\sum_i Z_{ij} K_H(Z_i) - E\{Z_{ij} K_H(Z_i)\}| = O(h\zeta) \tag{4.25}$$

$$\sup_x |n^{-1}\textstyle\sum_i \epsilon_i Z_{ij} K_H(Z_i) - E\{\epsilon_i Z_{ij} K_H(Z_i)\}| = O(h\zeta) \tag{4.26}$$

$$\sup_x |n^{-1}\textstyle\sum_i Z_{ij} Z_{ik} K_H(Z_i) - E\{Z_{ij} Z_{ik} K_H(Z_i)\}| = O(h^2\zeta) \tag{4.27}$$

$$\sup_x |n^{-1}\textstyle\sum_i e_{jk} Z_{ij} Z_{ik} K_H(Z_i) - E\{e_{jk} Z_{ij} Z_{ik} K_H(Z_i)\}| = O(h^2\zeta) \tag{4.28}$$

$$\sup_x \Big|n^{-1}\textstyle\sum_i e_{jk} Z_{ij} Z_{ik} Z_{il} K_H(Z_i) - E\{e_{jk} Z_{ij} Z_{ik} Z_{il} K_H(Z_i)\}\Big| = O(h^3\zeta) \tag{4.29}$$

Standard treatment of the expectation integrals reveals that

$$\begin{aligned} E\{K_H(Z_i)\} &= f(x) + O(h) && (4.30)\\ E\{K_H(Z_i)^2\} &= h^{-d}\{f(x)R(K) + O(h)\} && (4.31)\\ E\{Z_{ij}K_H(Z_i)\} &= O(h^2) && (4.32)\\ E\{\epsilon_i Z_{ij}K_H(Z_i)\} &= 0 && (4.33)\\ E\{Z_{ij}Z_{ik}K_H(Z_i)\} &= O(h^2) && (4.34)\\ E\{e_{jk}Z_{ij}Z_{ik}K_H(Z_i)\} &= O(h^2) && (4.35)\\ E\{e_{jk}Z_{ij}Z_{ik}Z_{il}K_H(Z_i)\} &= O(h^4) && (4.36) \end{aligned}$$

If $h \asymp n^{-1/(d+4)}$, as it will be under Lemma 4.5, then the asymptotic rates in the expectations (4.30)-(4.36) will dominate those of the deviations (4.23)-(4.29), with the execption of (4.33). We may then conclude that, uniformly on x,

$$\begin{aligned} n^{-1}\sum_i K_H(Z_i) &= f(x) + O(h) && (4.37)\\ n^{-1}\sum_i K_H(Z_i)^2 &= h^{-d}\{f(x)R(K) + O(h)\} && (4.38)\\ n^{-1}\sum_i \epsilon_i K_H(Z_i) &= O(h) && (4.39)\\ n^{-1}\sum_i Z_{ij}K_H(Z_i) &= O(h^2) && (4.40)\\ n^{-1}\sum_i \epsilon_i Z_{ij}K_H(Z_i) &= O(h^2) && (4.41)\\ n^{-1}\sum_i Z_{ij}Z_{ik}K_H(Z_i) &= O(h^2) && (4.42)\\ n^{-1}\sum_i e_{jk}Z_{ij}Z_{ik}K_H(Z_i) &= O(h^2) && (4.43)\\ n^{-1}\sum_i e_{jk}Z_{ij}Z_{ik}Z_{il}K_H(Z_i) &= O(h^4) && (4.44) \end{aligned}$$

Proof of Theorem 4.2: From Lemma 4.5 we know that an estimator of h that minimises mean integrated squared error will satisfy $h \asymp n^{-1/(d+4)}$. Theorem 4.1 then implies that

$$\sup_{x\in\mathcal{C},j=1,\ldots,d} |(\widehat{D^j g})(x) - (D^j g)(x)| = O(h\sqrt{\log n})\,.$$

Notice that the estimates $\hat{\gamma}_j$ at x in the minimisation (4.5) are exactly the estimates $(\widehat{D^j g})(x)$. The adjusted parameter estimates $\hat{\beta}_j$ in (4.8) therefore satisfy

$$\hat{\beta}_j = (\widehat{D^j g})(x)\{\textstyle\sum(X_{ij} - \bar{X}_x^{(j)})^2 K_H(X_i - x)\}^{1/2}\,. \qquad (4.45)$$

Let $\beta_j = (D^j g)(x)\{nh^2\mu_2(K)f(x)\}^{1/2}$. We aim to show that $\hat{\beta}$ converges to β sufficiently fast uniformly in x.

$$\begin{aligned}
\sup_{x\in\mathcal{C},\, j} |\hat{\beta}_j - \beta_j| &\leq \sup_{x,j} \left|(\widehat{D^j g})(x)[\{\sum (X_{ij} - \bar{X}_x^{(j)})^2 K_H(X_i - x)\}^{1/2} - \{nh^2\mu_2(K)f(x)\}^{1/2}]\right| \\
&+ \sup_{x,j} \left|\{nh^2\mu_2(K)f(x)\}^{1/2}\{(\widehat{D^j g})(x) - (D^j g)(x)\}\right| \\
&\leq A_1 \sup_{x,j} \left|\{\sum (X_{ij} - \bar{X}_x^{(j)})^2 K_H(X_i - x)\}^{1/2} - \{nh^2\mu_2(K)f(x)\}^{1/2}\right| \\
&+ O(h^2\sqrt{n\log n}) \qquad (4.46)
\end{aligned}$$

In the first term of the last line we use the fact that $(D^j g)$ is bounded and $(\widehat{D^j g})$ converges uniformly so may be bound be some constant A_1, and for the second term we use the boundedness of $f(x)$ and (4.45).

Focusing on the first term, note that

$$\sup_{x,\,j} |\bar{X}_x^{(j)} - x^{(j)}| = \sup_{x,\,j} \left|\frac{\sum (X_{ij} - x^{(j)}) K_H(X_i - x)}{\sum K_H(X_i - x)}\right| = O(h^2)\,,$$

using (4.37) and (4.40). Thus

$$\begin{aligned}
\sum_i (X_i - \bar{X}_x^{(j)})^2 K_H(X_i - x) &= \sum \{X_{ij} - x^{(j)} + O(h^2)\}^2 K_H(X_i - x) \\
&= O(nh^4) + \sum (X_{ij} - x^{(j)})^2 K_H(X_i - x)\,,
\end{aligned}$$

again using (4.37) and (4.40). Now we consider the expectation of $(X_{ij} - x^{(j)})^2 K_H(X_i - x)$ carefully,

$$\begin{aligned}
E\{(X_{ij} - x^{(j)})^2 K_H(X_i - x)\} &= \int (u^{(j)} - x^{(j)})^2 K_H(u - x) f(u) du \\
&= h^2 \int (z^{(j)})^2 K(z) f(x + hz) dz \\
&= h^2 \int (z^{(j)})^2 K(z)\{f(x) + hz^T \mathcal{D}_f(x) + O(h^2)\} dz \\
&= h^2\mu_2(K) f(x) + O(h^4)\,.
\end{aligned}$$

The differentiability assumptions in the statement of the Theorem ensure that this formulation is uniform over all x in $\mathcal{C}$. Using this and (4.27) in (4.46) and noting

that if $x \to 0$ then $(1+x)^{1/2} - 1 = O(x)$, we see that

$$\begin{aligned}
\sup_{x\in\mathcal{C},\, j} |\hat{\beta}_j - \beta_j| &\leq A_1 \sup_{x,j} |\{nh^2\mu_2(K)f(x) + O(nh^2\sqrt{\log n}))\}^{1/2} - \{nh^2\mu_2(K)f(x)\}^{1/2}| \\
&+ O(h^2\sqrt{n\log n}) \\
&= \sup_x A_1\{nh^2\mu_2(K)f(x)\}^{1/2}|\{1 + O(h^2\sqrt{\log n})\}^{1/2} - 1| \\
&+ O(h^2\sqrt{n\log n}) \\
&= \sup_x A_1\{nh^2\mu_2(K)f(x)\}^{1/2}O(h^2\sqrt{\log n}) + O(h^2\sqrt{n\log n}) \\
&= O(h^2\sqrt{n\log n})
\end{aligned}$$

We do not need to worry about small nonzero values of β_j by our assumption on $\mathcal{O}_\delta$, so the nonzero β_j grow at $O(n^{1/2}h)$. Further, the estimate error of $\hat{\beta}_j$ is $O(h^2\sqrt{n\log n})$ uniformly in x. A λ that grows at some rate between these two as, suggested in the Theorem, will be able to separate the true variables from the redundant ones with probability tending to 1. ■

Proof of Theorem 4.3: Let $\hat{\mu}_0$, $\hat{\mu}$ be the parameter estimates for the case where the jth variable is removed from consideration, so $\mu^{(j)} = 0$. Theorem 4.1 ensures that the maximum distance for the estimators $\hat{\gamma}_0, \hat{\mu}_0$ from $g(x)$ is $O(\zeta)$ and similarly $\hat{\gamma}, \hat{\mu}$ converge to the derivative $\mathcal{D}_g(x)$ at $O(\zeta h^{-1})$, with the exception of $\hat{\mu}^{(j)} = 0$. Thus we may expand the sum of squares difference and use results (4.37)–(4.44):

$$\begin{aligned}
&SS_j(x) - SS(x) \\
&= n^{-1}\sum\{Y_i - \hat{\mu}_0 - \hat{\mu}^T Z_i\}^2 K_H(Z_i) - n^{-1}\sum\{Y_i - \hat{\gamma}_0 - \hat{\gamma}^T Z_i\}^2 K_H(Z_i) \\
&= n^{-1}\sum K_H(Z_i)\Big[\{O(\zeta) + \epsilon_i + \mathcal{D}_g(x)^{(j)}Z_{ij} + T(x) + O(\zeta h^{-1})\textstyle\sum_k Z_{ik}\}^2 \\
&\quad -\{O(\zeta) + \epsilon_i + T(x) + O(\zeta h^{-1})\textstyle\sum_k Z_{ik}\}^2\Big] \\
&= n^{-1}\sum K_H(Z_i)\Big[O(\zeta^2) + \epsilon_i O(\zeta) + T(x)O(\zeta) + O(\zeta^2 h^{-1})\textstyle\sum_k Z_{ik} \\
&\quad +O(\zeta)\mathcal{D}_g(x)^{(j)}Z_{ij} + 2\epsilon_i\mathcal{D}_g(x)^{(j)}Z_{ij} + 2T(x)\mathcal{D}_g(x)^{(j)}Z_{ij} \\
&\quad +O(\zeta h^{-1})\mathcal{D}_g(x)^{(j)}\textstyle\sum_k Z_{ij}Z_{ik} + (\mathcal{D}_g(x)^{(j)}Z_{ij})^2 + O(\zeta h^{-1})\epsilon_i\textstyle\sum_k Z_{ik} \\
&\quad +O(\zeta h^{-1})T(x)\textstyle\sum_k Z_{ik} + O(\zeta^2 h^{-2})\textstyle\sum_{k,l} Z_{ik}Z_{il}\Big] \\
&= O(\zeta^2) + (\mathcal{D}_g^{(j)})^2 O(h^2)\,.
\end{aligned}$$

This shows the behaviour of the numerator in our expression. Note that our assumption on $\mathcal{O}_\delta$ ensures that when $|\mathcal{D}_g^{(j)}|$ is nonzero it is bounded away from 0, so true separation is possible. In a similar fashion to that above we may expand and

deal with the denominator $n^{-1}\sum(Y_i - \hat{\gamma}_0 - \hat{\gamma}^T Z_i)^2 K_H(Z_i)$. The dominating term here is the asymptotic expectation of $n^{-1}\sum \epsilon_i^2 K_H(Z_i)$, which tends to $\sigma^2 f(x)$, and everything else converges to zero at h or faster, uniformly in x. Therefore, so as long as λ shrinks faster than h^2 but slower than $\zeta^2 = h^4 \log n$, the variable selection will be uniformly consistent. ■

Before proving Theorem 4.4, we prove the following three lemmas. The first allows us to separate out the effects of various variables in the LABAVS procedure. The latter two are concerned with the change in estimation error for local and global variable redundance respectively.

Lemma 4.8. *Let $\mathcal{B}_1$ and $\mathcal{B}_2$ be disjoint subsets of $\{1,\ldots,d\}$ such that $\mathcal{B}_1 \cup \mathcal{B}_2 = \{1,\ldots,d\}$. The final estimates of the LABAVS procedure would be the same as applying the bandwidth adjustment, that is steps 3 and 4, in the procedure twice; the first time only expanding the bandwidths at x of those variables in $\hat{\mathcal{A}}^-(x) \cap \mathcal{B}_1$ to the edges of the maximal rectangle and shrinking those remaining, and the second time expanding the variables in $\hat{\mathcal{A}}^-(x) \cap \mathcal{B}_2$ and shrinking the variables in $\hat{\mathcal{A}}^+(x)$.*

Proof: Choose $x \in \mathcal{C}$. With some slight abuse of notation, since the bandwidths are possibly asymmetric, let $H_1(x)$ denote the adjusted bandwidths after the first step of the two-step procedure, with shrunken variables having bandwidth h_1. Similarly let $H_2(x)$ denote the bandwidths after the second step, with bandwidth on the shrunken variables h_2. Further, let $d_1(x)$ equal the cardinality of $\hat{\mathcal{A}}^+(x) \cup \mathcal{B}_1$ and $d_2(x)$ equal the cardinality of $\hat{\mathcal{A}}^+(x)$. The bandwidths for the redundant variables are expanded to the edges of the maximal rectangle, so we need only show that the resulting shrunken bandwidth is the same as when applying the one-step version of the algorithm. Using expression (4.14) we know that

$$\begin{aligned} h_1 &= M(H_1, H) = h\left[\frac{E\{d_1(X)\}E\{V[X,H_1]\}}{dE\{V[X,H]\}}\right], \text{ and} \\ h_2 &= M(H_2, H_1) = h_1\left[\frac{E\{d_2(X)\}E\{V[X,H_2]\}}{E\{d_1(X)\}E\{V[X,H_1]\}}\right]. \end{aligned}$$

Substituting the first expression into the second gives

$$h_2 = h\left[\frac{E\{d_2(X)\}E\{V[X,H_2]\}}{dE\{V[X,H]\}}\right],$$

which recovers the equation in the one-step bandwidth adjustment. Thus the bandwidths are unchanged for every $x \in \mathcal{C}$. ■

Lemma 4.9. *Suppose that h is chosen to minimise squared error as in (4.17). Also, suppose that the LABAVS procedure identifies that no variables are globally redundant*

but some (possibly all) variables are locally redundant and that the local redundancy takes place on a set of non-zero measure. Then the LABAVS procedure reduces the overall MISE of the estimation of g by a factor of $M[\{H^L(X), H^U(X)\}, H] < 1$.

Proof: We shall ignore the difficulties associated with incorrect selection on $\mathcal{O}_\delta$, as it only affects an arbitrarily small subset of the domain. With probability tending to one we have correct variable classification, so we work under this assumption. Since some variables are relevant in some regions, the choice of h' is well defined. Pick $x \in \mathcal{C}$ and let u^+ denote the components of d-vector u indexed by $\mathcal{A}^+(x)$ and u^- be the residual components. We can express the density $f(x)$ as $f(x^+, x^-)$ so the relevant and redundant components may be treated separately. From (4.37) and (4.38) we know that

$$V[x, H] = \frac{h^{-d}\{R(K)f(x) + O(h)\}}{\{f(x) + O(h)\}^2} = h^{-d}\left\{\frac{R(k)}{f(x)} + O(h)\right\}.$$

Taking an expectation over x we see that

$$E\{V[X, H]\} = h^{-d}R(K)A_{\mathcal{C}} + O(h^{-(d-1)}). \tag{4.47}$$

For convenience let $H^*(x)$ denote the asymmetric bandwidths $H^L(x)$ and $H^U(x)$. We now show that the factor $M\{H^*(X), H\}$ is less than 1. Firstly observe that $V[x, H] = V[x, H^*(x)]$ whenever $\mathcal{A}^+(x) = (1, \ldots, d)$. Consider the case when $\mathcal{A}^+(x) \neq (1, \ldots, d)$. In particular assume that k components are redundant at x. We see that

$$\begin{aligned}
E\{K_{H^*}(X_i - x)^2\} &= \int\int \Bigg\{ \prod_{j \in \mathcal{A}^-(x)}^{n} h_j^*(x)^{-2} K^*\{h_j^*(x)^{-1}(u^{(j)} - x^{(j)})\}^2 \\
&\quad \cdot (h')^{-(d-k)} \prod_{j \in \mathcal{A}^+(x)}^{n} K^*(z^{(j)})^2 \cdot f(x^+ + h'z^+, u^-) \Bigg\} dz^+ du^- \\
&= (h')^{-(d-k)} R(K)^{(d-k)/d} \Bigg[O(h') \\
&\quad + \int \prod_{j \in \mathcal{A}^-(x)}^{n} h_j^*(x)^{-2} K^*\{h_j^*(x)^{-1}(u^{(j)} - x^{(j)})\}^2 f(x^+, u^-) du^- \Bigg] \\
&= (h')^{-(d-k)}\{B_1(x) + O(h')\},
\end{aligned}$$

where $B_1(x)$ is a uniformly bounded and strictly positive number depending only on x. An argument using Bernstein's Theorem similar to that in Lemma 4.7 shows that the uniform bound of $n^{-1}\sum K_{H^*}(X_i - x)^2$ away from $E\{K_{H^*}(X_i - x)^2\}$ is

$O[(h')^{-(d-k)}\{n(h')^{(d-k)}\}^{-1/2}\sqrt{\log n}]$, so we may deduce that

$$n^{-1}\sum K_{H^*}(X_i - x)^2 = (h')^{-(d-k)}\{B_1(x) + O(h')\}\,.$$

In a similar fashion we can show that

$$n^{-1}\sum K_{H^*}(X_i - x) = f_{X^+}(x^+)B_2(x) + O(h')\,,$$

where $B_2(x)$ is a uniformly bounded and strictly positive number depending only on x. This leads to

$$V[x, H^*] = (h')^{-(d-k)}\left\{\frac{B_1(x)}{B_2(x)^2} + O(h')\right\}\,. \tag{4.48}$$

Let $\mathcal{E}$ denote the event that $\mathcal{A}^+(X) = (1, \ldots, n)$ and $\mathcal{E}^c$ the complement. We know $P(\mathcal{E}^c) > 0$ by assumption and also that given $\mathcal{E}^c$ is true, $V[X, H^*] = O\{(h')^{-(d-1)}\}$ from (4.48). Thus as $n \to \infty$, for some strictly positive constants B_3, B_4 and B_5,

$$\begin{aligned}\frac{E\{V[X, H^*]\}}{E\{V[X, H]\}} &= \frac{P(\mathcal{E})E\{V[X, H^*]\ |\mathcal{E}\} + P(\mathcal{E}^c)E\{V[X, H^*]\ |\mathcal{E}^c\}}{P(\mathcal{E})E\{V[X, H]\ |\mathcal{E}\} + P(\mathcal{E}^c)E\{V[X, H]\ |\mathcal{E}^c\}} \\ &= \frac{h^{-d}\{B_3 + O(h)\} + O\{(h')^{-(d-1)}\}}{h^{-d}\{B_3 + O(h)\} + h^{-d}\{B_4 + O(h)\}}\,,\end{aligned}$$

where B_3, B_4 are constants satisfying $B_3 \geq 0$ and $B_4 > 0$. But from our definition of h' in (4.14) and the definition of $M(H^*(X), H)$, we may deduce that

$$\left(\frac{h'}{h}\right)^4 \asymp \frac{B_3 + O\{h^d(h')^{-(d-1)}\}}{B_3 + B_4} + O(h)\,.$$

From this expression it follows that both sides must be less than 1 in the limit. Thus we have $M(H^*, H) < 1$ asymptotically, as required.

Ruppert and Wand (1994) show that for a point x in the interior of $\mathcal{C}$, $\mathrm{Var}(\hat{g}(x)|X_1, \ldots, X_n)$, using a bandwidth matrix H, is equal to $\sigma^2 e_1^T(\ddot{X}^T W \ddot{X})^{-1}\ddot{X}^T W^2 \ddot{X}(\ddot{X}^T W \ddot{X})^{-1}e_1$, where $\ddot{X}$ is the $n \times (p+1)$ matrix $(\mathbf{1}, X)$, e_1 is a p-vector with first entry 1 and the others 0, and W is an $n \times n$ diagonal matrix with entries $K_H(X_i - x)$. This variance may be reexpressed as

$$\sigma^2 \frac{\sum K_H(X_i - x)^2}{\{\sum K_H(X_i - x)\}^2} e_1^T(\ddot{X}^T \ddot{X})^{-1}\ddot{X}^T \ddot{X}(\ddot{X}^T \ddot{X})^{-1}e_1\,.$$

Taking ratios of the expectations for the variance factors under the adjusted and initial bandwidths recovers the expression $M(H^*, H)$ in (4.12). Thus the variance term in the MISE is reduced by a factor of $M(H^*, H)$. Furthermore, the bias term in the MISE, in which we may ignore the zero bias contributed be the nth variable where it is redundant, is reduced by a factor of $(h'/h)^4$ which, from (4.14), is strictly

less than the factor $M(H^*, H)$. Thus the MISE is reduced by the factor $M(H^*, H)$ as required. ■

Lemma 4.10. *Suppose that h is chosen to minimise squared error as in (4.17). Also, suppose that the LABAVS procedure finds that all variables are relevant everywhere in $\mathcal{C}$ except for a single variable $X^{(j)}$, which is globally irrelevant. Then the LABAVS procedure reduces the overall MISE of the estimation of g by a factor of $M(H^*, H) < 1$. Furthermore the resulting bandwidth h' is asymptotically optimal, in the sense that it minimises the $d-1$ dimensional MISE expression.*

Proof: Let $\mathcal{C}'$ denote the $d-1$ dimensional space formed by removing the irrelevant variable and denote the volume of this space by $A_{\mathcal{C}'}$. We know that our initial h satisfies (4.21). By similar reasoning it follows that we are required to show that our adjusted bandwidth is asymptotically equal to

$$h_{\text{opt}} = \left[\frac{(d-1)\sigma^2 R(K)^{(d-1)/d} A_{\mathcal{C}'}}{n\mu_2(K)^2 A_{\mathcal{H}_g}} \right]^{1/(d+3)}, \tag{4.49}$$

which is the bandwidth the minimises MISE in the reduced dimension case. Here $A_{\mathcal{C}'}$ denotes the volume of the $d-1$ dimensional case. Equivalently, combining (4.21) and (4.49), it is sufficient to show in the limit that

$$\frac{(h')^{d+3}}{h^{d+4}} = \frac{(d-1)A'_{\mathcal{C}}}{dR(K)^{1/d} A_{\mathcal{C}}}. \tag{4.50}$$

Arguments similar to those in the previous Lemma can be made to show

$$n^{-1} \sum K_{H^*}(X_i - x)^2 = (h')^{-(d-1)} \{R(K)^{(d-1)/d} f_{X^{(-n)}}(x^{(-n)}) + O(h')\}, \text{ and}$$

$$n^{-1} \sum K_{H^*}(X_i - x) = \{R(K)^{(d-1)/d} f_{X^{(-n)}}(x^{(-n)}) + O(h').$$

Thus

$$\begin{aligned} E\{V[X, H^*]\} &= (h')^{-(d-1)} \Big\{ R(K)^{(d-1)/d} \int\int f_{X^{(-n)}}(x^{(-n)})^{-1} f_{X^{(-n)}}(x^{(-n)}) \\ &\quad \cdot f_{X^{(n)}|X^{(-n)}}(u^{(n)}) du^{(n)} du^{(-n)} + O(h') \Big\} \\ &= (h')^{-(d-1)} \{A_{\mathcal{C}'} R(K)^{(d-1)/d} + O(h')\} \end{aligned}$$

Combining this with (4.47) and (4.14) gives

$$\left(\frac{h'}{h}\right)^4 = \frac{d-1}{d} M(H^*, H) = \frac{d-1}{d} \frac{h^d}{(h')^{d-1}} \left\{ \frac{A_{\mathcal{C}'}}{A_{\mathcal{C}}} R(K)^{-1/d} + O(h) \right\}.$$

Rearranging this last expression and letting $n \to \infty$ leads to the required expression

(4.50). Note that (4.50) also implies that $(h'/h)^{d+3}h^{-1}$ is asymptotically constant, so $h'/h \to 0$. This in turn implies that $(h'/h)^4 = M(H^*, H)(d-1)/d$ tends to zero so asymptotically $M(H^*, H) < 1$ as required. The argument that the MISE is in fact reduced by the factor $M(H^*, H)$ is entirely analogous to the previous Lemma. ■

Proof of Theorem 4.4: Correct variable selection at every point $x \in \mathcal{C}$ with probability tending to 1 on the set $\mathcal{C} \setminus \mathcal{O}_\delta$ for locally redundant variables, and $\mathcal{C}$ for globally redundant variables, is guaranteed by Theorem 4.2 or Theorem 4.3. For a given point x, repeated application of Lemma 4.8 allows us to consider the eventually result by adjusting the bandwidths for any partition of variables in any order. Choose an order in which globally redundant variables are treated first, one at a time, followed by a final adjustment for those variables that are locally redundant. Lemma 4.10 ensures that when allowing for each globally redundant variable, the resulting bandwidths in the remaining variables is asymptotically optimal. This means that the strong nonparametric oracle property is satisfied after the global bandwidth adjustments. Lemma 4.9 provides the quantification of the additional benefit resulting from the local variable removal. ■

Chapter 5

Bootstrap assessment of an empirical ranking

5.1 Background

We have seen in Chapter 2 that attempting to find key variables in a high-dimensional context will often amount to a ranking of the components. More broadly, the ordering of a sequence of random variables is often a major aspect of contemporary statistical analyses. For example, data on the comparative performance of institutions (e.g. local governments, or health providers, or universities) are frequently summarised by reporting the ranking of empirical values of a performance measure; and the relative influence of genes on a particular response is sometimes indicated by ranking the values of the weights that are applied to them after the application of a variable selector, such as the lasso. It is reasonable to argue that, especially in contentious situations, no ranking should be unaccompanied by a measure of its authority (Goldstein and Spiegelhalter, 1996). The bootstrap is a popular approach to developing such a measure.

This chapter focuses on both the theoretical and the numerical properties of bootstrap estimators of the distributions of rankings. We show that the standard n-out-of-n bootstrap, introduced in Section 2.3.2 generally fails to give consistency when comparisons between components are close, and in fact may not produce distribution estimators that converge either almost surely or in probability. The m-out-of-n bootstrap overcomes these difficulties, but requires empirical choice of m. We suggest a tuning approach to solving this problem. This technique remains appropriate in cases where the number, p say, of populations is very large, although in that context one could also regard m as a means of setting the level of sensitivity of the bootstrap

to near-ties among ranks, rather than as a smoothing parameter.

In some contemporary prediction problems the empirical rank is quite highly variable. We develop mathematical models in this setting, and explore the validity of bootstrap methods there. In particular, we show that the inherent inconsistency of the standard n-out-of-n bootstrap does not prevent that method from correctly capturing the order of magnitude of the expected value of rank, or the expected length of prediction intervals, although it leads to errors in estimators of the constant multiplier of that order of magnitude.

Another issue is that of adequately reflecting, in the bootstrap algorithm, dependence among the datasets representing the different populations — e.g. data on the performances of different health providers, or on the expression levels of different genes. In examples of the first type, where different institutions are being ranked, the assumption of independence is often appropriate; it can usually be accommodated through conditioning. In such cases, resampling can be implemented in a way that explicitly reflects population-wise independence.

However, in the genomic example, data on expression levels of different genes from the same individual are generally not independent. In this setting, using the standard nonparametric bootstrap to assess the authority of ranking would seem to be a good choice, since in more conventional problems it captures well the dependence structure of data vectors. However, we show that, even when the number of variables being ranked is much less than sample size, the standard approach can give unreliable results in some problems. This is largely because knowing the composition of a resample for the jth population (e.g. for the jth gene, in the genomic example) identifies exactly the resamples for other genes. Therefore, the resamples for different populations are hardly independent, even conditional on the original data.

This has a variety of repercussions. For example, it implies that standard bootstrap probabilities, when computed conditional on the information we have in the resample about the jth gene, degenerate to indicator functions. Conditional inference is attractive in ranking problems, since it can lead to substantial reductions in variability. To overcome the problem we suggest using an "independent component" version of the bootstrap, where the bootstrap is applied as though the ranked variables were statistically independent. This approach can be valid even in the case of non-independence. (In order to make it clear that in this setting we use the term "standard bootstrap" to mean the resampling of p-vectors of data, we shall refer to this bootstrap method as the "synchronous" bootstrap; the standard bootstrap results in vector components being synchronised with one another in each resampling step.)

It is possible to generalise our treatment to cases where several rankings are undertaken jointly, for example where universities are ranked simultaneously in terms of

the quality of their graduate programs and the career prospects of their undergraduates. Our main conclusions about the relative merits of different bootstrap methods persist in this more general setting, although a detailed treatment of that case would be significantly longer and more complex.

Work on the bootstrapping of statistics related to ranks includes that of Srivastava (1987), who introduced bootstrap methods for a class of ranking and slippage problems (although not directly related to the problems discussed in this chapter); Tu et al. (1992), who discussed bootstrap methods for canonical correlation analysis; Langford and Leyland (1996), who address bootstrap methods for ranking the performance of doctors; Larocque and Léger (1994), Steland (1998) and Pelin et al. (2000), who developed bootstrap methods for quantities such as rank tests and rank statistics; Goldstein and Spiegelhalter (1996), who discussed bootstrap methods for constructing interval estimates; Langford and Leyland (1996), who addressed bootstrap methods for ranking the performance of doctors; Cesário and Barreto (2003), Hui et al. (2005) and Taconeli and Barreto (2005), who discussed bootstrap methods for ranked set sampling; and Mukherjee et al. (2003), who developed methods for gene ranking using bootstrapped p-values.

The problem treated in this chapter is the same one addressed independently by Xie et al. (2009). While the setup and use of the bootstrap are similar, the methods for addressing the possible degeneracy of the standard bootstrap are very different; Xie et al. focus on soft-thresholding methods, while our work uses the m-out-of-n bootstrap.

5.2 Methodology

5.2.1 Model. Assume we have datasets $\mathcal{X}_1, \ldots, \mathcal{X}_p$ drawn from populations $\Pi_1, \ldots, \Pi_p$, respectively, and that for the jth population there is an associated parameter θ_j which measures, for example, the strength of an attribute in the population, or the performance of an individual or an organisation related to the population, or the esteem in which an institution or a program is held. If the θ_j's were known then our ranking of the populations would be

$$r_j = 1 + \sum_{k \neq j} I(\theta_k \geq \theta_j)\,, \quad \text{for } j = 1, \ldots, p\,, \tag{5.1}$$

say, signifying that r_j is the rank of the jth population. Here, tied rankings can be considered to have been broken arbitrarily, for example at random.

We wish to develop an empirical version of the ranking at (5.1). For this purpose we compute from $\mathcal{X}_j$ an estimator $\hat{\theta}_j$ of θ_j, for $1 \leq j \leq p$, and we rank the populations

in terms of the values of $\hat{\theta}_j$. In particular, if we have $\hat{\theta}_1, \ldots, \hat{\theta}_p$, then we write

$$\hat{r}_j = 1 + \sum_{k \neq j} I(\hat{\theta}_k \geq \hat{\theta}_j)\,, \quad \text{for } j = 1, \ldots, p\,, \tag{5.2}$$

to indicate the empirical version of (5.1). Again ties can be broken arbitrarily, although in the case of (5.2) the noise implicit in the estimators $\hat{\theta}_j$ often means that there are no exact ties.

We shall treat two cases: "fixed p" and "large p," distinguished in theoretical models by taking p fixed and allowing n to diverge, and by permitting p to diverge, respectively. Cases covered by the latter model include instances where $\mathcal{X}_0$ is a set of p-vectors, say $\mathcal{X}_0 = \{X_1, \ldots, X_n\}$ where $X_i = (X_{i1}, \ldots, X_{ip})$. There, $\mathcal{X}_j = \{X_{1j}, \ldots, X_{nj}\}$ is the set of jth components of each data vector, and in particular each $\mathcal{X}_j$ is of the same size. This example arises frequently in contemporary problems in genomics, where X_i is the vector of expression-level data on perhaps $p = 5\,000$ to $20\,000$ genes for the ith individual in a population. In such cases n can be relatively small, for example between 20 and 200. The vectors X_i can generally be regarded as independent, but not so the components $\mathcal{X}_1, \ldots, \mathcal{X}_p$. However, as we shall argue in Section 5.4.1, there may be advantages in conducting inference as though the components were independent, even when that assumption is incorrect.

5.2.2 Basic bootstrap methodology. The authority of the ranking at (5.2), as an approximation to that at (5.1), can be queried. A simple approach to quantifying the authority is to repeat the ranking many times in the context of bootstrap resamples $\mathcal{X}_1^*, \ldots, \mathcal{X}_p^*$, which replace the respective datasets $\mathcal{X}_1, \ldots, \mathcal{X}_p$. In particular, for each sequence $\mathcal{X}_1^*, \ldots, \mathcal{X}_p^*$ we can compute the respective versions $\hat{\theta}_1^*, \ldots, \hat{\theta}_p^*$ of the estimators of θ_j, and calculate the bootstrap version of (5.2):

$$\hat{r}_j^* = 1 + \sum_{k \neq j} I(\hat{\theta}_k^* \geq \hat{\theta}_j^*)\,, \quad \text{for } j = 1, \ldots, p\,. \tag{5.3}$$

The bootstrap here can be of conventional n-out-of-n type, either parametric or nonparametric, or it can be the m-out-of-n bootstrap (again either parametric or nonparametric), where the resamples $\mathcal{X}_j^*$ are of smaller size than the respective samples $\mathcal{X}_j$. For definiteness, in Section 5.4, where we need to refer explicitly to the implementation of bootstrap methods, we shall use the nonparametric bootstrap. However, our conclusions also apply to parametric bootstrap methods. More generally, the way in which the bootstrap resamples $\mathcal{X}_j^*$ are constructed can depend on the nature of the data. See Section 5.4.1 for discussion.

One question in which we are obviously interested is whether the bootstrap cap-

tures the distribution of $\hat{r}_j$ reasonably well, for example whether

$$P(\hat{r}_j^* \le r \,|\, \mathcal{X}) - P(\hat{r}_j \le r) \to 0\,, \tag{5.4}$$

in probability for each integer r, as $n \to \infty$. The answer to this question, if we use the familiar n-out-of-n bootstrap, is generally "only in cases where the limiting distribution of $\hat{r}_j$ is degenerate." However, the answer is more positive if we employ the m-out-of-n bootstrap. There, if the populations $\Pi_1, \ldots, \Pi_p$ are kept fixed in an asymptotic study then

$$\begin{aligned}&\text{the limiting distribution of } \hat{r}_j \text{ is supported on the set of integers } \{k_1 + \\ &1, k_1 + 2, \ldots, k_2\}, \text{ where } k_1 = \textstyle\sum_k I(\theta_k > \theta_j) \text{ and } k_2 = \sum_k I(\theta_k \ge \theta_j),\end{aligned} \tag{5.5}$$

and the m-out-of-n bootstrap consistently estimates this distribution. In particular, (5.4) holds; see Section 5.3 for details. However (still in the case of fixed p), if we are more ambitious and permit the population distributions to vary with n in such a way that the limiting distribution is more complex than that prescribed by (5.5), then even the m-out-of-n bootstrap may fail to give consistency.

Having computed a bootstrap approximation $P(\hat{r}_j^* \le r \,|\, \mathcal{X})$ to the probability $P(\hat{r}_j \le r)$, we can calculate an empirical approximation to a prediction interval, specifically an interval $[r_1, r_2]$ within which $\hat{r}_j$ lies with given probability, for example 0.95. Goldstein and Spiegelhalter (1996) refer to such intervals as "overlap intervals," since they are generally displayed in a figure which shows the extent to which they overlap. Particularly when p is relatively small, the discrete nature of the distribution of $\hat{r}_j$ makes it a little awkward to discuss the accuracy of bootstrap prediction intervals, and so we focus instead on measures of the accuracy of distributional approximations, for example (5.4) and (5.5).

5.3 The case of p distinct populations

5.3.1 Preliminary discussion. Write n_j for the size of the sample $\mathcal{X}_j$. The values of n_j may differ, but we shall assume that they all of the same order. That is, writing $n = p^{-1} \sum_j n_j$ for the average sample size, we have:

$$n^{-1} \sup_{1 \le j \le p} n_j = O(1)\,, \quad 1 = O\Big(n^{-1} \inf_{1 \le j \le p} n_j\Big)\,. \tag{5.6}$$

When interpreting (5.6) it is convenient to think of n as the "asymptotic parameter," i.e. the quantity which we take to diverge to infinity, and to consider $n_1, \ldots, n_p$ as functions of n.

When using the m-out-of-n bootstrap, where a resample of size $m_j < n_j$ is drawn

either from the population distribution with estimated parameters (the parametric case) or by with-replacement resampling from the sample $\mathcal{X}_j$ (the case of the non-parametric bootstrap), and $\mathcal{X}_j$ is of size n_j, we assume that the average resample size, $m = p^{-1} \sum_j m_j$, satisfies the analogue of (5.6):

$$m^{-1} \sup_{1 \le j \le p} m_j = O(1)\,, \quad 1 = O\Big(m^{-1} \inf_{1 \le j \le p} m_j\Big)\,. \tag{5.7}$$

Furthermore, we ask that m be large but m/n be small.

In the cases of both fixed and divergent p the properties of $\hat{r}_j$ and $\hat{r}_j^*$ are strongly influenced by the potential presence of tied values of θ_j. However, it is perhaps unreasonable to assume, in practice, that two values of θ_j are exactly tied, although there might be cases where two values are so close that, for most practical purposes, the properties of $\hat{r}_j$ for small to moderate n are similar to those that would occur if the values were tied. The borderline case is that where two values of θ_j differ by only a constant multiple of $n^{-1/2}$, with n denoting average sample size. (This requires the distribution of the populations Π_j to vary with n.) If the constant is sufficiently large then, practically speaking, the two values of θ_j are not tied, but if the constant is small then a tie might appear to be present.

To reflect this viewpoint we shall, for any particular j and for all $k \neq j$, write

$$\theta_k = \theta_j + n^{-1/2}\, \omega_{jk}\,, \tag{5.8}$$

where the ω_{jk}'s are permitted to depend on n. Of course, (5.8) amounts to a definition of ω_{jk}, and if the quantities θ_k, for $1 \le k \le p$, are all fixed then (5.8) implies that ω_{jk} either vanishes or diverges to either $+\infty$ or $-\infty$, in the latter two cases in proportion to $n^{1/2}$. However, since we shall permit the distributions of the populations Π_k, and hence also the θ_k's, to depend on n, then the problem can be set up in such a way that the ω_{jk}'s have many different modes of behaviour.

In the case of the m-out-of-n bootstrap, where $m \to \infty$ but $m/n \to 0$, sensitivity is somewhat reduced by using a smaller resample size. Reflecting this restriction, in the m-out-of-n bootstrap setting we use the following formula to define quantities ω'_{jk}, in place of the ω_{jk}'s at (5.8):

$$\theta_k = \theta_j + m^{-1/2}\, \omega'_{jk}\,. \tag{5.9}$$

It can be proved that, under regularity conditions, the sum over r of the squared distance between the m-out-of-n bootstrap approximation to the distribution function of $\hat{r}_j$, and the limiting form G_j of that distribution (see (5.14) below), equals $C_1\, m^{-1} + C_2\, m\, n^{-1} + o(m^{-1} + m\, n^{-1})$, where C_1 and C_2 are positive constants. This result implies that the asymptotically optimal choice of m equals the integer part

of $(C_1\, n/C_2)^{1/2}$. However, this limit-theoretic argument is not always valid when p is large, and even in the case of small p it is not straightforward to estimate the ratio C_1/C_2. In Section 5.3.4 we suggest an alternative, relatively flexible, method for choosing m.

In most cases where there are p distinct populations it is reasonable to argue that the datasets $\mathcal{X}_1, \ldots, \mathcal{X}_p$ are independent. For example, $\mathcal{X}_j$ might represent a sample relating to the performance of the jth of p health providers that are being operated essentially independently (see e.g. Goldstein and Spiegelhalter, 1996), and the data in $\mathcal{X}_j$ would be gathered in a way that is largely independent of data for other health providers. To the extent to which the data are related, for example through the common effects of government policies, or shared health-care challenges such as epidemics, we might interpret our analysis as conditional on those effects.

If the assumption of independence is valid then it is straightforward to reflect the assumption during the resampling operation, obtaining bootstrap parameter estimators $\hat\theta_1^*, \ldots, \hat\theta_p^*$ that are independent conditional on $\mathcal{X} = \cup_j \mathcal{X}_j$. If the independence assumption is not appropriate then resampling is generally a more complex operation, and may be so challenging as to be impractical. In the remainder of this Section we shall assume that $\mathcal{X}_1, \ldots, \mathcal{X}_p$ are independent, and that $\hat\theta_1^*, \ldots, \hat\theta_p^*$ are independent conditional on $\mathcal{X}$.

Sections 5.3.2 and 5.3.5 will outline theoretical properties in the case of fixed p and increasingly large p, respectively. To simplify and abbreviate our discussion we shall state our main results only for one j at a time, but joint distribution properties can also be derived, analogous to those in Theorem 5.4.

5.3.2 Theoretical properties in the case of fixed p. To set the scene for our results we note first that, under mild regularity conditions, it holds true that for fixed p, for each $1 \le j \le p$ and for each real number x,

$$P\bigl\{n^{1/2}\,(\hat\theta_j - \theta_j) \le \sigma_j\, x\bigr\} \to \Phi(x)\,, \quad P\bigl\{m^{1/2}\,(\hat\theta_j^* - \hat\theta_j) \le \sigma_j\, x \,\big|\, \mathcal{X}\bigr\} \to \Phi(x)\,, \quad (5.10)$$

where the asymptotic standard deviations $\sigma_j \in (0, \infty)$ do not depend on n, Φ denotes the standard normal distribution function, and the convergence in the second part of (5.10) is in probability. In that second part the value of m equals n if we are using the conventional bootstrap, and equals m if we are using the m-out-of-n bootstrap.

The first formula in (5.10) is the conventional statement that the statistics $\hat\theta_j$ are asymptotically normally distributed, and the second is the standard bootstrap form of that assumption. It asserts only that the bootstrap estimator of the distribution of $n^{1/2}\,(\hat\theta_j - \theta_j)$ is consistent for the normal distribution with zero mean and variance σ_j^2.

In this section we keep p fixed as we vary n, although we permit the distributions of the populations $\Pi_1, \ldots, \Pi_p$ to depend on n. Let $N_1, \ldots, N_p$ denote independent

standard normal random variables and, given constants $c_1, \ldots, c_p$, let $F_j(\,\cdot\,|\,c_1, \ldots, c_p)$ denote the distribution function of the random variables

$$1 + \sum_{k\,:\,k \neq j} I(\sigma_j\, N_j \leq \sigma_k\, N_k + c_k)\,.$$

The value of c_j has no influence on F_j, but it is cumbersome to reflect this in notation.

Theorem 5.1. *Assume that p is fixed and the datasets $\mathcal{X}_1, \ldots, \mathcal{X}_p$ are independent, that $\hat\theta_1^*, \ldots, \hat\theta_p^*$ are independent conditional on $\mathcal{X}$, and that* (5.6), (5.7) *(if using the m-out-of-n bootstrap) and* (5.10) *hold. (In* (5.10) *we take $m = n$ unless using the m-out-of-n bootstrap.) (i) For each integer r,*

$$P(\hat r_j \leq r) - F_j(r\,|\,\omega_{j1}, \ldots, \omega_{jp}) \to 0 \tag{5.11}$$

as $n \to \infty$. (ii) Using the standard n-out-of-n bootstrap, either parametric or nonparametric, define the ω_{jk}'s by (5.8). *Then there exists a sequence of random variables $Z_1, \ldots, Z_p$, depending on n and being, for each choice of n, independent and having the standard normal distribution, such that*

$$P(\hat r_j^* \leq r\,|\,\mathcal{X}) - F_j\Big(r \;\Big|\; \omega_{j1} + \sigma_1\, Z_1 - \sigma_j\, Z_j, \ldots, \omega_{jp} + \sigma_p\, Z_p - \sigma_j\, Z_j\Big) \to 0 \tag{5.12}$$

in probability as $n \to \infty$. (iii) In the case of the m-out-of-n bootstrap, again either parametric or nonparametric, and for which $m/n \to 0$ and $m \to \infty$, define the ω'_{jk}'s by (5.9). *Then* (5.12) *alters to:*

$$P(\hat r_j^* \leq r\,|\,\mathcal{X}) - F_j(r\,|\,\omega'_{j1}, \ldots, \omega'_{jp}) \to 0 \tag{5.13}$$

in probability as $n \to \infty$.

5.3.3 Interpretation of Theorem 5.1. To illustrate the implications of the theorem, let us assume that ω_{jk}, defined by (5.8), has (for each j and k) a well-defined limit (either finite or infinite) as $n \to \infty$, and that $\omega_{jk} \to +\infty$ for $k \in \mathcal{K}_+$, $\omega_{jk} \to -\infty$ for $k \in \mathcal{K}_-$, and ω_{jk} has a finite limit, ω_{jk}^0 say, for $k \in \mathcal{K}_j = \{1, \ldots, p\} \backslash (\{j\} \cup \mathcal{K}_+ \cup \mathcal{K}_-)$. (Both $\mathcal{K}_+$ and $\mathcal{K}_-$ may depend on j.) Define G_j to be the distribution function of

$$1 + (\#\mathcal{K}_+) + \sum_{k \in \mathcal{K}_j} I(\sigma_j\, N_j \leq \sigma_k\, N_k + \omega_{jk}^0)\,.$$

Then $F_j(r\,|\,\omega_{j1}, \ldots, \omega_{jp}) \to G_j(r)$, and so (5.11) implies that, as $n \to \infty$,

$$P(\hat r_j \leq r) \to G_j(r) \tag{5.14}$$

for each integer r.

Analogously to the argument leading from (5.11) to (5.14), result (5.12) implies that, in the case of the n-out-of-n bootstrap,

$$\begin{aligned} &P(\hat r_j^* \le r \,|\, \mathcal{X}) \text{ converges in distribution to the random variable}\,, \\ &P\Bigg[1 + (\#\mathcal{K}_+) + \sum_{k \in \mathcal{K}_j} I\Big\{\sigma_j\,(N_j + N_j') \\ &\qquad\qquad \le \sigma_k\,(N_k + N_k') + \omega_{jk}^0\Big\} \le r \;\Bigg|\; N_1, \ldots, N_p\Bigg]\,, \end{aligned} \tag{5.15}$$

where $N_1, \ldots, N_p, N_1', \ldots, N_p'$ are independent standard normal random variables. However, the convergence of $P(\hat r_j^* \le r \,|\, \mathcal{X})$ is not in probability.

If $\mathcal{K}_+ \cup \mathcal{K}_- = \{1, \ldots, p\} \setminus \{j\}$, which occurs (for example) if the θ_k's are fixed and there are no ties for the value of θ_j, then it follows from (5.11) and (5.12) that $P(\hat r_j = r_j) \to 1$ and $P(\hat r_j^* = r_j \,|\, \mathcal{X}) \to 1$ in probability, where r_j denotes the rank of θ_j in the set of all θ_k's. Therefore in this degenerate setting the standard n-out-of-n bootstrap correctly captures the asymptotic distribution of $\hat r_j$.

In all other cases, however, the limiting distribution of $\hat r_j$ (see (5.14)) does not equal the limit of the n-out-of-n bootstrap distribution of $\hat r_j^*$ (see (5.15)). Nevertheless, it is clear from (5.14) and (5.15) that:

> The support of the limiting distribution of $\hat r_j$, and the support of the weak limit of the distribution of $\hat r_j^*$ given $\mathcal{X}$, are identical, and both are equal to the set $\{\#\mathcal{K}_+ + 1, \ldots, \#\mathcal{K}_+ + \#\mathcal{K}_j + 1\}$.

(5.16)

To this extent the standard n-out-of-n bootstrap correctly captures important aspects of the distribution of $\hat r_j$.

Superficially, (5.13) seems to imply that the m-out-of-n bootstrap overcomes this problem. However, the ω_{jk}''s are now defined by (5.9), and are different from the ω_{jk}'s at (5.8). As a result, the m-out-of-n bootstrap does not, in general, correctly capture the limiting distribution at (5.14). Nevertheless, if

$$\text{for each } k \ne j, \text{ either } m^{1/2}\,(\theta_k - \theta_j) \to \pm\infty \text{ or } n^{1/2}\,(\theta_k - \theta_j) \to 0\,, \tag{5.17}$$

then $P(\hat r_j^* \le r \,|\, \mathcal{X}) - P(\hat r_j \le r) \to 0$ in probability, i.e. (5.4) holds. In particular, the m-out-of-n bootstrap consistently estimates the distribution of empirical ranks. Under condition (5.17) the following analogue of (5.16) holds for the m-out-of-n bootstrap:

> The limiting distributions of $\hat r_j$, and of $\hat r_j^*$ conditional on $\mathcal{X}$, are identical when using the m-out-of-n bootstrap, and the support of each equals the set $\{\#\mathcal{K}_+ + 1, \ldots, \#\mathcal{K}_+ + \#\mathcal{K}_j + 1\}$.

(5.18)

Property (5.17) holds if the θ_k's are all fixed (i.e. do not depend on n). Therefore, the m-out-of-n bootstrap correctly estimates the distribution of ranks in the presence of ties, when the populations are kept fixed as sample sizes diverge, and also in other cases where the differences $\theta_k - \theta_j$ are of either strictly larger order than $m^{-1/2}$ or strictly smaller order than $n^{-1/2}$. When (5.17) holds, the asymptotic distribution of $\hat{r}_j$ is supported on a set the size of $\#\mathcal{K}_j$, that is the number of integers k for which $m^{1/2}\,(\theta_k - \theta_j) \to 0$.

5.3.4 Methods for choosing m. Consider a comparison of two of the populations Π_j and Π_k, and focus on the probability of ranking one higher than the other using the m-out-of-n bootstrap. Assuming (5.10) and letting $c = (\sigma_j^2 + \sigma_k^2)^{-1/2}$, we see that

$$\begin{aligned} P(\hat{r}_j^* < \hat{r}_k^* | \mathcal{X}) - P(\hat{r}_j < \hat{r}_k) &= P(\hat{\theta}_j^* > \hat{\theta}_k^* | \mathcal{X}) - P(\hat{\theta}_j > \hat{\theta}_k) \\ &\approx \Phi\big\{m^{1/2}\,c\,(\hat{\theta}_j - \hat{\theta}_k)\big\} - \Phi\big\{n^{1/2}\,c\,(\theta_j - \theta_k)\big\} \\ &\approx \Phi\big\{m^{1/2}\,c\,(\theta_j - \theta_k) + c\,(m/n)^{1/2}\,Z\big\} - \Phi\big\{n^{1/2}\,c\,(\theta_j - \theta_k)\big\} \\ &= \Phi\big\{(m/n)^{1/2}\,(-c\,\omega_{jk} + Z)\big\} - \Phi(-c\,\omega_{jk})\,. \end{aligned}$$

Here Z denotes a realisation of a normal random variable, and Φ is the standard normal distribution function. Thus, choosing m to minimise the squared difference between the bootstrapped and true probabilities is approximately equivalent to choosing m to minimise the expression

$$\big[\Phi\big\{(m/n)^{1/2}\,(-c\,\omega_{jk} + Z)\big\} - \Phi(-c\,\omega_{jk})\big]^2\,. \tag{5.19}$$

If $\omega_{jk} \to \pm\infty$ then the expression is minimised as long as $(m/n)^{1/2}\,\omega_{jk} \to \pm\infty$ too, which guarantees that $m \to \infty$ as long as ω_{jk} is no larger than $O(n^{1/2})$. Alternatively if $\omega_{jk} \to 0$ then (5.19) is minimised provided $m/n \to 0$. This discussion motivates an approach for choosing m by tuning the bootstrapped probabilities to match the true probabilities. In reality however, we do not know ω_{jk}, c or Z so these must be estimated using $\hat{\omega}_{jk} = n^{1/2}\,(\hat{\theta}_k - \hat{\theta}_j)$, $\hat{c} = (\hat{\sigma}_j^2 + \hat{\sigma}_k^2)^{-1/2}$ and a random normal variable respectively. The situation is simplified if we have a "gap" between the orders of the diverging ω_{jk} and those converging, such as the following:

$$\text{For each pair } j, k, \text{ either } \omega_{jk} \to 0 \text{ or } |\omega_{jk}|(\log n)^{-1/2} \to \infty. \tag{5.20}$$

Thus we estimate m by choosing it to minimise the expression

$$\sum_{j,k;j\neq k} \int \Big(\Phi[(m/n)^{1/2}\{-\hat{c}\,\hat{\omega}_{jk}(\log n)^{-1/2} + z\}] \\ -\Phi\{-\hat{c}\,\hat{\omega}_{jk}(\log n)^{-1/2}\}\Big)^2 \phi(z)dz\,, \quad (5.21)$$

The following theorem, a proof of which is given in the PhD thesis of the second author, shows that choosing m in this fashion is consistent.

Theorem 5.2. *Assume p is fixed and that (5.6), (5.7), (5.10) and (5.20) hold. Choose m by minimising* (5.21)*. Then we have for each j:*

$$P(\hat{r}_j^* \leq r|\mathcal{X}) - P(\hat{r}_j \leq r) \to 0$$

in probability.

While this result suggests a way of determining m, there remains some uncertainty since the $(\log n)^{-1/2}$ factor used is not unique in generating good asymptotic performance. For example, replacing it with $(\log Cn)^{-1/2}$ for some constant C would yield a similar theoretic result. In practice, the dataset under consideration often suggests whether the adopted factor is appropriate, and the choice of m is reasonably robust against such changes.

5.3.5 Theoretical properties in the case of large p. The results above can be generalised to cases where p diverges with n but the support of the limiting distribution of $\hat{r}_j$ remains bounded. The defining features of those extensions are that values of $|\theta_k - \theta_j|$, for indices k that are not in the $\mathcal{K}_j$ of the previous section, should be at least as large as $(n^{-1}\log n)^{1/2}$; and values of $|\theta_k - \theta_j|$, for k in $\mathcal{K}_j$, should be at least as small as $n^{-1/2}$. We shall give results of this type in Section 5.4.2. In the present section we show how to capture, in a theoretical model, instances where both p and the support of the distribution of $\hat{r}_j$ are large. Real-data examples of this type are given by Goldstein and Spiegelhalter (1996).

Specifically, we assume the following linear model for θ_j:

$$\theta_j = a - \epsilon\, j \text{ for } 1 \leq j \leq p, \text{ where } a = a(n,p) \text{ does not depend on } j \text{ and } \epsilon = \epsilon(n) > 0. \quad (5.22)$$

This condition ensures the simple numerical ordering $\theta_1 > \ldots > \theta_p$, which in more general contexts we can impose without loss of generality. Assumption (5.22) also allows us to adjust the difficulty of the empirical ranking problem by altering the size of ϵ; the difficulty increases as ϵ decreases.

As in Theorem 5.1 we assume that the datasets $\mathcal{X}_j$ are independent, but now we

permit $p = p(n)$ to diverge with n. In order that Theorem 5.3 below may be stated relatively simply we assume that the quantities $Z_k = n^{1/2}\,(\hat\theta_k - \theta_k)$ all have the same asymptotic variance σ. Our main conclusion, that the standard n-out-of-n bootstrap correctly captures order of magnitude but not constant multipliers, remains valid as long as the limiting variances of the Z_k's are bounded away from zero and infinity.

We also assume conditions (5.23) and (5.24) below. In cases where each θ_j is a quantity such as a mean, a quantile, or any one of many different robust measures of location, those conditions follow from moderate-deviation properties of sums of independent random variables, provided the data have sufficiently many finite moments and p does not diverge too rapidly as a function of n:

$$\begin{aligned}&P\{n^{1/2}\,(\hat\theta_k - \theta_k) \le \sigma\,x\} = \Phi(x)\,\{1 + o(1)\} + o(p^{-1}\,n^{-1/2}\,\epsilon^{-1}),\\ &\text{uniformly in } |x| = O(p\,n^{1/2}\,\epsilon) \text{ and in } 1 \le k \le p, \text{ as } n \to \infty,\\ &\text{where } \sigma > 0;\end{aligned} \tag{5.23}$$

$$\begin{aligned}&P\{n^{1/2}\,(\hat\theta_k^* - \hat\theta_k) \le \sigma\,x \mid \mathcal{X}\} = \Phi(x)\,\{1 + o_p(1)\} + o_p(p^{-1}\,n^{-1/2}\,\epsilon^{-1}),\\ &\text{uniformly in } |x| = O(p\,n^{1/2}\,\epsilon) \text{ and in } 1 \le k \le p, \text{ as } n \to \infty, \text{ where}\\ &\sigma \text{ is as in (5.23).}\end{aligned} \tag{5.24}$$

In order for (5.23) and (5.24) to hold as p increases, the value of ϵ should decrease as a function of p, i.e. the empirical ranking problem should be made more difficult for larger values of p. Define $\delta = (n/2)^{1/2}\,\epsilon/\sigma$, where $\sigma > 0$ is as in (5.23) and (5.24), and put $\widetilde\omega_{jk} = n^{1/2}\,\{\hat\theta_k - \theta_k - (\hat\theta_j - \theta_j)\}/(2^{1/2}\,\sigma)$.

Theorem 5.3. *Assume that the datasets $\mathcal{X}_1, \ldots, \mathcal{X}_p$ are independent, that $\hat\theta_1^*, \ldots, \hat\theta_p^*$ are independent conditional on $\mathcal{X}$, that (5.22)–(5.24) hold, and that $p = p(n) \to \infty$ and $\epsilon = \epsilon(n) \downarrow 0$ as n increases, in such a manner that $n^{1/2}\,\epsilon \downarrow 0$ and $p\,n^{1/2}\,\epsilon \to \infty$. Then*

$$E(\hat r_j) = \delta^{-1} \int_{-j\delta}^{\infty} \Phi(-x)\,dx + o\big(\delta^{-1}\big)\,, \tag{5.25}$$

$$E(\hat r_j^* \mid \mathcal{X}) = \{1 + o_p(1)\} \sum_{k\,:\,k \ne j} \Phi\{\widetilde\omega_{jk} + \delta\,(j - k)\} + o_p\big(\delta^{-1}\big)\,, \tag{5.26}$$

uniformly in $1 \le j \le C/(n^{1/2}\,\epsilon)$ for any $C > 0$.

The implications of Theorem 5.3 can be seen most simply when j is fixed, although other cases are similar. For any fixed j it follows from (5.25) and (5.26) that

$$E(\hat r_j) \sim C\,\delta^{-1}\,, \tag{5.27}$$

$$E(\hat r_j^* \mid \mathcal{X}) \sim_p \delta^{-1} \int_{-\infty}^{\infty} d\Phi(z) \int_0^{\infty} \Phi\big(W_j - x - z\,2^{-1/2}\big)\,dx\,, \tag{5.28}$$

where $C = \int_{x>0} \Phi(-x)\,dx$, $W_j = -(n/2)^{1/2}\,(\hat\theta_j - \theta_j)/\sigma$, $a_n \sim b_n$ for constants a_n and b_n means that $a_n/b_n \to 1$, and $A_n \sim_p B_n$ for random variables A_n and B_n means that $A_n/B_n \to 1$ in probability. Results (5.27) and (5.28) reflect the highly variable character of $\hat r_j$ in the difficult cases represented by the model (5.22). For example, if $r_j = j$, which of course is fixed if j is fixed, then both $E(\hat r_j)$ and $E(\hat r_j^* \,|\, \mathcal{X})$ are of size δ^{-1}, which diverges to infinity as $n \to \infty$. That is, despite r_j being fixed, $\hat r_j$ tend to be so large that its expected value diverges. Similar arguments show that $\mathrm{var}(\hat r_j \,|\, \mathcal{X}_j)$ and $\mathrm{var}(\hat r_j^* \,|\, \mathcal{X}, \mathcal{X}_j^*)$ are both of size δ^{-1}.

It is clear from (5.27) and (5.28) that the standard n-out-of-n bootstrap correctly captures the order of magnitude, δ^{-1}, of $E(\hat r_j)$, but does not get the constant multiplier right. Similar arguments, based on elementary properties of sums of independent random variables, show that the standard bootstrap produces a prediction interval for $\hat r_j$ for which the length has the correct order of magnitude, but again the constant multiplier is not correct. The m-out-of-n bootstrap gets both the order of magnitude and the constant right, but at the expense of more restrictive conditions on ϵ; one could predict from Theorem 5.1 that this would be the case. It is also possible to establish a central limit theorem describing properties of $E(\hat r_j)$ and $E(\hat r_j^* \,|\, \mathcal{X})$. However, since the limitations of the bootstrap are clear at a coarser level than that type of analysis would address, then we shall not give those results here.

5.3.6 Numerical properties. We present numerical work which reinforces and complements the theoretical issues discussed above. In our first set of simulations we observe n independent data vectors $(X_1, \ldots, X_{10})$, where the X_j's are independent and respectively distributed as normal $\mathrm{N}(\theta_j, 1)$. First we consider the case where $\theta_j = 1 - (j/10)$, implying that the means are evenly spaced and do not depend on n. Although this model appears straightforward, the gaps between means are one tenth of the standard deviation of the noise, and so significant ranking challenges are present. However, Figure 5.1 shows that this is a case that the standard n-out-of-n bootstrap can handle satisfactorily, with the 90% prediction intervals for the estimated ranks shrinking as n grows.

Nevertheless our theory suggests that the n-out-of-n bootstrap will fail to correctly estimate the distribution in cases where the values of θ_j are relatively close. To investigate this issue we took $\theta_j = 1$ for $j \in \{1, 2, 3, 4, 5\}$, and $\theta_j = 0$ otherwise. Then, in our bootstrap replicates we would expect $\hat r_j^*$, conditional on the data, to be approximately uniformly distributed on either the top five positions (in the case $j \le 5$) or the bottom five (when $j \ge 6$). Figure 5.2 shows the difference in distributions for a simulation with $n = 1000$ and two choices of m. For each variable, the shading intensities in that column show the relative empirical distributions across ranks. Here the m-out-of-n bootstrap, with $m = 300$, produces distributions closer to the truth, where each of the top left and bottom right regions would have exactly

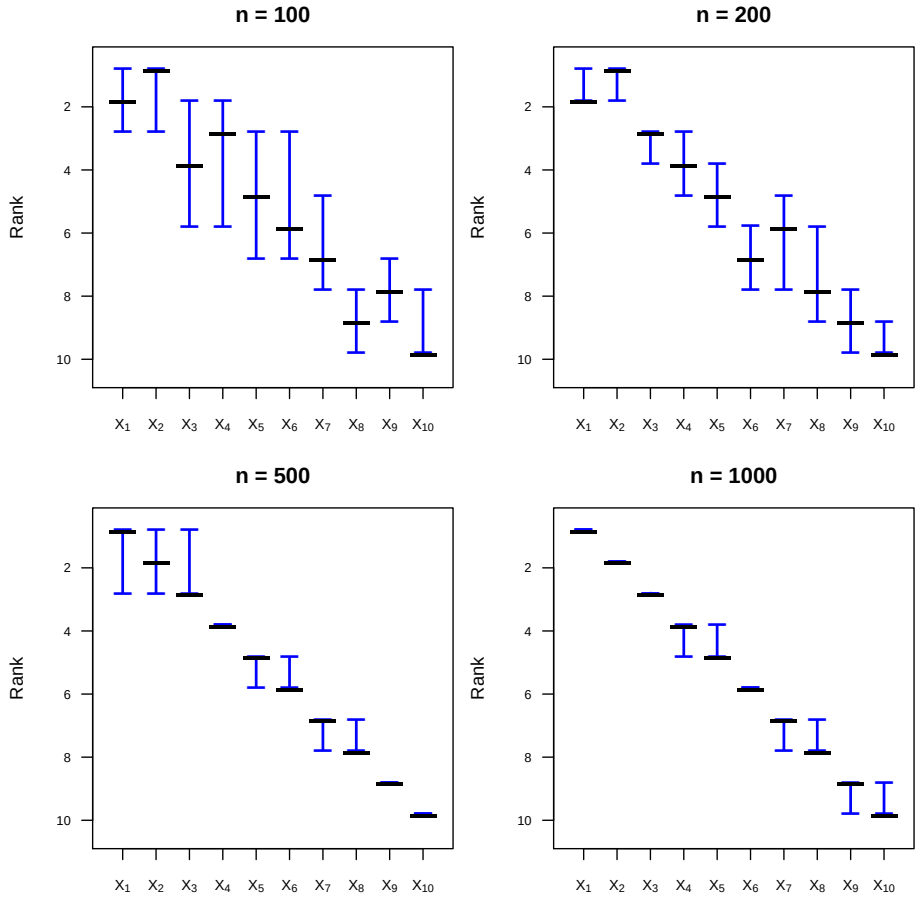


Figure 5.1: Ranking 90% prediction intervals for the case of fixed θ_j.

equal intensities everywhere.

The case of perfect ties demonstrates the advantages of the m-out-of-n bootstrap. In more subtle settings, when the θ_j's vary with n and are not exactly tied, we are interested in the ability of the bootstrap to distinguish θ_j's for which the absolute differences in $|\theta_j - \theta_k|$ are relatively small. The theory suggests considering differences of size $m^{-\alpha}$, where $\alpha = \frac{1}{2}$ is the critical value, lower values of α tend towards a (degenerate) perfect separation of ranks, and higher values asymptotically behave as though θ_j and θ_k were tied. Therefore the next set of simulations had the θ_j's equally spaced and uniformly decreasing, with $\theta_j - \theta_{j+1}$ equal to $0.2\,(10/m)^{\alpha}$. Here m was taken to be $\min(10n^{1/2}, n)$. Figure 5.3 shows, for a given pair (α, n), the average number of ranks contained within the 90% rank prediction interval. The results accord with the theory; cases where $\alpha < 0.5$ tend towards perfect separation (an average of 1), and cases where $\alpha > 0.5$ tend towards completely random ordering (an average of 10). Situations where α is closer to 0.5 diverge more slowly, and the behaviour when $\alpha = 0.5$ depends on the exact situation; in our simulations the degree of tuning has ensured that the case where $\alpha = 0.5$ does not show much

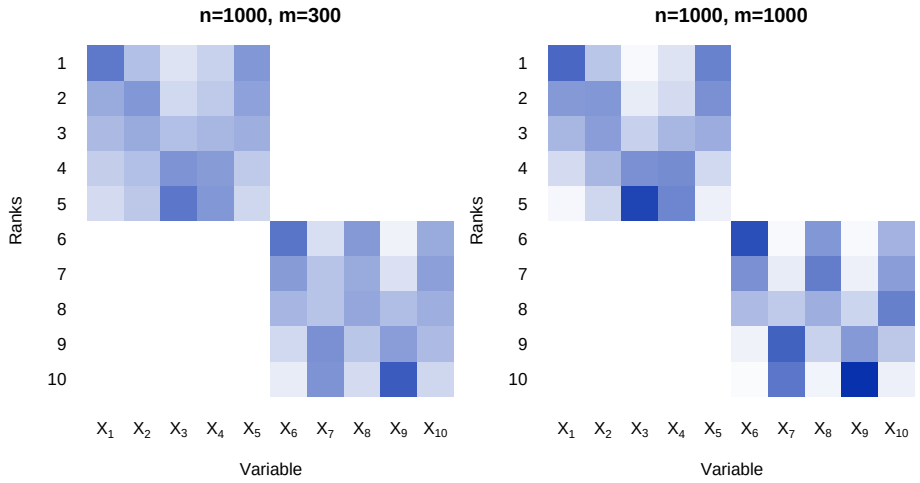


Figure 5.2: Distribution of ranks in the presence of ties.

tendency towards either extreme.

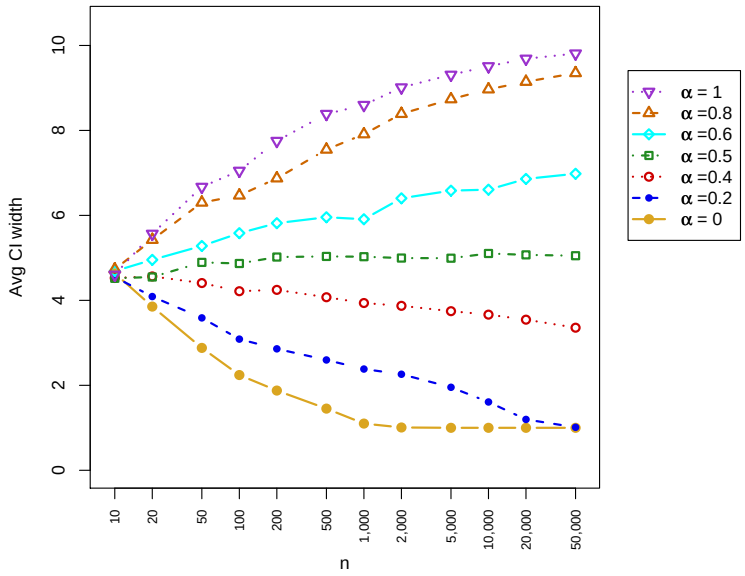


Figure 5.3: Behaviour of prediction interval widths for various α.

It is important to understand the distributional bias seen in the n-out-of-n bootstrap. One way this can be done is by exploring the distribution implied by (5.12). The distribution is dependent on the realisation of normal standard random variables $Z_1, \ldots, Z_p$. Figure 5.4 shows how the distribution of rankings varies with Z_1 for the special case of five variables, with $\omega_1 = \ldots = \omega_5 = 0$ and $Z_2 = Z_3 = Z_4 = Z_5 = 0$. Here, as $|Z_1|$ departs from 0, the ranking distribution is upset in two key ways. First, the average ranking is biased; for example, when $Z_1 = 1$ the average observed rank is 1.95 instead of 3, the average observed rank in the true underlying distribution

obtained when $Z_1 = 0$. Second, the variation of the observed rank is reduced; the variance is 1.4 when $|Z_1| = 1$ compared with 2 in the true distribution. These two effects combine to give overconfidence in the n-out-of-n bootstrap when it is not warranted.

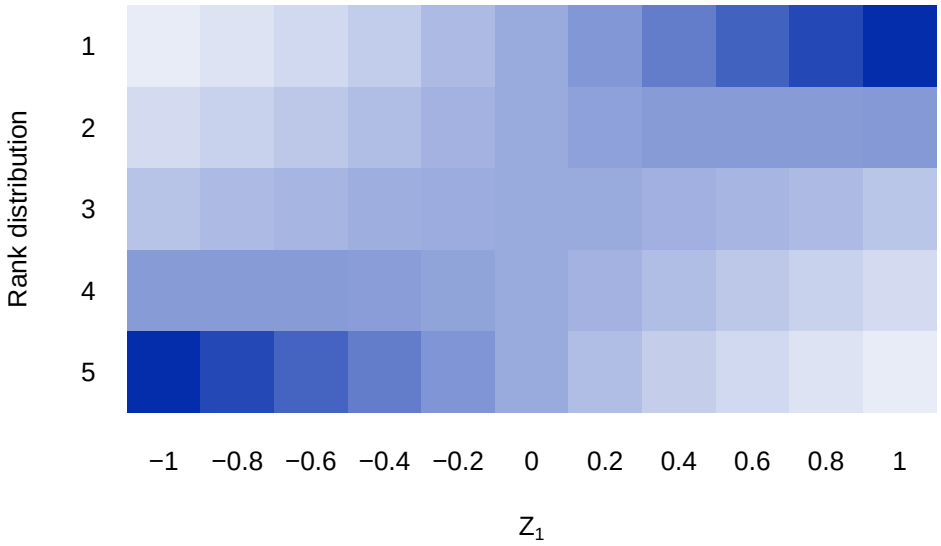


Figure 5.4: Distribution of ranks for various Z_1.

We now move to a real-data example. A service seeking to assist parents to choose secondary schools in the state of NSW, Australia, ranks 75 schools using the number of credits achieved in final year Higher School Certificate exams as a percentage of the number of exams sat. While there are clearly significant problems with such a simple statistic (see Goldstein and Spiegelhalter, 1996), the main one being that it ignores prior student ability, it would still be useful to give some indication of the variability of the rankings. Here n_j represents the number of exams sat at school j, and X_{ij}, for $1 \leq i \leq n_j$, is an indicator variable for whether a credit was achieved in exam i. Then $\theta_j = E(X_{ij})$ and $\hat{\theta}_j = n_j^{-1} \sum_i X_{ij}$. Figure 5.5 shows 95% prediction intervals for the ranks using the n-out-of-n bootstrap. It is clear that caution needs to be exercised when interpreting the intervals, the average width of which exceeds 14 places. However, we know that the n-out-of-n bootstrap ranking understates the true uncertainty, which would be better captured using the m-out-of-n bootstrap. Figure 5.6 shows the results using $m_j = \lfloor n_j \times 35.5\% \rfloor$. The percentage here was chosen using the approach discussed in Section 5.3.4, attempting to minimise the squared error between the bootstrap and real ranking distributions. Observe that the widths of the prediction intervals are now markedly longer (58% longer on average); the widest confidence interval now covers 81% of the possible rankings. Our theoretical results argue that these longer widths give a better indication of the true uncertainty associated with the ranking.

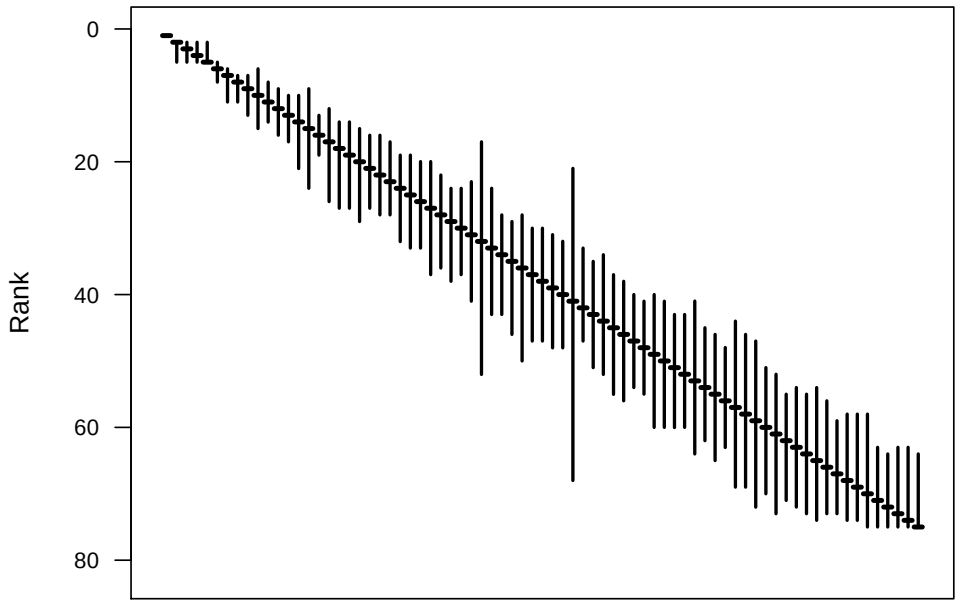


Figure 5.5: School ranking prediction intervals for n-out-of-n bootstrap.

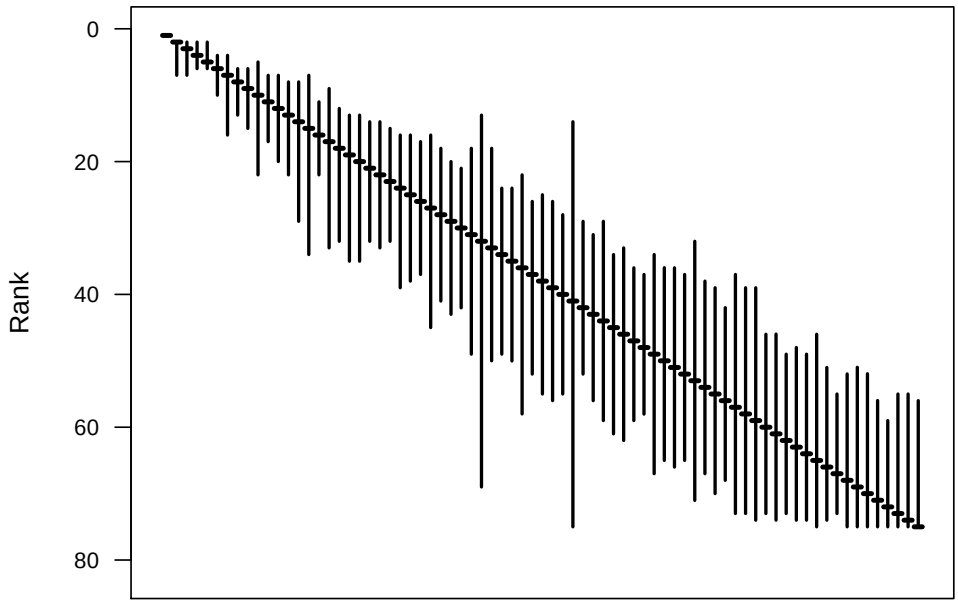


Figure 5.6: School ranking prediction intervals for m-out-of-n bootstrap with m_j equal to 35.5% of n_j.

5.4 Properties in cases where the data come as independent p-vectors

5.4.1 Motivation for the independent-component bootstrap. In this section we argue that, when vector components are not strongly dependent, the standard, "synchronous" bootstrap may distort relationships among components, particularly in the setting of conditional inference and when p is large. In such cases, even if the assumption of independent components is not strictly correct, it may be advantageous to apply the bootstrap as though independence prevailed. We refer to this working assumption as that of "component-wise independence."

We treat the case where the data arise via a sample $\mathcal{X}_0 = \{X_1, \ldots, X_n\}$ of independent p vectors. Here $X_i = (X_{i1}, \ldots, X_{ip})$, and $\mathcal{X}_j = \{X_{1j}, \ldots, X_{nj}\}$ denotes the set of jth components. The conventional, synchronous form of the nonparametric bootstrap involves the following resampling algorithm:

> Draw a resample $\mathcal{X}_0^* = \{X_1^*, \ldots, X_m^*\}$ by sampling randomly, with replacement, from $\mathcal{X}_0$, write $X_i^* = (X_{i1}^*, \ldots, X_{ip}^*)$ and take $\mathcal{X}_j^* = \{X_{1j}^*, \ldots, X_{mj}^*\}$. (5.29)

We can view $\mathcal{X}_j^*$ as the resample drawn from the jth "population." In (5.29) we take $m \le n$, thereby allowing for the m-out-of-n bootstrap.

We argue that this bootstrap method is not always satisfactory in problems where ranking is involved. One reason is that:

> If the data have a continuous distribution then knowing the dataset $\mathcal{X}_j^*$ conveys perfect information about which data vectors X_i are included in $\mathcal{X}_0^*$, defined in (5.29), and with what frequencies. Hence, knowing $\mathcal{X}_j^*$ tells us $\mathcal{X}_k^*$ for each k, and in particular the resamples $\mathcal{X}_1^*, \ldots, \mathcal{X}_p^*$ cannot be regarded as independent, conditional on $\mathcal{X}_0$, even if the vector components are independent. (5.30)

This result holds for the m-out-of-n bootstrap as well as for the standard, synchronous bootstrap, and so the problems to which it leads cannot be alleviated simply by passing to a smaller resample size.

To elucidate the consequences of (5.30), note that the jth empirical rank $\hat{r}_j$, and its bootstrap version $\hat{r}_j^*$, can be written as

$$\hat{r}_j = 1 + \sum_{k\,:\,k \ne j} I\bigl(\hat{\theta}_j \le \hat{\theta}_k\bigr)\,, \quad \hat{r}_j^* = 1 + \sum_{k\,:\,k \ne j} I\bigl(\hat{\theta}_j^* \le \hat{\theta}_k^*\bigr)\,, \tag{5.31}$$

respectively. Here, $\hat{r}_j$ and $\hat{r}_j^*$ are as at (5.2) and (5.3). We wish to estimate aspects of the distribution of $\hat{r}_j$. For example, we might seek an estimator of the variance of the conditional mean, $\hat{u}_j = E(\hat{r}_j \,|\, \mathcal{X}_j) = 1 + \sum_{k\,:\,k\neq j} \pi_{jk}$, of $\hat{r}_j$ given $\mathcal{X}_j$; or we might wish to approximate the variance of $\hat{r}_j$. (To derive the formula for $\hat{u}_j$ we used the first part of (5.31) and took $\pi_{jk} = P(\hat{\theta}_j \le \hat{\theta}_k \,|\, \mathcal{X}_j)$.) Undertaking conditional inference is particularly attractive in problems where p is large, because it has the potential to greatly reduce variability, from $O(p^2)$ (the order of the unconditional variance of $\hat{r}_j$) to $O(p)$ (the order of the variance of $\hat{r}_j$, conditional on $\mathcal{X}_j$, if the components are sufficiently weakly dependent).

The bootstrap version of $\hat{u}_j$ can be computed using the second formula in (5.31): $\hat{u}_j^* = E(\hat{r}_j^* \,|\, \mathcal{X}, \mathcal{X}_j^*) = 1 + \sum_{k\,:\,k\neq j} \hat{\pi}_{jk}^*$, where $\mathcal{X} = \cup_k \mathcal{X}_k$ and $\hat{\pi}_{jk}^* = P(\hat{\theta}_j^* \le \hat{\theta}_k^* \,|\, \mathcal{X}, \mathcal{X}_j^*)$. If we use the synchronous bootstrap algorithm at (5.29) then it follows from (5.30) that $\hat{\pi}_{jk}^* = I(\hat{\theta}_j^* \le \hat{\theta}_k^*)$. Since the probability has degenerated to an indicator function then, even when using the m-out-of-n bootstrap, and in the conventional setting of fixed p and increasing n, $\mathrm{var}(\hat{u}_j^* \,|\, \mathcal{X}) - \mathrm{var}(\hat{u}_j)$ fails to converge to zero except in degenerate cases.

The errors can become still more pronounced if p diverges with n. Indeed, in the problem of estimating

$$\mathrm{var}(\hat{u}_j) = \sum_{k_1\,:\,k_1\neq j} \sum_{k_2\,:\,k_2\neq j} \mathrm{cov}(\pi_{jk_1}\,, \pi_{jk_2})$$

using

$$\mathrm{var}(\hat{u}_j^* \,|\, \mathcal{X}) = \sum_{k_1\,:\,k_1\neq j} \sum_{k_2\,:\,k_2\neq j} \mathrm{cov}(\pi_{jk_1}^*\,, \pi_{jk_2}^* \,|\, \mathcal{X})\,, \tag{5.32}$$

and in the context of component-wise independence, the synchronous bootstrap at (5.29) introduces correlation terms of size $n^{-1/2}, n^{-1}, \ldots$; those terms would be zero if the bootstrap algorithm correctly reflected component-wise independence. If p is much larger than n then the impact of the extraneous terms is magnified by the summations over k_1 and k_2 in (5.32). These problems, too, persist when employing the m-out-of-n bootstrap.

The situation improves significantly if, instead of using the synchronous bootstrap at (5.29), we employ the following independent-component resampling algorithm:

> Compute $\mathcal{X}_j^* = \{X_{1j}^*, \ldots, X_{mj}^*\}$ by sampling randomly, with replacement, from $\mathcal{X}_j = \{X_{1j}, \ldots, X_{nj}\}$; and do this independently for each j. (5.33)

In this case, when using the m-out-of-n bootstrap and working under the assumption of component-wise independence, $\mathrm{var}(\hat{u}_j^* \,|\, \mathcal{X}) - \mathrm{var}(\hat{u}_j)$ converges to zero as n diverges, and the undesirable $n^{-1/2}$ terms that arise when estimating $\mathrm{var}(\hat{u}_j)$, using the syn-

chronous bootstrap, vanish. To summarise, under component-wise independence the independent-component bootstrap, defined at (5.33), corrects for significant errors that can be committed by the synchronous bootstrap algorithm at (5.29).

Importantly, similar conclusions are also reached in cases where p is large and the component vectors $(X_{1j}, \ldots, X_{nj})$ are not independent. In particular, if the dependence among components is sufficiently weak to ensure that the asymptotic distribution of $\hat{r}_j$ is identical to what it would be if the components were independent, then the independent-component bootstrap has obvious attractions. For example, in inferential problems involving conditioning on $\mathcal{X}_j$, it gives statistical consistency in contexts where the synchronous bootstrap does not. This can happen even under conditions of reasonably strong dependence, simply because the highly ranked components are lagged well apart. Details will be outlined in the first paragraph of Section 5.4.3, after exploring some theoretical properties of the approach.

5.4.2 Theoretical properties. We address only the j_0 highest-ranked populations, which for notational convenience we take to be those with indices $j = 1, \ldots, j_0$, and we take the ranks of these populations to be virtually tied, so that the limiting distribution of $\hat{r}_j$ is nondegenerate. Also, we allow both p and the distributions of $\Pi_1, \ldots, \Pi_p$ to depend on n. In particular, we assume that:

$$n^{1/2}\,(\theta_1 - \theta_j) \to 0 \quad \text{for} \quad j = 1, \ldots, j_0\,, \tag{5.34}$$

$$p = o\big(n^{C_1}\big) \quad \text{for some} \quad C_1 > 0\,. \tag{5.35}$$

To determine the limiting distribution of $\hat{r}_j$ we further suppose that:

$$(n/\log n)^{1/2} \inf_{j_0 < j \le p} (\theta_1 - \theta_j) \to \infty\,, \tag{5.36}$$

and

$$\begin{aligned}&\text{the random variables } n^{1/2}\,(\hat{\theta}_j - \theta_j), \text{ for } 1 \le j \le j_0, \text{ are asymptotically independent and normally distributed with zero means and respective variances } \sigma_j^2; \text{ and, for } C_2 > 0 \text{ sufficiently large,}\\ &\sup_{j \le p} P\{|\hat{\theta}_j - \theta_j| > C_2\,(n^{-1}\,\log n)^{1/2}\} = O(n^{-C_1}).\end{aligned} \tag{5.37}$$

When discussing the efficacy of the m-out-of-n bootstrap we ask, instead of (5.34), (5.36) and (5.37), that:

$$m^{1/2}\,(\theta_1 - \theta_j) \to 0 \quad \text{for} \quad j = 1, \ldots, j_0\,, \tag{5.38}$$

$$(m/\log m)^{1/2} \inf_{j_0 < j \le p} (\theta_1 - \theta_j) \to \infty\,, \tag{5.39}$$

conditional on $\mathcal{X}$, the random variables $m^{1/2}\,(\hat\theta_j^* - \hat\theta_j)$, for $1 \le j \le j_0$, are asymptotically independent and normally distributed with zero means and respective variances σ_j^2; and for $C_2 > 0$ sufficiently large, $\sup_{j \le p} P\{|\hat\theta_j^* - \hat\theta_j| > C_2\,(m^{-1}\,\log m)^{1/2}\} = O(n^{-C_1})$. (5.40)

For example, the last parts of (5.37) and (5.40) hold if θ_j and $\hat\theta_j$ are respectively population and sample means, if the associated population variances are bounded away from zero, and if the supremum over j of absolute moments of order C_3, for the population Π_j, is bounded for a sufficiently large $C_3 > 0$ (see Section 1.6). Likewise, (5.37) and (5.40) also apply in cases where each θ_j is a quantile or any one of many different robust measures of location. The first part of (5.37) is a standard central limit theorem for the estimators $\hat\theta_j$, and so is a weak assumption. In (5.40) we do not specify using the independent-component bootstrap (see (5.33)), but if we do impose that condition then the first part of (5.40) is a conventional central limit theorem for the m-out-of-n bootstrap, and in that setting we do not need to assume independence of the asymptotic normal distribution of the variables $m^{1/2}\,(\hat\theta_j^* - \hat\theta_j)$; it follows from the nature of the independent-component bootstrap.

Theorem 5.4. *Let $1 \le j \le j_0$. (i) If* (5.34)–(5.37) *hold then the ranks $\hat r_1, \ldots, \hat r_{j_0}$ are asymptotically jointly distributed as $R_1, \ldots, R_{j_0}$, where*

$$R_j = 1 + \sum_{k\,:\,k \le j_0,\, k \ne j} I(Z_j\,\sigma_j \le Z_k\,\sigma_k) \tag{5.41}$$

and $Z_1, \ldots, Z_{j_0}$ are independent and normal $N(0,1)$. (ii) Assume (5.35) *and* (5.37)–(5.40)*, and use the m-out-of-n bootstrap (where $m/n \to 0$ and $m \to \infty$ as $n \to \infty$), in either the conventional form at* (5.30) *or the component-wise from at* (5.33)*. Then the distribution of $(\hat r_1^*, \ldots, \hat r_{j_0}^*)$, conditional on the data $\mathcal{X}$, converges in probability to the distribution of $(R_1, \ldots, R_{j_0})$. (iii) Assume* (5.35) *and* (5.37)–(5.40)*, use the m-out-of-n bootstrap with $m/n \to 0$ and $m \to \infty$ as $n \to \infty$, and implement the bootstrap component-wise, as in* (5.33)*. Then the distribution of $\hat u_j^*$, conditional on $\mathcal{X}$, is consistent for that of $\hat u_j$. That is,*

$$P\{E(\hat r_j^* \,|\, \mathcal{X}, \mathcal{X}_j^*) \le x \,|\, \mathcal{X}\} \to P\{E(R_j \,|\, Z_j) \le x\} \tag{5.42}$$

in probability, for all continuity points x of the cumulative distribution function $P\{E(R_j \,|\, Z_j) \le x\}$. Moreover, $\mathrm{var}(\hat r_j^* \,|\, \mathcal{X}) \to \mathrm{var}(R_j)$.

5.4.3 Discussion. The assumptions underpinning Theorem 5.4 do not require the components of the data vectors $X_i = (X_{i1}, \ldots, X_{ip})$ to be independent, but they do ask that the empirical ranks $\hat\theta_j$, corresponding to the true θ_j's that are virtually

tied for the top j_0 positions, be asymptotically independent. Refer to the first part of (5.37). That condition holds in many problems where p is diverging but the components are strongly dependent, for example when θ_j is a mean and the common distribution of the vectors X_i is determined by adding θ_j's randomly to centred, although potentially strongly dependent, noise. For example, if the components of the noise process are q-dependent, where the integer q is permitted to diverge with increasing n and p, then in the case of fixed j_0 explored in Theorem 5.4, sufficient independence is ensured by the condition that $q/p \to 0$ as $p \to \infty$.

Parts (i) and (ii) of Theorem 5.4 together imply that (5.18) continues to hold in the present setting, provided j is in the range $1 \le j \le j_0$.

As noted in Section 5.4.1, the result in the first part of Theorem 5.4(iii) does not hold if the synchronous bootstrap is used. Likewise, while the independent-component, m-out-of-n bootstrap can be proved to consistently estimate the distribution of $\text{var}(\hat{r}_j \,|\, \mathcal{X}_j)$, neither the n-out-of-n bootstrap nor its m-out-of-n bootstrap form give consistency if applied using the conventional resampling algorithm at (5.29). The same challenges arise for a variety of other estimation problems; the problems treated in Theorem 5.4(iii) are merely examples.

In cases where p is very much larger than n, and the aim is to discover information concealed in a very high-dimensional dataset, choosing m for the m-out-of-n bootstrap might best be regarded as selecting the level of sensitivity rather than as choosing the level of smoothing in a more conventional, m-out-of-n bootstrap sense. Since the desired level of sensitivity depends on the unknown populations Π_j, and, in the most important marginal cases, is unknown, then it may not always be appropriate to use a standard empirical approach to choosing m. Instead, numerical results for different values of m could be obtained.

Results analogous to Theorem 5.3 can also be established in the present setting. In particular, in cases where $\hat{r}_j$ is highly variable, the standard n-out-of-n bootstrap correctly captures the order of magnitude, but not the constant multiplier, of characteristics of the distribution of $\hat{r}_j$, for example its expected value and the lengths of associated prediction intervals.

5.4.4 Numerical properties. To gain insight into the advantages of the independent-component bootstrap we consider the following setting: suppose we have p variables and n observations, and the jth variable X_j is modelled by $X_j = \theta_j + Z_j$, where θ_j is a constant, Z_j is a standard random normal variable, $\text{cor}(Z_j, Z_k) = \rho_n$ when $j \ne k$, and ρ_n decreases to 0 as n increases. We wish to compare performance of the standard and independent-component bootstraps in the task of ranking the

values of θ_j. As our performance measure we use the squared error criterion:

$$\sum_j \sum_r E\{P(\hat{r}_j^* = r \,|\, \mathcal{X}) - P(\hat{r}_j = r)\}^2 .$$

Figure 5.7 gives results for $n = 50$, $p = 200$ and $\theta_j = 1-\{j/(p-1)\}$, for various choices of ρ_n. It shows that the independent-component bootstrap consistently improves performance. Interestingly, performance of the independent-component case is at its best when a reasonable level of correlation present. This is apparently because, in the presence of correlation, the true ranking distribution becomes more "lumpy" or more degenerate.

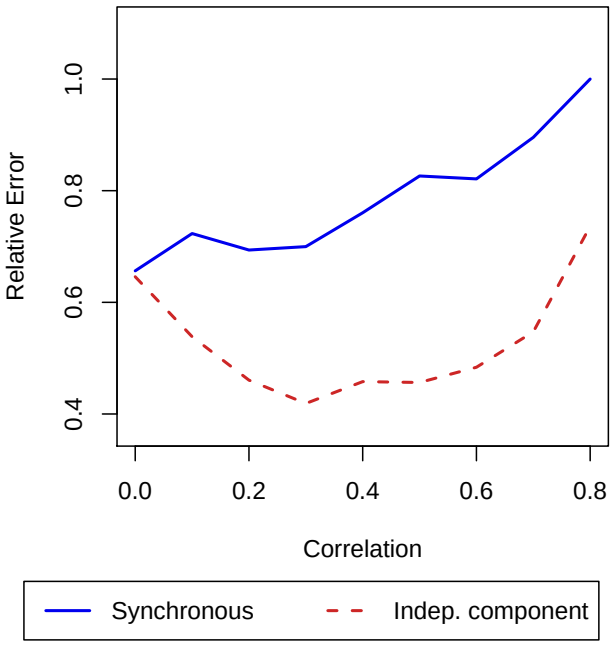


Figure 5.7: Relative error of synchronous and independent-component bootstrap distributions.

The Ro131 dataset was used by Segal et al. (2003) to compare a variety of genomic approaches, and was introduced in Sections 2.2.1 and 2.4.1. There, generalised correlation was measured between the observed Y and each set of gene expressions $\mathcal{X}_j$ in order to rank the genes. The results of the synchronous bootstrap to give indicative prediction intervals for these rankings was given in Figure 2.2 for the top 15 variables, and is reproduced for convenience in Figure 5.8. It should be observed that significant levels of correlation exist between pairs of influential genes. There are at least two possible reasons for this. First, if gene expression levels closely follow the movements of response variables then genes will share some of this correlation indirectly. Secondly, there may be intrinsic correlation between two genes if they are controlled by some common underlying process.

If the first reason is suspected to be the dominant one then the independent-component bootstrap should give a better indication of uncertainties in ranking. Figure 5.9 depicts results for the independent-component bootstrap. Notice that prediction interval widths are greater than in the synchronous case. This is because the positive correlations among values of $\hat{\theta}_j$ in the synchronous case reduce the variations in rankings.

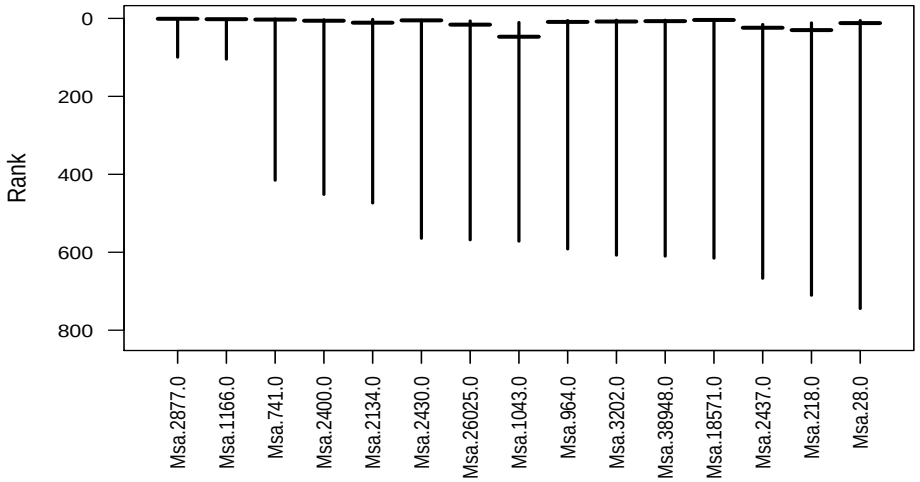


Figure 5.8: Synchronous bootstrap results for Ro131 dataset.

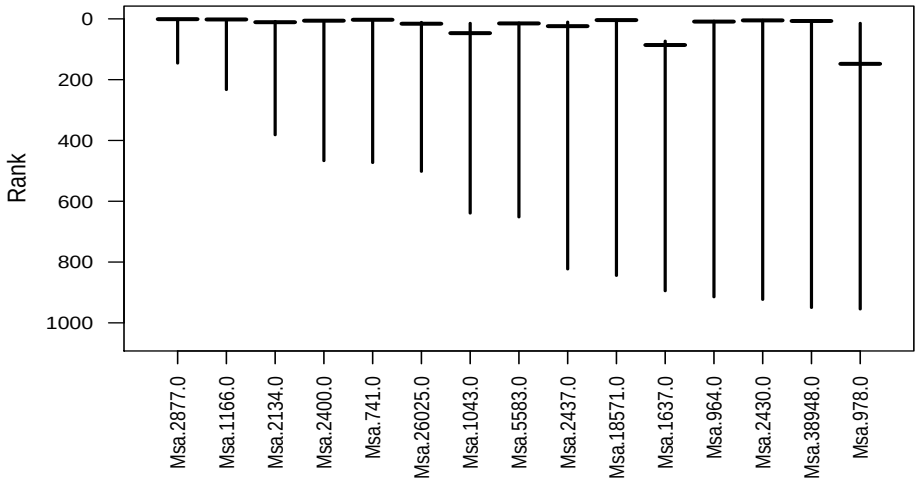


Figure 5.9: Independent-component bootstrap results for Ro131 dataset.

Another plot that is useful in understanding rankings is that of conditional rankings, the subject of Theorem 5.4. Figure 5.10 shows the rankings for the top genes, together with prediction intervals for $\hat{r}_j^*$, conditional on both $\mathcal{X}$ and $\hat{\theta}_j^*$. Thus, for a given gene we have held the observed generalised correlation for it constant and bootstrapped on all other genes, to estimate how the genes should be ranked given

the value of $\hat{\theta}_j$. The results for this analysis are highly dependent on whether the bootstrap is performed synchronously or independently. For reasons given in Sections 5.4.1–5.4.3 we prefer the independent-component bootstrap in this situation. Figure 5.10 displays the corresponding prediction intervals. Two features of the results are striking. First, the prediction intervals are very narrow compared to those seen in Figures 5.8 and 5.9, highlighting that the fact that most of the uncertainty in ranking the jth gene comes from the uncertainty of $\hat{\theta}_j$ itself. Secondly, the prediction intervals lie below the actual point estimate for the rank. This suggests that if the experiment were performed again we would be unlikely to see the top-ranked variables rank as highly as before. In fact, we would expect the top variable to rank outside the top twenty, even if it appeared as strongly as it did in our observed data. These two observations are interesting, and highlight the challenges of variable selection in such high-dimensional settings.

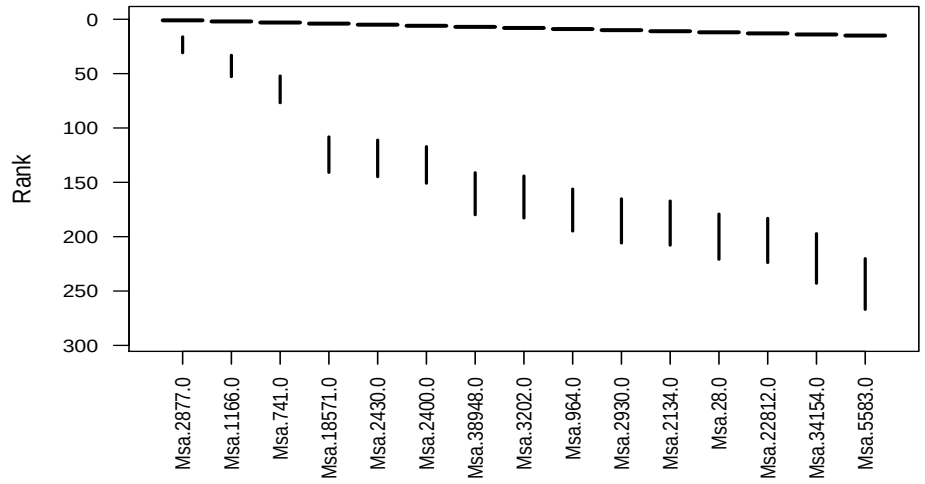


Figure 5.10: Independent reverse synchronous bootstrap results for Ro131 dataset.

We reiterate here one observation relevant to both the independent-component bootstrap and the discussion of the m-out-of-n bootstrap in Section 5.3. When constructing prediction intervals for ranks, the method that produces the shortest intervals is not necessarily the most powerful or the most accurate. Both the theoretical and numerical results suggest that the synchronous bootstrap will produce widths that are too narrow compared to the theoretical ranking distribution; the bootstrap ranks become "anchored" to the observed empirical ranks. Thus, interpreting ranking sensitivities for real datasets involves attempting to balance both maximising the power of an approach with the risks of overstating ranking accuracy. In many cases it will be simulation and experimentation that suggest the best balance in a given situation.

The final example comprises of a set of simulations that illustrate the results of

Theorem 5.4 in a high-dimensional setting. The aim here is to estimate the correct distribution for the top five ranked variables. For each of six scenarios, we start with the base case of $n = 20$, $p = 500$, which was constructed as follows: the mean is once more the statistic of interest. Each data point X_{ij} is normal with standard deviation 0.25 and the jth mean is $\theta_j = 1$ for $j = 1, \ldots, 5$ and is randomly sampled from the uniform distribution over $[0, 0.9]$ when $j > 5$. Once the data is generated, we may derive the ranking distribution using the independent component bootstrap with $m = 20$. We use the statistic

$$\text{Error} = \sum_{j=1}^{5} \sum_{r=1}^{5} \{P(\hat{r}^* \leq r|\mathcal{X}) - P(\hat{r} \leq r)\}^2 ,$$

to measure how accurately the rankings for the first five variables are estimated. Notice that $P(\hat{r} \leq r) = r/5$ for $r = 1, \ldots, 5$ and that the error statistic is 0 if and only if this distribution is matched exactly in the bootstrapped distribution. We repeat this experiment 100 times and report the average error along with 90% confidence intervals for this average. From here the simulation grows by increasing n and increasing m at rate $n/log(n)$. In each scenario p is constant or grows at a linear or quadratic rate relative to n. Also, the gap between the mean of the top five variables and the upper range of the uniform sampling distribution is either left constant or shrunk at a square rooted logarithmic rate. This results in six scenarios, the results of which are plotted in Figure 5.11. The error has been scaled so that 100 denotes maximum possible error. Observe that the quadratic growth simulations in particular achieve very high dimensions; when $n = 140$, $p = 24,500$, which is competitive with the dimensionality for many genomic applications.

Theorem 5.4 establishes that under each of these scenarios the distribution of the top five variables should be estimated correctly, since p increases only polynomially and the gap is either constant or shrinks sufficiently slowly; compare with (5.35), (5.40). The results reinforce these findings, with error steadily decreasing in all cases except the quadratic ones. In these final cases, the error increases briefly until the stability of the means outweighs the effects of the increasing p and decreasing gap. The error then steadily decreases, albeit at a much slower rate than the constant and linear scenarios. We can see that the problem is noticeably more difficult when the gap shrinks, as well as when p grows at a faster rate.

This example was constructed to demonstrate that the theoretical results can hold while the data size remained computable. However, there are instances where very large n are needed before such distributional accuracy is obtained. For instance, if we tripled the standard deviation in final scenario, where we have quadratic growth in p and a shrinking gap, we would require $n >$1,000 before the error started to decrease and satisfactory results were obtained. In this case p would be over one million,

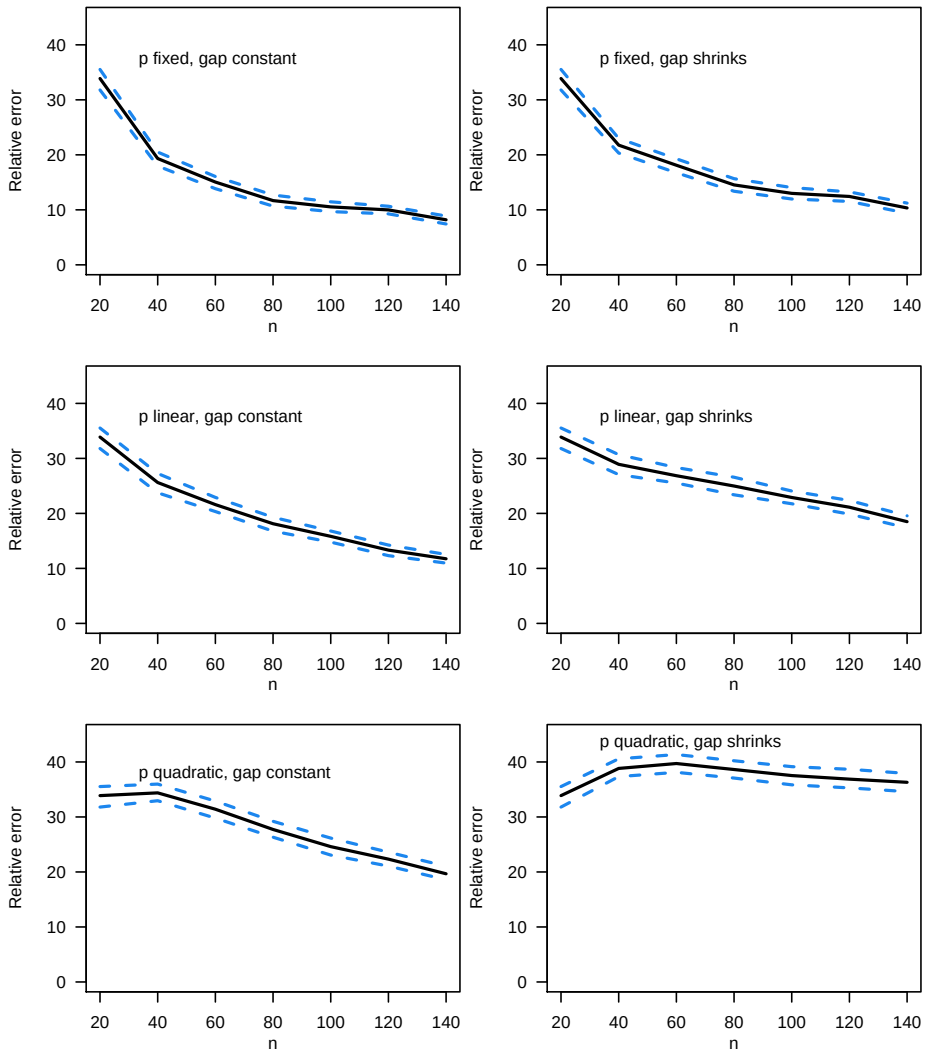


Figure 5.11: Average error with 90% confidence intervals for $p > n$ simulations.

which is in excess of feasible desktop computation.

5.5 Technical arguments

5.5.1 Proof of Theorem 5.1. (i) In view of the first part of (5.10) we may write

$$\hat{r}_j = 1 + \sum_{k\,:\,k\neq j} I\big(\hat{\theta}_j \le \hat{\theta}_k\big) = 1 + \sum_{k\,:\,k\neq j} I(\sigma_j\,\Delta_j \le \sigma_k\,\Delta_k + \omega_{jk})\,, \tag{5.43}$$

where the random variables $\Delta_k = n^{1/2}\,(\hat{\theta}_k - \theta_k)/\sigma_k$ are jointly independent and asymptotically standard normal. Result (5.11) can be proved from this quantity by considering the respective cases where values in the sequence ω_{jk}, for $1 \le k \le p$, are finite or infinite.

(ii) To derive (5.12) we note that, in view of the second part of (5.10),

$$\begin{aligned} \hat{r}_j^* &= 1 + \sum_{k\,:\,k \neq j} I\bigl(\hat{\theta}_j^* \leq \hat{\theta}_k^*\bigr) \\ &= 1 + \sum_{k\,:\,k \neq j} I\bigl(n^{-1/2}\,\sigma_j\,\Delta_j^* \leq n^{-1/2}\,\sigma_k\,\Delta_k^* + \hat{\theta}_k - \hat{\theta}_j\bigr) \\ &= 1 + \sum_{k\,:\,k \neq j} I\bigl(\sigma_j\,\Delta_j + \sigma_j\,\Delta_j^* \leq \sigma_k\,\Delta_k + \sigma_k\,\Delta_k^* + \omega_{jk}\bigr)\,, \end{aligned} \tag{5.44}$$

where, conditional on $\mathcal{X}$, the random variables $\Delta_k^* = n^{1/2}\,(\hat{\theta}_k^* - \hat{\theta}_k)/\sigma_k$ are jointly independent and asymptotically standard normal, and the Δ_k's are as in (5.43). Since, by the first part of (5.10), the Δ_k's are asymptotically independent and standard normal (in an unconditional sense), then, by Kolmogorov's extension theorem, we can (on a sufficiently large probability space) find random variables $Z_1, \ldots, Z_p$ which depend on n, are exactly independent and exactly standard normal for each n, and have the property that $\Delta_k = Z_k + o_p(1)$ for each k, as $n \to \infty$. Result (5.12) follows from these properties and (5.44).

(iii) Result (5.44) continues to hold in the case of the m-out-of-n bootstrap, except that to obtain the arguments of the indicator functions there we have to multiply throughout by $m^{1/2}$ rather than $n^{1/2}$. This means that to interpret (5.44) we should redefine $\Delta_k = m^{1/2}\,(\hat{\theta}_k - \theta_k)/\sigma_k$ and $\Delta_k^* = m^{1/2}\,(\hat{\theta}_k^* - \hat{\theta}_k)/\sigma_k$. Since $m/n \to 0$ then, on the present occasion, $\Delta_k \to 0$ in probability for each k, but, in view of the second part of (5.10), the conditional distribution of Δ_k^* continues to be asymptotically normal $\mathrm{N}(0, \sigma_k^2)$. Result (5.13) now follows from (5.44).

5.5.2 Proof of Theorem 5.2. Assume first that there are some ω_{jk} that tend to 0 and some that tend to $\pm\infty$ (the remaining cases are discussed towards the end of the proof). Observe that (5.10) ensures that when $\omega_{jk} \to \pm\infty$, $\hat{\omega}_{jk}$ will do so at the same rate, while when $\omega_{jk} \to 0$, we have $\hat{\omega}_{jk} b_n \to 0$ at rate of b_n. Since p is fixed, we can choose a positive sequence $a_n \to \infty$ to be the slowest of the diverging $|\omega_{jk}|$ (that is, those ω_{jk} that tend to $\pm\infty$ do so at a rate greater than or equal to a_n). Our assumption on b_n is that $a_n b_n \to \infty$. We shall also assume $\omega_{jk} = O(n^{1/2})$ (the gaps in the $\hat{\theta}_j$ are constant or shrinking), which is convenient (but not necessary). Now consider our minimisation expression (5.21). If $O(n/(a_n)^2) < O(m) < O(n)$ then we have:

- $(m/n)^{1/2} a_n \to \infty$, so $\Phi\{(m/n)^{1/2}(-c\,\hat{\omega}_{jk} + z)\}$ tends to 0 or 1 whenever $\Phi\{-c\,\hat{\omega}_{jk} b_n\}$ does (in the case of ω_{jk} diverging).
- $(m/n)^{1/2} \to 0$, so $\Phi\{(m/n)^{1/2}(-c\,\hat{\omega}_{jk} + z)\}$ tends to 0.5 whenever $\Phi\{-c\,\hat{\omega}_{jk} b_n\}$ does (the case of ω_{jk} converging to zero).

This implies that if $O(n/(a_n)^2) < O(m) < O(n)$ then (5.21) will be driven to zero.

Conversely, if $O(n/(a_n)^2) \geq O(m)$ then we know that for the slowest of the diverging $|\omega_{jk}|$, $\Phi\{(m/n)^{1/2}(-c\,\hat{\omega}_{jk} + z)\}$ does not tend to 0 or 1 while $\Phi\{-c\,\hat{\omega}_{jk} b_n\}$ does, so (5.21) is not sent to zero. Similarly if $m/n \to \alpha > 0$ then $\Phi\{(m/n)^{1/2}(-c\,\omega_{jk} + z)\}$ does not tend to 0.5 for all values of z when ω_{jk} converges to zero. Thus the suggested minimisation for choosing m guarantees that $O(n/(a_n)^2) < O(m) < O(n)$. Notice this ensures both $m/n \to 0$ and $m \to \infty$.

Now fix j. Let $\mathcal{K}_+$, $\mathcal{K}_-$ and $\mathcal{K}_j$ denote the $k \neq j$ satisfying $\omega_{jk} \to \infty$, $\omega_{jk} \to -\infty$ and $\omega_{jk} \to 0$ respectively, consistent with our earlier notation. Now by reasoning similar to that in the proof of Theorem 3.1,

$$\begin{aligned}
\hat{r}_j &= 1 + \sum_{k \neq j} I(\hat{\theta}_j \leq \hat{\theta}_k) \\
&= 1 + \sum_{k \neq j} I(\sigma_j \Delta_j \leq \sigma_k \Delta_j + \omega_{jk}) \\
&= 1 + \#\mathcal{K}_+ + \sum_{k \in \mathcal{K}_j} I(\sigma_j Z_j \leq \sigma_k Z_k + o_p(1)) + o_p(1)\,,
\end{aligned}$$

where the Z_j are independent standard normal random variables. But assuming $O(n/(a_n)^2) < O(m) < O(n)$ we similarly have

$$\begin{aligned}
\hat{r}_j^* &= 1 + \sum_{k \neq j} I(\hat{\theta}_j^* \leq \hat{\theta}_k^*) \\
&= 1 + \sum_{k \neq j} I\left\{\sigma_j \Delta_j^* + \left(\frac{m}{n}\right)^{1/2} \sigma_j \Delta_j \leq \sigma_k \Delta_k^* + \left(\frac{m}{n}\right)^{1/2} \sigma_k \Delta_j + \left(\frac{m}{n}\right)^{1/2} \omega_{jk}\right\} \\
&= 1 + \#\mathcal{K}_+ + \sum_{k \in \mathcal{K}_j} I(\sigma_j Z_j \leq \sigma_k Z_k + o_p(1)) + o_p(1)\,.
\end{aligned}$$

This shows we have asymptotic distributional consistency and thus completes the proof for this case.

When there are no ω_{jk} that converge to zero, we are only guaranteed by the above reasoning that $O(m) > O(n/(a_n)^2)$. However this is sufficient for distributional consistency since $\mathcal{K}_j$ is empty and the asymptotic distribution is degenerate for each j, $\hat{r}_j = 1 + \#\mathcal{K}_+ + o_P(1)$. Similarly if all ω_{jk} converge to zero, the minimisation of (5.21) only ensures $O(m) < O(n)$. This is all we need for distributional accuracy in this case, since $\mathcal{K}_+$ is empty and $\hat{r}_j = 1 + \sum_{k \in \mathcal{K}_j} I(\sigma_j Z_j \leq \sigma_k Z_k + o_P(1))$.

5.5.3 Proof of Theorem 5.3. Observe from (5.22), (5.23) and (5.43) that

$$\begin{aligned} E(\hat{r}_j) - 1 &= \sum_{k\,:\,k\neq j} P(\hat{\theta}_j \le \hat{\theta}_k) = \sum_{k\,:\,k\neq j} P\big\{\Delta_j \le \Delta_k + 2^{1/2}\,\delta\,(j-k)\big\} \\ &= \{1 + o(1)\} \sum_{k\,:\,k\neq j} \Phi\{\delta\,(j-k)\} + o\big(\delta^{-1}\big) \\ &= \delta^{-1} \int_{-j\delta}^{\infty} \Phi(-x)\,dx + o\big(\delta^{-1}\big)\,, \end{aligned}$$

where $\Delta_k = n^{1/2}\,(\hat{\theta}_k - \theta_k)/\sigma$. This gives (5.25). Similarly, (5.26) follows from:

$$\begin{aligned} E(\hat{r}^*_j \,|\, \mathcal{X}) - 1 &= \sum_{k\,:\,k\neq j} P(\hat{\theta}^*_j \le \hat{\theta}^*_k \,|\, \mathcal{X}) \\ &= \sum_{k\,:\,k\neq j} P\big\{\Delta^*_j \le \Delta^*_k + 2^{1/2}\,\widetilde{\omega}_{jk} + 2^{1/2}\,\delta\,(j-k)\big\} \\ &= \{1 + o_p(1)\} \sum_{k\,:\,k\neq j} \Phi\{\widetilde{\omega}_{jk} + \delta\,(j-k)\} + o_p\big(\delta^{-1}\big)\,, \end{aligned}$$

where $\Delta^*_k = n^{1/2}\,(\hat{\theta}^*_k - \hat{\theta}_k)/\sigma$.

5.5.4 Proof of Theorem 5.4. (i) By (5.37), the probability that $|\hat{\theta}_j - \theta_j| > C_2\,(n^{-1}\log n)^{1/2}$ for some $j = 1, \ldots, p$, equals $O(p\,n^{-C_1}) = o(1)$, where we used (5.35) to obtain the last identity. Therefore, by (5.34), (5.37) and (5.36), for each $C > 0$ the probability that $\hat{\theta}_j - \hat{\theta}_k > C\,(n^{-1}\,\log n)^{1/2}$ for all $j = 1, \ldots, j_0$ and all $k = j_0 + 1, \ldots, p$, converges to 1 as $n \to \infty$. From this result and the first part of (5.37) it follows that, for $1 \le j \le j_0$,

$$\hat{r}_j = 1 + \sum_{k\,:\,k\neq j} I\big(\hat{\theta}_j \le \hat{\theta}_k\big) = 1 + \sum_{k\,:\,k\le j_0,\,k\neq j} I(W_j\,\sigma_j \le W_k\,\sigma_k) + \Delta_j\,,$$

where the random variables $W_1, \ldots, W_{j_0}$ are asymptotically independent and distributed as normal $\mathrm{N}(0,1)$, and $P(\Delta_j = 0) \to 1$ as $n \to \infty$.

(ii) In the bootstrap case it follows from the second formula in (5.31) that

$$\hat{r}^*_j = 1 + \sum_{k\,:\,k\le j_0,\,k\neq j} I\big\{m^{1/2}\,(\hat{\theta}^*_j - \hat{\theta}_j) + \Delta_{jk} \le m^{1/2}\,(\hat{\theta}^*_k - \hat{\theta}_k)\big\} + \Delta^*_j\,, \qquad (5.45)$$

where, if n is so large that $\inf_{1\le j\le j_0}\,\inf_{j_0<k\le p}(\theta_j-\theta_k) > 4\,C_2\,(m^{-1}\,\log m)^{1/2}$, then

$$\begin{aligned}\sup_{1\le k\le j_0}|\Delta_{jk}| \;&\le\; 2\,m^{1/2}\Big(\sup_{1\le j\le j_0}\big|\hat\theta_j-\theta_j\big|+\sup_{1\le j_1,j_2\le j_0}\big|\theta_{j_1}-\theta_{j_2}\big|\Big)\to 0\,, \quad (5.46)\\ P(\Delta_j^*\ne 0) \;&\le\; p\,\sup_{1\le k\le p}\Big[P\big\{|\hat\theta_k-\theta_k|>C_2\,(m^{-1}\,\log m)^{1/2}\big\}\\ &\qquad +P\big\{|\hat\theta_k^*-\hat\theta_k|>C_2\,(m^{-1}\,\log m)^{1/2}\big\}\Big]\to 0\,. \quad (5.47)\end{aligned}$$

The convergence in (5.46) is in probability and is a consequence of (5.37), (5.38) and the fact that $m/n\to 0$, and (5.47) follows from (5.35) and the second parts of (5.37) and (5.40). Part (ii) of Theorem 5.4 follows from (5.45)–(5.47).

(iii) Note that

$$E(\hat r_j^*\,|\,\mathcal{X},\mathcal{X}_j^*)-1=\sum_{k\,:\,k\ne j}P(\hat\theta_j^*\le\hat\theta_k^*\,|\,\mathcal{X},\mathcal{X}_j^*)=S_1^*+(S_2+S_3+S_4^*)\,\Omega\,,$$

where $P(0\le\Omega\le 1)=1$,

$$\begin{aligned}S_1^* \;&=\; \sum_{k=1}^{j_0}P(\hat\theta_j^*\le\hat\theta_k^*\,|\,\mathcal{X},\mathcal{X}_j^*)\,,\\ S_2 \;&=\; \sum_{k=j_0+1}^{\infty}I\big\{\theta_j-\theta_k\le 4\,C_2\,(m^{-1}\,\log m)^{1/2}\big\}\,,\\ S_3 \;&=\; \sum_{k=1}^{p}I\big\{|\hat\theta_k-\theta_k|>C_2\,(m^{-1}\,\log m)^{1/2}\big\}\,,\\ S_4^* \;&=\; \sum_{k=1}^{p}P\big\{|\hat\theta_k^*-\hat\theta_k|>C_2\,(m^{-1}\,\log m)^{1/2}\,\big|\,\mathcal{X},\mathcal{X}_j^*\big\}\,.\end{aligned}$$

In view of (5.38) and (5.39), $S_2=0$ for all sufficiently large n; by (5.35) and the second part of (5.37), $E(S_3)=o(1)$; and by (5.35) and the second part of (5.40), $E(S_4^*)=o(1)$. Therefore $E(S_2+S_3+S_4^*)=o(1)$; call this result (R). Since, using the independent-component bootstrap, $\mathcal{X}_j^*$ and $\mathcal{X}_k^*$ (for $k\ne j$) are independent conditional on $\mathcal{X}$; and since

$$P(\hat\theta_j^*\le\hat\theta_k^*\,|\,\mathcal{X},\mathcal{X}_j^*)=P\Big\{m^{1/2}\,(\hat\theta_j^*-\hat\theta_j)+m^{1/2}\,(\hat\theta_j-\hat\theta_k)\le m^{1/2}\,(\hat\theta_k^*-\hat\theta_k)\;\Big|\;\mathcal{X}\Big\}\,;$$

then it follows from (5.34), the first parts of (5.37) and (5.40), and Kolmogorov's extension theorem, that the joint distribution function of $P(\hat\theta_j^*\le\hat\theta_k^*\,|\,\mathcal{X},\mathcal{X}_j^*)$, for $1\le k\le j_0$ and $k\ne j$ (and conditional on $\mathcal{X}$), minus the joint distribution function

of $P(Z_j\,\sigma_j \le Z_k\,\sigma_k \,|\, Z_j)$ for $1 \le k \le j_0$ and $k \ne j$ (for independent standard normal random variables Z_k defined on an enlarged probability space), converges to zero in probability in any integral metric on a compact set. Therefore the distribution function of $S_1^* + 1$, conditional on $\mathcal{X}$, minus the distribution of $E(R_j \,|\, Z_j)$, converges in probability to zero. (Here, R_j is the function of $Z_1, \ldots, Z_{j_0}$ defined at (5.41), and the construction of $Z_1, \ldots, Z_{j_0}$ involves them being measurable in the sigma-field generated by $\mathcal{X} \cup \mathcal{X}_j^*$.) This property, and result (R), together imply (5.42).

To derive the final portion of part (iii) of Theorem 5.4, note that the argument leading to (5.42) implies that

$$E\Big\{ \sum_{k=j_0+1}^{\infty} P(\hat\theta_j^* \le \hat\theta_k^* \,|\, \mathcal{X}, \mathcal{X}_j^*) \Big\} = o(1)\,.$$

Therefore,

$$\begin{aligned} E(\hat r_j^{*2} \,|\, \mathcal{X}) &= \sum_{k_1,k_2\,:\,k_1,k_2 \ne j} E\Big\{ P(\hat\theta_j^* \le \hat\theta_{k_1}^* \,|\, \mathcal{X}, \mathcal{X}_j^*)\, P(\hat\theta_j^* \le \hat\theta_{k_2}^* \,|\, \mathcal{X}, \mathcal{X}_j^*) \;\Big|\; \mathcal{X} \Big\} \\ &= \sum_{k_1,k_2\,:\,k_1,k_2 \ne j,\, 1 \le k_1,k_2 \le j_0} E\Big\{ P(\hat\theta_j^* \le \hat\theta_{k_1}^* \,|\, \mathcal{X}, \mathcal{X}_j^*) \\ &\qquad\qquad \times P(\hat\theta_j^* \le \hat\theta_{k_2}^* \,|\, \mathcal{X}, \mathcal{X}_j^*) \;\Big|\; \mathcal{X} \Big\} + o_p(1) \\ &= T_2 + o_p(1)\,, \end{aligned}$$

where, for $\ell = 1, 2$,

$$T_\ell = E\left[\left\{ \sum_{k\,:\,k \ne j,\, 1 \le k \le j_0} P(\hat\theta_j^* \le \hat\theta_k^* \,|\, \mathcal{X}, \mathcal{X}_j^*) \right\}^\ell \;\middle|\; \mathcal{X} \right] + o_p(1)\,.$$

More simply, $E(\hat r_j^* \,|\, \mathcal{X}) = T_1 + o_p(1)$. The argument in the previous paragraph can be used to show that T_1 and T_2 converge in probability to $E\{E(R_j - 1 \,|\, Z_j)^2\}$ and $E(R_j - 1)$, respectively. Since $E\{E(R_j - 1 \,|\, Z_j)^2\} = E\{(R_j - 1)^2\}$ then $\mathrm{var}(\hat r_j^* \,|\, \mathcal{X})$ converges in probability to $\mathrm{var}(R_j)$, as required.

Chapter 6

The accuracy of extreme rankings

6.1 Background

6.1.1 Discussion. In this chapter we continue to explore the characteristics of rankings, given their important role in a variety of contexts. We have seen that in these situations a given ranking can carry a high degree of uncertainty, with this effect particularly pronounced in high dimensional cases; that is, where there are very many populations or institutions to be ranked. Diagnosing the extent of this uncertainty has been the focus of the previous chapter.

One interesting feature of many rankings reported over time is that the ordering at the extreme top or bottom remains relatively invariant. To rephrase, the uncertainty of a ranking is more of an issue in the middle ranks. For example, in the THE-QS university rankings[1], Harvard University ranked first for each of the years 2005-2008, while New York University's rankings are 56, 43, 49 and 40. If we believe that the observed data used for ranking are measures of true underlying values, distorted by noise, then we can reinterpret this behaviour as a tendency to obtain correct rankings at extremes, but not otherwise. It is this phenomenon that we explore in this chapter, using both theoretical and numerical arguments.

Intuitively this behaviour has a natural explanation. Those scores at the extreme of a range are more likely to be sufficiently "spaced out" to overcome the problems of data noise, whereas less extreme scores are likely to be bunched more closely together. We introduce models that describe this behaviour, and explore their properties. Related to this, it turns out that one important consideration for correct ranking at the extremes is whether the possible scores used for ranking have infinite support but

[1] www.topuniversities.com

nevertheless have light tails. If this is the case and the tail of the distribution of the underlying scores is smooth, we can expect accurate ranking of the top portion of the institutions, even when dimension is very large. Moreover, even when the support is bounded, there remains potential for correct ranking at extremes, although now there is greater likelihood that the ranking will change if new institutions are added. Such results have a variety of practical implications; we briefly present two of these here, with more detail provided in the numerical section.

6.1.2 Example 1: University rankings. Suppose we attempt to rank universities and other research institutions by counting how many papers their faculty members publish in *Nature*[2] each year. This is a high dimensional example due to the large number of institutions competing to be published. Figure 6.1 shows the ranking of the top 50 institutions on this measure. The institutions are aligned along the horizontal axis, with the each dot denoting the point estimate of the rank and the vertical line a corresponding estimated 90% prediction interval. The four plots show how the confidence intervals change as we increase the number of years, n, of data used for the ranking.

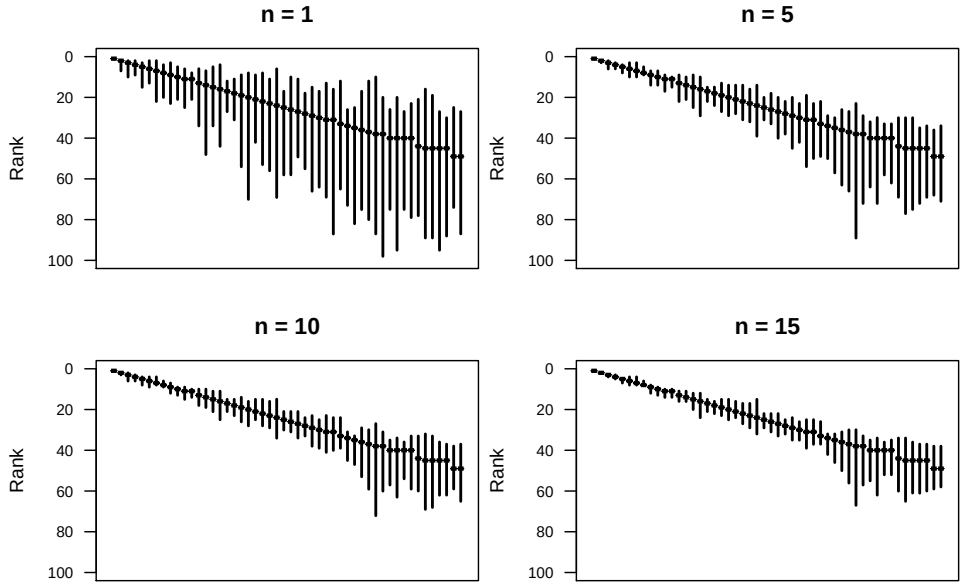


Figure 6.1: Prediction intervals for top-ranked universities based on publications in Nature, averaged over various numbers of years

The two main observations are that the prediction intervals are widest when a smaller number of years are considered and that the prediction intervals for the highest ranked universities are the smallest. In fact the intervals are small enough in the extremes to give us genuine confidence in that aspect of the ranking. Even

[2]www.nature.com/nature/index.html

when $n = 1$ we can be reasonably sure that the top ranked institution (Harvard University) is in fact ranked correctly. When $n = 15$ the top four universities are known with a high degree of certainty, and the next set of ten or so is fairly stable too. Thus it is possible to have correctness in the upper extreme of this ranking, even when the lower ranks remain highly variable. In the present work we model this phenomenon by addressing the underlying stochastic properties of the institutions; the data provide only a noisy measure of this random process, and we assess the impact of the noise on the ranking.

6.1.3 Example 2: Colon microarray data. We take the colon microarray data first analysed by Alon et al. (1999). It consists of 62 observations in total, each of which indicates either a normal colon or a tumor (the binary response). For each observation there are also expression levels for $p = 2{,}000$ genes. It is of interest to determine which genes are most closely related to the response, so that they can be investigated further. This of course amounts to a ranking and we are interested in stability at the extreme, since we seek only a small number of genes. Here the genes are ranked based on the Mann-Whitney U test statistic, which is a nonparametric assessment of the difference between the two distributions.

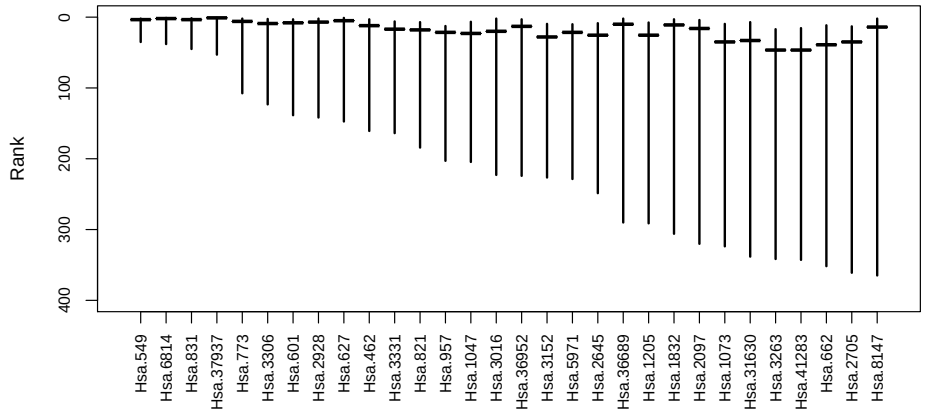


Figure 6.2: Prediction intervals for top-ranked genes in Colon dataset

Figure 2 plots the top 30 genes, ranked by the lower tail of an estimated 90% prediction interval, rather than the point estimate of the rank. In this situation we cannot authoritatively conclude that any of the top genes are ranked exactly correctly, but the top four genes appear much more stable than the others. This stability is highly important; if the length of all prediction intervals were roughly the same as the average length (1,400 genes), then there would be little hope of discovering useful genes from such datasets.

There is a literature on the bootstrap in connection with rankings, discussed in Section 5.1. More generally there is a vast literature on ranking problems in

statistics, and we cite here only the more relevant items since 2000. Joe (2000, 2001) discusses ranking problems in connection with random utility models, and points to connections to multivariate extreme value theory. Murphy and Martin (2003) develop mixture-based models for rankings. Mease (2003) and Barker et al. (2005) treat methods for ranking football players. McHale and Scarf (2005) study the problem of ranking immunisation coverage in US states. Brijs et al. (2006, 2007) introduce Bayesian models for the ranking of hazardous road sites, with the aim of better scheduling road safety policies. Chen et al. (2006) discuss ranking accuracy in ranked-set sampling methods, and Opgen-Rhein and Strimmer (2007) examine the accuracy of gene rankings in high-dimensional problems involving genomic data. Nordberg (2006) addresses the reliability of performance rankings. Corain and Salmaso (2007) and Quevedo et al. (2007) discuss ways of constructing rankings.

Section 6.2 describes our model for the ranking problem, and discusses the main properties of this framework. The formal theoretical results which underpin the discussion in Section 6.2 are given in Section 6.3. Section 6.4 presents simulated and real-data numerical work, including details on the examples presented above. Technical proofs are deferred to Section 6.5.

6.2 Methodology

As in the previous chapter, we consider a set of underlying parameters $\theta_1, \ldots, \theta_p$ corresponding to the objects to be ranked, hereafter referred to as items. The error in the estimation is controlled by the number of observed data points, n. In our analysis we take $p = p(n)$ to diverge with n as the latter increases. An obvious difficulty here is in establishing where the newly added items should fit into the ranking. A natural solution is to take the θ_js to be randomly generated from some distribution function. In the setup below we interpret the Θ_js as values of means; see the end of this Section for generalisations.

Let $\Theta_1, \ldots, \Theta_p$ denote independent and identically distributed random variables, and write

$$\Theta_{(1)} \leq \ldots \leq \Theta_{(p)} \tag{6.1}$$

for their ordered values. There exists a permutation $R = (R_1, \ldots, R_p)$ of $(1, \ldots, p)$ such that $\Theta_{(j)} = \Theta_{R_j}$ for $1 \leq j \leq p$. If the common distribution of the Θ_js is continuous then the inequalities in (6.1) are all strict and the permutation is unique.

We typically do not observe the Θ_js directly, only in terms of noisy approximations which can be modelled as follows. Let $Q_i = (Q_{i1}, \ldots, Q_{ip})$ denote independent and identically distributed random p-vectors with finite variance and zero mean, independent also of $\Theta = (\Theta_1, \ldots, \Theta_p)$. Suppose we observe

$$X_i = (X_{i1}, \ldots, X_{ip}) = Q_i + \Theta \tag{6.2}$$

for $1 \le i \le n$. The mean vector,

$$\bar{X} = (\bar{X}_1, \ldots, \bar{X}_p) = \frac{1}{n} \sum_{i=1}^{n} X_i = \bar{Q} + \Theta \,, \tag{6.3}$$

is an empirical approximation to Θ. (Here $\bar{Q} = n^{-1} \sum_i Q_i$ equals the mean of the p-vectors Q_i.) The components of $\bar{X}$ can also be ranked, as

$$\bar{X}_{(1)} \le \ldots \le \bar{X}_{(p)} \,, \tag{6.4}$$

and there is a permutation $\widehat{R}_1, \ldots, \widehat{R}_p$ of $1, \ldots, p$ such that $\bar{X}_{(j)} = X_{\widehat{R}_j}$ for each j. If the common distribution of the Θ_js is continuous then, regardless of the distribution of the components of Q_i, the inequalities in (6.4) are strict with probability 1.

The permutation $\widehat{R} = (\widehat{R}_1, \ldots, \widehat{R}_p)$ serves as an approximation to R, and we wish to determine the accuracy of that approximation. In particular, for what values of $j_0 = j_0(n, p)$, and for what relationships between n and p, is it true that

$$P\big(\widehat{R}_j = R_j \ \text{ for } \ 1 \le j \le j_0\big) \to 1 \tag{6.5}$$

as n and p diverge? That is, how deeply into the ranking can we go before the connection between the true ranking and its empirical form is seriously degraded by noise?

The answer to this question depends to some degree on the extent of dependence among the components of each Q_i. To elucidate this point, let us consider the case where all the components of Q_i are identical; this is an extreme case of strong dependence. Then the components of $\bar{Q}$ are also identical. Clearly, in this setting $\widehat{R}_j = R_j$ for each j, and so (6.5) holds in a trivial and degenerate fashion. Other strongly dependent cases, although not as clear-cut as this one, can also be shown to be ones where $\widehat{R}_j = R_j$ with high probability for many values of j.

The case which is most difficult, i.e. where the strongest conditions are needed to ensure that (6.5) holds, occurs when the components of Q_i are independent. To emphasise this point we give sufficient conditions for (6.5), and show that when the components of each Q_i are independent, those conditions are also necessary. Our arguments can be modified to show that the conditions continue to be necessary under sufficiently weak dependence, for example if the components are m-dependent where $m = m(n)$ diverges sufficiently slowly as n increases.

The assumptions under which (6.5) holds are determined mainly by the lower tail of the common distribution of the Θ_js. If that distribution has an exponentially light left-hand tail, for example if the tail is like that of a normal distribution, then a sufficient condition for (6.5) is that j_0 should increase at a strictly slower rate than $n^{1/4} (\log n)^c$, where the constant c, which can be either positive or negative, depends

on the rate of decay of the exponential lower tail of the distribution of Θ. For example, $c = 0$ if the distribution decays like $e^{-|x|}$ in the lower tail, and $c = -\frac{1}{4}$ if it is normal. As indicated in the previous paragraph, the condition $j_0 = o\{n^{1/4}\,(\log n)^c\}$ is also necessary for (6.5) if the components of the Q_is are independent.

These results have several interesting aspects, including: (a) The exponent $\frac{1}{4}$ in the condition $j_0 = o\{n^{1/4}\,(\log n)^c\}$ does not change among different types of distribution with exponential tails; (b) the exponent is quite small, implying that the empirical rankings $\widehat{R}_j$ quite quickly become unreliable as predictors of the true rankings R_j; and (c) the critical condition $j_0 = o\{n^{1/4}\,(\log n)^c\}$ does not depend on the value of p. (We assume that p diverges at no faster than a polynomial rate in n, but we impose no upper bound on the degree of that polynomial.)

The condition on j_0 such that (6.5) holds changes in important ways if the lower tail of the distribution of the Θ_js decays relatively slowly, for example at the polynomial rate $x^{-\alpha}$ as $x \to \infty$. Examples of this type include Pareto, non-normal Stable, and Student's t distributions, and more generally, distributions with regularly varying tails. Here a sufficient condition for (6.5) to hold is $j_0 = o\{(n^{\alpha/2}\,p)^{1/(2\alpha+1)}\}$, and this assumption is necessary if the components of the Q_is are independent. In this setting, unlike the exponential case, the value of dimension, p, plays a major role in addition to the sample size, n, in determining the number of reliable rankings.

In practical terms, a major way in which this heavy-tailed case differs from the light-tailed setting considered earlier is that if a polynomially large number of new items are added to the competition in the heavy-tailed case, and all items are re-ranked, the results will change significantly and the number of correct rankings will also alter substantially. By way of contrast, if a polynomially large number of new items are added in the light-tailed, or exponential, case then there will again be many changes to the rankings, but now there will be relatively few changes to the number of items that are correctly ranked.

The exponential case can be regarded as the limit, as $\alpha \to \infty$, of the polynomial case. More generally, note that as the left-hand tail of the common distribution of the Θ_js becomes heavier, the value of j_0 can be larger before (6.5) fails. That is, if the distribution of the Θ_js has a heavier left-hand tail then the empirical rankings $\widehat{R}_j$ approximate the true rankings R_j for a greater number of values of j, before they degenerate into noise.

The analysis above has focused on cases where the ranks of the X_js are estimated by ranking empirical means of noisy observations of those quantities; see (6.4). However, similar results are obtained if we rank other measures of location. Such a measure need only satisfy moderate deviation properties similar to (6.19) and (6.20) in the proof of Theorem 6.1. Thus, the results are applicable to a wide range of ranking contexts. For example, L_q location estimators for general $q \geq 1$ enjoy moderate

deviation properties under appropriate assumptions. Therefore if we take the variables Q_{ij} to have zero median, rather than zero mean, and continue to define X_i by (6.2) but replace the ranking in (6.4) by a ranking of medians, then the results above and those in Section 6.3 continue to hold, modulo changes to the regularity conditions. Other suitable measures include the Mann-Whitney test used in the genomic example, quantiles, and some correlation-based measures.

The model suggested by (6.2), where data on Θ arise in the form of p-vectors $X_1, \ldots, X_n$, is attractive in a number of high-dimensional settings, for example genomics. There, the jth component X_{ij} of X_i would typically represent the expression level of the jth gene of the ith individual in a sample. However, in other cases the means $\bar{X}_1, \ldots, \bar{X}_p$ at (6.3), or medians or other location estimators, might be computed from quite different datasets, one for each component index j. Moreover, those datasets might be of different sizes, $n_1, \ldots, n_p$ say, and then the argument that they arise naturally in the form of vectors would be inappropriate. This can happen when data are used to rank items, for example schools where the ranking is based on individual student performance. The conclusions discussed earlier in this Section, and the theoretical properties developed in Section 3 below, continue to apply in this case provided there is an "average" value, n say, of the n_js which represents all of them, in the sense that

$$n = O\Big(\min_{1 \le j \le p} n_j\Big) \quad \text{and} \quad \max_{1 \le j \le p} n_j = O(n) \tag{6.6}$$

as n diverges. Additionally, in such cases it is often realistic to make the assumption that the corresponding centred means (or medians, etc) $\bar{Q}_j = n^{-1} \sum_i Q_{ij}$ are stochastically independent of one another, and so the particular results that are valid in this case are immediately available.

The distribution of the Θ_j's has been taken to be continuous. This is usually appropriate although there can be contexts in which the distribution is discrete. Note that assumption of discreteness of the Θ_js is different from that of discreteness of the observations X_{ij}. In such cases the analysis still holds, except that allowance must be made for ties (any reordering of tied Θ_js is still "correct"), and the tail density assumptions should be characterised in integral form.

The model has been set up so that it focuses on the populations with lowest parameters Θ_j. Obviously similar arguments apply to the largest parameters too, so the results are applicable to both the most highly and lowly ranked populations.

6.3 Theoretical properties

For the most part we shall assume one of two types of lower tail for the common distribution function, F, of the random variables Θ_j: either it decreases exponentially

fast, in which case we suppose that $F(-x) \asymp x^{\beta} \exp(-C_0\, x^{\alpha})$ as $x \to \infty$, where $\alpha > 0$ and $-\infty < \beta < \infty$; or it decreases polynomially fast, in which case $F(-x) \asymp x^{-\alpha}$ as $x \to \infty$, where $C_0, \alpha > 0$. (The notation $f(x) \asymp g(x)$, for positive functions f and g, will be taken to mean that $f(x)/g(x)$ is bounded away from zero and infinity as $x \to \infty$.) The former case covers distributions such as the normal, exponential and Subbotin; the latter, distributions such as the Pareto, Student's t and non-normal stable laws (e.g. the Cauchy).

It is convenient to impose the shape constraints on the densities, which we assume to exist in the lower tail, rather than on the distribution functions. Therefore we assume that one of the following two conditions holds as $x \to \infty$:

$$(d/dx)\, F(-x) \;\asymp\; (d/dx)\, x^{\beta} \exp\bigl(-C_0\, x^{\alpha}\bigr)\,, \tag{6.7}$$

$$(d/dx)\, F(-x) \;\asymp\; (d/dx)\, x^{-\alpha}\,. \tag{6.8}$$

In both (6.7) and (6.8), α must be strictly positive, but β in (6.7) can be any real number. The constant C_0 in (6.7) must be positive. We assume too that:

$$\begin{aligned}&\text{for fixed constants } C_1, \ldots, C_5 > 0, \text{ where } C_2 > 2\,(C_1 + 1) \text{ and}\\ &C_4 < C_5,\, p = O(n^{C_1}) \text{ as } n \to \infty, \text{ and, for each } j \geq 1,\, E|Q_j|^{C_2} \leq\\ &C_3,\, E(Q_j) = 0, \text{ and } E(Q_j^2) \in [C_4, C_5].\end{aligned} \tag{6.9}$$

Recall from Section 6.1 that we wish to examine the probability that the true ranks R_j, and their estimators $\widehat{R}_j$, are identical over the range $1 \leq j \leq j_0$. We consider both j_0 and p to be functions of n, so that the main dependent variable can be considered to be n. With this interpretation, define

$$\nu_{\rm exp} = \nu_{\rm exp}(n) = n^{1/4}\,(\log n)^{\{(1/\alpha)-1\}/2}\,, \quad \nu_{\rm pol} = \nu_{\rm pol}(n) = (n^{\alpha/2}\, p)^{1/(2\alpha+1)}\,, \tag{6.10}$$

where the subscripts denote "exponential" and "polynomial," respectively, and refer to the respective cases represented by (6.7) and (6.8). In the theorem below we impose the additional condition that, for some $\epsilon > 0$,

$$n = O(p^{4-\epsilon})\,. \tag{6.11}$$

This restricts our attention to problems that are genuinely high-dimensional, in the sense that, with probability converging to 1, not all the rankings are correct. (That property fails to hold if p diverges sufficiently slowly as a function of n.) Assumption (3.5) is also very close, in both the exponential and polynomial cases, to the basic condition $j_0 \leq p$, as can be seen via a little analysis starting from (3.6) and (3.7) in the respective cases; yet, at the same time, (3.5) is suitable to both cases, and so helps to unify our account of their properties. Note too that (3.5) implies that, in

both the exponential and polynomial cases, $\nu_{\rm exp} = O(p^{1-\delta})$ and $\nu_{\rm pol} = O(p^{1-\delta})$ for some $\delta > 0$.

Theorem 6.1. *Assume (6.9), (6.11) and that either (a) (6.7), or (b) (6.8) holds. In case (a), if*

$$j_0 = o(\nu_{\rm exp}) \tag{6.12}$$

as $n \to \infty$ then (6.5) holds. Conversely, when the components of the vectors Q_i are independent, (6.12) is necessary for (6.5). In case (b), if

$$j_0 = o(\nu_{\rm pol})\,, \tag{6.13}$$

then (6.5) holds. Conversely, when the components of the vectors Q_i are independent, (6.13) is necessary for (6.5).

It can be deduced from Theorem 6.1 that when a new item (e.g. an institution) enters the competition that leads to the ranking, we are still able to rank the top j_0 institutions correctly. In this sense the institutions that make up the cohort of size j_0 do not need to be fixed.

It is also of interest to consider cases where the common distribution, F, of the Θ_js is bounded to the left, for example where $F(x) \asymp x^\alpha$ as $x \downarrow 0$. However, it can be shown that in this context, unless p is constrained to be a sufficiently low degree polynomial function of n, very few of the estimated ranks $\widehat{R}_j$ will agree with the correct values R_j.

To indicate why, we first recall the model introduced in Section 6.1, where the estimated ranks $\widehat{R}_j$ are derived by ordering the values of $\bar{Q}_j + \Theta_j$. Here $\bar{Q}_j = n^{-1} \sum_{1 \le i \le n} Q_{ij}$ is the average value of n independent and identically distributed random variables with zero mean. Therefore the means, $\bar{Q}_j$, are of order $n^{-1/2}$. By way of contrast, if we take $\alpha = 1$ in the formula $F(x) \asymp x^\alpha$ as $x \downarrow 0$, for example if F is the uniform distribution on $[0, 1]$, then the spacings of the order statistics $\Theta_{(1)} \le \ldots \le \Theta_{(p)}$ are approximately of size p^{-1}. (More concisely, they are of size Z/p where Z has an exponential distribution; an independent version of Z is used for each spacing.) Therefore, if p is of larger order than $n^{1/2}$ then the errors of the "estimators" $\bar{Q}_j + \Theta_j$ of Θ_j, for $1 \le j \le p$, are an order of magnitude larger than the spacings among the Θ_js. This can make it very difficult to estimate the ranks of the Θ_js from the ranks of values of $\bar{Q}_j + \Theta_j$. Indeed, it can be shown that, in the difficult case where the components of the Q_is are independent, and even for fixed j_0, if $\alpha = 1$ and p is of larger order than $n^{1/2}$ then in contrast to (6.5),

$$P\big(\widehat{R}_j = R_j \ \text{ for } \ 1 \le j \le j_0\big) \to 0\,. \tag{6.14}$$

This explains why, when $F(x) \asymp x^\alpha$, it can be quite rare for the estimated ranks

$\widehat{R}_j$ to match their true values. Indeed, no matter what the value of α and no matter what the value of j_0, property (6.5) will typically fail to hold unless p is no greater than a sufficiently small power of n, in particular unless $p = o(n^{\alpha/2})$, as the next result indicates. Thus, the differences between the cases of bounded and unbounded distributions are stark, as can be seen by contrasting Theorem 6.1 with the properties described below.

Theorem 6.2. *Assume that* $(d/dx)\, F(x) \asymp x^{\alpha-1}$ *as* $x \downarrow 0$, *where* $\alpha > 0$, *and that* (6.9) *holds. Part (a): Instances where* (6.5) *holds and* $p^2/n^\alpha \to 0$. *Under the latter condition, (i) if* $\alpha < \frac{1}{2}$ *then* (6.5) *holds even for* $j_0 = p$; *(ii) if* $\alpha = \frac{1}{2}$ *then* (6.5) *holds provided that*

$$(\log j_0)^{2\alpha}\, (p^2/n^\alpha) \to 0\,; \tag{6.15}$$

and (iii) if $\alpha > \frac{1}{2}$ *then* (6.5) *holds provided that*

$$j_0 = o\big\{(n^{\alpha/2}/p)^{1/(2\alpha-1)}\big\}\,. \tag{6.16}$$

Part (b): Converses to (a)(ii) and (a)(iii). If $p^2/n^\alpha \to 0$ *and the components of the vectors* Q_i *are independent then, if* (6.5) *holds, so too does* (6.15) *(if* $\alpha = \frac{1}{2}$*) or* (6.16) *(if* $\alpha > \frac{1}{2}$*). Part (c): Instances where* (6.14) *holds. If* $\alpha > 0$ *and* $p^2/n^\alpha \to \infty$, *and if the components of the vectors* Q_i *are independent, then* (6.14) *holds even for* $j_0 = 1$.

The proof of Theorem 6.2 is similar to that of Theorem 6.1, and so is omitted. Theorem 6.1 is derived in Section 6.5. Both results continue to hold if the sample from which $\bar{X}_j$ is computed is of size n_j for $1 \le j \le p$, rather than n, provided that (6.6) holds.

6.4 Numerical properties

This section discusses three real-data and three simulated examples linked to the theoretical properties in Section 6.3. The real-data examples make use of the bootstrap to create prediction intervals (Xie et al., 2009; Chapter 5 of this thesis). In each simulated example the error is relatively light-tailed, and any discussion of tails refers to the distribution of the Θ_js. In our real-data examples the noise has been averaged and so is also generally light-tailed. Thus, any heavy-tailed behaviour present in the real-data examples is likely to be due to heavy tails of the distribution of the Θ_js, rather than the noise.

6.4.1 Example: Continuation of Example 6.1.2. The originating institutions of *Nature* articles were obtained using the ISI Web of knowledge database[3] for each

[3] www.isiknowledge.com

of the years 1999 through 2008. A point ranking was obtained by taking the average number of articles published per year. Of course, there are implicit simplifying assumptions in doing this, most significantly concerning the independence of articles between years, and the stationarity of means time. These assumptions appear reasonable in context, and are consistent with most publication-based analyses.

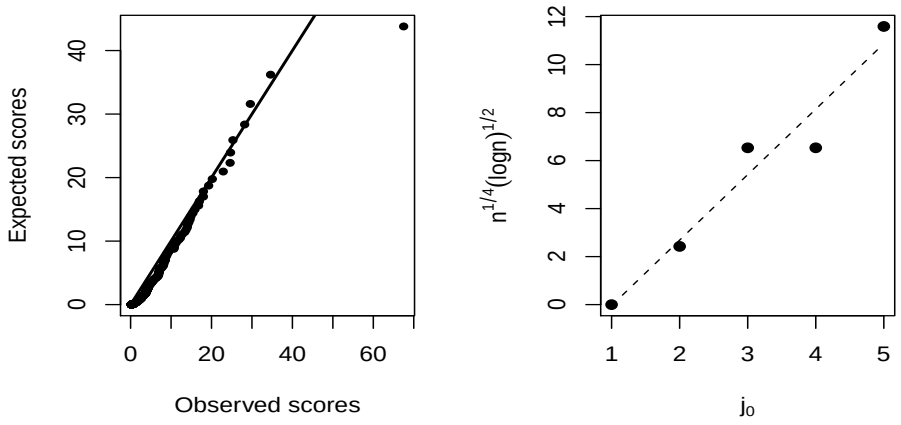


Figure 6.3: The left panel is a QQ plot for the Nature data against the exponential distribution. The right panel plots a transform of the number of years of data required to rank j_0 institutions correctly for various j_0.

When constructing prediction intervals the bootstrap resamples for each institution were drawn independently, conditional on the data, as in the independent component bootstrap of Section 5.4. The number of observations in the resample can be varied to create different time windows, as illustrated in Figure 6.1. The most natural question from a ranking correctness viewpoint is determining the behaviour at the right tail; there are many institutions with mean at or near the hard threshold of zero, so there is little hope for ranking correctness in the left tail. Furthermore, the right tail appears to be long. Harvard University has an average of 67.5 papers per year, followed by means of 34.6, 29.6 and 28.2 for Berkeley, Stanford and Cambridge respectively.

A natural question to ask is what the tail shape for this example might be. Approaches to estimating the shape parameter of a distribution with regularly varying tails, such as the method of Hill (1975), are unstable for these data; the number of extreme data for which a linear fit is plausible is very small, implying that the decay rate is faster than polynomial. Indeed, the left panel of Figure 6.3 shows the QQ plot of the observed data against a random variable with distribution function $F(x) = 1 - \exp(-0.85\, x^{1/2})$, which suggests that an exponential tail might be reasonable for the data. If this is the case then the number of institutions that we expect to be ranked correctly should depend, to first order, only on n, not on p, and be of order up to $n^{1/4}\,(\log n)^{1/2}$. One way to explore this further is to take j_0 as given, and to

resample from the data, seeking (for example) the number of years, n, needed to obtain correct ranking of the first j_0 institutions at least 90% of the time. A plot of j_0 against $n^{1/4}\,(\log n)^{1/2}$ should be roughly linear. The right-hand panel of Figure 6.3 plots results of this experiment and appears to support the hypothesis. The flatness between $j_0 = 3$ and $j_0 = 4$ indicates that these two institutions are quite difficult to separate from each other.

6.4.2 Example: Continuation of Example 6.1.3. The Mann-Whitney test statistic can be written as

$$\max\left\{\sum_{i,j} I(x_i < y_j), \sum_{i,j} I(x_i > y_j)\right\},$$

where the x_is and y_js are the observed values of the two samples. Notice that this statistic will have a hard lower threshold at $n_1 n_2/2$, where n_1 and n_2 are the sizes of the two classes. Here, like the previous example, when the distributions differ only in location the difference has to be quite large to be detectable. Figure 6.4 shows the estimated density as well as the truncated normal density, which is the distribution that the scores would have if none of the genes had systematically different means for the two classes. This suggests that an assumption that the majority of genes is unrelated to whether the tissue is tumourous is not valid here.

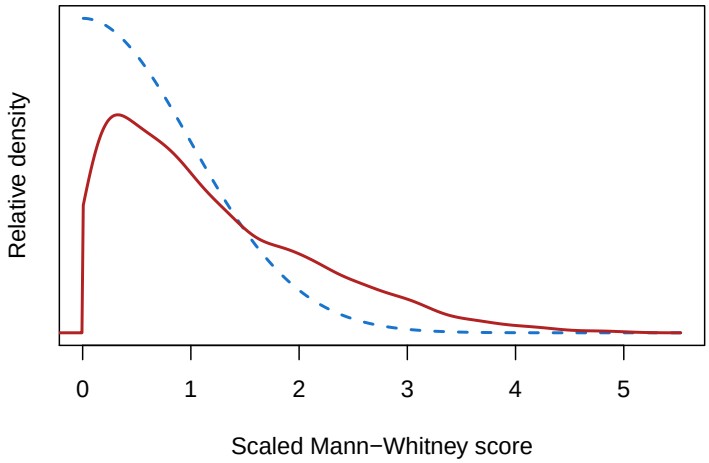


Figure 6.4: Estimated sampling density genes under the Mann-Whitney test for Colon data

Bootstrapped versions of the dataset with different choices for n were created to indicate how many observations we need to obtain reasonable confidence in a ranking. Table 6.1 shows the probability that the set of the top j genes is identified correctly out of the 2,000 for various j and n. Note that this is a slightly different statistic from the one in (6.5), since we allow any permutation of the top j genes to be detected.

The results suggest that we have nearly a 50% chance of detecting the top gene if $n = 250$, and a 20% chance of correctly choosing the top four. The upper tail for this dataset again appears relatively light; the model $F(x) = 1 - \exp\{-0.19\,(x-1)^2\}$, for $x > 1$, produces a good fit to the upper tail.

j	n				
	62	100	150	200	250
1	0.251	0.326	0.437	0.446	0.490
2	0.067	0.109	0.166	0.218	0.277
4	0.022	0.054	0.094	0.163	0.193
6	0.007	0.018	0.035	0.040	0.068

Table 6.1: Probability that set of top j genes is correct for Colon data

Theorem 6.1 suggests that these probabilities should not depend on the choice of p. We can obtain a sense of this by randomly sampling, without replacement, $p = 500$ or $p = 1{,}000$ genes from the original $p = 2{,}000$, for each simulation; and recalculating the values in Table 1. For $j = 4$ and $n = 250$ the respective probabilities were 0.183 and 0.170, quite close to the value 0.193 observed for $p = 2{,}000$. While the equivalence appears good for $j \geq 4$, there are larger departures for $j = 1$ or 2, where the initial results for this particular realisation tend to distort the calculation.

6.4.3 Example: School rankings. A third example of accuracy in the extremes of a ranking is based on the student performance results introduced in Section 5.3.6. The results in Figure 6.5 indicate the increased confidence we can have in the upper extreme, with the top school identified with reasonable certainty. In this example the possible range of scores for ranking has finite support, being restricted to the interval $[0, 1]$; thus it is a context where Theorem 6.2 is applicable.

The estimator of α by Hill (1975), when (6.8) holds, is relatively stable in this example and suggests that $\alpha \approx 6$. From (6.16) we can calculate that $(n^{\alpha/2}/p)^{1/(2\alpha-1)} \approx 4$, which is consistent with a small number of schools being correctly ranked. If the number were large then we would expect a significant portion of the schools to be ranked with a high degree of accuracy. In the case of these data, however, the small value suggests that it might not be possible to obtain any correct ranks.

6.4.4 Example: Simulation with exponential tails and infinite support. Here we simulate increasing n and p in the case of exponential tails. For a given n, set $p = 0.0005\,n^2$, let the Θ_js be drawn from a standard exponential distribution and the Q_{ij}s be normal random variables with zero mean and standard deviation 3.5. Table 6.2 shows the results of 1,000 simulations for various values of n, approximating (6.5) for different choices of j_0. Theorem 6.1 suggests that the results should converge to 1 if $j_0 = o(n^{1/4})$, and degrade otherwise. This appears consistent with the results.

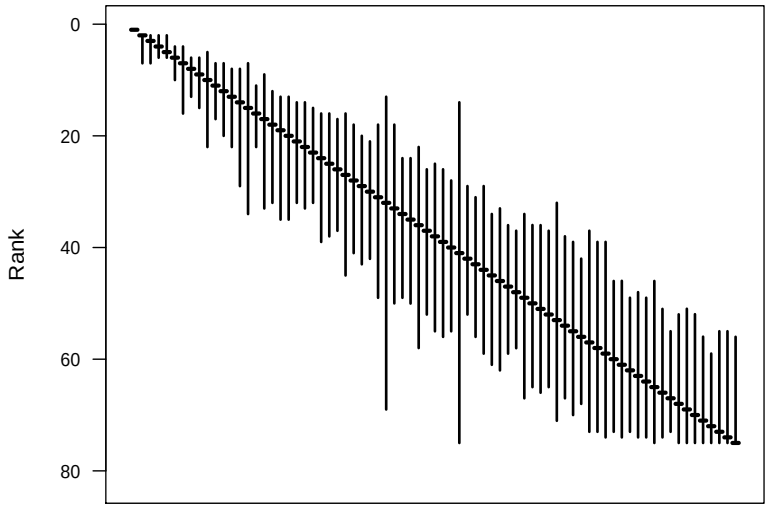


Figure 6.5: Rankings of schools by students' exam performance with prediction intervals

The difficulty of the problem due to the quadratic growth of p and the large error in Q_{ij} is also evident; even when $j_0 = 1$ and n is large, reliable prediction of the top rank is not assured.

j_0	n						
	500	1,000	2,000	5,000	10,000	20,000	50,000
1	0.909	0.9365	0.959	0.970	0.9745	0.9840	0.9910
$n^{0.15}$	0.764	0.823	0.767	0.844	0.897	0.872	0.890
$n^{0.20}$	0.591	0.700	0.655	0.683	0.667	0.664	0.743
$n^{0.25}$	0.420	0.406	0.424	0.383	0.334	0.402	0.428
$n^{0.30}$	0.183	0.188	0.180	0.116	0.101	0.079	0.069
$n^{0.35}$	0.056	0.030	0.021	0.004	0.002	0.000	0.001

Table 6.2: Probability that the first j_0 rankings are correct in the case of exponential tails

6.4.5 Example: Simulation with polynomial tails and infinite support. We use the same setup as in the previous example, except that the generating distribution for the Θ_js is Pareto, $F(x) = 1 - x^{-\alpha}$ for $x \geq 1$, with $\alpha = 4$. Theorem 6.1 and (6.10) suggest that the rate $n^{4/18}\, p^{1/9} = n^{4/9}$ is critical for j_0, and this is consistent with the results in Table 6.3. This is an easier problem than that in the previous example, because of the polynomial decay of the tail. For instance, the top right-hand result in the table suggests that the top nine ranks can be correctly ascertained more than 90% of the time when $p > 50{,}000$, whereas the figure 0.890 in the last column of Table 6.2 suggests that, for the distribution represented there, only the top five ranks have this level of reliability.

j_0	n						
	500	1,000	2,000	5,000	10,000	20,000	50,000
$(1/5)\,n^{0.35}$	0.884	0.832	0.908	0.920	0.898	0.921	0.945
$(1/5)\,n^{0.40}$	0.694	0.672	0.708	0.731	0.801	0.786	0.803
$(1/5)\,n^{4/9}$	0.477	0.510	0.586	0.568	0.569	0.520	0.540
$(1/5)\,n^{0.50}$	0.283	0.242	0.252	0.161	0.140	0.120	0.096
$(1/5)\,n^{0.55}$	0.071	0.086	0.031	0.020	0.006	0.002	0.001

Table 6.3: Probability that the first j_0 rankings are correct in the case of exponential tails

6.4.6 Example: Simulation with polynomial tails with finite support. Theorem 6.2 has many interesting consequences, but the present example focuses on case (iii), where $\alpha > \frac{1}{2}$. First let the Θ_js be uniformly distributed on $[0, 1]$, and consider a case where the entire ranking is correct. Using the notation of Section 6.3 and taking $\alpha = 1$, Theorem 6.2 implies that $p \asymp n^{1/4}$ defines the critical growth in dimension. For simulation we took $p = 2\,n^k$ for various k, and scaled the (normally distributed) error for each k such that the $n = 500$ case had probability approximately 0.5 of correctly identifying all ranks. Each simulation was repeated 10,000 times, with results summarised in Table 6.4. As predicted, growth rates in dimension slower than $n^{1/4}$ have probability of correct ranking tending to 1, while those faster than $n^{1/4}$ degrade.

k	n						
	500	1,000	2,000	5,000	10,000	20,000	50,000
1/6	0.502	0.494	0.525	0.593	0.635	0.658	0.701
1/5	0.498	0.511	0.471	0.558	0.568	0.578	0.606
1/4	0.497	0.478	0.492	0.505	0.517	0.496	0.502
1/3	0.500	0.457	0.395	0.343	0.289	0.259	0.212
1/2	0.502	0.369	0.249	0.107	0.046	0.011	0.000

Table 6.4: Probability all ranks identified correctly when Θ_j is uniformly distributed

k	n							
	5×10^3	1×10^4	2×10^4	5×10^4	1×10^5	2×10^5	5×10^5	1×10^6
0.05	0.500	0.539	0.553	0.583	0.603	0.609	0.628	0.641
0.07	0.502	0.532	0.506	0.546	0.558	0.580	0.555	0.591
1/11	0.497	0.486	0.489	0.516	0.489	0.463	0.513	0.496
0.11	0.497	0.481	0.471	0.432	0.461	0.447	0.452	0.421
0.13	0.506	0.492	0.461	0.481	0.445	0.427	0.387	0.385

Table 6.5: Probability that lowest $10n^k$ scores identified correctly

Next we examine the case $p = 5 \times 10^{-6}\, n^2$, where dimension grows at a quadratic rate; and $F(x) = x^\alpha$ on $[0, 1]$, with $\alpha = 6$, implying a reasonably severe tail. Theorem 6.2 suggests that if $j_0 = o(p^{1/22})$, or equivalently if $j_0 = o(n^{1/11})$, then (6.5) should hold. Table 6.5 shows the probability of ranking the smallest $j_0 = 10\, n^k$ scores correctly for various k and n, with 10,000 simulations. Again the normal error is tuned so that the $n = 5{,}000$ case has probability of close to $\frac{1}{2}$. The results suggest that $n^{1/11}$ indeed separates values of k for which correct ranking is possible.

6.5 Technical arguments

6.5.1 Sketch of proof and preliminary lemmas. We begin by giving a brief sketch of the proof of Theorem 6.1. Two steps in the proof are initially presented as lemmas, the first using moderate deviation properties to approximate sums related to the object of interest, and the second employing Taylor expansion applied to Rényi representations of order statistics to show that the gaps $\Theta_{(j+1)} - \Theta_{(j)}$ have a high probability of being of reasonable size. In the proof itself we use Lemma 6.3 to bound the probability in (6.5) from below (see (6.35)) and then show that the last two terms in this expression converge to zero, implying that the probability converges to 1 if (6.12) holds. For the converse, assuming independence, we find an upper bound to the probability in (6.36) and show that if this probability tends to one then the sum $s(n)$, introduced at (6.37), must converge to zero, which in turn implies (6.12). Only the exponential tail case is presented in detail; comments at the end of the proof describe the main differences in the polynomial tail case.

Throughout we let $\mathcal{E}(j_0)$ denote the event that $\bar{Q}_{R_j} + \Theta_{R_j} > \bar{Q}_{R_{j_0}} + \Theta_{R_{j_0}}$ for $j_0 + 1 \le j \le p$, we define $\mathcal{E}_j$ to be the event that $\Theta_{(j+1)} - \Theta_{(j)} \ge -(\bar{Q}_{R_{j+1}} - \bar{Q}_{R_j})$, and we take $\tilde{\mathcal{E}}(j_0)$ and $\tilde{\mathcal{E}}_j$ to be the respective complements. Also we let $\zeta_j = \Theta_{(j+1)} - \Theta_{(j)}$ denote the jth gap, where $\Theta_{(0)} = -\infty$ for convenience.

In Lemma 6.3 below we write $\mathcal{O}$ to denote the sigma-field generated by the Θ_js, N for a standard normal random variable independent of $\mathcal{O}$, δ_n for any given sequence of positive constants δ_n converging to zero, and Δ for a generic random variable satisfying $P(|\Delta| \le \delta_n) = 1$.

Lemma 6.3. *For any positive integer $j_0 < p$, let $\mathcal{J}$ denote the set of positive, even integers less than or equal to j_0. Put*

$$T_{1j} = \frac{\min(\zeta_{j-1}, \zeta_j)}{2\,(\mathrm{var}\bar{Q}_{R_j})^{1/2}}\,, \quad T_{2j} = \frac{\zeta_j}{\{\mathrm{var}(\bar{Q}_{R_{j+1}} - \bar{Q}_{R_j})\}^{1/2}}\,.$$

Then

$$\sum_{j=1}^{j_0} P\Big\{|\bar{Q}_{R_j}| > \tfrac{1}{2}\min\big(\zeta_{j-1},\zeta_j\big)\Big\}$$
$$= 2\,\{1+o(1)\}\sum_{j=1}^{j_0} P(|N| > T_{1j}) + o(1)\,. \tag{6.17}$$

If in addition the components of the Q_is are independent then

$$E\Big[\exp\Big\{-\sum_{j\in\mathcal{J}} P\big(\tilde{\mathcal{E}}_j \,|\, \mathcal{O}\big)\Big\}\Big]$$
$$\le \{1+o(1)\}\,E\Big[\exp\Big\{-(1+\Delta)\sum_{j\in\mathcal{J}} P(N > T_{2j}\,|\,\mathcal{O})\Big\}\Big] \tag{6.18}$$

Proof: Using the arguments of Rubin and Sethuraman (1965) and Amosova (1972) it can be shown that, if the constant C_2 in (6.9) satisfies $C_2 > B^2+2$ where $B>0$, then as n (and hence also p) diverges,

$$P\big\{|\bar{Q}_j| > x\,(\mathrm{var}\bar{Q}_j)^{1/2}\big\} = \{1+o(1)\}\,2\,\{1-\Phi(x)\}\,, \tag{6.19}$$
$$P\Big[-\big(\bar{Q}_{j_1}-\bar{Q}_{j_2}\big) \ge x\,\big\{\mathrm{var}\big(\bar{Q}_{j_1}-\bar{Q}_{j_2}\big)\big\}^{1/2}\Big]$$
$$= \{1+o(1)\}\,\{1-\Phi(x)\}\,, \tag{6.20}$$

uniformly in $0 < x < B\,(\log p)^{1/2}$ and $j, j_1, j_2 \ge 1$ such that $j_1 \ne j_2$. Expression (6.20) requires the independence assumption. Therefore, since $C_2 > 2\,(C_1+1)$ in (6.9), we can take $B = (2+\epsilon)^{1/2}$ for some $\epsilon > 0$, and then (6.19) and (6.20) hold uniformly in $0 < x < \{(2+\epsilon)\log p\}^{1/2}$. Thus as $n\to\infty$, they hold uniformly in all $x > 0$, modulo an $o(p^{-1})$ term. We use (6.19) to derive (6.17), while (6.20) implies that

$$\sum_{j\in\mathcal{J}} P\big(\tilde{\mathcal{E}}_j\big) = \{1+o(1)\}\sum_{j\in\mathcal{J}} P(N > T_{2j}) + o(1)\,,$$

which leads to (6.18).

Lemma 6.4. *If* (6.7)*, indicating the case of exponential tails, holds then there exist $B_4, B_5 > 0$ such that, for any choice of constants c_1, c_2 satisfying $0 < c_1 < c_2 < (4-\epsilon)^{-1}$ with ϵ as in* (6.11)*, and for all $B_6 > 0$,*

$$\inf_{j\in[1,n^{c_1}]} P\Big\{\zeta_j\, Z_{j+1}^{-1}\,(\log n)^{1-(1/\alpha)} \ge B_4\, n^{-c_1}\Big\} = 1 - O\big(n^{-B_6}\big)\,, \tag{6.21}$$
$$\inf_{j\in[n^{c_1},n^{c_2}]} P\Big\{B_4 \le j\,\zeta_j\, Z_{j+1}^{-1}\,(\log n)^{1-(1/\alpha)} \le B_5\Big\} = 1 - O\big(n^{-B_6}\big)\,. \tag{6.22}$$

Note further that the constraint on c_2 permits n^{c_2} to be of size $\nu_{\exp}\, n^{\epsilon_1}$ (where $\epsilon_1 > 0$).

Proof: If $U_{(1)} \leq \ldots \leq U_{(p)}$ denote the order statistics of a sample of size p drawn from the uniform distribution on $[0,1]$ then, for each p, we can construct a collection of independent random variables $Z_1, \ldots, Z_p$ with the standard negative exponential distribution on $[0,\infty]$, such that, for $1 \leq j \leq p$, $U_{(j)} = 1 - \exp(-V_j)$ where

$$V_j = \sum_{k=1}^{j} \frac{Z_k}{p-k+1} = w_j + W_j\,.$$

For details see Rényi (1953). Further, uniformly in $1 \leq j \leq \frac{1}{2}\,p$ and $2 \leq p < \infty$,

$$w_j = \sum_{k=p-j+1}^{p} \frac{1}{k} = \frac{j}{p} + O\big(j^2/p^2\big) = O(j/p)\,, \tag{6.23}$$

$$W_j = \sum_{k=p-j+1}^{p} k^{-1}\,(Z_{p-k+1} - 1)\,, \quad \sup_{1\leq j\leq p/2} j^{-1/2}\,|W_j| \leq p^{-1}\,W(p)\,, \tag{6.24}$$

$$\sup_{1\leq j\leq p/2} j^{-3/2}\left|W_j - \frac{1}{p}\sum_{k=p-j+1}^{p}(Z_{p-k+1}-1)\right| \leq p^{-2}\,W(p)\,, \tag{6.25}$$

where the nonnegative random variable $W(p)$, which without loss of generality we take to be common to (6.24) and (6.25), satisfies the expression $P\{W(p) > p^{\epsilon}\} = O(p^{-C})$ for each $C, \epsilon > 0$.

Using the second identity in (6.23), and (6.24), we deduce that

$$\begin{aligned} U_{(j+1)} - U_{(j)} &= (V_{j+1} - V_j)\Big\{1 - \tfrac{1}{2}\,(V_{j+1} + V_j) \\ &\qquad + \tfrac{1}{6}\,(V_{j+1}^2 + V_j\,V_{j+1} + V_j^2) - \ldots\Big\} \\ &= \frac{Z_{j+1}}{p-j}\left\{1 + \Psi_{j1}\left(\frac{j}{p} + \frac{S_{j1}}{p^{1/2}}\right)\right\}, \end{aligned} \tag{6.26}$$

uniformly in $1 \leq j \leq \frac{1}{2}\,p$, where the random variable Ψ_{j1} satisfies, for $k = 1$,

$$P\left(\max_{1\leq j\leq p/2} |\Psi_{jk}| \leq A\right) = 1\,, \tag{6.27}$$

$A > 0$ is an absolute constant, and for each $C, \epsilon > 0$ the nonnegative random variable S_{j1} satisfies, with $k = 1$,

$$P\left(\sup_{1\leq j\leq p/2} S_{jk} > p^{\epsilon}\right) = O\big(p^{-C}\big)\,. \tag{6.28}$$

Using the third identity in (6.23), and (6.25), we deduce that

$$0 \le U_{(j)} = w_j + W_j - \tfrac{1}{2}\,(w_j + W_j)^2 + \ldots = \frac{j}{p} + \Psi_{j2}\left(\frac{j^2}{p^2} + \frac{j^{1/2}\,S_{j2}}{p}\right), \qquad (6.29)$$

where Ψ_{j2} and $S_{j2} \ge 0$ satisfy (6.27) and (6.28), respectively.

Define $D_j = U_{(j+1)} - U_{(j)}$ and without loss of generality, $C_0 = 1$ in (6.7). If the common distribution function of the Θ_js is F then, by Taylor expansion,

$$\begin{aligned} \zeta_j &= F^{-1}(U_{(j)} + D_j) - F^{-1}(U_{(j)}) \\ &= D_j\left(F^{-1}\right)'(U_{(j)} + \omega_j\,D_j)\,, \\ &= \Psi_j\,\frac{D_j}{U_{(j)} + \omega_j\,D_j}\,\left\{-\log(U_{(j)} + \omega_j\,D_j)\right\}^{(1/\alpha)-1}, \end{aligned} \qquad (6.30)$$

where $0 \le \omega_j \le 1$ and the last line makes use of (6.7). The random variable Ψ_j satisfies, for constants B_1, B_2 and B_3 satisfying $0 < B_1 < B_2 < \infty$ and $0 < B_3 < 1$,

$$P\Big(B_1 \le \Psi_j \le B_2 \quad \text{for all } j \text{ such that } \quad U_{(j+1)} < B_3\Big) = 1\,.$$

The required result then follows from (6.26), (6.29) and (6.30).

6.5.2 Proof of Theorem 6.1. Take $j_0 < p$ a positive integer. Note that, taking $\mathcal{E}(j_o)$, $\mathcal{E}_j$, $\tilde{\mathcal{E}}(j_o)$, $\tilde{\mathcal{E}}_j$, $\mathcal{O}$ and $\mathcal{J}$ as for Lemma 6.3,

$$\begin{aligned} &\{\widehat{R}_j = R_j \ \text{ for } \ 1 \le j \le j_0\} \\ &\quad \supseteq \Big\{|\bar{Q}_{R_j}| \le \tfrac{1}{2}\,\min\left(\zeta_{j-1}, \zeta_j\right) \ \text{ for } \ 1 \le j \le j_0\Big\} \cap \mathcal{E}(j_0)\,, \end{aligned}$$

where we define $\Theta_{(j-1)} = -\infty$ if $j = 1$ as before. Therefore, defining $\pi(j_0) = P(\widehat{R}_j = R_j \ \text{ for } \ 1 \le j \le j_0)$, we deduce that

$$\pi(j_0) \ge 1 - \sum_{j=1}^{j_0} P\Big\{|\bar{Q}_{R_j}| > \tfrac{1}{2}\,\min\left(\zeta_{j-1}, \zeta_j\right)\Big\} - P\{\tilde{\mathcal{E}}(j_0)\}\,. \qquad (6.31)$$

Also,

$$\begin{aligned} \{\widehat{R}_j &= R_j \ \text{ for } \ 1 \le j \le j_0\} \\ &= \Big\{\bar{X}_{R_1} \le \ldots \le \bar{X}_{R_{j_0}} \ \text{ and } \ \bar{X}_j > \bar{X}_{R_{j_0}} \ \text{ for } \ j \notin \{R_1, \ldots, R_{j_0}\}\Big\} \\ &= \Big\{\zeta_j \ge -\big(\bar{Q}_{R_{j+1}} - \bar{Q}_{R_j}\big) \ \text{ for } \ 1 \le j \le j_0 \\ &\qquad\qquad \text{and } \ \Theta_j - \Theta_{(j_0)} \ge -\big(\bar{Q}_j - \bar{Q}_{R_{j_0}}\big) \ \text{ for } \ j \notin \{R_1, \ldots, R_{j_0}\}\Big\}, \end{aligned}$$

and so

$$\pi(j_0) \le P\Big\{\zeta_j \ge -\big(\bar{Q}_{R_{j+1}} - \bar{Q}_{R_j}\big) \ \text{ for } \ 1 \le j \le j_0\Big\}\,. \tag{6.32}$$

Letting $\pi_1(j_0)$ denote the probability that $\mathcal{E}_j$ holds for all $j \in \mathcal{J}$, by (6.32),

$$\pi(j_0) \le \pi_1(j_0)\,. \tag{6.33}$$

Note that if the components of each Q_i are independent, then the events $\mathcal{E}_j$, for $j \in \mathcal{J}$, are independent conditional on $\mathcal{O}$. Therefore,

$$\begin{aligned} \pi_1(j_0) = E\Big\{P\Big(\bigcap_{j\in\mathcal{J}} \mathcal{E}_j \,\Big|\, \mathcal{O}\Big)\Big\} &= E\Big[\prod_{j\in\mathcal{J}} \{1 - P(\mathcal{E}_j \,|\, \mathcal{O})\}\Big] \\ &\le E\Big[\exp\Big\{-\sum_{j\in\mathcal{J}} P\big(\tilde{\mathcal{E}}_j \,|\, \mathcal{O}\big)\Big\}\Big]\,. \end{aligned} \tag{6.34}$$

Using Lemma 6.3 we have the following inequalities regarding $\pi(j_0)$:

$$\pi(j_0) \ge 1 - 2\,\{1 + o(1)\} \sum_{j=1}^{j_0} P(|N| > T_{1j}) - P\{\tilde{\mathcal{E}}(j_0)\} + o(1) \tag{6.35}$$

$$\pi(j_0) \le \{1 + o(1)\}\, E\Big[\exp\Big\{-(1+\Delta)\sum_{j\in\mathcal{J}} P(N > T_{2j} \,|\, \mathcal{O})\Big\}\Big]\,. \tag{6.36}$$

To show that (6.12) implies (6.5), by (6.35) it is sufficient to show that $P\{\tilde{\mathcal{E}}(j_0)\}$ and $\sum_{j=1}^{j_0} P(|N| > T_{1j})$ are both $o(1)$, which we shall do in turn.

Define $\ell = (\log n)^{(1/\alpha)-1}$, let N be a standard normal random variable independent of $\mathcal{O}$, and let Z be independent of N and have the standard negative exponential distribution. Let K_1 be a positive constant. If a_n is a sequence of positive numbers and f_n is a sequence of nonnegative functions, write $a_n \doteq f_n(K)$ to mean that, for constants $L_1, L_2 > 1$, either (a) $a_n \le L_1\, f_n(K)$ whenever $K \ge L_2$ and n is sufficiently large, and $a_n \ge L_1^{-1}\, f_n(K)$ whenever $K \le L_2^{-1}$ and n is sufficiently large, or (b) $a_n \ge L_1^{-1}\, f_n(K)$ whenever $K \ge L_2$ and n is sufficiently large, and $a_n \le L_1\, f_n(K)$ whenever $K \le L_2^{-1}$ and n is sufficiently large. Let $0 < c_1 < c_2 < \frac{1}{2}$ and $c_1 < \frac{1}{4}$, and let j_0 and j_1 denote integers satisfying $|j_1 - n^{c_1}| \le 1$, $j_1 \le j_0 \le n^{c_2}$ and $j_1/j_0 \to 0$.

When (6.7) holds with $C_0 = 1$, Lemma 6.4 implies that, for each $B_6 > 0$ and

letting $\gamma_j = n^{-1/2}\, j\, \ell^{-1}$,

$$s(n) \equiv \sum_{j=1}^{j_0} P\{|N| > K_1\, n^{1/2}\, (\zeta_j)\} \tag{6.37}$$

$$\begin{aligned}
&\doteq O\big\{j_1\, P\big(|N| > K_2 Z \gamma_{j_1}^{-1}\big) + n^{-B_6}\big\} + \sum_{j_1 < j \le j_0} P\big(|N| > K Z \gamma_j^{-1}\big) \\
&\doteq O\Big\{j_1 \Big(P\big(Z \le \gamma_{j_1}\big) + E\Big[Z^{-1}\gamma_{j_1}\, \exp\big\{-\tfrac{1}{2}\,(K\,Z\gamma_{j_1}^{-1})^2\big\}\, I\big(Z > \gamma_{j_1}\big)\Big]\Big)\Big\} \\
&\qquad + \sum_{j_1 < j \le j_0} \Big(P\big(Z \le \gamma_j\big) + E\Big[Z^{-1}\gamma_j\, \exp\big\{-\tfrac{1}{2}\,(K\,Z\gamma_j^{-1})^2\big\}\, I\big(Z > \gamma_j\big)\Big]\Big) \\
&\doteq O\Big\{j_1 \Big(\gamma_{j_1} + E\Big[Z^{-1}\gamma_{j_1} \exp\big\{-\tfrac{1}{2}\,(K\,Z\gamma_{j_1}^{-1})^2\big\}\, I\big(Z > \gamma_{j_1}\big)\Big]\Big)\Big\} \\
&\qquad + \sum_{j_1 < j \le j_0} \Big(\gamma_j + E\Big[Z^{-1}\gamma_j\, \exp\big\{-\tfrac{1}{2}\,(K\,Z\gamma_j^{-1})^2\big\}\, I\big(Z > \gamma_j\big)\Big]\Big)\,.
\end{aligned}$$

Now,

$$\begin{aligned}
&E\Big[Z^{-1}\gamma_j\, \exp\big\{-\tfrac{1}{2}\,(KZ\gamma_j^{-1})^2\big\}\, I\big(Z > \gamma_j\big)\Big] \\
&\quad = \int_{\gamma_j}^{\infty} z^{-1}\gamma_j\, \exp\big\{-\tfrac{1}{2}\,(KZ\gamma_j^{-1})^2 - z\big\}\, dz \\
&\quad = \gamma_j \int_1^{\infty} u^{-1}\, \exp\big\{-\tfrac{1}{2}\,(K\,u)^2 - \gamma_j\, u\big\}\, du \asymp \gamma_j = n^{-1/2}\, j\, \ell\,.
\end{aligned}$$

(Here we have used the fact that $j \le j_0 \le n^{c_2}$ where $c_2 < \frac{1}{2}$.) Therefore,

$$\begin{aligned}
s(n) &\asymp j_1 \cdot n^{-1/2}\, j_1\, \ell^{-1} + \sum_{j_1 < j \le j_0} n^{-1/2}\, j\, \ell^{-1} \\
&\asymp n^{-1/2}\, j_1^2\, \ell^{-1} + n^{-1/2}\, j_0^2\, \ell^{-1} \\
&\asymp n^{-1/2}\, j_0^2\, \ell^{-1}\,.
\end{aligned} \tag{6.38}$$

(Here we have used the fact that $j_1/j_0 \to 0$.)

The right-hand side of (6.38) converges to zero if and only if (6.12) holds. Moreover, in view of the fact that

$$P(|N| > T_{1j}) \le P\left(|N| > \frac{\zeta_{j-1}}{2\,(\mathrm{var}\bar{Q}_{R_j})^{1/2}}\right) + P\left(|N| > \frac{\zeta_j}{2\,(\mathrm{var}\bar{Q}_{R_j})^{1/2}}\right)\,,$$

and depending on the choice of K_1 in the definition of $s(n)$ at (6.37), $s(n)$ can be an upper bound to the series $\sum_{j=1}^{j_0} P(|N| > T_{1j})$ on the right-hand side of (6.17).

Hence,

$$\sum_{j=1}^{j_0} P(|N| > T_{1j}) = o(1)\,. \tag{6.39}$$

This deals with the second term on the right-hand side of (6.35). Similarly, if $r \in [2,\infty)$ is a fixed integer, and if $j_0 = o(n^{1/4}\,\ell^{1/2})$, then

$$s_1(n) \equiv \sum_{j=j_0+1}^{j_0+r-1} P\big\{|N| > K_1\, n^{1/2}\,(\zeta_j)\big\} = o(1)\,. \tag{6.40}$$

Moreover, if j_1 denotes the integer part of $n^{c_2} - j_0$ then, for constants K_2 and K_3 satisfying $K_1 > K_2 > K_3 > 0$, and for any $B > 0$,

$$\begin{aligned} s_2(n) &\equiv \sum_{j=j_0+r}^{j_0+j_1} P\big\{|N| > K_1\, n^{1/2}\,(\Theta_{(j+1)} - \Theta_{(j_0)})\big\} \\ &\le \sum_{j=r}^{j_1} P\Big\{|N| > K_2\, n^{1/2}\,\ell \sum_{k=1}^{j} (j_0+k)^{-1}\, Z_k\Big\} + O\big(n^{-B}\big) \\ &\le j_1\, P\big\{|N| > K_2\, n^{1/4}\,\ell^{1/2}\,(Z_1 + \ldots + Z_r)\big\} + O\big(n^{-B}\big) \\ &= O\Big\{j_1\,\big(n^{1/2}\,\ell^2\big)^{-r}\Big\}\,, \end{aligned} \tag{6.41}$$

where we have assumed that $j_0 = o(n^{1/4}\,\ell^{1/2})$ and also used the fact that $Z_1 + \ldots + Z_r$ has a gamma$(r,1)$ distribution. If we choose r so large that $p\,n^{-r/2} = O(n^{-\epsilon})$ for some $\epsilon > 0$ then we can deduce from (6.40) and (6.41) that $s_1(n) + s_2(n) \to 0$, and hence, by (6.22), that

$$\sum_{j=j_0+1}^{n^{c_2}} P(\bar{Q}_{R_j} + \Theta_{R_j} > \bar{Q}_{R_{j_0}} + \Theta_{R_{j_0}}) \to 0\,. \tag{6.42}$$

A more crude argument can be used to prove that if r is so large that $p^2\,n^{-r/2} = O(n^{-\epsilon})$ for some $\epsilon > 0$, and if $j_0 = o(n^{1/4}\,\ell^{1/2})$, then

$$\sum_{n^{c_2} < j \le p} P(\bar{Q}_{R_j} + \Theta_{R_j} > \bar{Q}_{R_{j_0}} + \Theta_{R_{j_0}}) \to 0\,. \tag{6.43}$$

Together, (6.42) and (6.43) imply that if $j_0 = o(n^{1/4}\,\ell^{1/2})$ then

$$P\{\tilde{\mathcal{E}}(j_0)\} \to 0\,. \tag{6.44}$$

Thus in light of (6.35), we see (6.39) and (6.44) imply that (6.12) is sufficient for (6.5).

We next show that (6.5) implies (6.12) in the independent case. If (6.5) holds,

then by (6.36),

$$\sum_{j \in \mathcal{J}} P(N > T_{2j} \,|\, \mathcal{O}) \to 0$$

in probability. Therefore, by Lemma 6.4, with j_0 and j_1 as above, there exists $K_1 > 0$ such that

$$\sum_{j_1 < j \le j_0} P\{|N| > n^{1/2} K_1 (\zeta_j) \,|\, \mathcal{O}\} \to 0$$

in probability. (We can take the sum over all $j \in [j_1 + 1, j_0]$, rather than just over even j, since (6.18) holds for sums over odd j as well as over even j.) Hence, arguing as in the lines below (6.37), we deduce that for sufficiently large $K_2 > 0$,

$$\sum_{j_1 < j \le j_0} f(Z_j / \delta_j) \to 0 \tag{6.45}$$

in probability, where the random variables Z_j are independent and have a common exponential distribution, $\delta_j = n^{-1/2} j \ell^{-1}$ and

$$f(z) = z^{-1} \exp(-K_2 z^2) I(z > 1) .$$

We claim that this implies that the expected value of the left-hand side of (6.45) also converges to 0:

$$\sum_{j_1 < j \le j_0} E\{f(Z_j / \delta_j)\} \to 0 , \tag{6.46}$$

or equivalently that $\sum_{j_1 < j \le j_0} \delta_j \to 0$, and thence (using the argument leading to (6.38)) that $s(n) \asymp n^{-1/2} j_0^2 \ell^{-1} \to 0$, which is equivalent to (6.12). Therefore, if we establish (6.46) then we shall have proved that (6.5) implies (6.12).

It remains to show that (6.45) implies (6.46). This we do by contradiction. If (6.46) fails then, along a subsequence of values of n, the left-hand side of (6.46) converges to a nonzero number. For notational simplicity we shall make the inessential assumptions that the number is finite and that the subsequence involves all n, and we shall take $K_2 = 1$ in the definition of f. In particular,

$$t(n) \equiv \sum_{j_1 < j \le j_0} E\{f(Z_j / \delta_j)\} \to t(\infty) , \tag{6.47}$$

where $t(\infty)$ is bounded away from 0. Now, $t(n) = \{1 + o(1)\} \mu(1) \delta(n)$, where $\delta(n) = \sum_{j_1 < j \le j_0} \delta_j$ and, for general $\lambda \ge 1$, $\mu(\lambda) = \int_{z > \lambda} z^{-1} \exp(-z^2) \, dz$. Therefore,

$$\delta(n) \to \delta(\infty) \equiv t(\infty) / \mu(1) . \tag{6.48}$$

For each $\lambda > 1$ the left-hand side of (6.45) equals $\Delta_1 + \Delta_2$, where, in view of (6.47),

$$E(\Delta_2) = \sum_{j_1 < j \le j_0} E\{f(Z_j/\delta_j)\, I(Z_j > \lambda\, \delta_j)\} = \{1 + o(1)\}\, \mu(\lambda)\, \delta(n) \tag{6.49}$$

and

$$\Delta_1 = \sum_{j_1 < j \le j_0} f(Z_j/\delta_j)\, I(Z_j \le \lambda\, \delta_j) = \sum_{j_1 < j \le j_0} f(W_j)\, I_j\,,$$

with $W_j = Z_j/\delta_j$ and $I_j = I(\delta_j \le Z_j \le \lambda\, \delta_j)$. However,

$$\sum_{j_1 < j \le j_0} P(I_j = 1) = \mu_1(\lambda)\, \delta(n) + o(1) = \delta(\infty)\, \mu_1(\lambda) + o(1)\,,$$

where $\mu_1(\lambda) = \int_{1<z<\lambda} z^{-1} \exp(-z^2)\, dz$. Therefore, in the limit as $n \to \infty$, Δ_1 equals a sum, S_λ say, of N independent random variables each having the distribution of $f(W)$, where W is uniformly distributed on $[1, \lambda]$, N has a Poisson distribution with mean $\delta(\infty)\, \mu_1(\lambda)$, and N and the summands are independent. The distribution of S_λ is stochastically monotone increasing, in the sense that $P(S_\lambda > s)$ increases with λ. On the other hand, since $\mu(\lambda) \to 0$ as $\lambda \to \infty$ then, by (6.48) and (6.49),

$$\lim_{\lambda \to \infty} \limsup_{n \to \infty} E(\Delta_2) = 0\,.$$

Combining these results we deduce that $\Delta_1 + \Delta_2$, i.e. the left-hand side of (6.45), does not converge to zero in probability. This contradicts (6.45) and so establishes that $t(\infty)$ must equal zero; that is, (6.46) holds.

6.5.3 Comments on proving the polynomial case. The proof for the case of polynomial tails proceeds similarly. The main difference is that in the proof of Lemma 6.4 we use (6.8) instead of (6.7), which forces a factor of $p^{-1/\alpha}$ into the results of the lemma, rather than $(\log n)^{1-(1/\alpha)}$. This in turn implies that $s(n) \asymp n^{-1/2} j_0^{2+1/\alpha} p^{-1/\alpha}$, entailing that convergence occurs if (and, in the case of independence, only if) $j_0 = o(\nu_{\rm pol})$, as required.

Chapter 7

Confidence intervals for parameter extrema

7.1 Background

Chapter 5 explored in detail the use of the bootstrap to assess the uncertainty of a ranking. One of the key results was that the standard n-out-of-n bootstrap fails to give asymptotic consistency when the measures for various populations are close. Thus, bootstrap methods can face serious difficulties when used to estimate the distributions of extrema of parameter estimators, for example of $\max(\hat{\theta}_1, \ldots, \hat{\theta}_p)$ where $\hat{\theta}_1, \ldots, \hat{\theta}_p$ are estimators of the respective values of parameters $\theta_1, \ldots, \theta_p$. The reason is that the asymptotic distribution of $\max_j \hat{\theta}_j$ can be non-normally distributed. This is consistent with a bootstrap metatheorem which argues that, in a range of settings, bootstrap methods give consistent results for estimating distributions of parameter estimators "if and only if" the limiting distribution is normal (see e.g. Mammen, 1992). Consequently, since the joint distribution of $\hat{\theta}_1 - \theta_1, \ldots, \hat{\theta}_p - \theta_p$ generally is asymptotically normal, that distribution can typically be estimated accurately even though the limiting distribution of $\max_{j \le p} \hat{\theta}_j$ might not be estimable by any method, be it the bootstrap or another approach. This property underpins the methodology introduced in the present chapter.

While the m-out-of-n bootstrap explored in Chapter 5 (see also e.g. Swanepoel, 1986; Hall, 1990; Bickel and Ren, 1996; Bickel et al., 1997; Politis et al., 1999) appears to overcome some the problems faced by the standard bootstrap in estimating the distribution of quantities such as $\max_{j \le p} \hat{\theta}_j$, practical difficulties can still exist (See e.g. Andrews, 2000). Even in problems where the m-out-of-n bootstrap enjoys attractive asymptotic properties, it can exhibit very poor finite-sample performance

because the noise introduced through estimating the tuning parameter, m, in the m-out-of-n bootstrap can seriously degrade performance. As a result, the m-out-of-n bootstrap can produce confidence intervals and hypothesis tests with serious anticonservative level inaccuracies.

Although these challenges stand in the way of accurate statistical methodology, the problem of making inference about extrema of parameters is important because it arises in a variety of contexts, in fields ranging from frontier analysis (see Berger and Humphrey, 1997; Kim et al., 2007) to methodology based on empirical eigenvalues (e.g. Ringrose and Benn, 1997; Schott, 2006). Contributors to the area aside from those already mentioned in Chapter 5 include Beran (1982), Bretagnolle (1983), Beran and Srivastava (1985), Dümbgen (1993), Hall et al. (1993) and Andrews (1999, 2000). Nevertheless, there still do not exist methods that overcome effectively the difficulties discussed above.

In the present chapter we show that, in a variety of problems involving hypothesis tests and confidence intervals for extrema of general linear combinations of parameters, an indirect application of the bootstrap can overcome the difficulties. Our approach involves implicitly constructing a bootstrap confidence interval for the joint distribution of $\hat{\theta}_1 - \theta_1, \ldots, \hat{\theta}_p - \theta_p$, and using simple monotonicity arguments to construct tests or confidence intervals that are guaranteed to be conservative except for a small bootstrap error. We suggest using a double bootstrap approach to ensure good accuracy. The conservatism of our methodology derives from the fact that our bootstrap methods, for both confidence intervals and hypothesis tests, are based on resampling from the least favourable null distribution, interpreted in a nonparametric context.

An interesting feature of the problems treated in this chapter is that, in the cases that cause most difficulty, the distributions that we seek to approximate are asymmetric even in the asymptotic limit. In consequence, even two-sided confidence regions and two-sided hypothesis tests have level errors of order $n^{-1/2}$, not n^{-1}. Use of the double bootstrap is necessary to ensure that the bootstrap error is of order n^{-1} in two-sided, as well as one-sided, cases.

In today's computing environment, in cases there the $\hat{\theta}_j$s are relatively simple functions of the data, and when only a single sample is involved, it is often feasible to use triple bootstrap methods and thereby reduce error to order $n^{-3/2}$. However, it would be impractical to explore the effectiveness of the triple bootstrap in a simulation study, since this would increase computational labour several hundred fold. In principle, analytical rather than bootstrap methods, based on theoretical expressions for high-order terms in asymptotic formulae, might be used to effect further corrections. However, we feel that the complexity of the formulae involved, and the fact that they change from one problem to another, make this type of correction

unattractive.

Although the intervals and tests tend towards conservatism, they are constructed so that, in difficult cases where there are ties for unknown parameter values, the procedures are less conservative, and in some cases asymptotically exact, modulo the bootstrap approximation. We explore the extent of this conservatism in our numerical work; it varies from being vanishingly small, when ties present, to being more significant in other cases.

We suggest a general approach that enables us to construct conservative tests and confidence intervals for a very wide variety of statistics based on parameter extrema. The quantities that can be addressed using our methodology include confidence intervals for, and hypothesis tests about, functions of the parameters $\theta_1, \ldots, \theta_p$ such as the following:

$$\begin{aligned}&\max_{1\le j\le p} \theta_j\,, \quad \min_{1\le j\le p} \theta_j\,, \quad \max_{j\in\mathcal{J}_1} \theta_j \;\pm\; \max_{j\in\mathcal{J}_2} \theta_j\,, \quad \min_{j\in\mathcal{J}_1} \theta_j \;\pm\\ &\max_{j\in\mathcal{J}_2} \theta_j\,, \quad \max\Bigl(\max_{1\le j\le p} \theta_j\,, C\Bigr), \quad \max\Bigl(\min_{1\le j\le p} \theta_j\,, C\Bigr),\end{aligned} \tag{7.1}$$

where $\mathcal{J}_1$ and $\mathcal{J}_2$ are arbitrary subsets of $1, \ldots, p$, and C is any known constant. In practice $\mathcal{J}_1$ and $\mathcal{J}_2$ would usually be disjoint, but even this condition is not required for the general problem solved in Section 7.2.

7.2 Methodology

7.2.1 Problem setup. First we give notation used throughout. Let $\mathcal{J}_1$ and $\mathcal{J}_2$ be nonempty subsets of $\{1, \ldots, p\}$, and, for $k = 1$ and 2, put $\max_{(k)} = \max_{j\in\mathcal{J}_k} \theta_j$ and $\widehat{\max}_{(k)} = \max_{j\in\mathcal{J}_k} \hat\theta_j$. Define $\min_{(k)}$ and $\widehat{\min}_{(k)}$ analogously. Our general treatment allows either extreme of each population to be considered, so write $\mathrm{mm}_{(k)}$ to denote either $\min_{(k)}$ or $\max_{(k)}$, and define $\widehat{\mathrm{mm}}_{(k)}$ to equal $\widehat{\max}_{(k)}$ if we took $\mathrm{mm}_{(k)}$ to equal $\max_{(k)}$, and to equal $\widehat{\min}_{(k)}$ otherwise. We wish to test H_0 against H_1, where

$$H_0 : \mathrm{mm}_{(1)} \ge \mathrm{mm}_{(2)}\,, \quad H_1 : \mathrm{mm}_{(1)} < \mathrm{mm}_{(2)}\,, \tag{7.2}$$

and again the test errs on the conservative side; or we wish to construct a confidence interval with the property:

$$P\Bigl\{\mathrm{mm}_{(2)} - \mathrm{mm}_{(1)} \in \bigl[\widehat{\mathrm{mm}}_{(2)} - \widehat{\mathrm{mm}}_{(1)} - c_\alpha, \infty\bigr)\Bigr\} \ge 1 - \alpha\,.$$

In practice, c_α will not be known and must be estimated. For this step we use the bootstrap to compute an empirical approximation, $\hat c_\alpha$ say, to c_α. The error committed here is quite low, particularly if the double bootstrap is employed, since our methodology ensures that c_α is defined in terms of the joint distribution of the

centred variables $\hat{\theta}_j - \theta_j$; that distribution is relatively accessible.

Our discussion above, and our development of methodology below, focuses on one-sided conservative procedures. However, there is no difficult extending our methodology to two-sided approaches in the usual way, by taking the intersection of two one-sided procedures.

7.2.2 Obtaining conservative tests. Extend the definition of $\mathrm{mm}_{(k)}$ in above, by writing $\mathrm{mm}_{j\in\mathcal{J}_k}\,\theta_j$ for $\min_{j\in\mathcal{J}_k}\,\theta_k$ or $\max_{j\in\mathcal{J}_k}\,\theta_j$, according to whether $\mathrm{mm}_{(k)}$ denotes the former or latter. Also define $\mathrm{mm}_{j\in\mathcal{J}_k}\,x_j$ for a general quantity x_j, and note that:

$$\begin{aligned}
P\big(\mathrm{mm}_{(2)} \;-\; & \mathrm{mm}_{(1)} + c_\alpha \ge \widehat{\mathrm{mm}}_{(2)} - \widehat{\mathrm{mm}}_{(1)}\big) \\
&= P\Big\{\mathrm{mm}_{(2)} - \mathrm{mm}_{(1)} + c_\alpha \ge \mathop{\mathrm{mm}}_{j\in\mathcal{J}_2}(\hat{\theta}_j - \theta_j + \theta_j) - \mathop{\mathrm{mm}}_{j\in\mathcal{J}_1}(\hat{\theta}_j - \theta_j + \theta_j)\Big\} \\
&\ge P\Big[\mathrm{mm}_{(2)} - \mathrm{mm}_{(1)} + c_\alpha \ge \mathop{\mathrm{mm}}_{j\in\mathcal{J}_2}\Big\{\max_{k\in\mathcal{J}_2}(\hat{\theta}_k - \theta_k) + \theta_j\Big\} \\
&\qquad\qquad - \mathop{\mathrm{mm}}_{j\in\mathcal{J}_1}\Big\{\min_{k\in\mathcal{J}_1}(\hat{\theta}_k - \theta_k) + \theta_j\Big\}\Big] \\
&= P\Big\{c_\alpha \ge \max_{k\in\mathcal{J}_2}(\hat{\theta}_k - \theta_k) - \min_{k\in\mathcal{J}_1}(\hat{\theta}_k - \theta_k)\Big\} = 1-\alpha\,, \qquad (7.3)
\end{aligned}$$

where the final identity holds provided that we define c_α by

$$P\Big\{\min_{k\in\mathcal{J}_1}(\hat{\theta}_k - \theta_k) - \max_{k\in\mathcal{J}_2}(\hat{\theta}_k - \theta_k) \ge -c_\alpha\Big\} = 1-\alpha\,.$$

Hence, if we take $\hat{c}_\alpha$ to be an empirical approximation to c_α, then, no matter what our choice of min and max in the definitions of $\mathrm{mm}_{(1)}$ and $\mathrm{mm}_{(2)}$, the confidence interval

$$\big[\widehat{\mathrm{mm}}_{(2)} - \widehat{\mathrm{mm}}_{(1)} - \hat{c}_\alpha, \infty\big)$$

covers $\mathrm{mm}_{(2)} - \mathrm{mm}_{(1)}$ with probability at least $1-\alpha$, modulo any error in the approximation of c_α by $\hat{c}_\alpha$. Analogously, the hypothesis test that rejects $H_0 : \mathrm{mm}_{(1)} \ge \mathrm{mm}_{(2)}$, in favour of $H_1 : \mathrm{mm}_{(1)} < \mathrm{mm}_{(2)}$, if $\widehat{\mathrm{mm}}_{(2)} - \widehat{\mathrm{mm}}_{(1)} - \hat{c}_\alpha > 0$, has level at most α, except for the error in the bootstrap approximation:

$$\begin{aligned}
P\big(\widehat{\mathrm{mm}}_{(2)} \;-\; & \widehat{\mathrm{mm}}_{(1)} - \hat{c}_\alpha > 0 \,\big|\, \mathrm{mm}_{(2)} \le \mathrm{mm}_{(1)}\big) \\
&\le P\big(\widehat{\mathrm{mm}}_{(2)} - \widehat{\mathrm{mm}}_{(1)} - \hat{c}_\alpha > \mathrm{mm}_{(2)} - \mathrm{mm}_{(1)} \,\big|\, \mathrm{mm}_{(2)} \le \mathrm{mm}_{(1)}\big) \\
&= P\Big\{\mathrm{mm}_{(2)} - \mathrm{mm}_{(1)} \notin \big[\widehat{\mathrm{mm}}_{(2)} - \widehat{\mathrm{mm}}_{(1)} - \hat{c}_\alpha, \infty\big) \;\Big|\; \mathrm{mm}_{(2)} \le \mathrm{mm}_{(1)}\Big\} \\
&\le 1 - (1-\alpha) = \alpha\,, \qquad (7.4)
\end{aligned}$$

where the inequality in (7.3) follows from (7.4).

It is worth mentioning again that our focus on confidence intervals and hypothesis

tests based on the difference $\text{mm}_{(2)} - \text{mm}_{(1)}$ does not restrict us to quantities such as $\max_{j\in\mathcal{J}_2}\theta_j - \min_{j\in\mathcal{J}_1}\theta_j$. By taking one or more of the components of θ to be null we can treat any of the quantities in (7.1), and others, in the way discussed above. To make this explicit, Table 7.1 below shows a broad range of hypotheses that may be treated, along with the corresponding confidence interval form and the equation defining c_α. As noted before, conservative two-sided confidence intervals may be created by intersection. Notice that in the case of ties and when $\text{mm}_{(1)}$ and $\text{mm}_{(2)}$ denote minimum and maximum respectively the inequality in derivation (7.3) is in fact an equality; that is, the test is exact. This implies cases 1, 4 and 5 in Table 1 are potentially exact. In other cases, the tied case is conservative rather than exact, and this conservatism will tend to grow with the number of populations under consideration.

Case	H_0	H_1	Conf. Interval
1	$\min_{(1)} \geq C$	$\min_{(1)} < C$	$\min_{(1)} \in (-\infty, \widehat{\min}_{(1)} + c_\alpha]$
2	$\min_{(1)} \leq C$	$\min_{(1)} > C$	$\min_{(1)} \in [\widehat{\min}_{(1)} - c_\alpha, \infty)$
3	$\max_{(1)} \geq C$	$\max_{(1)} < C$	$\max_{(1)} \in (-\infty, \widehat{\max}_{(1)} + c_\alpha]$
4	$\max_{(1)} \leq C$	$\max_{(1)} > C$	$\max_{(1)} \in [\widehat{\max}_{(1)} - c_\alpha, \infty)$
5	$\min_{(1)} \geq \max_{(2)}$	$\min_{(1)} < \max_{(2)}$	$\max_{(2)} - \min_{(1)} \in [\widehat{\max}_{(2)} - \widehat{\min}_{(1)} - c_\alpha, \infty)$
6	$\min_{(1)} \leq \max_{(2)}$	$\min_{(1)} > \max_{(2)}$	$\min_{(1)} - \max_{(2)} \in [\widehat{\min}_{(1)} - \widehat{\max}_{(2)} - c_\alpha, \infty)$
7	$\max_{(1)} \geq \max_{(2)}$	$\max_{(1)} < \max_{(2)}$	$\max_{(2)} - \max_{(1)} \in [\widehat{\max}_{(2)} - \widehat{\min}_{(1)} - c_\alpha, \infty)$
8	$\min_{(1)} \geq \min_{(2)}$	$\min_{(1)} < \min_{(2)}$	$\min_{(2)} - \min_{(1)} \in [\widehat{\max}_{(2)} - \widehat{\min}_{(1)} - c_\alpha, \infty)$

Case	Equation for c_α
1	$P\{\min_{k\in\mathcal{J}_1}(\hat{\theta}_k - \theta_k) \geq -c_\alpha\} = 1-\alpha$
2	$P\{\max_{k\in\mathcal{J}_1}(\hat{\theta}_k - \theta_k) \leq c_\alpha\} = 1-\alpha$
3	$P\{\min_{k\in\mathcal{J}_1}(\hat{\theta}_k - \theta_k) \geq -c_\alpha\} = 1-\alpha$
4	$P\{\max_{k\in\mathcal{J}_1}(\hat{\theta}_k - \theta_k) \leq c_\alpha\} = 1-\alpha$
5	$P\{\max_{k\in\mathcal{J}_2}(\hat{\theta}_k - \theta_k) - \min_{k\in\mathcal{J}_1}(\hat{\theta}_k - \theta_k) \leq c_\alpha\} = 1-\alpha$
6	$P\{\max_{k\in\mathcal{J}_1}(\hat{\theta}_k - \theta_k) - \min_{k\in\mathcal{J}_2}(\hat{\theta}_k - \theta_k) \leq c_\alpha\} = 1-\alpha$
7	$P\{\max_{k\in\mathcal{J}_2}(\hat{\theta}_k - \theta_k) - \min_{k\in\mathcal{J}_1}(\hat{\theta}_k - \theta_k) \leq c_\alpha\} = 1-\alpha$
8	$P\{\max_{k\in\mathcal{J}_2}(\hat{\theta}_k - \theta_k) - \min_{k\in\mathcal{J}_1}(\hat{\theta}_k - \theta_k) \leq c_\alpha\} = 1-\alpha$

Table 7.1: Possible hypothesis tests of extremes, along with corresponding confidence intervals and equations for obtaining c_α.

7.3 Approximating distributions of extrema of estimators

7.3.1 Models, and the challenges of distribution approximations. For specificity in this section we treat the case where the quantity of interest is $\min_{j\in\mathcal{J}_1} - \max_{j\in\mathcal{J}_2}$, where $\mathcal{J}_1 = \{1,\ldots,r\}$, $\mathcal{J}_2 = \{r+1,\ldots,p\}$ and $1 \leq r \leq p-1$. In regular problems the estimators $\hat{\theta}_j$ are root-n consistent and asymptotically normally distributed, where n denotes sample size. Therefore it is reasonable to suppose that

we can write

$$\hat{\theta}_j = \theta_j + n^{-1/2}\, N_j \quad \text{for} \quad 1 \le j \le p\,,$$

where $N_1, \ldots, N_p$ have a joint limiting normal distribution with zero mean. In this context, we explore properties of

$$\widehat{\omega} \equiv \min_{j \in \mathcal{J}_1} \hat{\theta}_j - \max_{j \in \mathcal{J}_2} \hat{\theta}_j = \min_{j \in \mathcal{J}_1} \left(\theta_j + n^{-1/2}\, N_j\right) - \max_{j \in \mathcal{J}_2} \left(\theta_j + n^{-1/2}\, N_j\right). \tag{7.5}$$

The root-n consistency condition admits a wide variety of potential statistics, including all of those previously introduced for ranking variables; means, quantiles and various types of correlations all fall into this category.

The difficulties arise even under very generous assumptions, for example if $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)$ is the mean of a sample of size n from a normal $\mathrm{N}(\theta, \Sigma)$ distribution where Σ is known. We shall show that, even in this case, difficulties with near ties can make it impossible to estimate consistently the asymptotic distribution of $\widehat{\omega}$, at (7.5), no matter whether we use the bootstrap or any other method.

To give a simple, specific example, assume that

$$\theta_2 = \theta_1 + n^{-1/2}\, \nu \text{ and } \theta_j = \theta_1 \text{ for } 3 \le j \le p, \text{ where } \nu \text{ is a fixed constant}\,. \tag{7.6}$$

Standard information-theoretic arguments show that ν is not identifiable, in the sense that it cannot be estimated consistently from data. Now, (7.5) and (7.6) imply that

$$n^{1/2}\, \widehat{\omega} = \min(N_1, N_2 + \nu, N_3, \ldots, N_r) - \max(N_{r+1}, \ldots, N_p)\,,$$

where, when $\hat{\theta}$ is the estimated mean of an n-sample from a normal $\mathrm{N}(\theta, \Sigma)$ population, $(N_1, \ldots, N_p)$ is distributed as normal $\mathrm{N}(0, \Sigma)$. The shape of the distribution of $n^{1/2}\, \widehat{\omega}$, not just its location and scale, depend in detail on the non-estimable quantity ν.

The null hypothesis H_0, at (7.2), holds if $\nu \ge 0$, and the alternative obtains if $\nu < 0$. A conventional approach to testing H_0 would be to estimate the null distribution of $\widehat{\omega}$ for $\nu \ge 0$, and to reject H_0 if the value of $\widehat{\omega}$ were less than an estimator of a lower critical point for the distribution of $\widehat{\omega}$. However, since the distribution of $\widehat{\omega}$ cannot be estimated consistently then this approach is not viable.

7.3.2 Using the bootstrap to estimate the distribution of the centred version of $\widehat{\omega}$.

Recall the definition of $\widehat{\omega}$ at (7.5). In the case of the example suggested in Section 7.3.1 the centred version of $\widehat{\omega}$ with which we work is

$$\widetilde{\omega} \equiv \min_{j \in \mathcal{J}_1} \left(\hat{\theta}_j - \theta_j\right) - \max_{j \in \mathcal{J}_2} \left(\hat{\theta}_j - \theta_j\right).$$

If $\hat\theta_j$ is constructed from a random sample vector $\mathcal{X} = \{X_1, \ldots, X_n\}$, and if $\mathcal{X}^* = \{X_1^*, \ldots, X_n^*\}$ denotes a resample drawn from $\mathcal{X}$ by sampling randomly, with replacement, then we write $\hat\theta_j^*$ for the version of $\hat\theta_j$ computed from $\mathcal{X}^*$ rather than $\mathcal{X}$. The bootstrap form of $\widetilde\omega$ is

$$\widetilde\omega^* = \min_{j \in \mathcal{J}_1} \left(\hat\theta_j^* - \hat\theta_j\right) - \max_{j \in \mathcal{J}_2} \left(\hat\theta_j^* - \hat\theta_j\right).$$

A percentile-bootstrap estimator of the distribution function F of $\widetilde\omega$, defined by $F(x) = P(\widetilde\omega \le x)$, is given by $\widehat F(x) = P(\widetilde\omega^* \le x \,|\, \mathcal{X})$.

The theoretical critical point c_α, corresponding to the confidence interval in case 5 of Table 7.1, is defined by $F(-c_\alpha) = \alpha$. We could define its estimator, $\hat c_\alpha$, simply as the solution, $c = \tilde c_\alpha$ say, of $\widehat F(-c) = \alpha$. However, a greater degree of accuracy is obtained by using the double bootstrap in this step, as follows. Given the resample $\mathcal{X}^*$, let $\mathcal{X}^{**} = \{X_1^{**}, \ldots, X_n^{**}\}$ denote a re-resample drawn by sampling randomly, with replacement, from $\mathcal{X}^*$, and write $\widetilde\omega^{**}$ for the version of $\widetilde\omega$ computed from $\mathcal{X}^{**}$ rather than $\mathcal{X}$. Put $\widehat F^*(x) = P(\widetilde\omega^{**} \le x \,|\, \mathcal{X}^*)$, and let $c = \tilde c_\alpha^*$ be the solution of $\widehat F^*(-c) = \alpha$. Let $\beta = \hat\beta(\alpha)$ be the solution of $P(\widehat\omega^* \le -\tilde c_\beta^* \,|\, \mathcal{X}) = \alpha$. In this notation the double-bootstrap estimator of c_α is $\hat c_\alpha = \tilde c_{\hat\beta(\alpha)}$. Of course, in practice the probabilities $P(\widetilde\omega^{**} \le x \,|\, \mathcal{X}^*)$ and $P(\widehat\omega^* \le -\tilde c_\beta^* \,|\, \mathcal{X})$ usually cannot be computed exactly. They are instead calculated by simulation over many simulated versions of $\mathcal{X}^*$ and $\mathcal{X}^{**}$.

7.3.3 Accuracy of the bootstrap. In the Appendix we shall show that the single-bootstrap critical point $\tilde c_\alpha$ satisfies:

$$\tilde c_\alpha = c_\alpha + O_p\big(n^{-1}\big) \tag{7.7}$$

as $n \to \infty$. The fact that the error here is n^{-1}, rather than $n^{-1/2}$, reflects only the fact that we have not normalised when defining c_α. Indeed, in asymptotic terms,

$$n^{1/2}\, c_\alpha = u(\alpha) + O\big(n^{-1/2}\big)\,, \tag{7.8}$$

where $u(\alpha)$ is defined in terms of a p-variate normal distribution; see (7.16).

The $O_p(n^{-1})$ error in (7.7) is comprised primarily of errors incurred in estimating the variance matrix, and in fact (7.7) can be written in more detail as:

$$\tilde c_\alpha = c_\alpha + n^{-1/2} \left(\widehat\Sigma - \Sigma\right)^{\mathrm{T}} w + O_p\big(n^{-3/2}\big)\,, \tag{7.9}$$

where $\widehat\Sigma$ denotes the bootstrap estimator of the variance matrix of the p-vector $n^{1/2}\,(\hat\theta - \theta)$, and is approximated implicitly in the process of calculating $\tilde c_\alpha$; and w is a fixed vector of length equal to the number of components of Σ. Result (7.9)

is derived in the Appendix. There we also outline a proof that the coverage error of a confidence interval, or level error of a hypothesis test, if we use the single-bootstrap critical point $\tilde{c}_\alpha$, rather than its double-bootstrap counterpart, is of size $n^{-1/2}$, not n^{-1}. We show too that this level of accuracy prevails in both one-and two-sided procedures; the parity properties that are familiar in more conventional cases do not, in this context, lead to a reduction to $O(n^{-1})$ accuracy in two-sided problems.

7.4 Numerical properties

7.4.1 Example: university rankings. The relative ranking of universities continue to attract interest both within academic communities and in broader society. Our methodology allows comparisons to be made between groups of universities. For example, suppose we wanted to explore whether Switzerland or The Netherlands has the best university for scientific research. One statistic of interest could be the average number of articles published in the journals *Science* or *Nature* per year. This is one of the pieces of information used in the popular Shanghai Jiao Tong University rankings[1]. Figure 7.1 shows the distribution of the number of papers for two of the leading universities in Switzerland and the top four universities in The Netherlands for the 12 years from 1997 to 2008. Note that the other leading university in Switzerland, ETH Zurich, has not been considered here due to its non-stationarity over this time period; for instance, it had five papers in both 2003 and 2004, while produced 20 and 37 in 2007 and 2008 respectively. Aside from this omission, a university was included in the comparison if it had the most number of papers published for that country in an individual year.

Our statistic $\hat{\theta}_j$ is the mean number of papers published over the twelve years. The Utrecht University in The Netherlands has the highest overall mean, so we use the test of case 7 in Table 7.1 with Swiss universities as the first population and Dutch universities the second. The double bootstrap with 999 resamples in each layer resulted in a p-value of 0.14, suggesting that there is moderate evidence for The Netherlands having the better university.

The most important observation is that this significance test is more appropriate than pairwise comparisons between universities, since it recognises that we are not entirely sure which institution is the best performing for each country. For instance, if we calculated a p-value in a similar fashion but only compared the University of Zurich and the Utrecht University (having the observed maximum mean for Switzerland and The Netherlands respectively), we would obtain a value of 0.054. The significance level is misleadingly high because we have ignored the role of other universities.

[1] www.arwu.org

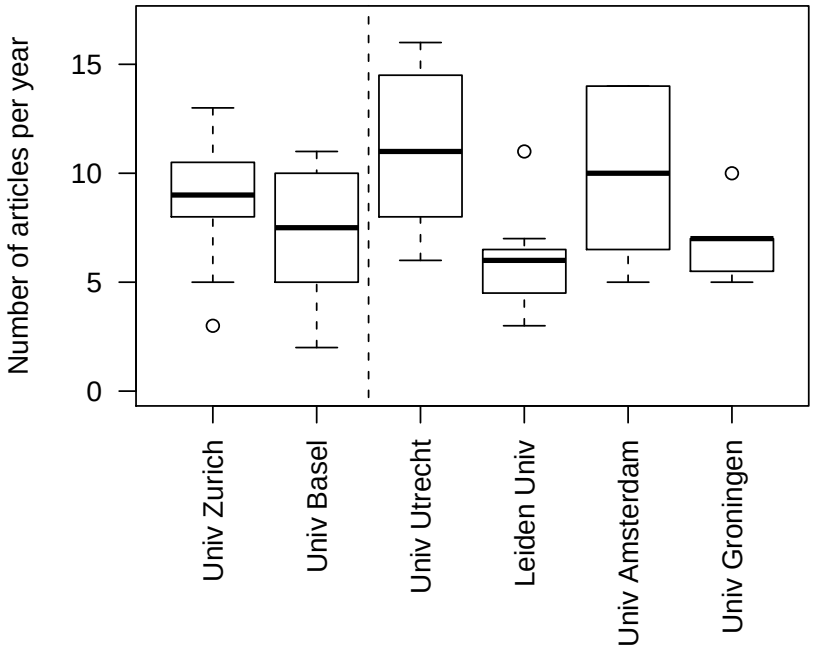


Figure 7.1: Boxplots of number of articles published per year in Science *or* Nature *for Swiss and Dutch institutions.*

While our test does not guarantee maximum power, it does give a better indication of certainty.

We make a few further comments about the result. Firstly, the comparison statistic is obviously somewhat simplified. We ignore whether or not the author from a given institution is listed first or otherwise, any relationships between articles and any changes in performance over the twelve year periods. Many of these issues are common to similar analyses. Secondly, this particular example is interesting in our context because of the uncertainty regarding the best university from each country. If a similar study was to be completed comparing the USA to Japan say, a pairwise test without conservatism would be appropriate, since each country has a clear leader for mean number of papers published (Harvard University and the University of Tokyo respectively).

7.4.2 Example: tennis player performance. Figure 7.2 shows the winning proportion of the top ten ranked mens tennis players; that is, the number of matches won divided by the number of matches played. In the figure they are ordered according to their official ranking, current as at 20 August 2009, and the proportion was calculated based on matches of the Association of Tennis Professionals, commonly referred to as the ATP, in the 2009 calendar year up to the same date. Eighty percent confidence intervals for these percentages are included in the figure as well. The most notably feature is Simon, ranked 9, who has a much lower winning percentage

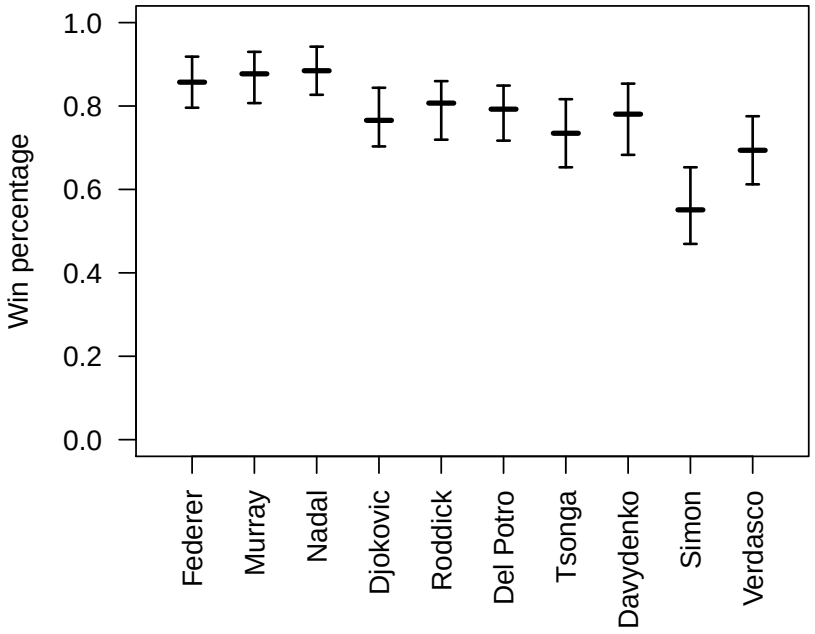


Figure 7.2: Winning percentages for the world top ten male tennis players

compared to the other players. Using the test from case 6 in Table 7.1, we can find p-values under the null hypothesis that Simon's performance is at least as good as the worst of the other top t players. Table 7.2 shows the p-values, estimated by means of the double bootstrap with 999 resamples in each layer, for various choices of t. Note that $t = 9$ is an irrelevant case since Simon himself occupies that ranking. The results suggest there is some weak evidence ($p = 0.21$) that Simon is below everyone else in the top 10, but increasingly strong evidence for smaller values of t. For instance, the corresponding p-value comparing Simon to the minimum of the top six players is 0.027. The bootstrap resample was conducted independently, and so ignores any dependence introduced by players having matches against each other.

	t									
	1	2	3	4	5	6	7	8	9	10
p-value	0.000	0.001	0.001	0.019	0.024	0.027	0.059	0.082	N/A	0.214

Table 7.2: Estimated p-values for the hypothesis that Simon's winning rate is as good as the minimum of the top t players, excluding himself.

In this example being able to test the multiple hypothesis is important, since it is not at all clear who has the lowest winning percentage after Simon. Besides perhaps the top three players, the remaining six have heavily overlapping confidence intervals.

7.4.3 Example: Wisconsin breast cancer. The final real data example is included to demonstrate the utility of this approach in cases where conservatism is

not an issue and introducing other possible statistics besides the mean which may be of interest. The Wisconsin dataset was first introduced in Wolberg and Mangasarian (1990) and was also used earlier in Section 3.1. It has 699 observations, each with 9 variables regarding tumor characteristics, along with an assignment of malignancy. While the main emphasis is usually the prediction of malignancy, an important part of any model is ensuring relationships between predictors are understood. Thus here we focus on determining which two of the nine predictor variables have the highest pairwise Pearson correlation. The correlation statistic is asymptotically normal and so our bootstrap methodology is appropriate. Variables two and three have the highest pairwise correlation by a fair margin; 0.91 compared to the next highest, 0.76. We test the hypothesis that the correlation between variables two and three is larger than any of the other 35, using the test in case 7 of Table 7.1. The resulting p-value is less than 0.001. Thus we can apply the method easily to problems that might otherwise be difficult to test formally.

7.4.4 Simulation of conservatism. The following example shows the increasing conservatism as the differences in the true means θ_j diverge. Suppose we have $p = 10$ populations and the parameter of interest is the mean, which take values $\{t, 2t, \ldots, pt\}$ on the populations, for some scalar t. We are interested in constructing an upper confidence interval for the maximum of these, as in the case 4 of Table 7.1. We assume the underlying observations are standard normal and each population has n observations from which to estimate the mean. In this example we can find c_α analytically, allowing us to better focus on the conservatism rather than estimation error. We know that the $(1-\alpha)$th quantile, d_α say, of the distribution of the maximum of p standard normal random variables is given by $F_p\{\Phi(d_\alpha)\} = 1-\alpha$, where $F_p(x) = x^p$ for $0 \leq x \leq 1$ and Φ is the standard normal cumulative distribution function. Thus, for each simulation we may set $c_\alpha = n^{-1/2} d_\alpha$. Table 7.3 shows the estimated coverage probabilities over 20,000 simulations for the $1-\alpha = 90\%$ confidence interval for various choices of t and n. Standard errors are shown in brackets.

	t					
n	0	0.2	0.4	0.6	0.8	1
10	0.901 (0.002)	0.946 (0.002)	0.965 (0.001)	0.977 (0.001)	0.981 (0.001)	0.983 (0.001)
20	0.899 (0.002)	0.958 (0.001)	0.976 (0.001)	0.981 (0.001)	0.985 (0.001)	0.985 (0.001)
50	0.902 (0.002)	0.968 (0.001)	0.983 (0.001)	0.986 (0.001)	0.987 (0.001)	0.989 (0.001)
100	0.901 (0.002)	0.975 (0.001)	0.985 (0.001)	0.986 (0.001)	0.988 (0.001)	0.990 (0.001)

Table 7.3: Simulated coverage probabilities exploring conservatism in Section 7.4.4.

The trend of increasing conservatism is evident as we move across the table from left (tied populations and no conservatism) to right. Another evident feature which is undesirable is that if anything, the conservatism is increases with the number of observations. This is perhaps not unsurprising, since by treating all observed $\hat{\theta}_j$ as

equal in the hypothesis test we lose any benefit associated with increased sample size.

The issue of conservatism being independent of sample size can be addressed by adding a preceding step to our analysis, again using our conservative hypothesis test. For a given simulation and $k < p$, we can perform a hypothesis test on whether the maximum of the k populations with smallest means is below the maximum of the other $p - k$ (as in the seventh line of Table 7.1). If we find the maximum k such that we reject the null at some suitably high confidence, say $\alpha = 0.02$, then we can construct a confidence interval for the overall maximum using only the remaining $p - k$ populations. The results of such an approach is presented in Table 7.4.

	t					
n	0	0.2	0.4	0.6	0.8	1
10	0.897 (0.002)	0.943 (0.002)	0.961 (0.001)	0.973 (0.001)	0.974 (0.001)	0.974 (0.001)
20	0.896 (0.002)	0.955 (0.001)	0.972 (0.001)	0.974 (0.001)	0.974 (0.001)	0.971 (0.001)
50	0.898 (0.002)	0.964 (0.001)	0.974 (0.001)	0.973 (0.001)	0.971 (0.001)	0.968 (0.001)
100	0.897 (0.002)	0.971 (0.001)	0.974 (0.001)	0.968 (0.001)	0.966 (0.001)	0.961 (0.001)

Table 7.4: Simulated coverage probabilities for example in Section 7.4.4 with additional initial hypothesis test.

The table shows that for a fractional loss of conservatism when all observations are in fact tied (in the first column of the table), we have significantly reduced the conservatism in situations where the $\hat{\theta}_j$ are well separated. In fact, if either t or n grows sufficiently large, the coverage will again approach the target coverage of 0.9.

7.4.5 Illustration of the accuracy of the double bootstrap. We give two illustrative examples comparing the coverage accuracy of the double bootstrap to that for the single bootstrap, where the αth percentile of the bootstrapped means is used, and the normal approximation, where the estimated mean is assumed to follow its asymptotic t-distribution. In the first the means are sampled from the exponential distribution with mean 1 and we test the coverage of one-sided 80% confidence. The second has means sampled from a Pareto distribution with mean equal to 2, scale parameter of 1 and shape parameter of 2, tested at 90% confidence. In each case $p = 10$ and we tested a range of n. We used $B = 599$ resamples for each bootstrap layer, and averaged over 2,000 simulations. Results are presented in Table 7.5. In the exponential case, the double bootstrap enjoys good coverage accuracy for all n, with results lying in natural variation levels around 0.80. The single bootstrap underestimates the interval width, although gives reasonable results for $n \geq 20$. The normal approximation overestimates the interval width, and this effect persists for n of moderate size. In the Pareto case all approaches overestimate the confidence interval width, with the double bootstrap clearly preferred at all n tested. Note that the double bootstrap is usually computationally manageable; in the university and tennis examples our computation time was 38.3 and 14.5 seconds respectively. Thus

if the dataset is sufficiently large and well-behaved the user may find the accuracy of the single bootstrap or normal approximation sufficient, but otherwise the double bootstrap is recommended over the competing approaches.

	Exponential distribution			Pareto distribution		
	Single	Normal	Double	Single	Normal	Double
n	Bootstrap	Approximation	Bootstrap	Bootstrap	Approximation	Bootstrap
10	0.750 (0.010)	0.858 (0.008)	0.797 (0.009)	0.955 (0.005)	0.981 (0.003)	0.913 (0.006)
15	0.780 (0.009)	0.840 (0.008)	0.807 (0.009)	0.960 (0.004)	0.967 (0.004)	0.916 (0.006)
20	0.792 (0.009)	0.821 (0.009)	0.804 (0.009)	0.964 (0.004)	0.969 (0.004)	0.909 (0.006)
25	0.805 (0.009)	0.822 (0.009)	0.812 (0.009)	0.968 (0.004)	0.967 (0.004)	0.909 (0.006)

Table 7.5: Simulated coverage probabilities comparing interval estimation approaches in Section 7.4.5. Targeted coverage was 80% for the exponential case and 90% for the Pareto distribution.

7.5 Technical arguments for Section 7.3

Under the smooth-function model (Bhattacharya and Ghosh, 1978) an estimator, or a vector of estimators such as $\hat{\theta} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)$, computed from a sample of size n, is represented as a smooth function of a mean of n independent random vectors all distributed as V, say. If the distribution of V has sufficiently many finite moments and satisfies Cramér's condition, i.e. $\limsup_{\|t\|\to\infty} |E\{\exp(it^{\mathrm{T}}V)\}| < 1$, for which it is sufficient that the distribution of V be nonsingular; if the function has sufficiently many derivatives; and if the limiting variance-covariance matrix of $\hat{\theta}$, Σ say, is nonsingular; then, for $r \geq 1$,

$$\begin{aligned} P\Big\{n^{1/2}\left(\hat{\theta}_j - \theta_j\right) &\leq z_j \text{ for } 1 \leq j \leq p\Big\} \\ &= \Phi_\Sigma(z) + \sum_{k=1}^{r} n^{-k/2}\, P_k(z)\, \phi_\Sigma(z) + O\big(n^{-(r+1)/2}\big)\,, \quad (7.10) \end{aligned}$$

uniformly in $z = (z_1, \ldots, z_p)$, where ϕ_Σ and Φ_Σ are respectively the density and distribution functions of the normal $\mathrm{N}(0, \Sigma)$ distribution, $P_1, \ldots, P_k$ are polynomials, not depending on n, with coefficients depending on derivatives of the smooth functions in the smooth-function model, evaluated at moments of V. The number of moments required of V increases with r. See Bhattacharya and Ranga Rao (1976) and Bhattacharya and Ghosh (1978).

Analogously, (7.10) has an empirical version:

$$\begin{aligned} P\Big\{n^{1/2}\left(\hat{\theta}^*_j - \hat{\theta}_j\right) &\leq z_j \text{ for } 1 \leq j \leq p \;\Big|\; \mathcal{X}\Big\} \\ &= \Phi_{\widehat{\Sigma}}(z) + \sum_{k=1}^{r} n^{-k/2}\, \widehat{P}_k(z)\, \phi_{\widehat{\Sigma}}(z) + O_p\big(n^{-(r+1)/2}\big)\,, \quad (7.11) \end{aligned}$$

where $\mathcal{X}$ denotes the dataset from which $\hat\theta_1, \ldots, \hat\theta_p$ were computed, $\widehat\Sigma$ is the bootstrap estimator of Σ calculated from $\mathcal{X}$, and $\widehat P_k$ is the version of P_k when moments of V, appearing in P_k, are replaced by their empirical counterparts. See, for example, Chapter 3 of Hall (1992).

Next we discuss using the single bootstrap to construct an estimator, $\tilde c_\alpha$ say, of c_α, the latter defined by:

$$P\Big\{ \max_{j\in\mathcal{J}_2} (\hat\theta_j - \theta_j) - \min_{j\in\mathcal{J}_1} (\hat\theta_j - \theta_j) \le c_\alpha \Big\} = 1-\alpha\,. \tag{7.12}$$

See Table 7.1. We define $\tilde c_\alpha$ by

$$P\Big\{ \max_{j\in\mathcal{J}_2} (\hat\theta^*_j - \hat\theta_j) - \min_{j\in\mathcal{J}_1} (\hat\theta^*_j - \hat\theta_j) \le \tilde c_\alpha \;\Big|\; \mathcal{X} \Big\} = 1-\alpha\,. \tag{7.13}$$

Put $d_\alpha = n^{1/2}\, c_\alpha$, $\tilde d_\alpha = n^{1/2}\, \tilde c_\alpha$, $Z_j = n^{1/2}\,(\hat\theta_j - \theta_j)$ and $Z^*_j = n^{1/2}\,(\hat\theta^*_j - \hat\theta_j)$. In this notation, (7.10)–(7.13) imply that:

$$\begin{aligned}
1-\alpha &= P\Big\{ \max_{j\in\mathcal{J}_2} Z_j - \min_{j\in\mathcal{J}_1} Z_j \le d_\alpha \Big\} \\
&= \int_{\mathcal{A}(d_\alpha)} \Big[\phi_\Sigma(z) + n^{-1/2}\, \frac{d}{dz} \{ P_1(z)\, \phi_\Sigma(z) \} \Big]\, dz + O\big(n^{-1}\big)\,, \qquad (7.14) \\
1-\alpha &= P\Big\{ \max_{j\in\mathcal{J}_2} Z^*_j - \min_{j\in\mathcal{J}_1} Z^*_j \le \tilde d_\alpha \;\Big|\; \mathcal{X} \Big\} \\
&= \int_{\mathcal{A}(\tilde d_\alpha)} \Big[\phi_{\widehat\Sigma}(z) + n^{-1/2}\, \frac{d}{dz} \{ \widehat P_1(z)\, \phi_{\widehat\Sigma}(z) \} \Big]\, dz + O_p\big(n^{-1}\big)\,, \qquad (7.15)
\end{aligned}$$

where $\mathcal{A}(d)$ is the set of $z \in \mathbb{R}^p$ such that $z_{j_2} - z_{j_1} \le d$ for all $j_1 \in \mathcal{J}_1$ and all $j_2 \in \mathcal{J}_2$, and we have taken $r = 1$ in (7.10) and (7.11). In this notation $u(\alpha)$, at (7.8), is the solution of the equation

$$1-\alpha = \int_{\mathcal{A}\{u(\alpha)\}} \phi_\Sigma(z)\, dz\,. \tag{7.16}$$

Define $e_\alpha = e_\alpha(\Sigma)$ to be the solution of the equation: $\int_{\mathcal{A}(e_\alpha)} \phi_\Sigma(z)\, dz = 1-\alpha$. If Σ_1 is a general nonsingular covariance matrix then

$$\big| e_\alpha(\Sigma_1) - e_\alpha(\Sigma) - v(\Sigma_1 - \Sigma)^{\mathrm{T}}\, \dot e_\alpha(\Sigma) \big| \le C_1\, \|\Sigma_1 - \Sigma\|^2 \tag{7.17}$$

whenever $\|\Sigma_1 - \Sigma\| \le C_2$, where $v(M)$ is the vector of length $\frac12\, p\,(p+1)$ defined as a concatination of the distinct components of a general symmetric matrix M, $\dot e_\alpha(\Sigma)$ is the vector of derivatives of $e_\alpha(\Sigma)$ with respect to the components of $v(\Sigma)$, $\|\cdot\|$ denotes any given matrix norm, and C_1 and C_2 are positive constants depending only on Σ.

In view of (7.14), $d_\alpha = e_{1-\beta}(\Sigma) + O(n^{-1})$ where

$$\beta = 1 - \alpha - n^{-1/2} \int_{\mathcal{A}(d_\alpha)} \frac{d}{dz} \{P_1(z)\, \phi_\Sigma(z)\}\, dz\,.$$

Since $\widehat{\Sigma} = \Sigma + O_p(n^{-1/2})$ and $(d/dz)\,\{\widehat{P}_1(z) - P_1(z)\}\,\phi_\Sigma(z) = O_p(n^{-1/2})$, the latter identity holding uniformly in z, then (7.15) implies that $\tilde{d}_\alpha = e_{1-\beta}(\widehat{\Sigma}) + O_p(n^{-1})$. Hence, by (7.17) and the fact that $\dot{e}_{1-\beta}(\widehat{\Sigma}) = \dot{e}_{1-\beta}(\Sigma) + O(n^{-1/2})$, we have: $\tilde{d}_\alpha = d_\alpha + v(\widehat{\Sigma} - \Sigma)^{\mathrm{T}}\, \dot{e}_{1-\beta}(\Sigma) + O_p(n^{-1})$. This result implies that:

$$\tilde{c}_\alpha = n^{-1/2}\, c_\alpha + n^{-1/2}\, v\big(\widehat{\Sigma} - \Sigma\big)^{\mathrm{T}}\, \dot{e}_{1-\beta}(\Sigma) + O_p\big(n^{-3/2}\big)\,, \tag{7.18}$$

and hence entails (7.9).

The coverage error in a confidence interval, or level error in a hypothesis test, that we incur when using the single-bootstrap approximation, $\tilde{c}_\alpha$, to c_α is the value we obtain when substituting $\tilde{c}_\alpha$ for c_α in (A.4) and subtracting $1 - \alpha$. That is, the error equals:

$$P\Big\{ \max_{j \in \mathcal{J}_2} Z_j - \min_{j \in \mathcal{J}_1} Z_j \le n^{1/2}\, \tilde{c}_\alpha \Big\} - (1 - \alpha)\,.$$

Substituting for $\tilde{c}_\alpha$ using (7.18), or equivalently (7.9), we deduce that:

$$P\Big\{ \max_{j \in \mathcal{J}_2} Z_j - \min_{j \in \mathcal{J}_1} Z_j \le d_\alpha + v\big(\widehat{\Sigma} - \Sigma\big)^{\mathrm{T}}\, \dot{e}_{1-\beta}(\Sigma) + O_p\big(n^{-1}\big) \Big\} - (1 - \alpha)\,. \tag{7.19}$$

Unsurprisingly, standard arguments (see e.g. Chapter 5 of Hall, 1992) show that the $O_p(n^{-1})$ inside the probability in (7.19), if dropped, produces a remainder term of order n^{-1}, and likewise that the term in $v(\widehat{\Sigma} - \Sigma)^{\mathrm{T}}\, \dot{e}_{1-\beta}(\Sigma)$, being exactly of size $n^{-1/2}$, if ignored gives a remainder of exact size $n^{-1/2}$. Therefore the coverage error or level error of the single-bootstrap procedure is genuinely of size $n^{-1/2}$.

Analogously to (7.18), the critical point $\tilde{c}^*_\alpha$, introduced in Section 7.3.2, is given by:

$$\tilde{c}^*_\alpha = n^{-1/2}\, \tilde{c}_\alpha + n^{-1/2}\, v\big(\widehat{\Sigma}^* - \widehat{\Sigma}\big)^{\mathrm{T}}\, \dot{e}_\alpha(\Sigma) + O_p\big(n^{-3/2}\big)\,. \tag{7.20}$$

Here we have used the fact that, owing to the smoothness of $\dot{e}_\alpha(\Sigma)$ as a function of Σ, $\dot{e}_\alpha(\widehat{\Sigma}) = \dot{e}_\alpha(\Sigma) + O_p(n^{-1/2})$. In particular, the double bootstrap correctly captures the main cause of error, arising from the difference $\widehat{\Sigma} - \Sigma$ in (7.18) and represented by $\widehat{\Sigma}^* - \widehat{\Sigma}$ in (7.20), in the single-bootstrap approximation. That is, the double bootstrap correctly captures the first-order terms that describe departures of size $n^{-1/2}$ from the limiting distribution. A rigorous proof of this result follows using standard arguments given in Chapter 5 of Hall (1992).

Chapter 8

Recursive variable selection in high dimensions

8.1 Background

We now return to the problem of building a sparse model, but in contrast to Chapters 2 and 4, we focus on the binary classification problem, where the response Y takes only the value 0 or 1. The main task for a high-dimensional problem remains the same: finding a good, relatively small collection of variables on which to base a final model. This is clearly not a wholly new problem; for instance, projection pursuit (see e.g. Friedman and Tukey, 1974) tackles the problem of variable selection by seeking the linear combination of dimensions that is "most interesting" in some sense, for example because it is the least Gaussian or has greatest entropy, among all projections that are orthogonal to those that have already been chosen. Other classical approaches, for example Asimov's (1985) grand tour and the N-land tool suggested by Ward et al. (1994), also involve searching through many possibilities. This is often feasible when dimension, p, is smaller than sample size, n, and also on some occasions where n and p are broadly similar. However, in contemporary problems where p is much larger than n, approaches of this type are ruled out on several grounds. One is their computational complexity. For example, even if each coefficient in a linear combination can take only two values, searching for the most appropriate one among all possibilities involves $O(2^p)$ calculations if each combination has to be explored. This is infeasible when, as is commonly the case today, p is in the thousands or tens of thousands.

These considerations motivate alternative approaches for solving contemporary high-dimensional classification problems. The latter include methods based on linear

(or logistic) prediction, as discussed in Hall et al. (2010), penalised discriminant methods, distance-based methods such as the support vector machine and centroid classifiers, and also techniques that involve ranking relevance measure for each of the p components such as those explored in Chapter 2. Algorithms such as these require at most $O(p \log p)$ operations, in terms of their dependence on p, and so are feasible even in many ultra-high dimensional settings.

Prediction-based approaches are generally top-down, in that they start with the full p-dimensional problem and successively reduce dimension. They and ranking-based methods usually do not take into account the sort of classifier that will be used. For instance, the set of components that minimise the error for one classifier might be different from those that are optimal for another, but that fact will typically not influence the variable selection step. More generally, the methods discussed above are relatively insensitive to interactions.

An alternative, bottom-up approach involves sequentially and explicitly building a model. Generally this is done by some form of forward stagewise selection where variables are sequentially added to the model according to which of the variables best improves an objective function. Such approaches are termed "wrapper" methods when applied to genetic microarray datasets. There are clear advantages to this recursive approach. In particular, it addresses all potential arrangements (e.g. permutations) of the vector components, but requires only $O(p)$ calculations. It is highly adaptive to the classifier type, and places no restriction on the nature of interactions among components that can be permitted. Indeed, in this respect it merely reflects the classifier; if the latter is responsive to highly nonlinear combinations of vector components then the recursive variable selector is too.

In this chapter we propose approaches of this type. Their main feature is that they explicitly target the leave-one-out misclassification rate, which leads to more robust variable selection and heightened protection against overfitting. Our theoretical results demonstrate that these methods produce good asymptotic performance, even in very high dimensional situations. We also investigate bootstrap tools that give insight into the stability of variable set selections. Further, we demonstrate the use of a double-layer of cross-validation to produce more reliable accuracy rates.

Knowing which components have greatest influence on correct classification can greatly enhance scientific interpretation of the results of classification. A method that simply assigns new data to different populations is not nearly so useful. The approach that we suggest is inherently of the former type. It has few peers in terms of the explicitness with which it selects variables that have greatest leverage on the successful performance of a particular classifier. It also offers new opportunities for practitioners to compare classifiers, for a particular dataset and on the basis of the variables or features that they select. This includes comparing the emphases that

different classifiers give to different variables.

Related work relating to linear methods for classification was introduced in Section 1.7. More generally, the literature on classification problems, particularly with respect to variable selection, is now vast. Duda et al. (2001), Hastie et al. (2001) and Shakhnarovich et al. (2005) provide book-length treatments of classification and related problems. Dudoit et al. (2002) discuss the performance of different classifiers. Fields of application of high-dimensional classifiers are as diverse as image analysis (e.g. Cootes et al., 1994), forestry (e.g. Franco-Lopez et al., 2001), speech recognition (e.g. Bilmes and Kirchhoff, 2003) and chemometrics (e.g. Schoonover et al., 2003), and of course, genomics (e.g. Moon et al., 2006; Clarke et al., 2008; Hua et al., 2009). Discussion of wrapper methods for microarray data include the review by Saeys et al. (2007), as well as work of Xiong et al. (2001) and Inza et al. (2004).

8.2 Model and Methodology

8.2.1 Estimator of error rate. Assume that a population is a mixture of two sub-populations, Π_0 and Π_1. For a given individual from Π_j we observe a data pair (X, Y), where $X = (X^{(1)}, \ldots, X^{(p)})$ is a p vector and $Y = j$ denotes the sub-population type. Suppose too that we have training data, in the form of random samples $\mathcal{S} = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, with each (X_i, Y_i) coming from either Π_0 or Π_1, known through the value of Y_i. Let $\mathcal{S}_0$ and $\mathcal{S}_1$ denote the training points belonging to classes 0 and 1 respectively, with corresponding cardinalities n_0 and n_1. Also, let $\mathcal{C}(X \,|\, k_1, \ldots, k_t)$ denote the result of applying a particular classifier $\mathcal{C}$ to a data vector X that has been dimension-reduced to just the components with distinct indices $k_1, \ldots, k_t$. That is, $\mathcal{C}(X \,|\, k_1, \ldots, k_t)$ denotes the sub-population type, either 0 or 1, to which the classifier assigns X. Our estimator of the error rate, computed for these indices and based on the training data, is

$$\begin{aligned} \widehat{\mathrm{err}}(k_1, \ldots, k_t) \;=\; & \frac{\pi}{n_0} \sum_{i:\, Y_i = 0} I\{\mathcal{C}_{-i}(X_i \,|\, k_1, \ldots, k_t) = 1\} \\ & + \frac{1-\pi}{n_1} \sum_{i:\, Y_i = 1} I\{\mathcal{C}_{-i}(X_i \,|\, k_1, \ldots, k_t) = 0\}\,, \end{aligned} \tag{8.1}$$

where π denotes the prior probability of sub-population Π_0, the notation $\mathcal{C}_{-i}$ means that the classifier is constructed by omitting the ith observation from the training sample, and I denotes the indicator function; $I(\mathcal{E}) = 1$ if the event $\mathcal{E}$ holds, and $I(\mathcal{E}) = 0$ otherwise. Section 8.2.4 will define $\mathcal{C}(X \,|\, k_1, \ldots, k_t)$ and $\mathcal{C}_{-i}(X_i \,|\, k_1, \ldots, k_t)$ in the case of centroid-based classifiers, and show how these definitions are altered for other classifier types. In the case of priors set equal to the observed frequencies in the observed data, (8.1) precisely equals the leave-one-out cross-validated error rate.

Under mild assumptions on the classifier, the estimator of error rate at (8.1) converges to the true error rate,

$$\begin{aligned}\mathrm{err}(k_1,\ldots,k_t) &= \pi\, P\{\mathcal{C}(X_i\,|\,k_1,\ldots,k_t) = 1 \text{ and } Y_i = 0\} \\ &\quad +(1-\pi)\, P\{\mathcal{C}(X_i\,|\,k_1,\ldots,k_t) = 0 \text{ and } Y_i = 1\}\,, \quad (8.2)\end{aligned}$$

as n_0 and n_1 increase. This gives the approach a major advantage over other recursive methods, since it provides automatic protection against overfitting. While most other approaches quickly drive the error to zero in the training set, we shall see that the leave-one-out error plateau is comparable to the true error rate.

8.2.2 Algorithm. First we describe the initial step of the algorithm, where we select $k = k_1$ from among $1,\ldots,p$ to minimise $\widehat{\mathrm{err}}(k)$. At this point, and in similar situations below, the manner in which we deal with ties is important, since in high-dimensional problems it is often the case that k can take many more distinct values than $\widehat{\mathrm{err}}(k)$. We suggest determining the set $\mathcal{K}$ of values of k for which $\widehat{\mathrm{err}}(k)$ achieves its minimum, and choosing k_1 to be the element of $\mathcal{K}$ for which the classifier produces, in an average sense, the most authoritative classification, over all training data and incorporating prior probabilities where appropriate. For example, many classifiers assign a new data value X to Π_j on the basis of the sign of a score function S, computed from the training data. In particular, in the case of distance-based classifiers we can take $S(X)$ to be a function of the distance from X to $\mathcal{S}_1$, minus the same function of the distance from X to $\mathcal{S}_0$. Section 8.2.4 gives details in the case of the centroid, as well as other, classifiers. Then, choose k_1 to be the value of $k \in \mathcal{K}$ that maximises the prior-weighted median, or mean, or a similar measure of "average," of the values taken by $S(X)$ for variables X in the training data, based on only the kth component. In this paragraph and the next, $X \in \mathcal{S}_0 \cup \mathcal{S}_1$ and the function S itself actually depends on X, because it is computed from the data in $\mathcal{S}_0 \cup \mathcal{S}_1 \setminus \{X\}$. However, to avoid unwieldy notation we do not express this dependence explicitly.

Now we describe how to apply the algorithm recursively. Given distinct integers $k_1,\ldots,k_t$ between 1 and p, we choose k_{t+1} from $\{1,\ldots,p\} \setminus \{k_1,\ldots,k_t\}$ to minimise $\widehat{\mathrm{err}}(k_1,\ldots,k_{t+1})$. We use the procedure suggested above for breaking ties. That is, we choose k_{t+1} to be the value of k that, among those that minimise $\widehat{\mathrm{err}}(k_1,\ldots,k_t,k)$, maximises the average of the values taken by $S(X)$ for variables X in the training data, when they are stripped of all their components except those indexed by $k_1,\ldots,k_t,k$. However, we terminate the algorithm at the sequence $k_1,\ldots,k_t$ if the operation of adjoining any other component index k_{t+1} would lead to a deterioration in estimated error rate. Here we can define "deterioration" in a nonstrict sense, meaning that $\widehat{\mathrm{err}}(k_1,\ldots,k_{t+1}) \le \widehat{\mathrm{err}}(k_1,\ldots,k_t)$, or in a strict sense, where $\le$ is replaced by $<$. If the algorithm does not terminate itself within a reasonable number

of steps, it can be concluded in other ways. The permitted maximum value of t can also be determined differently, for example by terminating when the improvement in error rate,

$$\widehat{\mathrm{err}}(k_1,\ldots,k_t) - \widehat{\mathrm{err}}(k_1,\ldots,k_{t+1})\,, \tag{8.3}$$

falls below a given, positive level, or when t reaches a ceiling beyond which scientific interpretation (e.g. in a biological sense, if t denotes the number of genes used in the classifier) is difficult, or when the level of computation reaches a practical limit.

In practice we could determine, in advance, an upper bound t_0 chosen on the basis of computational resources. We would run the algorithm as suggested above, stopping at an empirically determined step $\hat{t}$ if $\hat{t} \le t_0$; and terminating the algorithm at t_0, or at least reconsidering it at that point, if at the t_0th step it was clear that algorithm would continue. The computational labour required to determine the indices $k_1,\ldots,k_t$ would then be bounded above by a constant multiple of n^2pt_0. In practice n_0 and n_1 are often very much less than p, for example being in the tens whereas p is in the thousands or tens of thousands.

8.2.3 Extensions of the algorithm. A handicap of the approach described above is that if, for some reason, we produce poor initial choices of k, we must keep them in the index set, potentially disadvantaging the final classifier. Further, different variable choices at some early stage will often result in completely different variable selections later. These problems can be mitigated by using a jittered algorithm. This generates a pool of possibilities for k, for example consisting of all the values that minimise $\widehat{\mathrm{err}}(k)$ and for which the average of the leave-one-out versions of $(-1)^j\,S(X)$, when classification is based solely on the kth component, is among the ℓ_1 largest, where ℓ_1 is fixed (or, in theoretical terms, does not depend on p). We can also adjoin the set of indices k for which, given k_1 from the aforementioned set, $\widehat{\mathrm{err}}(k_1,k)$ is minimised and the average value of $(-1)^j\,S(X)$ is among the ℓ_2 largest. Write $\mathcal{L}$ for the set of integers k derived using these or related methods. Once $\mathcal{L}$ has been determined, the algorithm can be re-run so that it starts from any member of $\mathcal{L}$, rather than from k_1; or alternatively, it could start from a subset of $\mathcal{L}$.

8.2.4 Example: Centroid-based classifier. The standard centroid-based classifier assigns a new data value $X = (X^{(1)},\ldots,X^{(p)})$ to Π_0 or Π_1 according to whether the statistic

$$S(X) = \sum_{k=1}^{p} \left\{ \bigl(X^{(k)} - \bar{X}_1^{(k)}\bigr)^2 - \bigl(X^{(k)} - \bar{X}_0^{(k)}\bigr)^2 \right\} \tag{8.4}$$

is positive or negative, respectively. Here, $\bar{X}_j = (\bar{X}_j^{(1)},\ldots,\bar{X}_j^{(p)}) = n_j^{-1}\sum_{i:\,Y_i=j} X_i$ is the average value of the vectors in the training sample $\mathcal{S}_j$. We can equivalently, and more conventionally, interpret this rule as assigning X to Π_0 if X is closer to $\bar{X}_0$

than to $\bar{X}_1$, and assigning it to Π_1 otherwise. That is, X is deemed to be from Π_0 if $\|X - \bar{X}_0\| \le \|X - \bar{X}_1\|$, and from Π_1 otherwise, where $\|\cdot\|$ denotes the conventional Euclidean metric. This approach to classification is popular in genomics; see, for example, Tibshirani et al. (2002), Dabney (2005), Dabney and Storey (2005, 2007) and Wang and Zhu (2007).

When constructing the error estimator at (8.1) we take X to be from one of the training samples $\mathcal{S}_j$, and we delete it from that sample when constructing $S(X)$ at (8.4). For example, if $X = X_{i_1}$ with $Y_{i_1} = 0$, then $S(X)$ becomes:

$$S_{-i_1}(X_{i_1}) = \sum_{k=1}^{p} \left\{ \bigl(X_{i_1}^{(k)} - \bar{X}_1^{(k)}\bigr)^2 - \left(X_{i_1}^{(k)} - \frac{1}{n_0 - 1} \sum_{i\,:\,Y_i = 0,\, i \ne i_1} X_i^{(k)} \right)^2 \right\}, \quad (8.5)$$

and a similar formula applies, for $S_{-i_1}(X_{i_1})$, if $Y_{i_1} = 1$ instead. Likewise, the classifier $\mathcal{C}_{-i}(X_i \,|\, k_1, \ldots, k_t)$ introduced in Section 8.2.1 is defined to be the rule that assigns X_i to Π_0 if $S_{-i}(X_i \,|\, k_1, \ldots, k_t) \ge 0$, and assigns it to Π_1 otherwise, where, for example when $Y_{i_1} = 0$,

$$S_{-i_1}(X_{i_1} \,|\, k_1, \ldots, k_t) = \sum_{k = k_1, \ldots, k_t} \left\{ \bigl(X_{i_1}^{(k)} - \bar{X}_1^{(k)}\bigr)^2 - \left(X_{i_1}^{(k)} - \frac{1}{n_0 - 1} \sum_{i\,:\,Y_i = 0,\, i \ne i_1} X_i^{(k)} \right)^2 \right\}. \quad (8.6)$$

The classifier $\mathcal{C}(X \,|\, k_1, \ldots, k_t)$ assigns X to Π_0 if $S(X \,|\, k_1, \ldots, k_t) \ge 0$, where $S(X \,|\, k_1, \ldots, k_t)$ has the definition of $S(X)$ at (8.4) except that the sum on the right-hand side there is taken only over $k = k_1, \ldots, k_t$.

Next we give an explicit definition of the tie-breaking procedure discussed in Section 8.2.2. Suppose $k_1, \ldots, k_t$ have been determined, and we seek k_{t+1}. Compute the set $\mathcal{K}_t$ of values of k for which $\widehat{\mathrm{err}}(k_1, \ldots, k_t, k)$ achieves its minimum. For each $k \in \mathcal{K}_t$ and each $i \in \{1, \ldots, n_j\}$, calculate $S_{-i}(X_i \,|\, k_1, \ldots, k_t, k)$, and put

$$T(k) = \frac{\pi}{n_0} \sum_{i\,:\,Y_i = 0} S_{-i}(X_i \,|\, k_1, \ldots, k_t, k) - \frac{1 - \pi}{n_1} \sum_{i\,:\,Y_i = 1} S_{-i}(X_i \,|\, k_1, \ldots, k_t, k)\,, \quad (8.7)$$

where π denotes the prior probability of Π_0. (We place a minus sign in front of the second term because $S_{-i}(X_i \,|\, k_1, \ldots, k_t, k)$ is negative if $Y_i = 1$ and the classifier $\mathcal{C}_{-i}(\cdot \,|\, k_1, \ldots, k_t, k)$ correctly assigns X_i to Π_1. Recall from Section 8.2.2 that the basic classifier has the property: Assign a new data value X to Π_j if $(-1)^j\, S(X) > 0$.) Choose $k \in \mathcal{K}_t$ to maximise $T(k)$.

The definition at (8.7) uses the mean to assess the average authority of classification decisions when k is included among components. If using the median we would redefine

$$T(k) = \pi \operatorname*{med}_{i\,:\,Y_i = 0} S_{-i}(X_i \,|\, k_1, \ldots, k_t, k) - (1 - \pi) \operatorname*{med}_{i\,:\,Y_i = 1} S_{-i}(X_i \,|\, k_1, \ldots, k_t, k)\,,$$

where $\text{med}_{i\in\mathcal{I}}\ u_i$ denotes the median of values the u_i indexed by set $\mathcal{I}$.

These constructions have close analogues for other classifiers, for example the support vector machine, where, in high-dimensional settings, $S(X)$ equals the square of the nearest distance from X to the convex hull formed by $\mathcal{S}_1$, minus its counterpart for the convex hull formed by $\mathcal{S}_0$; the nearest-neighbour classifier, where $S(X) = \min_{i:\,Y_i=1} \|X - X_i\|^2 - \min_{i:\,Y_i=0} \|X - X_i\|^2$; the average distance classifier, where $S(X) = n_1^{-1} \sum_{i:\,Y_i=1} \|X-X_i\|^2 - n_0^{-1} \sum_{i:\,Y_i=0} \|X-X_i\|^2$; the median-based classifier, an analogue of the centroid-based classifier, where $S(X) = \sum_k (|X^{(k)} - \text{med}\, X_1^{(k)}| - |X^{(k)} - \text{med}\, X_0^{(k)}|)$; and the discriminant classifier, where $S(X) = \log(p_0/p_1)$ and p_0, p_1 are the posterior probabilities of being in group $0, 1$ respectively. In each of these cases the definitions of $S_{-i}(X_i)$ and $S_{-i}(X_i \,|\, k_1, \ldots, k_t)$, analogous to those at (8.5) and (8.6), follow directly.

8.3 Numerical properties

8.3.1 Preliminary discussion of real-data analysis. We shall use two genetic microarray datasets introduced in earlier chapters to demonstrate our approach. The leukemia dataset was described in Section 4.2.2, while the colon dataset of Alon et al. (1999) was introduced in Section 6.1.3.

For each dataset we applied the methods listed in Table 8.1, using the recursive variable selection framework. Most of these methods were introduced in Section 8.2. The 5 nearest neighbour method classifies on the basis of the majority of the closest 5 observations. The score function $S(X)$ used is number of zeros minus number of ones. This creates the possibility of ties in the measure of authority, T, which must be ranked, for instance at random. However, we found this was not a major concern on implementation. We used both linear and quadratic discriminant analysis, an introduction to which may be found in Chapter 4 of Hastie et al. (2001).

Name	Description
Cent	Centroid-based classifier
Cent.med	Centroid-based classifier with median authority measure
Med	Median-based classifier
Dist	Average distance classifier
1-NN	Nearest neighbour classifier
5-NN	5 nearest neighbour classifier
LDA	Linear discriminant analysis
QDA	Quadratic discriminant analysis
SVM1	Linear support vector machine

Table 8.1: Approaches included in numerical comparisons

The main set of results compares prediction accuracy for each of these methods

as a function of the number of genes in the model. Accuracy was measured using a double layer of cross-validation, with the inner layer leave-one-out and the outer layer 10-fold. Thus, for each classifier we first divided the data into ten subsets of equal size. Then, for each subset we used the other nine to build a recursive model which involved looping through the inner layer of cross-validation to select variables using the leave-one-out method. A series of models with increasing numbers of genes resulted. Once the model variables were selected this was applied to the remaining 10% of data, to assess accuracy.

There is a strong case for the use of a double cross-validation method in such situations, since it avoids over-optimistic accuracy measurements caused by overfitting to the data. The resulting estimates give us a true indication of performance on an unseen dataset, as in each case the 10% of data set aside have not been used to fit the model at all.

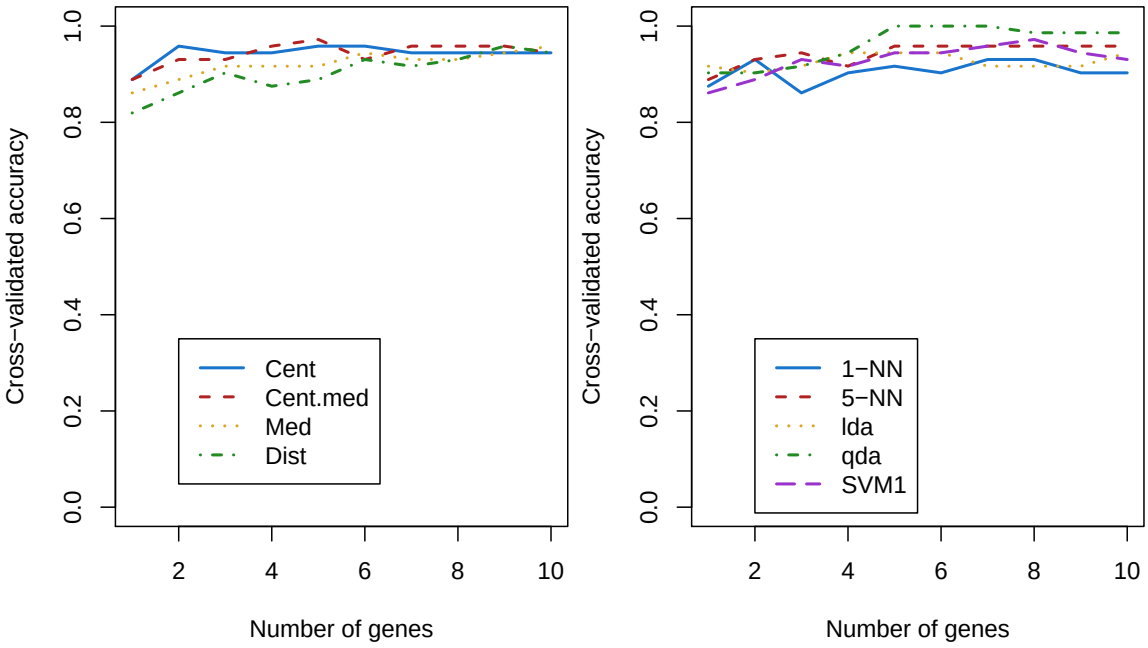


Figure 8.1: Accuracy curves for leukemia data.

8.3.2 Example: Leukemia data. Figure 8.1 shows the accuracy curves for the array of methods, measured using double layered cross-validation. Table 8.2 shows the obtained accuracy in the experiments, using the stopping rule (failure to decrease $\widehat{\text{err}}$) for model size selection. An average model size is reported since each outer validation fold requires its own stopping rule, so model size may differ across folds. The variables listed for the "final model" are those selected when we ran a single layer of cross-validation. We make the following observations:

1. The accuracy curves as well as the stopping rule results suggest that only a

Name	Avg Model Size	Accuracy (10-fold) CV	Final model variables
Cent	2.9	0.931	4847, 804, 6281
Cent.med	2.9	0.917	4847, 807, 6225
Med	2.5	0.889	4847, 804, 4951
Dist	3.5	0.889	3252, 1144, 2111
1-NN	2.8	0.861	3252, 4472, 6041
5-NN	3.4	0.931	4847, 804, 1685
LDA	3.1	0.944	4847, 2295, 1796, 2642
QDA	2.7	0.861	1882, 4342, 4582
SVM1	3.4	0.931	4847, 804, 4680

Table 8.2: Accuracy of models using suggested model size for leukemia data

small number of genes are required in the recursive model. In fact, accuracy does not improve beyond a few genes for most of the methods, with QDA and 5-NN the main exceptions. One explanation for this is that the small sample size does not allow reliable detection of effects that give small improvements in prediction accuracy.

2. Most approaches show comparable accuracy, here around 90%, despite the significant variation in model structure. The 1-NN method was generally worst. Also, QDA had equal worst performance in Table 8.2 when using a small number of genes, but had the best performance when taking a larger number of genes.

3. There appears to be a fair amount of stability in the first variable selected across the various methods, with variable 4847 chosen first in six of the nine experiments. Also, given that 4847 was selected first, four of the six then selected 804 second. However, there was no consistency in the selection of third variables, suggesting that there was not a clear signal across all methods at this level of the model.

4. One computational consideration is whether an initial feature selection could be effected to reduce the dimensionality, and hence improve computability, while still leaving enough variables so that the final models were unchanged. This would involve discarding all variables that performed poorly in the initial variable selection step. In the case of the leukemia data, it turns out that any significant pruning would impact on some of the models. Variable 4342 in the QDA model was initially ranked 6746th and variable 1144 in the distance model was initially ranked 1219th. At the other extreme, the 1-NN model needed only the top 50 variables to construct the classifier. In practice such pre-screening should be avoided where possible. The colon dataset, considered

below, illustrates another situation where significant initial pruning negatively affects the final models.

One way to gain an understanding of the reliability of variable selection is through the bootstrap. As described in Section 8.2.3, we take resamples of the data and see which variable is selected first in each replication. Similarly, given a first variable we can use bootstrap replications to investigate which variable was selected second, and so on. Results for the leukemia dataset are presented in Figure 8.2, using the centroid version of the recursive method. The leftmost plot shows that there are two main contenders for top selection, variables 4847 and 1834. The second plot shows the range of choices if we choose 4847 as the first variable. We can see that there was much more variability in this second choice, with lower proportions for the strongest variables and a much greater proportion of "other". The third plot gives the results when we choose variable 1834 first, again with a fairly large spread of possibilities. There is not a great deal of overlap in the second and third lists, suggesting that variables 1834 and 4847 are different enough for the subsequent pathways to be distinct. Further, variable 4847 does not appear on the third list, even though it is the best individual predictor, and conversely variable 1834 does not appear on the second list. This suggests that 4847 and 1834 contain fairly similar information, and thus better gains in accuracy can be obtained by choosing other variables.

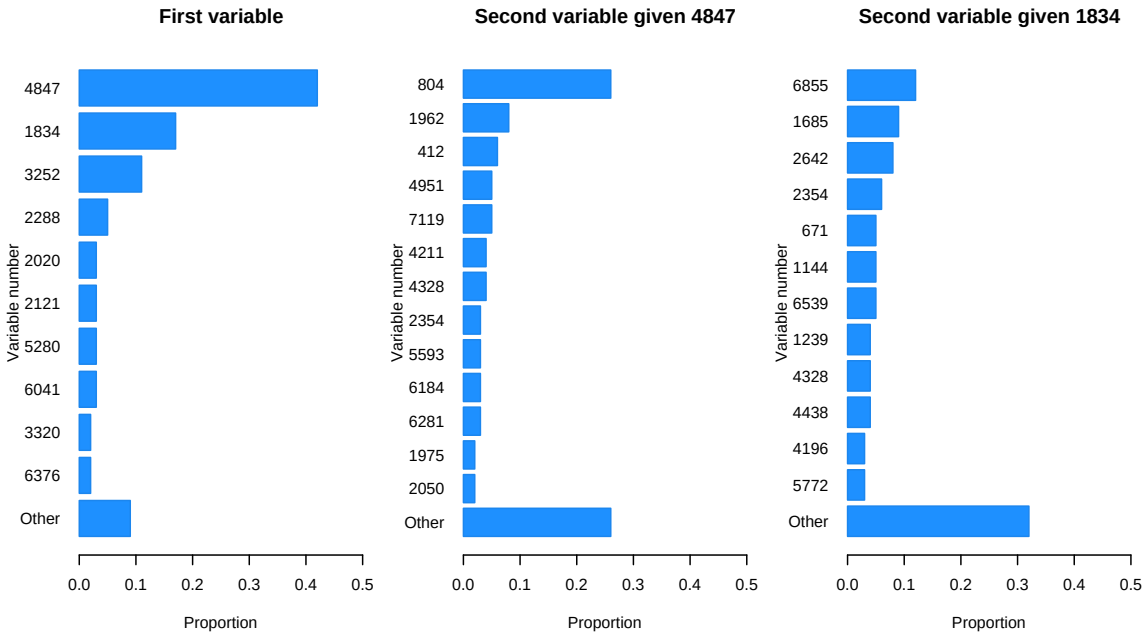


Figure 8.2: Variable selection frequency under bootstrap resampling.

A relevant question is how these classifiers compare to other contemporary approaches. Figure 8.3 provides an answer in the case of penalty-based methods. The model labeled "SVM1" denotes the L1 penalised support vector machine (Zhu et al.,

2004), while "HHSVM" represents a hybrid Huberised support vector machine with L1 penalty (Wang et al., 2006, 2008). Adjusting the L2 penalty in the latter case did not appear to have a significant impact on the presented results. Finally, "GLM1" refers to the logistic regression method with L1 penalty. The results here are not conclusive; all approaches produce similar maximum accuracy when optimal gene sets are selected. However, the centroid-based classifier seems to do this the most efficiently, needing only a few genes for good accuracy. Questions of relative performance are pursued further in simulation work below.

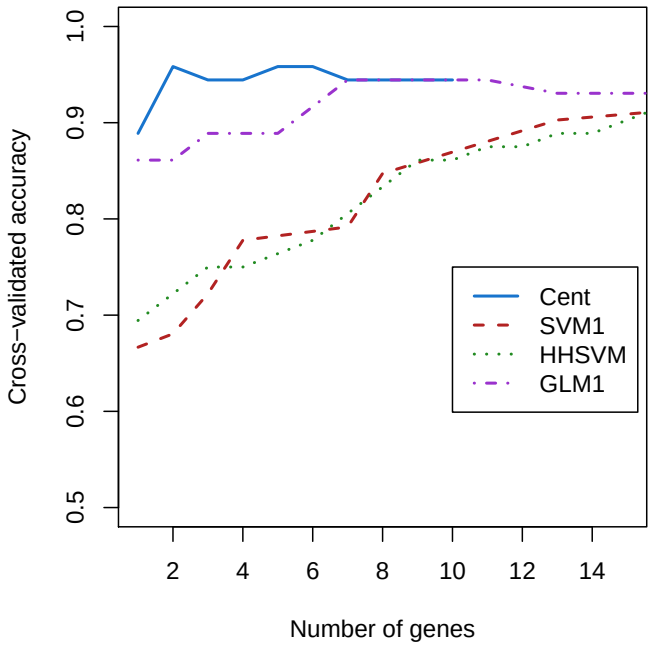


Figure 8.3: Plots for top variable by recursion and feature selection, respectively.

Name	Avg Model Size	Accuracy (10-fold) CV	Final model variables
Cent	3.1	0.806	249, 1346, 799
Cent.med	2.9	0.819	249, 1346, 799
Med	2.9	0.861	249, 32
Dist	3.6	0.819	245, 1772, 206
1-NN	4.8	0.792	1042, 883, 1900, 1414
5-NN	3.7	0.792	1671, 1365
LDA	4.4	0.792	1423, 1870, 678, 137, 1769
QDA	4.8	0.792	249, 1757, 377, 1042
SVM1	2.7	0.778	249, 1935, 1976

Table 8.3: Accuracy of models using suggested model size for colon data

8.3.3 Example: Colon data. The accuracy curves for these estimates are plotted in Figure 8.4 and Table 8.3 presents the results using the model size suggested by

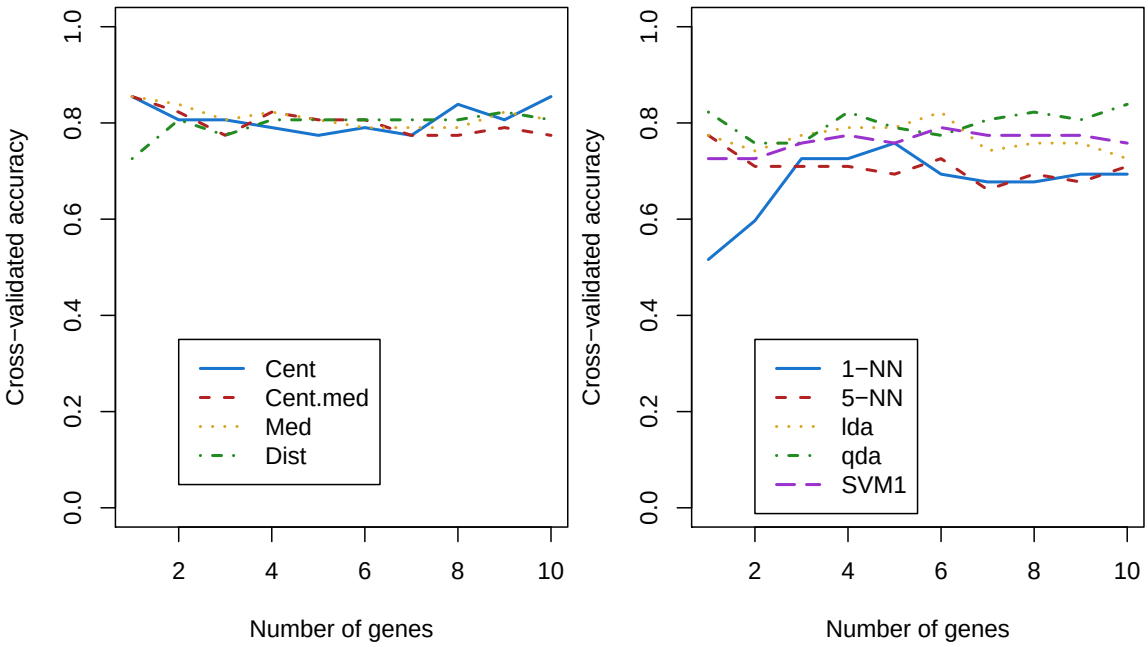


Figure 8.4: Accuracy curves for colon data.

our stopping rule. The results prompt observations similar to those for the leukemia data, although it is clear that the overall accuracy is lower in this case. For instance, the selection of the first variable is reasonably stable across methods, with decreasing stability for later additions. Also, the suggested model sizes are still small, although one of the methods (LDA) selected five genes. As before, an initial variable pruning would be inadvisable here since in many instances a variable initially ranked poorly appears in the model; in the 1-NN model, 1883 is initially ranked 1870th, and in the distance model, variable number 206 is initially ranked 1056th.

8.3.4 Comparison with alternative approaches under simulation. Here we compare the performance of the cross-validated centroid-based classifier with the L1 penalised Huberised support vector machine and L1 penalised logistic regression. Two alternative setups were used. In the first, we took $n = 100$ and $p = 5{,}000$. Each variable had zero mean, except for those related to the binary response, Y, and 100 variables had mean μ_j when $Y = 1$ and mean 0 otherwise, with the μ_j randomly generated from a uniform distribution on $[0, 2]$. Each class had the same number of corresponding observations. The error for each variable was independent and normally distributed with zero mean. The standard deviation of the error was allowed to vary for each variable and each class, and was sampled from a uniform distribution on $[1, 3]$. The models built were fitted on a separate test set of 500 observations, and the simulation was repeated 50 times and the results averaged.

The results for this simulation are shown in the left panel of Figure 8.5. While

the centroid-based classifier strongly outperforms its competitors when the number of genes in the model is low (three or less), it is generally inferior at larger model sizes. The reason for this appears to be that while the centroid-based classifier correctly targets the variables that maximise accuracy early on, for larger models it regularly has to use the tie-breaking rule to decide on variable choice. This rule makes no distinction between the score on observations classified well and those that are more marginal. By contrast, the loss functions for the competing methods cause them to focus on data that are misclassified. Thus, the discreetness of the cross-validated measure seems to be a factor. One possible means of addressing this is to modify the score function (8.4), although details are not pursued here.

The second simulation differed from the first in the following respect: There were 500 features that had different means in the two classes, making the simulation less sparse. These differences in means were sampled from those observed in the Leukemia dataset, and the error had a fixed standard deviation of 1. The results of this simulation are presented in the right hand panel of Figure 8.5. In this case the centroid-based classifier dominates across all model sizes, reaching a plateau at a higher accuracy level. This suggests that the modelling approach has superior performance in less sparse situations. In particular, the centroid-based classifier includes fewer redundant variables, improving the accuracy. This also illustrates the idea that a single approach will not perform best in all situations, as discussed in Section 1.3.

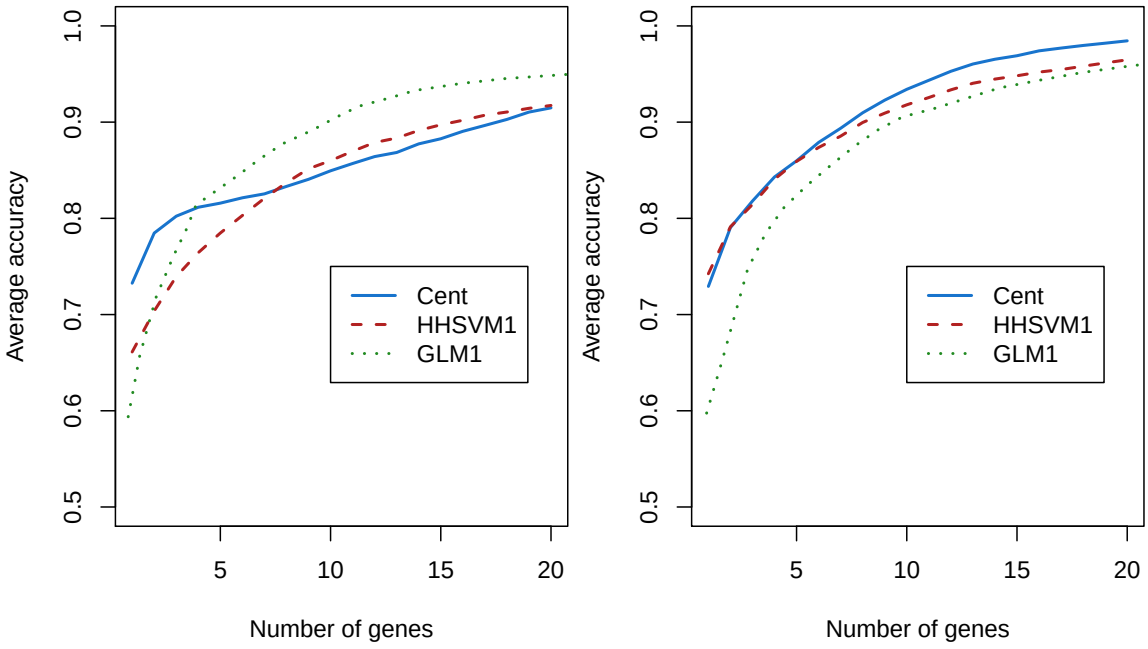


Figure 8.5: Comparison of accuracy results under simulation

8.3.5 Numerical work supporting theoretical results. Theorem 8.1 argues that if the classifier is constructed so as to be sensitive to the differences between Π_0 and Π_1, in particular to focus on extreme components if the principal differences are in terms of extrema, then our recursive approach can give particularly good performance and be more effective than, say, a linear classifier which does not acknowledge the context of the data.

p	u_p	v_p	avg gen.	avg irrel.	avg err.
10	0.126	0.063	3.55	1.36	0.350
20	0.179	0.089	3.54	3.10	0.255
50	0.283	0.141	3.31	3.62	0.166
100	0.400	0.200	3.30	3.08	0.133
200	0.566	0.283	3.45	1.76	0.110
500	0.894	0.447	4.14	0.21	0.062
1000	1.265	0.632	3.89	0.03	0.021
2000	1.789	0.894	3.18	0.00	0.004
5000	2.828	1.414	2.05	0.00	0.000

Table 8.4: Detection rates of genuine and irrelevant variables plus misclassification rates for data as in Theorem 4.1

To illustrate this we suppose that the data are generated as in the model addressed by Theorem 8.1. We keep $n_0 = n_1 = 30$ fixed, we take the components of X (when (X, Y) is drawn from Π_0) to be independent normal $\mathrm{N}(0, 1)$, we set $r = 5$, and we take $u_p = v_p/2$ and $v_p = p^{1/2}$. Table 8.4 shows the average results over 100 simulations each for various values of p. The fourth column in the table shows the average number of the five genuine variables detected. As Theorem 4.1 suggests, we do not consistently detect all five, but this does not impact on asymptotic classification performance. The fifth column of Table 8.4 shows the average number of variables selected that are not genuine, and that this number decreases to zero as p and u_p grow. Finally, the sixth column shows the misclassification rate, which, in accordance with the theoretical results, is also driven to zero.

To illustrate the results of Theorem 4.2 we take the components of X to be independent standard normal when drawing from Π_0, and we let $n_0 = n_1 = 30$ and $r = 5$, as before. In the base case we set $p = 20$ and $\mu = 1$, and allow μ to increase at rate $p^{1/8}$. We repeated each simulation 100 times for various p and recorded the number of genuine and redundant variables selected, as well as the error rates for the recursive model and for a model that used all variables. The error rates were calculated using a separate test set of data generated in the same manner. The results are presented in Table 8.5. As expected we see that as p increases the average number of redundant variables mistakenly selected reduces, and the error for the recursive model also shrinks. However, the error for the full model increases, approaching 0.5

p	μ	avg gen.	avg irrel.	avg err. recursive model	avg err. full model
20	1.00	3.27	0.43	0.189	0.157
50	1.12	2.70	0.31	0.187	0.160
100	1.22	2.58	0.23	0.173	0.162
200	1.33	2.47	0.28	0.155	0.176
500	1.50	2.69	0.28	0.124	0.207
1000	1.63	2.64	0.30	0.108	0.231
2000	1.78	2.85	0.23	0.083	0.260
5000	1.99	3.14	0.22	0.056	0.304
10000	2.17	3.22	0.13	0.038	0.323

Table 8.5: Detection rates of genuine and irrelevant variables plus misclassification rates for data as in Theorem 4.2

since μ is growing too slowly compared to p.

8.3.6 Computational considerations. In the past the large sizes of some datasets meant that comprehensive cross-validation was undesirable due to the computational labour required. This is not the case today, however. In support of this claim, Table 8.6 shows the "raw" computation time, without any attempt made to optimise the cross-validation step, taken to select $t = 10$ genes for a single layer, leave-one-out cross-validated fit to the Leukemia data ($n = 72$, $p = 7{,}129$) for each of the methods used. The fits were implemented using R running on a typical desktop computer with a 2.66Ghz processor. The number of classifiers fitted for each method was over five million.

Name	Raw fitting time (mins)
Cent	19.9
Cent.med	19.9
Med	121.3
Dist	28.3
1-NN	27.5
5-NN	23.5
LDA	4.7
QDA	5.9
SVM1	399.0

Table 8.6: Computer time taken to fit recursive models on Leukemia data

With the possible exception of the median and SVM approaches, the computation times given in Table 8.6 are very reasonable. In the case of microarrays, many hours of laboratory time are needed to produce the data, so 20 minutes to fit a robust model is not unduly onerous. Of course, part of the reason the modelling times are so reasonable is that we have used relatively simple classifiers, although the good

prediction performance above suggests that this is not a significant issue.

Note too that there are often ways to improve model fitting times, taking advantage of the fact that calculations of moments and other similar statistics do not change greatly as individual observations are excluded. For instance, in the case of the centroid-based classifier we can rewrite (8.5) as:

$$\begin{aligned} S_{-i_1}(X_{i_1}) &= \sum_{k=1}^{p} \left\{ \left(X_{i_1}^{(k)} - \bar{X}_1^{(k)}\right)^2 - \left(X_{i_1}^{(k)} - \frac{1}{n_0 - 1} \sum_{i\,:\,Y_i=0,\, i\neq i_1} X_i^{(k)} \right)^2 \right\} \\ &= \sum_{k=1}^{p} \left\{ \left(X_{i_1}^{(k)} - \bar{X}_1^{(k)}\right)^2 - \left(1 + \frac{1}{n_0 - 1} \right) \left(X_{i_1}^{(k)} - \bar{X}_0^{(k)}\right)^2 \right\} \end{aligned}$$

Thus, rather than calculating the class means separately when omitting each of the n observations, we may calculate the overall class means and modify these slightly. Since calculation of the mean is the most computer-intensive stage in fitting the centroid model, this simplification offers substantial performance improvements. In the case of the centroid-based classifier applied to the Leukemia dataset, using this approached reduces the fitting time from 19.9 minutes to only 3.8 minutes. A similar gain can be made in the case of the median method since removing a single observation will only move the overall median up or down half a rank. In this case the computation time reduced from over two hours to 8.3 minutes. Similarly, computation time for the distance method can be reduced from 28.3 to 4.8 minutes. The nearest neighbour methods are not as amenable to this type of optimisation. The remaining methods, despite being somewhat more involved, also admit streamlined approaches to computation.

We also note the advent of relatively cheap and accessible parallel computing, where independent tasks may be distributed across multiple cores of a computer or group thereof. Our recursive method is ideally suited to benefit from this technology, as the task of computing $T(K)$ for each k is easily distributed, allowing further dramatic gains in computational speed.

8.4 Theoretical illustrations

8.4.1 Example where Π_0 and Π_1 differ in terms of a small number of components taking extreme values. We show here that using a classifier that is tuned to the differences between Π_0 and Π_1, and employing the recursive variable selection algorithm given in in Section 8.2.2, can result in particularly accurate identification of those vector components that have greatest leverage for classification. On the other hand, using a conventional approach to variable selection, in particular one based on a linear model, can produce poor results. Therefore, an attempt at classification using variables chosen by a standard method can be quite ineffective.

First we characterise Π_0 and Π_1. A random vector (X, Y) from Π_1, i.e. a vector for which $Y = 1$, is constructed by drawing (X, Y) from Π_0 and replacing r specific components, with indices $k = k_{01}, \ldots, k_{0r}$ say, by random variables all of whose absolute values exceed a given number u_p. We keep the training sample sizes, n_0 and n_1, fixed as p increases, reflecting the high dimension and small sample size of many contemporary problems. The value of u_p is taken to increase with p, and r is held fixed. The vectors $X = (X^{(1)}, \ldots, X^{(p)})$ in the training data are assumed to be independent, but no assumptions are made about dependence among the components of any given X.

The classifier that we shall use reflects characteristics of the data, as would ideally be the case in practice. In particular, a sequence $k_1 < \ldots < k_t$ of distinct integers between 1 and p is chosen empirically using the training data, as suggested in Section 8.2, and a new data vector X, for which the corresponding Y is not known, is classified as type 1 if $|X^{(k_s)}| > v_p$ for $1 \le s \le t$, where $v_p \in (0, u_p)$ is given, and as type 0 otherwise.

In this example the actual construction of the classifier does not depend on the training data; for the sake of simplicity we are assuming that v_p is not a function of the data in $\mathcal{S}_0 \cup \mathcal{S}_1$. Therefore the leave-one-out aspect of the definition (8.1) of estimated error rate can be ignored, and we can define instead:

$$\begin{aligned}
\widehat{\mathrm{err}}(k_1, \ldots, k_t) &= \frac{\pi}{n_0} \sum_{i:\, Y_i = 0} I\{\mathcal{C}(X_i \,|\, k_1, \ldots, k_t) = 1\} \\
&\qquad + \frac{1 - \pi}{n_1} \sum_{i:\, Y_i = 1} I\{\mathcal{C}(X_i \,|\, k_1, \ldots, k_t) = 0\} \\
&= \frac{\pi}{n_0} \sum_{i:\, Y_i = 0} I\big(\big|X_i^{(k_s)}\big| > v_p \text{ for all } s,\, 1 \le s \le t\big) \\
&\qquad + \frac{1 - \pi}{n_1} \sum_{i:\, Y_i = 1} I\big(\big|X_i^{(k_s)}\big| \le v_p \text{ for some } s,\, 1 \le s \le t\big)\,.
\end{aligned}$$

Moreover, in the asymptotic regime considered in Theorem 4.1 below, the probability that there is a tie for the minimising value of k_t, between two indices in the respective sequences $1, \ldots, r$ and $r + 1, \ldots, p$, converges to zero as p diverges. Hence, when stating the theorem there is no need to consider the tie-breaking scheme discussed in Section 8.2.

The theorem shows that, with probability converging to 1 as $p \to \infty$, and provided that p does not increase too rapidly relative to v_p, the recursive variable selector correctly chooses at least a subset of the components where the distributions of the sub-populations Π_0 and Π_1 differ; it does not choose any other components; and it results in zero classification error. We take the prior probability π of the sub-population Π_0 to lie in the interval $(0, 1)$ and to not depend on p. Define $\alpha_p =$

$\max_{k \le p} P(|X^{(k)}| \ge v_p \,|\, Y = 0)$; that is, α_p is the maximum over k of the probability that $|X^{(k)}|$ exceeds v_p when (X, Y) is drawn from Π_0.

Theorem 8.1. *Assume that* $0 < v_p < u_p$, *that* $n_0, n_1 \ge 1$ *are kept fixed as* p *increases, that* p *and* v_p *increase together in such a manner that* $p\,\alpha_p^{n_1} \to 0$ *as* $p \to \infty$, *and that the models for* Π_0 *and* Π_1, *given above, apply. Then with probability converging to 1 as* $p \to \infty$, (*i*) *the algorithm terminates at an integer* $t = \hat{t} \le r$, *and* (*ii*) *the values of* $k_1, \ldots, k_{\hat{t}}$ *chosen by the recursive algorithm prior to termination are all among the special indices* $1, \ldots, r$ *for which the distribution of the vector component differs between the sub-populations* Π_0 *and* Π_1. *Moreover,* (*iii*) *the error rate of the classifier, given by* (8.2) *with* t *replaced by* $\hat{t}$, *converges to 0 as* $p \to \infty$, *and with probability converging to 1 the classifier based on the reduced dimensions* $k_1, \ldots, k_{\hat{t}}$ *gives correct classification for data vectors drawn from either* Π_0 *or* Π_1.

In many of the cases covered by Theorem 4.1, conventional linear variable selection can be expected to perform very poorly. For example, if the components of X, when the data come from Π_0, have a symmetric distribution, then Y is uncorrelated with each component of $X = (X^{(1)}, \ldots, X^{(p)})$. This follows from the fact that the conditional distribution of $X^{(j)}$, given Y, is symmetric. Therefore, a model that depended linearly on the variables would have little opportunity for expressing the influence that any $X^{(j)}$ has on Y. This argument was illustrated in Section 8.3.5 of the numerical work.

8.4.2 Example where Π_0 and Π_1 differ in location. Here we show that even the simple algorithm in Section 8.2.2 can substantially improve the performance of a conventional classifier. We treat the standard centroid-based method, the performance of which can be quite poor when applied to classification problems where the distributions of the two sub-populations differ by a relatively large amount in only a small number of components, rather than by a relatively small amount in a large number of components. The latter context is often referred to as having low sparsity, since information is available in a relatively high proportion of components. Although the centroid-based approach has optimality properties, it demonstrates them only when the degree of sparsity is low, not (as in the examples we give here) in the case of high sparsity; and it shares this feature with related methods, such as the support vector machine. Theorem 4.2, below, shows that recursive variable selection adapts well to high-sparsity settings, ensuring relatively good performance there. A similar result can be proved in the case of support vector machine classifiers (see Hall and Pham, 2010).

We assume that a random vector $X = (X^{(1)}, \ldots, X^{(p)})$, when (X, Y) is drawn from Π_1, is constructed by taking the vector from Π_0 and then adding the constant μ to r specific components of X, in particular those with indices $k_{01}, \ldots, k_{0r}$. The

algorithm in Section 8.2.2 is used to construct the variable selector.

The theorem below shows that, with probability converging to 1 as $p \to \infty$, the recursive classifier correctly assigns a new data value to either Π_0 or Π_1, provided that $|\mu|$ is of larger order then $\log p$; and that, on the other hand, the standard centroid-based classifier fails to give correct classification unless $|\mu|$ is at least as large as $p^{1/4}$. This establishes the extent to which the recursive algorithm can improve performance of the classifier in cases where information is sparse. The theorem also demonstrates that, with probability converging to 1 as $p \to \infty$, the recursive approach correctly chooses at least a subset of the components where the distributions of the sub-populations Π_0 and Π_1 differ, and does not choose any other components.

We suppose that that, for some $c > 0$,

$$\sup_{p \geq 1} \max_{1 \leq k \leq p} E\bigl\{ \exp\bigl(c \, |X^{(k)}|\bigr) \, \big| \, Y = 0 \bigr\} < \infty \,. \tag{8.8}$$

That is, we ask that the component-wise moment generating functions of X, when (X, Y) is drawn from Π_0, be uniformly bounded in some neighbourhood of the origin.

Theorem 8.2. *Assume that* (8.8) *holds, that* $n_0, n_1 \geq 2$, *that* $|\mu| / \log p \to \infty$, *and that the above models for* Π_0 *and* Π_1 *apply. Then with probability converging to 1 as* $p \to \infty$, *results* (*i*) *and* (*ii*) *from Theorem 4.1 hold. Moreover, result* (*iii*) *from that theorem obtains, and with probability converging to 1 the classifier based on the reduced dimensions* $k_1, \ldots, k_{\hat{t}}$ *gives correct classification for data vectors drawn from either* Π_0 *or* Π_1. *Also,* (*iv*) *if the components of* X *when* (X, Y) *is drawn from* Π_0 *are independent and identically distributed; if we employ the standard centroid-based classifier, in which the sign of the statistic at* (8.4) *is used to assign* X *to* Π_0 *or* Π_1, *without any dimension reduction; and if* $n_0 = n_1$ *and* $|\mu| / p^{1/4} \to 0$ *as* $p \to \infty$; *then the probability of correct classification converges to* $\frac{1}{2}$ *as* $p \to \infty$.

The requirement in Theorem 8.2 that $n_0 \geq 2$ and $n_1 \geq 2$ ensures that the leave-one-out approach to estimating error rate is feasible. The assumption of independence in part (iv) of the theorem makes the classification problem relatively difficult, since then the noise can differ markedly from one vector component to another. On the other hand the condition $n_0 = n_1$, also in part (iv), actually gives a result that is relatively favourable to the standard centroid-based classifier. It permits a critical cancellation at one point in the argument. Without the condition $n_0 = n_1$, and in the case of fixed n_0 and n_1, the value of $|\mu|$ generally has to be as large as $p^{1/2}$, not $p^{1/4}$, before the centroid-based classifier can distinguish between Π_0 and Π_1.

8.5 Technical arguments

8.5.1 Proof of Theorem 8.1. In the next paragraph we treat the problem of empirical choice of the first component index, k_1, using the algorithm in Section 8.2. Without loss of generality the special components $k_{01}, \ldots, k_{0r}$ are just $1, \ldots, r$. We show that:

$$\text{the probability that } k_1 \in \{1, \ldots, r\}, \text{ and that no data in } \mathcal{S}_0 \cup \mathcal{S}_1 \text{ are misclassified by } \mathcal{C}(\cdot \,|\, k_1), \text{ converges to 1 as } n \to \infty. \tag{8.9}$$

In the subsequent paragraph we note that the same argument extends to choices of other component indices.

If $k_1 \in \{1, \ldots, r\}$ is fixed then, with probability 1, each data pair (X_i, Y_i) in $\mathcal{S}_1$ is correctly classified by the classifier $\mathcal{C}(\cdot \,|\, k_1)$, and with probability converging to 1, each data pair in $\mathcal{S}_0$ is correctly classified. From this property, and the fact that r is fixed, we deduce that the probability of a misclassification of one or more of the training data converges to zero uniformly in choices $k_1 \in \{1, \ldots, r\}$. Next we consider the case where $k_1 \in \{r+1, \ldots, p\}$. We can achieve zero misclassification of data in $\mathcal{S}_1$ by choosing $k_1 \in \{r+1, \ldots, p\}$, if and only if, for some k in the range $r+1 \le k \le p$, the event $\mathcal{E}_k$ that $\inf_{i:\, Y_i = 1} \bigl|X_i^{(k)}\bigr| > v_p$ holds. Recall that $\alpha_p = \max_{k \le p} P(|X^{(k)}| \ge v_p \,|\, Y = 0)$. Then, $P(\mathcal{E}_k) = \alpha_p^{n_1}$ for each such k, and so

$$P\Bigl(\bigcup_{k=r+1}^{p} \mathcal{E}_k\Bigr) \le \sum_{k=r+1}^{p} P(\mathcal{E}_k) = p\,\alpha_p^{n_1} = o(1)\,,$$

where we used an assumption in the Theorem statement to obtain the final identity. Therefore (8.9) holds.

Next we extend (8.9) to general sequences $k_1, \ldots, k_{t+1}$. Let $t \in [1, r]$ denote a fixed integer, and let $\mathcal{A}$ denote the class of all subsets $A = \{k_1, \ldots, k_{t+1}\}$ of distinct elements of $\{1, \ldots, p\}$ of which k_{t+1} is the only value exceeding r. The number of elements of $\mathcal{A}$ is at most $2^r\, p$, and so the argument leading to (8.9) implies that

$$P\Bigl(\text{for some } A \in \mathcal{A}\,, \quad \inf_{i:\, Y_i = 1} \bigl|X_i^{(k)}\bigr| > v_p \quad \text{for all} \quad k \in A\Bigr) \le 2^r\, p\, \alpha_p^{n_1} = o(1)\,.$$

Therefore, by induction from (8.9), if the sequence $k_1, \ldots, k_t$ contains only numbers between 1 and r, then the probability that a classifier that results from adjoining some index k_{t+1} between $r+1$ and p, and confining attention to vector coordinates with indices in the set $k_1, \ldots, k_{t+1}$, misclassifies at least one data value in $\mathcal{S}_1$, converges to 1 as $p \to \infty$. This result implies properties (i)–(iii) in the theorem.

8.5.2 Proof of Theorem 8.2. Again we may assume that $\{k_{01}, \ldots, k_{0r}\} = \{1, \ldots, r\}$. First we prove that (8.9) holds if $|\mu| / \log p \to \infty$. Since notation becomes quite complex if we address specifically the leave-one-out setting, then we

shall initially treat the general case where we have training samples of sizes n_0 and n_1 and use them to classify a new data value, X. Then we shall specialise this result to its counterpart in leave-one-out settings.

Consider the case where (X, Y) is drawn from Π_0. Then (8.8) holds, and by Markov's inequality,

$$\begin{aligned} P\big(\big|X^{(k)}\big| > \log p\big) &= P\big(\big|X^{(k)}\big| > \log p \,\big|\, Y = 0\big) \\ &\le \exp(-c \log p \,\big|\, Y = 0\big)\, E\big\{\exp\big(c\,\big|X^{(k)}\big|\big) \,\big|\, Y = 0\big\} = O\big(p^{-c}\big)\,. \end{aligned}$$

Therefore, if $r+1 \le k \le p$ then, no matter whether $X = (X^{(1)}, \ldots, X^{(p)})$ is from Π_0 or Π_1, and for $j = 0, 1$,

$$\begin{aligned} P\big(\big|X^{(k)} - \bar{X}_j^{(k)}\big| > 2\,\log p\big) &\le P\big(\big|X^{(k)}\big| > \log p\big) + P\big(\big|\bar{X}_j^{(k)}\big| > \log p\big) \\ &\le (n_j + 1)\, P\big(\big|X^{(k)}\big| > \log p\big) = O\big(p^{-c}\big)\,, \end{aligned}$$

uniformly in $r+1 \le k \le p$. Hence, defining $\Delta_k = (X^{(k)} - \bar{X}_1^{(k)})^2 - (X^{(k)} - \bar{X}_0^{(k)})^2$, we deduce that

$$P\big\{|\Delta_k| > 8\,(\log p)^2\big\} \le \sum_{j=1}^{2} P\big(\big|X^{(k)} - \bar{X}_j^{(k)}\big| > 2\,\log p\big) = O\big(p^{-c}\big)\,,$$

uniformly in $r+1 \le k \le p$. Therefore,

$$P\Big\{\max_{r+1 \le k \le p} |\Delta_k| > 8\,(\log p)^2\Big\} = O\big(p^{1-c}\big) = o(1)\,. \tag{8.10}$$

If $1 \le k \le r$ and $|\mu| > c_1 \log p$, where c_1 is arbitrarily large but fixed, and if $0 < c_2 < c_3 < c_1$, then

$$\begin{aligned} P_0\big(\big|X^{(k)} - \bar{X}_0^{(k)}\big| > c_2\,\log p\big) &\le P_0\big(\big|X^{(1)}\big| + \big|\bar{X}_0^{(1)}\big| > c_2\,\log p\big) \to 0\,, \\ P_0\big(\big|X^{(k)} - \bar{X}_1^{(k)}\big| > c_3\,\log p\big) &\ge P_0\big(|\mu| - \big|X^{(k)}\big| - \big|\bar{X}_1^{(k)} - \mu\big| > c_3\,\log p\big) \\ &\ge P_0\big\{\big|X^{(1)}\big| + \big|\bar{X}_1^{(1)} - \mu\big| < (c_1 - c_3)\,\log p\big\} \to 1\,, \end{aligned}$$

where, here and below, P_j denotes probability measure under the assumption that X comes from Π_j. It follows that, for $1 \le k \le r$, $P_0\{\Delta_k > (c_3^2 - c_2^2)\,(\log p)^2\} \to 1$. From this result, and its counterpart when X is drawn from Π_1, we deduce that if $1 \le k \le r$ then for each $c_4 > 0$,

$$P_0\big\{\Delta_k > c_4\,(\log p)^2\big\} \to 1\,, \quad P_1\big\{\Delta_k < -c_4\,(\log p)^2\big\} \to 1\,. \tag{8.11}$$

Combining (8.10) and (8.11) we see that, with probability converging to 1 as $p \to \infty$, and for each $C > 0$, if X is from Π_0 then $\inf_{1 \le k \le r}\, \Delta_k - \max_{r+1 \le k \le p}\, |\Delta_k| > C$,

and if X is from Π_1 then $\inf_{1 \le k \le r} (-\Delta_k) - \max_{r+1 \le k \le p} |\Delta_k| > C$. (Here we have used the fact that r is fixed.) Therefore the least value taken by $|S(X)|$ when the classifier is confined to an index $k \in \{1, \ldots, r\}$, divided by the largest value taken by $|S(X)|$ when $k \in \{r+1, \ldots, p\}$, diverges to infinity in probability as $p \to \infty$; and moreover, with probability converging to 1, if X is from Π_j then $(-1)^j \, S(X) > 0$ for each $k \in \{1, \ldots, r\}$.

Since these results hold for each choice of n_0 and n_1 then they immediately translate to the case of the leave-one-out classifier, for which the corresponding values of n_0 and n_1 can be $n_0 - 1$ or $n_1 - 1$ but are never less than 1. Since n_0 and n_1 are kept fixed as p increases then it follows that, with probability converging to 1 as $p \to \infty$, the minimum value of the leave-one out versions of $|S(X)|$ for all different choices of the omitted data value, and whenever the classifier is confined to an index $k \in \{1, \ldots, r\}$, divided by the largest value taken by leave-one out versions of $|S(X)|$ over all values of the omitted value, and all values of $k \in \{r+1, \ldots, p\}$, diverges to infinity in probability as $p \to \infty$; and moreover, with probability converging to 1, $(-1)^j \, S(X) > 0$ in all all cases where the omitted training data value X is from Π_j and for each $k \in \{1, \ldots, r\}$. These results establish (8.9), and as in the proof of Theorem 4.1, a similar argument can be used to give properties (i)–(iii) in Theorem 4.2.

To establish part (iv) of the theorem, note that $S(X) = S_1(X) + S_2(X)$ where $S_1(X) = \sum_{1 \le k \le r} \Delta_k$ and $S_2(X) = \sum_{r+1 \le k \le \infty} \Delta_k$. For all $k \ge r+1$, $E(\Delta_k) = 0$ (here we used the fact that $n_0 = n_1$), and so, since $X^{(k)}$ (when (X, Y) is from Π_0) has four finite moments (by virtue of the assumption that it has finite moment generating function), S_2 is asymptotically normal $\mathrm{N}(0, p\,\sigma^2)$ where $0 < \sigma < \infty$. A simpler argument shows that $S_1(X) = O_p(\mu^2) = o_p(p^{1/2})$ as $p \to \infty$. (Here we used the fact that $|\mu| = o(p^{1/4})$.) Both these results hold regardless of whether X comes from Π_0 or Π_1. Therefore, $S(X)$ is asymptotically normal $\mathrm{N}(0, p\,\sigma^2)$, regardless of whether X comes from Π_0 or Π_1. It follows that the probability that the classifier assigns X to the wrong population converges to $\frac{1}{2}$ as $p \to \infty$.

Bibliography

Alon, U., Barkai, N., Notterman, D., Gish, K., Ybarra, S., Mack, D. and Levine, A. (1999). Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc. Natl. Acad. Sci. USA*, **96** 6745–6750.

Amosova, N. (1972). On limit theorems for probabilities of moderate deviations. *Vestnik Leningrad. Univ*, **13** 5–14.

Andrews, D. W. K. (1999). Estimation when a parameter is on a boundary. *Econometrica*, **67** 1341–1383.

Andrews, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, **68** 399–405.

Asimov, D. (1985). The grand tour: a tool for viewing multidimensional data. *SIAM J. Sci. Statist. Comput.*, **6** 128–143.

Barker, L., Smith, P., Gerzoff, R., Luman, E., McCauley, M. and Strine, T. (2005). Ranking states' immunization coverage: an example from the National Immunization Survey. *Stat. Med.*, **24** 605–613.

Becker, R., Cleveland, W. and Shyu, M. (1996). The visual design and control of trellis display. *J. Comput. Graph. Statist.*, **5** 123–155.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)* 289–300.

Beran, R. (1982). Estimated sampling distributions: the bootstrap and competitors. *Ann. Statist.*, **10** 212–225.

Beran, R. and Srivastava, M. S. (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. *Ann. Statist.*, **13** 95–115.

Berger, A. and Humphrey, D. (1997). Efficiency of financial institutions: International survey and directions for future research. *European J. Operational Res.*, **98** 175–212.

Bertin, K. and Lecué, G. (2008). Selection of variables and dimension reduction in high-dimensional non-parametric regression. *Electron. J. Stat.*, **2** 1224–1241.

Bhattacharya, R. N. and Ghosh, J. K. (1978). On the validity of the formal Edgeworth expansion. *Ann. Statist.*, **6** 434–451.

Bhattacharya, R. N. and Ranga Rao, R. (1976). *Normal approximation and asymptotic expansions*. John Wiley & Sons, New York-London-Sydney. Wiley Series in Probability and Mathematical Statistics.

Bickel, P. J., Götze, F. and van Zwet, W. R. (1997). Resampling fewer than n observations: gains, losses, and remedies for losses. *Statist. Sinica*, **7** 1–31.

Bickel, P. J. and Ren, J.-J. (1996). The m out of n bootstrap and goodness of fit tests with double censored data. In *Robust statistics, data analysis, and computer intensive methods (Schloss Thurnau, 1994)*, vol. 109 of *Lecture Notes in Statist.* Springer, New York, 35–47.

Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.*, **37** 1705–1732.

Bilmes, J. A. and Kirchhoff, K. (2003). Generalized rules for combination and joint training of classifiers. *PAA Pattern Anal. Appl.*, **6** 201–211.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics*, **37** 373–384.

Breiman, L. (2001a). Random forests. *Mach. Learn.*, **45** 5–32.

Breiman, L. (2001b). Statistical modeling: the two cultures. *Statist. Sci.*, **16** 199–231. With comments and a rejoinder by the author.

Bretagnolle, J. (1983). Lois limites du bootstrap de certaines fonctionnelles. *Ann. Inst. H. Poincaré Sect. B (N.S.)*, **19** 281–296.

Brijs, T., Karlis, D., Van den Bossche, F. and Wets, G. (2007). A Bayesian model for ranking hazardous road sites. *J. Roy. Statist. Soc. Ser. A*, **170** 1001–1017.

Brijs, T., Van den Bossche, F., Wets, G. and Karlis, D. (2006). A model for identifying and ranking dangerous accident locations: a case study in Flanders. *Statist. Neerlandica*, **60** 457–476.

Buhlmann, P. (2006). Boosting for high-dimensional linear models. *Annals of Statistics*, **34** 559–583.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, **35** 2313–2351.

Cesário, L. C. and Barreto, M. C. M. (2003). Study of the performance of bootstrap confidence intervals for the mean of a normal distribution using perfectly ranked set sampling. *Rev. Mat. Estatíst.*, **21** 7–20.

Cestnik, G., Konenenko, I. and Bratko, I. (1987). Assistant-86: A knowledge-elicitation tool for sophisticated users.

Chambers, J. and Hastie, T. (1992). *Statistical Models in S*. Wadsworth/CRC, Pacific Grove, CA.

Chen, H., Stasny, E. A. and Wolfe, D. A. (2006). An empirical assessment of ranking accuracy in ranked set sampling. *Comput. Statist. Data Anal.*, **51** 1411–1419.

Chen, S. S., Donoho, D. L. and Saunders, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.*, **20** 33–61 (electronic).

Clarke, R., Ressom, H., Wang, A., Xuan, J., Liu, M., Gehan, E. and Wang, Y. (2008). The properties of high-dimensional data spaces: implications for exploring gene and protein expression data. *Nature Reviews Cancer*, **8** 37–49.

Cootes, T., Hill, A., Taylor, C. and Haslam, J. (1994). The use of active shape models for locating structures in medical images. *Image and vision computing*, **12** 355–366.

Corain, L. and Salmaso, L. (2007). A non-parametric method for defining a global preference ranking of industrial products. *J. Appl. Stat.*, **34** 203–216.

Csörgő, S. and Hall, P. (1982). Estimable versions of Griffiths' measure of association. *Austral. J. Statist.*, **24** 296–308.

Dabney, A. (2005). Classification of microarrays to nearest centroids. *Bioinformatics*, **21** 4148–4154.

Dabney, A. and Storey, J. (2005). Optimal feature selection for nearest centroid classifiers, with applications to gene expression microarrays. *UW Biostatistics Working Paper Series* 267.

Dabney, A. and Storey, J. (2007). Optimality driven nearest centroid classification from genomic data. *PLoS One*, **2** (electronic).

Davison, A. and Hinkley, D. (1997). *Bootstrap methods and their application*. Cambridge Univ Pr.

De Boor, C. (2001). *A practical guide to splines*. Springer Verlag.

Dempster, A. (1972). Covariance selection. *Biometrics*, **28** 157–175.

Dettling, M. (2004). BagBoosting for tumor classification with gene expression data. *Bioinformatics*, **20** 3583–3593.

Diaconis, P. and Efron, B. (1983). Computer-intensive methods in statistics. *Scientific American*, **248** 116–130.

Dobson, A. (2001). *An introduction to generalized linear models*. CRC Pr I Llc.

Donoho, D. L. (2006a). For most large underdetermined systems of equations, the minimal l_1-norm near-solution approximates the sparsest near-solution. *Comm. Pure Appl. Math.*, **59** 907–934.

Donoho, D. L. (2006b). For most large underdetermined systems of linear equations the minimal l_1-norm solution is also the sparsest solution. *Comm. Pure Appl. Math.*, **59** 797–829.

Donoho, D. L. and Elad, M. (2003). Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization. *Proc. Natl. Acad. Sci. USA*, **100** 2197–2202 (electronic).

Donoho, D. L. and Huo, X. (2001). Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inform. Theory*, **47** 2845–2862.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia? *J. Roy. Statist. Soc. Ser. B*, **57** 301–369. With discussion and a reply by the authors.

Duda, R. O., Hart, P. E. and Stork, D. G. (2001). *Pattern classification*. 2nd ed. Wiley-Interscience, New York.

Dudoit, S., Fridlyand, J. and Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. *J. Amer. Statist. Assoc.*, **97** 77–87.

Dümbgen, L. (1993). On nondifferentiable functions and the bootstrap. *Probab. Theory Related Fields*, **95** 125–140.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.*, **32** 407–499. With discussion, and a rejoinder by the authors.

Efron, B. and Tibshirani, R. (1997). *An introduction to the bootstrap*. Chapman & Hall.

Fan, J. and Fan, Y. (2008). High-dimensional classification using features annealed independence rules. *Ann. Statist.*, **36** 2605–2637.

Fan, J. and Gijbels, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Roy. Statist. Soc. Ser. B*, **57** 371–394.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96** 1348–1360.

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **70** 849–911.

Farcomeni, A. (2008). A review of modern multiple hypothesis testing, with particular attention to the false discovery proportion. *Statistical Methods in Medical Research*, **17** 347–388.

Franco-Lopez, H., Ek, A. and Bauer, M. (2001). Estimation and mapping of forest stand density, volume, and cover type using the k-nearest neighbors method. *Remote Sensing of Environment*, **77** 251–274.

FRIEDMAN, J. and TUKEY, J. (1974). A Projection Pursuit Algorithm for Exploratory Data Analysis. *IEEE Trans. Comput.*, **100** 881–890.

FUCHS, J. (2005). Recovery of exact sparse representations in the presence of bounded noise. *IEEE Trans. Inform. Theory*, **51** 3601–3608.

GAO, H.-Y. (1998). Wavelet shrinkage denoising using the non-negative garrote. *J. Comput. Graph. Statist.*, **7** 469–488.

GOLDSTEIN, D. (2009). Common genetic variation and human traits. *New England J. Med.*, **360** 1696–1698.

GOLDSTEIN, H. and SPIEGELHALTER, D. (1996). League tables and their limitations: statistical issues in comparisons of institutional performance. *J. Roy. Statist. Soc. Ser. A*, **159** 385–443.

GOLUB, T., SLONIM, D., TAMAYO, P., HUARD, C., GAASENBEEK, M., MESIROV, J., COLLER, H., LOH, M., DOWNING, J., CALIGIURI, M. ET AL. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286** 531–537.

GRIFFITHS, R. C. (1972). Linear dependence in bivariate distributions. *Austral. J. Statist.*, **14** 182–187.

GRINDEA, S. and POSTELNICU, V. (1977). Some measures of association. In *Proceedings of the Fifth Conference on Probability Theory (Braşov, 1974)*. Editura Acad. R.S.R., Bucharest., 197–203.

HALL, P. (1990). Using the bootstrap to estimate mean squared error and select smoothing parameter in nonparametric problems. *J. Multivariate Anal.*, **32** 177–203.

HALL, P. (1992). *The bootstrap and Edgeworth expansion*. Springer Series in Statistics, Springer-Verlag, New York.

HALL, P., HÄRDLE, W. and SIMAR, L. (1993). On the inconsistency of bootstrap distribution estimators. *Comput. Statist. Data Anal.*, **16** 11–18.

HALL, P., JIN, J. and MILLER, H. (2010). Feature selection when there are many influential features. *Manuscript*.

HALL, P. and MILLER, H. (2009a). Using generalized correlation to effect variable selection in very high dimensional problems. *J. Comput. Graph. Statist.*, **18** 533–550.

HALL, P. and MILLER, H. (2009b). Using the bootstrap to quantify the authority of an empirical ranking. *Ann. Stat.*, **37** 3929–3959.

HALL, P. and MILLER, H. (2010a). Bootstrap confidence intervals and hypothesis tests for extrema of parameters. *Biometrika, to appear*.

HALL, P. and MILLER, H. (2010b). Determining and depicting relationships among components in high-dimensional variable selection. *Manuscript*.

Hall, P. and Miller, H. (2010c). Modelling the variability of rankings. *Ann. Statist., to appear.*

Hall, P. and Miller, H. (2010d). Sequential, bottom-up variable selection for high dimensional classification. *Manuscript.*

Hall, P. and Pham, T. (2010). Optimal properties of centroid-based classifiers for very high-dimensional data. *Ann. Statist.*, **38** 1071–1093.

Hall, P., Racine, J. and Li, Q. (2004). Cross-validation and the estimation of conditional probability densities. *J. Amer. Statist. Assoc.*, **99** 1015–1026.

Hall, P., Titterington, D. and Xue, J. (2009). Tilting methods for assessing the influence of components in a classifier. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **71** 783–803.

Hastie, T., Tibshirani, R. and Friedman, J. (2001). *The elements of statistical learning.* Springer Series in Statistics, Springer-Verlag, New York. Data mining, inference, and prediction.

Hill, B. M. (1975). A simple general approach to inference about the tail of a distribution. *Ann. Statist.*, **3** 1163–1174.

Hirschhorn, J. (2009). Genomewide association studies–illuminating biologic pathways. *New England J. Med.*, **360** 1699–1701.

Hoerl, A. and Kennard, R. (1970). Ridge regression: applications to nonorthogonal problems. *Technometrics*, **12** 69–82.

Hosmer, D. and Lemeshow, S. (2000). *Applied logistic regression.* Wiley-Interscience, New York.

Hua, J., Tembe, W. and Dougherty, E. (2009). Performance of feature-selection methods in the classification of high-dimension data. *Pattern Recognition*, **42** 409–424.

Hui, T. P., Modarres, R. and Zheng, G. (2005). Bootstrap confidence interval estimation of mean via ranked set sampling linear regression. *J. Stat. Comput. Simul.*, **75** 543–553.

Ibragimov, I. A. and Linnik, Y. V. (1971). *Independent and stationary sequences of random variables.* Wolters-Noordhoff Publishing, Groningen. With a supplementary chapter by I. A. Ibragimov and V. V. Petrov, Translation from the Russian edited by J. F. C. Kingman.

Inglot, T., Kallenberg, W. C. M. and Ledwina, T. (1992). Strong moderate deviation theorems. *Ann. Probab.*, **20** 987–1003.

Inza, I., Larrañaga, P., Blanco, R. and Cerrolaza, A. (2004). Filter versus wrapper gene selection approaches in DNA microarray domains. *Artif. Intell. Med.*, **31** 91–103.

Joe, H. (2000). Inequalities for random utility models, with applications to ranking and subset choice data. *Methodol. Comput. Appl. Probab.*, **2** 359–372.

Joe, H. (2001). Multivariate extreme value distributions and coverage of ranking probabilities. *J. Math. Psych.*, **45** 180–188.

Kim, M., Kim, Y. and Schmidt, P. (2007). On the accuracy of bootstrap confidence intervals for efficiency levels in stochastic frontier models with panel data. *J. Productivity Anal.*, **28** 165–181.

Kraft, P. and Hunter, D. (2009). Genetic Risk Prediction–Are We There Yet? *New England J. Med.*, **360** 1701–1703.

Lafferty, J. and Wasserman, L. (2008). Rodeo: sparse, greedy nonparametric regression. *Ann. Statist.*, **36** 28–63.

Langford, I. H. and Leyland, A. H. (1996). Discussion of Goldstein and Spiegelhalter. *J. Roy. Statist. Soc. Ser. A*, **159** 427–428.

Larocque, D. and Léger, C. (1994). Bootstrap estimates of the power of a rank test in a randomized block design. *Statist. Sinica*, **4** 423–443.

Lin, Y. and Zhang, H. (2006). Component selection and smoothing in smoothing spline analysis of variance models. *Ann. Statist.*, **34** 2272–2297.

Loader, C. (1999). *Local regression and likelihood*. Statistics and Computing, Springer-Verlag, New York.

Mammen, E. (1992). *When does bootstrap work?: asymptotic results and simulations*, vol. 77 of *Statistics and Computing*. Springer-Verlag New York.

Mardia, K., Kent, J., Bibby, J. et al. (1979). *Multivariate analysis*. Academic Press London.

Masry, E. (1996). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *J. Time Series Anal.*, **17** 571–600.

McCullagh, P. and Nelder, J. (1989). *Generalized linear models*. Chapman & Hall/CRC.

McHale, I. and Scarf, P. (2005). Ranking football players. *Significance*, **2** 54–57.

Mease, D. (2003). A penalized maximum likelihood approach for the ranking of college football teams independent of victory margins. *Amer. Statist.*, **57** 241–248.

Meier, L., van de Geer, S. and Bühlmann, P. (2008). The group Lasso for logistic regression. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **70** 53–71.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, **34** 1436–1462.

Meinshausen, N., Rocha, G. and Yu, B. (2007). A tale of three cousins: Lasso, L_2 Boosting and Dantzig. Discussion of Candes and Tao (2007). *Ann. Statist*, **35** 2373–2384.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, **37** 246–270.

Miller, H., Clarke, S., Lane, S., Lonie, A., Lazaridis, D., Petrovski, S. and Jones, O. (2009). Predicting customer behaviour: The university of melbourne's kdd cup report. *Proceedings of the KDD Cup, to appear*.

Miller, H. and Hall, P. (2010). Local polynomial regression and variable selection. *Manuscript*.

Moon, H., Ahn, H., Kodell, R., Lin, C., Baek, S. and Chen, J. (2006). Classification methods for the development of genomic signatures from high-dimensional data. *Genome Biology*, **7** R121.1–R121.7.

Mukherjee, S., Roberts, S., Sykacek, P. and Gurr, S. (2003). Gene ranking using bootstrapped P-values. *ACM SIGKDD Explorations Newsletter*, **5** 16–22.

Murphy, T. B. and Martin, D. (2003). Mixtures of distance-based models for ranking data. *Comput. Statist. Data Anal.*, **41** 645–655.

Nadaraya, E. (1964). On estimating regression. *Theor. Probab. Appl.*, **9** 141–142.

Ng, A. (1998). On feature selection: learning with exponentially many irrelevant features as training examples. In *Proceedings of the Fifteenth International Conference on Machine Learning*. Citeseer, 404–412.

Nordberg, L. (2006). On the reliability of performance rankings. In *Festschrift for Tarmo Pukkila on his 60th birthday*. Dep. Math. Stat. Philos. Univ. Tampere, Tampere, 205–216.

Opgen-Rhein, R. and Strimmer, K. (2007). Accurate ranking of differentially expressed genes by a distribution-free shrinkage approach. *Stat. Appl. Genet. Mol. Biol.*, **6** Art. 9, 20 pp. (electronic).

Pelin, P., Brcich, R. and Zoubir, A. (2000). A bootstrap technique for rank estimation. In *Statistical Signal and Array Processing, 2000. Proceedings of the Tenth IEEE Workshop on*. 94–98.

Peng, J., Wang, P., Zhou, N. and Zhu, J. (2009). Partial Correlation Estimation by Joint Sparse Regression Models. *J. Amer. Statist. Assoc.*, **104** 735–746.

Politis, D. N., Romano, J. P. and Wolf, M. (1999). *Subsampling*. Springer Series in Statistics, Springer-Verlag, New York.

Quevedo, J. R., Bahamonde, A. and Luaces, O. (2007). A simple and efficient method for variable ranking according to their usefulness for learning. *Comput. Statist. Data Anal.*, **52** 578–595.

Rényi, A. (1953). On the theory of order statistics. *Acta Math. Acad. Sci. Hungar.*, **4** 191–231.

Ringrose, T. and Benn, D. (1997). Confidence regions for fabric shape diagrams. *J. Structural Geol.*, **19** 1527–1536.

Rubin, H. and Sethuraman, J. (1965). Probabilities of moderate deviations. *Sankhyā: The Indian Journal of Statistics, Series A*, **27** 325–346.

Ruppert, D. and Wand, M. P. (1994). Multivariate locally weighted least squares regression. *Ann. Statist.*, **22** 1346–1370.

Saeys, Y., Inza, I. and Larranaga, P. (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23** 2507–2517.

Schechtman, E. and Yitzhaki, S. (1987). A measure of association based on Gini's mean difference. *Comm. Statist. Theory Methods*, **16** 207–231.

Schoonover, J., Marx, R. and Zhang, S. (2003). Multivariate curve resolution in the analysis of vibrational spectroscopy data files. *Applied spectroscopy*, **57** 154A–170A.

Schott, J. (2006). A high-dimensional test for the equality of the smallest eigenvalues of a covariance matrix. *J. Multivariate Anal.*, **97** 827–843.

Segal, M., Dahlquist, K. and Conklin, B. (2003). Regression approaches for microarray data analysis. *J. Comput. Biol.*, **10** 961–980.

Shakhnarovich, G., Darrell, T. and Indyk, P. (2005). *Nearest-neighbor methods in learning and vision: Theory and practice*. The MIT Press, Cambridge, Mass.

Simonoff, J. S. (1996). *Smoothing methods in statistics*. Springer Series in Statistics, Springer-Verlag, New York.

Srivastava, M. S. (1987). Bootstrap method in ranking and slippage problems. *Comm. Statist. Theory Methods*, **16** 3285–3299.

Steland, A. (1998). Bootstrapping rank statistics. *Metrika*, **47** 251–264.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. Roy. Statist. Soc. Ser. B*, **36** 111–147. With discussion by G. A. Barnard, A. C. Atkinson, L. K. Chan, A. P. Dawid, F. Downton, J. Dickey, A. G. Baker, O. Barndorff-Nielsen, D. R. Cox, S. Giesser, D. Hinkley, R. R. Hocking, and A. S. Young, and with a reply by the authors.

Street, W., Wolberg, W. and Mangasarian, O. (1993). Nuclear feature extraction for breast tumor diagnosis. In *IS&T/SPIE 1993 International Symposium on Electronic Imaging: Science and Technology*, vol. 1905. Citeseer, 861–870.

Swanepoel, J. (1986). A note on proving that the(modified) bootstrap works. *Comm. Statist. A—Theory Methods*, **15** 3193–3203.

Taconeli, C. A. and Barreto, M. C. M. (2005). Evaluation of a bootstrap confidence interval approach in perfectly ranked set sampling. *Rev. Mat. Estatíst.*, **23** 33–53.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B*, **58** 267–288.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proc. Natl. Acad. Sci. USA*, **99** 6567–6572.

Tropp, J. (2004). Greed is good: Algorithmic results for sparse approximation. *IEEE Trans. Inform. Theory*, **50** 2231–2242.

Tropp, J. A. (2005). Recovery of short, complex linear combinations via l_1 minimization. *IEEE Trans. Inform. Theory*, **51** 1568–1570.

Tu, X. M., Burdick, D. S. and Mitchell, B. C. (1992). Nonparametric rank estimation using bootstrap resampling and canonical correlation analysis. In *Exploring the limits of bootstrap (East Lansing, MI, 1990)*. Wiley Ser. Probab. Math. Statist. Probab. Math. Statist., Wiley, New York, 405–418.

Wahba, G. (1990). *Spline models for observational data*. SIAM.

Wand, M. P. and Jones, M. C. (1995). *Kernel smoothing*, vol. 60 of *Monographs on Statistics and Applied Probability*. Chapman and Hall Ltd., London.

Wang, L., Zhu, J. and Zou, H. (2006). The doubly regularized support vector machine. *Statist. Sinica*, **16** 589–615.

Wang, L., Zhu, J. and Zou, H. (2008). Hybrid huberized support vector machines for microarray classification and gene selection. *Bioinformatics*, **24** 412–419.

Wang, S. and Zhu, J. (2007). Improved centroids estimation for the nearest shrunken centroid classifier. *Bioinformatics*, **23** 972–979.

Ward, M., LeBlanc, J. and Tipnis, R. (1994). N-land: a graphical tool for exploring n-dimensional data. In *Proc. Computer Graphics International Conference*. Citeseer, Melbourne, Australia.

Wasserman, L. and Roeder, K. (2009). High dimensional variable selection. *Ann. Statist.*, **37** 2178–2201.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A*, **26** 359–372.

Wolberg, W. and Mangasarian, O. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proc. Natl. Acad. Sci. USA*, **87** 9193–9196.

Xie, M., Singh, K. and Zhang, C. (2009). Confidence Intervals for Population Ranks in the Presence of Ties and Near Ties. *J. Amer. Statist. Assoc.*, **104** 775–788.

Xiong, M., Fang, X. and Zhao, J. (2001). Biomarker identification by feature wrappers. *Genome Res.*, **11** 1878–1887.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **68** 49–67.

Zhao, P. and Yu, B. (2007). Stagewise lasso. *J. Mach. Learn. Res.*, **8** 2701–2726.

Zhu, J., Rosset, S., Hastie, T. and Tibshirani, R. (2004). 1-norm support vector machines. In *Advances in Neural Information Processing Systems* (S. Thrun, L. Saul and B. Schölkopf, eds.), vol. 16. The MIT Press, Boston, 49–56.

Zou, H. (2006). The adaptive lasso and its oracle properties. *J. Amer. Statist. Assoc.*, **101** 1418–1429.